IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125146

Credit Card Fraud Detection Using Machine Learning

Dr. Kiran[1], Sanchitha L Anand[2], Samudyata S[3], Raju Poovarsha[4], Soujanya G V[5]

Assistant Professor, Department of Electronics and Communication Engineering, Vidyavardhaka College of Engineering, Mysuru, India[1]

Department of Electronics and Communication Engineering, Vidyavardhaka College of Engineering, Mysuru, India[2-5]

Abstract— All economic opportunities were made possible by digitalization, which also confused the system with illegal activity. One improvement in the banking system is credit cards. Credit cards were able to draw new users every day because of how simple they were to use. Due to its popularity, there have been more fraudulent users, erroneous transactions, and card theft over time. Systems for detecting fraud were developed in order to stop these illicit activities. Our suggested article seeks to establish the truth or falsity of the completed transaction. To extract the results, we employed ML methods like logistic regression and random forest. It has been demonstrated that the Random Forest algorithm technique delivers an accurate generalisation error estimate. It was discovered the Random Forest algorithm technique. The Random Forest algorithm technique was found to be relatively stable, to resist overfitting, and to give a decent estimate of the generalisation error. Based on their precision, specificity, and accuracy, the results are evaluated.

Keywords— credit card, fraud detection, logistic regression, random forest, machine learning

I. INTRODUCTION

The majority of financial institutions have improved the public's access to business services through online banking in the twenty-first century. Electronic payment systems are essential in today's financial environment, which is highly competitive. The purchase of goods and services has been made very simple. Financial organisations routinely issue cards to customers so they can purchase without carrying cash. Similar to debit cards, credit cards offer consumers protection from lost, damaged, or stolen products. Before using a credit card, customers must confirm the transaction with the merchant. According to statistics, Visa and Mastercard distributed 2287 million credit cards throughout the world in 2017. Visa issued 1131 million, compared to 1131 million issued by MasterCard.

These figures show how card-based transactions gained acceptance among consumers. Fraudsters are building the basis for manipulating this group because they account for a sizable number of foreign transactions. Furthermore, social engineering people can occasionally be simple. Customers can profit greatly from credit cards, but there are drawbacks as well, including security and fraud issues. Banks and other financial institutions are addressing this issue. the credit card's issuance fraud is one that banks and other financial organisations are addressing. Through unprotected internet platforms and websites, credit card information is susceptible to theft. They could also be obtained as a result of identity theft. fraudsters may unlawfully access users' credit and debit card numbers without their knowledge or consent.. One of the main reasons for financial losses in the finance industry, according to "U.K. finance," is due to fraudulent usage of credit and debit cards. It is a significant risk that results in significant financial losses on a global scale as a result of technological development. Therefore, detection is essential to minimising financial setbacks. Machine learning is good at identifying the difference between honest and dishonest transactions. One of the biggest problems with detection methods is the difficulty in exchanging ideas about fraud detection.

The majority of academics and researchers are now paying attention as credit card fraud detection has substantially increased recently. The purpose of this research paper is to analyse and evaluate several aspects of identifying credit and debit fraud. The article examines numerous strategies for spotting fraudulent credit card transactions before offering a more practical approach to preventing credit card fraud. Some methodological obstacles that are preventing the use of ML in real-time applications are being addressed by researchers. A few of the studies that have been conducted in a variety of fields include the detection of abnormal patterns, biometric identification, diabetes prediction, happiness prediction, water quality prediction, accident prevention at Heathrow, timely diagnosis of bone diseases, and prediction of informational efficiency using deep neural networks. Researchers are making an effort Despite these constraints, we aim to improve the fraud detection performance of ML.

