iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** X **Month of publication:** October 2022

DOI: https://doi.org/10.22214/ijraset.2022.47028

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue X Oct 2022- Available at www.ijraset.com

Credit Card Fraud Detection Using Hybrid Machine Learning Algorithms

Parth Chaturvedi[1], Suraj Mishra[2], Suraj Agrawal[3]
[1, 2, 3]*Department of Computer Science & Engineering, Maulana Azad National Institute of Technology, Bhopal, India*

Abstract: ***Credit cards have a high frequency of transactions taking place daily. There are approximately 36.4 percent fraud related to commercial cards which include credit card, debit card, etc. In 2022 there are 64 million people who use credit cards to initiate the transaction, therefore they are also prone to card fraud. To evaluate the model efficacy, a publicly available credit card data set is used. Investigation on different hybrid algorithms for the given dataset were studied. This research investigates seven different hybrid machine learning models to detect fraudulent activities. Our findings indicated that the hybrid model of Adaboost and LGBM is the most efficient model due to its high performance***.
Keywords: credit card fraud, machine learning, dataset, transactions, Classification, hybrid

I. INTRODUCTION

There is a rapid growth in the frequency of credit card transactions which has led to a considerable rise in fraudulent activities. Credit card fraud is a broad term for theft and fraud committed using a credit card as a fraudulent source of funds in each transaction. This problem is particularly challenging from the perspective of learning, as it is characterized by various factors such as class imbalance. The number of valid transactions far outnumber fraudulent ones. Also, the transaction patterns often change their statistical properties over the course of time. This is a very relevant problem that demands the attention of communities such as machine learning and data science where the solution to this problem can be automated. The dataset is highly unbalanced, the positive class (frauds) account for 0.17% of all transactions. Most banks and financial firms use rule-based systems, in which an expert will use historical fraud data to define a set of rules, and a system will raise an alarm if a new transaction match one of the rules [1,2]. The main limitations of this manual process are that it is reactive, lacks flexibility and consistency as well as the fact that it is time-consuming [3]. Oftentimes, these datasets may have many attributes that could have a negative impact on the performance of the classifiers during the training process. To solve the issue of a high feature dimension space, we implement principal component analysis for dimensionality reduction.

Machine learning is the science of getting computers to learn without being explicitly programmed [4]. It has been commonly used in a wide range of disciplines such as; Chemistry, Bioinformatics, Manufacturing industries, the Medical Field, Biology and in Finance. For fraud detection, machine learning is mainly used to help organizations and financial institutions detect fraudulent transactions. However, fraud detection can pose a challenge for machine learning for several reasons such as the distribution of the data is highly imbalanced as the number of fraudulent transactions is very small, the data is continually evolving over time and the lack of real-world dataset due to privacy concerns.

In this paper, we will focus on credit card fraud and its detection measures. Our research focuses on the application of the following supervised ML algorithms for credit card fraud detection: Random Forest Classifier, Adaptive Boost Classifier, eXtreme Gradient Boosting Classifier, LightGBM model.

The main aim of this paper is to implement the above-mentioned models on a data set in order to determine the most accurate and effective algorithm for detecting credit card fraud.

II. LITERATURE SURVEY

It is imperative for any banking or financial institution that issues credit and debit cards to put in place an effective measure to detect any cases of fraudulent transactions. Some of the notable methods identified to help detect fraud in credit cards include Random Forest Classifier, Adaptive Boost Classifier, eXtreme Gradient Boosting Classifier, LightGBM mode.

Random Forest Classifier: Random Forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Adaptive Boost Classifier: An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

eXtreme Gradient Boosting Classifier: Extreme Gradient Boosting is a tree-based algorithm, which sits under the supervised branch of Machine Learning. It can be used for both classification and regression problems.

LightGBM model: LightGBM is a gradient boosting framework based on decision trees to increase the efficiency of the model and reduce memory usage.

Fraud detection has received much attention in the past decade. A growing body of literature has proposed hybrid approaches to enhance fraud detection.

A hybrid model for improving fraud detection accuracy by combining supervised and unsupervised methods was presented by the authors in [5]. They displayed several criteria for calculating outlier scores at various levels of granularity (from high granular card-specific outlier scores to low granular global outlier scores). Then, they evaluated their added value in terms of precision once integrated as characteristics in a supervised learning approach. Unfortunately, in terms of local and global methods, the results are unconvincing. However, the model provides a more considerable result in terms of Area Under the Precision–Recall Curve.

More recent research was conducted to develop a hybrid model to detect credit card fraud using credit card datasets and utilizing machine learning classifiers with LR, Gradient Boosting (GB), RF and voting classier [6]. The author found that RF and GB gave maximum detection rates of 99.99 percent.

Furthermore, a twelve-machine learning algorithm in conjunction with the AdaBoost and majority voting methods using a real credit card dataset obtained from a financial institution has been used to investigate the performance of the used classifiers [7]. Their result for the highest Matthews correlation coefficient (MCC) score was 0.823, which was obtained by a majority of the votes. However, when using AdaBoost and majority voting procedures, a perfect MCC score of 1 was obtained. To further assess the hybrid models, noise ranging from 10 percent to 30 percent was added to the data samples. When 30 percent noise was added to the data set, the majority voting procedure produced the best MCC score of 0.942. Therefore, the authors reported that the majority vote method performs well in the presence of noise.

