Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020)
IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281-4108-4

Fraud Detection in Credit Card Data using Unsupervised Machine Learning Based Scheme

Arun Kumar Rai*, Rajendra Kumar Dwivedi+

*+Department of IT & CA, MMMUT Gorakhpur, India
*arun.akrai@gmail.com
+rajendra.gkp@gmail.com

***Abstract*— Development of communication technologies and e-commerce has made the credit card as the most common technique of payment for both online and regular purchases. So, security in this system is highly expected to prevent fraud transactions. Fraud transactions in credit card data transaction are increasing each year. In this direction, researchers are also trying the novel techniques to detect and prevent such frauds. However, there is always a need of some techniques that should precisely and efficiently detect these frauds. This paper proposes a scheme for detecting frauds in credit card data which uses a Neural Network (NN) based unsupervised learning technique. Proposed method outperforms the existing approaches of Auto Encoder (AE), Local Outlier Factor (LOF), Isolation Forest (IF) and K-Means clustering. Proposed NN based fraud detection method performs with 99.87% accuracy whereas existing methods AE, IF, LOF and K Means gives 97%, 98%, 98% and 99.75% accuracy respectively.**

***Keywords*— Unsupervised Learning, Anomaly Detection, Fraud Detection, Auto-Encoder, Credit Card**

I. INTRODUCTION

Falsification of the credit card can be defined as the unapproved use of a customer's card data to create purchases or to dismiss funds from the cardholder's record. The misconduct extortion starts from the credit card when somebody incorrectly acquires the number printed on card or the essential records for the card to be operated [9,10]. The owner of the card, the agent by whom card is issued and even guarantor of a card might not be informed of the fraud until the record is used to create purchases. As shopping through internet-based applications and paying bills online has been come into practice, there is no longer requirement of a physical card to create purchases.

Figure 1 shows the taxonomy of frauds. Frauds can be categorized in three ways: financial frauds, communication frauds and online marketing frauds. Credit card frauds come under financial frauds. These frauds must be prevented and detected in time. In this direction, many researches are carried out by various researchers to devise the effective and efficient techniques [14, 15]. Hackers and intruders are trying different new approaches to breach the security [13]. Therefore, there should always be a safety alert against such frauds. Several machine learning based algorithms have been proposed in this direction. A learning based technique is proposed for detecting the credit card frauds.

Figure 1: Taxonomy of frauds

Rest of this paper is organized as follows. Section 2 presents the literature review of the related work. Section 3 presents the proposed work and section 4 provides the performance analysis of various unsupervised learning schemes used for fraud detection in credit card information. Finally, section 5 concludes the work with some future directions.

II. RELATED WORK

Here we present a review on the unsupervised learning approaches used to identify the frauds and a comparative analysis of this study is also given in table 1. There are two important categories of machine learning techniques to identify the frauds in credit card transactions: supervised and

unsupervised learning model. In supervised approach, early transactions of credit card are labelled as genuine or frauds. Then, the scheme identifies the fraud transaction with credit card data.

Altyeb Altaher Taha et al. [1] employed an intelligent fraud detection method for this problem. They worked on optimized light gradient boosting machine (LightGBM) and demonstrated good results as compared to the other existing schemes.

Surya Narain Kalid et al. [2] proposed the Multiple Classifier System (MCS) to tackle the problems that is based on their analysis using single classifiers. They presented through their results of experiments that MCS model outperformed than existing works.

Sara Makki et al. [3] have done the experimental work with a rigorous experimental study and the results to challenge the problem of imbalance arrangement. Their results of detecting frauds were outstanding than the existing schemes.

Kuldeep Randhawa et al. [4] used machine learning approach for fraud detection in this system. They first used the common methods and then the hybrid methods of AdaBoost and majority voting methods.

Andrea Dal Pozzolo et al. [5] has also contributed with the industrial partner towards the fraud detection issues that realistically describe the operating conditions of fraud detection systems by analysing massive streams of credit card transactions.

Lorenzo Meneghetti et al. [6] have used the isolation forest and local outlier factor for anomaly detection. It works better on unlabelled data set. The algorithm allows avoiding the subtasks of detection.

