Phishing Website Detection

|  |  |
| --- | --- |
| Kavya Paliwal  *Computer Science with AI and Robotics*  *Vellore Institute of Technology*  Chennai, India | Dr. A Sheik Abdullah  *School of Computer Science and Engineering*  *Vellore Institute of Technology*  Chennai, India |

***Abstract—*** ***Phishing sites are designed to trick Internet users into entering sensitive information and are difficult to detect because they look legitimate. Techniques such as machine learning, feature-based analysis and heuristics have been developed, but due to the evolving nature of these sites, further research is needed. Despite detection techniques, phishing sites are still a serious threat. Internet users must be careful when surfing the web, because even the most advanced technologies cannot fully protect against phishing attacks. Despite the advances in these technologies, phishing sites are still a serious threat. This is partly due to the constant evolution of these sites, which makes them difficult to detect with static techniques. Therefore, further research is needed to develop more effective techniques to detect phishing sites. In conclusion, detecting phishing sites is a complex and ongoing challenge. Various techniques have been developed to identify phishing sites, but new techniques are needed to keep up with the ever-evolving nature of these sites. Finally, it is important for Internet users to be alert and cautious when browsing the Internet, as even the most advanced detection technologies cannot fully protect against phishing attacks.***

***Keywords—phishing; logistic regression; feature selection; classification models; random forest; prediction model***

# I. INTRODUCTION

Phishing is one of the most important and dangerous online threats in cyber security today. The use of social networks, e-commerce, e-banking and other online services has grown tremendously thanks to the rapid development of internet technologies. We Are Social (Global Overview Report 2021) released data "A Digital Report in 2021", according to which the number of Internet users increased to 4.66 billion, which is 7.3 percent (316 million new users) more than in January 2020. . Currently, Internet penetration remains unchanged at 59.5 percent, which gives a phishing attacker the opportunity to make money by extorting and stealing confidential information from Internet users. The attacker develops a fraudulent website and sends links to online platforms such as Facebook, Twitter, e-mails, etc., conveying a message of panic, urgency or a financial offer, prompting the recipient to act immediately. If a user accidentally clicks on a link and updates sensitive credentials, cyber attackers gain access to user information such as financial information, personal information, username, password, etc. Cybercriminals use this stolen information for various illegal activities, including blackmail. Victims

Phishing attacks are now used to distribute malware such as ransomware. The Anti-Phishing Working Group (APWG) studies phishing attacks and published a report showing that the number of phishing attacks identified by APWG members more than doubled in 2020 (APWG Q4 2020 Report 2020). In October alone, 225,304 new phishing sites were discovered during the COVID19 situation, breaking all previous monthly records. In 2020, the Internet Crime Complaint Center (IC3) received a record number (241,342) of complaints from Americans about phishing scams, with a reported loss of more than $54 million. Therefore, in this article, we will focus on effectively identifying phishing websites so that unsuspecting Internet users do not fall into a phisher's trap and thus reduce emotional and financial damage.

In the context of information technology, "data science" has become a hot field these days, because almost everything in our daily lives is digitally stored as data, and the insights gathered from the data are key to providing intelligent solutions. Such data solutions can be used for effective modeling and intelligent decision-making in various real-world application areas such as business or financial analytics, cyber security, IoT applications, etc. Thus, this article aims to create an effective data solution using machine learning techniques to detect if a website is phishing. Most machine learning based phishing detection methods extract features from URL, search engine, third parties, web traffic, DNS etc. These types of approaches may not be suitable for real-time phishing detection due to complexity and time constraints. According to APWG 1H2014 statistics, the average life cycle of phishing sites is less than 10 hours, and half of phishing sites were shut down in less than a day. However, most phishing pages using compromised domains remain on the Internet for more than a day. Therefore, the research question addressed in this paper is: "How can we develop an effective and intelligent phishing detection model considering the above issues?"

This research presents the accuracy improvement using the feature selection algorithm used, as well as the prediction model that uses ensemble learning, where most of the results contribute to the final prediction. In summary, the most important results of all the models used in combination are discussed.

# II. LITERATURE SURVEY

[1] authors Rishikesh Mahajan and Irfan Siddavatam selected decision trees, random forests, and support vector machines as their three classification algorithms. 17,058 non-phishing URLs and 19,653 phishing URLs were included in their dataset, which was gathered via the Alexa website and PhishTank. Both had sixteen properties. In the ratios of 50:50, 70:30, and 90:10, the data were split into training and test sets. As performance evaluation metrics, accuracy scores, false negatives, and false positives were taken into consideration. They achieved the lowest false negative proportion with a Random Forest algorithm accuracy of 97.14 percent. The study discovered that using additional data during training increases accuracy.

