DETECTION OF PHISHING WEBSITES USING MACHINE LEARNING

Vedavyas J[1], Bhupathiraju Deepthi[2], Harini Hardageri[3], G Veda Samhitha[4], H Niveditha[5]

[1]Assistant Professor, Department of CSE, Ballari Institute of Technology & Management, Ballari
[2,3,4,5]Final Year Students, Department of CSE, Ballari Institute of Technology & Management, Ballari

---***---

Abstract - *Phishing is a technique widely employed to deceive unsuspecting people into exposing their personal information by means of fake websites. Phishing website URLs are made with the intention of collecting user data, such as usernames, passwords, and details of online financial activities. Phishers use websites that are semantically and visually identical to those authentic websites. By using anti-phishing technologies to recognize phishing, we may inhibit the rapid evolution of phishing strategies caused by the rapid advancement of technology. To prevent phishing efforts, machine learning is used as a powerful tool. The four techniques used in this paper are AdaBoost Classifier, XGBoost Classifier, Random Forest Classifier, Gradient Boosting Classifier, and Support Vector Machine (SVM).*

Key Words: **AdaBoost Classifier, XGBoost Classifier, Random Forest Classifier, Gradient Boosting Classifier and Support Vector Machine (SVM).**

1. INTRODUCTION

Phishing is the risky illegal behavior in online world. Phishing efforts have considerably increased over many years as many people utilize the online services given by governmental and private institutions. When con artists discovered a profitable business model, they did so. Phishers use a range of methods to target the unwary, including voice over IP (VOIP), messaging, spoof links, and fake websites. It's simple to create a fraud website that is similar to real website, but it's not. Even the information on these websites would be identical to that on the genuine versions. The target of these websites is to gather user information, such as account numbers, login credentials, debit and credit card passwords, etc. Attackers also pose as high-level security measures and ask users to respond to security questions. Those who respond to those inquiries are more likely to fall victim to phishing scams. Many investigations have been done to stop phishing assaults by various groups throughout the world. By identifying the websites and educating people to recognize phishing websites, phishing assaults can be stopped. Machine learning techniques are the best ways to spot phishing websites.

One of the key techniques assisting artificial intelligenceis machine learning (AI). It is founded on algorithms designed to comprehend and recognize patterns from massive amounts of data to build a system that can forecast anomalous behavior and occurrences. It changes over time as it picks up on typical behavioral tendencies.

2. LITERATURE REVIEW

A thoughtful piece of writing known as a literature review communicates the information that is currently available, including significant findings and theoretical and methodological commitments to a particular subject.

M. Somesha et al. investigated the architecture of a system that comprises of feature collection, feature picking, and classification procedures. A list of website URLs are used as input to the feature collector, and it pulls the required features from three sources (URL obscuring, anchoring text and other sources based). The obtained features are then fed into the IG attribute positioning algorithm. The proposed model's drawback is that it depends on external services, which means that if these services aren't available, work performance would be limited.Moreover, the suggested model could be unable to identify phishing websites that replace textual content with embedded objects [1].

A very successful phishing website detection model (OFS-NN) based on neural network technology and an appropriate feature selection method was addressed by Erzhou Zhu et al. A sign termed feature validity value (FVV) has been constructed in this suggested model to evaluate the effects of each of those parameters on the identification of such websites. An algorithm is now being built to find the best features on the phishing attacks based on this recently created sign. The issue with the neural network greatly lessened by the selected strategy. The issue of the neural network's over-fitting will be greatly reduced by the chosen algorithm. The neural network is trained using these ideal attributes to create an ideal classifier that can identify phishing URLs. Nevertheless, the problem is thatthe OFS must continually gather additional features due to the expanding number of features that are vulnerable to phishing attempts [2].

Derek Doran and Mahdieh Zabihimayvan talked about the Fuzzy Rough Set (FRS) theory, which was developed into a tool that selects the best features from a small number of standardized datasets. Afterwards, a few classifiers receive these features in order to detect phishing. A dataset of 14,500 website models is used to train the models in order to examine the feature identification for FRS in developing a common detection of phishing. Nevertheless, disadvantage is that the method's unique properties are not stated [3].

Even though the efficiency of the system will be heavily dependent on idea of the features, Peng Yang et al. offered a theory that discussed feature engineering is a crucial part in discovering answers for detecting fake websites. The constraint is in the amount of time it takes to gather these features, Even the attributes obtained from all of the different aspects are acknowledgeable. The researchers have suggested many aspects of fraud detection feature perspective that focuses on a quick detection method by utilizing deep learning to address this flaw (MFPD) [4].

