Credit Card Fraud Detection System

Jatin Choudhary
AIT-CSE
Chandigarh University
India
saraanjatin91@gmail.com

Jatin Kumar Saini
AIT-CSE
Chandigarh University
India
jatinkumarsaini2000@gmail.com

Mriganka Das
AIT-CSE
Chandigarh University
India
Mriganka3222@gmail.com

Saksham Bhatia
AIT-CSE
Chandigarh University
India
sakshambhatia2018@gmail.com

***Abstract*—We are continuously moving toward the online banking system. All the things were converted into an online mode where users can do transactions anywhere at any time. But the rate of cybercrime and fraud is increasing day by day. Our project's main purpose is to make people aware of the ongoing online credit card fraud. The primary goal of the system for detecting credit card fraud is the necessity to safeguard our transactions and improve security. With the solution provided by our research, fraudsters won't have a chance to make multiple transactions from stolen or counterfeit cards. Instead, the cardholder will get aware of the fraud before any activity by the fraudster. The model can also detect whether any whether a new transaction is fraudulent. Our goal here will be to provide a model to detect 100 percent thereby minimising the erroneous fraud categories, of the fraudulent transactions. When a credit card is lost physically or when private information about a credit card is lost, it is considered to be fraudulent. For detection, there are numerous machine-learning techniques available. This study demonstrates a number of methods that can be used to categorize transactions as real or fraudulent. The research made use of the Credit Card Fraud Detection dataset. SMOTE technique was employed for oversampling due to the dataset's extreme imbalance. Also, the dataset was divided into training data and test data after feature selection and dataset splitting. Naive Bayes, Multilayer Perceptron, Random Forest, and Logistic Regression were the algorithms employed in the experiment. The results show that each algorithm is quite accurate in spotting credit card fraud. The suggested model can be used to discover further anomalies.**

I. Introduction

Having a credit card is like having work as an instrument through which users can do online transactions. It is provided by a financial institution or by an organization, it allows users to borrow funds. The limit of the credit card is determined by the credit score, income, and credit history of the user. It can be used for shopping, electricity bills, restaurants, electronic devices, etc.

A. Credit Card Fraud

It refers to a scammer using your card number and pin for transactions without your knowledge, or they have stolen the card for financial transactions from your account. Fraud involving the use of a payment card of any kind is just one of the many various shapes and forms that credit card fraud can take. The causes of credit card fraud are likewise diverse. Some are intended to take money out of your accounts, while others want free products. It's also critical to understand the connections between credit card fraud and identity theft. According to the Federal Trade Commission, some 5 percent of all people over 16 in this country have been or will be the victim of identity theft. Additionally, it was discovered that the prevalence of identity theft has increased by 21 percent since the last count in 2008. The percentage of identity theft cases connected to credit card fraud, on the other hand, dropped, which is encouraging and a tribute to law enforcement officials and the general public as a whole. .

Identify applicable funding agency here. If none, delete this.

B. Types of credit card fraud:

1) Application Fraud:- Identity theft frequently occurs in tandem with fraud on applications. It occurs when a person else requests credit or a new credit card in your name. They typically steal supporting documents first, which are then used to support they made a fake application. To avoid this kind banks have a history put in place a number of safety precautions. The most crucial one is requiring only original documentation. Additionally, they frequently call employers to verify their identification. Unfortunately, fraudsters frequently supply fake phone numbers for employers and falsify paperwork. Unfortunately, there are always methods to get around some safety precautions.
2) Electronic or Manual Credit Card Imprints **-** Credit card imprints represent a second type of credit card fraud. This indicates that data stored on the card's magnetic strip has been skimmed. Following that, this is utilized to encode a bogus card or carry out fraudulent activities.
3) CNP (Card Not Present) Fraud **-** Someone can commit CNP fraud against you if they know your card's expiration date and account number. You can do this over the phone, via mail, or online. In essence, it indicates that someone uses your card without having it in their actual possession. The card verification code is becoming more and more frequently required by retailers, which makes CNP fraud significantly more challenging. However, if a fraudster can obtain your account number, they probably also have that number. The verification code's possible permutations are also limited to 999. As a result, many crooks attempt to place very low-value orders until they determine the correct quantity. Therefore, keep an eye out for tiny payments on your statements.

