IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 7, July 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

||Volume 10, Issue 7, July2021||

| DOI:10.15680/IJIRSET.2021.1007134 |

Credit Card Fraud Detection Using Supervised Machine Learning Model

Harshitha P[1], Sahana D Gowda[2]

P.G. Student, Department of Computer Science and Engineering, BNM Institute of Technology, Banashankari II Stage, Karnataka, Bengaluru, India[1]

Professor and Head of the Department, Department of Computer Science and Engineering, BNM Institute of Technology, Banashankari II Stage, Karnataka, Bengaluru, India[2]

ABSTRACT: Today use of credit card is common among people in urban cities. This facility eases the problem of carrying large amounts of currency in hand. In the physical card purchase, the cardholder presents his card physically to a merchant for making a payment. If the cardholder does not realize the loss of card it can lead to financial loss to credit card company. In case of online payment mode attackers need only little information for doing fraudulent transaction(secure code ,card number, expiration date etc.…).Another problem is that there are vast number of similar transaction happening at the same time and it is difficult to monitor each and every transaction individually and identify the fraud .So that the fraud detection model should be able to distinguish between genuine transaction and fraud transaction .These kinds of frauds result in millions being lost every year .Thus existing approaches of supervised and unsupervised ML techniques can be applied to the data.

KEYWORDS: Credit card,machine learning,Random Forest(RF),Support Vector Machine (SVM).

I. INTRODUCTION

A fraud is defined as a deliberate deception carried out for monetary gain. This is an immoral and unethical practise that is really growing rapidly day by day. Digital payment methods such as card transactions have become increasingly popular, resulting in an increase in card fraud. Credit cards is used to make purchases both digitally and in store. The card may not need to be physically given in the area of digital payment.In this kind of situations, hackers or cyber thieves may be able to obtain card data. Huge amounts of money is lost every year as a consequence of these types of frauds. Many algorithms have been created and are being developed to overcome this barrier. Various detecting approaches are being developed in order to solve this problem as quickly as possible.

Card transactions are very frequent these days, however they come with their own set of issues. When it comes to fraud detection, there are also plenty of issues to deal with. A transaction's evaluation occurs in a fraction of a second, ranging from microseconds to milliseconds. As a consequence, the technique for detecting a fraudulent transaction must be extremely fast and precise.

Additional problem seems to be that a large number of similar transactions are taking place on a regular basis. To distinguish between a genuine and a fraudulent transaction, an effective Fraud Detection System must be implemented. This type of technology is based on the idea of understanding customer card consumption patterns. As an outcome, the data can be analysed using existing supervised and unsupervised learning methods.

Today, data is readily available all over the world; small and large businesses alike store information with a great volume, diversity, speed, and value. This information is acquired from a variety of sources, including user purchasing behavior and followers on social media, likes, and comments. The hidden data pattern was analysed and visualised using all of this information. Big data analysis used to be primarily focused on data quantities, such as general public databases, biometrics, and financial analysis.Credit cards have become a simple and attractive target for criminals that's because a large number of money may be obtained fast and without risk.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

||Volume 10, Issue 7, July2021||

| DOI:10.15680/IJIRSET.2021.1007134 |

As fraudsters attempt to make each fraud appear real, detecting fraud is tough. According to an increase in credit card payments, over 70% of consumers in the United States are vulnerable to these criminals. Because it includes many legitimate transactions than fraudulent ones, the credit card dataset is very unbalanced. Which indicates that without identifying a fraudulent transaction, the prediction will have a really high accuracy score.

Credit card Fraud

- Inner Card Fraud: This is accomplished by assuming a phoney identity.
- External Card Fraud: Whenever a credit card is stolen, it is utilised for fraudulent purposes.

Obtaining important information from a card, such as the card number, CVV code, kind of card, and other parameters, is one of the issues. A fraudster may steal a big chunk of money or conduct a significant transaction using stolen credit card information before the cardholder realises it. Because it accounts for the majority of credit card frauds, external card fraud has attracted a lot of attention. Detecting fraudulent transactions with outdated manual approaches is time-consuming and wasteful, and manual approaches are becoming increasingly impractical as big data becomes more common.

