ISSN: 2278-4632
Vol-13, Issue-12, December 2023

Phishing Detection using Enhanced Multilayer Stacked Ensemble Learning Model

Vadla Dheeraj Kumar,
Student, Master of Computer Applications, Chaitanya Bharathi Institute of Technology(A), Hyderabad, Telangana, India, dheerajofficial3292@gmail.com

Ramesh Ponnala
Assistant Professor, Department of Master of Computer Applications, Chaitanya Bharathi Institute of Technology (A), Hyderabad, Telangana, India, pramesh_mca@cbit.ac.in

P. Krishna Prasad
Assistant Professor, Department of Master of Computer Applications, Chaitanya Bharathi Institute of Technology (A), Hyderabad, Telangana, India, pkrishnaprasad_mca@cbit.ac.in

Abstract:
Phishing attacks is a digital attack where fraudsters use false websites in order to trick users into giving important information, create a serious threat in the digital age. Anti-phishing strategies and technologies still exist, but these attacks are still always a worry. We have employed an enhanced multi-layered stacked ensemble learning model which performs EDA, Class balancing and outlier removal, Feature selection and finally uses multiple machine learning algorithms at different layers. The predictions from the algorithms in one layer are used as input in the next layer. Implementing this process can improve overall performance of the model. The model we used has detected the URLs of different websites with best accuracy. Additionally, it performed better than baseline models, showing significant improvements in accuracy and F-score metrics.

Keywords: Phishing, fraudsters, multi-layered stacked ensemble learning, estimators

I.INTRODUCTION

In order to fight cyber criminals and safeguard internet users, it is crucial to find phishing websites. Building strong barriers is essential because phishing attacks focus on innocent people by copying reputable websites. In this study, we provide an innovative strategy to address this problem by using an advanced machine learning algorithms and feature selection techniques. By selecting the essential features from the provided datasets, we try to enhance the performance of our model.

In [13] the authors have used a Multi-layer stacked ensemble learning model we are going to enhance it.

To do this, we will explore different feature selection techniques and examine how well they are able to isolate important features for phishing website identification. In order to further increase the precision and predictive strength of our model, we will look into the combination of feature selection techniques. We expect to increase the accuracy of detection and decrease the errors by combining these strategies.

By creating a powerful and accurate model for phishing website detection, our studyintends to advance cyber security. The suggested approach improves the precision ofpresent methods and offer valuable data for upcoming research projects. We work to improve online security and protect consumers from falling prey to these criminalpractices by dealing with the problems brought on by phishing attacks.

ISSN: 2278-4632
Vol-13, Issue-12, December 2023

II.LITERATURE SURVEY

Shatha Ghareeb et al [1] focused on finding the proper set of characteristics by using pre- processing techniques to the dataset. The behavior of each model's phishing detection accuracy in relation to each feature selection method is also examined in this study. A classification methodology is put out that determines whether a website is real or a phishing site. Logistic Regression, Random Forest, and an ensemble model comprising LR, RF, and XGBoost classifiers are used for this work.

Kishwar Sadaf et al [2] has evaluated the XGBoost and Catboost tree-based ensemble classifiers. Without hyper-parameter adjustment in this work, XGBoost and Catboost showed notable performance. Better results are produced when parameters are properly set to take full use of these classifiers. Both classifiers outperformed traditional classifiers in terms of performance. They noticed that XGBoost outperformed Catboost by a small margin.

Rabab Alayham Abbas Helmi et al [3] has utilized Agile Unified Process (AUP). Scott Ambler developed a well-liked methodology referred to as a hybrid modeling technique. AUP is the combination of Rational Unified Process (RUP) and Agile Methods (AM). AUP will consist of the following four steps: Inception, Elaboration, Construction, and Transition.

Somil Tyagi et al [4] the authors have employed a client-side framework in the form of a browser plugin that is suitable for all kinds of contemporary issues. The author has created a dataset using a model and an algorithm that gathers the features mostly used to find out phishing websites. For the execution phase, a Chrome extension written in JavaScript was created to collect the URL. For backend, a set with features was created and it is supplied to the classifiers for prediction. As a result, an automatic Chrome plug-in has been created that serves as a one-stop shop for identifying web URLs and classifying them as harmful or benign.

Basant Subba et al [5] the author has employed a ensemble-based architecture with three first-level classifiers and a meta-level classifier has been used by the author. Their methodology extracts distinct features from a given corpus of URLs.

