Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 46 (2020) 941–948

Procedia MANUFACTURING

www.elsevier.com/locate/procedia

13th International Conference Interdisciplinarity in Engineering (INTER-ENG 2019)

Selection Features and Support Vector Machine for Credit Card Risk Identification

Naoufal Rtayli[a,*], Nourddine Enneya[a]

[a]*Laboratory of Informatics Systems and Optimization, Ibn Tofail University, Kenitra, Morocco*

Abstract

For identifying credit card risk in massive and high dimensionality data, feature selection is considered very important to improve classification performance and fraud identification process. One of the commonly used feature selection methods is Random Forest Classifier (RFC), which is very suitable for large dataset. RFC has a good performance; it tends to identify the most predictive features, which may provide a significant improvement in classification performance of credit card risk identification model. In this paper, we propose an enhanced Credit Card Risk Identification (CCRI) method based on the features selection algorithm as Random Forest Classifier and Support Vector Machine to detecting fraud risk. Our experimental results show that the proposed algorithm outperforms the Local Outlier Factor, Isolation Forest and Decision Tree in term of classification performance on a larger dataset.

Keywords: Credit card risk; Fraud identification;Imbalanced data; Support Vector Machine; Machine learning.

1. Introduction

The focus of this paper is to address the problem of detecting Credit Card Risk (CCR) using machine-learning methods. A credit card is among the most successful payment processes for online transactions in many advanced

* Corresponding authors. Tel.: +212 619 818 822.
E-mail address: naoufal.rtayli@uit.ac.ma

10.1016/j.promfg.2020.05.012

countries. With the evolution of advanced technologies such as the Internet of Things, mobile computing and the Internet; credit cards have brought online transactions simpler and more useful. Nevertheless, it has additionally opened up new fraud possibilities for criminals, thereby increasing the rate of fraud[1].

There are several definitions of fraud. Based on the definition provided by the Certified Fraud Examiner Association (CFEA)[2], fraud is the use of one's occupation for personal enrichment through misapplication of resources or assets that are owned by organizations or companies. The main reason behind the commitment of fraud is to realize a gain on the incorrect ground by an illicit means. This has a negative influence on the economics, law and even the moral values of humanity[3]. Table 1 shows the number of claims received by the IC3 and the number of dollar losses for each year between 2015 and 2018. We can observe from this table, the amount of loss regularly grows while the number of claims declines; that is because fraud is causing more losses now than in the past.

Table 1. IC3 Report on internet crime[4].

Year	Claims Received	Dollar Loss
2015	288,012	$1,070.7 M
2016	298,728	$1,450.7 M
2017	301,580	$1,418.7 M
2018	351,937	$2,706.4 M

The negative effect of CCR is frightening; many companies have estimated millions of dollars (US) in losses [5]. In addition, cybercriminals regularly develop advanced methods. Therefore, the critical task is to create an enhanced and powerful method, which must be able to adapt to rapidly evolving fraudulent models [1]. Reaching this task is very difficult because of the dynamic nature of fraud and lack of data set for researchers. In this context, this paper proposes a credit card risk detection model based on machine learning algorithms that have the capacity to identify small anomalies in huge data with a high level of precision. Precisely, this paper focused on extracting the most relevant features by using Random Forest Classifier method combined with Support Vector Machine method (SVM) in order to identifying the fraudulent transactions, the proposed model is tested on a large dataset.

The detection of frauds is the subject of numerous investigations and review articles; that can be based on topics such as areas of fraud, types of fraud, the approaches and techniques of fraud identification. M. Pejic-Bach et al. investigated the identification of fraud in different domains based on data mining techniques and statistics[6]. In addition, M. Behdad et al. studied the detection of fraud using nature-inspired techniques[7]. Nature-inspired techniques, as the name indicates, are artificial intelligence techniques inspired by the workings of natural systems. For example, the neural network is inspired by the central nervous system of an animal (especially the brain) capable of learning and recognizing itself. In the same field, S.B. E. Raj et al. analyzed diverse types of methods used to identify credit card risk [8]. Y. Rebahi et al. presented the problem of VoIP fraud and examined the fraud detection systems offered in various areas, as well as their usability in the context of VoIP[9]. These Fraud identification techniques are categorized into two approaches supervised and unsupervised[10][11][12] .