Machine learning is the answer for this century's problem of replacing these methodologies and working with massive datasets that are tough for humans to handle.Fraud can be detected in a number of different ways, and the approach selected is completely up to the individual. When to use it based on the dataset Prior classification of anomalies is required for supervised learning. In recent times, various supervised algorithms are being implemented to identify fraud.The imbalance dataset used in this experiment was chosen from the Kaggle database. Imbalanced denotes that the distribution of classes is unbalanced. The idea of this project is to develop a machine learning-based fraud detection model. We can increase the model's sensitivity and accuracy by using Random Forest (RF) and Support Vector Machine (SVM) approaches.

The dataset includes purchases made by a cardholder over the course of two days. here are 284,807 transactions in total, with 492 of them being fraudulent transactions. This is a really uneven dataset.There are 31 numerical features in the dataset. The PCA transformation of some of the input variables was employed to ensure that the data remained anonymous because they contained financial information. The feature "Time" displays the time duration of the first and subsequent transactions in the dataset. The amount of credit card transactions is displayed in the "Amount" feature. he classification is represented by the feature "Class," which only has two values: 1 in case of fraud ,and 0 in case of not fraud.

II. RELATED WORK

L.Zheng[1] et al., The Behaviour Certificate, which reflects cardholder transaction behaviours, is the basis for this document. The association between features and some unique instances, such as festivals, weekends, and so on, is called a Behaviour Certificate. The author will collect a set of behaviour characteristics from each cardholder's transaction and use these to construct his or her Behaviour Certificate. Finally, he uses his Behaviour Certificate to calculate the risk level of each cardholder's incoming transaction. It is deemed fraud if the degree is higher than the criterion.

According to a survey of more than 160 businesses, the internet scams is 12 times larger than the number of offline scams each year. Fraudsters are used to copy and replicate card information, or sensitive information is obtained through phishing (cloned websites) or dishonest credit card company workers. The fraud can be identified by looking at each account's spending patterns and looking for any deviations from the "normal" transaction patterns. In this work, the author provides a revolutionary credit card fraud detection method based on the cardholder's transaction records. He includes four attributes in a transaction, and thirteen related features are used as behaviour features, instead of single unique attribute such as the amount or time . He also took into account the impact of unusual circumstances, such as festivals and weekends, on cardholders' transaction habits. Furthermore, investigations show that this method's accuracy is greater than 90% across a variety of input datasets.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

||Volume 10, Issue 7, July2021||

| DOI:10.15680/IJIRSET.2021.1007134 |

S. Mittal [2] et al., The author of this work assesses a number of machine learning algorithms in terms of their capacity to detect fraudulent activity in a real-world unbalanced dataset. The sample process, variable selection, and identification algorithms used in the dataset all have a significant result on the accuracy of credit card fraud detection. There are a few obstacles that these algorithms must overcome in order to produce accurate categorization results. This study examines the performance of neural networks, deep learning, nave bayes, k-nearest neighbour, and logistic regression on strongly imbalanced credit card fraud data. On the extremely unbalanced data, a hybrid method of under-sampling and oversampling is used. The accuracies obtained using neural network, deep learning, naive bayes, k-nearest neighbour, and logistic regression classifiers were 58.1 %, 98.0 %, 97.92 %, and 54.86 %, respectively.

A. Thennakoon [3] et al., This article focuses on identifying credit card fraud in real time. To accomplish so, the author employs predictive analytics based on machine learning models and an API module to find out whether genuinity of the transaction.The information in this study is broken down into two categories: category data and numerical data. The data was initially categorical in nature. After converting categorical data to numerical data, valid and reliable processes can be applied. Machine learning techniques are then used to identify the best strategy using categorical data. The fundamental purpose of this research is to compare machine learning methodologies and develop a useful performance score for detecting fraudulent behaviour in order to identify the best algorithms for each of the four forms of fraud.

