Journal of Electrical Engineering and Computer (JEECOM)
Vol. xx, No. xx (20xx), DOI: 10.33650/jeecom.v4i2
p-ISSN: 2715-0410; e-ISSN: 2715-6427

Classification of Credit Card Frauds Detection using machine learning techniques

Rasha Rokan Ismail [1], Farah Hatem Khorsheed [2]
[1] Department of Computer Science, Diyala University President, Diyala, Iraq
[2] Electronic Computer Center, University of Diyala

Article Info

Article history:

Received Juli 24, 2023
Revised Agustus 01, 2023
Accepted Oktober 27, 2023

Keywords:

Fraud Credit
SVM
TREE
Regression

ABSTRACT

Credit card fraud refers to the illegal activities carried out by criminals. In this research paper, we delve into the topic by exploring four different approaches to analyze fraud, namely decision trees, logistic regression, support vector machines, and Random Forests. Our proposed technique encompasses four stages: inputting the dataset, balancing the data through sampling, training classifier models, and detecting fraud. To analyze the data, we utilized two methods: forward stepwise logistic regression analysis (LR) and decision tree analysis (DT), in addition to Random Forest and support vector machine. Based on the outcomes of our analysis, the decision tree algorithm produced the highest AUC and accuracy value, achieving a perfect score of 1. On the other hand, logistic regression yielded the lowest values of 0.33 and 0.2933 for AUC and accuracy, respectively. Moreover, the implementation of forest algorithms resulted in an impressive accuracy rate of 99.5%, which signifies a significant advancement in automating the detection of credit card fraud.

Corresponding Author: Rasha Rokan Ismail
Coresponding Author Name, Rasha Rokan Ismail
Department of Computer Science, Diyala University President, Diyala, Iraq
Email: rasha_rokan@uodiyala.edu.iq

1. INTRODUCTION

The unauthorized and undesired use of a credit score card account through a person who isn't always the account's proprietor is known as credit score card fraud. To placed prevent to this misuse, right preventative measures have to be installed vicinity, and the conduct of such fraudulent movements may be studied so as to limit and save you destiny occurrences. Credit card fraud takes place while someone makes use of some other person's credit score card for non-public advantage at the same time as the cardboard proprietor and issuing government are blind to the transaction. This is an important trouble that necessitates the eye of companies along with device studying and information science, wherein the answer can be automated. Because it's far described through severa functions along with elegance imbalance, this trouble is mainly tough to deal with from an academic approach. Legitimate transactions far surpass the number of fraudulent transactions by a significant margin. Furthermore, the statistical traits of transaction styles extrade through the years. However, those are not the simplest demanding situations that include installing vicinity a real-international fraud detection system. In practice, automated systems carefully review a large number of price requests before deciding which ones to approve. They utilize machine learning algorithms to examine all approved transactions and report any suspicious ones. Professionals look into those proceedings and call cardholders to decide whether or not or now no longer the transaction became authentic. The information became despatched into the automated system, which became used to educate and refine the set of rules through the years so as to enhance fraud detection efficacy [1].

In 2004, Germany saw a rise in the issuance of playing cards, including debit cards, credit cards and payment cards. The total number of playing cards issued reached a staggering 120 million, which was an increase of nearly 4% compared to the previous year. Additionally, the overall value of card transactions in Germany reached €375 billion in 2004, marking an almost 4% increase from the previous year. This included transactions such as cash withdrawals. The increased usage of cards for payments has had a significant impact on spending habits. The amount spent on purchases and online transactions using any type of card has risen by 5%, reaching a total of €170 billion. Cash withdrawals, on the alternative hand, accelerated greater slowly. If clients transfer from coins to card payments, the ones new styles in purchaser charge conduct are probable to be linked. When it involves credit score playing cards, in Germany, the term "Kreditkarte" refers to each fee and credit score playing cards. Unlike English, wherein distinct gadgets have their very own nomenclature, there may be no seen difference among the 2. Credit card companies have provided customers with the option to revolve their credit by using credit cards, in order to differentiate between debit cards and credit cards. This kind of provider or credit score is probably applied to tempt them withinside the future. Customers have get right of entry to to revolving credit score, however now no longer all people takes use of it. Credit playing cards, on the alternative hand, rose at a quicker charge than fee playing cards in 2004 [2].

