ISSN: 2249-7196
IJMRR/April. 2025/ Volume 15/Issue 2s/410-418
Kapaka Hareesh / International Journal of Management Research & Review

Detection of Fraudulent Transactions from Highly Imbalanced Dataset using Different ML Classifiers

Kapaka Hareesh
PG scholar, Department of MCA, CDNR collage, Bhimavaram, Andhra Pradesh.
A.DURGA DEVI
(Assistant Professor), Master of Computer Applications, DNR collage, Bhimavaram, Andhra Pradesh.

Abstract *Nowadays, there is faster rate of increase in the use of card payment and online payments. There is most common issue with such card or online transaction is fraud possibility and there is need to understand cybersecurity concern. Globally it is observed that 43-billion-dollar loss got in 5 years. So, financial organizations such as banks need to identify such fraud transaction details and performance of machine learning algorithms. In this application there is performance analysis of different machine learning classifiers including unsupervised and supervised machine learning classifiers. The dataset used has highly imbalanced data as supervised transactions are very less compared to normal/no-fraud transactions. Unsupervised machine learning algorithms performs superior as compare to supervised machine learning classifiers for credit card fraud detection from highly imbalanced dataset. So, this machine learning based fraud transaction helps in preventing/stopping abnormal transactions and allowing normal transactions. Performance can be measured using different parameters such as balanced accuracy, precision, recall, negative prediction rate, etc. It is observed with result analysis that unsupervised machine learning classifiers such as K-means algorithm, isolation forest algorithm, local outlier algorithm performs much superior than supervised machine learning algorithms such as Support vector machine algorithm, random forest algorithm, naïve bayes algorithms, etc.*

Keywords: *Fraud Detection, cybersecurity, credit card, machine learning, supervised and unsupervised machine learning.*

Introduction

1.1 Credit Card Transactions

A credit card is a payment card that allows the user to pay for products and services from merchants. A credit card is a financial instrument issued by a bank whose credit limit is preset. Credit cards are used to make a transaction cashless [2]. A credit card shows a payment mechanism that facilities both the consumer and commercial business. It is an option for cash or checks payments.

At the time of the transaction, a credit card puts a condition on it that pays the return of the borrowed money, which is taken by cardholders and interest on that as well as the agreed charges on it. It provides the facilities like online payment. Using a credit card, you can do shopping directly without needing to carry the money around. We have to get different discounts and offers on the credit card. If the card is not used correctly, it becomes a debt trap.

A **credit card** is a card which allows people to buy items without cash. Each card has a unique number. Using this number, plus other details on the card (such as the validity date, or a code), the client can buy goods or services. The issuer of the card automatically transfers the money to the seller. The person using the card gets a credit. The customer has a certain amount of time to pay the credit card bill. If the bill is left unpaid for some time, the customer will have to pay interest for the amount that is left unpaid. Payment using a credit card is one of the most common methods of electronic payment. Credit cards are usually small plastic cards with a unique number attached to an account. Most are magnetic stripe cards and many have an EMV chip for use by card readers.

A credit card is a payment card issued to users (cardholders) to enable the cardholder to pay a merchant for goods and services based on the cardholder's accrued debt (i.e., promise to the card issuer to pay them for the amounts plus the other agreed charges).[1] The card issuer (usually a bank or credit union) creates a revolving account and grants a line of credit to the cardholder, from which the cardholder can borrow money for payment to a merchant or as a cash advance. There are two credit card groups: consumer credit cards and business credit cards.

Most cards are plastic, but some are metal cards (stainless steel, gold, palladium, titanium), and a few gemstone-encrusted metal cards.

A regular credit card is different from a charge card, which requires the balance to be repaid in full each month or at the end of each statement cycle. In contrast, credit cards allow the consumers to build a continuing balance of debt, subject to interest being charged. A credit card differs from a charge card also in that a credit card typically involves a third-party entity that pays the seller and is reimbursed by the buyer, whereas a charge card simply defers payment by the buyer until a later date. A credit card also differs from a debit card, which can be used like currency by the owner of the card. In 2018, there were 1.12 billion credit cards in circulation in the U.S., and 72% of adults had at least one card

Necessary prevention measures can be taken to stop this abuse and the behaviour of such fraudulent practices can be studied to minimize it and protect against similar occurrences in the future. In other words, Credit Card Fraud can be defined as a case where a person uses someone else's credit card for personal reasons while the owner and the card issuing authorities are unaware of the fact that the card is being used. Fraud detection involves monitoring the activities of populations of users in order to estimate, perceive or avoid objectionable behaviour, which consist of fraud, intrusion, and defaulting.