According to T. Nathezhtha et al., a triple-phase identification system dubbed the Web Crawler based Phishing Attack Detector (WC-PAD) has been proposed to accurately find the instances of phishing. The classification of phishing and legitimate websites is done using the web's content, web traffic, and website URL as input attributes. Nevertheless, the disadvantage is that it takes time because there are three phases, and each website must go through them [5].

To find which website is real or a fake site, C. Emilin Shyni et al. discussed. This is a creative way to find these websites by using the Google API to intercept all of the hyperlinks on the current page and creating a parse tree out of all of the hyperlinks that were intercepted. Here, parsing starts at the root node. If a child node has the identical value as the root node, it uses the Depth-First Search technique to find it. The disadvantage, however, is that both false positive and false negative rates are high [7].

3. PROBLEM STATEMENT

To design and develop a system that should take less time to detect phishing websites so that it can accurately and effectively classify websites as real or fraudulent.

4. EXISTING SYSTEM

To stop phishing assaults, a method based on web crawling was created. It employs three phases for detection, using the input elements of URL, traffic, and online content. The disadvantage is that it takes time because each website must go through three stages.

A method called Fuzzy Rough Set (FRS) was developed to help users to choose the best attributes from a small number of standardized datasets. Nevertheless, the disadvantage is that the method's unique properties are not stated.

5. PROPOSED SYSTEM

Machine learning is an innovative and popular technology that has a wide range of applications in society and can handle massive amounts of data as well as refined and updated algorithms.

The dataset is submitted to the proposed system, where attributes like IP Address, Tiny URL, URL Length, URL Depth and others are extracted from the dataset. Among the machine learning techniques, the system uses AdaBoost, Random Forest Classifier, XGBoost, Gradient Boosting Classifier, and Support Vector Machine (SVM). Random Forest Classifier is the algorithm that most accurately determines whether a website URL is legitimate or fraudulent.

6. OBJECTIVES

- To create a model that can identify phishing websites from legitimate websites.
- To use different machine learning methods to train model to produce accurate and effective results.

7. METHOLOGY

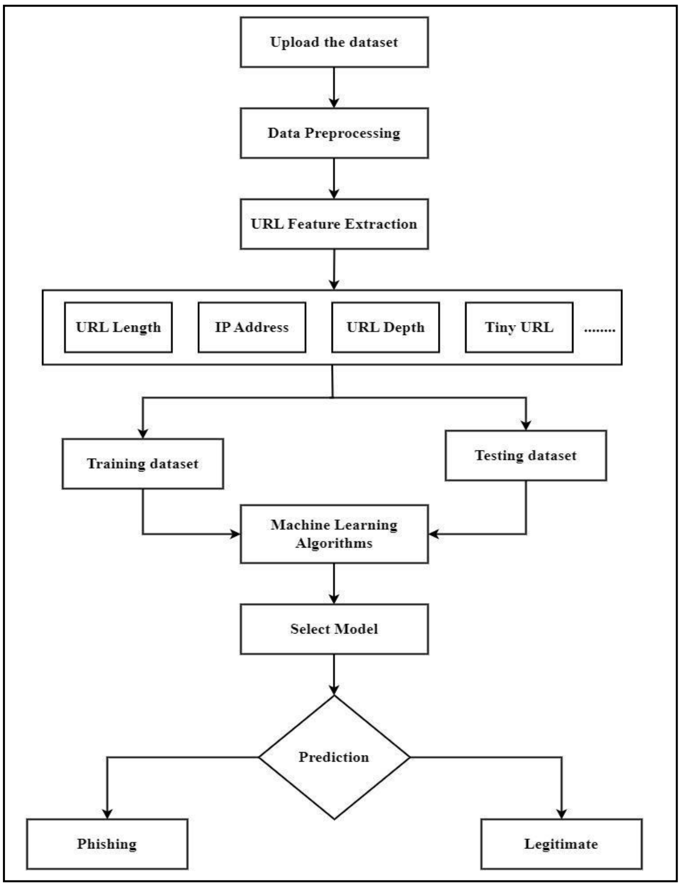


Fig. 1: System Architecture

I. AdaBoost Classifier Algorithm:

The AdaBoost classifier, commonly referred to as adaptive boosting, is a machine learning ensemble technique. Adaptive boosting is the process of redistributing the weights to each instance, giving larger weights to instances that were incorrectly identified. Boosting is used to reduce bias and variation in supervised learning. It is based on the idea that learners make progress in sequential manner and in stages. Except for the first, every learner after that is created from a previous learner. Iterative construction is used to integrate a number of weak classifiers to produce a strong classifier with high accuracy.