Although all the mentioned studies were concerned with fraud detection, different algorithms were used depending on the nature of the dataset. As evident from previous efforts, various approaches were used to detect fraudulent transactions in the financial sector especially the credit card domain either using a single machine learning algorithm or hybrid models. However, these hybrid models only utilized a single model without consideration of the performance of other models to confirm that the selected model is the optimal choice for the chosen dataset. Therefore, this might lead to inaccurate results and a lack of generalization for the proposed model. Therefore, a comprehensive comparison of hybrid models using the same datasets is still needed to understand the relative performance of the proposed technique. The aim of this paper is to find an apt hybrid model to resolve the above-mentioned problem statement.

III.MATERIALS AND METHODOLOGY

A. Data Collection

The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions [8].

Initially, the original dataset included many transactions, thus, to avoid computational cost and model training delay, a smaller random subset sample was created from the original dataset to prove our concept of the study (POC). However, the POC dataset still suffers from the same imbalance ratio as the original dataset.

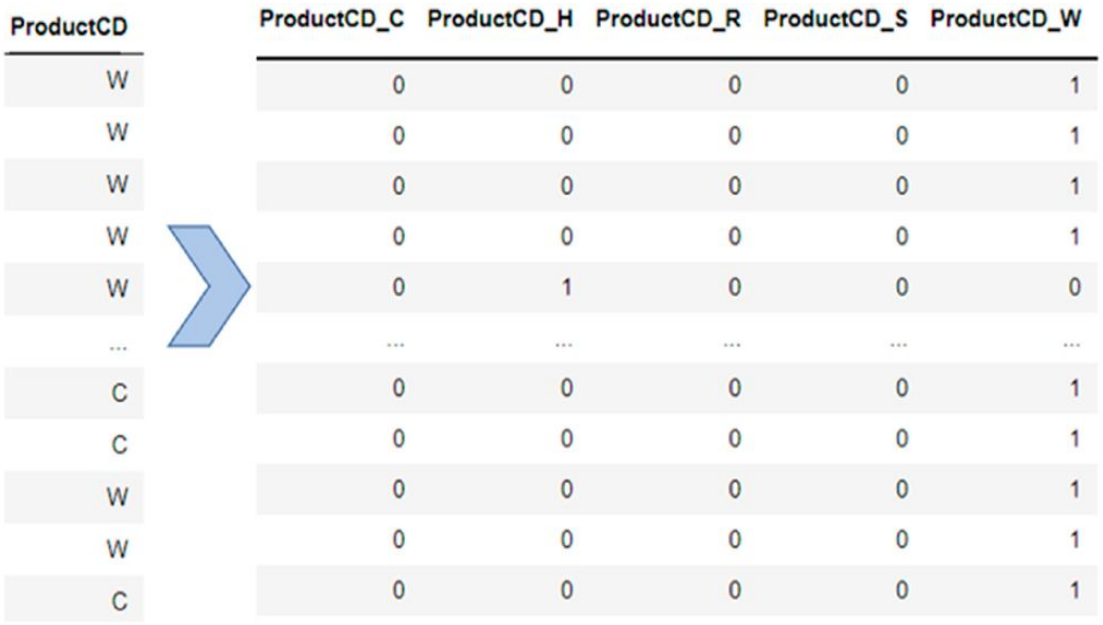
B. Data Preparation

In any predictive analytics study, this phase is the most critical one as it defines the success of the study. Real-world datasets are well-known to be chaotic because they contain a massive number of outliers, missing values, irregular cardinality, etc. Such phenomena can lead to the failure of the research if not handled correctly. In this paper, our preparation includes handling missing values, transforming categorical features, feature scaling, feature selection.

1) *Missing and Null Values:* We first read the data and scan for missing and null values. This step is critical to data preparation as such values can tamper with the accuracy achieved by the models. In this data set it was noticed that there were about values 45 percent values that were missing, these missing values can lead to the creation of an unpredictive or inaccurate model.

Therefore, if the missing values reaches above 60 percent, then the features were removed as the values stored in the feature would be insufficient and lead to no contribution to the predictive model. The rest of the features having less than 50 percent missing data were imputed with mode for categorical data and median for numerical features.

2) *Encoding Categorical Features:* Most machine learning algorithms require the input and output features to be in a numerical format. This implies that categorical data must be converted to numbers before the development of a prediction model. After removing features with high missing values only 15 features remained. Ten of them contained only two levels of cardinality (with attributes true and false), they were replaced by 0 and 1, respectively. For the remaining dataset having greater than 2 levels, one hot encoding technique was used. In one hot encoding each attribute value is converted into new categorical column and given a binary value. An illustrative representation is given in the figure (1). In this example, the categorical attribute Product CD with the categories C, H, R, S, and other as W is encoded with 5-dimensional feature vectors [1, 0, 0, 0, 0], [0, 1, 0, 0, 0], [0, 0, 1, 0, 0], [0, 0, 0, 1, 0] and [0, 0, 0, 0, 1]. This technique is implemented using get dummies in Panda's software library.



ProductCD	ProductCD_C	ProductCD_H	ProductCD_R	ProductCD_S	ProductCD_W
W	0	0	0	0	1
W	0	0	0	0	1
W	0	0	0	0	1
W	0	0	0	0	1
W	0	1	0	0	0
...	...	...	...	...	...
C	0	0	0	0	1
C	0	0	0	0	1
W	0	0	0	0	1
W	0	0	0	0	1
C	0	0	0	0	1

Figure 1. One-hot encoding.

3) *Imbalance Data:* Imbalanced data refers to those types of datasets where the target class has an uneven distribution of observations, i.e., one class label has a very high number of observations and the other has a very low number of observations. In this dataset we see that only 0.172% of the transactions are fraudulent, that means the data is highly unbalanced.