DOI: 10.17148/IJARCCE.2023.125146

II. LITERATURE SURVEY

Saiju, Sanisa, S. Akshaya Jyothy, Christeena Sebastian, Liss Mathew, and Tintu Sabu [1] In terms of application domain, the supervised algorithm like random forest stands first in the literature. Likelihood fraudulent transactions can be identified easily as soon as the algorithms are integrated into the fraud detection system of a bank.In contrast to past classification problems, new approach was taken to the study's objective by implementing a variable penalty for misclassification.The suggested system's performance is assessed using precision, f1-score, and accuracy.

Arafath, Yeasin, Animesh Chandra Roy, M. Shamim Kaiser, and Mohammad Shamsul Arefin [2] The sequence is put together during the detection phase after the credit card holder's shopping habits are assessed during the training phase using the kmeans clustering method. Using sequence alignment based on real cardholder transaction history and transaction behavioural changes.If the gap between the good and poor scores is greater than a specific limit, the transaction is unlawful; otherwise, it is allowed.

Awoyemi, John O., Adebayo O. Adetunmbi, and Samuel A. Oluwadare [3] The experiments are presented and discussed in two steps. Eight classification methods are compared in the first stage. Three factors were taken into consideration for the comparison: sensitivity, accuracy, and the area under the precision recall curve.The second phase then compares various imbalance classification methodologies. Additionally, the Auto Associative Neural Network is used and contrasted with the ANN.

Ileberi, Emmanuel, Yanxia Sun, and Zenghui Wang [4] In this study, a feature selection approach for a machine learning (ML)-based credit card fraud detection engine is proposed. It uses the genetic algorithm. With the help of a dataset, the effectiveness of the suggested fraud detection engine was assessed. The suggested detection engine uses the ML classifiers Decision Tree, Random Forest, Logistic Regression, Artificial Neural Network and Naive Bayes after choosing the optimal features. Using the Synthetic Minority Oversampling Technique (SMOTE) oversampling technique, the researcher solved the issue of class imbalance in the dataset. The major performance indicator for evaluating the effectiveness of each ML method was accuracy.

Sadineni, Praveen Kumar [5] Through the use of Artificial Neural Networks (ANN), Decision Trees, Support Vector Machines (SVM), Logistic Regression, and Random Forest, the current study aimed to identify fraudulent transactions. Accuracy, precision, and false alarm rate are used to assess how well each technique performs. Decision Tree excels with sampled and pre processed data, whereas Logistic Regression excels with unprocessed, raw data. With categorical and continuous data, random forest performs well.

Sanobar khan, Sanovar, Suneel Kumar , Mr Hitesh Kumar [6] 28 of the 31 columns in the datasets under examination have the labels v1v28 to protect sensitive data, making a total of 31 columns. Time, Amount, and Class are represented in the remaining columns. A fraudulent transaction is one that is a valid class 0.After these datasets have been analysed, a histogram is shown for each column that was taken into consideration. A graph of the datasets is created as a result.

Sharma, Pratyush, Souradeep Banerjee, Devyanshi Tiwari, and Jagdish Chandra Patni [7] After obtaining the dataset, the data was separated into train, validation, and test sets. The 70/30 guideline was adhered to, with test data making up 15%, validation data being 15, and training data being 70. Several machine learning models were trained using the dataset. The performance of each model is evaluated. The macro averages of the F 1 score, recall, and precision are used.

Meenakshi, B. Devi and N. Indira [8] An approach for classification and regression is called Random Forest. It is essentially a group of decision tree classifiers. Decision trees perform worse than random forests since the former breaks the bad habit of overfitting the training set. Following, a decision tree is built, with each node splitting on a feature selected at random from the entire feature set. Each tree is trained independently of the others. The algorithm has been found to be resistant to overfitting and to provide a reliable estimate.

