Detection of Credit Card Fraud using a Hybrid Ensemble Model

Sayali Saraf[1], Anupama Phakatkar[2]
Department of Computer Engineering
SCTR's Pune Institute of Computer Technology
Pune, India

***Abstract*—The rising number of credit card frauds presents a significant challenge for the banking industry. Many businesses and financial institutions suffer huge losses because card users are reluctant to use their cards. A primary goal of fraud detection is to identify prior transaction patterns to detect future fraud. In this paper, a hybrid ensemble model is proposed to combine bagging and boosting techniques to distinguish between fraudulent and legitimate transactions. During the experimentation two datasets are used; the European credit card dataset and the credit card stimulation dataset which are highly imbalanced. The oversampling method is used to balance both datasets. To overcome the problem of unbalanced data oversampling method is used. The model is trained to predict output results by combining random forest with Adaboost. The proposed model provides 98.27 % area under curve score on the European credit cards dataset and the stimulation credit card dataset gives 99.3 % area under curve score.**

***Keywords*—Credit card; hybrid ensemble model; bagging; boosting; data imbalance**

I. Introduction

There is a growing issue of financial fraud in the government, businesses, and financial sector with significant implications [1]. In credit card fraud, purchases occur on a cardholder's account without the cardholder's knowledge or consent. It is crucial to prevent fraud by taking all necessary precautions when carrying out these transactions. Bank regulators must also employ snipping technology to anticipate these thefts. Predicting the transactions that account holders will make but which will be completed by other people with access to the account is a fraud detection method for our dataset. It is a complex issue that needs to be resolved by both the account holder and the bank so that other customers don't face the same issue. However, there is a problem of class inequality with this issue. An individual consumer will complete many more legitimate transactions than fraudulent ones, or even none at all. A transaction that differs from a customer's previous purchases might be considered fraud. As credit card transactions increase in popularity for payment, academics are focusing on several strategies for fighting credit card fraud. The most common yet challenging issue is credit card fraud detection. As a result of the limited amount of credit card data, it is challenging to match a pattern for a dataset. Second, many records in the collection could include fraudulent transactions that follow a pattern of honest activity [2]. There are also some limitations to the issue. Firstly, study results are often classified and regulated, making them unavailable. Additionally, classified data sets are not readily available to the general public. Due to this, benchmarking specific models may be challenging. It is also difficult to develop solutions due to the security issue, which limits the exchange of concepts and techniques for detecting fraud, particularly credit card fraud [3]. The last point is that data sets are continually changing and evolving. It produces profiles of legitimate and fraudulent behavior separate from current valid transactions that may have been fraudulent in the past. In this paper, we will use a variety of machine learning algorithms, including logistic regression, random forest, and AdaBoost, to evaluate the performance of our proposed model. Two credit card datasets are used in the experiment, one of which is very skewed and unbalanced. The hybrid ensemble model is used to differentiate between fraudulent and legal transactions.

The work presented in the paper can be summarized as follows:

1) A hybrid ensemble model is proposed to classify fraudulent and legitimate transactions. The system uses an Adaboost, random forest, and Logistic regression to build a classifier.

2) The oversampling method and the removing outliers' approach are two methods used to address the problem of imbalanced data.

3) The train and test datasets are used to conduct the experiments on the proposed model.

The structure of the paper is organized as follows: The related work of existing algorithms is described in section II, while section III refers proposed hybrid model for fraud detection. Experimental credit card fraud detection, results, and discussion are presented in Section IV. The paper's conclusion is discussed in Section V in the final part.

II. Related Work

The performance of machine learning and data mining to prevent credit card fraud has been examined by the authors in [1]. On the other hand, most researchers used some classification measures to assess the solutions. A credit card detection model was used to extract the right attributes from transactional data. The aggregate approach was utilized to observe the customer's spending behavioral pattern. The author of this research proposes to construct a new set of features based on the periodic behaviors of transaction time

using an aggregation technique. An actual credit card fraud detection dataset from a large European cardholder company was used by the author. To examine the results, the author compared state-of-the-art credit card fraud models and weighed the pros and cons of various feature sets.

Credit cards are becoming more widely used in financial transactions, and at the same time, fraud is also increasing. The author presented a convolutional neural network framework to capture the pattern of fraud data in this research [2]. The author has proposed a trading entropy model to identify more complex consuming behaviors. Aside from that, the author merges the trending features into feature matrices for convolutional neural networks. As a result, the CNN model outperforms state-of-the-art approaches.

Supervised fraud classification algorithms for credit card fraud detection were proposed in [3]. The author has used two bank datasets to test these methods. Aggregation methods were suitable in many situations, but not all. SVM, logistic regression, random forest, and KNN were some of the classification algorithms used by the author. Out of this, the random forest gives better accuracy. Credit card transactions, as well as the fraud linked with them, are becoming more popular today. When credit card information is obtained unlawfully and used to make purchases, credit card fraud occurs. If credit card data is available and sufficient for a company or service, the author used a different machine learning technique to tackle the problem.

In [4], several popular methods in supervised, unsupervised, and ensemble classification were evaluated. The authors have applied different algorithms to identify fraudulent and legitimate transactions. Because unsupervised algorithms handle the skewness of datasets better than supervised algorithms, they outperform supervised algorithms in terms of performance measures. In future work, the author wants to contribute to the re-sampling techniques that will help us to balance data.

In [5], the authors have proposed a method to identify fraudulent and legitimate transactions. Because of the rapid progress of e-commerce and online banking, the usage of credit cards has increased dramatically, resulting in a large number of fraud instances. The author proposed a novel fraud detection method that has three stages. The first phase involves initial user authentication and card details verification. After the initial state, the transaction proceeds to the following step, where a fuzzy c-means clustering method was applied to determine the new pattern of credit card users based on their previous transactions. The authors used fuzzy c-means clustering algorithms to group similar datasets and a neural network to reduce misclassification based on the amount, timing, and kind of items purchased. For analyzing the proposed model, the author used stochastic models. The authors concluded that the application of fuzzy clustering and learning was the solution to a real-world problem based on the findings.

The main objective is to determine whether a transaction is legitimate or fraudulent. Various techniques, such as supervised and unsupervised procedures, were used to detect fraud [6]. Numerous methods identify fraud when utilizing supervised techniques. The author combined supervised and unsupervised techniques to classify credit card fraud to build a hybrid approach to improve system accuracy. This based on the results using the hybrid model gives better accuracy.

In [7], the authors have proposed long short-term memory networks as a method to aggregate the new pattern of data purchase behavior of cardholders, to improve the accuracy of the credit card fraud system. The comparison of baseline random forest to long short-term memory in this research improves the detection of accuracy as offline transactions, where the cardholder was physically present at the merchant. The author looks at both sequential and non-sequential learning systems that benefit from aggregation strategies in this paper.