4) *Feature Scaling:* Most features in the world have varied quantitative units, this can lead to one feature naturally dominating others. As a result, these features need to be scaled to eliminate the impact of various quantitative units. In this research, MinMaxScaler technique was used to rescale features between 0 and 1. This technique benefits due its robustness against outliers and uses statistical techniques that does not affect variance of data (Equation 1). As shown in the below equation x indicates the original value, x' is the scaled value, max is the upper bound of the original value and min is the lower bound. For data with numerous 0 values, MinMaxScaler preserves the sparsity of the input data.

$$x' = (x - \min(x))/\max(x) - \min(x) \quad (1)$$

5) *Feature Selection:* Feature selection works by removing irrelevant and unnecessary values from a feature. The most popular techniques include filter and wrapper approaches, but each of these methods has its own disadvantages as well. In the filter approach, the selection of features is independent of any machine learning algorithms. Instead, features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable. The downside of this approach is may bot have the ability to select the best features. In wrapper selection, the feature selection algorithm exits as a wrapper around the predictive model algorithm and uses the same model to select best features. The disadvantage in this technique is that it is prone overfitting and computationally expensive. These methods have their demerits due to which an alternative approach, which is less exhaustive and has fewer shortcomings needs to be used. In this research, a hybrid feature selection model is used, which combines both filter and wrapper techniques. As shown in figure [2], some group features showed high correlation results.

Accordingly, strongly positive features that corelated with each other were removed because they would make no significant contribution to the prediction model, prevent overfitting and doing so would save computational resources. The result obtained was that there was no co-relation between features. There were only some certain feature co-relation present and **Time** (inverse correlation with **V3**) and **Amount** (direct correlation with **V7** and **V20**, inverse correlation with **V1** and **V5**).

6) *Data Resampling:* The data in this research is highly imbalanced, this can lead to the assumption of equal distribution for both minority and majority classes and which can provide misinterpreted results and poor predictive modelling performance. Also, the imbalance problem appears to be related to learning too few minorities class examples in the presence of other factors, such as overlapping. The solution to this problem is using the Synthetic Minority Oversampling technique with edited nearest neighbours [SMOTE-ENN]. In this technique SMOTE is applied over the oversampling phase followed by ENN as a data cleaning method to eliminate the overlapping between classes to produce better defined results. Table [1] illustrates the number of observations for each class (fraud and non-fraud) before and after SMOTE-ENN. Figure [3] displays the sample data set before pre-processing phase.