Abdul Karim et al. [6] conducted tests and used machine learning algorithms, like naive Bayes, decision trees, linear regression, etc and a hybrid model combining LR, SVC, and DT with soft and hard voting, to achieve the best performance results. The LSD Ensemble model employs algorithms for grid search hyper parameter optimization and canopy feature selection with cross-fold validation.

Upendra Shetty DR et al [7] the author has used three ML algorithms Random Forest, LightGBM and XGBoost. Out of all, the random forest algorithm has given the best and most accurate results.

P.Chinnasamy et al [8] the authors utilized the Random forest, Support vector machine(SVM) and Genetic Algorithm. During their observation, it was noted that a genetic algorithm with a very low false positive rate achieved an accuracy of 94.73%. Additionally, it was found that the performance improves as the input training data increases.

Swarangi Uplenchwar et al [9] to identify phishing in text messages, the author employed PADSTM (phishing attack detection system for text messaging). This work's main contribution is its ability to identify phishing utilizing specific text message keywords, URL verification using a blacklist, and machine learning approaches. The best phishing attack detection is achieved with the proposed PADSTM by comparing the text message content to the blacklist of URLs prior to classification.
Mohammad Nazmul et al. [1 0], the author has used a machine learning-based method to detect phishing attacks. Several strategies were used to recognize phishing attacks. To analyze and choose

ISSN: 2278-4632
Vol-13, Issue-12, December 2023

datasets for classification and detection purposes, two well-known machine learning approaches, decision trees and random forests, were used. The components of the datasets were identified and categorized using principal component analysis (PCA). Decision trees (DT) and random forest (RF) approaches were used to classify websites. After that, a confusion matrix was created to evaluate how well these algorithms performed. Due to its capacity to address overfitting issues and lower variance, random forest was chosen over choice trees. The random forest model's accuracy rate was 97%.

III. METHODOLOGY

A. Phishing Data:

The dataset utilized in this study was obtained from Mendely [12] and consists of approximately 58,000 samples of phishing and legitimate data, with 111 features. Each URL within the dataset is segmented into a set of features indicating the legitimacy of the corresponding website. In the target variable, phishing samples are represented by 1, while legitimate samples are denoted by 0. This dataset is suitable for training machine learning algorithms and has a size of approximately 15MB.

B. Exploratory Data Analysis (EDA):

In this study, the exploratory data analysis (EDA) was done to understand more about details of the dataset. Initially commencing, the EDA process, we examined the dataset for general understanding. We examined the head of the dataset by using*df.head( )* method from pandas to observe a few initial rows to view the format and data structure.

By using various python modules and methods the dataset's dimensions are known to us, as shown in fig-1 dataset consists of 112 rows and nearly 58000 samples. We next looked at the characteristics along with their data types to further understand the parameters. By examining the data categories, such as numerical, categorical, or textual, we got to learn more about the various kinds of information present in the dataset. For use in further evaluation, we created summary statistics of the dataset. Include statistics like the mean, median, standard deviation, and quartiles for each feature. These statistics gave useful details about the main patterns, range, and distribution of data.

To identify the connections and patterns present in the dataset we produced a heat-map using the seaborn module of python. Using the heat-map we can find out how the different variables are correlated with each other, and dropped few highly correlated features which highlighted potential dependencies and connections.

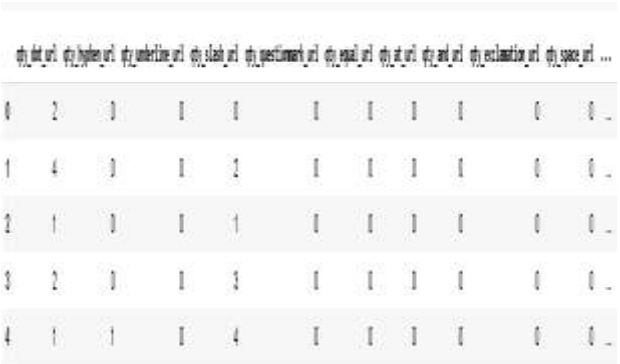

Figure-1: A look into the dataset using df.head()

