Turkish Journal of Computer and Mathematics Education Vol.12 No.14 (2021),6195-6204
Research Article

Credit Card Fraud Detection on Class Imbalance Dataset

Neha Purohit and Dr. Rajeev G. Vishwakarma
Department of Computer Science & Engineering, Dr. A.P.J. Abdul Kalam University, Indore (M.P.) 452010, India
Corresponding Author Email : nehapurohit059@gmail.com

Abstract— India is growing day by day and a number of enhancements to banking and finance are performed by the government. In this context, the government is frequently supporting digital payments for large as well as small transactions. However, it increases the transparency in payments but in the same ratio, the financial fraud cases are increasing. Among them, credit card fraud is a very common and frequent fraud in the banking system. However, there are a number of automated systems for credit card fraud detection available, but most of them are suffering from the class imbalance problem. The imbalanced training samples are misleading the Machine Learning (ML) algorithm, which leads to an increase in false alarm rates. In this paper, our aim is to contribute an ML method, which is able to deal with the class imbalance issue. Additionally, accurately identify fraud cases. In this context, first, we discuss the class imbalance issue and its available solutions. Then, adopt two appropriate over-sampling methods for handling the class imbalance i.e. ADASYN and SMOTE. Finally, a Binary Convolutional Neural Network has been implemented to classify the over-sampled dataset to classify transactions into fraud and legitimate. The experimental analysis of the model has been carried out based on the Kaggle dataset. The performance results of the proposed technique in terms of accuracy and Area Under the Precision-Recall Curve (AUPRC) are evaluated. According to the obtained results, we found the proposed methodology is enhancing the results and produce up to 99% accurate results.

Keywords—Machine Learning, Classification, Credit Card Fraud Detection, Machine Learning Application, Supervised and Unsupervised Learning, Comparison.

I. INTRODUCTION

India is a country where a large number of transactions are done using digital platforms. In everyday life, more than 60% of the population utilizes digital transactions. Among various payment channels like Unified Payments Interface (UPI), debit cards, and online banking, the credit card is also a frequently used payment method in India. However, all the discussed payment methods are secure but there is a significant amount of possibility to get fraud. In this context, we need a security technique that is able to capture the patterns of credit card usage and estimate possible fraud cases [3]. In recent years a number of contributions are placed to preventing fraud most of them based on machine learning techniques. However, machine learning techniques have the ability to deal with a large amount of data and perform accurate data analysis but the class imbalance problem on the training datasets is a complex issue to deal with for accurate data classification.

In this paper, the main aim is to study the class imbalance problems in credit card fraud detection. The class imbalance problem is a classical machine learning issue, which misleads the classifier and increases the misclassification rate of the machine learning classifiers. Therefore, we need an enhanced method, which will help to deal with this issue and improve the fraud detection rate. In this context, in the next section, we provide an overview of the class imbalance dataset problem. Additionally, the available solution to this problem is also discussed. The next section described the proposed model for credit card fraud detection and their working process in detail. Further, the experimental results are discussed and the measured performance is described. Finally, the conclusion will be made on the basis of experiments and design observation. Additionally, future work is also discussed.

II. RESEARCH BACKGROUND

This section involves the study of the class imbalance problem in machine learning, additionally, the different available solution to deal with it is also discussed.

During data analysis when we are working with binary classification we encounter an imbalanced classification problem. Imbalanced data means the number of rows of a class is more than the other class. In other words, the ratio of the counts of classes is much higher. Such a data set is known as an imbalanced dataset. The imbalanced dataset makes the classification more challenging when we build a classifier with such data it works well with the majority class but gives a poor performance with the minority class. Some Machine Learning algorithms are more sensitive toward imbalanced data, such as Logistic Regression. However, some algorithms tackle this issues self, such as Random Forest. There are two techniques available to handle this issue:

- Under Sampling
- Over Sampling

A. What is sampling

The sampling is method to create new samples or select appropriate samples from the given data set.