According to Jitendra Kumar et al. in [2], they trained a number of classifiers on variables taken from the lexical structure of URLs, including logistic regression, naïve Bayesian classifier, random forest, decision tree, and KLearest neighbour. In order to address the issues of data imbalance, biassed training, variation, and overfitting, they developed the URL dataset. The Naive Bayes classifier turned out to be more useful because it had the greatest AUC value, even though all the classifiers had about the same AUC (area under the ROC curve). The Naive Bayes model had a 98 percent accuracy rate.

Using eight different algorithms across three different datasets, Mehmet Korkmaz et al. [3] suggest a machine learning-based phishing detection system. Logistic regression (LR), K-nearest neighbor, support vector machine, decision tree, naive Bayes, XGBoost, random forest, and artificial neural networks were among the methods utilised. It was discovered that the accuracy of the LR, SVM, and NB models is poor. They came to the conclusion that, because of their excellent accuracy and quick training times, RF and ANN algorithms can be applied.

Using Random Forest and Decision Tree, Mohammad Nazmul Alam et al. [4] devised a method to identify phishing attacks. With feature selection algorithms like Principal Component Analysis (PCA), a 32-feature Kaggle dataset was employed. The dataset's redundant useless or superfluous data is reduced by the select function. Prior to applying PCA, the suggested model uses the feature selection methods REF, Relief-F, IG, and GR. Random Forest got an accuracy of 97%. It avoided overfitting and had reduced variance.

Using the UCI dataset, Abdulhamit Subasi et al. published an intelligent phishing detection method in [5]. As classifiers to identify phishing sites, a variety of machine learning algorithms were utilized, including Artificial Neural Networks (ANN), K-Nearest Neighbor Networks (K-NN), Support Vector Machine (SVM), C4.5 Decision Tree, Random Forest (RF), and Rotational Forest (RoF). In terms of accuracy, F-measure, and AUC, the proposed RF classifier performed better than the others. RF was superior to other classifiers in terms of speed, reliability, and accuracy.

Caballet et al. (Qabajeh et al., 2018) recently used traditional automated phishing detection techniques. Traditional anti-phishing methods include raising awareness, educating users, holding regular training sessions and workshops, and using legal perspectives. Computer-based or automated anti-phishing approaches refer to list-based and machine learning-based techniques. More importantly, we compared the similarities, strengths, and weaknesses of these approaches from a user and performance perspective. According to this research, machine learning and rule extraction are well suited to combat phishing attacks. The limitations of this work are:This review is based on 67 research topics, and this research does not include deep learning techniques for detecting phishing sites.

Zuraiq Alkasassbeh (Zuraiq and Alkasassbeh, 2019) provided a comprehensive overview of current phishing detection methods. This research covers anti-phishing techniques including heuristic, content-based, and fuzzy rule-based approaches. In our research, we found that there are better ways to identify phishing sites. The background to the work is based on research conducted between 2013 and 2018. A drawback of this work is that it analyzed only 18 studies and did not include machine learning, list-based, or deep learning techniques for identifying phishing sites.

Athulya Praveen talked about various phishing attacks, the latest phishing tactics and anti-phishing strategies. Additionally, this article aims to raise awareness of phishing attacks and strategies for detecting phishing. According to this research, the best way to prevent phishing attacks is to educate users about the different types of phishing attacks. Users can choose the best security software tools or applications, such as anti-phishing browser extensions, to detect phishing attacks. The literature in this work is based on nine research topics. This research does not include deep learning techniques to identify phishing sites.

Basit et al. (Basit et al., 2020) report on AI-based phishing detection techniques. The authors used statistical phishing reports to investigate phishing attack failures and trends. In this article, he divides the anti-phishing ratings into four categories:

Machine learning, hybrid learning, scenario-based, deep learning. Research shows that machine learning techniques yield the best results compared to other approaches. This work is based on literature published in the last decade and analyzes only 21 research subjects.

Kathryn et al. (Kathrine et al., 2019) presented a framework for detecting and preventing various phishing attacks. According to this study, machine learning-based algorithms are effective in identifying true positives. The limitations of this study are: Only 11 studies were considered in the literature for this study, and this study does not include deep learning techniques to reduce phishing sites.