AdaBoost must adhere to two requirements:

1. Interactive training of the classifier using a variety of weighed training samples is recommended.

2. By reducing training error, it seeks to offer a superb fit for these samples in each iteration.

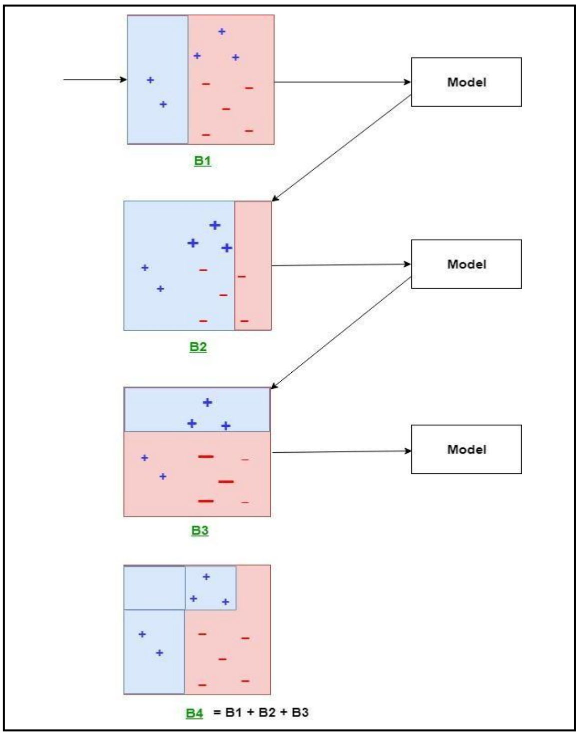


Fig. 2: AdaBoost Classifier Algorithm

II. XGBoost Classifier Algorithm:

The acronym XGBoost stands for Extreme Gradient Boosting. It is an elaborated gradient boosting library developed to be extremely accurate, versatile. It uses machine learning methods using Gradient Boosting framework. One of the important features of XGBoost is its efficient handling of missing values, which enables it to handle real-world data with missing values without necessitating a lot of pre-processing. Moreover, it has built-in parallel processing functionality, making it possible to train models quickly on large datasets. Additionally, by allowing for the fine-tuning of multiple model parameters, it is very versatile and facilitates performance optimization.

XGBoost is an extended version which was developed expressly to boost speed and performance. One of the important features of XGBoost is its efficient handling of missing values, which enables it to handle real-world data with missing values without necessitating a lot of pre-processing. Furthermore, XGBoost has built-in parallel processing capabilities that lets to train models on huge datasets quickly. It has ability to work with large datasets and offer cutting-edge performance in numerous machine learning tasks including classification and regression.

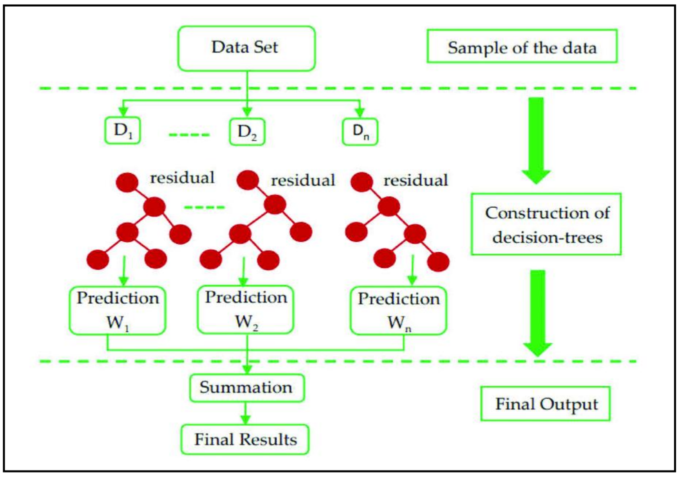


Fig. 3: XGBoost Classifer Algorithm

III. Random Forest Algorithm:

The Random Forest Algorithm, which is far less sensitive to training data, consists of a number of random decision trees. Creating a new dataset from the original data is the first step in building a Random Forest. The act of creating new data is known as boot strapping. Afterwards, a random selection of features will be utilized to train each tree. Each decision tree is constructed and then given a new data point. The next step is to merge all of the trees. As it's a classification problem, decision of the majority is followed. Aggregation is the phrase used to describe the process of integrating all decision tree outputs. In Random forest, aggregation occurs after bootstrapping.

