2021 5th International Conference on Information Systems and Computer Networks (ISCON)
GLA University, Mathura, India. Oct 22-23, 2021

Phishing Detection using Extra Trees Classifier

Arathi Krishna V*, Anusree A, Blessy Jose, Karthika Anilkumar, Ojus Thomas Lee†
Department of Computer Science & Engineering
College of Engineering Kidangoor
Kerala, India
*arathikrishna499@gmail.com, †ojustl@ce-kgr.org

2021 5th International Conference on Information Systems and Computer Networks (ISCON) | 978-1-6654-0341-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISCON52037.2021.9702372

***Abstract*—With a rapid growth in global networking, the online users are vulnerable to different kinds of attacks, phishing being prevalent among them. Phishing is the type of attack where the attacker aims to steal critical information by tricking the user to click on phishing links. There already exists several anti-phishing software and computational methods for actively detecting phishing activities. However, new methods of cybercrimes are evolved by the attackers that surpass the existing detection models. So, there is a constant need to research and improvise the ways to detect phishing. The proposed system develops a web-based application to detect phishing URLs using a machine learning model. Two ensemble classifiers, Random Forest (RF) and Extra Trees (ET) are compared to find the one with higher performance measures. The models are trained on the UCI dataset with 30 features. Hyperparameter Tuning is performed on the models to check whether it enhances their predictive performance. The Extra Trees classifier without tuning achieved the highest accuracy of 97.47% on the test dataset with the least false positive rate.**

Keywords—phishing detection, url features, ensemble classifier, hyperparameter tuning

I. INTRODUCTION

There has been a tremendous increase in the number of security threats to web services on the Internet over the last few years. Phishing is one of the several risks that users come across while using the Internet. It is the most commonly used technique by the cybercriminals to obtain confidential information from users by impersonating a legitimate entity such as a bank, an organization, a social network etc.

According to the 2021 Verizon Data Breach Investigations Report, phishing is responsible for a majority of the breaches in social engineering kind of attacks [1]. The Phishing Activity Trends Report by Anti-Phishing Working Group (APWG) for the first quarter of 2021 stated that financial institutions, webmail and social media sectors were the most frequently victimized by phishing in this quarter [2]. APWG's records shows that, January 2021 had an unprecedented 245,771 phishing attacks in a single month. According to Barracuda Networks, with the onset of the pandemic, malicious mails rose by 667%. In 2020, Google claimed to block more than 100 million scam emails everyday, out of which 18 million were related to Covid-19 [3].

There are various approaches for detecting phishing websites that have been discussed over the years. There is Rule based or Heuristics based approach [4], Blacklisting approach [5], Content based approach [6] etc. The existing studies focuses on applying supervised machine learning algorithms to identify phishing websites. These algorithms predict whether websites are phishing or legitimate. It learns the characteristics of existing phishing websites and predicts new phishing characteristics. The performance of the machine learning models varies based on the measures used for evaluating them as discussed in the paper [7]. The major highlights of this work are: comparison of ensemble classifiers Random Forest and Extra Trees or Extremely Randomized Trees, Hyperparameter Tuning of the classifiers to check whether it enhances predictive performance, developing a machine learning based web application for phishing detection.

The rest of the paper is structured as follows: some of the recent research works in this field are discussed in Section II. Section III discusses the methodology used in this paper. The results obtained, final observations after comparing the models and web application design are discussed in Section IV. The paper is concluded in Section V.

II. RELATED WORKS

In literature, there are various models to detect phishing. The existing works apply Naive Bayes, Decision Tree, Random Forest, Support Vector Machine, k-Nearest Neighbor and Artificial Neural Network (ANN) for machine learning based phishing detection.

Smita Sindhu et al. [8] proposed a browser extension to detect phishing. They used three algorithms, Support Vector Machine (SVM), Random Forest (RF) and Neural Network with backpropagation on the UCI phishing dataset. They concluded SVM as the best classifier as it gave better frequency than Neural Networks. Even though RF had high accuracy, the accuracy rates were not constant most of the time. They performed lexical feature extraction on the dataset before passing it to the classifiers.