ISSN: 2278-4632
Vol-13, Issue-12, December 2023

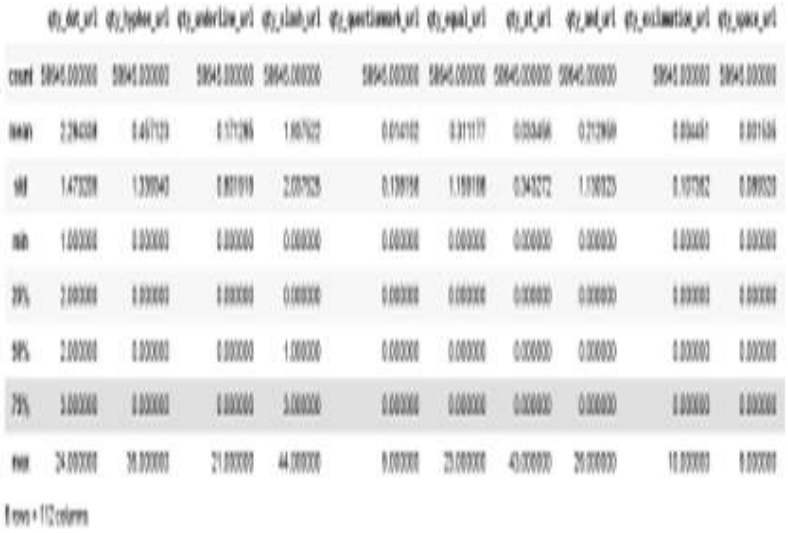

Figure-2: Statistical details of the dataset

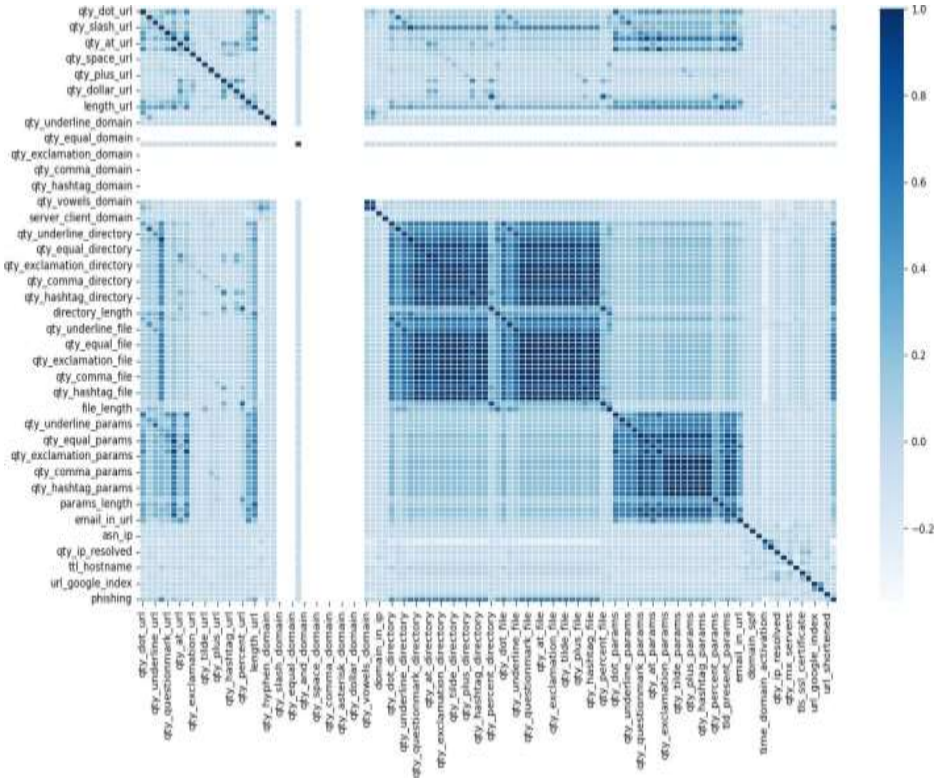

Figure-3: Heat map of dataset

We also checked for any missing data to verify that the dataset contained precise and full information. Along with that We did a duplicate check, locating and managing any duplicate records to protect data integrity.