Lutao Zheng et al. [7] concluded the behaviour profile of logical graphs that are total ordering based methods to present the logical relation of attributes of transaction records. It is based on path transition probability from an attribute to another one. They described a data diversity coefficient based on entropy in order to depict the variety of transaction behaviour of a user.

Sonali Bakshi et al. [8] used the hidden Markov model, fuzzy bunching, neural system and data mining to identify the frauds. They described the card challenges of card holder and card guarantor. They identified the falsification performed by the intruder and provide the corrective actions procedures.

F. Carcillo et al. [11] presented a hybrid method that is the combination of unsupervised and supervised approaches to increase the accuracy of the fraud detection. Outlier score are calculated on several stages of granularity and found outperforming the existing schemes.

Table 1: Summary of the related works

Ref	Authors	Year	Issues	Techniques Used	Remarks
[1]	Altyeb Altaher Taha et.al.	2020	Fraud detection	OLightGBM	Better results
[2]	Surya Narain Kalid et.el.	2020	Anomaly detection	Multiple Classifier System	Works better
[3]	Sara Makki et.al.	2019	Fraud detection	C5.0, SVM, ANN	Devoted ecosystem based on big-data
[4]	Kuldeep Randhawa et.al.	2018	Fraud detection	AdaBoost, Majority voting	Noise is removed
[5]	Andrea Dal Pozzolo et.al.	2018	Fraud detection	Supervised learning	Feedbacks system interaction
[6]	Lorenzo Meneghetti et.al.	2018	Anomaly detection	Isolation forest and local outlier factor	Isolation forest works better on unlabelled data set
[7]	Lutao Zheng et.al.	2018	Fraud detection	Markov-chain model, path based model	Good metrics for accuracy
[8]	Sonali Bakshi	2018	Fraud detection	Hidden Markov model, fuzzy bunching, neural system and data mining	Finds the fraudsters well
[11]	F. Carcillo et.al	2019	Anomaly detection	Hybrid Algorithms using unsupervised learning	Performance is moderate

Now we can observe that machine learning techniques are very good for outlier detection. The proposed work is focused on unsupervised learning approaches. It is noticed that K-means and NN are the algorithms which generally perform better than the other existing schemes. But this performance can be further improved. Therefore, Neural Network is used widely for fraud detection. The proposed work is compared with the existing schemes namely AE, IF, LOF, K-means.

III. PROPOSED APPROACH

Here, the unsupervised learning approaches are used for fraud detections. Proposed work is presented through different flow charts which are shown in figure 2, 3 and 4. The flow charts show the steps to detect the frauds. It demonstrates that pre-processing is done to clean the data and to extract the features. Then we train the model and apply testing with various unsupervised learning models. Then performance is measured on various metrics. For identifying the outliers, we used five algorithms (NN, AE, IF, LOF, K-Means) one by one and evaluated the performance. We found that NN outperforms the other schemes. Figure 2 shows the first flowchart that represents the fraud detection model. Figure 3

presents the second flow chart that shows how the attacks are performed by the intruders. Figure 4 shows the third flow chart which describes how the transactions are performed by our system.