B. Under Sampling

Under-sampling techniques help in balancing the class distribution for distorted class distribution. Under-sampling techniques eliminate some examples from the training data set belonging to the majority class. It is to better balance the class distribution by reducing the skewness. It can also be used with the combination of the Over-sampling technique. The combination of both techniques provides better results than utilizing them individually. The Under sampling technique removes the samples randomly from the majority class, which is known as 'random under sampling'. But, there is a risk of losing useful information, which is able to distinguish between the classes. Thus, we need a heuristic approach to select samples for non-deletion. Some under-sampling techniques use such heuristics. Some of them are discussed in this section.

- **Near Miss Under sampling:** This technique selects the data based on the distance between majority and minority class samples. It has three types, and each considers the different neighbours from the majority class.

- ✓ Type 1 keeps samples with a minimum average distance to the nearest samples of the minority class.
- ✓ Type 2 selects instances with a minimum average distance to the samples of the minority class.
- ✓ Type 3 keeps examples from the majority class of closest record in the minority class.

Among them type 3 is more accurate since it considers examples of the majority class that are on decision boundary.

- **Condensed Nearest Neighbor (CNN):** This technique is aimed to find a subset of samples, which minimizes the loss. This technique store those samples that consist of samples from the minority class and incorrectly classified from the majority class.
- **Tomek Links Under sampling:** It is a modified version of CNN in which the redundant majority class examples are selected for deletion. These examples are near the decision boundary.
- **Edited Nearest Neighbours Under sampling:** This technique uses the nearest neighbours approach to delete the misclassified samples. It computes three nearest neighbors of each instance if the sample of a majority class is misclassified by these three neighbors then it is removed. If the instance is of the minority class and misclassified by the three nearest neighbors, then its neighbors from the majority class are removed.
- **One-Sided Selection Under sampling:** This technique combines Tomek Links and the CNN. Tomek links remove the noisy examples, and CNN removes the isolated examples from the majority class.
- **Neighborhood Cleaning Under sampling:** It is a combination of CNN and ENN. It selects all the minority examples. Then ENN identifies the ambiguous samples to remove from the majority class. Then CNN deletes the misclassified examples.

C. Oversampling

Oversampling focuses on increasing minority class samples. We can also duplicate the examples to increase the minority class samples. However it balances the data, but it does not provide information for classification. Therefore synthesizing new examples is necessary.

- **Synthetic Minority Oversampling Technique (SMOTE):** It selects the samples in the feature space, then draws a line between them, and at a point along the line, it creates a new sample. In other words, it picks an instance randomly from the minority class and finds its k nearest neighbours from minority class. Then one of the neighbours gets selected randomly between these two instances to generate a combination of these two instances.
- **Borderline-SMOTE:** This method selects minority class instance that is misclassified with a k-nearest neighbor (KNN). Since borderline examples are tend to misclassified.
- **Borderline-SMOTE SVM:** This method selects the misclassified instances of Support Vector Machine (SVM).
- **Adaptive Synthetic Sampling (ADASYN):** This approach works according to the density of the minority class instances. Generating new samples is inversely proportional to the density of the minority class. It generates more samples where minority class examples density is low and fewer samples in the high-density space.

III. PROPOSED WORK

The proposed work is aimed to develop an accurate credit card fraud detection model. In this context, it is essential to obtain a suitable experimental dataset. Therefore we have identified and downloaded a number of datasets from different sources. During this, we found all the dataset has similar attributes and data samples. Among them, the dataset available in Kaggle has the most proper and explained set of information. This dataset contains credit card transactions of European cardholders. This dataset has the transactions made in September 2013 and has 492 fraud transactions among a total of 284,807 transactions. Due to security reasons, it contains the result of Principle Component Analysis (PCA) to retain undisclosed original features. The only features which have not been transformed are 'Time' and 'Amount'. The consequences of transactions are given as a 'Class' variable. It has a value of 1 for fraud and 0 for legitimate.
Due to PCA-based transformed values, the data has "+" positive as well as negative "-" values. Therefore, we scale the entire dataset values between 0 to 1. In this context, we utilized min-max normalization. Additionally, the dataset has a total of 30 attributes means the dataset has large dimensional data. Therefore, we have performed a chi-square test between the class attribute and the other dataset attributes. The chi-square test provides two resultant variables chi-square score and p values of the attributes. The sorted p values-based attribute ranking is demonstrated in figure 1. According to the obtained p values the attributes 'V11', 'V4', 'V14', 'V12', 'V17', 'V16', 'V18', and 'V10' has less significant as compared to others. Therefore, we have removed the less significant attributes, and the remaining 22 attributes are going to be used for further processing.