There are two types of datasets:

1.Fraud transaction log file

2.File containing all transaction logs

The fraud transaction all transaction logs contains all cases of digital card fraud, whereas the all files containing all transaction log file contains all transactions maintained by the bank over a certain period of time. The structure of the data was substantially skewed while examining the combined datasets due to an uneven number of genuine transactions and fraudulent occurrences. The API module connects the Fraud detection model, the GUI, and the data warehouse in real time. The machine learning models' live transactions, expected findings, and other critical data were stored in a Data Warehouse. Graphical user interfaces that display real-time transactions, fraud alerts, and historical fraud data so that the user can establish a connection with the fraud detection system. The API module receives a signal when the fraud detection model detects a fraudulent transaction. After that, the API module will send a notification and feedback to the end user. As a result, the developed machine learning models have high accuracy.

D. Varmedja [4] et al., This article is based on identifying credit card fraud including Logistic Regression (LR), Random Forest (RF), Nave Bayes (NB), and Multilayer Perceptron (MLP). The relationship between continuous, binary, and categorical predictors is represented by the logistic regression model. It's feasible to have binary dependent variables. Based on some forecasts, we can anticipate if something will happen or not.

The Naive Bayes method ignores the interdependence of attributes. The Bayes theorem is employed in this situation. Random forest is a classification and regression technique that can be used in conjunction. There are many decision trees in it. This method delivers better results and keeps the model from overfitting when there are more trees in the forest. A feedforward artificial neural network which has three layers of nodes is known as a multilayer perceptron. Each node makes use of the activation function. The weighted sum of the activation function's inputs is given a bias.This paper's main purpose was to examine several machine learning techniques for detecting fraudulent transactions. The Random Forest method, as a result of the comparison, produced the best accuracy, i.e it will identify whether transactions are fraudulent or genuine.

Pawan Kumar [5] et al.,Online shopping has become a vital part of everyone's lives as a result of the internet's exponential expansion. MasterCard is the most used method of payment for internet purchases. n order to limit their losses, all MasterCard supply banks were compelled to instal fraud detection systems. Neural networks (NN), rule-based approaches, fuzzy systems, call trees, Support Vector Machines (SVM), Logistic Regression, Local Outlier Factor (LOF), Isolation Forest, K-Nearest Neighbour, and Genetic algorithms are the most often used fraud detection technologies. Outlier detection is a data analysis technique that is frequently employed in the identification of fraud. Outliers are data points that contradict the dataset's memory or depart dramatically from alternative observations,

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

||Volume 10, Issue 7, July2021||

| DOI:10.15680/IJIRSET.2021.1007134 |

implying that they were generated by a different process.Techniques such as neural networks, local outlier factors, and isolation forests, among others, are frequently used to discover outliers. Anomaly detection is the process of identifying unusual things, circumstances, or findings that stand out from the rest of the data (also known as outlier detection).

The Isolation forest,Local outlier forest and SVM discovered mistakes as 73,94 and 8516 respectively. LOF (99.65%) and SVM (99.74%) are both more precise than Isolation Forest (70.09 %).When we compare the error precision and recall of the three variations, we can see that the Isolation Forest outperforms the LOF, with a fraud detection rate of around 27% vs. 2% for the LOF and an SVM of 0%. As a result, the Isolation Forest methodology did a far better job of detecting the 30 fraud incidents.

III. METHODOLOGY

The methodology focuses on machine learning processes for detecting, classifying, and predicting Credit Card Fraud, hence solving the problem. I will discover the most essential characteristics lead to increased accuracy in fraudulent transaction detection using Random Forest (RF) and Support Vector Machine (SVM).

1. Random Forest

Random forests is a learning process that is supervised. It is very user-friendly and customizable algorithm. A forest is formed by trees. The more trees in a forest, the stronger it is considered.

Random Forest follows a four-step process:

1. Take randomly selected samples from a dataset.
2. Build a decision tree for each data and use it to create a result.
3. For each expected outcome, cast a vote.
4. Choose the answer with the most votes as the final classification.