Finally, we looked into the existence of outliers as part of the EDA procedure. By employing statistical methods (Using quartile ranges) and visualization tools, outliers were located and handled independently. Outliers were handled properly to make sure they did not unreasonably influence later studies

ISSN: 2278-4632
Vol-13, Issue-12, December 2023

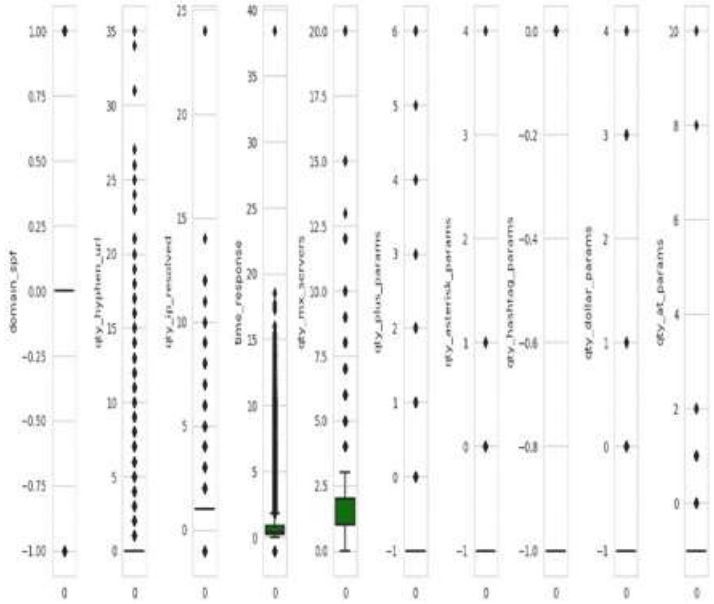

Figure-4: outliers present in few rows of dataset

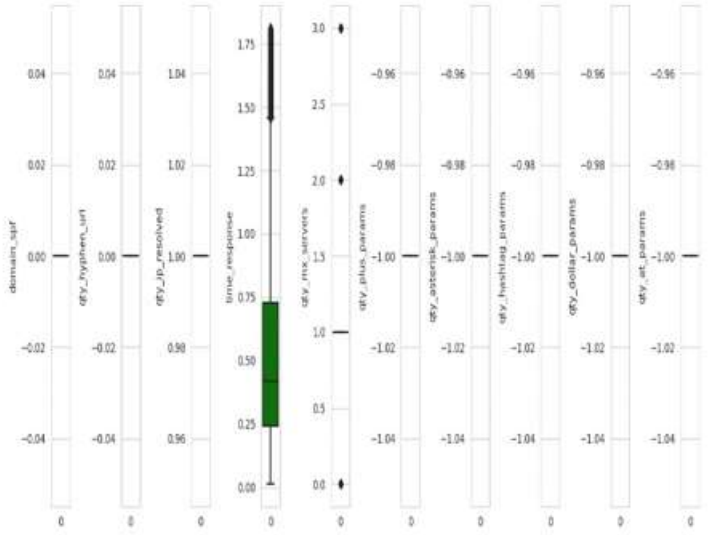

Figure-5: After handling the outliers

C. Class Balancing

The class balancing process is essential to make predictions unbiasedly [11]. When there is a class imbalance in any important feature, and if the number of samples in the various classes vary in considerable numbers, then the model performance may be skewed. In this work, we used the Synthetic Minority Over-sampling Technique (SMOTE) to evaluate the distribution of classes in our target variable and address any difficulties with class imbalance.

To understand the level of imbalance among the majority and minority classes we initially displayed the class distribution of the dataset using a bar graph. As we can observe from fig-6 there are around 30000 samples of class-1 and 28000 samples of class-0 in our target variable it means there is a bias in the target variable.