Korkmaz et al. We have proposed an overview paper on choosing features to use in URL-based phishing detection systems. The purpose of this study is to provide a general research resource for researchers studying website classification or web security. A limitation of this study is that the article only considered her five literature reviews.

Arshad et al. (Arshad et al., 2021) presented various phishing and anti-phishing techniques in their research. According to SLR, the most commonly used phishing techniques are phone phishing, email spoofing, phishing, and email spoofing. According to this study, the highest accuracy was achieved using machine learning techniques. This study is limited by the fact that it is based on only 20 studies.

Catarrhal etc. (Catal et al., 2022) undertook a systematic literature review that answered nine research questions. The main purpose of this study is to identify, evaluate, and integrate the results of deep learning approaches to phishing detection. According to this study, supervised ML algorithms were used in 42 of his 43 studies. The most commonly used algorithm was DNN, and the best performance was obtained from DNN and hybrid DL algorithms. This article deals exclusively with research related to deep learning for phishing detection.

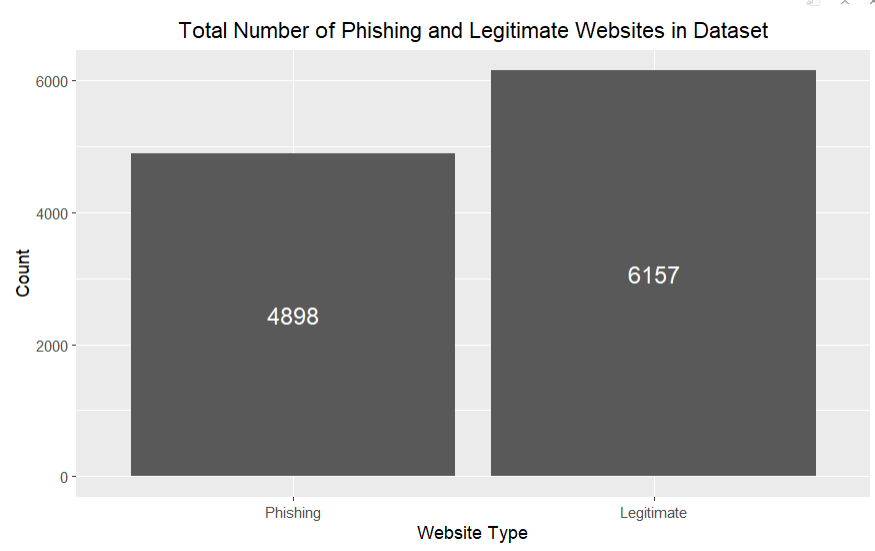
# III. PROPOSED METHODOLOGY

## The proposed solution model improves accuracy by using a feature selection algorithm. After filtering the original data to 30 features, the algorithm selects those that are critical to influence the prediction result. Therefore, irrelevant features do not affect the accuracy of the model and its prediction because they have few features. In addition, the prediction model is trained using ensemble learning, which uses multiple learning models. By using multiple models to make predictions, the results are not biased towards just one model. Thus, we show that the results of all models are used and calculated to determine the majority. For example, if most of the models indicate that a website is phishing, the final prediction of the set indicates that the website is indeed phishing.