Fraud is defined as a deception this is unlawful or crook in nature and is perpetrated for financial or non-public benefit. It is a planned act done in violation of a law, rule, or coverage so one can obtain a bootleg financial benefit. In this topic, a extraordinary wide variety of articles on anomaly or fraud detection had been posted and are to be had to the overall audience. According to a complete analysis, statistics mining applications, automatic fraud detection, and antagonistic detection are many of the processes hired on this subject. A new work, [3], become posted. [4] To come across credit score card fraud, techniques which include Supervised and Unsupervised Learning had been developed. These strategies and algorithms, in spite of their sudden efficacy in a few areas, didn't offer a long-time period and constant solution to fraud detection. In a recent experiment conducted by Wen Fang YU and Na Wang [5], they utilized outlier mining, outlier detection mining and distance sum algorithms to identify fraudulent transactions within a specific commercial bank's credit card transaction dataset. Outlier mining is a data mining technique commonly employed in the financial and internet industries to uncover anomalies or irregularities, such as instances of fraud. Various methods like supervised and unsupervised learning have been developed to combat credit card fraud. However, despite their effectiveness in certain areas, these approaches and algorithms have not provided a consistent long term solution for detecting fraud. Na Wang and Wen Fang YU [5] also explored the usage of emulation in their experiment, employing outlier detection mining and distance sum algorithms to effectively identify fraudulent transactions within the given dataset from an industrial bank. Outlier mining serves as a valuable tool in data analysis within the financial and internet domains. Its motive is to music out matters which have grow to be indifferent from the primary system, which include fraudulent transactions.

On the subject of detecting fraud in credit card transactions, several studies have been conducted. With the use of sampling strategies and incremental learning, Dal Pozzolo A et al. performed final prediction using an ensemble of those models [10]. They achieved excellent outcomes by utilizing the Random Forest classifier combined with the Synthetic Minority Oversampling approach (SMOTE) sampling technique. In their comparison of several decision tree splitting criteria, Najadat H et al. developed the Hellinger distance [11], a novel skew measure. They suggested employing Hellinger distance to create decision trees for improved performance. The two sampling strategies were given additional insight by Drummond C and Holte RC [12]. The instructions suggest that we need to achieve a balance in the data to effectively adjust the machine learning algorithms for situations where the data is unbalanced. The researchers discovered that under sampling yielded better results compared to oversampling. They found that using a random forest classifier and data balance improved prediction accuracy. For accuracy and the detection of fraudulent transactions, Kumar MS et al. employed random forest [13]. They used a confusion matrix to analyze the performance, and they found that 90% of the time it was accurate. To identify fraudulent transactions, Sadineni PK worked with a variety of machine learning approaches [14In the study conducted by Sailusha R et al., they found that different models were used to identify fraudulent transactions. The random forest model achieved an accuracy rate of 99.21%, the decision tree model had a rate of 98.47%, the logistic regression model achieved 95.55% accuracy, while the support vector machine model had an accuracy of 95.16%. The artificial neural network model performed exceptionally well with an impressive accuracy rate of 99.92%. used the Adaboost algorithm and random forest [15]. Accuracy and the F1-score were used to evaluate performance. They achieved 99.9% accuracy for Adaboost and 100% accuracy for Random Forest. To address the imbalance in the data, we sampled our datasets in our research using three distinct sampling approaches. The dataset with an equal distribution of samples was used to train four distinct classifier models. To validate the accuracy of these models, a publicly available dataset was utilized. In terms of algorithmic approaches, [16] employed a technique that involved a combination of visual tuning and Bayesian based optimization for hyper parameters. This was made possible by merging data from two different public databases—one containing fraudulent transactions and the other consisting of real world valid transactions. Their proposed method outperformed previous approaches in terms of Accuracy, Precision and F1 Score. Referring to a study conducted by [16], they developed an effective learning algorithm to detect credit card fraud considering the relatively high ratio of fraudulent to genuine transactions. The study found that random forests are more successful than neural networks in identifying fraud incidents. Another factor considered was the inclusion of large credit card transactions. Ensemble learning, which combines various machine learning techniques like neural networks and random forests, was employed. The research conducted by [16] highlighted the increasing prevalence of credit card fraud in recent years. To combat this issue, several machine learning algorithms have been utilized for detecting fraudulent transactions and preventing their completion. Two novel data driven strategies were proposed based on the best anomaly detection methods for credit card fraud detection. Selecting appropriate kernel settings and employing a T2 control chart. In this paper, we discuss four different approaches for analyzing fraud; decision trees, logistic regression, support vector machines and Random Forests.