Fraud has been increasing drastically with the progression of state-of-art technology and worldwide communication. Fraud can be avoided in two main ways: prevention and detection. Prevention avoids any attacks from fraudsters by acting as a layer of protection. Detection happens once the prevention has already failed. Therefore, detection helps in identifying and alerting as soon as a fraudulent transaction is being triggered.

Recently, cardnot-present transactions in credit card operations have become popular among web payment gateways. According to the Nilson Report in October 2016, more than $31 trillion were generated worldwide by online payment systems in 2015, increasing 7.3% than 2014. Worldwide losses from credit card fraud have been rising to $21 billion in 2015, and will possibly reach $31 billion by 2020. However, there has been an extreme increase in fraudulent transactions that affect the economy dramatically.

Credit card fraud can be classified into several categories. The two types of frauds that can be mainly identified in a set of transactions are Card-not-present (CNP) frauds and Card-present (CP) frauds. Those two types can be described further by bankruptcy fraud, theft/counterfeit fraud, application fraud, and behavioural fraud. Our study aims at addressing four fraud natures that belong to the CNP fraud category described above and we propose a method to detect those frauds real time. Machine learning is this generation's solution which replaces such methodologies and can work on large datasets which is not easily possible for human beings.

Literature survey

The aim of the data analytics is to uncover hidden patterns and use them to make better decisions in various situations. For fraudulent with the advancement of modernised technology the credit card fraud has increased significantly and has a become an easy target. In the financial industry Credit card fraud detection is major issue also the coting billions of dollars each year. Due to a lack of real worlds transection datasets the design of a fraud detection algorithm is very difficult task. Due to continentality and highly imbalanced publicly available datasets.

In this paper for the detection of the credit card fraud we use a real- world dataset to apply different supervised machine learning algorithms. Also, using ensemble learning technique to build a super classifier we use these algorithms. The detection of the credit card fraud transections we identify the most important variables that may lead to more accuracy. Furthermore, in this aper we discuss and compare the performance of the different supervised machine learning algorithms that have been published in the literature versus the super classifier that we implemented [1].

The widespread adoption of the EMV (Europay-MasterCard-VISA) chip card is design in the industry of credit card in largely resolved that issue posed by the old Magnetic Stripe card

technology. In the various papers discussing a EMV design and implementation into question. To capture the potentially anomalous transections there is a detection model is available as a backup in case the technology is failed. There are various classifiers are determined at the time of model development but the Random Tree and J48 produced the more accuracy values of 94.32% and the 93.50% respectively. To understanding the transaction logs data after the examination the two classifiers reveals that the J48 is better suited [2].

Now the transactions with Credit card are very common, as its fraud associated with them. One of the most common methods of committing the fraud is obtaining card information illegally and using it to make an online purchase. To detect these fraud transections from the thousands of the normal transections it is very difficult for the credit card companies and merchants. To solve this problem the machine learning algorithm is used if enough data is collected and made available. In this project the popular supervised machine learning algorithm is used to detect the fraud of credit card in a highly imbalanced dataset[3].

The transections on Credit Card are very common now as are the frauds that accompany them. One of the most common methods of committing the fraud is obtaining card information illegally and using it to make an online purchase. To detect these fraud transections from the thousands of the normal transections it is very difficult for the credit card companies and merchants. To solve this problem the machine learning algorithm is used if enough data is collected and made available. To detect a credit card fraud which is highly unblanched so detection of frudness we use the popular supervised and unsupervised machine learning algorithms. Th unsupervised a machine learning algorithm was discovered to eb capable of handling skewness and produce a result of the classification [4].