## A. Dataset

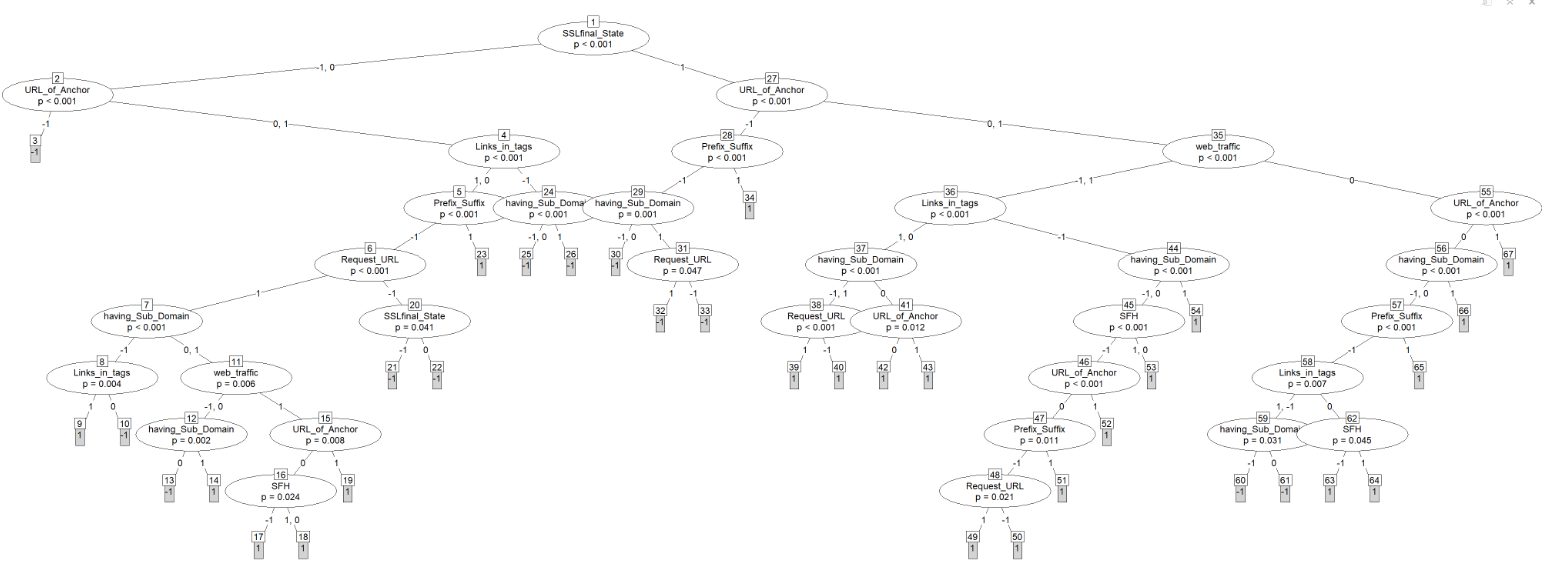
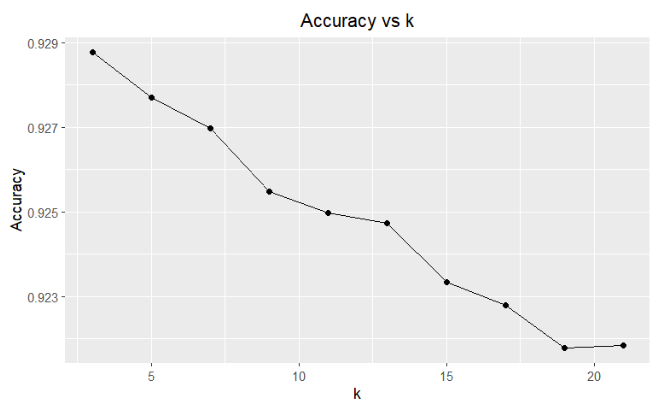
The data is collected from the PhishTank dataset, which is publicly available. The dataset consists of 11,055 instances with 30 attributes. We split the dataset 1:1. We used 50 percent for training and 50 percent for testing.

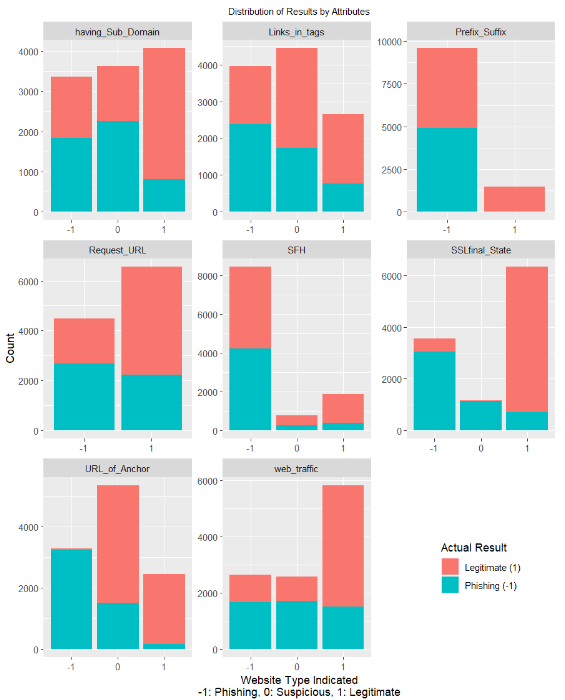
There are 31 attributes in this dataset. Attribute 31 is "Result", which is the result (indicating whether the site is actually phishing or legitimate). There are 11,055 observations in the database. All values ​​are categorical in nature with two/three levels indicating whether a website can be phishing (-1), suspicious (0) or legitimate (1).