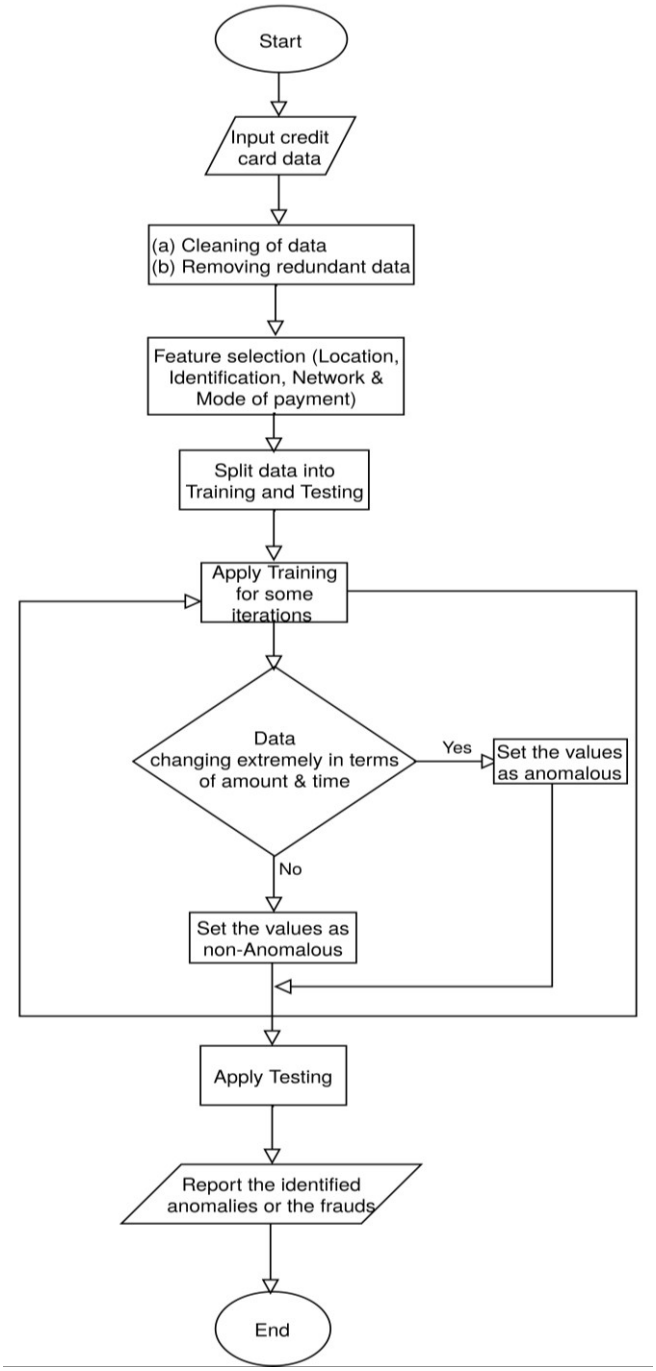


Figure 2: Flowchart of Fraud Detection

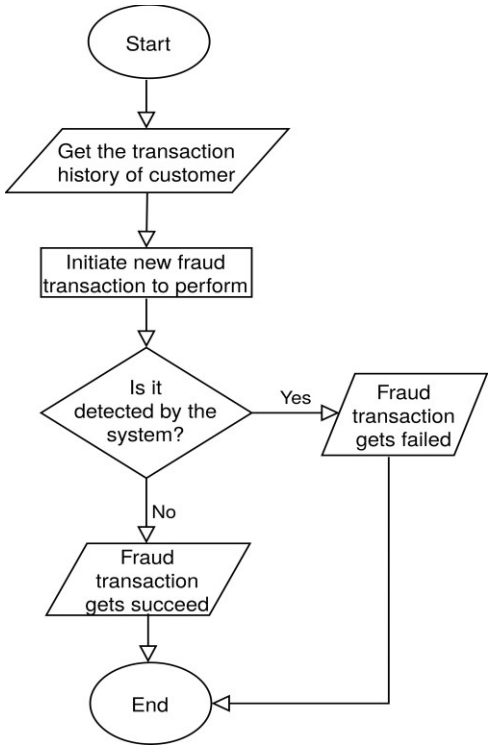


Figure 3: Flowchart for Performing Attacks (Performed by Intruder)