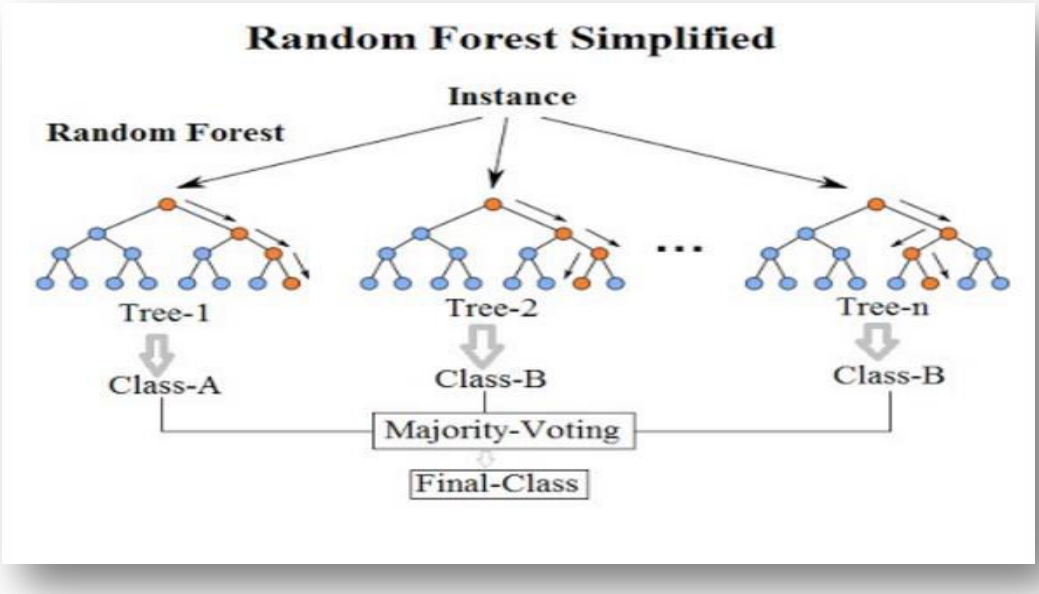

Figure 1:Random Forest Simplified

Random forests is considered a highly accurate and robust strategy due to the vast number of decision trees involved in the process. The problem of overfitting has no influence on it. Because it averages all of the forecasts, all biases are cancelled out. This method is used to solve both classification and regression issues.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

||Volume 10, Issue 7, July2021||

| DOI:10.15680/IJIRSET.2021.1007134 |

2. Support Vector Machine

Support vector Machines, or SVMs, are used to analyse data for classification and regression using various learning methods. A separating hyperplane formally defines it. In other words, the algorithm produces an ideal hyperplane that categorises fresh samples given labelled training data (supervised learning). The scenario is analysed by generating a feature vector, which is a finite-dimensional vector space in which each dimension represents a "feature" of a specific object. The frequency or importance of a certain keyword is each "feature". Common kernels in SVM libraries includes Polynomial, Radial Basis Function (rbf), and Sigmoid. SVC, which is utilised in Machine Learning, is the classification function of SVM. The SVC function appears to be as follows:

sklearn.svm.SVC (C=1.0, kernel= 'rbf', degree=3)

Hyperplane:

A hyperplane is a line that divides and categorises a set of data in a linear manner.

Support Vector:

Support vectors are the points that are closer to the hyperplane if eliminated, would modify the position of the dividing hyperplane.