2. METHOD

This is a retrospective study that was carried out with the use of a database of 869 people who lived in the years 2018 and 2020.

We analyzed the data using two different methods, forward step wise logistic regression analysis (LR) and decision tree analysis (DT), as well as Random Forest and support vector machine. For each approach, we calculated the ROC. The dataset was divided into two groups; one for creating the model (70 percent) and the other for validation (the remaining 30 percent). Our suggested technique consists of four stages; inputting the dataset, balancing the data through sampling, training classifier models and detecting fraud. Figure 1 provides an overview of this process.

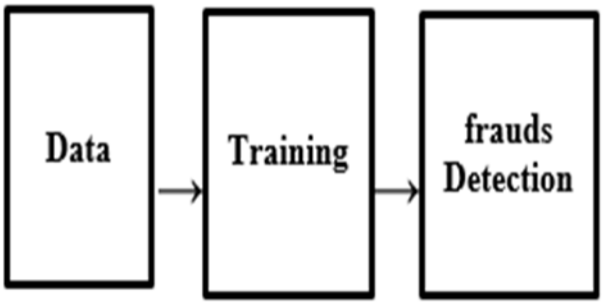

Figure 1. Proposed Approach

3. DATA ANALYSIS

This study proposes a method that takes advantage of the latest advancements in machine learning to identify outliers or unusual activities. According to reference [23], machine learning (ML) is employed to teach computers how to handle data more efficiently. By analyzing the gathered data, ML can extract valuable insights. The demand for ML has significantly increased with the availability of recent data. Machine learning encompasses both supervised and unsupervised learning techniques. While unsupervised learning focuses on discovering internal patterns in input data, supervised learning utilizes known input and outcome data to train models for predicting future outputs.

ISSN: 2715-6427

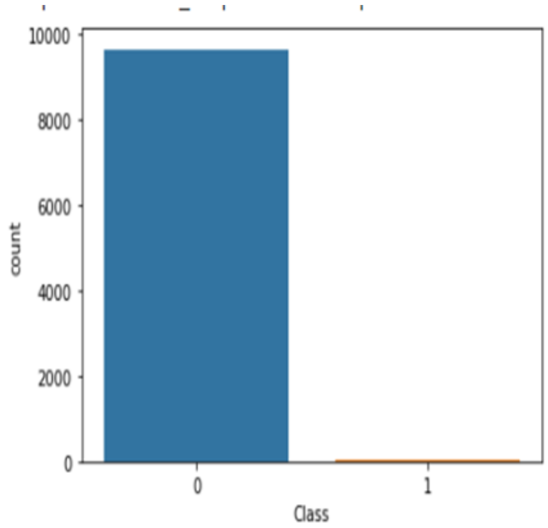

Figure 2. It is evident that the count of fraudulent transactions is significantly lower in comparison to the number of legitimate ones