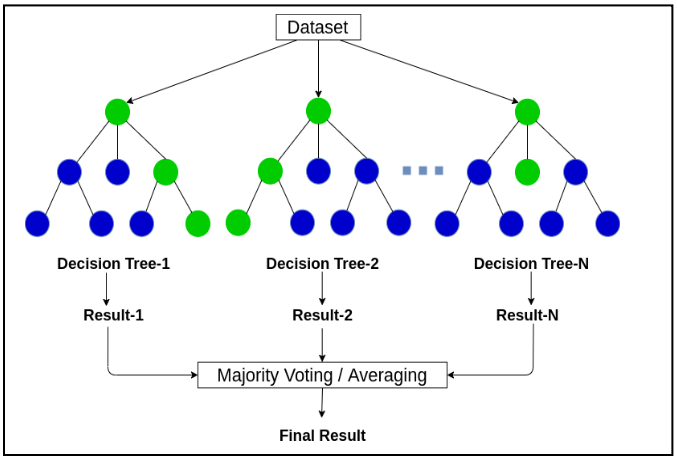


Fig. 4: Random Forest Algorithm

IV. Gradient Boosting Classifier Algorithm:

This gradient boosting algorithm is the most used algorithm in the machine learning field because it makes less mistakes and is widely known. It is used to accurately predict errors in the system and also used to lessen the errors.

The base estimator in the gradient boosting process cannot be identified, unlike the AdaBoost algorithm. The Gradient Boost algorithm's default finder, Decision Stump, is fixed. The gradient boosting algorithm's n estimator can be adjusted, just like AdaBoost. The default value of n_estimator for this algorithm, however, is 100 if we do not specify a number for it. It will forecast categorical variables and also continuous variables.

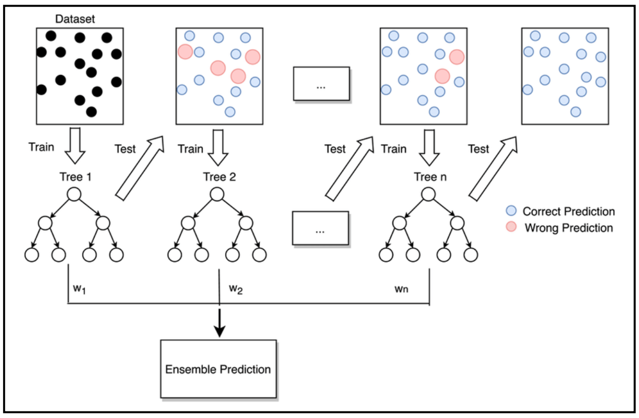


Fig. 5: Gradient Boosting Classifier Algorithm

V. Support Vector Machine (SVM) Algorithm:

Support Vector Machine is liked by everyone and it is used everywhere to solve the problems. The main purpose of SVM is to create a boundary that divides the system into subparts which will help us to find answers and the boundary is called as Hyperplane.

Using Hyperplanes and Support vectors in the SVM algorithm:

1. Hyperplane: It is feasible to divide classes into a variety of lines or decision boundaries into multi-dimensions, but it is necessary to find the decision boundary which is best for categorizing the data points. This ideal boundary is known as the SVM hyperplane. Given that the dataset's features define the hyperplane's dimensions, a straight line will be the hyperplane if there are just two features. In addition, the hyperplane will only have two dimensions if there are three features.

2. Support vectors: The closest data points or vectors near the hyperplane and those that have an impact on the plane's position are referred to as Support vectors.