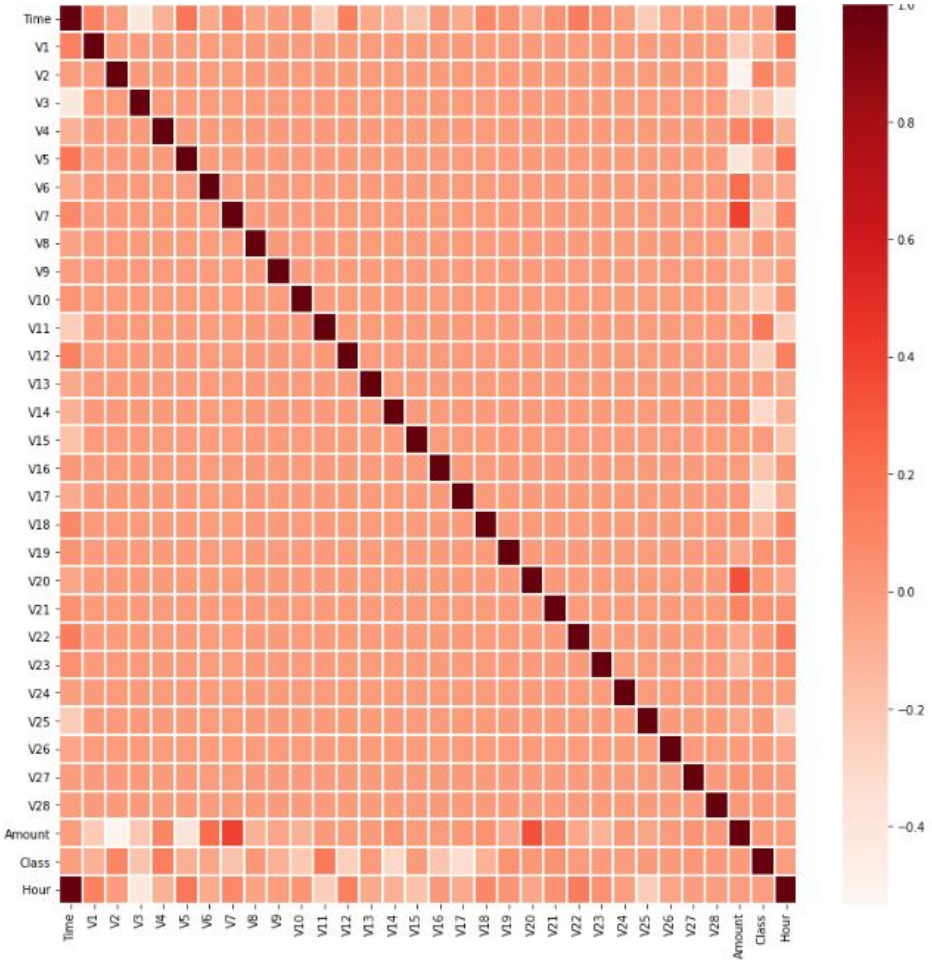


Figure 2. Co-relation Based Filter

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
2	0	-1.35981	-0.07278	2.536347	1.378155	-0.33832	0.462388	0.239599	0.098698	0.363787	0.090794	-0.5516	-0.6178	-0.99139	-0.31117
3	0	1.191857	0.266151	0.16648	0.448154	0.060018	-0.08236	-0.0788	0.085102	-0.25543	-0.16697	1.612727	1.065235	0.489095	-0.14377
4	1	-1.35835	-1.34016	1.773209	0.37978	-0.5032	1.800499	0.791461	0.247676	-1.51465	0.207643	0.624501	0.066084	0.717293	-0.16595
5	1	-0.96627	-0.18523	1.792993	-0.86329	-0.01031	1.247203	0.237609	0.377436	-1.38702	-0.05495	-0.22649	0.178228	0.507757	-0.28792
6	2	-1.15823	0.877737	1.548718	0.403034	-0.40719	0.095921	0.592941	-0.27053	0.817739	0.753074	-0.82284	0.538196	1.345852	-1.11967
7	2	-0.42597	0.960523	1.141109	-0.16825	0.420987	-0.02973	0.476201	0.260314	-0.56867	-0.37141	1.341262	0.359894	-0.35809	-0.13713
8	4	1.229658	0.141004	0.045371	1.202613	0.191881	0.272708	-0.00516	0.081213	0.46496	-0.09925	-1.41691	-0.15383	-0.75106	0.167372
9	7	-0.64427	1.417964	1.07438	-0.4922	0.948934	0.428118	1.120631	-3.80786	0.615375	1.249376	-0.61947	0.291474	1.757964	-1.32387

Figure 3. Sample of data before pre-processing

	No. of observed Fraud Cases	No. of observed non-Fraud cases
Before SMOTE-ENN	1,157	32,060
After SMOTE-ENN	28,689	27,155

Table 1. The number of observations for each class (fraud and non-fraud) before and after SMOTE-ENN.

IV. MODEL DEVELOPMENT

Different machine learning classification techniques have been applied to detect fraudulent transactions as discussed earlier. Yet, there is no optimal algorithm for a specific problem. Therefore, eight different linear and nonlinear algorithms were selected from the literature as they indicated promising performance in the context of fraud detection, including LR, RF, DT, XGBOOST, SVM, NB, Adaboost and LGBM. The development phase of the hybrid models is divided into two phases. In the first phase, a single baseline machine learning classification model was developed using the following eight machine learning algorithms: LR, RF, DT, XGBOOST, SVM, NB, Adaboost and LGBM where their performance was investigated. Even though algorithm parameter tuning can be useful, a consideration of default parameters is more common in practice. The need for considerable work and time for tuning can dissuade people from implementing the step and could also lead to issues of overfitting for specific datasets. Appropriately, there were no deliberate efforts to fine-tune the parameters of the methods.

Subsequently, in the second phase of the proposed model, the algorithm with the best performance from the previous experiment based on the highest Area Under the Receiver Operating Characteristic (AUROC) metric served as a baseline model and was used to train the rest of the seven algorithms. The correctly classified data points—true positive (TP) and true negative (TN)—that are generated by the single machine learning algorithm with the highest performance in level one was used to train the hybrid models separately. Consequently, seven hybrid models were constructed and are as follows: The best baseline single model + LR; The best baseline single model + RF; The best baseline single model + DT; The best baseline single model + XGBOOST; The best baseline single model + SVM; The best baseline single model + NB; The best baseline single model + LGBM.

The utilization of the algorithms, which is derived by its score in AUROCmetric for detecting the correct classes, will assist the hybrid models to precisely detect fraudulent and legitimate activities. The proposed hybrid models will be compared with state-of-the-art algorithms to check their effectiveness. Figure [4] presents the details of the proposed flowchart. According to the No Free Lunch Theorem, no single model or algorithm can handle all classification problems. Furthermore, each different algorithm has its advantages and disadvantages as illustrated in Table [2]. Consequently, the combination of several algorithms exploits the weaknesses of one, such as overfitting. This combination of several algorithms will be beneficial if the algorithms are substantially different from each other. Combining these algorithms together will result in optimal performance and help to overcome the limitation of a single classifier and therefore enhance the detection of fraudulent cases. This distinction could be in the algorithm, or the data used in that algorithm.