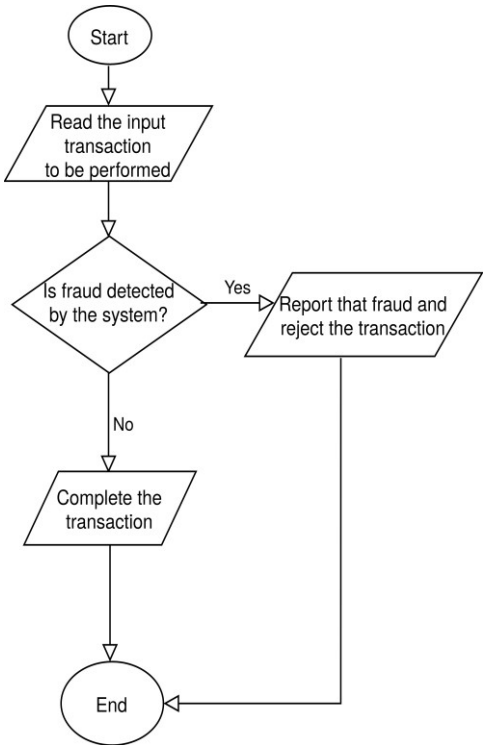


Figure 4: Flowchart for Performing Transactions (Performed by our System)

IV. Performance Evaluation

Proposed work is implemented in Python on a credit card system dataset [12] and the performance is evaluated on certain metrics which are explained in this section.

A. *Unsupervised learning approaches to detect the frauds*

We have used unsupervised learning approaches for detection of fraud in credit card data as discussed below.

Neural Network (NN): The idea of this method is to train the model using neural network methods with an enormous of some well-known samples then apply testing.

Auto Encoder (AE): It is an unsupervised Artificial Neural Network scheme that is very effective in learning from the data. The purpose of an auto-encoder is learning a presentation of a set of data, usually on behalf of dimensionality reduction. It trains the network by ignoring the noise.

Isolation Forest (IF): They are used to isolate the observations through randomised selection of a feature. It selects the values randomly.

Local Outlier Factor (LOF): LOF is a model of unsupervised learning technique which calculates the local compactness deviation of specified data point with respect to situation neighbours.

K-Means Clustering: This is an algorithm of unsupervised machine learning used for unlabelled data. It is a clustering based learning approach.

B. *Confusion metrics*

Classification matrix of the NN, AE, IF, LOF and K-Means algorithms is presented in Table 2. Here we can see that NN is outperforming the other algorithms. Its visual representation for NN, AE, IF, LOF and K-means methods is presented in Figure 5.

Table 2: Confusion Matrix for various unsupervised learning techniques

Approaches Used	True Negative	False Positive	False Negative	True Positive
NN	56851	13	18	80
AE	55491	1373	20	78
IF	55789	1075	48	50
LOF	55091	773	52	46
K-Means	55734	130	13	85

C. *Performance* measurements

Various performance metrics are presented in equation 1 to equation 7:

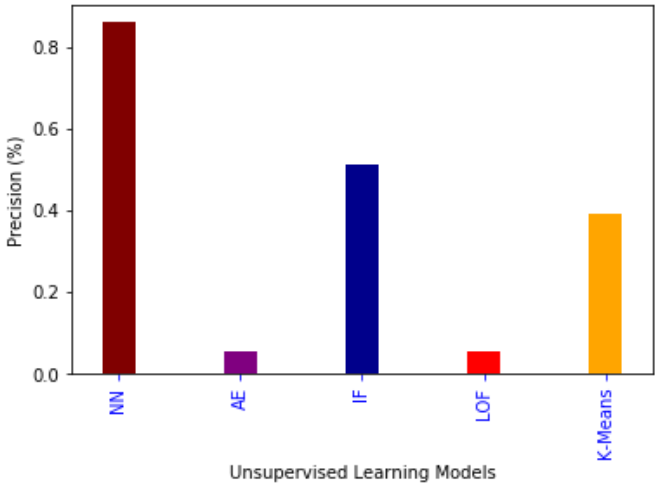
Precision: TP / (TP + FP) (1)
Recall: TP / (TP + FN) (2)
False Positive Rate: FP / (FP + TN) (3)
Specificity = TN / (TN + FN) (4)
Balance Accuracy: (Recall + Specificity) / 2 (5)
F1 Score: 2 * (Precision * Recall) / (Precision + Recall) (6)
Accuracy: (TP + TN) / (TP + FP + TN + FN) (7)