Shinelle Hutchinson et al. [9] created five different subsets of features from the UCI dataset and compared them by using Random Forest algorithm. As it was evident from the results, the subset having only the most important features were enough to increase the accuracy, precision and recall. By focusing on feature importance, they reduced the feature count by over half compared to the total number of features. Random Forest was chosen because it runs efficiently on large datasets and it can handle missing values.

The authors Mahajan Mayuri Vilas et al. [10] used 30 features in the UCI dataset to classify phishing URLs. They applied the Extreme Learning Machine (ELM) algorithm on the dataset. ELM algorithm reduces the time-consuming training speed and overfitting issues. The process of ELM is different from that of ANN as it renews its parameters and input weights are accidentally chosen while output weight is

calculated analytically. ELM also avoids local minimization and multiple iterations.

Almaha Abuzuraiq et al. [11] used a balanced dataset that contained 5000 phishing and 5000 legitimate websites. The proposed model had 2 stages. In the first stage, machine learning algorithms were applied on the dataset along with feature selection algorithms such as Infogain and Relief-F. Out of the 48 features, combining 20 features with Random Forest gave the best accuracy of 98.11%. In the second stage, various fuzzy logic algorithms were applied on the same dataset. The results obtained were incredible, with the accuracy rates close to 100% by using only five features. For machine learning algorithms, higher number of features resulted in higher accuracy rate whereas for fuzzy logic systems, lower number of features leads to a higher accuracy rate. It was also concluded that the time taken to build a fuzzy logic model was lower.

The authors Amani Alswailem et al. [12] aimed to create a higher performance classifier by studying the features of URLs and choosing a better combination. They used the Random Forest algorithm for classification. A combination of 26 features obtained the highest accuracy of 98.8%. The dataset had 6116 instances with a total of 36 features. The dataset was split into training and test data in the ratio 80:20. They proposed to train and test all possible combinations of 36 features to get the strongest features that enhance the detection accuracy. The final classifier was executed with the minimum number of features to get the maximum accuracy.

III. METHODOLOGY

In this section, an overview of the proposed method which includes the execution flow as shown in Fig. 1 is presented along with the system architecture given in Fig. 2. It is followed by a brief discussion about the datasets used and an insight into how the model evaluation happens.

A. Overview

In this paper, the process of phishing detection requires a machine learning model to identify whether an input URL is phishing or legitimate by extracting its features [13]. In the existing works, the studies mostly compare the performance of ensemble classifiers like Random Forest against single classifiers. Single classifiers mostly produce less effective models as compared to ensemble classifiers [14]. The proposed approach will therefore compare the performance of two ensemble models on a dataset. The models used in this paper are two similar ensemble classifiers, Random Forest (RF) and Extra Trees (ET). The difference is that RF chooses an optimum split while ET chooses a split at random. The experiment is broken into 4 stages as depicted in Fig. 1.

All the chosen classifier models are evaluated on the collected dataset. The classifier model that gives the best performance metrics is selected for implementation after comparing all the models. The architecture of the proposed system is given in Fig. 2.

B. Dataset

The first stage involves collection of dataset and pre-processing of the data for the experimental setup. The phishing dataset used in this study is the UCI phishing website dataset that has been widely used in the existing studies. The dataset contains 11,055 entries, of which 6,157 are legitimate URLs and 4,898 are phishing URLs. There are 30 independent features with two labels, '-1' representing a phishing URL and '1' representing a legitimate URL [15], [16], [17].