Credit card transactions have become common
place today and so is the frauds associated with it. One of the
most common modus operandi to carry out fraud is to obtain the
card information illegally and use it to make online purchases.
For credit card companies and merchants, it is infeasible to
detect these fraudulent transactions among thousands of normal
transactions. If sufficient data is collected and made available,
machine learning algorithms can be applied to solve this
problem. In this work, popular supervised and unsupervised
machine learning algorithms have been applied to detect credit
card frauds in a highly imbalanced dataset. It was found that
unsupervised machine learning algorithms can handle the
skewness and give best classification results.

In the credit card transactions, there is one of the testbeds to detect a fraud is computational intelligence algorithm. In fact, this issue involves a number of pertinent challenges, including the concept drift class imbalance and verification latency. For the fraud detection the vast majority of the learning algorithms are proposed on the basis of the assumptions that are unlikely hold in Fraud detection system. This lack of realism is manifested in two ways such as,

I. The timing and manner with which supervised information is provided.
II. To evaluate fraud detection performance the metrices are used.

In this paper there are three contributions. 1st we propose with the assistance of our industrial partner, a formalization of the fraud detection problem that accurately describes the operating conditions of the FDS that analyze massive streams of the credit card transaction on the daily basis. For the fraud detection we demonstrate the most appropriate performance measures.2nd we develop and evaluate the novel leaning strategy that address concept drift, verification latency and class imbalance. 3rd we demonstrate the impact of the concept drift and class unbalance in a real-world stream containing more than 75millions of transactions authorized over a 3-year period of experiments[5].

Detecting frauds in credit card transactions is perhaps one of the best testbeds for computational intelligence algorithms. In fact, this problem involves a number of relevant challenges, namely: concept drift (customers habits evolve and fraudsters change their strategies over time), class imbalance (genuine transactions far outnumber frauds) and verification latency (only a small set of transactions are timely checked by investigators). However, the vast majority of learning algorithms that have been proposed for fraud detection, relies on assumptions that hardly hold in a real-world Fraud Detection System (FDS). This lack of realism concerns two main aspects: i) the way and timing with which supervised information is provided and ii) the measures used to assess fraud-detection performance.

This paper has three major contributions. First we propose, with the help of our industrial partner, a formalization of the fraud-detection problem which realistically describes the operating conditions of FDSs that everyday analyze massive streams of credit card transactions. We also illustrate the most appropriate performance measures to be used for fraud-detection purposes.

Second, we design and assess a novel learning strategy which effectively address class imbalance, concept drift and verification latency. Third, in our experiments we demonstrate the impact of class unbalance and concept drift in a real-world data stream containing more than 75 millions transactions, authorized over a time window of three years

As an online shopping becomes a more popular so the fraud transection is goes on increasing**.** As a result, the research on the fraud detection is both important and interesting. An important method of detecting the fraud transactions to extract the user's behavior profiles form their previous historical transaction records and then use their history to determine whether an incoming transection is fraudulent or not. To represent the user BPs (Behavior profile) the Markov chain models are commonly used. It is effective for the users whose transection behaviors are relatively stable.

Nowadays we see that the online shopping becomes a more popular and it becomes a more suitable for the users to consume via the internet, which shows the diversities of the transection behavior of the user. As a result, for the representing the behaviors the Markov chain models are unsuitable. In this paper we propose a logical graph on the based model for representing the logical relationship of record of transection attributes. Simultaneously we define an information entropy-based diversity coefficient to characteristics to users' diversity [5] of transection behaviors. We create a BP for each user and use it to determine whether the incoming transection is fraudulent or not. Our experiment on real data shows our method outer performs three state of the art one oneness [6].

Nowadays the peoples in cities are used a credit card. This service alleviates the burden

of carry out the large amount cash on one's person. In the physical card purchase the cardholder physically present his card to merchant to make a payment. If the cardholder does not realize the card has been lose the company of credit card suffers a financial loss. In case of the online payment, attackers require only a small amount of information to conduct the fraudulent transactions such as a card number, secure code, expiration date and so on.