The authors have used different algorithms on real-time datasets such as nearest neighbors, random forest, naive Bayes, multiple Perceptron, ad boost, quadrant discriminative analysis, pipelining, and ensemble learning [8]. The sample consists of European cardholders who were present for two days in September 2013. The dataset is highly unbalanced, so the ADASYN method has been used to correct it. As for performance measures, the author used precision, recall, accuracy, F1-measure, Matthew's correlation coefficient, and Balanced Classification Rate. Depending on a variety of parameters the pipelining gives better accuracy.

The authors have proposed Fraud-BNC, a customized Bayesian Network Classifier (BNC) algorithm on a real-time credit card fraud dataset in [9]. The Hyper Heuristic Evolutionary Algorithm was used to create BNC automatically (HHEA). A categorization dataset provided by Pag Seguro, a well-known Brazilian online payment provider, caused this difficulty. The author deals with two issues: a skewed dataset and misclassified fraud costs. As a result of Fraud-BNC, the method’s economic efficiency was evaluated and tested against seven alternative classification algorithms. When it comes to accuracy, Fraud-BNC outperforms other algorithms.

In business and banking, credit card fraud has become an issue. Credit card fraud occurs when a fraudster employs modern techniques and technology to complete credit card information without the owner's permission. The author proposed an intelligent approach for detecting credit card fraud using an upgraded light gradient boosting machine to address this issue (OLightGBM) in [10]. The author goes through numerous stages to establish this framework, including data collection, data pre-processing, model development, and model evaluation. The researcher had to use a Bayesian-based hyperparameter to maximize the parameter in the suggested approach. To assess the performance of the intelligent technique, the author employed two real-time datasets for detecting credit card fraud transactions. The first dataset comes from credit card fraud transactions made by European Cash Holders in 2013. The second dataset came from the UCSD-FICO Data Mining Contest in 2009. To compare with the provided technique, the author used a variety of machine learning algorithms. As a result, OLGBM exceeds confusion matrices, accuracy, precision, and recall, among other performance metrics.

In today's technology world and internet banking, credit card usage is rapidly increasing. Credit cards have become the most frequent payment method for online purchases. As a result, the number of cases of fraud increased. Stopping fraud was critical since it hurts the economy. To solve this problem, the author [11], has employed several techniques, including logistic regression (LR), Naive Bayes (NB), Support Vector Machines (SVM), decision tree (DT), and K-nearest neighbors (KNN), as well as random forest (RF). The author proposed a new deep learning architecture based on Spark to detect fraud. After that, the author compared the proposed deep learning architecture and the machine learning algorithm. The author used accuracy, precision, and recall performance metrics to classify fraudulent and legitimate transactions. As a result, random forest generates more precise outcomes.

In [12], the authors have proposed a novel fraud detection system that evaluates customers' previous transaction records and extracts pattern behavior. At the start, the authors used the clustering method to separate the cardholders into groups. To determine cardholder behavior, researchers employ the sliding window method to organize transactions. The dataset contained the European cardholder dataset. To balance the credit card fraud dataset, the author applied SMOTE techniques. Another option for dealing with unbalanced datasets is to employ the single class SVM. The authors used a variety of algorithms, both with and without statistical methodologies, to determine the dataset's correctness. Local Outlier factor, Isolation Forest, Support vector machine, logistic regression, decision tree, and random forest are some of the algorithms used. The main objective of this paper is to classify fraud and legitimate transactions.

In [13], has developed a deep learning and machine learning method to detect credit card fraud transactions. For developed a model author performed data pre-processing, normalization, and under-sampling techniques using a European cardholder imbalanced dataset. Then compare the machine learning methods such as Support Vector Machine and K-Nearest neighbors. After that, to train, the model artificial neural network was used by the author. As a result, the artificial neural network gives better accuracy.

In [14], the authors have focused on a new ensemble learning algorithm that combined bagging and boosting. As a result, detecting credit card fraud is a difficult task. The author proposed an ensemble hybrid model with the bagging and boosting method. The authors perform steps like pre-processing and feature engineering with ad-boost divide the data between train and test groups, classify the test data set using bagging-based ensemble classifiers like random forest and extra tree approaches, and generate results. The dataset UCSD-FICO was used as an input, and it is a severely unbalanced dataset. As a result, the author of the balancing data set utilized various strategies. The original feature space to the next feature space mapping approach was utilized in the first step, followed by the generation of the feature space. The next step is to use a tree-based classifier to solve the classification and regression problems. The author employed false negative and false positive rates, detection rates, and accuracy rates in the proposed model.

To detect credit card fraud, the authors [15] have used a different machine learning method. The results of a benchmark and a real-world dataset were compared by the author. A hybrid method combining ad boost and majority voting has also been developed by the researcher. The author compared the performance of a single model and a hybrid model on the same dataset. Researchers used naive Bayes, random forest, Decision tree, Neural Network, Linear Regression, Deep Learning, Logistic Regression, SVM, and Multilayer Perceptron as machine learning techniques. As a result, the methods were utilized to assess using ad boost and majority voting, which improved the accuracy with benchmark and real-time datasets.

In [16], a semi-supervised technique to detect credit card fraud, in which user profile clusters were created and used to construct classifiers. Users were profiled and grouped based on their patterns of conduct. Consumer segments were spread and further divided based on transaction factors such as volume, frequency, and distance. Random forest and XGBoost classifiers were trained on the total sample and compared to transaction-level classifiers in each cluster. This study finds that classifiers trained at the cluster level need not improve classifiers trained in the sample group in terms of overall weighted performance. The clustering method was used to identify groups of account holders. Moreover, some classifiers trained in specific groups do significantly better than the baseline, whereas classifiers learned in other groups do not perform as well. The optimum classifier for a given cluster varies by cluster and demonstrates the potential for new classifiers to perform well on groups that currently use underperforming models.

For credit card fraud detection, a Decision tree and random forest are used [17]. The author has used public data as sample data to test the model's efficiency. The finding was similar to a set of real-world credit card data obtained from a financial institution. Furthermore, some clatter was introduced to the data samples as a secondary check on the system's endurance. The study's methods were significant in that the first method created a tree against the user's activity, and frauds were detected using this tree. A user activity-based forest will be generated in a second way, and an attempt will be made to identify the suspect using this forest.

Artificial intelligence techniques were used to classify a fraudulent transaction as a routine transaction [18]. The author was to compare and contrast the results of several machine learning algorithms in detecting credit card fraud. The algorithm's rank and performance are of primary interest to the author. The model for identifying bad transactions in the e-commerce dataset was analyzed using the UCSD-FICO Data mining content 2009 dataset; Performance measures used by the author, such as classification accuracy and fraud detection rate.