The supervised detection algorithm is a method based on supervised machine learning algorithms that are trained on some labelled data to build predictive models, which will allow us to predict the output of new unseen data; it is named supervised in view of the fact that the learning process is performed under the supervision of an output variable. Several supervised methods have been studied in the literature[13][14][15]. For example, Wei et al. built a fraud identification framework called ContrastMiner which merges the neural network application for exploring contrast patterns and the decision forest with the aim of obtaining high precision in the identification of fraud[16]. Moreover, Zareapoor et al. proposed a new strategy for handling data imbalance in credit card transactions in non-stationary environments[17]. A contrast vector is created for each client, based on his historical behavior. The model suggests a balancing approach that frequently exploits the extraction of sets of elements to provide effective predictions. Finally, a model includes Fraud Miner proposed by Hegazy et al. uses the Apriori algorithm for prediction, while Enhanced Fraud Miner uses Lingo, a clustering-based data mining method to identify frequent patterns[18].

The unsupervised detection algorithm is a machine learning method that occurs when there is no target variable and the learning algorithm looks for hidden structures in the data. It only exploits features to detect similar

patterns that are unknown[19]. It does not need any labels, nor is there any distinction between training and test dataset. The important advantage of using the unsupervised algorithm is that this kind of technique does not rely on labelled data that could be in short supply or non-existent[19]. For example, there is Principal Component Analysis (PCA), it is often used as a process of transformation[20][21]. It is helpful to convert the original dataset in a way that facilitates the detachment of natural instances and anomalies, In addition, the author Khan et al. proposed a method to detecting credit card fraud based on Hidden Markov Model (HMM)[22]. The method performs detection using spending patterns of cardholders. Moreover, Ghobadi et al. have proposed a prediction model for cost-sensitive neural network fraud[23]. Cost-sensitive models are gaining importance because of the high levels of correspondence between their predictions and their decision-making process. The model proposed in [23] identifies fraud rules by creating a cost-sensitive neural network. However, additional training for handling the conceptual drift is impossible because of the high computing requirements of neural networks.

The paper is structured as follow: Section 2 comprises the proposed model, section 3 shows the definition of evaluation measures, section 4 contains the results obtained and finally, section 5 closes the article.

2. The proposed model

The quantity of transactions identified as fraud is usually a very small portion of the total transactions. Consequently, the identification of fraudulent transactions is actually difficult. Hence, the development of an adequate method that can detect fraud transactions from millions of normal transactions appears primordial. In this regard, we propose our model that is formed of two steps:

- Step1: Selection of relevant features based on Random Forest Classifier Method in order to increase the performance of the model.

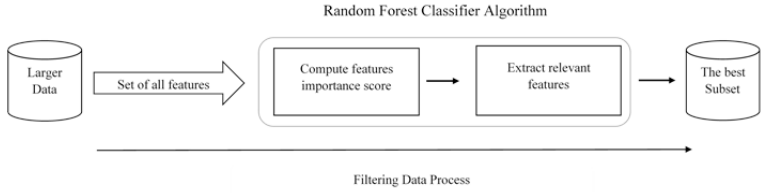

Fig. 1. Schematic illustration of filtration of data and extract relevant features.

- Step2: Identification of fraudulent transactions although Support Vector Machine method.

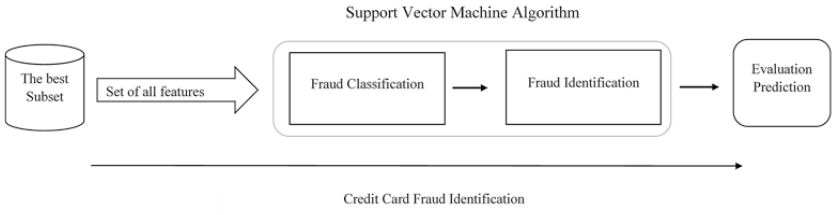

Fig. 2. Schematic illustration of Credit Card Fraud Identification.