*Fig 1.* Bar Graph of the Dataset used for Website Phishing. the Dataset Contains 45% Phishing and 55% Legitimate Websites.

## B. Modeling

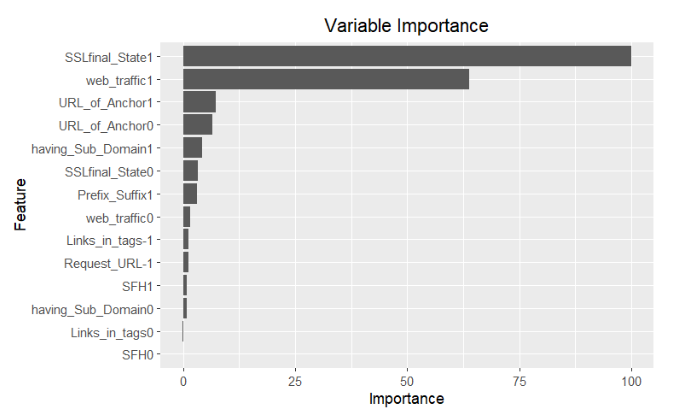
* SVM: Support vector machine (SVM) is a supervised learning algorithm that classifies data points into two parts and predicts new data points belonging to each part. It is suitable for linear binary classification with two labeled classes and the classifier is a hyperplane with N dimensions related to the number of features. The main idea of ​​this algorithm is to maximize the distance between the data point and the segmentation hyperplane.
* Decision tree: a decision tree is a machine learning algorithm and the model is a tree structure. Each node of the decision tree is a feature; each root represents a property value and an option, and the last node represents the result. A straighter tree structure generally has better performance. If the trees grow very deep, this is likely to cause overfitting of the training datasets.
* Random Forest: A random forest is a collection of decision trees for classification and regression. Random forests reduce the overfitting problem by ranking or averaging the output of individual trees in the training process. Therefore, random forests have higher accuracy than decision tree algorithms.
* k-NN: The k-nearest neighbors (k-NN) algorithm is a non-parametric classification algorithm that makes predictions by finding similar data points by calculating the distance between an object and its nearest neighbors. For continuous data, there are a number of methods to calculate the distance, the Euclidean distance and the Hamming distance for discrete values.
* XG Boost: It is a gradient-boosted decision tree implementation designed for speed and efficiency. Boosting is a general learning strategy that includes additional techniques to correct flaws in previously presented models. Models will be added in order until further improvements are not possible. It uses a gradient descent method to reduce losses by adding new models. This approach is used to ensure fast computation time and memory. That approach aimed to make the most of the available resources to train the model. The two main reasons to work with XG Boost are execution speed and model performance.



*Fig 1.* Distribution of the features of the data set

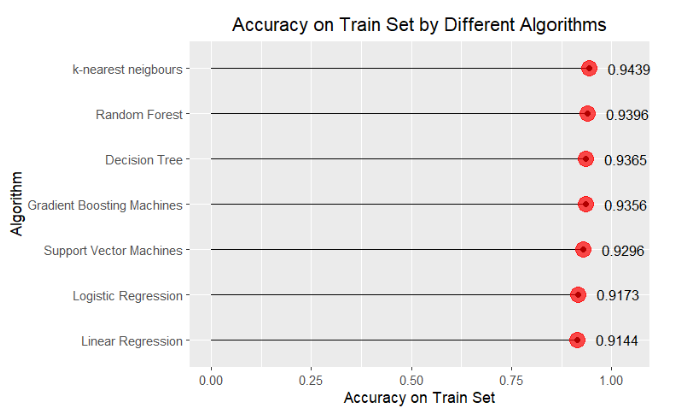
## C. Prediction Model (Ensemble Learning)

The prediction model reads a newly created CSV file containing only the output data and the selected features identified by the feature selection algorithm. We set SEED to 8888, where our model tests and train sizes are 0.2 and 0.8. The concept of ensemble learning is when two or more models are used to make a final prediction from the data. In this project we combined several models namely Gaussian Naive Bayes, Support Vector Machine, K Nearest Neighbor, Logistic Regression, Multilayer Perceptron NN, Gradient Boosting and Random Forest Classifiers. Finally, each of these models is evaluated individually based on their predictions. Predictions made are compared with experimental data. After that, all predictions for each model are listed. So, each model has its own list of results. The list is compared and then compared with the test data list to obtain accuracy scores of the combined models relative to the exact result.