To deal with anomalous transactions and develop a cardholder behavior model was proposed in [19]. To classify fraudulent and legitimate activities, the author used classification algorithms such as naive Bayes, Bayes Net, random forest, j48, libSVM, and MOLEM, as well as the Weka tool. Initially, the data was created and tested using the

random forest and j48 models. The author used a real-time dataset to test the efficacy, and random forests performed better.

In [20], the authors have proposed decision trees, random forests, and logistic regression as machine learning algorithms for fraud detection. The analytical model was put to the test using the benchmark dataset. The most accurate models are the random forest and decision tree. A confusion matrix was employed to assess accuracy Table I.

TABLE I. LITERATURE SURVEY

Sr.no	Title	Dataset	Methodology & tools	Advantages and limitation
1	Feature engineering strategies for credit card fraud detection.	European cardholder dataset	Decision Tree, Logistic Regression, Random Forest	The author proposed a method to improve the performance results of credit card fraud. The author has a problem with the system it takes a long time to make a decision.
2	Credit card fraud detection using convolutional neural networks." In International conference on neural information processing	Real-time credit card dataset	Convolution Neural network cost Estimation method	The advantage of implementing a convolutional neural network is to capture neural patterns of fraud activity discovered from a labeled dataset. Because there are far fewer fraud transactions in real life than in non-fraud situations, the major drawback when implementing it is the issue of an unbalanced dataset.
3	Transaction aggregation as a strategy for credit card fraud detection	Bank A and Bank B	SVM, logistic regression, Random Forest and KNN, Cart,	The advantage of aggregation is data do not need to be properly classified, it may be more resistant to the impacts of population drift.
4	Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection	European cardholder dataset	NB, RF, KNN LR, XGBT, SVM ANN, DL	The main advantage of unsupervised algorithms performs better throughout all measures both in absolute terms and in comparison, to other approaches since they are better at handling the dataset skewness.
5	Credit Card Fraud Detection: A Hybrid Approach Using Fuzzy Clustering & Neural Network	Real-time credit card dataset	Fuzzy Clustering and Neural Network	In this paper, fuzzy clustering is used to decrease the misclassified rates of transactions and also find new patterns based on past transaction data. The authors further used the different attributes to correctly classify the transactions.
6	Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection	Credit card fraud detection dataset	K-means clustering, ensemble learning,	In this paper main advantage is a combination of supervised techniques and unsupervised techniques to improve accuracy.
7	Sequence Classification for Credit-Card Fraud Detection	Real-world fraud detection dataset.	Random Forest, long short-term memory	The advantage of implementing a neural network to identify credit card fraud is that it can identify credit card activity and use patterns in a significant amount of customer and transactional data.
8	Credit Card Fraud Detection using Pipeline and Ensemble Learning	European cardholder dataset	Logistic Regression, Naive Bayes, K nearest neighbors, Multi-Layer Perceptron, Ada Boost, Quadrant Discriminant Analysis, Random Forests, Pipelining, and Ensemble Learning	The author compared different algorithms in which pipelining works best as compared to another algorithm. The benchmark dataset was highly imbalanced so the author was able to the balanced dataset.
9	A customized classification algorithm for credit card fraud detection	Pag Seguro dataset	customized Bayesian Network Classifier (BNC) algorithm for a real credit card fraud detection problem	In this paper, the authors were given High processing and detection speed on the Pag Seguro dataset for credit card fraud detection.
10	An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine	Real-world credit transaction	Optimized light gradient boosting machine (OLightGBM), light gradient boosting machine (LightGBM), KNN, SVM, NB, DT.	In this paper, OLightGBM gives better accuracy than other machine learning algorithms. The proposed model identifies a useful pattern of credit card fraud.
11	An Enhanced Secure Deep Learning Algorithm for Fraud Detection in Wireless Communication	European cardholders	Logistic regression (LR), Naive Bayes (NB), Support Vector Machines (SVM), decision tree (DT), and K-nearest	The credit card fraud dataset is highly imbalanced but in this system that imbalanced problem was solved. Credit card fraud prevention was a very important task but the author was unable to achieve it.

			neighbor (KNN) as well as random forest (RF)	
12	Credit card fraud detection using machine learning.	European cardholder dataset	DT, Local Outlier factor, Isolation forest, LR, RF	The advantage is that the author balanced the dataset using SMOTE techniques. In this paper, the authors need a balanced dataset for achieving high precision and recall.
13	Credit card fraud detection using artificial neural network	European cardholder dataset	SVM, KNN, and ANN	Using an artificial neural network model turns out to be the best for detecting fraud. The author was also unable to balance data using normalizing, under-sampling, or pre-processing methods.
14	Credit Card Fraud Detection by Modelling Behavior Pattern using Hybrid Ensemble Model	Brazilian bank data and UCSD-FICO data.	Ensemble learning techniques such as boosting and bagging.	Combination of Bagging and boosting ensemble learning method for distributing credit card detection. Dataset is highly imbalanced. For analyzing the behavior of the customer drift method
15	Credit card fraud detection using AdaBoost and majority voting	Benchmark Dataset. Real-time dataset	AdaBoost and majority voting methods	In this paper author used Majority and AdaBoost. The majority voting gives better accuracy. The author wants to use online learning methods where online learning methods prevent fraud it informs before fraud happens.
16	Improving Credit Card Fraud Detection by Profiling and Clustering Accounts	Credit card bank dataset	Random forest and XGBoost, k-mean clustering,	In this paper, k means clustering used for the effective and fast result of data. The disadvantage of k means is selected k values. The authors used Clustering to improve the detection of credit card fraud
17	A Novel Approach for Credit Card Fraud Detection using Decision Tree and Random Forest Algorithms	Credit card fraud Dataset	Bayesian network, Gaussian network, Random Forest, Decision Tree	The advantages of decision trees and random forests where random forests are utilized for preventing overfitting problems. The disadvantage of a decision tree, it gives the problem of overfitting.
18	An Evaluation of Computational Intelligence in Credit Card Fraud Detection	UCSD-FICO Data Mining Contest 2009 dataset	This paper analyses and compares various popular classifier algorithms that have been most commonly used in detecting credit card fraud	Classification accuracy and fraud detection rate are high in an evaluation of computational intelligence in credit card fraud detection. For credit card, and anomaly detection author want to propose a reliable expert system.
19	Credit Card Fraud Detection Based on Transaction Behavior	Real-time dataset	Random Tree and J48	Random forest gives the highest accuracy on the real-time dataset. The author had been unable to solve a random forest problem due to the slowness of the algorithm in a large number of trees.
20	Predictive Modelling for Credit Card Fraud Detection Using Data Analytics	German credit card fraud dataset	Logistic Regression, Decision Tree, Random Forest, Decision Tree.	The author compares various algorithms and concludes that random forest gives better accuracy. The author occurred a problem during the testing of random forest speed during the predictive model.