D. *Results*

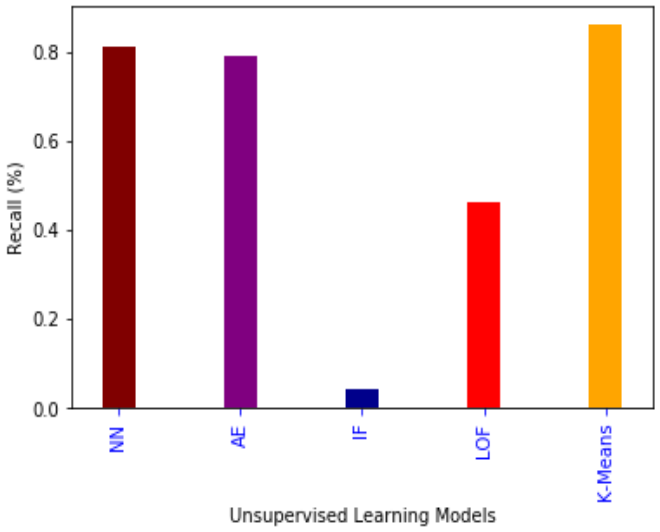
Fraud Detection algorithms are implemented in Python. The performance is compared in Table 3. Figure 5 presents various results of performance analysis of NN, AE, IF, LOF, K-Means clustering algorithms on different metrics.

Table 3: Performance comparison

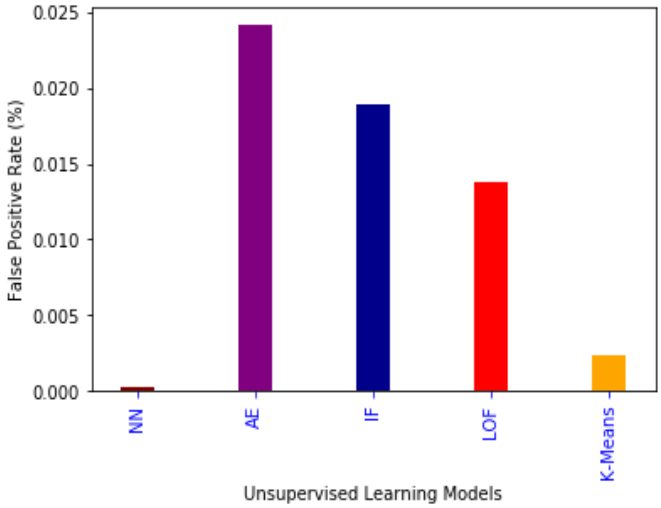
Metrics	Unsupervised learning models				
	NN	AE	IF	LOF	K-Means
Precision	0.86	0.053	0.51	0.056	0.39
Recall	0.81	0.79	0.044	0.46	0.86
False Positive Rate	0.00022	0.02414	0.01890	0.01383	0.00232
Specificity	0.99	0.97	0.83	0.98	0.090
Balanced Accuracy	0.90	0.88	0.67	0.72	0.47
F1 Score	0.83	0.100	0.081	0.100	0.54
Accuracy	0.99	0.97	0.98	0.98	0.99