Priya, G. Jaculine, and S. Saradha [9] The suggested device because there are fewer minority class data in the dataset. SMOTE synthesises minority class elements based on those that already exist. It operates by choosing any unspecified point from the minority class, then calculating its k-nearest neighbours. The created nodes are then sent through a splitting function using the chosen features. The best split function is typically employed when using a decision tree To limit the amount of nodes generated for each tree, there should be a limit defining how many of these nodes should be created.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125146

Azhan, Mohammed, and Shazli Meraj [10] Study throws light on the use of machine learning and neural networks to spot future fraudsters by examining their past wrong doings and data on previous fraudsters is explored. Support Vector Machines, Logistic Random Forest Regression, Multinomial Naive Bayes, and a Simple Neural Network are also used. A confusion matrix and classification reports generated by the Sklearn software were used to assess the models.

III. METHODOLOGY

Classification of transactions in the dataset which are fraudulent or non-fraudulent by making use of algorithms like random forest algorithm and logical regression is the main objective of this paper. We can determine fraud transactions more accurately by comparing these two algorithms. The diagram with the complete architecture of fraud detection system contains numerous steps.

Work Process

1. Data Collection
2. Data Preprocessing
3. Feature Extraction
4. Evaluation Model

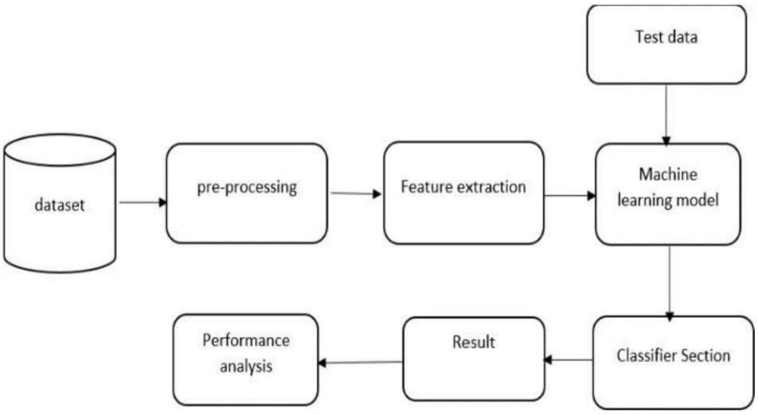


Figure1: General operating flow chart of system

1. DATA COLLECTION

Data used in this paper is a set of product reviews collected from credit card transactions records. This step is concerned with selecting the subset of all available data that you will be working with. ML problems start with data preferably, lots of data (examples or observations) for which you already know the target answer. Data for which you already know the target answer is called labelled data.

A dataset for machine learning typically consists of a collection of structured or unstructured data points that are used to train a machine learning model. The dataset may include various types of data, such as numerical, categorical, or text data, depending on the problem being addressed. In supervised learning, the dataset contains both input data and corresponding output labels, which are used to train the model to make predictions on new data. In unsupervised learning, the dataset may only contain input data, and the model is trained to identify patterns or relationships in the data.The dataset may also include features or variables that are used to describe each data point, such as transaction amount, date and time, location, or customer information. In addition, the dataset may contain metadata, such as data source, data quality, or data format.To ensure the quality and reliability of the dataset, it is often subjected to various preprocessing steps, such as cleaning, normalization, and feature engineering, before being used to train a machine learning model. Finally, the dataset is typically split into training, validation, and testing sets, to evaluate and optimize the performance of the model.

2. DATA PREPROCESSING

Preprocessing data is required before implementing a machine learning algorithm, considering various models produce diverse specifications to the predictors, and data training can affect predictive production. Data preprocessing purposes are to clean and prepare the data to a spot that comprises more concise prejudice, checking for missing values, and more variation. Data contains both numerical and categorical, which means encoding the categorical data is necessary before

DOI: 10.17148/IJARCCE.2023.125146

using them for modeling. Outlier detection and removal was performed. We have the independent variables in the same range by performing feature scaling. To reduce feature skewness, a box-cox transformation was carried out. Resampling method such as undersampling and oversampling was performed on the imbalanced original dataset to avoid any form of bias and overfitting in our training model. We have adopted Python data manipulation library pandas and machine learning library sci-kit learn to achieve these preprocessing responsibilities.