III. Proposed Hybrid Model for Fraud Detection

Fig. 1 shows the primary steps of the proposed models, which include data preparation, EDA, and a hybrid ensemble model. For the credit card fraud system, the result has been given as a categorization of genuine and fraudulent transactions. Data preparation employs the Smote technique, outlier removal, and null value deletion. After preprocessing, the hybrid ensemble model has been proposed for differentiating between legitimate and fraudulent transactions. A hybrid model reduces the risk of fraudulent transactions compared with a single model before applying SMOTE.

A. Data Preprocessing

Data pre-processing is an essential step because without it, the model can generate inaccurate results and it helps to preserve the integrity of the data. Data distribution, outlier identification, and noise reduction have been part of the data preprocessing stage.

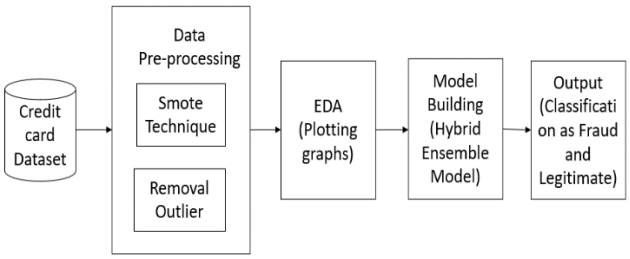

Fig. 1. Proposed Model of Credit Card Fraud Detection.

Anomaly detection is a method for finding outliers or odd patterns that deviate from anticipated behavior. In the proposed model Interquartile range (IQR) has been used to remove the outliers [21]. The interquartile range, or IQR, is the space between the first and third quartiles, or Q1 and Q3:

IQR = Q3 - Q1. Outliers are data points that are either below or above the median (Q1 - 1.5 IQR or Q3 + 1.5 IQR).

Q1 is the median.

Q2 is the average of the n smallest data points.

Q3 is the average of the n highest data points.

B. Bagging-Based Ensemble Learning

Bootstrap aggregation, sometimes known as "bagging," is a common strategy applied in ensemble learning-based models that integrate both classification and regression techniques, hence increasing accuracy and other associated metrics. The principle of bagging is the combination of weak learners with a strong learner [14]. For our experimentation, we implemented decision tree-based bagging classifiers such as random forest-based classifiers.

In bootstrap sampling, replacement sampling is used to produce a bootstrap sample B Si that is equal to D, where D is the input data. When D is big enough, B Si acts as an independent version of D, and the assumed empirical distributions resemble D [15]. Therefore, B Si might be viewed as a distinct and comparable variant of D. At bagging, in the i_{th} iteration, the model's predictions are averaged to suit the bootstrap sample B Si.

In the end, bootstrap sampling wants to remove a classifier's potential for overfitting.

1) Random forest: A supervised machine learning approach based on ensemble learning is known as a random forest. To create a more efficient prediction model, you can combine several algorithms or use the same technique more than once in ensemble learning [3, 4]. The term" Random Forest" comes from the fact that the random forest method mixes several algorithms of the same type or different decision trees into a forest of trees. Both regression and classification tasks may be performed using the random forest approach.

The basic steps of the random forest method are as follow [21]:

a) Choose N records at random from the dataset.

b) Based on these N records, construct a decision tree.

c) Repeat steps 1 and 2 after choosing how many trees you.

d) Want in your algorithm.

e) Each tree in the forest can forecast the category to which the new record belongs in a classification issue. The category that receives the majority of the votes is finally given a new record.

2) Adaboost algorithm: Adaboost is one boosting technique, which is similar to Random Forest Classifier. The Ada-boost classifier combines weak and strong classifier algorithms to create a large classifier [8]. A single algorithm may incorrectly categorize the items; however, by combining many classifiers, selecting the training set at each iteration, and assigning the appropriate amount of weight in the final vote, we can achieve a high accuracy score for the entire classifier. It keeps the algorithm repeatedly by selecting the training set depending on prior training accuracy. At any iteration, the weighting of each trained classifier is determined by the accuracy achieved. Adaboost provides weight to each training item after training a classifier at any level. The weight of a misclassified item is increased so that it is more likely to appear in the training subset of the Adaboost classifier. The basic steps of the Adaboost algorithm [21] are:

a) Initialize M, the maximum number of models to be fit, and set the iteration counter m=1.

b) Initialize the observation weights w_i= 1 N for i= 1, 2, N. Initialize the ensemble model $\hat{f}_b$= 0.

c) Train a model using $\widehat{f_m}$the observation weights that minimize the weighted error ECM defined by summing the weights for the misclassified observations is shown in (1).

d) Add the model to the ensemble:

$$\widehat{f_m} = (\widehat{f_{m-1}}) + (\widehat{\alpha_m})(\widehat{f_m}) \quad (1)$$

Where $(\widehat{\alpha_m}) = \frac{\log(1-e)m}{e_m}$ (2)

e) Update the weights $w_1, w_2, w_3, \ldots w_N$ so that the weights are increased for the observations that were misclassified. The size of the increase depends on $(\widehat{\alpha_m})$ with larger values $(\widehat{\alpha_m})$ leading to bigger weights as mentioned in (2).

f) Increment the model counter m=m+1 if m_i=M, go to step 1. The boosted estimate is given below:

$$\hat{f} = (\widehat{\alpha_1})\ (\widehat{f_1}) + (\widehat{\alpha_2})\ (\widehat{f_2}) + \ldots (\widehat{\alpha_m})\ (\widehat{f_m}) \quad (3)$$

g) The factor $(\widehat{\alpha_m})$ has a lower error and higher weight.

C. Logistic Regression

One of the most widely used Machine Learning algorithms, within the category of Supervised Learning, is logistic regression. The categorical dependent variable is predicted using a collection of independent factors. In a categorical dependent variable, the output is predicted using logistic regression [3]. As a result, the result must be a discrete or classifying value. Instead of providing the exact values of 0 and 1, it gives the probabilistic values that fall between 0 and 1. It can be either Yes or No, 0 or 1, true or false. Except for how they are applied, logistic regression and linear regression are very similar [4]. While logistic regression is used to solve classification difficulties, linear regression is used to solve regression problems. In logistic regression, we fit an" S" shaped logistic function, which predicts two maximum values, rather than a regression line (0 or 1) is shown in Fig. 2. The logistic function's curve shows the possibility of several things, like whether or not the cells are malignant, and whether or not a rat is fat depending on its weight [11]. Because it can classify new data using both continuous and

discrete datasets, logistic regression is a significant machine learning approach.