# *Fig 1.* Importance given to the different predictors in this algorithm can be visualized through the plot

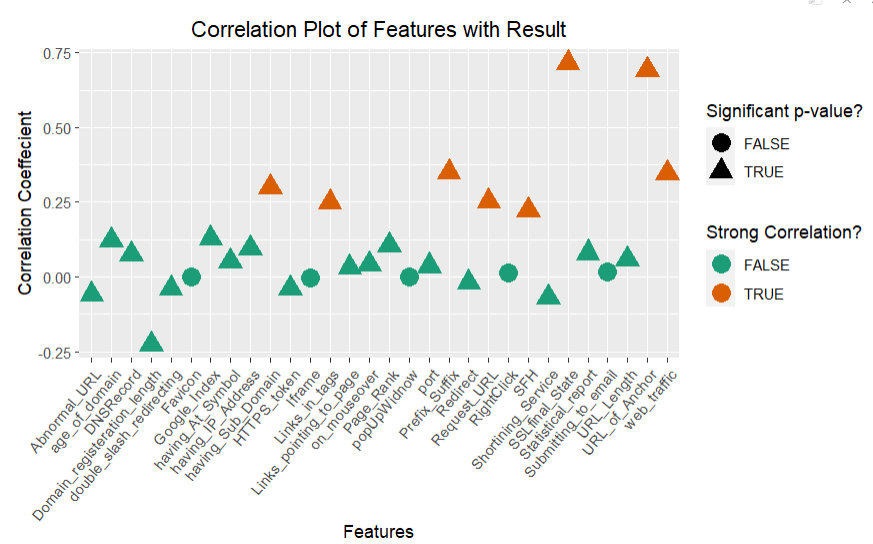
Ensembling is the last strategy used in this project to maximize the accuracy of the test suite. For the ensemble, four algorithms are used that are as accurate as possible in the train set (because the result of the test set is supposed to be unknown).



# *Fig 2.* Accuracy of different algorithms on train set

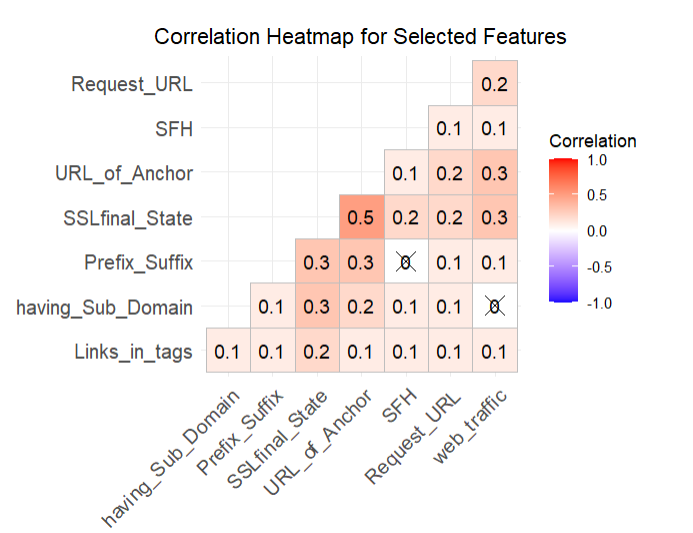
# IV. EXPERIMENTAL RESULT

Before analysis, the dataset is pre-processed to include only those attributes that show a strong correlation with the result (>=0.2) and an associated significant p-value (<=0.05). This allows us not only to select the best predictors, but also to avoid long computation times due to multiple predictors. Figure 1 shows the correlation coefficient of the 30 attributes in relation to the outcome.



# *Fig 1.* Correlation plot of features

In this graph, the 8 orange triangles represent features that show high correlation and indicate p-values. These are the variables selected from the phishing dataset for analysis and prediction. The final dataset is called phishing. It has 9 properties: *Prefix\_Suffix, having\_Sub\_domain, SSLfina\_State, Request\_URL, URL\_of\_Anchor, Links\_in\_tags, SFH, web\_traffic, Result*.