ISSN: 2278-4632
Vol-13, Issue-12, December 2023

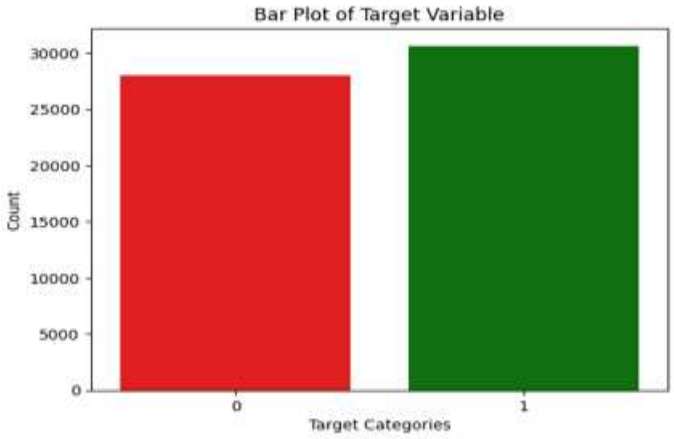

Figure-6: Data distribution of each class

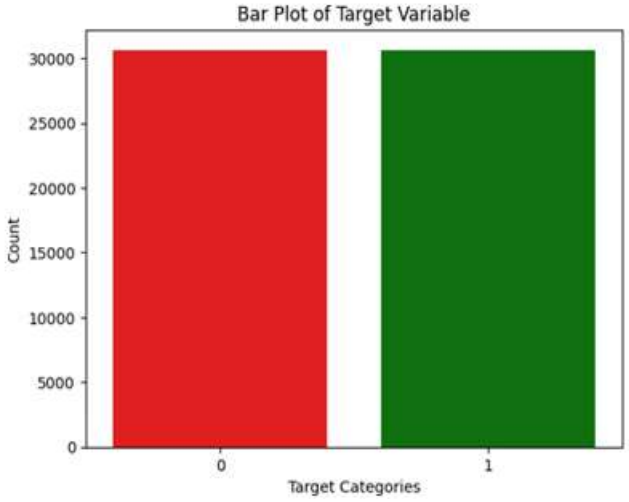

Figure-7: After applying the SMOTE algorithm

After applying the SMOTE we can observe that there is a proper split in the minority and majority classes through bar graph. For handling this issue, we used the SMOTE algorithm, which creates the artificial data samples for the minority class i.e. class-0, producing a more balanced dataset, fig-10 represents the same.

D. Feature Selection

This process helps us to select the most important and unique features, it helps in differentways by eliminating noise, reduce dimensionality, and focus on the most relevant aspects of the data. This process not only improves computational efficiency but also enhances the generalization capability of the model by eliminating irrelevant or redundant features.

In this study, we utilized various feature selection techniques to identify the informative features for our analysis. The chosen methods included random forest feature importance, L1-based feature selection, and correlation coefficient and PCA.

And by using those 68 features we created dataset. This refined dataset makes sure that we mostly focus on important features, which reduces noise and enhances the efficiency of our model.

Table-1: After feature selection

ISSN: 2278-4632
Vol-13, Issue-12, December 2023

Feature selectionstechniques	Selected features
PCA	33
Random forest featureimportance	38
L1 based feature selection	54
Correlation coefficient	94
Repeated features	67

IV. Enhanced Multi-Layer Stacked Ensemble Model

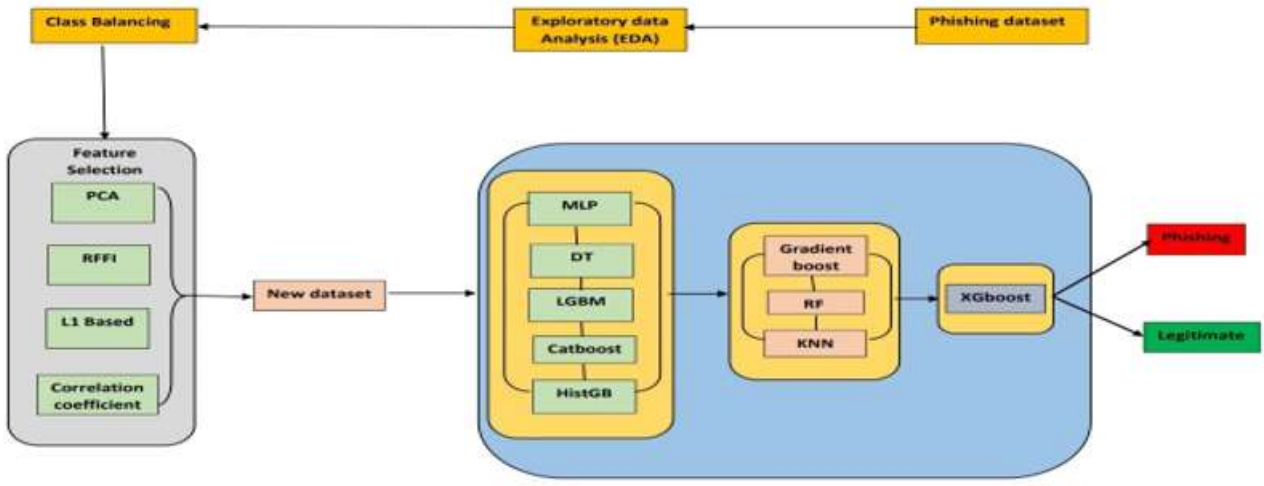

Figure-8: The Overall architecture of the Enhanced Multilayer Stacked Ensemble learning model

Three-layer architecture is used in the Enhanced multi-layer stacked ensemble learning model for phishing detection. In the layer-1, 5 different machine learning algorithms are used which include MLP classifier, Decision Tree, Histogram Gradient boosting, cat boost and Light-gradient boosting to trainour dataset. We assess each algorithm using various performance metrics such as accuracy, precision, recall, and F1-score.