[A] DATA CLEANING

The credit card dataset was imported using the python import command, and the data cleaning process was done. During data cleaning we perform two tasks; 1. Remove null values and missing values, and 2. Handle outliers. The dataset contains 1048575 transactions in total. There were no null values in the dataset. Also, our dataset does not have any missing value. Hence, next we look for outliers in the dataset. Outliers are known as the observations that are numerically distant from the rest of the data. The boxplot technique was adopted to detect the presence of outliers in all the independent features. An outlier is a data point located outside the box plot's whiskers. However, for simplicity we only show the box plot for the feature “amount.

Although the box plots show the presence of outliers in the data, the outliers were removed using the Inter Quantile Range (IQR) technique which is one of the most popular techniques for handling outliers as it is more robust to outliers. In this technique, any value that is outside the Q3 + 1.5 IQR boundary is considered to be an outlier and, any outlier is discarded to make the machine learning models more robust and accurate.

[B] ENCODING CATEGORICAL VARIABLES

After cleaning the dataset, we convert any categorical features to a numeric value as most machine learning algorithms perform better with numeric inputs. There are few ways to convert categorical values into numeric values with each approach having its own tradeoffs and impact on the feature set. In the study, we have used One-Hot Encoder to convert the categorical variables to numeric values. For a feature with two categories, the categories 18 are assigned a numeric value of 1 or 0.

[C] FEATURE SCALING

This is another stage of the data preprocessing method used to normalize the range of independent variables within a dataset. Depending on the adopted scaling technique, it is centered around 0 or in the range of 0 and 1. If input variables have tremendous values applicable to the additional input variables, these large values can overlook or skew some machine learning algorithms. We have performed feature scaling using the Robust Scaler technique, also known as robust standardization. Scaling can be achieved by calculating the median 50th percentile, the 25th, and 75th percentiles. The values of each variable then have their median subtracted and are divided by the interquartile range (IQR), which is the difference between the 75th and 25th percentiles.

[D] DATASET RE-SAMPLING

Data resampling is a technique of inexpensively using a data sample to improve the accuracy and measure the unpredictability of a population variable. The nested resampling method has been used to carry out dataset resampling. The dataset used for this study was highly imbalanced; that is why we have carried out resampling methods like Undersampling and Oversampling.

[E] UNDERSAMPLING

Since most of the instances in the dataset belong to the majority class, the dataset was under-sampled randomly, by reducing the numbers of instances of the majority class, which means that some essential data instances are not captured for training purposes in the data.

[F] OVERSAMPLING

This method duplicates new or sometimes simulates examples in the minority class. It increases the instances, which makes the training of the model to perform better.

3. FEATURE EXTRACTION

Each of the features we obtain in the dataset might not be beneficial in building a machine learning model to execute the necessary prediction. Using some of the features might improve the prediction accuracy. So, feature correlation performs a tremendous purpose in creating a better machine learning model. Features with high correlation are more likely to be linearly dependent and have almost the same impact on the dependent variable. Therefore, when two features produce a high correlation, we can drop one of the two features. The heatmap for the correlation of the original dataset, and resampled dataset (both undersampled, and the oversampled) is shown in Fig. 9, and 10. It can be observed that the

DOI: 10.17148/IJARCCE.2023.125146

heatmap is not revealing too much information because it's a huge dataset, and that is why we performed feature selection to help select the important features. Feature selection is one of the important stages in data preprocessing, and it is known as a path to capture relevant features for use in the implementation of the machine learning model to expedite the training period and improve the learning interpretability and decrease the model over-fitting when there are many unnecessary features contributing no more helpful information than the current subset of variables. The excessive and verbose information in the dataset may hugely influence the performance of our model. In this study, we have performed feature selection using the lasso technique, which is a tool that helps minimize the cost function. Lasso regression will automatically choose the features that are beneficial to our model, discarding the redundant features. So, the purpose of using Lasso regression for feature selection goals is straightforward: we apply a Lasso regression on our scaled dataset, and we admit only those features that produce a coefficient different from 0.