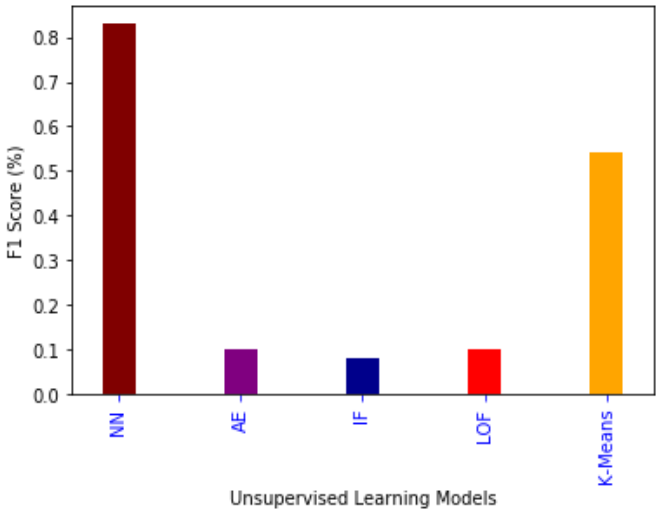
(a) Comparison of Precision



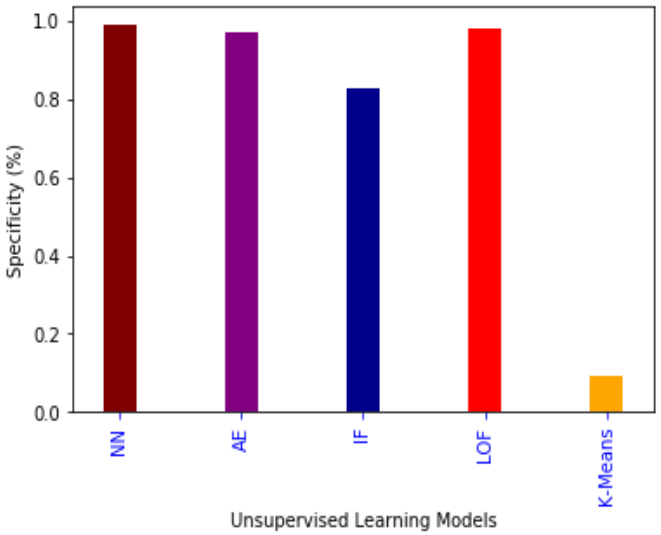
(b) Comparison of Recall



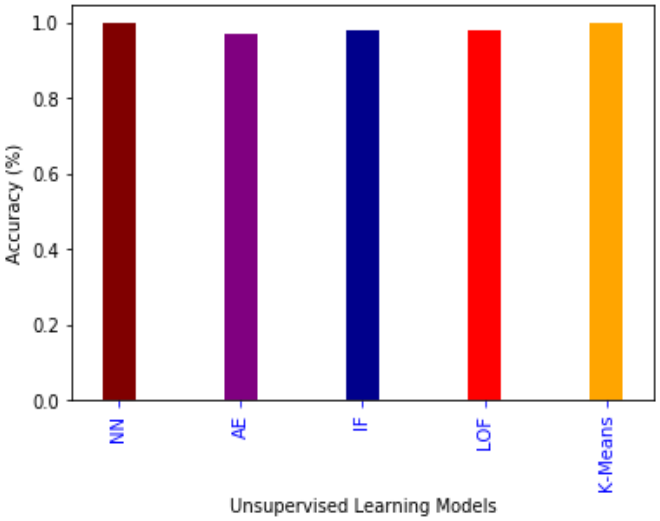
(c) Comparison of False Positive Rate



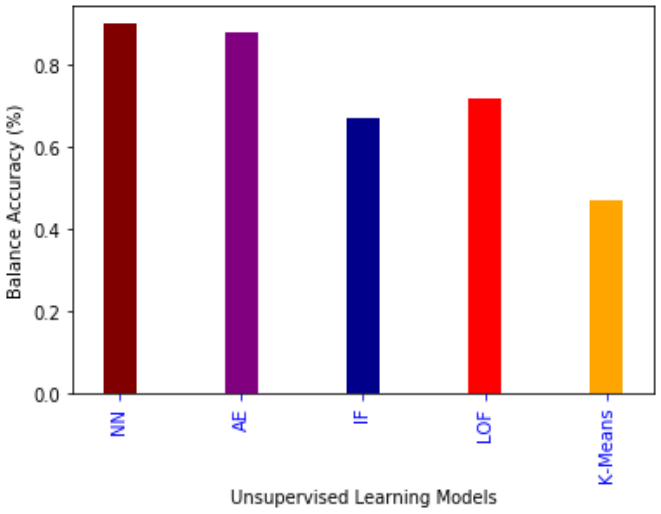
(f) Comparison of F1 Score



(d) Comparison of Specificity



(g) Comparison of Accuracy



(e) Comparison of Balance Accuracy

Figure 5: Performance measurements of NN, AE, IF, LOF and K-Means on different metrics

V. Conclusions and Future Directions

Here we presented a neural network based fraud detection scheme for fraud detection in credit card data in which unsupervised machine learning is used. Performance comparison of proposed work with the existing schemes viz., Auto Encoder, Isolation Forest, Local Outlier Factor and K means clustering is done on a credit card dataset. It can be observed that neural network based approach performs better than the existing schemes. Experimental results show that neural network based proposed approach provides 99.98% accuracy while accuracies of Auto Encoder, Isolation Forest, Local Outlier Factor and K Means clustering are 97%, 98%, 98 % and 99.97% respectively.