Algorithms	Performance Metrics				
	Accuracy	Precision	Recall	F1-score	Average
MLP	0.937	0.945	0.935	0.940	0.939
DT	0.934	0.929	0.948	0.938	0.937
LGBM	0.954	0.953	0.962	0.957	0.956
Catboost	0.959	0.960	0.963	0.961	0.961
HistGB	0.941	0.935	0.955	0.945	0.944

Table-2: Performance metrics of layer 1

ISSN: 2278-4632
Vol-13, Issue-12, December 2023

Algorithms	Performance Metrics				
	Accuracy	Precision	Recall	F1-score	Average
Gradient boost	0.954	0.952	0.960	0.956	0.955
RandomForest	0.953	0.953	0.957	0.955	0.954
KNN	0.952	0.953	0.955	0.954	0.954

Table-3: Performance metrics of layer 2

Similarly in layer-2 we used three distinct machine learning algorithms they are Random Forest, Gradient boost and CNN. We feed the predictions made by the

previous layer as input to the present layer and train the algorithms using that predictions data. And as used in the previous layer. We assess each algorithm by using various performance metrics like accuracy, precision, recall, and F1-score.

Finally in layer-3 which is also called as meta layer, we use XGBoost, predictions of the previous layer are used to train the algorithm. The performance of the meta layer is considered as the performance of the model.

Algorithms	Performance Metrics				
	Accuracy	Precision	Recall	F1-score	Average
XGBoost	0.970	0.971	0.971	0.971	0.971

Table-4: Performance metrics of layer 3

V. Results

In the phishing detection using enhanced multi-layer stacked ensemble learning model, the final predictions are obtained from the meta-model. The meta-model combines the output of the second layer models and leverage their collective knowledge to make the ultimate decision on whether a website is a phishing attempt or not.

In our study, we indicated the presence for phishing attack as positive (1) and Legitimate as negative (0). And also, few others as

a. Number of (N): The total number of cases
b. Positive (P): The Phishing cases
c. Negative (N): The legitimate cases
d. True Positive (TP): The phishing case predicted as phishing
e. True Negative (TN): The legitimate case predicted as legitimate
f. False positive (FP): The legitimate case predicted as phishing
g. False negative (FN): The phishing case predicted as legitimate

ISSN: 2278-4632
Vol-13, Issue-12, December 2023

The metrices can be calculated using the below formulas:

$$Accuracy = \frac{N(TP+TN)}{N(Samples\ in\ dataset)} \quad (1)$$

$$Precision = \frac{N(TP)}{N(TP+FP)} \quad (2)$$

$$Recall = \frac{N(TP)}{N(TP+FN)} \quad (3)$$

$$F1\text{-}Score = 2* \frac{Precision*Recall}{Precision+Recall} \quad (4)$$

To evaluate the performance of the first two layers and also the final layer for detection, a comprehensive assessment using various valuation metrics is conducted. These metrics include accuracy, recall, precision, F1-score, and the ROC (Receiver Operating Characteristic) curve and also confusion matrix is used.

We can observe the above performance metrics of 3 Layers used in our model from Table-1, Table-2, Table-3 respectively. As said earlier, we have also used ROC curve and confusion matrix to visualize the performance. The Receiver operating curve (ROC Curve) helps us to find the binary outcome. It plots based on the true positive and false positive rate as shown in fig-8.