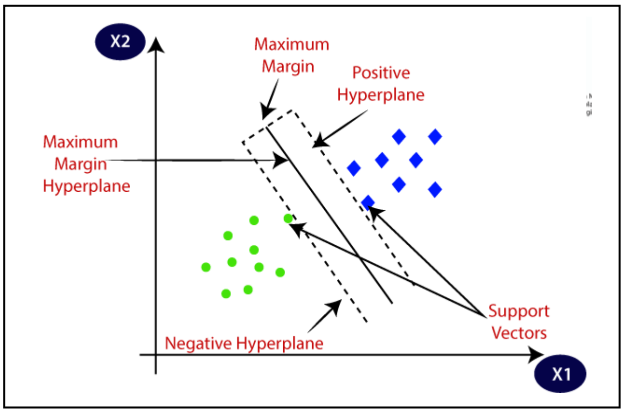
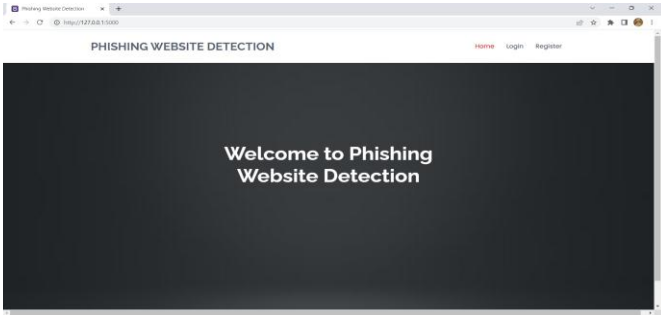


Fig. 6: SVM Algorithm

8. EXPERIMENTAL RESULTS

In order to assess whether a given URL is real or phishing, the system is given a dataset of URL attributes including IP Address, URL Length, URL Depth, Tiny URL, etc. and the dataset is trained using various machine learning algorithms. The system processes, analyzes the given URL and gives the result whether the website is phishing or legitimate.



Welcome to Phishing Website Detection

Fig. 7: Home Page

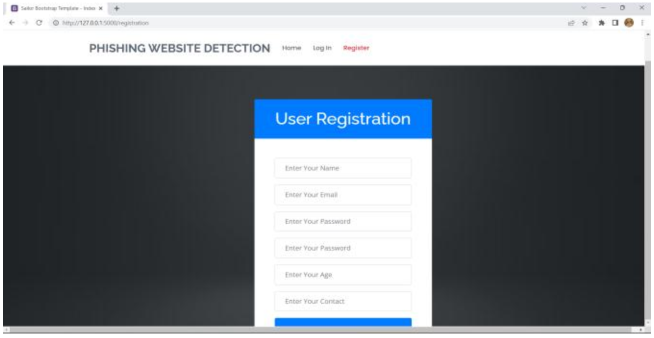
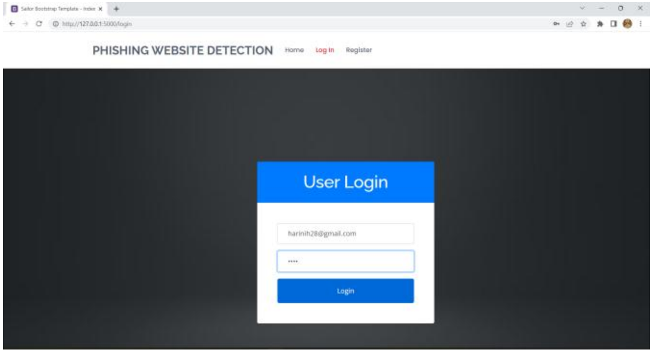
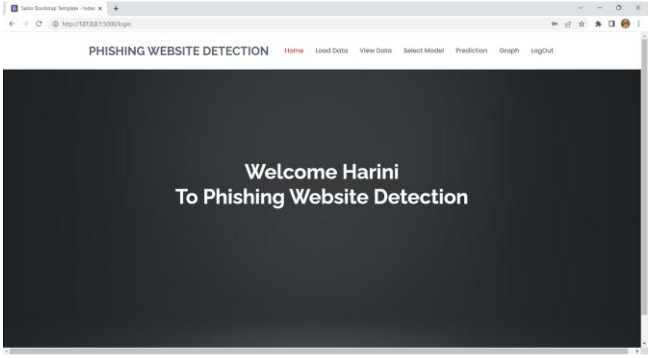