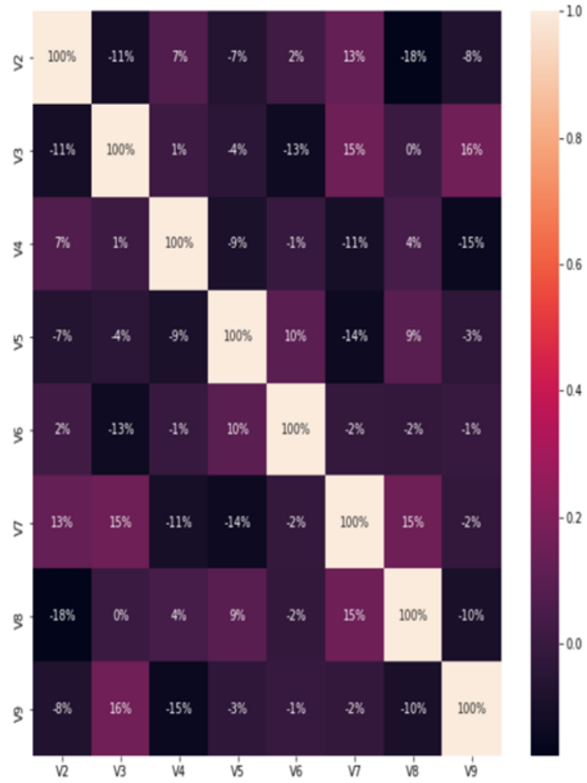

Figure 3. Fraud correlation matrix

Every feature variable in the data was used during the training phase, just like in figure (3). The anticipated occurrences are listed and the results of the confusion matrix are revealed. The "True-True" field contains the vast majority of estimated transactions, as can be seen. Logistic Regression. Figure 4 shows a sigmoid function that represents it. To make predictions for the output value, we combine the input values in a linear manner using weights or coefficients. This represents a binary variable that can only take on two possible values; 0 or 1. The goal is to create a mathematical equation that gives us a score between 0 and 1.

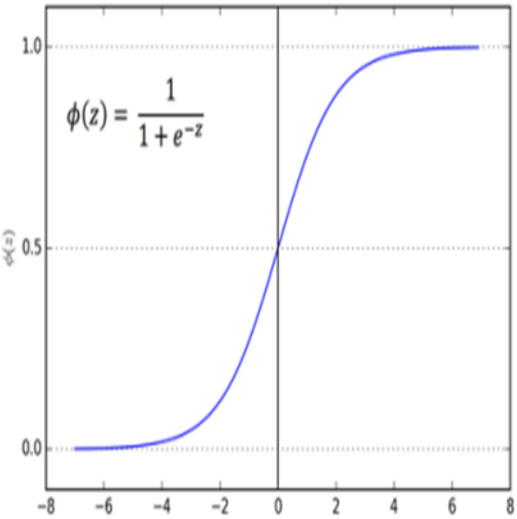

Figure 4. Logistic Regression [19].

SVMs are a well-known data modeling approach. They combine generalization control with a solution to the dimensionality curse. SVM finds the separation hyperplane with the greatest feasible margin between two classes. The hyperplane is situated halfway between the two classes to optimize the margin [6.7]. As an added benefit, the SVM approach identifies support vectors. Support vectors are data points that are placed on or near the margin and are entirely responsible for the solution. If the support vectors were simply fed into the SVM algorithm as data points, the same separation hyperplane would be generated. The data is not necessarily linearly separable in some circumstances. Nonlinearities in the underlying system that generates the data points, as well as noise in the observations, might be at fault. Even if a "few" data points are misclassified and fall inside the margin, the SVM algorithm can still find the largest margin separating hyperplane. Non-linear curves can also be used by the SVM to differentiate data classes. This is accomplished by nonlinearly translating the input data into a new space where the data may still be separated by a linear hyperplane. The curve of the dividing hyperplane becomes visible when it is mapped back to the original space (refer to Figure 5). When it comes to detecting fraud, there are several commonly used technologies. These include rule induction approaches, decision trees, neural networks, Support Vector Machines (SVM), logistic regression and meta heuristics like genetic algorithms, k means clustering and nearest neighbor algorithms. These techniques can be employed individually or combined together to create classifiers using ensemble or meta learning methods. In previous studies [24 29], well known decision tree techniques such as ID3, C4.5 and C&RT were utilized to detect credit card fraud.