The confusion matrix is a matrix used to assess the performance of a trained machine learning model using a dataset. Figure 9 illustrates the confusion matrix, which is generated by evaluating the predictions made by the model and assessing the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

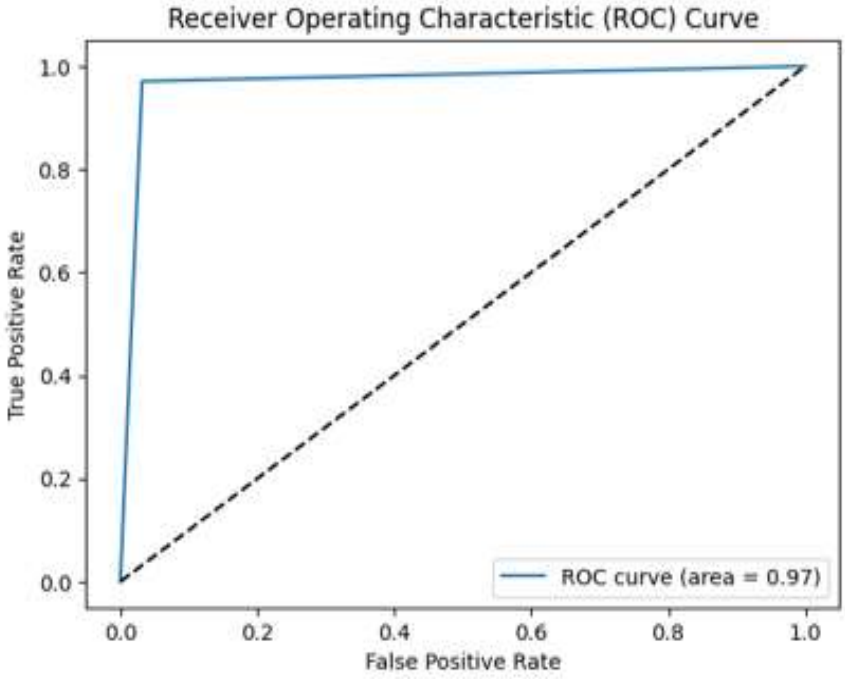

Figure-9: ROC curve of our predictions

ISSN: 2278-4632
Vol-13, Issue-12, December 2023

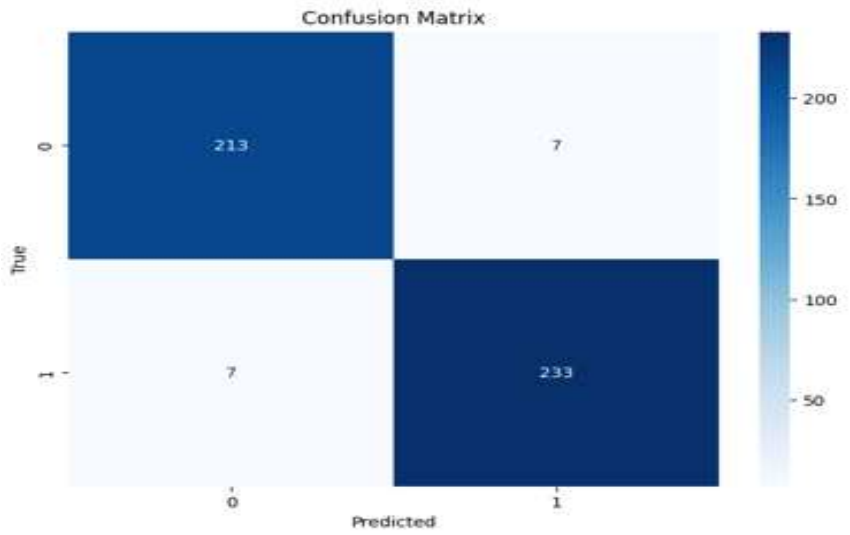

Figure-10: Confusion matrix of our model

	Performance Metrics				
	Accuracy	Precision	Recall	F1-score	Average
Our findings	0.970	0.971	0.971	0.971	0.971
Existing work	0.967	0.968	0.967	0.967	0.967

Figure-11: Comparison with the existing work

We can analyze our research with earlier works that has used the same dataset.

We took Lakshmana Rao K. Alabarige et al [13] for comparison as existing work. The findings are presented in Fig-10. We can observe that our model performed better than the existing work, with respectable accuracy, precision, and F1-score values of 97%, 97.10%, and 97.10% respectively.

VI. CONCLUSION

In our study we have used an enhanced multi-layer stacked ensemble learning model for phishing detection, where we have utilized the various methods mentioned in the EDA section for analyzing the dataset and partial removal of unwanted data. And then we have addressed the class balancing problem which is really necessary for accurate and unbiased predictions. And then we used the 4 feature selection methods to select the important features and created a new dataset with selected features. The new dataset is used to train the different machine learning algorithms in 3 different layers. Their performance is measured with various metrics and achieved accuracy, precision, and F1-score values of 97%, 97.10%, and 97.10% respectively. The average performance metric is 97.10%, which is considered very good. And also outperformed the existing work with a decent difference.