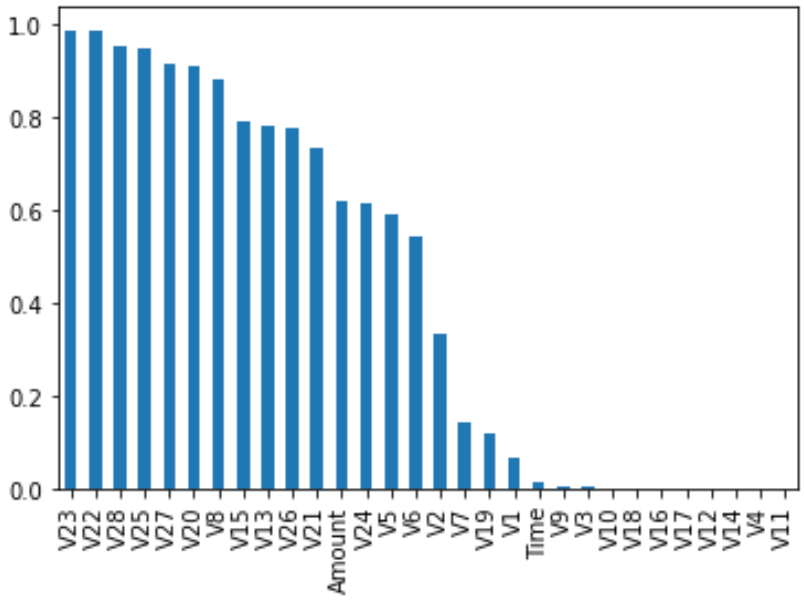

Figure 1: Shows the ranking of attributes based on p value

Now, we check the balance of the dataset in terms of class labels. During the evaluation of the dataset, we found 99.83% of data samples are belongs to legitimate class '0', and 0.17% of data samples are belongs to class '1'. This data is highly imbalanced. Figure 2 demonstrates the majority and minority classes available in the dataset.

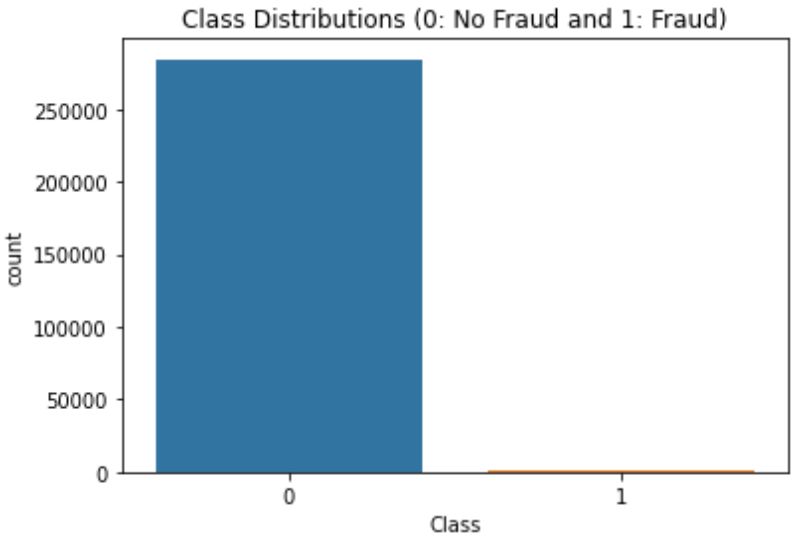

Figure 2: Dataset class distribution ratio

Therefore, we need to perform balancing of the class distribution. In this context we aimed to balance the classes using the sampling technique. In this scenario we can use under sampling approach but due to risk of information loss in place of under sampling we are implemented here the over sampling technique. Thus we have implemented Adaptive Synthetic Sampling (ADASYN) and Synthetic Minority Oversampling Technique (SMOTE). After applying the ADASYN technique the total samples 564546 instances of samples which is further divided into 75% of training samples and 25% of testing samples we got 426409 samples for training and 142137 samples for testing. Additionally, using the SMOTE technique the total samples are becomes 568630 instances. Additionally the training samples 75% becomes 426472 instances and 25% of samples are becomes 142158 instances. The class distribution generated by ADASYN is demonstrated by figure 3.