C. Model Evaluation

In this stage, we implement the ensemble classifier models with and without hyperparameter tuning on the UCI dataset

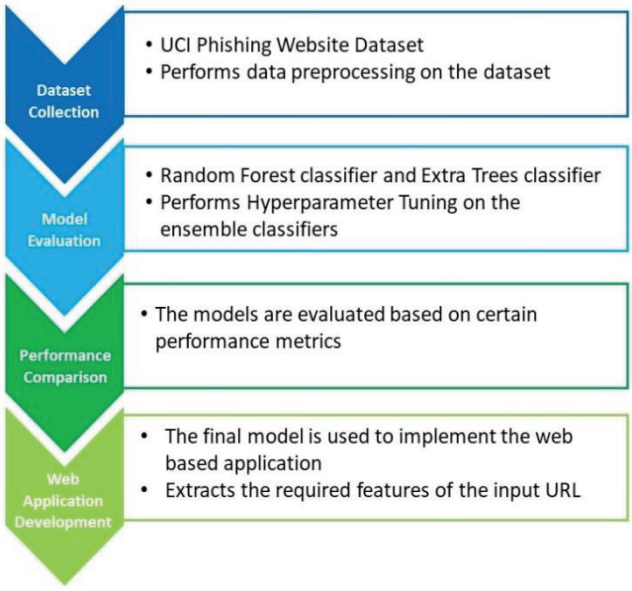

Fig. 1. Execution flow of the proposed model

and measure their performance based on certain parameters like test accuracy, precision, recall and false positive rate. Both RF and ET build multiple decision trees and merge them together to get a more accurate and stable prediction. The two models without hyperparameter tuning are to be called 'baseline models' in this paper. Then hyperparameter tuning is applied on the two classifiers.

Hyperparameters describes the properties of a model that can affect the model accuracy and computational efficiency. Hyperparameter Tuning is a technique used to find ideal parameters for a model and thus improve the performance of baseline models. It is also known to reduce overfitting [18]. This paper analyses the possibility of tuning the parameters of RF and ET, in order to check whether it improves the performance of the models when applied on this phishing dataset. The six hyperparameters tuned are as follows [19], [20]:

1) n_estimators = number of trees in the forest
2) max_features = maximum number of features considered for splitting a node
3) max_depth = maximum number of levels in each decision tree
4) min_samples_split = minimum number of data points placed in a node before the node is split
5) min_samples_leaf = minimum number of data points allowed in a leaf node
6) bootstrap = method for sampling data points (with or without replacement)