The extraction of strong predictors and non-redundant features to improve classification performance is a critical task where researchers attempt to extract the least number of relevant features that can improve CCRI method[24]. In this context, the proposed model utilizes the **importance scores** to choose the extremely discriminative features.

2.1. Computing Feature Importance Score

To calculate feature importance score, Random Forest Classifier uses the Gini index to construct decision trees and to determine final class in each tree. Therefore, the Gini index at node v,$g(v)$, measures the impurity of v, as:

$$g(v) = \sum_{j=1}^{J} Pj(1 - Pj) \quad (1)$$

Where Pj is the fraction of class – j (j is an index) records at node v.

The Gini information gain of feature X_i (i is an index) in purpose to divide a tree node v,g(X_i, v), is defined as follows:

$$g(X_i, v) = g(X_i, v) - (f_l\, g(X_i, v^l) + f_r\, g(X_i, v^r)) \quad (2)$$

Where $g(X_i, v)$ is the impurity at node v, $v^l$and v^r are the left and right child node of parent node v, respectively; f_l and f_r are the fraction of examples assigned to the left and right child node. Furthermore, the feature that maximizes the reduction in impurity is adopted as the splitting feature. Finally, the importance score for feature X_i is calculated from the $g(X_i, v)$ as:

$$Impt_i = \frac{1}{N_{tree}} \sum_{k \epsilon SX_i} g(X_i, v) \quad (3)$$

N_{tree} is the number of the tree in the random forests, and $k \epsilon SX_i$ is the set of split nodes. Then, the normalization of the importance score is defined as:

$$Impt_{norm} = \frac{Impt_i}{Impt_{max}} \quad (4)$$

$Impt_i$ represents the importance score of X_i from RF and $Impt_{max}$ represents the maximum importance, and the normalized importance score lies between $0 \leq Impt_{norm} \leq 1$.

2.2. Identifying credit card risk

In the application of CCRI, we adopt Support Vector Machine (SVM) to construct detection model where negative class means fraudulent transaction, which is more concerned. SVM is one of the most powerful methods. It is a supervised method that can be used for making classification decisions by transforming feature vectors into high-dimensional space, and identifying hyperplanes (lines separating the data points) to divide the space so as to separate the behavioral characteristics belonging to different classes. Multiple hyperplanes can be used and the optimal hyperplane will be the line that maximizes class separation and minimizes misclassification. In this model, we use each transaction as a point in a k-dimensional space (where k is a number of entities), the value of each entity is the value of a certain coordinate. Then, we carry out the classification by determining the hyperplane which adequately distinguishes the two classes (Legit / Fraud) very well.

Using hyperplane to perform binary analysis on the provided data of training, and the sampling features are $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots, (X_m, Y_m)$, where $X_i \in R_k$, $Y_i \in \{-1, +1\}$, and the vector is the vector formed by certain characteristics of the sample [25]. The key to the SVM method is to determine the F function. So that X independently of the sample can obtain the corresponding Y by F after sample formation, then a hyperplane designated by F can be determined after formation; it can divide the formation samples into positive and negative classifications, and then divide the other X from the sample. If the data are not linearly separable, the algorithm operates by matching the data to a higher dimensional entity space using a nonlinear kernel function Φ (X), then an optimized hyperplane is established in the same space. The algorithm can be written as below[25].

$$\min_{w,b,\xi} \ \frac{1}{2} w^T w + C \sum_{i=1}^{p} \xi_i$$
$$Subject\ to\ Y_i(w^T \emptyset(X_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \dots, m \quad (5)$$

In this algorithm, $w^T \emptyset(X_i) + b = 0$ specifies the hyperplane of separation, w s the usual the hyperplane vector, b is their offset. The $C > 0$ is the penalty parameter of the error term and w are the weight coefficient of the hyperplane, $\boldsymbol{\xi}_i = \{\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_m\}$ is a vector which gives the measure of sanction for a poorly classified sample. The proposed model is detailed in the algorithm below:

Algorithm : SVM based on RFC	
1:	First step:
2:	employ RFC to build decision trees using Gini index
3:	determinate final class in each tree
4:	measures the impurity of features
5:	$g(v) = \sum_{j=1}^{j} Pj(1 - Pj)$
6:	Gini information gain of feature in order to split a tree node
7:	$g(X_i, v) = g(X_i, v) - \left(f_l\, g(X_i, v^l) + f_r\, g(X_i, v^r)\right)$
8:	Compute importance score for feature
9:	$Impt_i = \frac{1}{N_{tree}} \sum_{k \epsilon SX_i} g(X_i, v)$
10:	normalization of the importance score
11:	$Impt_{norm} = \frac{Impt_i}{Impt_{max}}$
12:	Compute Accuracy score per number of variables
13:	Choose the threshold that maximize the performance of the model and minimize the number of variables
14:	Threshold: 0.1096
15:	Extract the important features using threshold = 0.1096
16:	Create a small subset composed from the predictive features
17:	Final step:
18:	employ SVM non-linear kernel function Φ(X)
19:	define the separating hyperplane
20:	$w^T \emptyset(X_i) + \mathrm{b} = 0$
21:	divide the training samples into positive (Normal) and negative (Fraud)
22:	evaluate prediction

3. Evaluation metric

In order to evaluate the performance of the proposed model, we use some techniques available to determine the predictability of the model, Accuracy metric, Area Under Curve and Recall[26]. The most important, there are:

- **Accuracy**: is a rate of correctly predicted transactions to the total transactions. It is defined as:

$$Accuracy = TP+TN/TP+FP+FN+TN \quad (6)$$

- **Recall**: is the rate of correctly predicted positive transactions to all transactions in actual class – True. It is defined as :

$$\text{Recall} = \text{TP/TP+FN} \quad (7)$$

- **Area Under the Curve** : is a single scalar value that measures the overall performance of the model. The value of AUC is in the range [0.5–1.0], the where the smallest value represents the performance of a random classifier and the highest value corresponds to a perfect classifier [26].

where :

- True Positive (TP): The fraudulent cases that the algorithm detected as fraud.
- False Positive (FP): The non-fraud cases that the algorithm detected as fraud.
- True Negative (TN): The non-fraud cases that the algorithm detected as non-fraud.
- False Negative (FN): The fraudulent cases that the algorithm detected as non-fraud.

4. Results

4.1. The used technology

We used the python language, because of its popularity in data science[27][28]. It has high-quality data science libraries with good documentation and a practical development environment - jupyter notebooks - which is great for quick and easy visualization and exploration of the data[27]. To effectively manipulate the data, we use the numpy and pandas libraries[28]. Numpy allows us to use much less memory for arrays than the default python lists and also makes matrix operations much more efficient. Pandas is built on numpy and provides a higher-level interface for manipulating datasets with named rows and columns. To measure the performance of our detection model, we used some helper functions from the sklearn library[28]. Finally, we used matplotlib[28] to create the figures presented throughout this paper. The model is tested by using the following materiel requirements: the mobile workstation Dell Precision 5522 equipped with the CPU i7 7850H, RAM memory of 32 Go, SSD M2 of 512 Go and the GPU Nvidia Quadro M1200.

4.2. Exploration of data

The dataset is produced of European transactions cardholders[29]. It has been made through a two-day period in September 2013. It was basically collected by Libre Brussels University with the aim of analyzing massive data and transactions that are frauds. This dataset is formed only of numerical input variables that are the outcome of a PCA conversion. It includes 492 fraudulent transactions out of 284,807 transactions. The dataset is extremely unbalanced as shown in the following Fig. 3.

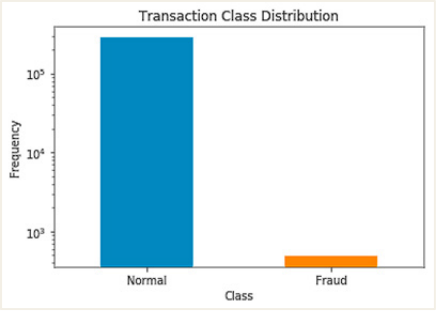

Fig. 3. Presents the distribution of transactions.

4.3. Results Analysis

This part of the paper analyses the implementation efficiency and effectiveness of SVM based on RFC. In order to validate their performance, it is compared with Isolation Forest (iForest), Decision Tree (DTree) and local Outlier Factor (LOF)[30][31], we utilize Accuracy, Recall and Area Under Curve as evaluation measures that are defined in the evaluation metrics section.