In future, we may extended this work to get some interesting scheme using supervised and semi-supervised models. We can test our model on the datasets of other applications too.

REFERENCES

[1] A. A. Taha, S. J. Malebary, "Intelligent Approach to Credit Card Fraud Detection Using an OLightGBM", IEEE Access (2020), pp. 25579-25587

[2] S. N. Kalid, K. H NG, G. K Tong, K. C Khore., "A Multiple Classifiers System for Anomaly Detection in Credit Card Data With Unbalanced and Overlapped Classes", IEEE Access (2020), Vol. 8, pp. 28210-28221

[3] S. Makki, Z. A Assaghir, Y. Taher, R. Haque, M. S Hacid, H. Zeineddine, "An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection", Special Section On Advanced Software And Data Engineering For Secure Societies, IEEE Access (2019), Vol 7, pp. 93010-93022

[4] K. Randhawa, C. K Loo, M. Seera, C. P Lim, A. K. Nandi, "Credit Card Fraud Detection Using Adaboost And Majority Voting", Ieee Access, (2018) Vol 6, pp 14277-14284

[5] Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi, Gianluca Bontempi, "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy", Ieee Transactions On Neural Networks And Learning Systems, (2018) Vol. 29, No. 8, pp. 3784-3794

[6] L. Meneghetti, M. Terzi, S. Del Favero, G. A Susto, C. Cobelli, "Data-Driven Anomaly Recognition for Unsupervised Model-Free Fault Detection in Artificial Pancreas", Ieee Transactions On Control Systems Technology, (2018) pp. 1-15

[7] L. Zheng, G. Liu, C. Yan, C Jiang, "Transaction Fraud Detection Based on Total Order Relation and Behavior Diversity", Ieee Transactions On Computational Social Systems (2018), pp. 1-11

[8] S. Bakshi, "Credit Card Fraud Detection A classification analysis", Second International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) IEEE (2018), ISBN 978-1-5386-1442-6, pp. 152-156

[9] R. K Dwivedi, A. K Rai, R. Kumar, "Outlier Detection in Wireless Sensor Network using Machine Learning Techniques", International Conference on Electrical and Electronics Engineering (ICE3), IEEE (2020), pp. 1-6

[10] R. K Dwivedi, A. K Rai, R. Kumar, "A Study on Machine Learning Based Anomaly Detection Approaches in Wireless Sensor Network", 10th International Conference on Cloud Computing, Data Science & Engineering, (Confluence). IEEE (2019), pp. 200-205.

[11] F. Carcillo, Y.-A. Le Borgne and O. Caelen et al., "Combining unsupervised and supervised learning in credit card fraud detection", Information Sciences, Elsevier (2019), pp. 1-15

[12] Dataset for credit card fraud. (2020). https://www.kaggle.com/mlg-ulb/creditcardfraud

[13] Rajendra Kumar Dwivedi, Sonali Pandey, Rakesh Kumar "A study on Machine Learning Approaches for Outlier Detection in Wireless Sensor Network" IEEE International Conference Confluence, (2018), pp. 189-192.

[14] Shantanu Rajora, Dong-Lin Li, Chandan Jha, Neha Bharill, Om Prakash Patel, Sudhanshu Joshi, Deepak Puthal, Mukesh Prasad et al., "A Comparative Study of Machine Learning Techniques for Credit Card Fraud Detection Based on Time Variance", IEEE Symposium Series on Computational Intelligence SSCI 2018, pp. 1958-1963.

[15] John O. Awoyemi, Adebayo O. Adetunmbi, Samuel A. Oluwadare et al., "Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis", IEEE, 2017, pp. 1-9.