4. EVALUATION MODEL

Model Evaluation is an essential part of the model development process. It helps to find the best model that represents our data and how well the selected model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can effortlessly generate overoptimistically and over fitted models. To avoid overfitting, evaluation methods such as hold out and cross-validations are used to test to evaluate model performance. The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is well-defined as the proportion of precise predictions for the test data. It can be calculated easily by mathematical calculation i.e. dividing the number of correct predictions by the number of total predictions.

Evaluated parameters:

1. ACCURACY

Accuracy is the ratio of the correct prediction number to the total number of input samples. It functions admirably just if there are an equivalent number of samples having a place with each class. For instance, consider 98% examples of class A and 2% examples of class B in our training set. Then, at that point, our model can undoubtedly get 98% accuracy by basically anticipating each training sample to be allied to class A. When a similar model is tried on a test set with 60% examples of class A and 40% examples of class B, then, at that point, the test accuracy would be reduced to 60%. Classification Accuracy is extraordinary; however, it gives us the misguided feeling of accomplishing high precision.

$Accuracy = Number\ of\ Correct\ predictions\ /\ Total\ number\ of\ predictions\ made$

2. RECALL

Recall can be calculated when the correct positive number results are divided by the number of all samples, which should have been recognized as a positive value.

$Recall = True\ Positives/True\ Positives+False\ Negatives$

3. PRECISION

Precision is dividing the correct positive number results by the number of positive results that the classifier predicted.

$Precision = True\ Positives/True\ Positives+False\ Positive$

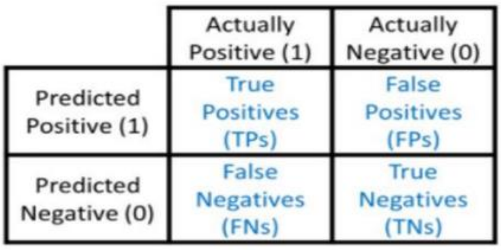
4. F1-SCORE

F1-score is used to evaluate the test's accuracy. It is the consonant mean between recall and precision. It allows a report on how precise the Classification is and how strong it can be. If a result gives high precision but low recall, it means we have incredibly high accuracy but note; it may miss a very high number of possibilities that are hard to classify. In short, it means the higher the F1 score, the best the model performed. It can be calculated using

$F1 = 2 \times 1\ 1\ precision+\ 1\ recall$

5. CONFUSION MATRIX

Confusion Matrix gives us a complete breakdown of the model performance in terms of matrix output. It evaluates well, especially when working with a binary classification where we have samples that belong to two classes: TRUE or False, YES or NO



	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure2: Confusion matrix

6. ROC AUC Score:

ROC (Receiver Operating Characteristics) AUC (Area Under Curve) is a widely used metric for model evaluation. AUC is the degree of measurement for separability, which reports how the model can differentiate between classes. Classification problems should measure performance with different thresholds been set. A better model can predict 0 classes as 0 and 1 classes as 1, while this can be confirmed if the AUC score is high. ROC is the curve probability [26]. This ROC curve plots the TPR (True Positive Rate) y-axis against the FPR (False Positive Rate) x-axis.

TPR (True Positive Rate) / Recall /Sensitivity = TP TP+FN Specificity = TN TN+FP FPR = 1 − Specificity = FP TN+FN

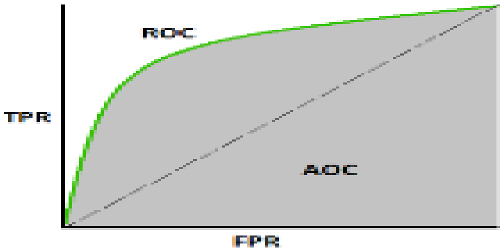
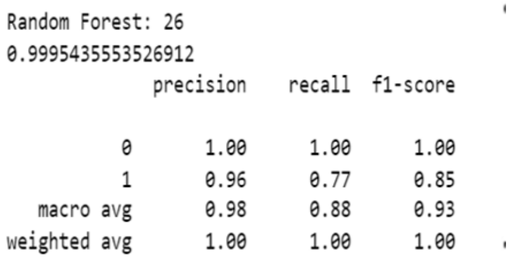