Fig. 8: User Register Page



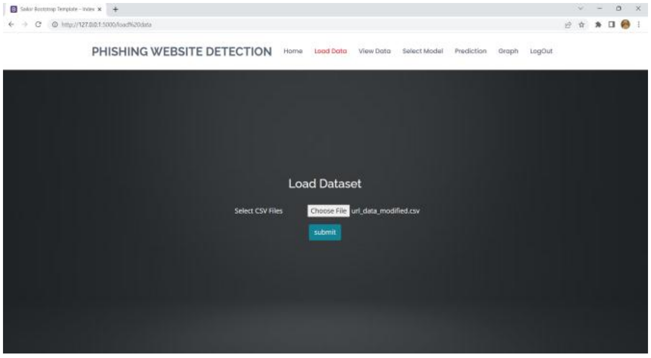
harinih28@gmail.com

Fig. 9: User filling the Login Page



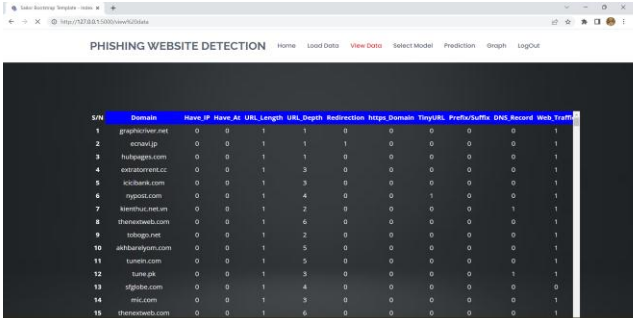
Welcome Harini
To Phishing Website Detection

Fig. 10: Welcome page is displayed when user logins



Select CSV Files: url_data_modified.csv

Fig. 11: Load URL Dataset Page



S/N	Domain	Have_IP	Have_At	URL_Length	URL_Depth	Redirection	https_Domain	TinyURL	Prefix/Suffix	DNS_Record	Web_Traffi
1	graphicriver.net	0	0	1	1	0	0	0	0	0	1
2	ecnavi.jp	0	0	1	1	1	0	0	0	0	1
3	hubpages.com	0	0	1	1	0	0	0	0	0	1
4	extratorrent.cc	0	0	1	3	0	0	0	0	0	1
5	icicibank.com	0	0	1	3	0	0	0	0	0	1
6	nypost.com	0	0	1	4	0	0	1	0	0	1
7	kienthuc.net.vn	0	0	1	2	0	0	0	0	1	1
8	thenextweb.com	0	0	1	6	0	0	0	0	0	1
9	tobogo.net	0	0	1	2	0	0	0	0	0	1
10	akhbarelyom.com	0	0	1	5	0	0	0	0	0	1
11	tunein.com	0	0	1	5	0	0	0	0	0	1
12	tune.pk	0	0	1	3	0	0	0	0	1	1
13	sfglobe.com	0	0	1	4	0	0	0	0	0	0
14	mic.com	0	0	1	3	0	0	0	0	0	1
15	thenextweb.com	0	0	1	6	0	0	0	0	0	1

Fig. 12: View URL Dataset page

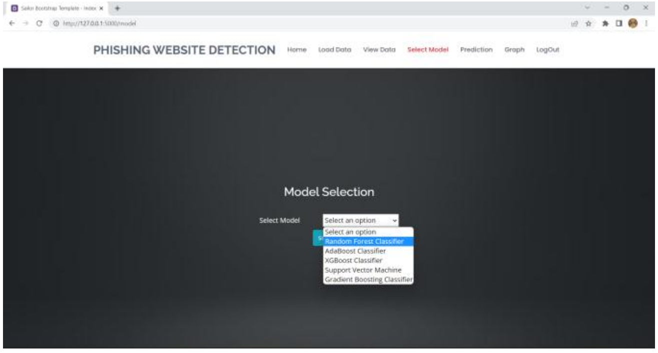
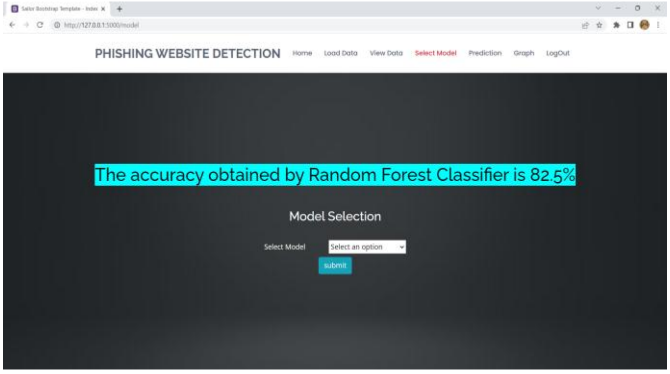
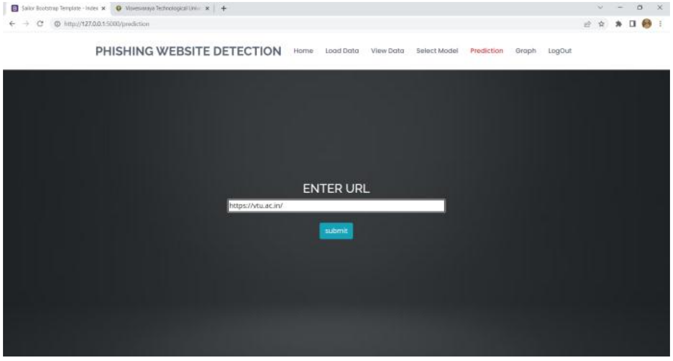