4) Counterfeit Card Fraud- The most typical scheme for fraudulent use of fake cards is skimming. Therefore, a magnetic swipe card that is fake could contain all of your credit card details. Then, a completely functional counterfeit card is made using this false strip. Since it is essentially an exact clone, fraudsters can just swipe it in a machine to pay for certain items. The use of your card information for this type of fraud is also possible. They can produce a so-called ”fake plastic” using this knowledge. Here, the card’s chip or magnetic stripe doesn’t actually function. However, it is frequently simple enough to persuade a merchant that there is a problem with the card, in which case they will manually process the transaction.
5) Mail Non-Receipt Card Fraud - Never received issue fraud and intercept fraud are other names for this kind of scam. You were hoping to receive a fresh card or a replacement in this instance, but a criminal was able to intercept these. After registering the card, the criminal will use it to make purchases and other uses.
6) Lost and Stolen Card Fraud - The following sort of fraud involves involving lost or stolen cards. Your card will be seized from you in this situation, either through theft or loss. Once they get it, the thieves will utilize it to make payments. Due to the need for a PIN, doing this through machines is challenging. To make online transactions, it is simple enough to utilize a recovered or stolen card. You must cancel your cards as soon as you notice they are gone because of this.

There are also two sorts of credit card fraud. The first is stealing the actual card, while the second entails collecting private information from the card, such as the card number, CVV code, kind of card, and others. Before the victim is informed, a fraudster can steal a significant amount of money or use stolen credit card information to make pricey transactions. Businesses use a number of machine-learning approaches to distinguish between fraudulent and legitimate transactions. This study will evaluate various machine learning techniques, such as Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), and Multilayer Perceptron (MLP), in order to determine which machine learning algorithm is most effective for identifying credit card fraud. ing, utility costs, dining out, electronics, etc.