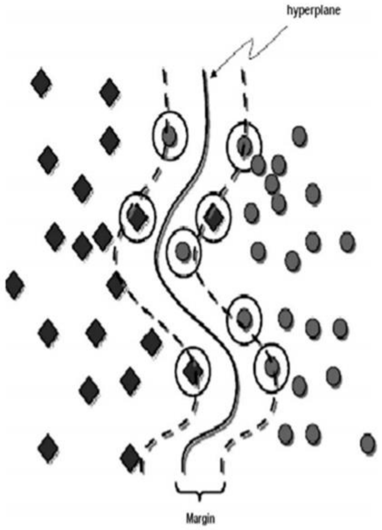

Figure 5. Classification of data using non-linear curves.

Random forests, also known as random selection forests, are a machine learning technique that uses a collection of decision trees to perform classification, regression and other tasks. These trees are trained and then combined to produce an overall prediction that represents the average or common outcome. The purpose of random selection forests is to address the issue of decision trees overfitting their training data. Random forests outperform preference bushes in maximum cases, however they're much less specific than gradient improved bushes. Data features, on the alternative hand, may also have an effect on their performance. [8,9]. The BCR prediction gadget became written in Python to provide prediction fashions for 4 exclusive set of rules types.

4. RESULTS AND DISCUSSION

The results for each algorithm as shown in table 1.

Table 1. Performance of Algorithms.

Analysis	Method	Accuracy	Sensitivity
Decision tree	Decision tree	1	1
Decision tree	Random Forest	0.99	0.95
Support Vector Machi	Linear	0.77	66.8

According to the result of analysis Table 1, the highest AUC and accuracy value for the four algorithm predictions was 1 for the decision tree while the lowest for value is 0. 33 and 0.2933 for logistic regression. And comparing our method of logistic with ref [20] the accuracy is .66 where our accuracy is 0.77 which show better results and comparing random forest with both [21,22] as it's the same tis first and higher than the former.

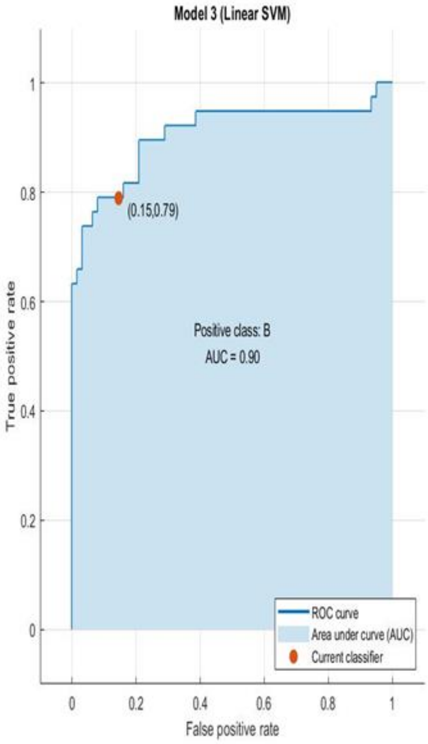

Figure 6. ROC curve for SVM.

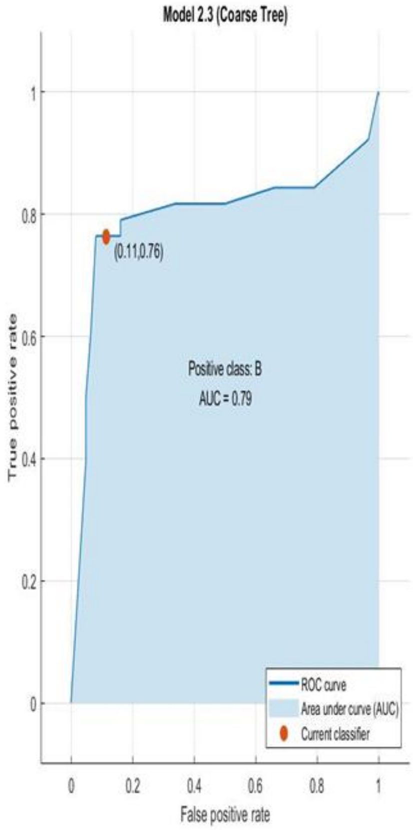

Figure 7. ROC curve for Tree.