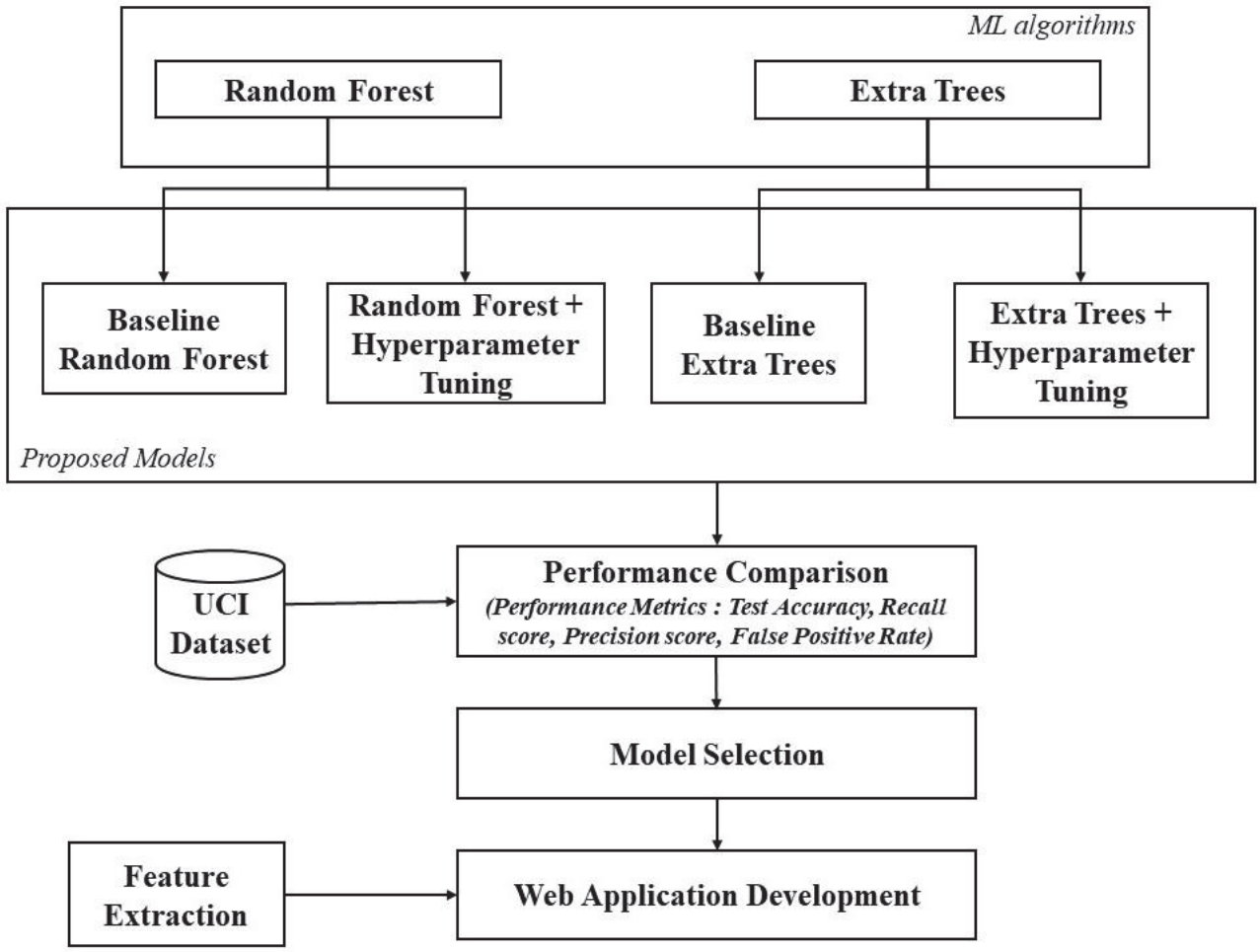

Fig. 2. System Architecture of the proposed model

The process of hyperparameter tuning consists of two steps, random search followed by grid search. Random search involves not trying every combination, but selecting at random to sample a wide range of values. The important arguments in random search are: n_iter, which is the number of different combinations to try, and cv, which is the number of folds to use for cross validation. Grid search, on the other hand, rather than sampling randomly from a distribution, it evaluates all the combinations that we define. In order to use grid search, we have to make another grid based on the best values obtained from random search.

IV. RESULTS AND DISCUSSIONS

In this section, the whole set of results of the proposed system obtained from evaluating all models, identification of the best performing model and the accuracy of the final model chosen for implementation are presented.

A. Performance Comparison

Performance comparison of RF and ET are done for baseline models followed by hyperparameter tuned RF and ET.

When applying random search, the arguments are set as n_iter = 100 and cv = 5 with total number of fits as 500. In case of grid search, we take n_iter = 400, cv = 5 and total fits is 2000. In each case of baseline and hyperparameter tuned models of RF and ET, confusion matrix is used to identify the best model. The set of values for grid search in case of RF is shown in Fig. 3, while that for ET is shown in Fig. 4.

The confusion matrix of baseline RF is given in Table I and Test Accuracy, Recall, Precision, False Positive Rate values due to the confusion matrix are presented in Table II.