VII. REFERENCES

[1] Shatha Ghareeb , Mohamed Mahyoub and Jamila Mustafina “Analysis of Feature Selection and Phishing Website Classification Using Machine Learning”. 2023 15th International conference on Developments in eSystems Engineering (DeSE) ©2023 IEEE | DOI: 10.1109/DESE58274.2023.10099697

[2] Kishwar Sadaf “Phishing Website Detection using XGBoost and Catboost Classifiers” 023 International Conference on Smart Computing and Application (ICSCA) | 979-8-3503-4705-23660/23/$31.00©2023 IEEE | DOI: 10.1109/ICSCA57840.2023.10087829

[3] Rabab Alayham Abbas Helmi,Md. Gapar Md. Johar and Muhammad Alif Sazwan bin Mohd. Hafiz “Online Phishing Detection Using Machine Learning”.2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC) | 978-1-6654-7275-3/23/$31.00©2023IEEE|DOI: 10.1109/ICAISC56366.2023.10085377

[4] Somil Tyagi and Dr. Rajesh Kumar Tyagi ,Dr. Pushan Kumar Dutta,Dr. Priyanka Dubey “Next Generation Phishing Detection and Prevention System using Machine Learning ”.2023 1st International Conference on Advanced Innovations in Smart Cities(ICAISC)|978-1-6654-7275-3/23/$31.00©2023IEEE|DOI:10.1109/ICAISC56366.2023.10085529

[5] Basant Subba “A heterogeneous stacking ensemble-based security framework for detecting phishing attacks”.2023 National Conference on Communications (NCC) | 978-1-6654-5625-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/NCC56989.2023.10068026

[6] Abdul Karim, Mobeen Shahroz, Khabib Mustofa, Samir Brahim Belhaouri and S. Ramana Kumar Joga. ”Phishing Detection SystemThrough Hybrid Machine Learning Based on URL”. DOI 10.1109/ACCESS.2023.325

[7] Upendra Shetty D R,Anusha Patil and Mohana “Malicious URL Detection and Classification Analysis using Machine Learning Models”.2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT) | 978-1-6654-7451-1/23/$31.00©2023 IEEE | DOI: 10.1109/IDCIoT56793.2023.10053422

[8] P.Chinnasamy, N.Kumaresan, R.Selvaraj, S. Dhanasekaran, K.Ramprathap, Sruthi Boddu ”An Efficient Phishing Attack Detection using Machine Learning Algorithms ”.2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC) | 978-1-6654-6109-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ASSIC55218.2022.10088399

[9] Swarangi Uplenchwar,Varsha Sawant,Prajakta Surve,Shilpa Deshpande,Supriya Kelkar “Phishing Attack Detection on Text Messages Using Machine Learning Techniques”.2022 IEEE Pune Section International Conference (PuneCon) | 978-1- 6654-9897-/22/$31.00©2022IEEE|DOI:10.1109/PUNECON55413.2022.1001487

[10] Mohammad Nazmul Alam,Dhiman Sarma,Farzana Firoz Lima,Ishita Saha,Rubaiath-E-Ulfath and Sohrab Hossain “Phishing Attacks Detection using Machine Learning Approach”.The Third International Conference on Smart Systems and Inventive Technology (ICSSIT 2020) IEEE Xplore Part Number: CFP20P17-ART; ISBN: 978-1-7281-5821-1

[11] R. Ponnala and C. R. K. Reddy, "Hybrid Model to Address Class Imbalance Problems in Software Defect Prediction using Advanced Computing Technique," 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023, pp. 1115-1122, doi: 10.1109/ICAAIC56838.2023.10141379.

[12] G.Vrbancic,‘‘Phishingwebsitesdataset,’’MendeleyData,vol.1,2020.[Online].Available:https://data.mendeley.com/datasets/72ptz43s9v/1

[13] Lakshmana Rao Kalabarige, Routhu Srinivasa Rao, Ajith Abraham and Lubna Abdelkareim Gabralla “Multilayer Stacked Ensemble Learning Model to Detect Phishing Websites” Digital Object Identifier 10.1109/ACCESS.2022.319467