Fig. 13: Select Algorithm Model Page



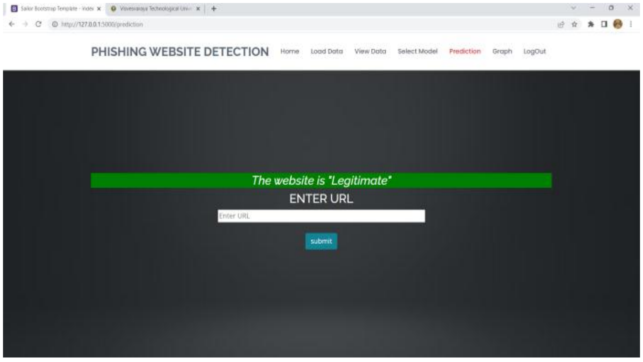
The accuracy obtained by Random Forest Classifier is 82.5%

Fig. 14: Accuracy of the Selected Random Forest Algorithm



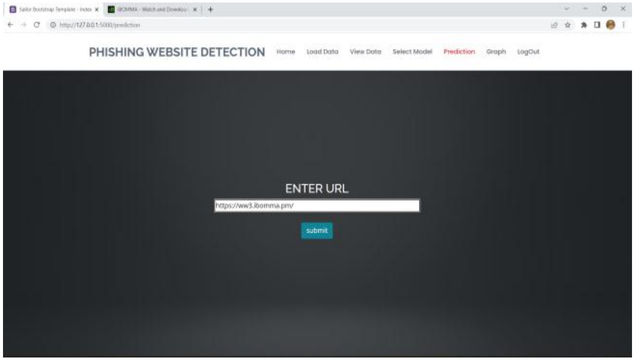
ENTER URL: https://vtu.ac.in/

Fig. 15: Legitimate website URL is given by the user


Sailor Bootstrap Template - Index
Visvesvaraya Technological Univ
http://127.0.0.1:5000/prediction
PHISHING WEBSITE DETECTION
Home
Load Data
View Data
Select Model
Prediction
Graph
LogOut
ENTER URL
Enter URL
submit

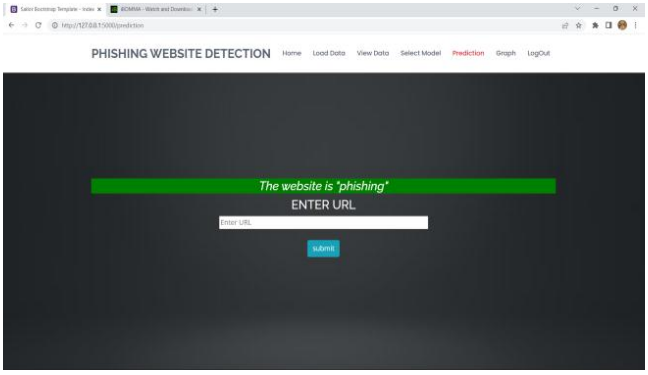
The website is "Legitimate"

Fig. 16: The given URL is classified as Legitimate


Sailor Bootstrap Template - Index
iBOMMA - Watch and Download
http://127.0.0.1:5000/prediction
PHISHING WEBSITE DETECTION
Home
Load Data
View Data
Select Model
Prediction
Graph
LogOut
ENTER URL
submit

https://ww3.ibomma.pm/

Fig. 17: Phishing website URL is given by the user


Sailor Bootstrap Template - Index
iBOMMA - Watch and Download
http://127.0.0.1:5000/prediction
PHISHING WEBSITE DETECTION
Home
Load Data
View Data
Select Model
Prediction
Graph
LogOut
ENTER URL
Enter URL
submit