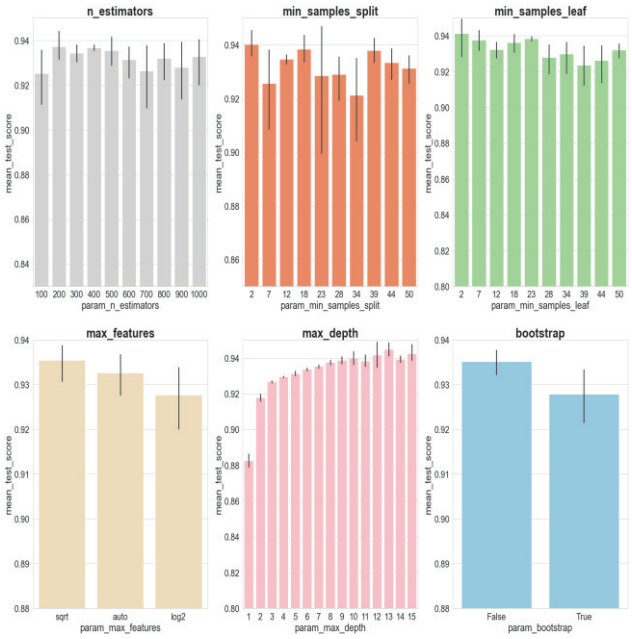

Fig. 3. Grid values after random search in Random Forest classifier

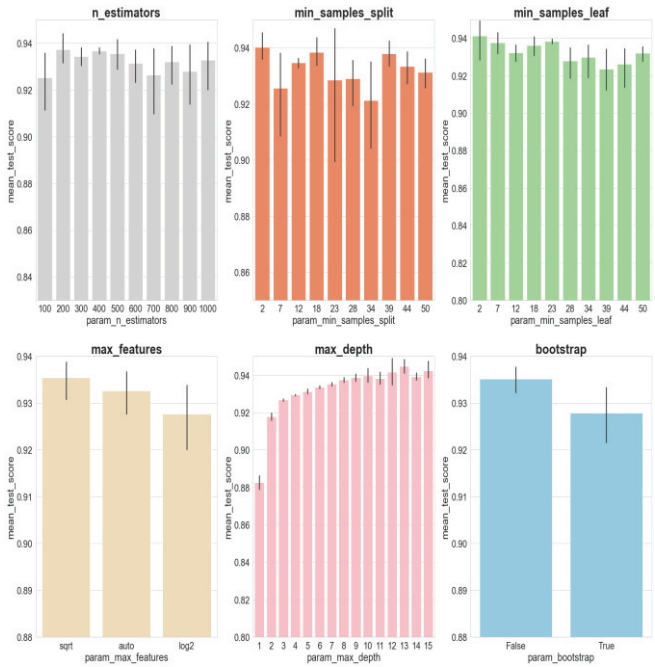

Fig. 4. Grid values after random search in Extra Trees classifier

TABLE I. CONFUSION MATRIX OF BASELINE RF

	Predicted Phishing	Predicted Legitimate
Actual Phishing	1174	51
Actual Legitimate	20	1519

TABLE II. PERFORMANCE METRICS OF BASELINE RF

Test Accuracy	Recall	Precision	False Positive Rate
97.43	98.70	96.75	4.16

TABLE III. CONFUSION MATRIX OF BASELINE ET

	Predicted Phishing	Predicted Legitimate
Actual Phishing	1179	46
Actual Legitimate	24	1515

TABLE IV. PERFORMANCE METRICS OF BASELINE ET

Test Accuracy	Recall	Precision	False Positive Rate
97.47	98.44	97.05	3.76

TABLE V. CONFUSION MATRIX OF RF AFTER HYPERPARAMETER TUNING

	Predicted Phishing	Predicted Legitimate
Actual Phishing	1173	52
Actual Legitimate	26	1513

TABLE VI. PERFORMANCE METRICS OF RF AFTER HYPERPARAMETER TUNING

Test Accuracy	Recall	Precision	False Positive Rate
97.18	98.31	96.68	4.24

TABLE VII. CONFUSION MATRIX OF ET AFTER HYPERPARAMETER TUNING

	Predicted Phishing	Predicted Legitimate
Actual Phishing	1164	61
Actual Legitimate	26	1513