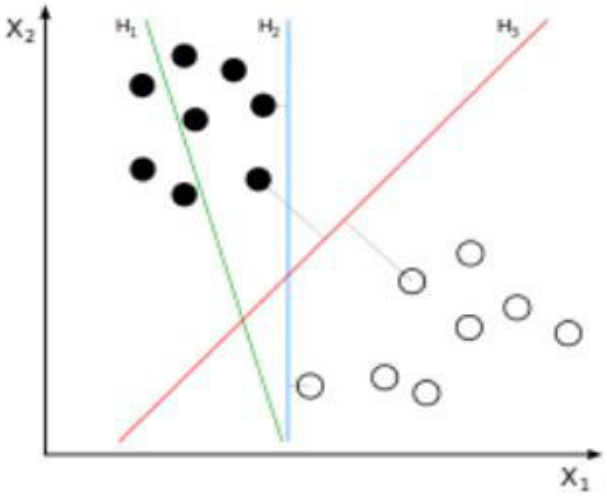

Figure 2: Support Vector Machine

The classes are not divided in H1. H2 has a slight advantage, but only by a slight bit. H3 puts the greatest distance between them.

Margin:

The space between the hyper - plane and the nearest target value from either collection called the margin.

3. Dealing with Imbalanced data

One of the most typical techniques of dealing with an imbalanced dataset is to resample the data. Imbalanced classification is when there is no equal distribution of classes in the training dataset. The degree of class imbalance varies, but a severe imbalance is more difficult to model and may necessitate specialist methodologies. Undersampling

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

||Volume 10, Issue 7, July2021||

| DOI:10.15680/IJIRSET.2021.1007134 |

and oversampling are the two most common ways for this. Oversampling techniques are preferred over undersampling techniques in most circumstances.

- SMOTE (Synthetic Minority Over-sampling Technique) is a form of over-sampling approach for correcting group imbalances.
- This methodology duplicates current minority data instances and makes modest changes to them, resulting in new minority data instance.
- In the SMOTE class, we can utilise the SMOTE implementation from the imbalanced-learn Python library.

The SMOTE class is similar to a scikit-learn data transform object it will not duplicate the data rather it will generate new record with slight changes.

SMOTE algorithm works in 4 simple steps:

1. Select an input vector for the minority class.
2. Find its k closest neighbours (theSMOTHE() function takes a k neighbours argument).
3. Choose one of the neighbours and draw a synthetic point on the line between the point in question and the chosen neighbour.
4. Repeat the process until the data is evenly distributed.

IV. EXPERIMENTAL RESULTS

The result of the implementation is examined using a large number of test scenarios. The best scenarios in which the model performs well are examined. The worst-case scenarios for the created model were also examined. The model is created by combining the two sets of data.

1.Random Forest:

Overall accuracy of Random Forest model using test-set is:99.9917

	precision	recall	f1-score	support
0	1.00	1.00	1.00	85149
1	1.00	1.00	1.00	85440
accuracy			1.00	170589
macro avg	1.00	1.00	1.00	170589
weighted avg	1.00	1.00	1.00	170589

Figure 3:Classification report of Random Forest

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

||Volume 10, Issue 7, July2021||

| DOI:10.15680/IJIRSET.2021.1007134 |

2.Support Vector Machine

Overall accuracy of Support Vector Machine model using test-set is:99.8267

	precision	recall	f1-score	support
0	0.99	1.00	0.99	17064
1	1.00	0.99	0.99	17054
accuracy			0.99	34118
macro avg	0.99	0.99	0.99	34118
weighted avg	0.99	0.99	0.99	34118