4.3.1. Accuracy and Sensitivity

First, we start by reporting the accuracy and the sensitivity as results of the proposed model. We can observe in Fig. 4. that SVM based on RFC shows good accuracy. It has a 95.12% isn't more accurate than LOF of 99.6 %, iForest of 99.7% and Decision Tree of 99.8%. Nevertheless, in Credit Card fraud detection field, the sensitivity of the model is decisive and considered as the most important metric than accuracy. Recall (or sensitivity) parameter permit us to evaluate CCRI performance[26] [26]. It compares the number of transactions correctly classified as fraud to the number wrongly recorded as fraud[26]. The fig. 4. presents also the sensitivity of each method in identifying fraudulent transactions. It is can be observed that the SVM based on RFC performed much better than the iForest, LOF and Dicision Tree; 87 %, 34%, 5% and 0 % respectively.

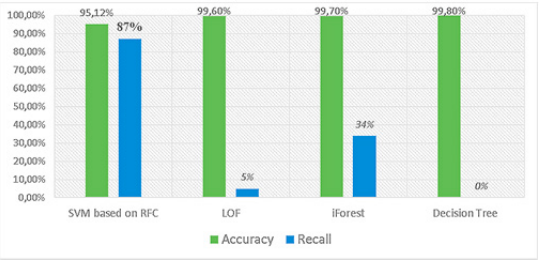

Fig. 4. Overall Accuracy and Sensitivity for each Method.

4.4. Area Under Curve:

The AUC is a good global metric to assess the performance of score classifiers for the reason that its calculation is based on the complete ROC curve and therefore implies all possible classification thresholds[26]. In Fig. 5, it is can be observed that the proposed model has a very good identification capability. It has got 91% that is more precise and faster in identifying Credit Card Fraud than LOF (52%), iForest (67%) and DTree (50%).

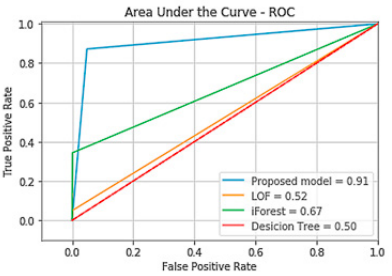

Fig. 5. Classification performance for each method.

The AUC-Roc measure also validates that the proposed method has the best performance and skills in identifying if the transaction presents a risk or not.

5. Conclusions

The proposed added some ameliorations in terms of CCRI that helps to increase both the sensitivity and the classification performance, which are the most important measures to evaluate Credit Card Risk Identification model. The main advantages of SVM based on RFC are: Firstly, the model has a good accuracy rate to **95**%. Secondly, it decreases the number of false positive transactions by improving the sensitivity rate to **87**% in a large and unbalanced dataset where the rate of frauds is very low (<**0.17**%), which is very convenient for the businesses to

minimize the high charge of investigation activity. Finally, it has got a high rate (**91**%) in term of classification performance.