TABLE VIII. PERFORMANCE METRICS OF ET AFTER HYPERPARAMETER TUNING

Test Accuracy	Recall	Precision	False Positive Rate
96.85	98.31	96.12	4.98

Table III gives the confusion matrix for the baseline ET and the Test Accuracy, Recall, Precision, False Positive Rate values of the baseline ET are presented in Table IV. In a similar fashion, Table V presents the confusion matrix and Table VI gives the parameter values of hyperparameter tuned RF. The confusion matrix and parameters for the hyperparameter tuned ET is given in Table VII and VIII respectively. On the dataset used in this paper, the baseline RF and ET models produce better performance measures than the hyperparameter tuned models. After the analysis it was found that the baseline ET has the best performance compared to the other three models. The baseline ET model has an accuracy of 97.47% and false positive rate is only 3.76. The performance comparison of the models is depicted in Fig. 5.

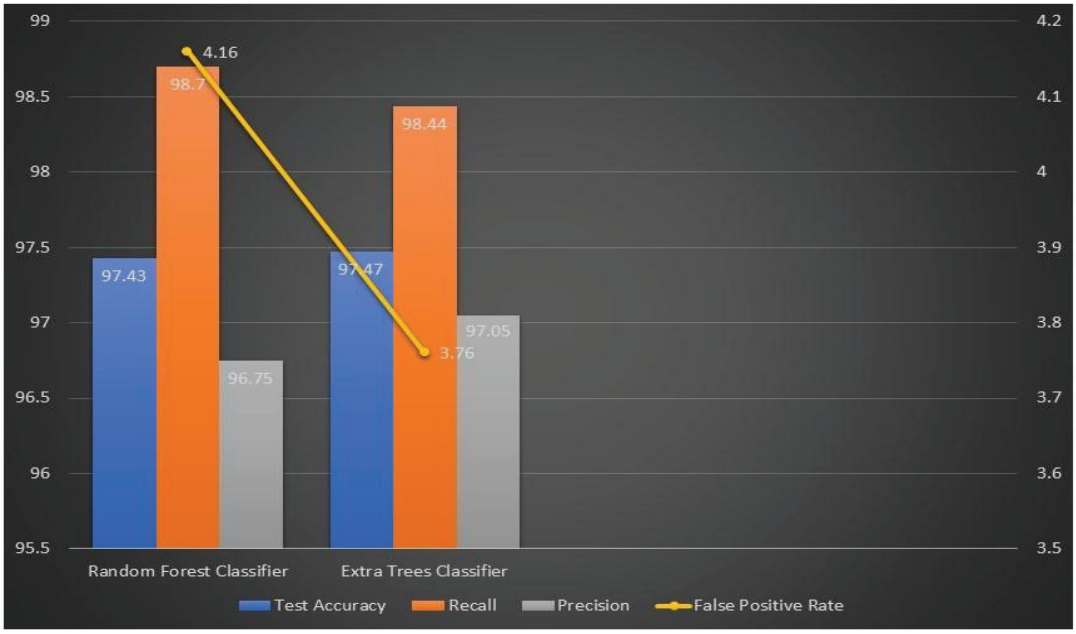

Fig. 5. Test Accuracy, Recall, Precision, False Positive Rate (axis on the right) values

B. *Web Application*

In this section, the implementation of the ET classifier and the process of feature extraction of an input URL that takes place in the web application development phase is discussed.

The dataset is divided into training data and test data in the ratio 75:25. The baseline ET classifier, with default hyperparameters is used to train the model on the UCI dataset. The test accuracy of the model is 97.47%. The proposed system prompts the user to input a URL on the screen as shown in Fig. 6. The system then extracts 30 features from the input URL and forms a feature vector. This feature vector is given to the classifier to predict the class to which the URL belongs, i.e., phishing or legitimate. The classification result along with certain properties is displayed to the user on the screen. This is depicted in Fig. 7. Finally the URL along with its prediction result and the feature vector is stored in a file.