II. Related work

Researchers were driven to develop a method to identify and stop fraud by the significant loss that fraudulent operations are creating. Many strategies have previously been put forth. and examined. [1] This is a brief review of some of them. Traditional methods have been shown to be effective, including Gradient Boosting (GB), Support Vector Machines (SVM), Decision Trees (DT), LR, and RF. In the study, the use of GB, LR, RD, SVM, and a mix of specific classifiers produced a high recall on a European dataset of over 91 percent. Only after balancing the dataset by undersampling the data were high precision and recall attained. [2] In the publication, the European dataset was also employed, and models based on LR, DT, and RF were compared. RF turned out to be the best model out of the three. k-Nearest neighbors (KNN) and outlier detection algorithms, according to them, can also be effective in detecting fraud detection. They have been shown to be effective in reducing false alarm rates and raising the fraud detection rate. In an experiment for their publication [3], the authors tested and compared KNN with other traditional algorithms, and found that it performed well. Unlike the previously listed publications, the study compared certain traditional methods and used machine learning methods. All of the strategies that were examined had an accuracy rate of about 80 percent. Using a European dataset, the authors of the paper compared the performance of the following algorithms: RF, GB, LR, SVM, DT, KNN, NB, XGBoost (XGB), MLP, and stacking classifier (a mixture of many machine learning classifiers). In-depth data preparation enabled all algorithms to achieve high accuracy of over 90 percent. The most effective classifier, [4] with 95 percent accuracy and 95 percent recall, was stacking. A neural network was tested using the European dataset in the paper. Back propagation neural network optimization using the Whale algorithm was part of the experiment. Twenty hidden layers, two output layers, and two input layers made up the neural network. Using 500 test samples, they got remarkable results thanks to the optimization algorithm: 96.40 percent accuracy and 97.83 percent recall. The authors of the papers employed neural networks to show how using ensemble approaches can improve results. [5] Autoencoder and Restricted Boltzmann Machine techniques were compared using three datasets in the paper, and the results showed that algorithms like MLP can be Mchine Learning fraud detection has been the subject of numerous articles. Although these models are expensive to compute and work better with larger datasets [6]. As we’ve seen in some articles, this strategy might produce excellent outcomes, but what if the same outcomes—or even better ones—could be obtained with fewer resources? Our major objective is to demonstrate that various machine learning algorithms may produce respectable results with the right preprocessing. While the majority of the authors of the papers mentioned employed the undersampling technique, utilizing the oversampling technique was a distinct strategy. [8] A comparison of the applicability of LR, RF, NB, and MLP for credit card fraud detection was made by the authors of this work in light of the available information. An experiment was carried out to achieve that [9]. Credit card fraud is a significant issue and comes at a significant cost to banks and card issuer businesses, according to the fraud detection system that has been proposed. Due to the extreme difficulty in transactions, banks take credit card fraud very seriously and have extremely intricate security systems to monitor transactions and spot fraud. [**?**]as quickly as possible one time it is dedicated. The aim of this survey is to achieve an overall review of different fraud detection methods and selects several innovative methods [10]. Engineering traits that are suggestive of fraudulent transactions are a key component of fraud detection. [11] This style of feature engineering has historically relied primarily on a manual creative process, where domain knowledge and experience serve as the main sources of inspiration for feature creation. [12] However, in recent years, computing power—in the form of autoencoders and other neural network mechanisms—has successfully replaced domain knowledge in other fields. It is a technique for unsupervised learning that uses neural networks to build features. By encoding the data using a function and then decoding the data into new features, it can rebuild new features, also known as predictors. [**?**] One of the main goals of this work is to compare features created to those that were manually built using domain expertise. In this paper, two methodologies—Decision Tree-based Algorithm and Logistic Regression Algorithm—are used to design [14]algorithmic approaches for credit card fraud/scam detection. As a result, an effort has been made to identify and address the problems of frauds in the credit card business system. [15]

III. Proposed Methodology

In this investigation, the Credit Card Fraud Detection dataset from Kaggle was utilized. The two-day transactions made by cards throughout Europe in September 2013 are included in this dataset. [16] In the dataset, there are 31 numerical features. As some of the input variables involve financial information, the PCA These input variables were changed in order to keep the data's confidentiality. Three of them did not convert the specified features. The "Time" feature shows the amount of time that has passed between the dataset's initial transaction and each following transaction. "Amount" designates a feature that shows the total value of credit card transactions. [17] The label is represented by the feature "Class," which only accepts the values 1 or 0, depending on whether the transaction is fraudulent.284,807 transactions total in the sample, 492 of which were frauds and the rest genuine. [18] With only 0.173 percent of transactions being categorized as frauds, we can see how highly skewed this dataset is by looking at the numbers. [19]. Data preprocessing is necessary because the distribution ratio of classes plays a crucial role in the precision and accuracy of the model. [20]