Figure3: ROC Curve

A. Random Forest Algorithm

One of the natural learning algorithm is the Random Forest algorithm. This algorithm is used for regression and classification problem solving. Classification problems are solved primarily using this algorithm. Decision tress are created from Random forest algorithm and each sample data is predicted from it. This algorithm performs single decision trees because it reduces overfitting by averaging the outcome. Hence it is called as the ensemble method.



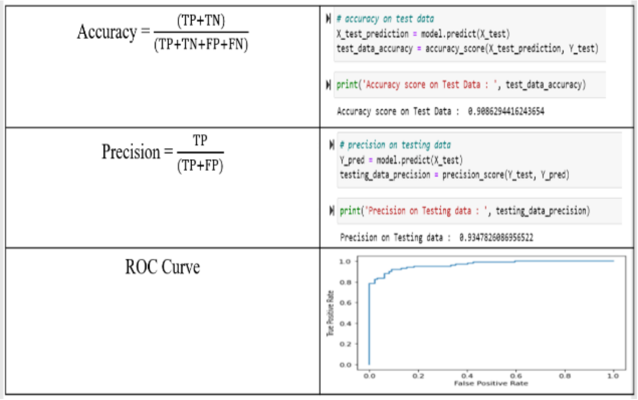
```
Random Forest: 26
0.9995435553526912
                precision    recall  f1-score

           0       1.00      1.00      1.00
           1       0.96      0.77      0.85
   macro avg       0.98      0.88      0.93
weighted avg       1.00      1.00      1.00
```

Figure4: Various parameters evaluated for random forest

B. Logistic Regression

Logistic regression performs both regression and classification tasks. Categorical variables are predicted by logistic regression using dependent variables. Sigmoid function or the logistic function employs logistic regression which is one of the most complex cost function. To be linearly related logistic regression does not require variables which are independent and also variance is equal within each group making it less constricted to statistical analysis procedure. As a result the likelihood credit card fraud transactions is employed by this algorithm.

DOI: 10.17148/IJARCCE.2023.125146



$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

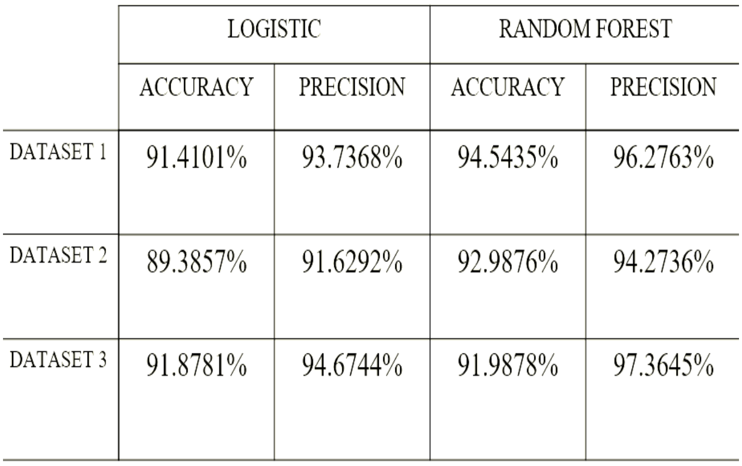
```
# accuracy on test data
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)

print('Accuracy score on Test Data : ', test_data_accuracy)

Accuracy score on Test Data :  0.9086294416243654
```

$$\text{Precision} = \frac{TP}{(TP+FP)}$$

```
# precision on testing data
Y_pred = model.predict(X_test)
testing_data_precision = precision_score(Y_test, Y_pred)

print('Precision on Testing data : ', testing_data_precision)

Precision on Testing data :  0.9347826086956522
```