When a URL is given as input, it searches for that URL in the file. If it is found there, the corresponding results are retrieved from the file. Thus the whole process of feature extraction and classification can be skipped. If the URL is not already present in the file, then only, feature extraction process is carried out on the URL. Since the websites are prone to updates, the results can only be stored for a short period, say fourteen days, after which the entry becomes invalid and the whole process has to be repeated.

V. CONCLUSION

A web based phishing detection model has been developed using the Extra Trees classifier. A small review of the existing works in the literature was conducted, all of which pointed out that an ensemble model like Random Forest performed better than single classifiers. Hence, in this paper, Random Forest and Extra Trees classifiers are being used on the dataset. Hyperparameter tuned classifiers' performance on the dataset is also measured. But it is concluded that hyperparameter tuning of Random Forest and Extra Trees classifiers doesn't improve their performance any further.

The model proposed in this paper has high performance metrics like the models discussed in section II and provides an efficient phishing detection mechanism. The objective of the study is to test the results of Extra Trees classifier against Random Forest classifier, understand the effect of hyper

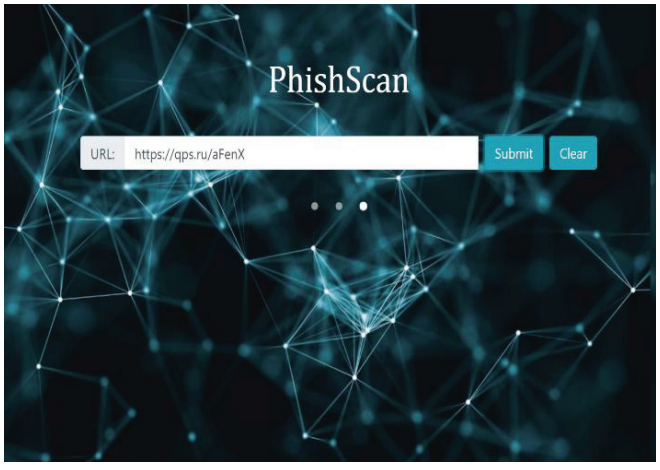

Fig. 6. User input screen

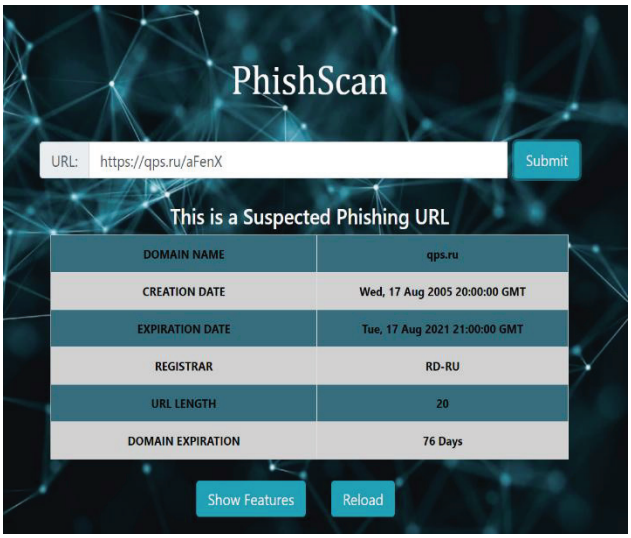

Fig. 7. Output screen

parameter tuning on the models and develop a system to identify phishing websites in order to help the users authenticate an unknown URL before clicking on it. Since the models' performance is restricted to the dataset being used in the experiment, in the future, we would like to explore the scope of extra trees classifier on a more extensive dataset that can identify spam URLs as well. We would also like to expand the application to generate predictions for multiple URLs at once.