Figure 4:Classification report of Support Vector Machine

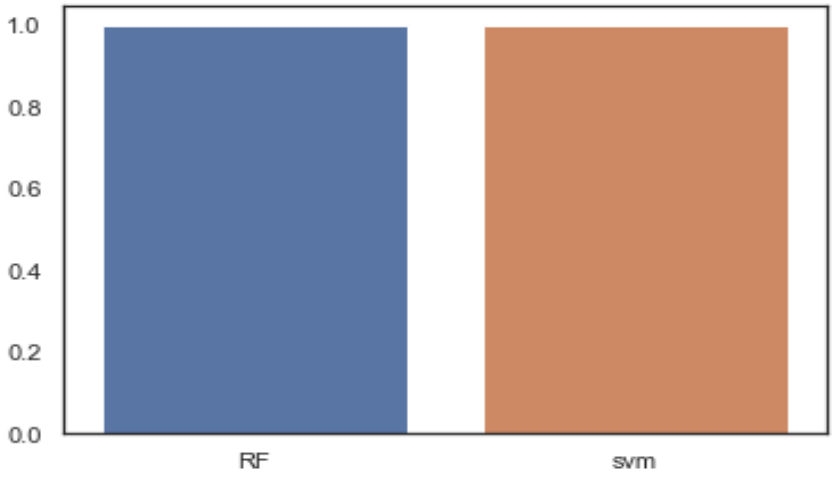

Figure 5: Comparison of Random Forest and Support Vector Machine

V. CONCLUSION

Credit card fraud is a big concern in the corporate world. Scams like these can cost you a lot of money and cause you to lose your identity. As a result, corporations are pouring money into developing innovative fraud detection and prevention concepts and procedures. To balance the unbalanced dataset, the SMOTHE approach was employed.The study's main goal was to compare numerous fraud detection machine learning systems. The Random Forest algorithm yields the most accurate results whether the transaction is genuine or fraudulent as a result of the comparison. Various factors such as accuracy, precision and recall are used to arrive at this conclusion. It is critical to have a high recall value for this type of situation. The importance of feature selection and dataset balancing in generating substantial outcomes has been demonstrate. The resampling approaches can be used by future researchers in this subject on the datasets they're working with. This methodology aids in the reduction of dataset imbalance ratios, resulting in improved classification results.

REFERENCES

[1] L. Zheng et al., ”A new credit card fraud detecting method based on behavior certificate,” 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), Zhuhai, 2018, pp. 1-6.
[2] S. Mittal and S. Tyagi, "Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection", 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2019.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

||Volume 10, Issue 7, July2021||

| DOI:10.15680/IJIRSET.2021.1007134 |

[3] "Real-time Credit Card Fraud Detection Using Machine Learning"AnuruddhaThennakoon, Chee Bhagyani , Sasitha Premadasa, ShalithaMihiranga, NuwanKuruwitaarachchi 2019.
[4]]Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., &Anderla, A. (2019). Credit Card Fraud Detection - Machine Learning methods. 2019 18th InternationalSymposiumINFOTEH-JAHORINA(INFOTEH). doi:10.1109/infoteh.2019.8717766
[5] Kumar, P., & Iqbal, F. (2019). Credit Card Fraud Identification Using Machine Learning Approaches. 2019 1st International Conference on Innovations in InformationandCommunicationTechnology(ICIICT). doi:10.1109/iciict1.2019.8741490
[6] J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, ''Credit card fraud detection using machine learning techniques: A comparative analysis," 2017 International Conference on Computing Networking and Informatics (ICCNI), Lagos, 2017.
[7] Kho, J. R. D., & Vea, L. A. (2017). Credit card fraud detection based on transaction behavior. TENCON 2017 - 2017 IEEE Region 10 Conference. doi:10.1109/tencon.2017.8228165
[8] Jiang, C., Song, J., Liu, G., Zheng, L., & Luan, W. (2018). Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism.IEEE Internet of Things Journal, 1–1. doi:10.1109/jiot.2018.2816007
[9] Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., & Jiang, C. (2018). Random forest for credit card fraud detection. 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC). doi:10.1109/icnsc.2018.8361343
[10] Samidha Khatri, Aishwarya Arora and Arun Prakash Agrawal," Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison" 2020.
[11] SurajPatil*, VarshaNemade, PiyushKumarSoni, Predictive Modelling for Credit Card Fraud Detection Using Data Analytics, International Conference on Computational Intelligence and Data Science (ICCIDS 2018)
[12] A. Roy, J. Sun, R. Mahoney, L. Alonzi, S. Adams and P. Beling, "Deep /learning detecting fraud in credit card transactions," 2018 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, 2018, pp. 129-134.
[13] Alex G.C.de S, Adriano C.M.Pereira, Gisele L.Pappa, A customized classification algorithm for credit card fraud detection, Engineering Applications of Artificial IntelligenceVolume 72, June 2018, Pages 21- 29.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details