ROC Curve

Figure4: Various parameters evaluated for logistic regression



	LOGISTIC		RANDOM FOREST	
	ACCURACY	PRECISION	ACCURACY	PRECISION
DATASET 1	91.4101%	93.7368%	94.5435%	96.2763%
DATASET 2	89.3857%	91.6292%	92.9876%	94.2736%
DATASET 3	91.8781%	94.6744%	91.9878%	97.3645%

Figure 6: Comparison table of algorithms with different datasets

III. CONCLUSION

This review study looks into the many methods used. Conclusion: ML approaches are a great tool to increase the precision of detection methodologies. The model must be trained on huge datasets in order to avoid data imbalance. Real-time datasets can give us access to a wider variety of data, but privacy is still an issue. In order to train the model while protecting privacy, we are taking into account the real-time datasets accessible. The suggested approach can help financial institutions and banks work together to use real-time datasets, which would be beneficial for everyone in terms of creating a system that is effective at detecting fraud. Despite its effectiveness, the proposed method has limits when it comes to real world deployment because it takes a lot of time and engineering resources to integrate, and even then, the outcome is still not ideal because it only uses a small portion of the total data available. Because each bank and finance institution have its own restrictions and relies on internal resources rather than a centralised strategy, adapting the suggested method will be challenging. As a result, even with the constraints still there, more needs to be done to convince banks and other financial organisations to adopt this technology.

REFERENCES

[1] Saiju, Sanisa, S. Akshaya Jyothy, Christeena Sebastian, Liss Mathew, and Tintu Sabu. "Credit Card Fraud Detection Using Machine Learning." *International Journal of Recent Advances in Multidisciplinary Topics* 2, no. 4 (2021): 31-34

[2] Arafath, Yeasin, Animesh Chandra Roy, M. Shamim Kaiser, and Mohammad Shamsul Arefin. "Developing a Framework for Credit Card Fraud Detection." In *Proceedings of the International Conference on Big Data, IoT, and Machine Learning*, pp. 637-651. Springer, Singapore, 2022.

DOI: 10.17148/IJARCCE.2023.125146

[3] Awoyemi, John O., Adebayo O. Adetunmbi, and Samuel A. Oluwadare. "Credit card fraud detection using machine learning techniques: A comparative analysis." In *2018 international conference on computing networking and informatics (ICCNI)*, pp. 1-9. IEEE, 2018.

[4] Ileberi, Emmanuel, Yanxia Sun, and Zenghui Wang. "A machine learning based credit card fraud detection using the GA algorithm for feature selection." *Journal of Big Data* 9, no. 1 (2022): 1-17.

[5] Sadineni, Praveen Kumar. "Detection of fraudulent transactions in credit card using machine learning algorithms." In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pp. 659-660. IEEE, 2020.

[6] Sanobar khan, Sanovar, Suneel Kumar , Mr Hitesh Kumar(2021);Credit Card Fraud Detection Using ML; International Journal of Scientific and Research Publications(IJSRP).

[7] Sharma, Pratyush, Souradeep Banerjee, Devyanshi Tiwari, and Jagdish Chandra Patni. "Machine learning model for credit card fraud detection-a comparative analysis." *Int. Arab J. Inf. Technol.* 18, no. 6 (2021): 789-796.

[8] Meenakshi, B. Devi, B. Janani, S. Gayathri, and N. Indira. "Credit card fraud detection using random forest." *International Research Journal of Engineering and Technology (IRJET)* 6, no. 03 (2019).

[9] Priya, G. Jaculine, and S. Saradha. "Fraud detection and prevention using machine learning algorithms: a review." In 2021 7th International Conference on Electrical Energy Systems (ICEES), pp. 564-568. IEEE, 2021.

[10] Azhan, Mohammed, and Shazli Meraj. "Credit card fraud detection using machine learning and deep learning techniques." In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pp. 514-518. IEEE, 2020.