REFERENCES

[1] 2021 Data Breach Investigations Report, Verizon https://enterprise.verizon.com/resources/reports/2021-data-breach-investigations-report.pdf

[2] Phishing Activity Trends Report, APWG, 1st Quarter 2021, https://docs.apwg.org/reports/apwg trends report q1 2021.pdf

[3] AI goes Phishing, https://analyticsindiamag.com/ai-goes-phishing/

[4] A. A. Ahmed and N. A. Abdullah, "Real time detection of phishing websites", 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, 2016

[5] A. Naga Venkata Sunil and A. Sardana, "A PageRank based detection technique for phishing web sites", 2012 IEEE Symposium on Computers & Informatics (ISCI), Penang, 2012

[6] N. Shrestha ; R. K. Kharel ; J. Britt ; R. Hasan, "High performance classification of phishing URLs using a multi-modal approach with MapReduce", 2015 IEEE World Congress on Services, New York, NY, 2015

[7] Charu Singh and Smt. Meenu, "Phishing website detection based on Machine Learning: A survey", 2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS), 2020

[8] Smita Sindhu ; Sunil Parameshwar Patil ; Arya Sreevalsan ; Faiz Rahman ; Ms. Saritha A. N., "Phishing detection using Random Forest, SVM and Neural Network with Backpropagation", 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), Bengaluru, 2020

[9] Shinelle Hutchinson ; Zhaohe Zhang ; Qingzhong Liu, "Detecting phishing websites with Random Forest", Third International Conference, MLICOM 2018 Proceedings, China, 2018

[10] Mahajan Mayuri Vilas ; Kakade Prachi Ghansham ; Sawant Purva Jaypralash ; Pawar Shila, "Detection of phishing website using Machine Learning approach", 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), India, 2019

[11] Almaha Abuzuraiq ; Mouhammd Alkasassbeh ; Mohammad Almseidin, "Intelligent methods for accurately detecting phishing websites", 2020 11th International Conference on Information and Communication Systems (ICICS), Jordan, 2020

[12] Amani Alswailem ; Bashayr Alabdullah ; Norah Alrumayh ; Dr.Aram Alsedrani, “Detecting phishing websites using Machine Learning”, 2019 2nd International Conference on Computer Applications & Information Security(ICCAIS), Saudi Arabia, 2019

[13] Rishikesh Mahajan and Irfan Siddavatam, “Phishing website detection using Machine Learning algorithms”, International Journal of Computer Applications (0975 – 8887) Volume 181 – No. 23, October 2018

[14] Mohammad Nazmul Alam ; Dhiman Sarma ; Farzana Firoz Lima ; Ishita Saha ; Rubaiath-E- Ulfath ; Sohrab Hossain, “Phishing attacks detection using Machine Learning approach”, Proceedings of the Third International Conference on Smart Systems and Inventive Technology (ICSSIT), India, 2020

[15] Arathi Krishna V ; Anusree A ; Blessy Jose ; Karthika Anilkumar ; Ojus Thomas Lee, “Phishing detection using Machine Learning based URL analysis: A survey”, International Journal of Engineering Research & Technology (IJERT), Special Issue, 2021

[16] Mehmet Korkmaz ; Ozgur Koray Sahingoz ; Banu Diri, “Feature selections for the classification of webpages to detect phishing attacks: A survey”, 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Turkey, 2020

[17] Rami M. Mohammad ; Fadi Thabtah ; Lee McCluskey, “Phishing websites features”

[18] Aung Kaung Myat and Myint Thu Zar Tun, “Predicting Palm oil price direction using Random Forest”, 2019 Seventeenth International Conference on ICT and Knowledge Engineering, Thailand, 2019

[19] Renan Soares de Andrades ; Mateus Grellert ; Mateus Beck Fonseca, “Hyperparameter Tuning and its effects on Cardiac Arrhythmia prediction”, 2019 8th Brazilian Conference on Intelligent Systems (BRACIS), Brazil, 2019

[20] Optimizing Hyperparameters in Random Forest Classification, https://towardsdatascience.com/optimizing-hyperparameters-in-randomforest-classification-ec7741f9d3f6