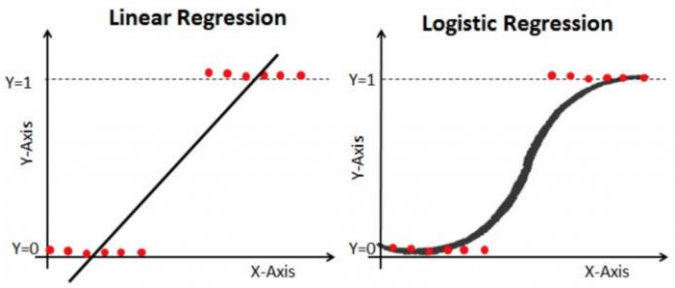

Fig. 2. Comparison between Linear Regression and Logistic Regression.

The equation of logistic regression of straight line written as [22]:

$$y=a_0+ a_1 * x_1 + a_2 * x_2 + \ldots\ldots a_k * x_k \quad (4)$$

In logistic regression, y can be between 0 and 1 only, so divide the above equation by (y-1):

$$\frac{y}{y-1} |0 \; for \; y = 0 \; and \; \infty \; for \; y = 1 \quad (5)$$

As a result, the logistic regression equation is defined as:

$$\log \frac{y}{y-1} =a_0+ a_1 * x_1 + a_2 * x_2 + \ldots\ldots a_k * x_k \quad (6)$$

D. Hybrid Ensemble Model

Fig. 3 shows how the hybrid ensemble model works. In a hybrid ensemble model, the first step is to train the model, and once it has generated individual results, the hybrid ensemble combines those outcomes with the help of majority voting to produce the final predicted results. The hybrid ensemble model is a combination of the bagging and boosting models. There are two well-known types of ensemble learning; bagging and boosting [14]. The widely used ensemble learning model for bagging is a random forest. Another well-liked ensemble learning approach that comes under the boosting category is AdaBoost. While the boosting models use the complete dataset, the bagging models only use a portion of the datasets. Random forest and Adaboost are employed as weak learners to create a hybrid ensemble model.

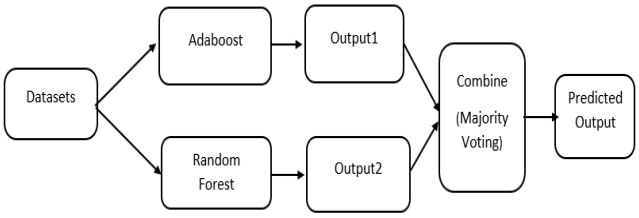

Fig. 3. Hybrid Ensemble Model.

IV. Experimental Detection of Credit Card Fraud Detection

This section explains the specificity and stability of our proposed models and compares them with the most recent research-based models. Our main objective is to increase the model's capacity for fraud detection. A better understanding of the data is required to do this. The experimental study was conducted on a simple Windows computer with a quad-core processor and 8 GB of RAM, and the results on the European credit card dataset and the Credit card stimulation dataset were acceptable. The proposed system has implemented the hybrid ensemble model for classification of the credit card fraud detection using python programming on a Jupyter Notebook. This system, primarily apply smote technique for data imbalanced problem. Further implementation of the hybrid ensemble model is on credit card fraud dataset. For checking the performance of the model, precision, recall, F1- score, and ROCAUC are calculated for every test case.

A. Data Description

Table II shows the instances, columns, and fraudulent and non-fraudulent cases of the European credit card dataset and credit card fraud stimulation dataset. Datasets are used to train and validate the efficacy of proposed approaches and hence play an essential part in research motivation. In this section, we'll go through two different datasets that have been used in our suggested approach's experiments.

1) European dataset: The first dataset, collected from www.kaggle.com, consists of credit card transactions performed by European cardholders within two days in September 2013, with 492 fraudulent transactions out of 284,807 as shown in Fig. 4. It has 31 features, including the time when a transaction occurred, the number of transactions, and 28 other qualities labeled V1 to V28, as well as the target label 'Class,' which uses a binary value of '1' or '0' to determine if a transaction is fraudulent or not [13].

TABLE II. The Credit card Dataset Description

Name	Instances	Features	Normal	Fraudulent
European Credit Card Dataset	284,807	31	248,315	492
Credit Card Stimulation Dataset	594,643	10	587,443	7200

V6	V7	V8	V9	...	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.128539	-0.189115	0.133558	-0.021053	149.62	0
-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.167170	0.125895	-0.008983	0.014724	2.69	0
1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.689281	-0.327642	-0.139097	-0.055353	-0.059752	378.66	0
1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.647376	-0.221929	0.062723	0.061458	123.50	0
0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.206010	0.502292	0.219422	0.215153	69.99	0

Fig. 4. European Credit Card Dataset.

2) Credit card bank stimulation dataset: The second dataset includes 594,643 transactions made across 180 simulated days, 7200 of which are considered fraudulent (1.2 percent). This dataset is a synthetic dataset developed with BankSim software, which is a simulation tool meant to simulate fraud data in Fig. 5, BankSim uses a multi-agent simulation methodology that is based on a sample of aggregated real-time transaction data provided by a Spanish bank. Thousands of transactional data records from November 2012 to April 2013 make up the initial bank data. To simulate this genuine bank data, BankSim employs many agents from three main categories: traders, customers, and fraudsters. These agents interact with one another for a duration of a few days, establishing a purchasing transaction log that strongly matches the original bank.

	step	customer	age	gender	zipcodeOri	merchant	zipMerchant	category	amount	fraud
0	0	'C1093826151'	'4'	'M'	'28007'	'M348934600'	'28007'	'es_transportation'	4.55	0
1	0	'C352968107'	'2'	'M'	'28007'	'M348934600'	'28007'	'es_transportation'	39.68	0
2	0	'C2054744914'	'4'	'F'	'28007'	'M1823072687'	'28007'	'es_transportation'	26.89	0
3	0	'C1760612790'	'3'	'M'	'28007'	'M348934600'	'28007'	'es_transportation'	17.25	0
4	0	'C757503768'	'5'	'M'	'28007'	'M348934600'	'28007'	'es_transportation'	35.72	0

Fig. 5. Credit Card Bank Stimulation Dataset.

B. Performance Parameter

The learning algorithm's performance measure showed unbalanced behavior in the imbalanced distribution of the classes. So, it is necessary to select suitable measures to assess the effectiveness of the categorization system. Precision, recall, F1 Score, and accuracy have been chosen as performance evaluation metrics for the proposed work because the learning algorithm exhibits an accuracy phenomenon in unbalanced scenarios.

True Positive (TP) - How many safe cases did our model properly predict.

False Negative (FN) - How many cases of our model incorrectly predicted.