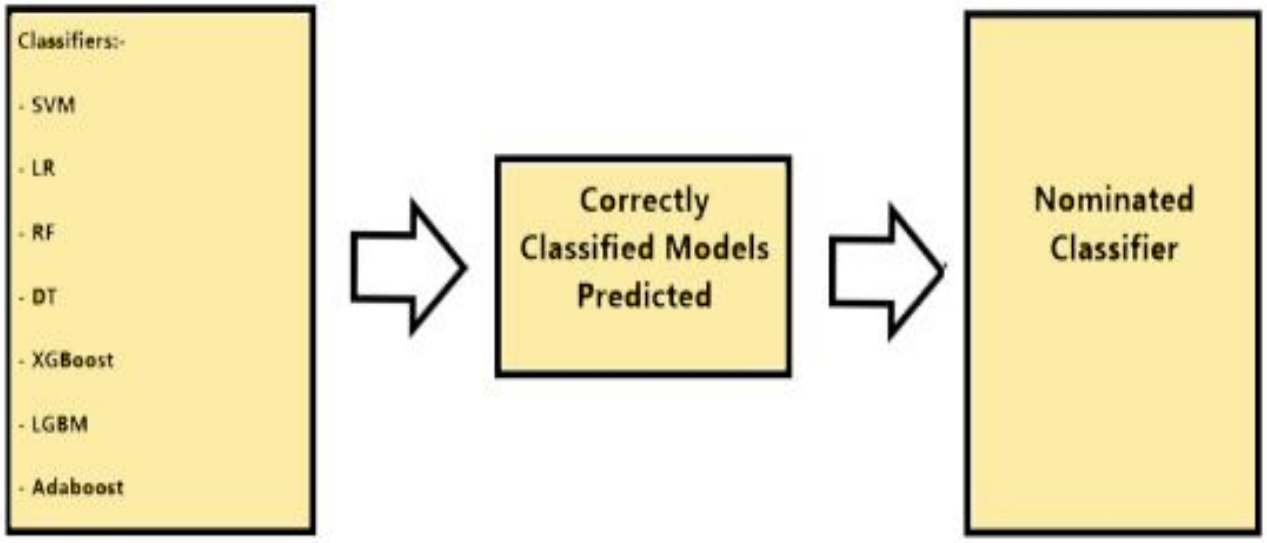


Figure 4. Flowchart of the proposed model

Algorithm	Strength	Weakness
LR	Linear regression is straightforward to understand and explain, and can be regularized to avoid overfitting	They are not naturally flexible enough to capture more complex patterns, and adding the right interaction terms can be tricky and time-consuming.
DT	Decision trees can learn non-linear relationships, and are fairly robust to outliers	Unconstrained, individual trees are prone to overfitting because they can keep branching until they memorize the training data
NB	They are easy to implement and can scale with your dataset	Due to their sheer simplicity, NB models are often beaten by models properly trained and tuned using the previous algorithms listed.
SVM	They are also fairly robust against overfitting, especially in high-dimensional space.	SVM's are memory intensive, trickier to tune due to the importance of picking the right kernel, and don't scale well to larger datasets
LGBM	High training speed, performance and low memory utilization	Has high chance of overfitting
Adaboost	Ease of use, less parameter tweaking and less prone to overfitting	Sensitive to outliers and noisiness

Table 2. Comparison between different algorithms and their strengths and weaknesses

V. MODEL EVALUATION

Stratified k-folds validation was applied to measure the efficiency of the proposed model in which it tries to ensure that both classes (fraud and non-fraud) are roughly evenly distributed in each fold. In this research, we randomly divided the validation set into five equal-sized subsets. At each phase of validation, a subset of 25 percent was set aside as the validation dataset to assess the result of the proposed method, while the remaining four subsets that compromise 75 percent were used as a training set. We employed various performance evaluation metrics that have been seen in this paper. It should be noted that the accuracy score is inadequate in the case of a highly imbalanced dataset. Consequently, different criteria are needed to evaluate the model's performance such as AUROC, AUC-PR, Type-I error, Type-II error F1-measure, recall, precision, misclassification rate and Specificity or True Negative Rate (TNR). The terms used in the applied metrics are defined as follows

1) *True Positive (TP):* the number of correctly classified data as fraudulent credit card transactions.
2) *True Negative (TN):* the number of correctly classified data as legitimate credit card transactions.
3) *False Positive (FP):* the number of legitimate credit card transactions classified as fraudulent.
4) *False Negative (FN):* the number of fraudulent credit card transactions classified as legitimate.

Although there is no ultimate individual evaluation metric that can be used to evaluate both negative and positive classes, it was decided that the best overall performance metric for the imbalanced fraud dataset was to use AUROC. AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. The ROC curve presents a compromise between the true positive rate (TPR) and false-positive rate (FPR) and it is calculated as follows:

$$TPR = TP / (TP + FN) \quad (2)$$

$$FPR = FP / (FP + TN) \quad (3)$$

The AUROC values vary from 0 to 1 where 1 represents desired prediction, 0 represents the worst prediction performance and 0.5 represents random performance. The advantage of AUROC is that it does not require a specific cut off value. Additionally, it provides valuable information on whether the model is indeed obtaining knowledge from the data or just providing random values. Additionally, it can be more readily understood compared with the numerical methods due to its visual representation method. In addition, recall and precision were also suitable to evaluate the predictive model to check if it is capable to identify fraudulent transactions accurately. A recall which is equivalent to TPR and sensitivity is the proportion of real credit card transactions predicted correctly by the model as fraudulent cases. On the other hand, precision is the proportion of predicted observations such as fraudulent credit card transactions predicted by the model that are accurate.

If the recall is equal to 1, it indicates that all the credit card transactions are classified as fraudulent. Conversely, precision will be low as many non-fraudulent credit card transactions will be falsely classified as fraud. Thus, performance measurements such as the F1-measure give equal consideration to precision and recall. Moreover, the misclassification rate or error rate will be used which determines the percentage of misclassified observations by the model. These measures were defined as follows.

$$\text{Recall} = TP / (TP + FN) \quad (4)$$

$$\text{Precision} = TP / (TP + FP) \quad (5)$$

$$F - \text{measure} = 2\,(\text{Precision} * \text{Recall})\,(\text{Precision} + \text{Recall}) \quad (6)$$

$$\text{Misclassification Rate} = (FP + FN) / (TP + TN + FP + FN) \quad (7)$$

False cases that are predicted as possible fraud are costly in fraud detection, as they are taken for further investigation. The precise detection of cases of fraud helps to avoid costs resulting from missing a fraudulent activity (Type-I error), which is usually greater than falsely alleging fraud (Type-II error). Therefore, a Type-I error and Type-II error were used. FP provides the total of nonfraudulent firms that are mistakenly labelled as fraudulent, whereas Type-II error (false negative) indicates the sum of nonfraudulent firms that are incorrectly labelled as fraudulent.