**

# *Fig 2.* Correlation heatmap for selected features

Performance evaluation was performed during the testing process. The original data would be divided into training data and test data, usually 80Evaluation of the behavior of the classifier in the test data had four statistical numbers: the number of correctly identified positive data points (TP), the number of correctly identified data. negative data points. (TN), the number of negative data points marked as positive by the classifier (FP) and the number of positive data points marked as negative by the model (FN).

There are several widely used metrics for evaluating performance. Classification accuracy is the ratio of correct predictions to total predictions:

*accuracy* = (*TP* + *TN*) */* (*TP* + *TN* + *FN* + *FP*) (1)

# TABLE I

CONFUSION MATRIX COMPARISON BETWEEN MODELS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classification | TP | FN | FP | TN |
| Logistic Regression | 2251 | 202 | 241 | 2834 |
| KNN | 2260 | 189 | 154 | 2925 |
| SVM | 2245 | 204 | 155 | 2924 |
| XG Boost | 2281 | 168 | 173 | 2906 |
| Decision Tree | 2243 | 206 | 151 | 2928 |
| Random Forest | 2256 | 193 | 143 | 2936 |

# TABLE 2

CONFUSION METRIC COMPARISON AMONG LEARNING MODELS

|  |  |  |
| --- | --- | --- |
| Classification | Accuracy | F1 Score |
| Logistic Regression | 91.99% | 0.9104 |
| KNN | 93.8% | 0.9294 |
| SVM | 93.51% | 0.9259 |
| XG Boost | 93.83% | 0.9304 |
| Decision Tree | 93.54% | 0.9262 |
| Random Forest | 93.92% | 0.9306 |

# TABLE I

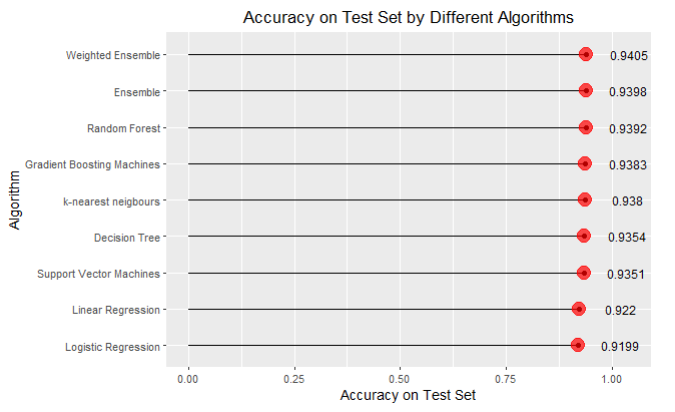
CONFUSION MATRIX FOR THE PROPOSED MODELS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classification | TP | FN | FP | TN |
| Ensemble Model | 2234 | 215 | 118 | 2961 |
| Weighted Ensemble Model | 2275 | 174 | 155 | 2924 |

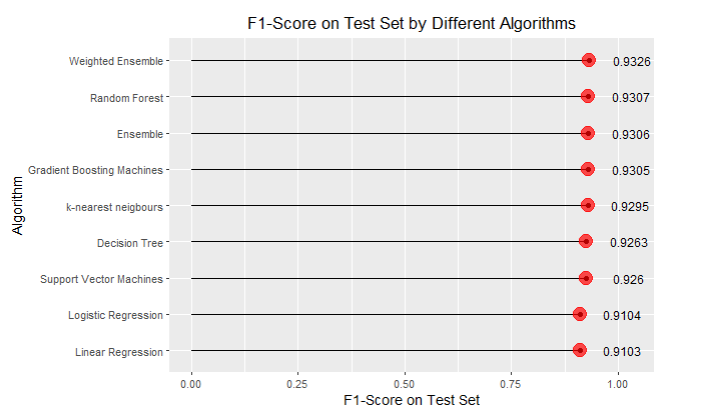
# TABLE I

CONFUSION MATRIX COMPARISON BETWEEN MODELS

|  |  |  |
| --- | --- | --- |
| Classification | Accuracy | F1 Score |
| Ensemble Model | 93.98% | 0.9306 |
| Weighted Ensemble Model | 94.05% | 0.9325 |

****

# *Fig 1.* Accuracy on Test Set by Different Algorithms

****

# *Fig 1.* F1-Score on Test Set by Different Algorithms

# V. CONCLUSION

The proposed method used four machine learning classifiers to achieve this and made a comparative study of the four algorithms. Accuracy results were also achieved. All four classifiers gave promising results, with the Random Forest Classifier being the best with 96.82 percent accuracy. Accuracy estimates may vary when using other datasets, and other algorithms may provide better accuracy than the random forest classifier. The random forest classifier is a general classifier and thus has high accuracy. This pattern can be deployed in real time to identify URLs as phishing or legitimate. Different algorithms were used during the project, giving different measures of accuracy, precision and recall. A weighted ensemble of different algorithms yielded 94.05 percent accuracy on the test set. Both sensitivity and specificity are high, indicating balanced activity.