False Positive (FP) - How many fraud cases are classified incorrectly.

True Negative (TN) - How many fraud cases are classified correctly.

Precision - TP/(TP+FP)

Recall - TP/(TP+FN)

F1 Score - Hormonal mean of precision and recall

F1 Score = (2*precision *. recall)/ (precision + recall)

C. Data Imbalanced

In this paper, two benchmark datasets have been used. In the given datasets there are few fraudulent incidence which makes data imbalanced.The Fig. 6 and Fig. 7, represents percentage ratio of fraudulent transaction of the European credit card dataset and the credit card stimulation dataset respectively. To increases the fraudulent cases Synthetic Minority Oversampling Technique (SMOTE) technique is used. The Fig. 8, representes the number of increses the fraudulent cases after applying smote apporach for the European credit card dataset and the Credit card stimulation dataset.

D. Exploratory Data Analysis

Exploratory data analysis is a data analysis process used to properly understand the data and discover its many characteristics, frequently using visual methods. Data analysis allows us to better understand and identify meaningful patterns in it.

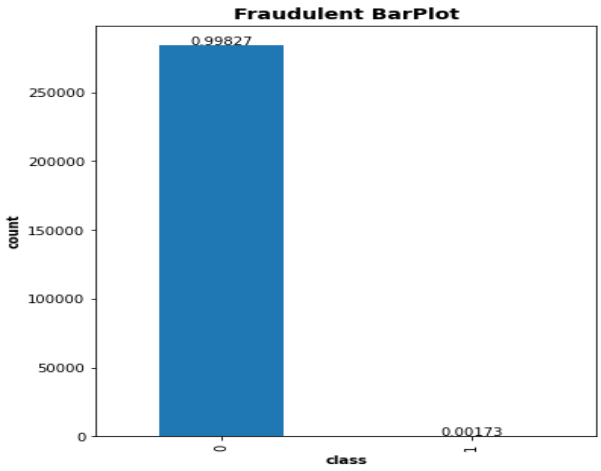

Fig. 6. European Credit Card Dataset before Applying Smote Technique.

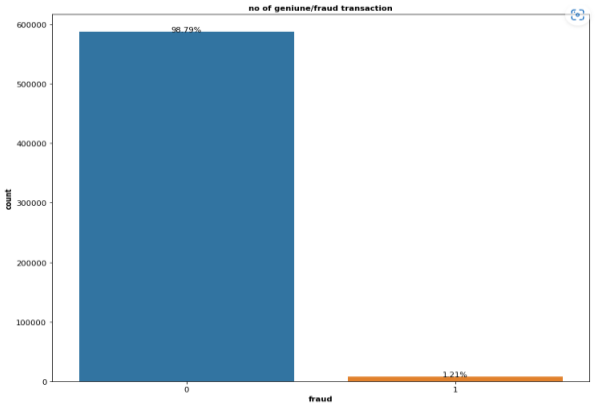

Fig. 7. Credit Card Stimulation Dataset before Applying Smote Technique.

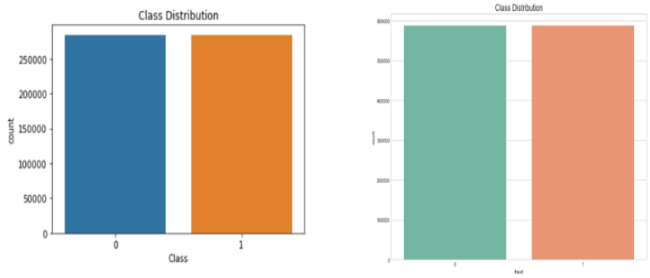

Fig. 8. European Credit card and Credit Card Stimulation Dataset after Smote Technique.

To analyze the time and amount in this paper, exploratory analysis is performed. Fig. 9 and 10, show how the time and number of transactions during the day and at night differ.

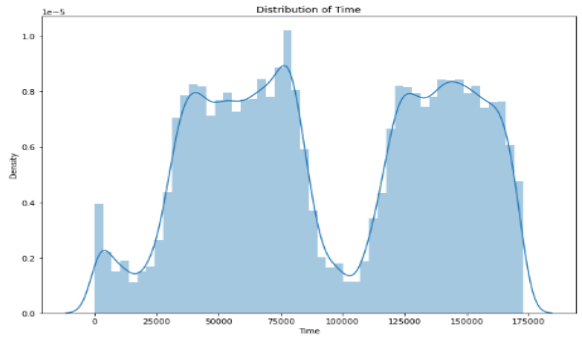

Fig. 9. Transaction Time of European Cardholder Dataset.

Fig. 9 shows the low peak value in the time distribution because there is a significant difference between the night and day transactions. In the density plot to x-axis shows the time of transaction and the y-axis shows the density of attributes.

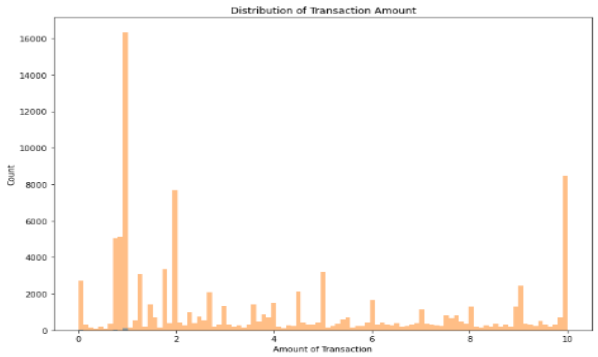

Fig. 10. Transaction Amount of European Cardholder Dataset.

Fig. 10 represents the total amount of money transacted. The majority of transactions are small, and just a few come close to reaching the maximum transaction value.

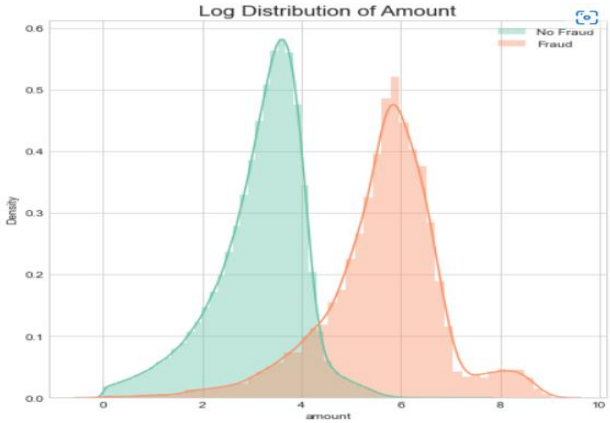

Fig. 11. Transaction Amount of Credit Card Stimulation Dataset.

Fig. 11 represents the total amount of money transacted. It shows that number of fraudulent transactions is less than a non-fraudulent transaction.