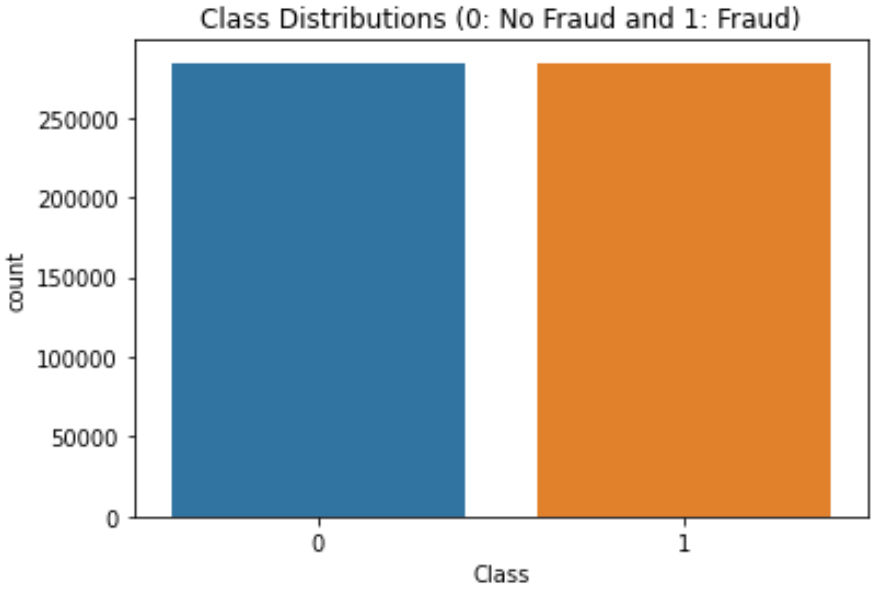

Figure 3: Class distribution after over sampling using ADASYN

Additionally, the samples generated by SMOTE based class distribution are given in figure 4.

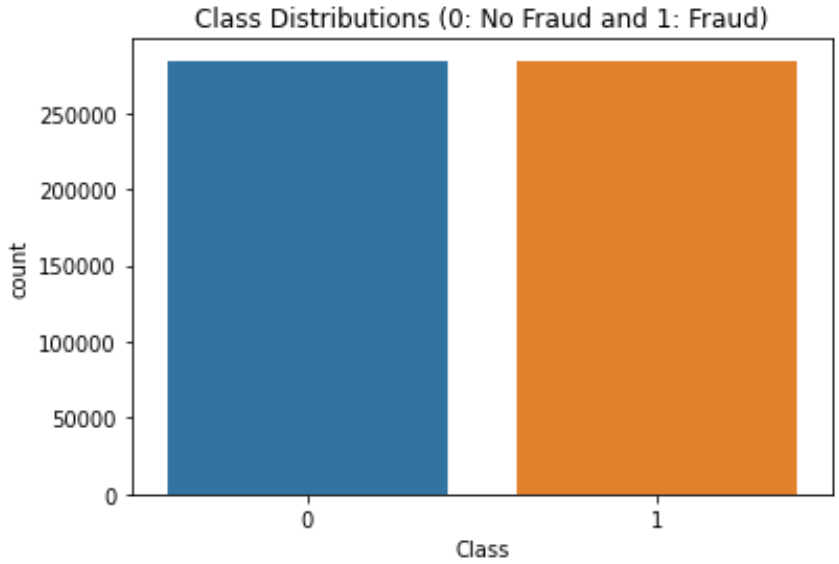

Figure 4: Class distribution after SMOTE sampling

After samples created we have configured a sequential Convolutional neural network. This network has the following layers:

- Input Layer: type dense, number of neurons = 100, input dimensions = 22, activation function = ReLu
- Hidden layer 1 : type dense, number of neurons = 64, activation function = sigmoid
- Hidden layer 2 : type dense, number of neurons = 32, activation function = ReLu
- Hidden layer 3 : type dense, number of neurons = 16, activation function = sigmoid
- Hidden layer 4 : type dense, number of neurons = 8, activation function = ReLu
- Hidden layer 5 : type dense, number of neurons = 4, activation function = sigmoid
- Output layer : type dense, number of neurons = 2, activation function = SoftMax

Additionally, to compile the configured network we utilize the following network properties:

- Loss Function = 'categorical_crossentropy'
- Optimizer = 'adam'
- Metrics = 'accuracy'

The same configuration of neural network will be used for both kinds of samples prepared by SMOTE and ADASYN over sampling technique. This section provides the details about the prepared credit card fraud detection technique. The next section discusses the experimental results of the prepared system.

IV. RESULTS & DISCUSSION

The proposed work is motivated to handle the class imbalance problem in the credit card fraud detection dataset. Thus, we have implemented two over-sampling techniques to deal with this issue. Additionally, a CNN model is trained to learn and identify credit card fraud transactions. In this section, we provide the performance of the implemented model by using both kinds of sampling methods. The obtained performance of the model is given in figure 5 in terms of

accuracy and loss. The accuracy is the ratio of correctly recognized information and total information produced for recognition. The accuracy can be measured using the following equation:

$$accuracy = \frac{correctly\ recognized}{total\ samples}$$

Additionally, loss is defined as the heuristics which provide the information about the convergence of the solution. In other words, the loss is a measurement of distance between the current solution and expected solution.

Figure 5(A) shows the accuracy of the model with the samples generated by both sampling methods for the training of the model. Similarly, figure 5(B) demonstrates the accuracy of the validation data. Next, figure 5(C) shows the loss obtained during the training additionally figure 5(D) shows the validation loss. According to the obtained results, we can found the performance of credit card fraud detection increases with the number of epoch cycles. Additionally, the samples generated by the SMOTE are better than the ADASYN method. However, the accuracy of the model is found near about 99%, but for evaluation of the class, imbalance datasets are sometimes biased against the predicted classes. Therefore, researchers are recommending evaluating the model's performance under the precision and recall curve.

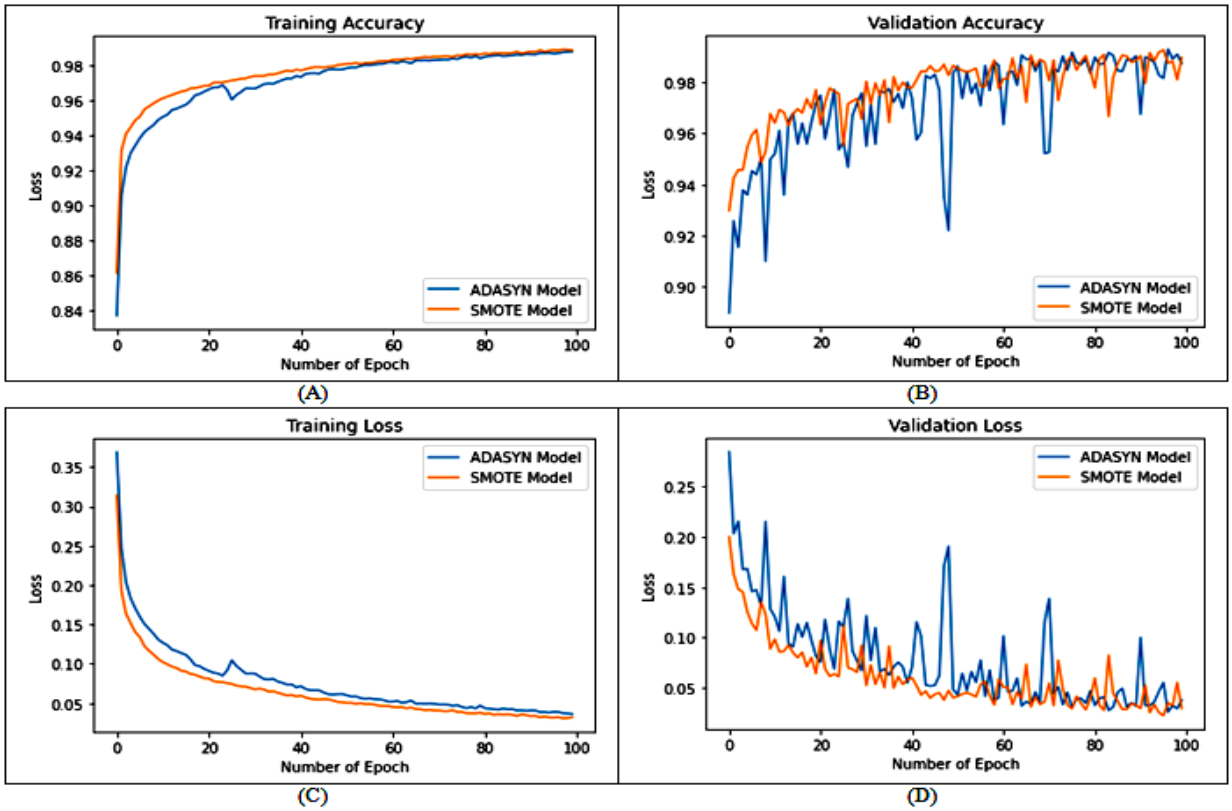

Figure 5: Experimental results of the implemented credit card fraud detection technique in terms of (A) training accuracy (B) validation accuracy (C) training loss and (D) validation loss

Precision is also known as positive predictive value. That is the fraction of relevant instances among the retrieved instances. Precision is defined as follows:

$$precision = \frac{TP}{TP + FP}$$

Recall is also known as sensitivity or true positive rate and is defined as follows:

$$recall = \frac{TP}{TP + FN}$$

The Area Under the Precision-Recall Curve (AUPRC) is demonstrated in figure 6. In this figure we can see the results of both kind of samples among them blue line shows the area of ADASYN and orange line shows the area of SMOTE. But we can see the orange line outlined on blue area which shows the superiority of SMOTE model.

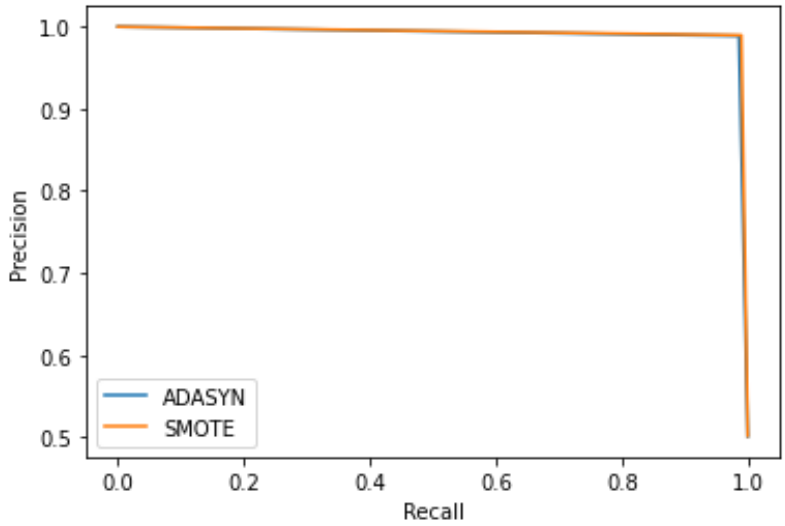

Figure 6: Area Under the Precision-Recall Curve (AUPRC) for credit card fraud detection

V. CONCLUSIONS

The credit card frauds are one of the crucial issues for the financial institutions. The loss causes by the fraud transactions are mostly fulfilled by the financial companies. Therefore, the companies monitor the credit card transactions. Additionally, to secure the transactions machine learning technologies are implemented. But, most of the techniques developed for this task have suffered from class imbalanced issue due to fewer amounts of fraud techniques are available on databases. Therefore, we need to deal with class imbalance issue. In this context, in this paper we provide the following key insights.

1. Provided a study on the different sampling technique to deal with the class imbalance issue
2. Provided detailed analysis of credit card dataset
3. Implement the over sampling techniques and performed comparative study
4. Configured deep neural network architecture for accurately classify the credit card fraud transaction detection

Based on the experimental analysis of the imbalance dataset we found the SMOTE is superior then ADASYN algorithm for dealing with this problem. Finally, in near future we provide more in depth analysis and study of the credit card fraud detection factors in real world scenario.

REFERENCES

1. Sangeeta Mittal and Shivani Tyagi, "Chapter 26: Computational Techniques for Real-Time Credit Card Fraud Detection", Handbook of Computer Networks and Cyber Security, © Springer Nature Switzerland AG 2020
2. G. Sasikala, M. Laavanya, B. Sathyasri, C. Supraja, V. Mahalakshmi, S. S. Sreeja Mole, Jaison Mulerikkal, S. Chidambaranathan, C. Arvind, K. Srihari, and Minilu Dejene, "An Innovative Sensing Machine Learning Technique to Detect Credit Card Frauds in Wireless Communications", Hindawi Wireless Communications and Mobile Computing Volume 2022, Article ID 2439205, 12 pages
3. Ibtissam Benchaji, Samira Douzi, and Bouabid El Ouahidi, "Credit Card Fraud Detection Model Based on LSTM Recurrent Neural Networks", Journal of Advances in Information Technology Vol. 12, No. 2, May 2021
4. Aishwarya Mujumdar, Dr. Vaidehi V, "Diabetes Prediction using Machine Learning Algorithms", Procedia Computer Science 165 (2019) 292–299
5. Naoufal Rtaylia, Nourddine Enneya, "Selection Features and Support Vector Machine for Credit Card Risk Identification", Procedia Manufacturing 46 (2020) 941–948
6. Praveen Kumar Sadineni, "Detection of Fraudulent Transactions in Credit Card using Machine Learning Algorithms", Proceedings of the Fourth International Conference on I-SMAC, 978-1-7281-5464-0/20/$31.00 ©2020 IEEE
7. Olawale Adepoju, Julius Wosowei, Shiwani lawte, Hemaint Jaiman, "Comparative Evaluation of Credit Card Fraud Detection Using Machine Learning Techniques", 2019 Global Conference for Advancement in Technology (GCAT), 978-1-7281-3694-3/$31.00 ©2019 IEEE
8. Dejan Varmedja, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, Andras Anderla, "Credit Card Fraud Detection - Machine Learning methods", 18th International Symposium Infoteh-Jahorina, 20-22 March 2019, 978-1-5386-7073-6/19/$31.00 ©2019 IEEE
9. Abdul Mohaimin Rahat, Abdul Kahir, Abu Kaisar Mohammad Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset", 8th International Conference on System Modeling & Advancement in Research Trends, 22nd–23rd November, 2019 Copyright © IEEE–2019 ISBN: 978-1-7281-3245-7
10. Sai Kiran, Jyoti Guru, Rishabh Kumar, Naveen Kumar, Deepak Katariya, Maheshwar Sharma, "Credit card fraud detection using Naïve Bayes model based and KNN classifier", International Journal of Advance Research, Ideas and Innovations in Technology, Volume 4, Issue 3, 2018
11. Imane SADGALI, Nawal SAEL, Nawal SAEL, "Fraud detection in credit card transaction using neural networks", SCA2019, October 2–4, 2019, CASABLANCA, Morocco, © 2019 Association for Computing Machinery

12. Tzu-Hsuan Lin and Jehn-Ruey Jiang, "Credit Card Fraud Detection with Autoencoder and Probabilistic Random Forest", Mathematics 2021, 9, 2683.
13. Dileep M R, Navaneeth A V, Abhishek M, "A Novel Approach for Credit Card Fraud Detection using Decision Tree and Random Forest Algorithms", Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV 2021). 978-1-6654-1960-4/21/$31.00 ©2021 IEEE
14. Ying Chen and Ruirui Zhang, "Research on Credit Card Default Prediction Based on k-Means SMOTE and BP Neural Network", Hindawi Complexity Volume 2021, Article ID 6618841, 13 pages
15. P. Shanmugapriya, R. Shupraja, V. Madhumitha, "Credit Card Fraud Detection System Using CNN", International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 10 Issue III Mar 2022