At the same time, the limitations of the project must be acknowledged. First, the calculation time of some algorithms, especially ensemble models, is quite long (about 15-20 minutes). Second, using these algorithms requires users to check websites carefully and in detail, which makes the tool inconvenient to use. Finally, the accuracy of the algorithm is limited because there are no predictors that have a significant correlation with the outcome.

# REFERENCES

1. Rishikesh Mahajan, and Irfan Siddavatam, “Phishing website detection using machine learning algorithms,” International Journal of Computer Applications(0975-8887), vol. 181, no. 23, 2018
2. Jitendra Kumar, A. Santhanavijayan, B. Janet, Balaji Rajendran, and Bindhumadhava BS, “Phishing website classification and detection using machine learning,” International Conference on Computer Communication and Informatics(ICCCI), 2020
3. Mehmet Korkmaz, Ozgur Koray Sahingoz, Banu Diri, “Detection of phishing websites by using machine learning-based URL analysis,” 11nth International Conference on Computing, Communication and Networking Technologies(ICCCNT), 2020
4. Mohammad Nazmul Alam, Dhiman Sarma et al., “Phishing attacks detection using machine learning approach,” 3rd International Conference on Smart Systems and Inventive Technology(ICSSIT), 2020
5. Abdulhamit Subasi, Esraa Molah, Fatin Almkallawi, Touseef J. Chaudhery, “Intelligent phishing website detection using Random Forest classifier,” International Conference on Electrical and Computing Technologies and Applications(ICECTA), 2017
6. Qabajeh, Issa, Fadi Thabtah, and Francisco Chiclana. "A recent review of conventional vs. automated cybersecurity anti-phishing techniques." *Computer Science Review* 29 (2018): 44-55.
7. Zuraiq, AlMaha Abu, and Mouhammd Alkasassbeh. "Phishing detection approaches." In *2019 2nd International Conference on new Trends in Computing Sciences (ICTCS)*, pp. 1-6. IEEE, 2019.
8. Kunju, Merlin V., Esther Dainel, Heron Celestie Anthony, and Sonali Bhelwa. "Evaluation of phishing techniques based on machine learning." In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pp. 963-968. IEEE, 2019.
9. Benavides, Eduardo, Walter Fuertes, Sandra Sanchez, and Manuel Sanchez. "Classification of phishing attack solutions by employing deep learning techniques: A systematic literature review." *Developments and Advances in Defense and Security: Proceedings of MICRADS 2019* (2020): 51-64.
10. Athulya, A. A., and K. Praveen. "Towards the detection of phishing attacks." In *2020 4th international conference on trends in electronics and informatics (ICOEI)(48184)*, pp. 337-343. IEEE, 2020.
11. Basit, Abdul, Maham Zafar, Xuan Liu, Abdul Rehman Javed, Zunera Jalil, and Kashif Kifayat. "A comprehensive survey of AI-enabled phishing attacks detection techniques." *Telecommunication Systems* 76 (2021): 139-154.
12. Kathrine, G. Jaspher Willsie, Paradise Mercy Praise, A. Amrutha Rose, and Eligious C. Kalaivani. "Variants of phishing attacks and their detection techniques." In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 255-259. IEEE, 2019.
13. Korkmaz, Mehmet, Ozgur Koray Sahingoz, and Banu Diri. "Feature selections for the classification of webpages to detect phishing attacks: a survey." In *2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pp. 1-9. IEEE, 2020.
14. Arshad, Ayesha, Attique Ur Rehman, Sabeen Javaid, Tahir Muhammad Ali, Javed Anjum Sheikh, and Muhammad Azeem. "A systematic literature review on phishing and anti-phishing techniques." *arXiv preprint arXiv:2104.01255* (2021).
15. Catal, Cagatay, Görkem Giray, and Bedir Tekinerdogan. "Applications of deep learning for mobile malware detection: A systematic literature review." *Neural Computing and Applications* (2022): 1-26.