E. Result and Discussion

Hybrid ensemble modeling is proposed to categorize fraudulent and legal transactions. The experiment is done on the European credit card and credit card stimulation dataset. The 70:30 % ratio is used for training and testing classifier. In both, the dataset fraudulent instances are less compared to non-fraudulent transactions. So, this is a serious issue that has been found with the dataset. In the European dataset 495 fraudulent transactions out of 284,807 non-fraudulent transactions, so the number of transactions needs to be increased. Same as credit card stimulation dataset 7200 fraudulent transactions out of 594,643. Our approach provides non-fraudulent transactions more weight when applied to an imbalanced dataset. Ensemble models are used to solve the main issue in credit card fraud detection, which is predicting future transaction behavior and finding the right solution.

The initial comparison between the single model and the original dataset is carried out in this study. But the single model has obtained less True positive value and more false positive value which indicates that more fraudulent transactions are presented in datasets because of unbalanced datasets. In order increases true positive value and handle unbalanced dataset oversampling strategies is used. To increases performance of the model hybrid ensemble model is proposed. The hybrid ensemble model is constructed by combining an Adaboost and random forest. This will improve the performance parameters of the system.

In this paper, the performance of the proposed hybrid ensemble model is compared to the machine learning algorithms, including logistic regression, random forest, and Adaboost. Table III and Table V shows the performance measure of European credit card and credit card bank stimulation on the imbalanced dataset. As we observed that the single model and ensemble model did not improve the true positive rate and true negative rate with the imbalanced dataset. So, we applied smote technique to the balanced dataset. Smote technique is used to increase fraudulent instances. After applying Smote oversampling method, the datasets are much more balanced.

The balanced dataset is added and tested with an ensemble model such as bagging and boosting and a predictive model such as logistic regression. Table IV is showing an improvement in precision, recall, and F1 score for the European dataset. Same for Credit card fraud detection dataset precision, recall, and F1 score slightly increased is observed in Table VI precision-recall (AUPR) curve is being used to analyses the performance measure of the proposed model. Table VII shows the comparison results of the European cardholder dataset and Credit card stimulation dataset, which show the AUPR score for boosting, LR, and Adaboost (+) random forest. Fig. 12 and Fig. 13 show the AUPR curve for random forest + Adaboost. On the European cardholder dataset and Credit card stimulation dataset, the proposed method shows a Random Forest +Adaboost which means the hybrid ensemble model gives better results than the single model.

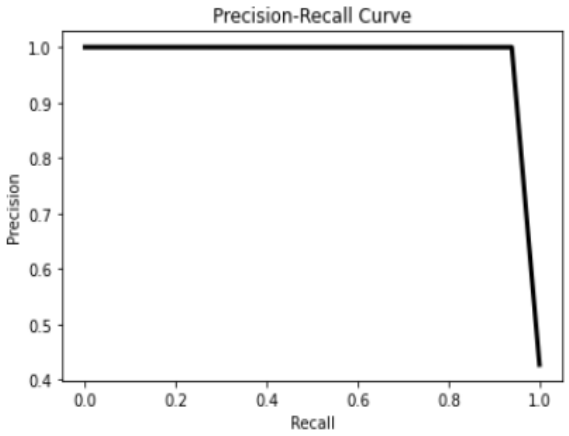

Fig. 12. Recall and Precision Curve for European Credit Card Dataset.

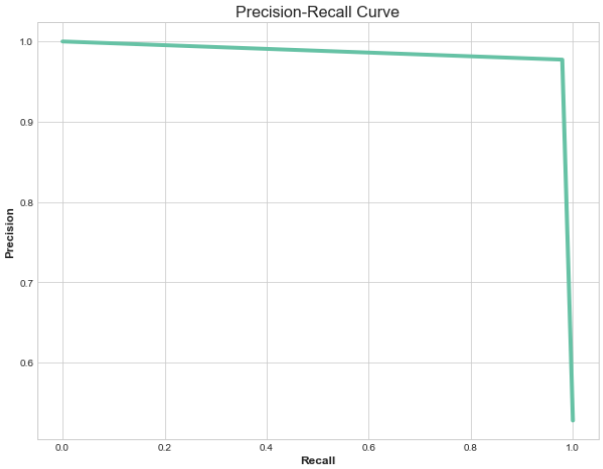

Fig. 13. Recall and Precision Curve for Credit Card Stimulation Dataset.

Table III and Table V shows the performance measure of classification algorithms on the European dataset and Credit card stimulation dataset before applying smote technique. So, we observed that after applying smote technique value of precision, recall, and F1 Score is improved rather to the without applying smote technique. Here hybrid ensemble model with random forest and Adaboost gives better precision, recall, and F1 Score.

Table IV and Table VI show the performance of measure of classification algorithm on European cardholder dataset and credit card stimulation dataset after applying smote technique. Here, the hybrid ensemble model with a combination of random forest and Adaboost gives better precision, recall, and F1- score than other algorithms.

TABLE III. BEFORE SMOTE TECHNIQUE ON EUROPEAN CREDIT CARD FRAUD DATASET

Algorithms	Precision	Recall	F1 Score
Logistic Regression	0.88	0.62	0.73
Adaboost	0.78	0.66	0.72
Random Forest	0.94	0.77	0.85
Random Forest+Adaboost	0.94	0.78	0.85

TABLE IV. AFTER SMOTE TECHNIQUE ON EUROPEAN CREDIT CARD FRAUD DATASET

Algorithms	Precision	Recall	F1 Score
Logistic Regression	0.96	0.90	0.93
Adaboost	0.97	0.94	0.95
Random Forest	0.97	0.98	0.95
Random Forest+Adaboost	1.00	0.94	0.97

TABLE V. BEFORE SMOTE TECHNIQUE ON CREDIT CARD FRAUD STIMULATION DATASET

Algorithms	Precision	Recall	F1 Score
Logistic Regression	0.88	0.62	0.73
Adaboost	0.78	0.66	0.72
Random Forest	0.94	0.77	0.85
Random Forest+Adaboost	0.91	0.78	0.84

TABLE VI. AFTER SMOTE TECHNIQUE ON CREDIT CARD FRAUD STIMULATION DATASET

Algorithms	Precision	Recall	F1 Score
Logistic Regression	0.92	0.99	0.97
Adaboost	0.97	0.99	0.98
Random Forest	0.98	0.97	0.98
Random Forest+Adaboost	0.99	0.99	0.99

The Predictive behavior of the proposed model is analyzed concerning the area under precision and recall curve. The result shown in Table VII is the area under precision and recall score for LR, boosting Adaboost +random forest. Here we observed that a hybrid ensemble model with random forest +Adaboost gives a better AUC Score than other algorithms.