A ROC curve, also known as a receiver operating characteristic curve, is a graph that illustrates the false positive rate (FPR) of a classification system. You can see this in figures 6 and 7. The FPR is determined by dividing the total number of test results by the number of false positives. On an ROC plot, the x axis represents the FPR while the y axis represents the true positive rate (TPR). This type of curve helps us assess the accuracy and bias of a classification system, which is valuable for evaluating model performance. As we move further away from the diagonal line on the ROC plot, it indicates higher accuracy for our model. Ideally, we want a classifier with a TPR of 1 and an FPR of 0, which you can observe at the upper left corner of the plot.

5. CONCLUSIONS

Credit card fraud is a type of behavior that criminals engage in. In this article we will delve into the advancements in combatting this problem. Shed light on the most prevalent forms of fraud as well, as techniques to detect them. Furthermore we will provide an explanation of how machine learning can improve the outcomes of fraud detection and discuss the implementation through experimentation. Notably employing forest algorithms yielded an accuracy rate of 99.5% which represents a significant achievement, in automating credit card fraud detection.

ISSN: 2715-6427

ACKNOWLEDGEMENTS

The authors thank all research contributors

REFERENCES

[1] PUMSIRIRAT, Apapan; LIU, Yan. Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine. International Journal of Advanced Computer Science and Applications, 2018, 9(1).

[2] Euromonitor International, 2006. Financial cards in Germany. Available at: http://www.euromonitor.com/Financial_Cards_in_Germany (Accessed: November 2006).

[3] CLIFTON PHUA1, VINCENT LEE1, KATE SMITH1 & ROSS GAYLER2. "A Comprehensive Survey of Data Mining-based Fraud Detection Research" published by School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800, Australia.

[4] "Survey Paper on Credit Card Fraud Detection by Suman", Research Scholar, GJUS&T Hisar HCE, Sonepat, published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 3 Issue 3, March 2014.

[5] K. Sivaraman, R. M. V. Krishnan, B. Sundarraj, and S. Sri Gowthem, “Network failure detection and diagnosis by analyzing syslog and SNS data: Applying big data analysis to network operations,” *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 9 Special Issue 3, pp. 883–887, 2019, doi: 10.35940/ijitee.I3187.0789S319.

[6] A. A. Aljumah, M. K. Siddiqui, and M. G. Ahamad, "Application of classification based data mining technique in diabetes care," Journal of Applied Sciences, vol. 13, no. 3, 2013.

[7] E. I. Georga, D. I. Fotiadis, and V. C. Protopappas, "Glucose prediction in type 1 and type 2 diabetic patients using data-driven techniques." INTECH Open Access Publisher, 2011.

[8] Piryonesi S. Madeh; El-Diraby Tamer E. (2020-06-01). "Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems." Journal of Transportation Engineering, Part B: Pavements, 146(2), 04020022. doi:10.1061/JPEODX.0000175.

[9] Piryonesi, S. Madeh; El-Diraby, Tamer E. (2021-02-01). "Using Machine Learning to Examine the Impact of the Type of Performance Indicator on Flexible Pavement Deterioration Modeling." Journal of Infrastructure Systems, 27(2), 04021005.

[10] Dal Pozzolo A, Caelen O, Le Borgne YA, Waterschoot S and Bontempi G 2014 Learned lessons in credit card fraud detection from a practitioner perspective. Expert Systems with Applications, 41, pp. 4915-28.

[11] Najadat H, Altiti O, Aqouleh AA, and Younes M 2020 Credit card fraud detection based on machine and deep learning. 11th Int. Conf. Information and Communication Systems (IEEE), pp. 204-08.

[12] Drummond C and Holte RC 2003 C4.5, class imbalance, and cost sensitivity: why undersampling beats oversampling. Workshop on learning from imbalanced datasets II (Washington DC: Citeseer), 11, pp. 1-8.

