Credit Card Fraud Detection System

Ankush Singh Bisht
AIT-CSE
Chandigarh University
India
21BCS6681@cuchd.in

Kunal Wadhwa
AIT-CSE
Chandigarh University
India
21BCS6631@cuchd.in

Prabjot Singh Bali
AIT-CSE
Chandigarh University
India
prabjot.e16592@cumail.in

***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 be aware of the fraud before any activity by the fraudster. The model can also detect whether a new transaction is fraudulent. Our goal here will be to provide a model to detect 100 percent thereby minimizing 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 many methods that can be used to categorize transactions as real or fraudulent. The research made use of the Credit Card Fraud Detection dataset. T h e SMOTE technique was employed for oversampling due to the dataset's extreme imbalance. Also, the dataset was split into training data and test data after feature selection and dataset splitting. Naive Bayes, Multilayer Perceptron, Logistic Regression, and Random Forest were the algorithms employed in the experiment. The results demonstrate how accurate each algorithm is in identifying credit card fraud. Additional anomalies may be found using the proposed model.**

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.

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 putting in place many 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 into 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 function. Nevertheless, 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 2 sorts of credit card fraud. The 1st is stealing the actual card, while the second entails collecting private data from the card, like CVV code, card number, card type, as well as others. Before the victim is informed, a fraudster could steal a substantial amount of money or use stolen credit card information to make pricey transactions. Businesses use many Machine-Learning (ML) approaches to differentiate between legitimate and fraudulent transactions. This study will evaluate different ML techniques, like MLP ("Multilayer Perceptron"), NB ("Naive Bayes"), RF ("Random Forest"), and LR ("Logistic Regression") to examine which ML algorithm is most effective for identifying credit card fraud. ing, utility costs, dining out, electronics, etc.

II. LITERATURE SURVEY

The substantial loss that fraudulent activities are causing has motivated investigators to create a technique to recognize and stop fraud. Numerous tactics have already been proposed and investigated. [1] This is a summary of some of them. It has been shown that traditional techniques like RF, LR, DT (Decision Trees), Support Vector Machines (SVM), and GB (Gradient Boosting), are successful. Using a European dataset, the study's use of SVM, LR, GB, and RD, a combination of specialized classifiers yielded a high recall of over 91%. High accuracy and recall were only reached once the dataset was balanced by under-sampling.[2] In the publication, the European dataset was also employed, and models on the basis of RF, DT, and LR were compared.

RF turned out to be the best model out of the three. KNN ("k-Nearest neighbors") and outlier detection algorithms, according to them, could also be effective in detecting fraud detection. They were indicated to be effective in reducing false alarm rates and raising the fraud detection rate. In an experiment for their publication [3], the authors examined and compared KNN with other traditional methods, and found that it performed well. Unlike the previously listed publications, the study compared certain traditional methods and used machine learning methods. Approximately 80 percent of the methods under examination were found to be accurate. The paper's authors evaluated the efficacy of the following algorithms using a European dataset: MLP, XGBoost (XGB), NB, KNN, DT, SVM, LR, GB, and RF stacking classifiers (a mixture of many ML classifiers). In-depth data preparation enabled all algorithms to achieve a high accuracy of over 90 percent. The most efficient classifier, [4] with 95 percent recall, and 95 percent accuracy was stacking. In the article, a NN ("Neural Network") was evaluated with a European dataset. The test included back propagation NN optimization using the Whale method. The NN consisted of 2 output layers, 2 input layers, and 20 hidden layers. Due to the optimization process, they were able to get impressive results with 500 test samples: 97.83 percent recall and 96.40 percent accuracy. The authors of the papers employed neural networks to show how using ensemble approaches can improve results. [5] Auto-encoder and Restricted Boltzmann Machine techniques were compared using three datasets in the paper, and the results showed that algorithms like MLP can be ML fraud detection was the subject of many papers. 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—can be obtained with fewer resources? Our major objective is to demonstrate that various ML methods may produce respectable results with the right preprocessing. While the majority of the authors of the papers mentioned employed the under-sampling technique, utilizing the oversampling technique was a distinct strategy. [8] A comparison of the applicability of RF, LR, NB, and MLP for credit card fraud detection was made by the authors of this work in light of the available data. An experiment was conducted to achieve that [9]. Credit card fraud is a substantial issue and comes at a significant cost to banks and card issuer businesses, according to the fraud detection system that has been proposed. To detect and prevent credit card fraud, banks have highly complex security systems in place to monitor transactions and identify fraudulent activity since credit card transactions are so complicated. One time it is devoted; it must be completed as soon as feasible. The objective of this study is to choose a few cutting-edge fraud detection techniques while achieving a comprehensive assessment of various approaches [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 type of autoencoders and other NN systems—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 primary goals of

this work is to compare features created to those that were manually built using domain expertise. In this paper, 2 methods—DT and LR Algorithm—are used to design [14] algorithmic approaches for scam detection /credit card fraud. As a result, an attempt was made to identify and address the issues of fraud in the
credit card business system. [15]