Algorithms for machine learning that are applied in fraud detection

1) Logistic Regression: The sigmoid function and logistic regression are compatible because the sigmoid function may be used to categorize the output that is a dependent feature and it uses probability to do so. Because the sigmoid function is utilized, this approach performs well with small data sets. If the sigmoid function's output value is larger than 0.5, the result will be 1, and if it is less than 0.5, the result will be 0. However, this sigmoid function is not appropriate for deep learning since, in deep learning, we must update the weights when going back from the output to the input in order to reduce the error in the weight update. We must. If the intermediate layer neuron's differentiation of the sigmoid activation function yields a value of 0.25, this will have an impact on the module's accuracy in deep learning.
2) Decision Tree:The categorization and regression issues can be solved using a decision tree. The working process is the same for both, although some formulas will vary. Entropy and information gain are used in the classification issue to develop the decision tree model. Information gain describes how much information we can extract from this feature, whereas entropy describes how random the input is. The decision tree model for the regression problem is built using the Gini and Gini indices. When solving classification issues, the root node is chosen based on its information gain, specifically its high information content and low entropy. Using Gini, which selects the feature with the least amount of data, the root node in regression issues is chosen. Here, Gini is chosen as the root. The parameter can be used to calculate the depth of the tree. Utilizing the grid search cv technique will provide optimization.
3) Random Forest: The random forest uses hyperparameter optimization to establish the decision tree's number and randomly selects the features that are independent variables as well as rows by rows. The output from each decision tree model inside the random forest, for the categorization problem statement, is the maximum occurrence output. In both deployed models and real-world scenarios, this is one of the extensively utilized machine learning algorithms. And this algorithm is utilized to answer the problem in the majority of the Kaggle computation challenges.
4) Naive Bayes: The machine learning algorithm for the classification problem, called Naive Bayes, relies on a feature of the Bayes theorem. The Naive Bayes theorem is used here to compute the probability of the dependent feature with respect to independent features. It may be applied by employing features in data sets with independent features as input and dependent features as output. The equation for a credit card fraud detection system can differ depending on the specific algorithm or method utilized to identify fraud. One common tactic is to use machine learning algorithms to look for patterns and irregularities in credit card transactions that could indicate fraud.

   EQUATION:- $p = 1 / (1 + e^{(} - z))$

where:

p is the predicted probability of fraud e is the mathematical constant approximately equal to 2.71828 z is a linear combination of the input features, weighted by learned coefficients: z = b0 + b1x1 + b2x2 + ... + bn*xn

where:

b0, b1, b2, ..., bn are the learned coefficients x1, x2, ..., xn are the input features

Flow Diagram

This flowchart represents the process

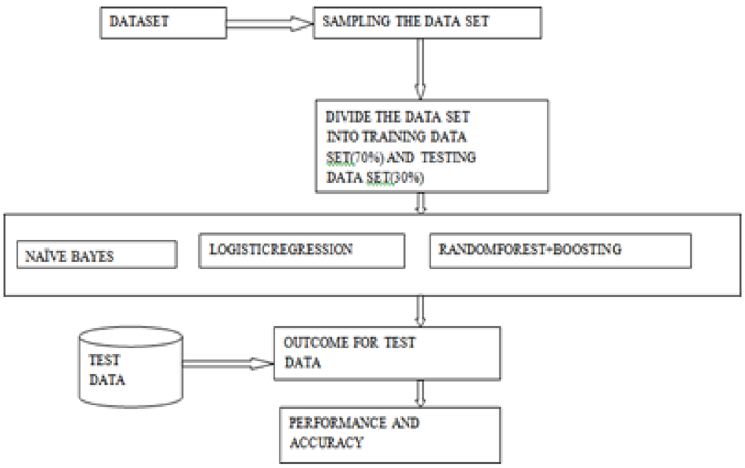

Fig. 1. Flowchat showing working of the model

IV. Comparision between logistic regression and decision tree

Both decision trees and logistic regression are commonly used algorithms for credit card fraud detection systems. But decision tree is Decision trees can be used in credit card fraud detection by analyzing historical transaction data and identifying patterns and rules that can distinguish between fraudulent and non-fraudulent transactions.

The decision tree algorithm works by recursively splitting the data into subsets based on the values of different features, with the goal of maximizing the difference in the outcome variable (fraudulent or non-fraudulent) between the resulting subsets. This process continues until a stopping criterion is met, such as a minimum number of samples in a subset or a maximum depth of the tree.In the context of credit card fraud detection, decision trees can be trained on a dataset of historical transactions that are labeled as fraudulent or non-fraudulent. The decision tree will then use the features of a new transaction (such as the transaction amount, location, and time) to classify it as either fraudulent or non-fraudulent. However, decision trees can be prone to overfitting if the tree is too complex or if there are too many features in the dataset. On fresh, unused data, this could result in subpar performance.. Therefore, it is important to use techniques such as pruning or ensemble methods (such as random forests) to improve the generalization ability of the model.