VI. RESULTS AND DISCUSSION

This section presents the results and discussion from our proposed approach and compares the performance of developed hybrid models to the state-of-the-art machine learning algorithms, namely LR, RF, DT, XGBOOST, SVM, NB, Adaboost and LGBM. The single algorithms were compared in terms of prediction performance using their AUC-ROC score to find which ones perform the best in this dataset and therefore are most suited for use as the first algorithm for the proposed hybrid models. Figure 1 illustrates that generally; all the single models (other than NB) gave relatively similar performance values (0.66–0.71). Adaboost achieved the highest score (0.71) in the first phase. The decision was made based on the highest TPR and lowest FPR achieved by Adaboost, while NB gives the worst performance with an AUROC score of 0.56. The low performance of NB relies completely on the independence assumptions, whereas the used dataset might have some dependence features. However, it demonstrated one of the highest performance rates in the AUC-PR (Figure 2) alongside SVM and LGBM. One the other hand, DT and LR has shown the worst performance with 0.22 and 0.28 AUC-PR measure, respectively. As a result of the superior performance of Adaboost in terms of its AUCROC measurement, it was selected as the optimum single baseline model and was be combined with the rest of the algorithms to determine the best hybrid model. The Adaboost algorithm was able to correctly classify 9023 credit card transactions as shown in Table 3. Next, to establish the correctly classified dataset, TP was added to TN to train and validate the hybrid models.

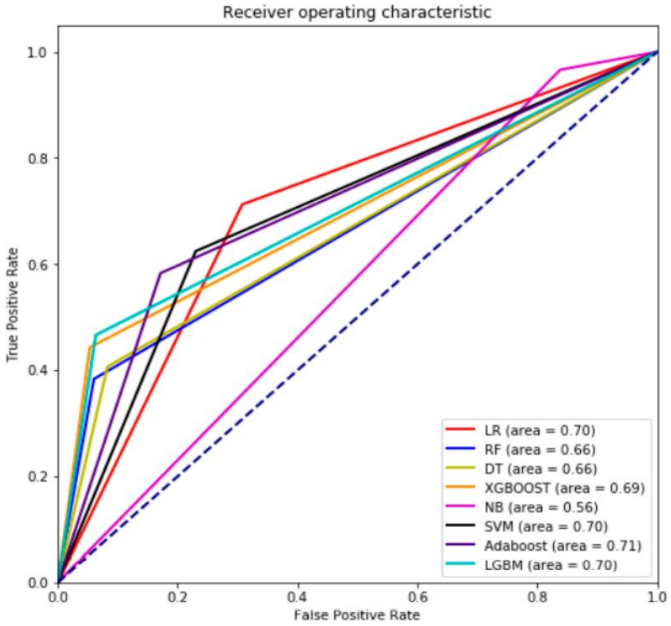


Fig. 5 AUCROC Curve of single models

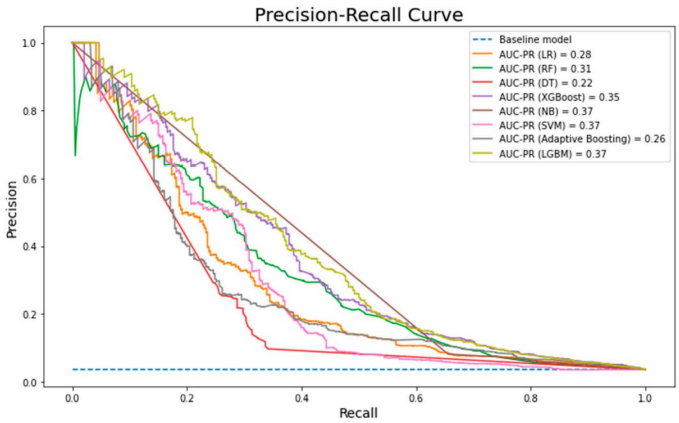


Fig. 6 AUC-PR of the single models.

	Predicted Positive	Predicted Negative
Actual Positive	8798	1889
Actual Negative	161	225

Table 3. Adaboost confusion matrix

In the second phase, seven hybrid machine learning models were developed (Figure 3). The predictive performance of the hybrid models has shown that the performance of the hybrid model Adaboost + LGBM excels in terms of its AUC-ROC measure when utilizing real world dataset (IEEE– CIS). As displayed in Table 4, the experimental results show that most of our proposed approaches outperformed the state-of-the-art machine learning algorithms in terms of AUC-1`ROC, Type-I error, Type-II error, F1-measure, precision, misclassification rate and TNR, although some of the hybrid models (Adaboost +LR, Adaboost + NB and Adaboost + SVM) had a higher Type-II error than the state-of-the-art algorithms. However, this will not be a server issue as Type-I error is more costly and being able to lower such an error will have a good impact on the bank system. Additionally, all the proposed hybrid models were able to detect the non-fraudulent cases that were identified as non-fraud at a rate of almost 0.99 percent.

state-of-the-Art Models	ROC	Recall	Precision	F-Measure	Misclassification Rate	TNR	Type-I Error	Type-II Error
LR	0.70	0.71	0.08	0.14	0.30	0.68	0.31	0.28
RF	0.66	0.38	0.19	0.25	0.07	0.93	0.06	0.59`
DT	0.66	0.41	0.15	0.22	0.10	0.91	0.08	0.57
XGBOOST	0.69	0.44	0.23	0.30	0.07	0.93	0.06	0.52
NB	0.56	0.97	0.04	0.08	0.81	0.15	0.84	0.03
SVM	0.70	0.62	0.09	0.16	0.23	0.74	0.25	0.35
ADABOOST	0.71	0.58	0.11	0.18	0.17	0.82	0.17	0.41
LGBM	0.70	0.47	0.21	0.29	0.07	0.92	0.07	0.52
Hybrid Models								
Adaboost+LR	0.67	0.36	0.83	0.50	0.004	0.999	0.0004	0.5
Adaboost+RF	0.74	0.50	0.97	0.66	0.003	0.999	0.0004	0.33
Adaboost+DT	0.76	0.54	0.51	0.52	0.006	0.990	0.0099	0.29
Adaboost+XGBOOST	0.79	0.59	0.94	0.73	0.002	0.996	0.0031	0.31
Adaboost+NB	0.76	0.96	0.05	0.10	0.105	0.579	0.5791	0.05
Adaboost+SVM	0.58	0.18	0.91	0.30	0.005	1.0	0.0000	0.66
Adaboost+LGBM	0.82	0.64	0.97	0.77	0.002	0.998	0.0018	0.25

Table 4. Comparison table of the state-of-the-art machine learning algorithms and the proposed hybrid models.