Another issue is that there are a large number of same transections occurring at the same time it makes difficult to monitor each transection individually and identify the fraud. As result, the fraud detection model should be able to distinguish between fraudulent transection and genuine. Every year millions of dollars are lost as a result of this

type of fraud thus the existing unsupervised and the supervised ML approaches can be applied to the data [7].

Proposed Methodology

In proposed methodology, multiple **unsupervised machine learning classifiers** such as Isolation Forest, K-means clustering algorithm and local outlier factor are used for credit card fraud transaction classification from highly imbalanced datasets.

Following steps are used for designing this application is,

a) Uploading selected dataset
b) Preprocess selected dataset
c) Training Machine learning algorithms
d) Testing Machine learning algorithms
e) Performance evaluation

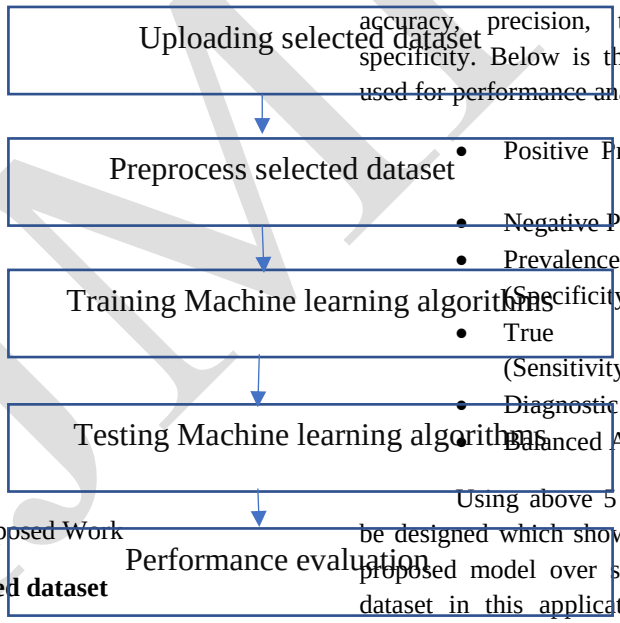


Fig. Block Diagram of proposed Work

a) Uploading selected dataset

The dataset selected for this application is highly imbalanced dataset with huge number of transactions for training and a smaller number of transactions for testing are available.

b) Preprocess selected dataset

Dataset has some empty values, in preprocessing that values get replaced and some more preprocessing is applied to get data in standard format.

Data splitting also applied which split our data in 80% training and 20 % testing.

c) Training Machine learning algorithms

Different machine learning classifiers are then trained with data splited and found accuracy of prediction which found superior for unsupervised machine learning classifiers than supervised machine learning classifiers.

d) Testing Machine learning algorithms

Machine learning classifiers first we will train and then we will use trained model for testing/prediction. As we used both unsupervised and supervised machine learning classifier for testing, we can find their prediction rate based on performance parameters.

e) Performance evaluation

There are multiple parameters such as balanced accuracy, precision, true negative rate, and specificity. Below is the performance parameters used for performance analysis,

- Positive Predictive Value (Precision)
- Negative Predictive Value
- Prevalence True Negative Rate (Specificity)
- True Positive Rate (Sensitivity/Recall)
- Diagnostic Odd Ratio (DOR)
- Balanced Accuracy

Using above 5 steps proposed method ca be designed which shows superior performance of proposed model over state of art technique. The dataset in this application is highly imbalanced dataset which contains normal transactions in lacs and fraud transactions in only hundreds. Proposed model even works with highly imbalanced dataset using machine learning techniques. Data cleaning and data splitting operations are performed as preprocessing operations for input dataset to make dataset in standard format. Machine learning classifiers are trained using training 80% of the samples and testing is performed on 20% of the dataset samples. Performance evaluation is applied

ISSN: 2249-7196
IJMRR/April. 2025/ Volume 15/Issue 2s/410-418
Kapaka Hareesh / International Journal of Management Research & Review

at the last using balanced accuracy, precision, recall, etc. Performance analysis parameters shows that unsupervised ML techniques works superior than supervised ML techniques.

RESUTLS

Overall Model (GUI) Designed using Python is as shown below,

Fig.4.1 Overall GUI of proposed Model

Here basic GUI structure is shown which uses 'Tkinter' prebuilt library from python for design. There is total six buttons we prepared, for every button in background we written coding part. Using GUI proposed model can be easily understandable. User can access this proposed model without complete knowledge of algorithm.