- Accuracy For Logistic Regression
  - precision: 58.82
  - recall: 91.84

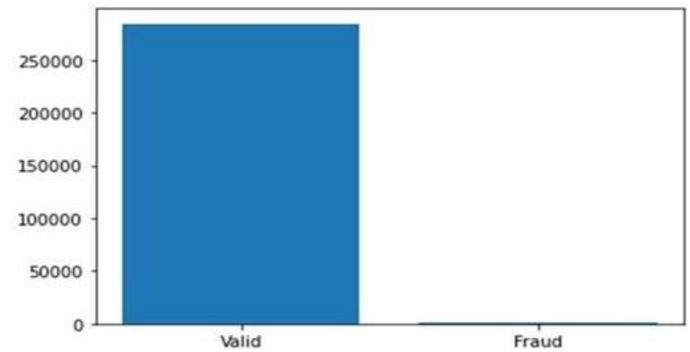

Fig. 2. Histogram showing valid and fraudulent results.

  - accuracy: 97.46
- Accuracy For Decision Tree
  - precision: 16.17
  - recall: 82.65
  - accuracy: 99.23

V. Result

Importing the Libraries In this, many libraries will be used for various purposes. The data will be loaded into a DataFrame object using the panda’s package, making it simpler to deal with as in Fig 3. For plotting purposes, the matplotlib and seaborn libraries will be utilized. While some of the data processing, model development, and model evaluation will be done using the sklearn library.

1) Performing Exploratory Data: Credit Card Fraud Detection Data contains the dataset that will be utilised for machine learning-based credit card fraud detection. The dataset has three types of features: amount, time, and predictors V1 through V28. Due to the likelihood that the columns V1 through V28 contain sensitive credit card information, they have been scaled and anonymised. The predicted column is the class column, where 0 denotes a legitimate transaction and 1 denotes a fraudulent one.

It is important to comprehend the data we are working with before using it to train the machine learning model. Finding the dataset’s shape, or the number of rows and columns, identifying the data sources, and performing an exploratory data analysis are all common tasks that are included in this step types of data items in each column, the presence or absence of values, correlation coefficients, etc.

With fraudulent transactions making up just 0.17 percent of all transactions, the dataset appears to be seriously out of balance. Unbalanced datasets should be treated carefully since they may introduce bias into machine learning models. Knowing how strongly the variables in our dataset are correlated is important information. This knowledge can be useful when deciding which machine or features to extract choosing a learning model. The correlation matrix can be plotted to give a visual representation of the correlation coefficients between the features and the result. There are no strong correlations between the predictor columns, according to the heat map above. No predictor column’s association with the

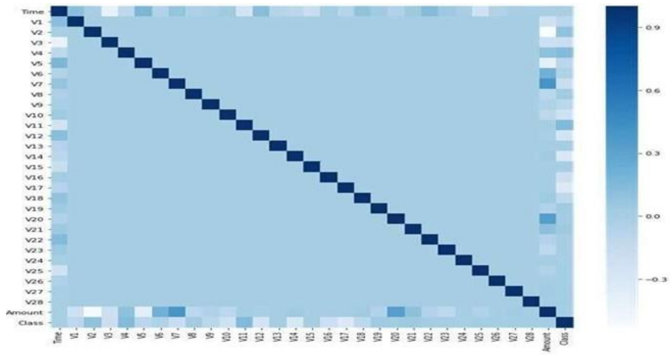

Fig. 3. Waterfall chart showing the number of changes.

Class column is very high. However, V2 and Amount have a negative association, whereas V7 and the Amount feature have a positive correlation.