III. Proposed Methodology

In this investigation, the dataset of "Credit Card Fraud Detection" from Kaggle was utilized. The two-day transactions made by cards throughout Europe in September 2013 are included in this dataset. [16] Thirteen numerical characteristics are included in the dataset. Given that some input variables include financial data, the PCA The secrecy of the data was maintained by altering these input variables. Not one of the three converted the designated characteristics. The time elapsed between the dataset's 1st transaction and every subsequent transaction is shown via the "Time" feature. The term "Amount" refers to a feature that shows the total amount 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% 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 accuracy and precision 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 to sigmoid function may be applied to categorize the O/P that is a dependent feature and its usage probability to do so. Because the sigmoid function is utilized, this approach performs well with small data sets. If the output value of the sigmoid function is higher than 0.5, the result will be 1, and if it is below 0.5, the result will be 0. Nevertheless, this sigmoid function is not appropriate for DL since, in deep learning, we must update the weights when going back from the output to the input to decrease the error in the weight update. We must. If the intermediate layer neuron's differentiation of the sigmoid activation function yields a 0.25 value, this will have an impact on the accuracy of the module in DL.
2) Decision Tree: It may be utilized to address the classification & regression problems. Though certain formulae may differ, both use the same working method. The DT model in the classification problem is developed using entropy and information gain. Entropy indicates the degree of randomness in the input, while information gain indicates the amount of information we can glean from this attribute. The Gini and Gini indices are used to construct the decision tree model for the regression issue. When solving classification issues, the root node is selected on the basis of 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 every DT model inside the RF, 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 ML algorithms. This algorithm is utilized to answer the problem in most of the Kaggle computation challenges.
4) Naive Bayes: The Bayes theorem feature is the foundation of the Naive Bayes ML technique for the classification issue. This calculation of the dependent feature's probability in relation to the independent features makes use of the Naive Bayes theorem. It may be used by using features in data sets where the dependent features are the output and the independent features are the input. 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: - $p = 1 / (1 + e^{(-z)})$ Here: p represents the predicted probability of fraud, e signifies the mathematical constant nearly equivalent to 2.71828, and z denotes the "linear combination" of the input features, weighted by learned coefficients: $z = b0 + b1x1 + b2x2 + ... +$
$bn*xn$ here:
b0, b1, b2, ..., bn indicates the learned coefficients x1, x2,
..., xn are the input features
Flow Diagram
This flowchart represents the process

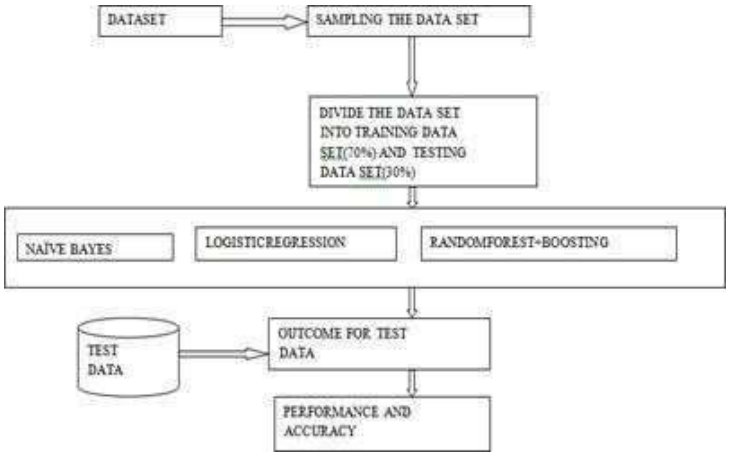

Fig. 1. Flow chart showing the working of the model

IV. Comparison between logistic regression and decision tree

Both decision trees and logistic regression are commonly used approaches for credit card fraud detection systems. Decision tree is Decision trees could be utilized in credit card fraud detection by analyzing historical transaction data and identifying patterns and rules that can differentiate between fraudulent as well as non-fraudulent transactions. The decision tree method maximizes the difference in the outcome variable (fraudulent or non-fraudulent) between the generated subsets by recursively dividing the data into subsets depending on the values of several attributes. This procedure keeps on until a stopping criterion—like the maximum depth of the tree or the minimum number of samples in a subset—is satisfied. Decision trees may be trained on a dataset of past transactions that have been classified as fraudulent or nonfraudulent in the context of credit card fraud detection. Next, the decision tree will categorize a new transaction as either fraudulent or non-fraudulent based on its attributes (like the transaction amount, location, and time). On the other hand, if a decision tree is very complicated or the dataset has an excessive number of characteristics, overfitting may occur. On fresh, unused data, this could result in subpar performance. Consequently, it is crucial to use strategies like pruning or ensemble approaches (like random forests) to increase the model's capacity for generalization.