The website is "phishing"

Fig. 18: The given website URL is classified as Phishing

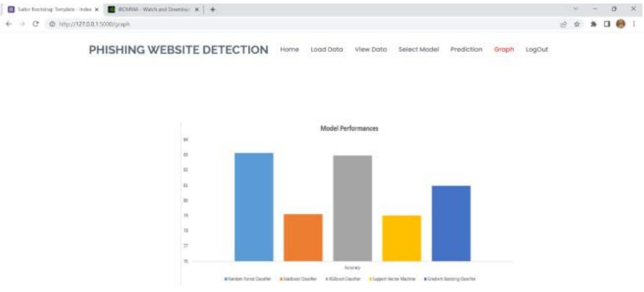


Fig. 19: Accuracy of Algorithms shown in Bar Graph

9. CONCLUSION

This paper included various machine learning techniques and methods to find phishing websites. We come to the inference that the larger part of the work is completed using well-known machine learning techniques like XGBoost Classifier, Random Forest Classfier etc. Some authors suggested an up to date system for fraud detection, similar to Phish Score and Phish Checker. In terms of exactness, correctness, recollect, etc., feature combinations were used. Phishing websites are becoming more prevalent every day, thus elements that are used to identify them may be added or replaced with new ones.

10. FUTURE SCOPE

There is always a space of improvement in every system. There are more classifiers such as the Bayesian network classifier, Neural Networks. Such classifiers can be included and this could be counted in future to give a more data to be compared with. The project can also include other variants of phishing like smishing, vishing, etc. to complete the system, so these can be implemented. Looking even further out, the methodology needs to be evaluated on how it might handle collection growth.

REFERENCES

[1] M Somesha, Alwyn Roshan Pais, Routhu Srinivasa Rao, Vikram Singh Rathour. "Efficient deep learning techniques for the detection of phishing websites", June 2020.

[2] E. Zhu, Y. Chen, C. Ye, X. Li, and F. Liu. OFS-NN: "An Effective phishing websites detection model based on optimal feature selection and neural network". IEEE Access, 7:73271–73284, 2019.

[3] Mahdieh Zabihimayvan and Derek Doran. "Fuzzy Rough Set feature selection to enhance phishing attack detection", 03 2019.

[4] P. Yang, G. Zhao, and P. Zeng. "Phishing website detection based on multi-dimensional features driven by deep learning". IEEE Access, 7:15196–15209, 2019.

[5] T. Nathezhtha, D. Sangeetha, and V. Vaidehi. "WC-PAD: Web crawling based phishing attack detection". In 2019 International Carnahan Conference on Security Technology (ICCST), pages 1–6, 2019.

[6] Y. Huang, Q. Yang, J. Qin, and W. Wen. "Phishing url detection via CNN and attention-based hierarchical RNN". In 2019 18th IEEE International Conference On 55 Trust, Security and Privacy in Computing and Communications/13th IEEE International Conference On Big Data Science and Engineering (TrustCom/BigDataSE), pages 112-119, 2019.

[7] C. E. Shyni, A. D. Sundar, and G. S. E. Ebby. “Phishing detection in websites using parse tree validation”. In 2018 Recent Advances on Engineering, Technology and Computational Sciences (RAETCS), pages 1–4, 2018.

[8] S. Parekh, D. Parikh, S. Kotak, and S. Sankhe. “A new method for detection of phishing websites: URL detection”. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), pages 949–952, 2018.

[9] J. Li and S. Wang. “Phishbox: An approach for phishing validation and detection”. In 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/ PiCom/DataCom/CyberSciTech), pages 557–564,2017.

[10] H. Shirazi, K. Haefner, and I. Ray. “Fresh-Phish: A framework for auto-detection of phishing websites”. In 2017 IEEE International Conference on Information Reuse and Integration (IRI), pages 137– 143, 2017.

[11] A. J. Park, R. N. Quadari, and H. H. Tsang. “Phishing website detection framework through web scraping and data mining”. In 2017 8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pages 680–684, 2017.

[12] Lisa Machado and Jayant Gadge. “Phishing sites detection based on C4.5 decision tree algorithm”. pages 1–5, 08 2017.

[13] S. Haruta, H. Asahina, and I. Sasase. “Visual similarity-based phishing detection scheme using image and CSS with target website finder”. In GLOBECOM 2017 - 2017 IEEE Global Communications Conference, pages 1–6, 2017.