2) Data Pre-Processing: Sjetting up the dataset for the machine learning model to be trained is known as data pre-processing. The key data pre-processing stage should change the data so that the machine learning algorithm of choice can process it. For instance, most classification algorithms won't be able to comprehend the text in the data, hence not preprocessing the data would result in mistakes. Imputing or removing records with missing values, label encoding categorical data, one hot encoding labeled data, scaling the data, and running train-test splits on the dataset are all common data pre-processing techniques. The majority of data pre-processing processes are not required because this dataset doesn't have any missing values or categorical data. First, the data is divided into the predictors (X) and the outcomes (Y). While Y has 284807 data records with one column and class, and X has 284807 data records with 30 features apiece. The dataset is split into a training set and a testing set using the train-test method. The testing set is used to evaluate the machine learning model after it has been trained using the training set. The 0.2 test size shows that 20 percent of the dataset has been chosen as the testing set. As a result, although the testing set has 56962 records, the training set has 227845 records.

3) Classification Model: Depending on the kind of task that needs to be completed, a machine-learning model should be chosen. Machine learning is capable of carrying out numerous tasks, including pattern extraction, grouping, regression, classification, and more. There are various algorithms that might be available for each assignment. Typically, two or more algorithms are tested to determine which one best fits the data and produce findings that are more reliable and accurate. As a credit card transaction must be classified as either valid or fraudulent, the problem of detecting credit card fraud is one of classification. As previously indicated, a variety of classification methods are available, including Decision Trees, Linear Classifiers, Naive Bayes Classifiers, Support Vector Machines, and Nearest Neighbour Classifiers. To solve this issue, a Random forest and The Decision Tree classifier is extended by the Forest Classifier, which is implemented

4) Model Evaluation: Each machine learning model needs to be judged according to the task that it completes. In order to evaluate a model, one must ask it to forecast values for records of data that have not yet been seen. This has already been completed and is saved in Y_pred. The values that must be compared to the genuine values, or Y_test, are those that the model anticipated, or Y_pred. Since this is a classification problem, metrics like accuracy, precision, and recall can be used to assess the model. How many data records the model correctly predicted values for is determined by the model's accuracy. This model's accuracy value is 0.9996, meaning that it correctly predicted outcomes 99.96 percent of the time. Precision proves that all of those records are accurate things were anticipated to be favorable. With a precision rating of 0.963, the model correctly predicted favorable outcomes 96.3 percent of the time. Last but not least, remembering a model shows how many values were truly constructive successfully identified. A 0.7959 recall value means that 79.59 percent from any wholesome ideals were correctly identified by model. On something called the confusion matrix, a heat map can be used to visualize the measurements stated above. The numbers of expectations between each class's true and expected values are shown in a confusion matrix.

VI. Conclusion

With an accuracy rate of 99.6 percent our algorithmic classifier was able to categorize a legitimacy of credit card purchases. For businesses, credit card fraud is a serious issue. These frauds may cause large losses in both the corporate and personal spheres. As a result, corporations spend an increasing amount of money on developing novel ideas and strategies that can help identify and avert fraud. The main objective of this article was to contrast various machine learning techniques for transaction fraud detection. Comparisons were made as a result, and it was discovered that the decision tree regression method delivers the best results, i.e., best classifies whether transactions are fraudulent or not. This was determined using a number of parameters, including as recall, accuracy, and precision. remember with this kind of issue, a high value is essential. It is now obvious that selecting the appropriate characteristics and balancing the dataset is essential for achieving noteworthy results. In order to better service outcomes, future studies should prioritise a wide range of Machine learning methods include a wide range of features, stacked classifiers of different kinds, and genetic algorithms.