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

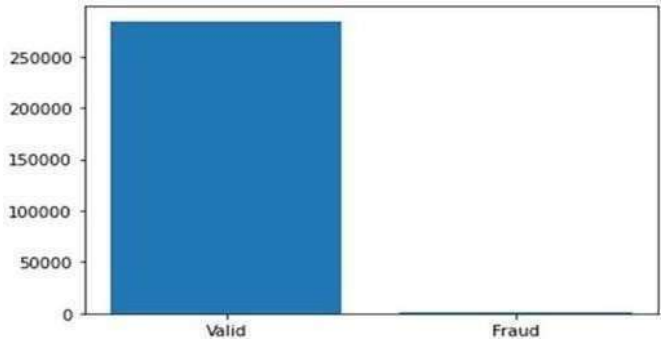

Fig. 2. Histogram showing valid and fraudulent results.

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

V. Result

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

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

It is important to understand the data we are working with before using it to train the ML model. Finding the dataset's shape, or the number of columns and rows, detecting 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 within our dataset are correlated is important information. This knowledge can be useful when deciding which machine or features to extract when choosing a learning model. The correlation matrix can be plotted to give a visual representation of the correlation coefficients between the features and results. There are no strong correlations between the predictor columns, according to the heat map above. No predictor column's association with the

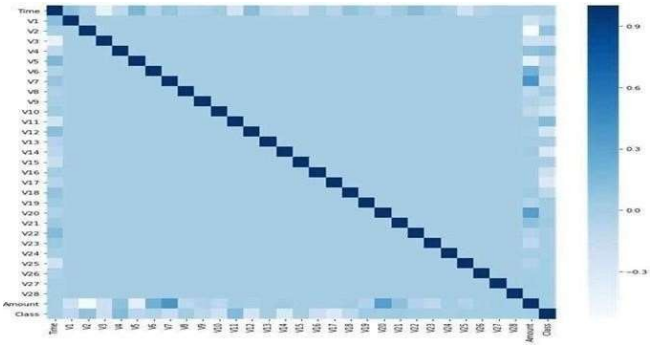

Fig. 3. Waterfall chart showing the number of changes.

The class column is very high. However, there is a negative link between V2 and Amount, whereas V7 and the Amount feature exhibit a positive correlation.

2) Data Pre-Processing: Setting up the dataset for the ML 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, the majority of classification algorithms will be unable to decipher the language inside the data; hence, failure to preprocess the data will lead to errors. Standard data preprocessing techniques include the following: approximating or removing records that have missing values, adjusting the data scale, one-hot encoding labelled data, label encoding categorical data, and doing train-test splits. The majority of data pre-processing processes are not required because this dataset doesn't have any categorical data or missing values. The predictors (X) and outcomes (Y) are first separated from the data. As opposed to X, which comprises 284807 data records with 30 features every, Y has 284807 data records with one column and class. Using the train-test approach, the dataset is split into a training set and a testing set. The ML model is trained on the training set, and then it is assessed on the testing set. According to the 0.2 test size calculation, 20% of the dataset has been selected 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 ML model should be chosen. Machine learning is capable of carrying out numerous tasks, including pattern extraction, grouping, regression, classification, and more. Various algorithms 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 issue of detecting credit card fraud is one of classification. As previously indicated, a variety of classification methods are available, including DTs, Linear Classifiers, NB Classifiers, SVMs, 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 ML model should be assessed based on the job that it completes. A model has to be asked to predict values for records of data that have not yet been observed to be evaluated. This is stored in Y_pred and has already been finished. Values that the model predicted, or Y_pred, are the ones that need to be compared to the actual values, or Y_test. Metrics such as accuracy, precision, and recall may be used to evaluate the model since this is a classification issue. The model accuracy determines the number of data records for which values were accurately predicted. With an accuracy rating of 0.9996, this model predicted results accurately 99.96 percent of the time. Precision demonstrates that all of those data are true and that good outcomes were expected. The model accurately predicted good outcomes at 96.3 %, with an accuracy rating of 0.963. Lastly, recalling a model demonstrates the number of truly constructive values that were accurately recognized. A 0.7959 recall value means that 79.59 percent of any wholesome ideals were correctly identified by the 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 the 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 the present article was to contrast various ML techniques for transaction fraud detection. Comparisons were made as a result, and it was discovered that the decision tree regression method delivers the best outcomes, such as classifying whether transactions are fraudulent or not. This was revealed using many parameters, like accuracy, recall, and precision. Remember with this type 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. To better service outcomes, future studies should prioritize a wide range of Machine learning methods including 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. Prodro- midis; "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 userbased model for credit card fraud detection based on the 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 Proceedings 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 Transformation 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