[13] Kumar MS, Soundarya V, Kavitha S, Keerthika ES, and Aswini E 2019 Credit card fraud detection using the random forest algorithm. 3rd Int. Conf. Computing and Communications Technologies (IEEE), pp. 149-53.

[14] Sadineni PK 2020 Detection of fraudulent transactions in credit cards using machine learning algorithms. 4th Int. Conf. IoT in Social, Mobile, Analytics, and Cloud (I-SMAC, IEEE), pp. 659-60.

[15] Sailusha R, Gnaneswar V, Ramesh R, and Rao GR 2020 Credit card fraud detection using machine learning. 4th Int. Conf. Intelligent Computing and Control Systems (IEEE), pp. 1264-70.

[16] Jiang, C.; Song, J.; Liu, G.; Zheng, L.; Luan, W. Credit card fraud detection: A novel approach using an aggregation strategy and feedback mechanism. IEEE Internet of Things J. 2018, 5, 3637–3647.

[17] Kumar, P.; Iqbal, F. Credit card fraud identification using machine learning approaches. In Proceedings of the 2019 1st International conference on innovations in information and communication technology (ICIICT), Chennai, India, 25–26 April 2019; pp. 1–4.

[18] Lamba, H. Credit Card Fraud Detection in Real Time. Ph.D. Thesis, California State University San Marcos, San Marcos, CA, USA, 2020

[19] "Comparative Study on Classic Machine learning Algorithms". Available at: https://towardsdatascience.com/comparative-study-on-classic-machine-learningalgorithms-24f9ff6ab222.

[20] MAHESH, Konduri Praveen; AFROUZ, Shaik Ashar; AREECKAL, Anu Shaju. Detection of fraudulent credit card transactions: A comparative analysis of data sampling and classification techniques. In: Journal of Physics: Conference Series. IOP Publishing, 2022, p. 012072.

[21] Sailusha R, Gnaneswar V, Ramesh R, and Rao GR 2020 Credit card fraud detection using machine learning. 4th Int. Conf. Intelligent Computing and Control Systems (IEEE), pp. 1264-70.

[22] HASAN, Fahim, et al. E-commerce Merchant Fraud Detection using Machine Learning Approach. In: 2022 7th International Conference on Communication and Electronics Systems (ICCES). IEEE, 2022, pp. 1123-1127.

[23] Xue, W.; Zhang, J. Dealing with imbalanced datasets: A resampling method based on the improved SMOTE algorithm. Commun. Stat. Simul. Comput. 2013, 45, 1160–1172.

[24] Chen, R., Chiu, M., Huang, Y., and Chen, L. 2004. Detecting credit card fraud by using a questionnaire-responded transaction model based on SVMs. In Proceedings of IDEAL2004.

[25] Brause, R., Langsdorf, T., and Hepp, M. 1999. Neural data mining for credit card fraud detection. In Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence.

[26] S. J. Stolfo, D. W. Fan, W. Lee, A. L. Prodromidis, and P. K. Chan, "Credit card fraud detection using meta-learning: Issues and initial results," in AAAI Workshop on AI Approaches to Fraud Detection and Risk Management, AAAI Press, Menlo Park, CA, 1997.

[27] S. Stolfo, W. Fan, W. Lee, A. L. Prodromidis, and P. Chan, "Cost-based modeling for fraud and intrusion detection: Results from the JAM Project," in Proceedings of the DARPA Information Survivability Conference and Exposition, IEEE Computer Press, New York, 1999.

[28] A. L. Prodromidis, P. Chan, and S. J. Stolfo, "Meta-learning in distributed data mining systems: issues and approaches," in Advances of Distributed Data Mining, H. Kargupta and P. Chan (Eds.), AAAI Press, 2000.

[29] R.-C. Chen, S.-T. Luo, X. Liang, and V. C. S. Lee, "Personalized approach based on SVM and ANN for detecting credit card fraud," in Proceedings of the IEEE International Conference on Neural Networks and Brain, Beijing, China, 2005.