References

[1] A. Thennakoon, C. Bhagyani, S. Premadasa, S. Mihiranga and N. Kuruwitaarachchi, "Real-time Credit Card Fraud Detection Using Machine Learning," 2019 9th International Conference on Cloud Computing, Data Science Engineering (Confluence), Noida, India, 2019, pp. 488-493, doi: 10.1109/CONFLUENCE.2019.8776942

[2] Satvik Vats, Surya Kant Dubey, Naveen Kumar Pandey, A Tool for Effective Detection of Fraud in Credit Card System, published in International Journal of Communication Network Security ISSN: 2231 1882, Volume-2, Issue-1, 2013

[3] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and G. N. Surname, Random forest for credit card fraud detection, IEEE 15th International Conference on Networking, Sensing and Control (ICNSC),2018..
[4] M. Hamdi Ozcelik, Ekrem Duman, Mine Isik, Tugba Cevik, Improving a credit card fraud detection system using genetic algorithm, published by International conference on Networking and information technology, 2010.
[5] Wen-Fang YU, Na Wang, Research on Credit Card Fraud Detection Model Based on Distance Sum, published by IEEE International Joint Conference on Artificial Intelligence, 2009
[6] Andreas L. Prodromidis and Salvatore J. Stolfo; "Agent-Based Distributed Learning Applied to Fraud Detection"; Department of Computer Science- Columbia University; 2000.
[7] Salvatore J. Stolfo, Wei Fan, Wenke Lee and Andreas L. Prodromidis; "Cost-based Modeling for Fraud and Intrusion Detection: Results from the JAM Project"; 0-7695-0490-6/99, 1999 IEEE
[8] . Soltani, N., Akbari, M.K., SargolzaeiJavan, M., A new user-based model for credit card fraud detection based on artificial immune system, Artificial Intelligence and Signal Processing (AISP), 2012 16th CSI International Symposium on., IEEE, pp. 029-033, 2012.
[9] S. Ghosh and D. L. Reilly, Credit card fraud detection with a neural- network, Proceedings of the 27th Annual Conference on System Science, Volume 3: Information Systems: DSS/ Knowledge Based Systems, pages 621-630, 1994. IEEE Computer Society Press.
[10] MasoumehZareapoor, Seeja.K.R, M.Afshar.Alam, Analysis of Credit Card Fraud Detection Techniques: based on Certain Design Criteria, International Journal of Computer Applications (0975 8887) Volume 52 No.3, 2012.
[11] Fraud Brief AVS and CVM, Clear Commerce Corporation, 2003, http://www.clearcommerce.com.
[12] All points protection: One sure strategy to control fraud, Fair Isaac, http://www.fairisaac.com, 2007. [13] Clear Commerce fraud prevention guide, Clear Commerce Corporation, 2002, http://www.clearcommerce.com
[13] Samaneh Sorournejad, Zahra Zojaji , Reza Ebrahimi Atani , Amir Hassan Monadjemi, A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective , IEEE 2016
[14] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and G. N. Surname, Random forest for credit card fraud detection, IEEE 15th International Conference on Networking, Sensing and Control (ICNSC),2018.
[15] T. Fawcett and F. Provost, "Adaptive fraud detection", Data Mining and Knowledge Discovery Journal, Kluwer Academic Publishers, Vol. 1, No. 3, 1997, pp. 291-316
[16] Y. Wang, H. Yang, X. Wang, and R. Zhang, "Distributed Intrusion detections Based on data fusion method.", in P roceedings of the 5th World Congress on Intelligent Control and Automation, 2004, pp. 4331– 4334.
[17] Yakub K. Saheed, Moshood A. Hambali, Micheal O. Arowolo, Yinusa A. Olasupo, "Application of GA Feature Selection on Naive Bayes Random Forest and SVM for Credit Card Fraud Detection", Decision Aid Sciences and Application (DASA) 2020 International Conference on, pp. 1091-1097, 2020.
[18] Zhang, R.; Zheng, F.; and Min, W. 2018. Sequential Behavioral Data Processing Using Deep Learning and the Markov Transition Field in Online Fraud Detection. FintechKDD
[19] Zhongfang Zhuang, Xiangnan Kong, E. R. J. Z. A. A. 2019. Attributed Sequence Embedding. arXiv:1911.00949.
[20] Bello-Orgaz, G., Jung, J. J., Camacho, D. (2016). Social big data: Recent achievements and new challenges. Information Fusion. 28 (Mar. 2016), 45–59