It is reflected in the AUC-PR (Figure 4) that the combination of Adaboost + XGBOOST outperforms the other six machine learning algorithms. Furthermore, Adaboost + XG1BOOST and Adaboost + LGBM have a high capability of accurately identifying fraudulent activities as they have the lowest misclassification error rate (0.002) of AUROC. Utilizing Adaboost as a pre-processing step gives us a cleaner dataset, which is expected to result in a more accurate and robust model that gives rise to a positive impact on the dataset via lowering the error rate for all the used algorithms. However, Adaboost + LGBM indicated a noticeable performance as it reached 0.82. On the contrary, LR and SVM showed diminishing performance for AUROC measures when hybridization with Adaboost took place. This indicates that hybridization between machine learning algorithms does not necessarily lead to higher performance. Additionally, looking into details for Adaboost + LGBM, a precision value of 0.97 indicates that when the model predicted a positive result it was correct 97 percent of the time and a recall value of 0.64 indicates that the model was able to identify 64 percent of all positive values correctly. In terms of the tradeoff between both measures, an F1-measure of 77 percent gives an equal consideration of both values. Having such a high result compared with other hybrid models in terms of its precision, ROC, F-measure, and misclassification rate, we conclude that Adaboost + LGBM is the best hybrid model for the given dataset in this study.

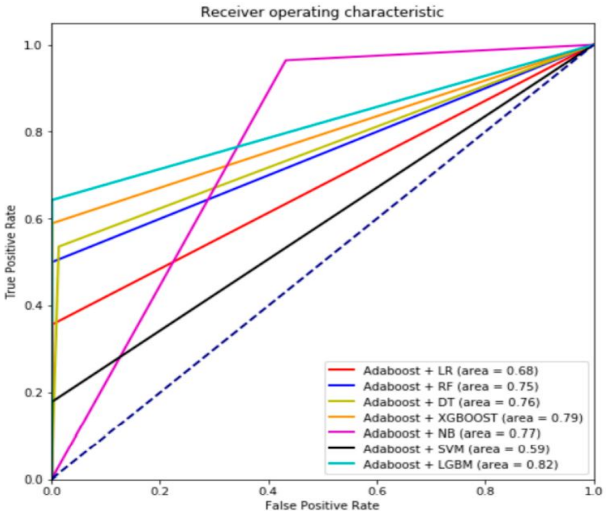


Fig. 7 AUROC curve of the hybrid models.