TABLE VII. AREA UNDER CURVE SCORE ON EUROPEAN CREDIT CARDATASET AND CREDIT CARD STIMULATION

The area under curve score (AUC Score)	European credit card dataset	Credit card bank stimulation dataset
Logistic Regression	95.10	95.80
Adaboost	96.79	98.43
Random Forest	97.32	98.09
Random Forest+Adaboost	98.26	99.37

V. CONCLUSION

In this paper, a hybrid ensemble-based model is proposed to classify fraudulent and legitimate transactions. At the beginning of the project, the data analysis technique is used to map the original feature. To overcome the imbalanced problem oversampling smote method is used to balance the dataset during data pre-processing. After pre-processing, logistic regression, random forest, and Adaboost are used to check whether a transaction is legitimate or fraudulent. The hybrid ensemble model before applying the smote technique gives 0.85 % F1-Score and after applying smote technique it gives 0.97% on the European credit card fraud dataset. So, the F1 -score of Smote technique with the hybrid ensemble model

gives more results. For the Credit card stimulation dataset before applying smote technique F1 score gives 0.84% and after applying smote technique F1 Score gives 0.99%. It is observed that a hybrid ensemble model that combines random forest and Adaboost gives better results. From Table VI, the hybrid ensemble model for the European dataset achieves 98.27 % area under the curve score, whereas the Credit card fraud stimulation dataset achieves 99.37 % area under the curve score. In future work, apply the under-sampling method to check the performance of the algorithm and also use deep learning techniques for the classification of fraudulent and legitimate transactions.

REFERENCES

[1] Bahnsen, Alejandro Correa, Djamila Aouada, Aleksandar Stojanovic, and Bj¨orn Ottersten.” Feature engineering strategies for credit card fraud detection.” Expert Systems with Applications 51 (2016): 134-142.

[2] Fu, Kang, Dawei Cheng, Yi Tu, and Liqing Zhang.” Credit card fraud detection using convolutional neural networks.” In International conference on neural information processing, pp. 483-490. Springer, Cham, 2016.

[3] Whitrow, Christopher, David J. Hand, Piotr Juszczak, David Weston, and Niall M. Adams.” Transaction aggregation as a strategy for credit card fraud detection.” Data mining and knowledge discovery 18, no. 1 (2009): 30-55.

[4] Mittal, Sangeeta, and Shivani Tyagi.” Performance evaluation of machine learning algorithms for credit card fraud detection.” In 2019 9th International Conference on Cloud Computing, Data Science Engineering (Confluence), pp. 320-324. IEEE, 2019.

[5] Behera, Tanmay Kumar, and Suvasini Panigrahi.” Credit card fraud detection: a hybrid approach using fuzzy clustering neural network.” In 2015 second international conference on advances in computing and communication engineering, pp. 494-499. IEEE, 2015.

[6] Carcillo, Fabrizio, Yann-A¨el Le Borgne, Olivier Caelen, Yacine Kessaci, Fr´ed´eric Obl´e, and Gianluca Bontempi.” Combining unsupervised and supervised learning in credit card fraud detection.” Information sciences 557 (2021): 317-331.

[7] Jurgovsky, Johannes, Michael Granitzer, Konstantin Ziegler, Sylvie Calabretto, Pierre-Edouard Portier, Liyun He-Guelton, and Olivier Caelen." Sequence classification for credit card fraud detection." Expert Systems with Applications 100 (2018): 234-245.

[8] Bagga, Siddhant, Anish Goyal, Namita Gupta, and Arvind Goyal." Credit card fraud detection using pipeline and ensemble learning." Procedia Computer Science 173 (2020): 104-112.

[9] de S´a, Alex GC, Adriano CM Pereira, and Gisele L. Pappa.” A customized classification algorithm for credit card fraud detection.” Engineering Applications of Artificial Intelligence 72 (2018): 21-29.

[10] Taha, Altyeb Altaher, and Sharaf Jameel Malebary.” An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine.” IEEE Access 8 (2020): 25579-25587.

[11] Sanober, Sumaya, Izhar Alam, Sagar Pande, Farrukh Arslan, Kantilal Pitambar Rane, Bhupesh Kumar Singh, Aditya Khamparia, and Mohammad Shabaz.” An enhanced secure deep learning algorithm for fraud detection in wireless communication.” Wireless Communications and Mobile Computing 2021 (2021).

[12] Sailusha, Ruttala, V. Gnaneswar, R. Ramesh, and G. Ramakoteswara Rao.” Credit card fraud detection using machine learning.” In 2020 4th international conference on intelligent computing and control systems (ICICCS), pp. 1264-1270. IEEE, 2020.

[13] Asha, R. B., and Suresh Kumar KR.” Credit card fraud detection using artificial neural network.” Global Transitions Proceedings 2, no. 1 (2021): 35-41.

[14] Karthik, V. S. S., Abinash Mishra, and U. Srinivasulu Reddy.” Credit Card Fraud Detection by Modelling Behaviour Pattern using Hybrid Ensemble Model.” Arabian Journal for Science and Engineering 47, no. 2 (2022): 1987-1997.

[15] Randhawa, Kuldeep, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim, and Asoke K. Nandi." Credit card fraud detection using AdaBoost and majority voting." IEEE Access 6 (2018): 14277-14284.

[16] Kasa, Navin, Andrew Dahbura, Charishma Ravoori, and Stephen Adams.” Improving credit card fraud detection by profiling and clustering accounts.” In 2019 Systems and Information Engineering Design Symposium (SIEDS), pp. 1-6. IEEE, 2019.

[17] Dileep, M. R., A. V. Navaneeth, and M. Abhishek.” A novel approach for credit card fraud detection using decision tree and random forest algorithms.” In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 1025-1028. IEEE, 2021.

[18] Mahmud, Mohammad Sultan, Phayung Meesad, and Sunantha Sodsee. ” An evaluation of computational intelligence in credit card fraud detection.” In 2016 International Computer Science and Engineering Conference (ICSEC), pp. 1-6. IEEE, 2016.

[19] Kho, John Richard D., and Larry A. Vea. ”Credit card fraud detection based on transaction behavior.” In TENCON 2017-2017 IEEE Region 10 Conference, pp. 1880-884. IEEE, 2017.

[20] Patil, Suraj, Varsha Nemade, and Piyush Kumar Soni." Predictive modeling for credit card fraud detection using data analytics." Procedia computer science 132 (2018): 385-395.

[21] Peter Bruce, Andrew Bruce, Peter Gedeck."Practical statistical for data scientists."O'Reilly Media,2017.

[22] Alenzi, Hala Z., and Nojood O. Aljehane. "Fraud detection in credit cards using logistic regression." International Journal of Advanced Computer Science and Applications 11, no. 12 (2020).