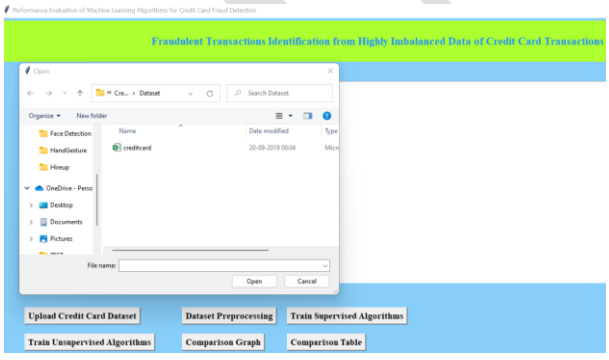


Fig. 4.2 Selection of Dataset

Selected dataset from user. The dataset is collected from github.com which is highly imbalanced dataset.

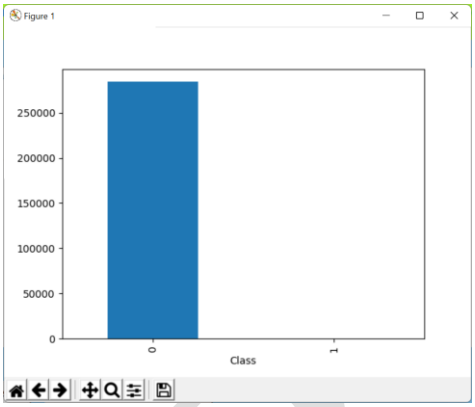


Fig.4.3 The Selected Dataset Shows Normal (0) And Fraud (1) Transactions

Fraud transactions are very less in comparison with normal transaction so graph is shown like above.

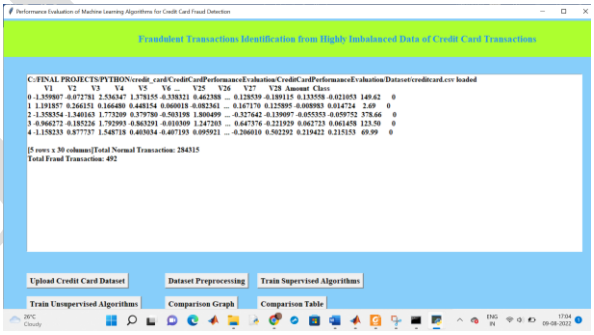


Fig.4.4 Uploaded dataset information

There are normal transaction 284315 which is very large number while fraud transaction is 492 which is very less in count. So this dataset is highly imbalanced.

Fig. 4.5 Dataset Preprocessing

Dataset preprocessing is applied here, and dataset is splited into training and testing parts.

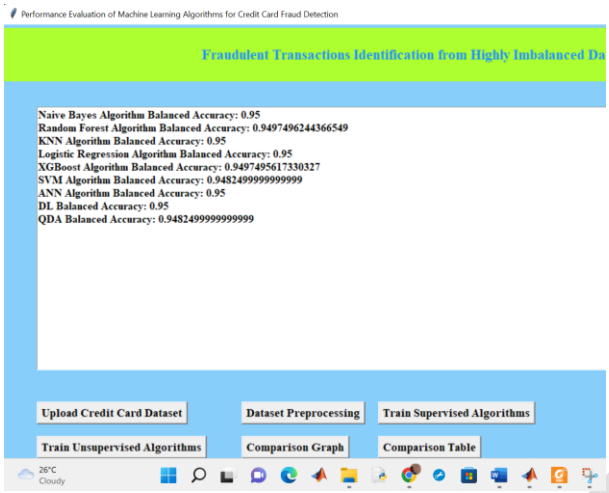


Fig.4.6 Training supervised ML classifier

Multiple machine learning classifiers such as supervised and unsupervised machine learning classifiers are used. Supervised machine learning classifier accuracies are displayed in above results screenshot.

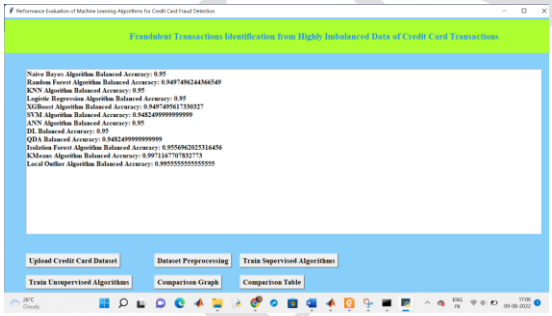


Fig. 4.7 Training unsupervised ML classifier

Here unsupervised ML classifier accuracy also displayed in same results.

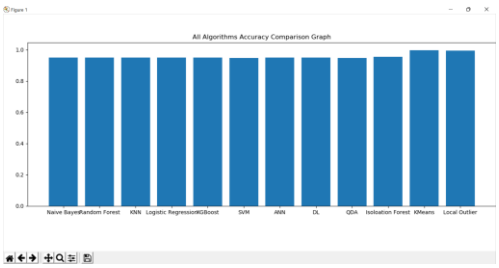
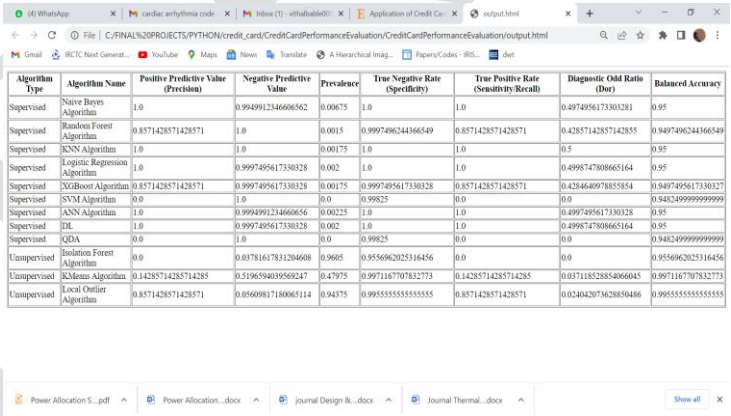


Fig. 4.8 Accuracy Bar graph for different ML classifiers

Bar graph is plotted for accuracy of all machine learning classifier.



Algorithm Type	Algorithm Name	Positive Predictive Value (Precision)	Negative Predictive Value	Prevalence	True Negative Rate (Specificity)	True Positive Rate (Sensitivity/Recall)	Diagnostic Odd Ratio (Dor)	Balanced Accuracy
Supervised	Naive Bayes Algorithm	1.0	0.9949912346606562	0.00675	1.0	1.0	0.4974956173303281	0.95
Supervised	Random Forest Algorithm	0.8571428571428571	1.0	0.0015	0.9997496244366549	0.8571428571428571	0.4285714285714285	0.9497496244366549
Supervised	KNN Algorithm	1.0	1.0	0.00175	1.0	1.0	0.5	0.95
Supervised	Logistic Regression Algorithm	1.0	0.9997495617330328	0.002	1.0	1.0	0.4998747808665164	0.95
Supervised	XGBoost Algorithm	0.8571428571428571	0.9997495617330328	0.00175	0.9997495617330328	0.8571428571428571	0.4284640978855854	0.9497495617330327
Supervised	SVM Algorithm	0.0	1.0	0.0	0.99825	0.0	0.0	0.9482499999999999
Supervised	ANN Algorithm	1.0	0.9994991234660656	0.00225	1.0	1.0	0.4997495617330328	0.95
Supervised	DL	1.0	0.9997495617330328	0.002	1.0	1.0	0.4998747808665164	0.95
Supervised	QDA	0.0	1.0	0.0	0.99825	0.0	0.0	0.9482499999999999
Unsupervised	Isolation Forest Algorithm	0.0	0.9378161783120608	0.9605	0.9556962025316456	0.0	0.0	0.9556962025316456
Unsupervised	KMeans Algorithm	0.14285714285714285	0.5196594039569247	0.47975	0.9971167707832773	0.14285714285714285	0.037118528854066045	0.9971167707832773
Unsupervised	Local Outlier Algorithm	0.8571428571428571	0.9560981718006514	0.94375	0.9955555555555555	0.8571428571428571	0.024042073628850486	0.9955555555555555

Fig. 4.9 Performance Analysis of ML classifiers using Different Metrics

Above we showed multiple performance metrics for different supervised and unsupervised classifiers.

CONCULSION

Credit card fraud detection from highly imbalanced data is implemented in this study. There are multiple supervised and unsupervised ML classifiers are used for performance analysis using different performance metrics such as precision, recall, balanced accuracy, etc which are calculated using confusion matrix. It is observed with performance analysis that unsupervised ML algorithms works better than supervised ML algorithms. Performance graphs are plotted using bar type of graph to understand different classifier performance.

References

[1] Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study, 2018 IEEE International Conference on Information Reuse and Integration (IRI), Sahil Dhankhad; Emad Mohammed; Behrouz Far.

[2] 'Credit card fraud detection based on transaction behavior', TENCON 2017 - 2017 IEEE Region 10 Conference, John Richard D. Kho; Larry A. Vea

[3] The importance of credit cards: https://budgeting.thenest.com/importancecredit-cards-29514.html

[4] Low and Slow Is How the Credit Card Fraudsters Roll: https://www.threatmetrix.com/digital-identity-blog/fraudprevention/low-and-slow-is-how-the-credit-card-fraudsters-roll/

[5] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi and G. Bontempi," Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 8, pp. 3784-3797, Aug. 2018.

[6] L. Zheng, G. Liu, C. Yan and C. Jiang," Transaction Fraud Detection Based on Total Order Relation and Behavior Diversity," in IEEE Transactions on Computational Social Systems, vol. 5, no. 3, pp. 796- 806, Sept. 2018

[7] 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.

[8] S . Xuan, G. Liu, Z. Li, L. Zheng, S. Wang and C. Jiang, "Random forest for credit card fraud detection," 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), Zhuhai, 2018, pp. 1-6.

[9] Vaishali. Article: Fraud Detection in Credit Card by Clustering Approach. International Journal of Computer Applications 98(3):29-32, July 2014.

[10] J. O. Awoyemi, A. O. Adewunmi and S. A. Oluwa dare," Credit card fraud detection using machine learning techniques: A comparative analysis," 2017 International Conference on Computing Networking and Informatics (ICCNI), Lagos, 2017, pp. 1-9.

[11] The chargeback process in a credit card: https://chargebacks911.com/chargeback-process/

[12] The importance of credit cards: https://budgeting.thenest.com/importancecredit-cards-29514.htm

[13] Suraj Patil*, VarshaNemade, PiyushKumarSoni, Predictive Modelling for Credit Card Fraud Detection Using Data Analytics, International Conference on Computational Intelligence and Data Science (ICCIDS 2018)

[14] 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.

[15] 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 Intelligence Volume 72, June 2018, Pages 21- 29

[16] S. Dhankhad, E. Mohammed and B. Far," Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study," 2018 IEEE International Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, 2018, pp. 122-125

[17] Zareapoor, Masoumeh K.R., SeejaAlam, Afshar. (2012). Analysis on Credit Card Fraud Detection Techniques: Based on Certain Design Criteria. International Journal of Computer Applications. 52. 35-42. 10.5120/8184-1538.

[18] Rajeshwari U and B. S. Babu," Real-time credit card fraud detection using Streaming Analytics," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, 2016, pp. 439-444.

[19] Credit Card Fraud Detection: What Payment Gateways Can Do for You:

https://www.chargebee.com/blog/credit-card-fraud-detection-tools/

[20] Sethi, Neha and Anju Gera. A Revived Survey of Various Credit Card Fraud Detection Techniques. (2014).

[21] Jain, Rajni Gour, Bhupesh Dubey, Surendra. (2016). 'A Hybrid Approach for Credit Card Fraud Detection using Rough Set and Decision Tree Technique'. International Journal of Computer Applications. 139. 1- 6. 10.5120/ijca2016909325.

[22] Pumsirirat, Apapan Yan, Liu. (2018). 'Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine'. International Journal of Advanced Computer Science and Applications.