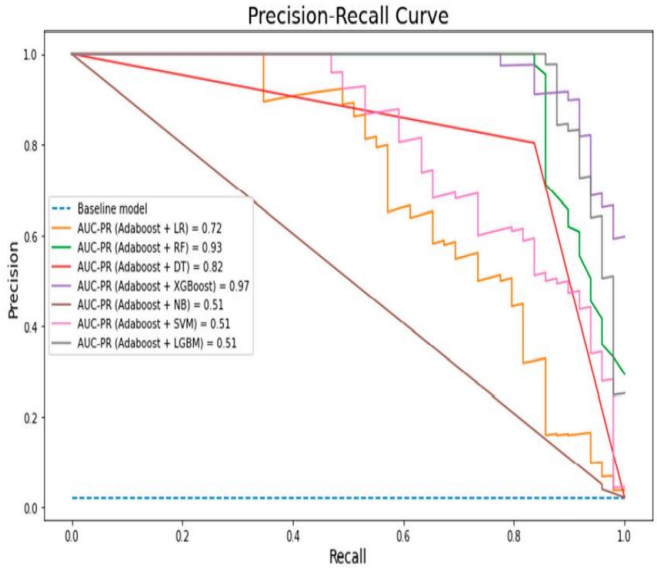


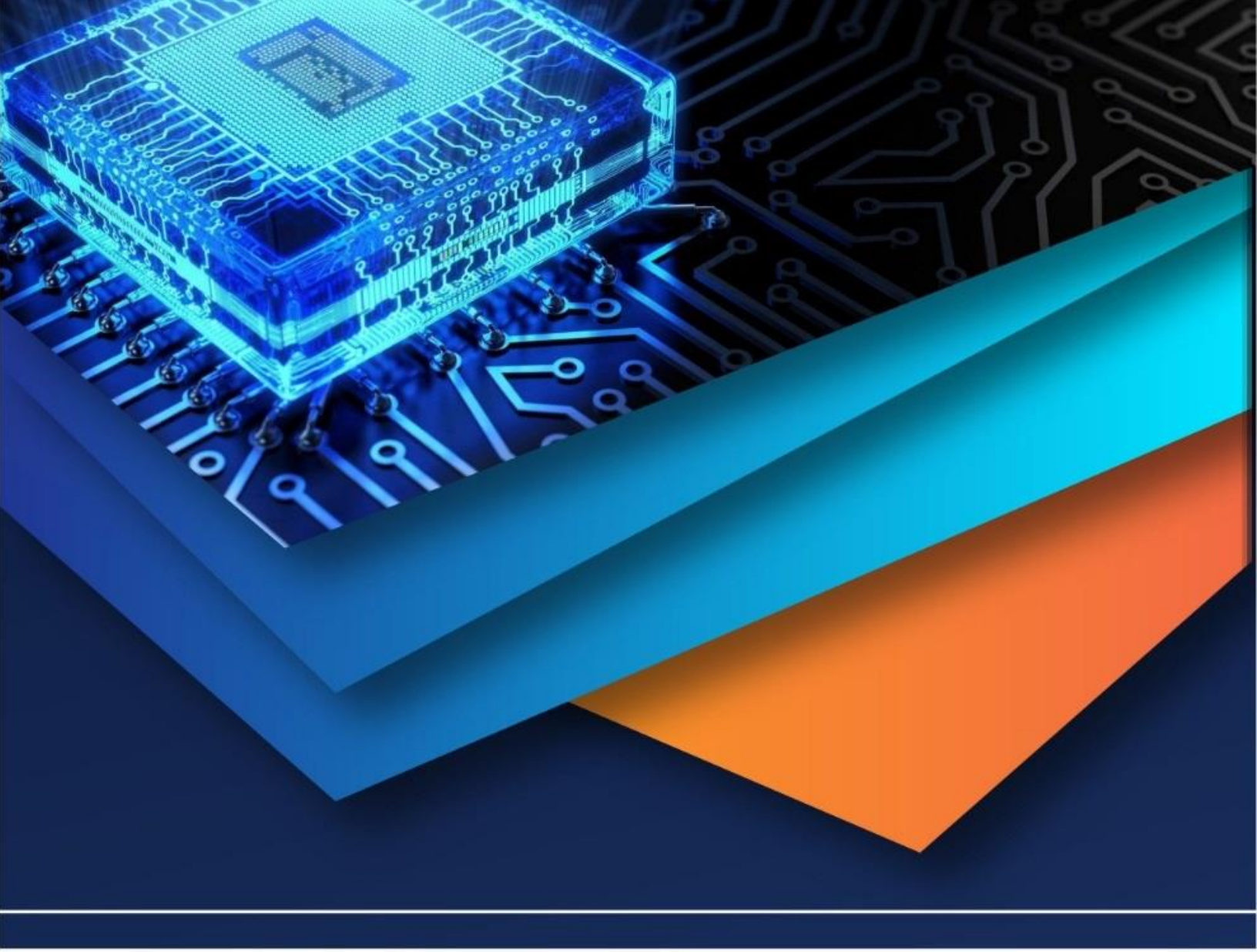
Fig. 8 AUR-PR curve of the hybrid models

VII. CONCLUSIONS

Credit card fraud has recently become a major concern worldwide, especially for financial institutions. Various approaches have been previously used to detect fraudulent activities; however, the need to investigate different reliable methods still exists to detect fraudulent credit card transactions, as was the aim in this work for a single case study. In this research, several hybrid machine learning models were developed and investigated based on the combination of supervised machine learning techniques as a part of a credit card fraud detection study. The hybridization of different models was found to have the ability to yield a major advantage over the state-of-the-art models. However, not all hybrid model worked well with the given dataset. Several experiments need to be conducted to examine various types of models to define which works the best. Comparing the performance of the hybrid model to the state-of-the-art and itself, we conclude that Adaboost and LGBM is the most accurate and precise model for this dataset. The result also illustrates that the use of hybrid methods has lowered the error rate. For future work, the hybrid models used in this study will be extended to other datasets in the credit card fraud detection domain. Future work may focus on different areas, starting by proposing data pre-processing techniques to overcome the drawback of the missing values. Additionally, different methods of feature selection and extraction should be studied and researched in the credit card domain and to determine the proposed model's impact on prediction accuracy. An investigation of the most appropriate hybrid model among the state-of-the-art machine learning algorithms to determine the most accurate model in the previously mentioned domain should be the main concern for future studies.

REFERENCES

[1] Kültür, Y.; Çaˇ glayan, M.U. Hybrid approaches for detecting credit card fraud. Expert Syst. **2017**, 34, 1–13.
[2] Kurshan, E.; Shen, H. Graph Computing for Financial Crime and Fraud Detection: Trends, Challenges and Outlook. Int. J. Semant. Comput. **2020**, 14, 565–589..
[3] West, J.; Bhattacharya, M. Intelligent Financial Fraud Detection: A Comprehensive Review. Comput. Secur. **2015**, 57, 47–66.
[4] Ethem, A. Introduction to Machine Learning, 2nd ed.; The MIT Press: Cambridge, MA, USA, 2014.R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.
[5] Angermueller, C.; Pärnamaa, T.; Parts, L.; Stegle, O. Deep learning for computational biology. Mol. Syst. Biol. **2016**, 12, 78.
[6] Carcillo, F.; Le Borgne, Y.A.; Caelen, O.; Kessaci, Y.; Oblé, F.; Bontempi, G. Combining unsupervised and supervised learning in credit card fraud detection. Inf. Sci. **2019**, 557, 317–331.
[7] Sivanantham, S.; Dhinagar, S.R.; Kawin, P.A.; Amarnath, J. Hybrid Approach Using Machine Learning Techniques in Credit Card Fraud Detection. In Advances in Smart System Technologies; Springer: Singapore, 2021.
[8] Credit Card Fraud Detection Database, Anonymized credit card transactions labeled as fraudulent or genuine, https://www.kaggle.com/mlg-ulb/creditcardfraud
[9] Cerda, P.; Varoquaux, G.; Kégl, B. Similarity encoding for learning with dirty categorical variables. Mach. Learn. **2018**, 107,1477–1494.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY