# Credit Card Fraud Detection using Machine Learning

B.Sai Charan Reddy
*Department of Computer Science and Engineering*
*Kalasalingam Academy of Research & Education*
Krishnankovil, TamilNadu
99210041146@klu.ac.in

N.V.Suneel Kumar Reddy
*Department of Computer Science and Engineering*
*Kalasalingam Academy of Research & Education*
KrishsnanKovil, TamilNadu
99210041597@klu.ac.in

B.V.Sudharshan
*Department of Computer Science and Engineering*
*Kalasalingam Academy of Research & Education*
Krishanakovil, TamilNadu
99210041454@klu.ac.in

B.Manideep Reddy
*Department of Computer Science and Engineering*
*Kalasalingam Academy of Research & Education*
Krishnankovil, TamilNadu
99210041018@klu.ac.in

Rajan Kumar Misra
*Department of Computer Science and Engineering*
*Kalasalingam Academy of Research & Education*
Krishnankovil, TamilNadu
99210041706@klu.ac.in

P. Anitha
*Assistant Professor*
*Department of Computer Science and Engineering*
*Kalasalingam Academy of Research & Education*
Krishnankovil, TamilNadu
p.anitha@klu.ac.in

***Abstract***—The modern world faces a serious problem with fraud. Losses resulting from payment fraud are rising as e-commerce develops further. Problems of fraud are occurring due to the loss of important information of user on online platform, online payments, loss of credit cards, fake email scams application fraud, cloned cards. Due to payment, businesses, governments, and people all suffered enormous losses. Credit card fraud causes the greatest loss of all payment fraud. As a result, we want to use machine learning to its best extent to address the issue of credit card fraud, which also applies to other types of fraud. Based on accuracy and f1_score, this study evaluates the effectiveness of logistic regression, decision trees, and random forest classifiers. In order to deal with the unbalanced composition of the data, we used a smote technique. We then compared the results of the supervised models on the oversampled data to the performance of the raw data.

***Keywords—*** *Credit card fraud, Logistic Regression, Decision Tree Classifier, Random Forest, Supervised algorithm, Fraud Detection.*

## I. INTRODUCTION

Rapid innovation in payment methods for products and services has resulted from technological advancement, allowing for quicker and more comfortable payment options. The era of only physical money being used around the world is long gone. A particular form of financing is a credit card. Credit cards are little cards that credit organizations offer to their authorized clients as a way of payment. It allows cardholders the convenience of making purchases and making payments to the business on a predetermined billing period [14].

When customers are cheated by the unknown, unrelated people then this comes under the fraud category. Every physical device that works for money transaction records the data of user such as balance, credit's, debits, and password etc. This is the information that is required to make scams on the credit cards. This can be identified and reduced by using some modern machine learning algorithm, among those random forest is highly preferable.

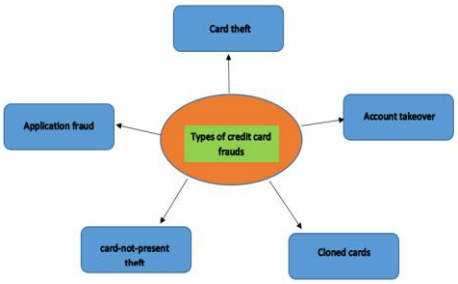


*Figure 1 Types of Credit Card Frauds*

In the present years, the majority of the population uses the credit card for the money transaction to purchase their necessaries. As middle-class population in high in the society the usage of credit cards is high when compared to the debit card. Credit cards is uses by customers when money is insufficient to buy their necessaries at the same time occurrence of fraud is happens. These frauds can be identified by using some machine learning algorithm and that can be prevented at the time of money transaction [15].

## II. RELATED STUDY

These are the related works done by other researchers.

1. Pratyush Sharma, Souradeep Banerjee, Devyanshi Tiwari, and Jagdish Chandra Patni research survey describes that ANN is the best algorithm that detects fraud transactions when compared

with the SVM and logistic regression machine learning algorithms. Paper used the SMOTE technique to increase the number of records in the dataset in balance way. By the comparison from literature research and analysis of the paper Random Forest showed to be more effective with F_1 score - 0.85 for training dataset while ANN with F_1 score of 0.91 Which is best for detecting fraudulent transactions also ANN can learn without need of reprogram.

2. C. Sudha, Dr.D. Akila proposed a system that extracts the features of users in first phase and then classify the features by using random forest classifier. In second phase, from user records extracts transaction features of users and classifies by using M-class SVM classifier. By the evaluation of the model's precision, accuracy, F_1 score, and recall concluded that both SVM and RF classifiers provides good accuracy to detect frauds.

3. Asha RB, Suresh Kumar KR proposed a method based on deep learning algorithm to predict credit card fraud transactions. It was challenging to train a model to identify the fraud transactions by using machine learning algorithms such as support vector machine and k-Nearest Neighbor. They developed a neural network model for optimising the credit card fraud detection with artificial neural network (ANN), of accuracy nearly 100%. Pre-processing, normalization and under-sampling are performed on the imbalanced data set to make it balance.

4. Ruttala Sailusha, V. Gnaneswar, R. Ramesh, G. Ramakoteswara Rao Research analysed that random forest and adaboost algorithms are of same accuracy while comparing. By considering precision, recall and F_1 score random forest achieved high value when compared to adaboost therefore concluded that random forest is best for fraud detection.

5. Yakub K. Saheed, Moshood A. Hambali, Micheal O. Arowolo, Yinusa A. Olasupo Research describes the efficiency of genetic algorithm is tested by using the naive bayes, random forest and support vector machine. This was done by taking two different attributes sets. the accuracy for naive Bayes is 94.3, random forest is 96.4 and support vector machine with 96.3 for first set of attributes. Where the accuracy of naive bayes is 64.2%, random forest is 60% and support vector machine with 59% for second set of attributes.

6. Yashvi Jain, Namrata Tiwari, ShripriyaDubey, Sarika Jain paper describes the idea of credit card fraud and its variants that have been discussed. Authors have described several methods for fraud detection systems, namely Bayesian Network, Artificial Neural Networks (ANN), Hidden Markov Model, K-Nearest Neighbor (KNN), Fuzzy Logic Based System, and Decision Trees. On the basis of statistical measure including accuracy, false alarm rate and detection rate, a complete analysis on the current and proposed models for credit card fraud detection has been conducted.

7. Naoufal Rtayli, Nourddine Enneya paper proposes hybrid method framework that provides three methods. The Synthetic Minority Oversampling Technique (SMOTE) to overcome the issue of unbalanced data, the Hyper-Parameters Optimization (HPO) approach to estimate the optimum hyperparameters to RFC, and the Recursive Features Elimination (RFE) method to minimize the number of features. Model is executed on three large datasets, and it shows the model's ability to perform with high accuracy during the CCFD process. Also, model ensures a very strong performance irrespective of the datasets used as compared to recent studies.

8. Rimpal R. Popat, Jayesh Chaudhary paper analyses that machine learning improved than previously used and prediction, grouping, outlier detection, etc. as there is high accuracy and detection rate, machine learning methods are mainly chosen in the fraud detection. Researchers continue hard to increase accuracy and detection rate. and the ultimate goal is to protect credit card transactions so that users of e-banking can secured.

9. S. Venkata Suryanarayana, G. N. Balaji, G. Venkateswara Rao paper describes the fraud detection models were developed to categorise a transaction as fraudulent or non-fraudulent using decision trees, K-Nearest Neighbor, logistic regression, and neural network methods. Their performances were assessed using metrics. The results indicated that no data mining technique is different than another. The development of a fraud detection model is performed with variety of data mining approaches could lead to performance improvement.

10. Andhavarapu Bhanusri, K. Ratna Sree Valli, P. Jyothi, G. Varun Sai, R. Rohith Sai Subash Since variable misclassification cost is present, this study goal was interpreted in different ways as how classification problems are typically approached. The performance of the suggested system is assessed using accuracy, precision, recall, the f1-score, and support. The random forest classifier with boosting technique outperformed the logistic regression and naive bayes methods, according to our comparison of the Naïve Bayes, Logistic regression, Random Forest approaches.

11. Vaishnavi Nath Dornadulaa, Geetha S paper descries The Matthews Correlation Coefficient was found to be the better parameter for handling imbalance datasets. There were other options besides MCC. Experimental with balancing the dataset using the SMOTE and observed that the classifiers were working more effectively than before. The use of one-class classifiers, such as one-class SVM, is an alternative method for handling imbalance datasets. Finally, observed that the algorithms that produced the best results were those that used logistic regression, decision trees, and random forests.

12. Ashraf Afifi and E.A. Zanaty, Said Ghoniemy paper describes that tested proposed kernel function with various data set sizes and attribute combinations. Experimental findings make it clear that RBF provides greater accuracy with smaller data sets than the Polynomial function. However, in large data sets, the polynomial kernel produces better results. because the proposed function combines the strength, the Gauss, and Polynomial functions, it achieves the best accuracy in almost all data sets and particularly in those with the most attributes. Therefore, for some particular datasets, the proposed kernel functions can be thought of as a good substitute for the Gaussian and polynomial kernel functions.

13. F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan and M. Ahmed, proposed a method which can predict the Credit Card Fraudulent. Random Forest algorithm performs better with large set of training data, for this it takes more pre-processing. While SVM algorithm has a problem of imbalanced data.

14. Amarachi Blessing Mbakwe and Sikiru Ademola Adewale paper describes that supervised algorithms are more used in Credit Card Fraud detection than the unsupervised algorithms. Based on the AUC, Random Forest algorithm is performing better than the other supervised algorithms like Logistic Regression, Decision Tree Classifier, etc. By oversampling the performance increasing.

## III. METHODOLOGY

We require a significant quantity of data on previous transactions made by bank clients in order to build a machine learning model that can predict credit card frauds just by analyzing the transaction. In order to protect the customer's confidential data, Principal Component Analysis (PCA), has now been performed to the dataset that is accessible from the bank. The flow of the Credit Card Fraud Detection looks like as shown in Fig. 2.

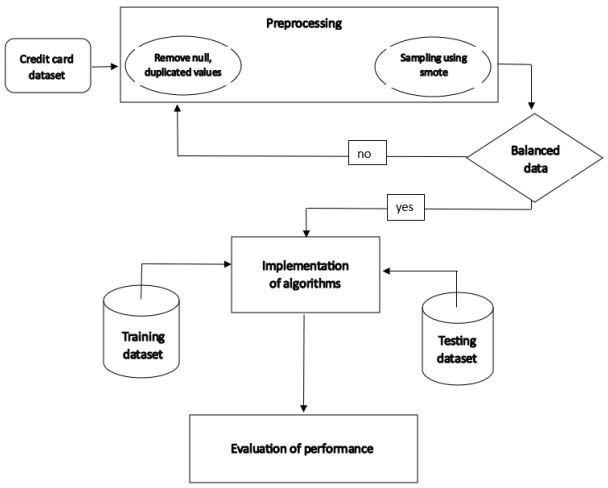


*Figure 2 Work flow*

### 3.1. Dataset Description

From Kaggle, an online platform for research competitions, the dataset is downloaded. To maintain user identities and sensitive features, the dataset's parameters comprise the quantity, time, class, and 28 PCA-transformed features. There are no empty rows, NAs, or missing data in the dataset.

For simplicity, the dataset's short statistics, interpretive statistical conclusions, and graphical representation are provided. We discovered that the dataset is unbalanced as a result of the data's structure. Principal component analysis transforms 28 columns, leaving out the modification of 3 columns. Time, Amount, and Class are the three columns that are not altered. The class is the response, and the other 30 columns serve as distinct variables, attributes, or variables of explanation after an in-depth review of the dataset. If the transaction is fraudulent, the outcome response has a value of 1, alternatively it has a value of 0.

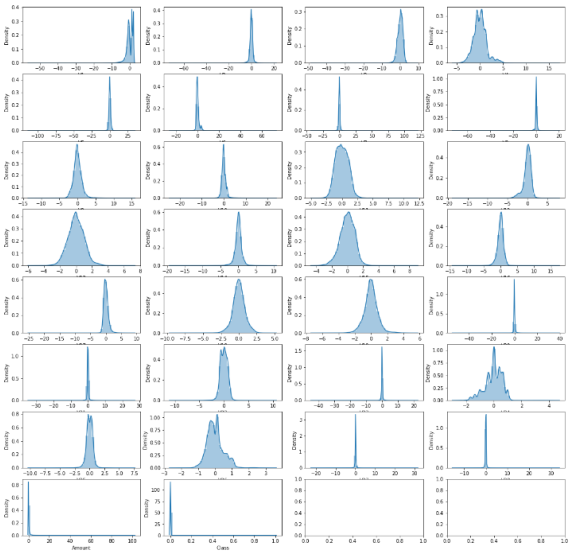

*Figure 3 Dist Plot of dataset*

3.2. Data Pre-processing

Pre-processing is the process of cleaning unwanted data, outliers etc. in the dataset. The main purpose of pre-processing is to increase the quality of the data and to enhance the performance of the model.

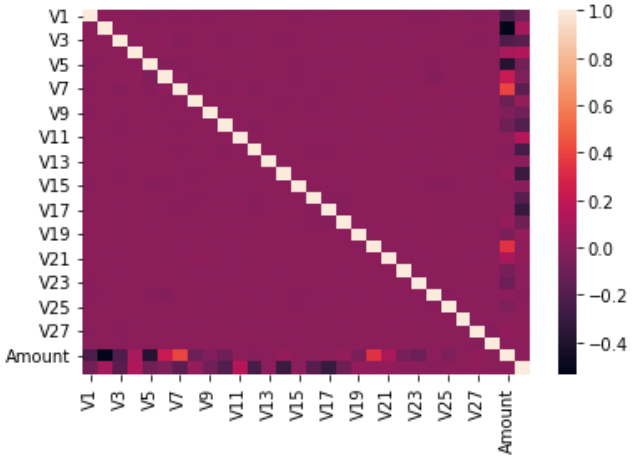


*Figure 4 Heat map of the dataset*

The Fig.4 shows heatmap that gives the information about dependencies or correlations among the attributes in the dataset. Removing these unrelated and negatively related attributes using drop function will make the model to perform effective in model evolution process.

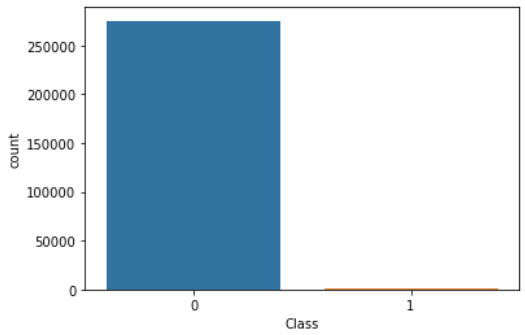


*Figure 5 Count plot of fraud and non-fraud*

The number of fraud transactions(class=1) in the dataset is much lower than the actual number of transactions(class=0), as shown in Fig. 5, the dataset after analysis and selection are highly unbalanced. Therefore, when training the model using the data set, it will achieve high accuracy, but the system will classify a transaction as safe even if it is fraudulent due to favorable towards classes with more data points. We must balance the data set in order to prevent such false negatives. By using the oversampling approach for balancing classes [16].

By using Standard Scalar function normalizes the features which are not in normalization, it eliminates the mean and scales to unit variance. Oversampling is the process of increasing the minor classes in the dataset in order to reduce the problem of imbalanced data.

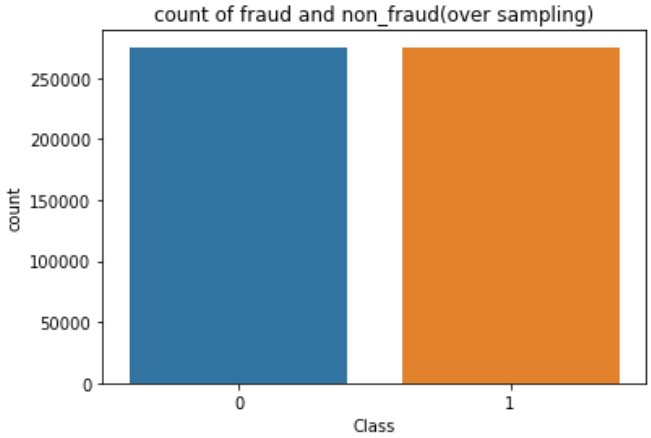


*Figure 6 Count plot of fraud and non-fraud after oversampling*

The data was balanced in the dataset by using the SMOTE () function in the oversampling technology as shown in Fig. 6. By this oversampling, the data is balanced and the performance of model evaluation increases. So, the probability of finding the fraud is equally likely. If the data is still imbalanced, then the pre-processing technique is used to make the data balance for the purpose of accurate results.

After the completion of preparing balanced data. The data was split into the training and testing samples in 8:2 ratio where 80% data comes under training phase and 20% under the testing phase.

3.3. Models

We explained three machine learning classification models that are Logistic Regression, Decision Tree Classifier, Random Forest.

1. Logistic Regression
   Logistic Regression is a statistical model that predicts the output of a categorical dependent variable. It is a supervised Machine learning algorithm [14]. The logistic model's parameters are estimated from the input data using a logistic regression curve. Through the output that is a continuous curve as the final result, it forecasts discrete categories [17]. Firstly, doing the pre-processing then fitting the logistic regression algorithm to the model for predicting the result. Next step is the predicting the test results and then visualization.

2. Decision Tree Classifier
   Using a tree-like depiction of alternatives and their potential outcomes, such as utility, resource costs, and chance event outcomes, a decision tree is a hierarchical decision-making model. This is a single method to demonstrate an algorithm that only utilizes statements of conditional control. Each leaf node of a decision tree represents a class's name, every inner node represents an examination on a property, and every link reflects the outcome of the test. The paths from roots to the leaf represent categorization rules [18]. Firstly,

choosing the main feature as a root node for the dataset and then divide that dataset into some subsets as branches. Next step is doing recursively the above process to get the result.

3. Random Forest
Random Forest is a supervised learning which is more suitable for classification problems. It contains different decision trees for classification, this works based on the majority voting of decision trees. The data set which contains continuous values are easily handled by the random forest. This is also used for the regression problems to predict numerical values [19]. Firstly, choosing some random points from the dataset and then building the decision trees with each of them. Next step is predicting results for the new data points and then visualization.

## IV. Results And Discussion

To analyze the credit card fraud detection using machine learning algorithms, we have compared the different classification models where accuracy, precision, recall and F1-score are the four standard performance measures used to evaluate model. The results are obtained as shown in Table 1.

*Table 1 Comparison of results with Different models*

| S.no | Model | Accuracy | Precision | Recall | F1_score |
|---|---|---|---|---|---|
| 1 | Logistic Regression | 94.47% | 97.35% | 91.42% | 94.47% |
| 2 | Decision Tree Classifier | 99.82% | 99.73% | 99.91% | 99.82% |
| 3 | Random Forest | 99.99% | 99.98% | 99.98% | 99.99% |

The result obtained by this can be displayed through the graphs. It is observed that the used model has given a very good accuracy for the desired dataset. By considering all the standard performance measures of classification models, Random Forest classifier is the best model for credit card fraud detection to predict the frauds. The bar graph indicates the accuracy level of all the classification models as shown in Fig. 7.

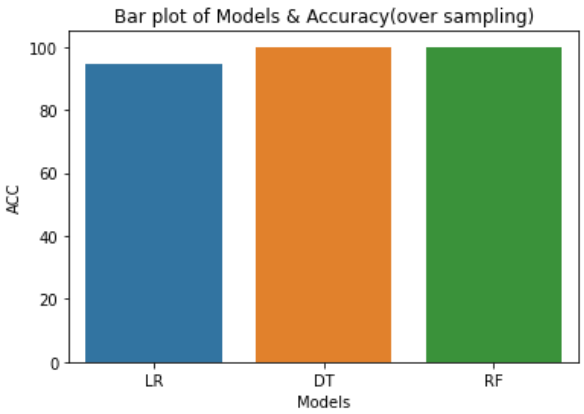


*Figure 7 Comparing the results using bar graph*

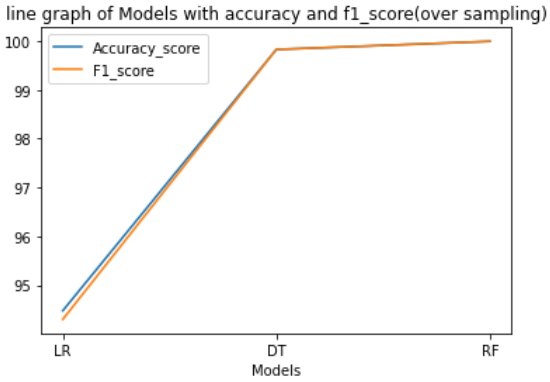


*Figure 8 Graphical analysis of accuracy and f1_score*

The Fig. 8 shows the line graph of accuracy and f1_score of all the classification models.

## V. Conclusion

In this study, we examined a range of machine learning classification methods, including Logistic Regression, Random Forest, Decision Tree Classifier, which express accuracy in identifying fraudulent transactions and reducing the number of false alarms. The likelihood of fraudulent transactions can be predicted shortly after credit card transactions if these algorithms are integrated into bank credit card fraud detection systems. Additionally, a number of anti-fraud methods can be implemented to lower risks and protect banks from significant losses. Recall, accuracy, f1-score, and precision are used to assess the performance of the classification models. We found that the random forest classifier methodology is superior to the other classification methods with 99.99% of accuracy after comparing all of them. The accuracies of other models are 94.47% and 99.82%.

## VI. References


[1] Sharma, P., Banerjee, S., Tiwari, D., & Patni, J. C. (2021). Machine learning model for credit card fraud detection-a comparative analysis. Int. Arab J. Inf. Technol., 18(6), 789-796.

[2] Sudha, C., & Akila, D. (2021, January). Credit card fraud detection system based on operational & transaction features using svm and random forest classifiers. In 2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM) (pp. 133-138). IEEE.

[3] Asha, R. B., & KR, S. K. (2021). Credit card fraud detection using artificial neural network. Global Transitions Proceedings, 2(1), 35-41.

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

[5] Saheed, Y. K., Hambali, M. A., Arowolo, M. O., & Olasupo, Y. A. (2020, November). Application of GA feature selection on Naive Bayes, random forest and SVM for credit card fraud detection. In 2020 international conference on decision aid sciences and application (DASA) (pp. 1091-1097). IEEE.

[6] Jain, Y., Tiwari, N., Dubey, S., & Jain, S. (2019). A comparative analysis of various credit card fraud detection techniques. Int J Recent Technol Eng, 7(5S2), 402-407.

[7] Rtayli, N., & Enneya, N. (2020). Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization. Journal of Information Security and Applications, 55, 102596.

[8] Popat, R. R., & Chaudhary, J. (2018, May). A survey on credit card fraud detection using machine learning. In 2018 2nd

international conference on trends in electronics and informatics (ICOEI) (pp. 1120-1125). IEEE.

[9] Suryanarayana, S. V., Balaji, G. N., & Rao, G. V. (2018). Machine learning approaches for credit card fraud detection. Int. J. Eng. Technol, 7(2), 917-920.

[10] Bhanusri, A., Valli, K. R. S., Jyothi, P., Sai, G. V., & Rohith, R. (2020). Credit card fraud detection using Machine learning algorithms. Journal of Research in Humanities and Social Science, 8(2), 04-11.

[11] Dornadula, V. N., & Geetha, S. (2019). Credit card fraud detection using machine learning algorithms. Procedia computer science, 165, 631-641.

[12] AbouSora, H., Ghoniemy, S., Banwan, S. A., Zanaty, E. A., & Afifi, A. (2013). Improved Fuzzy Possiblistic C-means (IFPCM) algorithms for magnetic resonance images segmentation. Journal of Global Research in Computer Science, 4(1).

[13] F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan and M. Ahmed, "Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms," in *IEEE Access*, vol. 10, pp. 39700-39715, 2022, doi: 10.1109/ACCESS.2022.3166891.

[14] Mbakwe, A. B., & Adewale, S. A. MACHINE LEARNING ALGORITHMS FOR CREDIT CARD FRAUD DETECTION.

[15] Pumsirirat, A. and Yan, L. (2018). Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine. International Journal of Advanced Computer Science and Applications, 9(1).

[16] A. S. Wheeler R, " Multiple algorithms for fraud detection. Knowledge-Based Systems," no. S0950-7051(00)00050-2, 2000.

[17] E. D. Yusuf Sahin, "detecting credit card fraud by ann and logistic regression," 2011.

[18] E. D. Y. Sahin, "Detecting credit card fraud by decision trees," in Proceedings of the international multiconference of engineers and computer science, Hong Kong, 2011.

[19] S. k. A. K. M. Ayushi agarwal, "Credit card fraud detection: A case study," in IEEE, New Delhi, India, 2015.