References

[1] W. Zhou and G. Kapoor, Detecting evolutionary financial statement fraud, *Decis. Support Syst.*, 2011.
[2] ACFE, Report to the Nations 2018 Global Study on Occupational Fraud and Abuse, 2019.
[3] P. Alexopoulos, K. Kafentzis, X. Benetou, T. Tagaris, and P. Georgolios, Towards a generic fraud ontology in e-government, in *ICE-B 2007 - Proceedings of the 2nd International Conference on e-Business*, 2007.
[4] ***, 2018 INTERNET CRIME REPORT, pp. 1–28. [Online], Available at: https://pdf.ic3.gov/2018_IC3Report.pdf
[5] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, Data mining for credit card fraud: A comparative study, *Decis. Support Syst.*, vol. 50, no. 3, pp. 602–613, 2011.
[6] M. Pejic-Bach, Profiling intelligent systems applications in fraud detection and prevention: Survey of research articles, in *ISMS 2010 - UKSim/AMSS 1st International Conference on Intelligent Systems, Modelling and Simulation*, 2010.
[7] M. Behdad, L. Barone, M. Bennamoun, and T. French, Nature-inspired techniques in the context of fraud detection, *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, 2012.
[8] S. Benson Edwin Raj and A. Annie Portia, Analysis on credit card fraud detection methods, in *2011 International Conference on Computer, Communication and Electrical Technology, ICCCET 2011*, 2011.
[9] Y. Rebahi, M. Nassar, T. Magedanz, and O. Festor, A survey on fraud and service misuse in voice over IP (VoIP) networks, *Inf. Secur. Tech. Rep.*, 2011.
[10] C. C. Aggarwal, *Outlier analysis*. 2013.
[11] R. Laxhammar, Anomaly Detection, in *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications*, 2014.
[12] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, A review of novelty detection, *Signal Processing*. 2014.
[13] A. Zafar and M. Sirshar, A Survey on Application of Data Mining Techniques; It's Proficiency In Fraud Detection of Credit Card, *Res. Rev. J. Eng. Technol.*, 2018.
[14] R. R. Popat and J. Chaudhary, A Survey on Credit Card Fraud Detection Using Machine Learning, in *Proceedings of the 2nd International Conference on Trends in Electronics and Informatics, ICOEI 2018*, 2018.
[15] P. Kumari and S. P. Mishra, Analysis of Credit Card Fraud Detection Using Fusion Classifiers, in *Advances in Intelligent Systems and Computing*, 2019.
[16] W. Wei, J. Li, L. Cao, Y. Ou, and J. Chen, Effective detection of sophisticated online banking fraud on extremely imbalanced data, pp. 449–475, 2013.
[17] M. Zareapoor and J. Yang, "A Novel Strategy for Mining Highly Imbalanced Data in Credit Card Transactions," *Intell. Autom. Soft Comput.*, 2017.
[18] M. Hegazy, A. Madian, and M. Ragaie, Enhanced Fraud Miner: Credit Card Fraud Detection using Clustering Data Mining Techniques, 2016.
[19] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, Deep clustering for unsupervised learning of visual features, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018.
[20] G. H. Li *et al.*, The flywheel fault detection based on Kernel principal component analysis, in *Proceedings of 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2019*, 2019.
[21] P. J. Rousseeuw and M. Hubert, Anomaly detection by robust statistics, *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 2018.
[22] A. Khan, T. Singh, and A. Sinhal, Implement credit card fraudulent detection system using observation probabilistic in hidden Markov model, in *3rd Nirma University International Conference on Engineering, NUiCONE 2012*, 2012.
[23] F. Ghobadi and M. Rohani, Cost sensitive modeling of credit card fraud using neural network strategy, in *Proceedings - 2016 2nd International Conference of Signal Processing and Intelligent Systems, ICSPIS 2016*, 2017.
[24] K. H. Hu, F. H. Chen, and W. J. Chang, Application of correlation-based feature selection and decision tree to detect earnings management and accounting fraud relationship, *ICIC Express Lett. Part B Appl.*, 2016.
[25] Q. Lu and C. Ju, "Research on credit card fraud detection model based on class weighted support vector machine, *J. Converg. Inf. Technol.*, 2011.
[26] J. Lever, M. Krzywinski, and N. Altman, Classification evaluation, *Nat. Methods*, 2016.
[27] A. Watson, S. Bateman, and S. Ray, PySnippet: Accelerating exploratory data analysis in Jupyter Notebook through facilitated access to example code, *CEUR Workshop Proc.*, vol. 2322, pp. 6–9, 2019.
[28] I. Stancin and A. Jovic, An overview and comparison of free Python libraries for data mining and big data analysis, *2019 42nd Int. Conv. Inf. Commun. Technol. Electron. Microelectron.*, pp. 977–982, 2019.
[29] ***, Credit card Fraud data - dataset by raghu543 | data.world. [Online]. Available: https://data.world/raghu543/credit-card-fraud-data.
[30] G. S. Na, D. Kim, and H. Yu, DILOF: Effective and memory efficient local outlier detection in data streams," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018.
[31] A. Singh and A. Jain, Adaptive Credit Card Fraud Detection Techniques Based on Feature Selection Method, in *Advances in Intelligent Systems and Computing*, 2019.