Fraudulent Credit Card Transactions Classification using Randomized Search CV with XGB Classifier

Mr. Kapil Dev Tripathi
M.Tech. Scholar
ShriRam College of Engineering & Management
Gwalior, India
kapil.dev.tripathi24@gmail.com

Mr.Vikash Singh Rajput
Assistant Professor
ShriRam College of Engineering & Management
Gwalior, India
Vickysngh135@gmail.com

Abstract—**The amount of credit card transactions (CCTs) is increasing exponentially with the rapid development of electronic retail. When the most common type of transaction is shopping online, purchase fraud is growing as well. Fraudulent CCTs regularly cause companies & consumers large financial costs so fraudsters are constantly trying to discover innovative techniques or solutions to fraudulent purchases. The prevention of fraudulent transactions has been an significant element in increasing the usage of online payments. Effective & efficient solutions to fraud identification in CCTs are also required. Fraudulent transactions can take place in various forms and can be categorized as being. This research work has done by Randomized Search CV with XGB Classifier for accurate prediction of fraudulent transactions. The large volume of CCF data is applied for the experiment which is taken from UCI repository. The simulation is done by Jupiter notebook of Python. The significance of proposed model is measured by different performance parameters those are accuracy, precision, recall, F1 score, MCC and ROC. From the results we achieved 83.02% accuracy**.

Keywords—Machine Learning, Fraud, Fraudulent Transactions, Credit Card Fraud,Randomized Search CV, XGB Classifier.

I. INTRODUCTION

The credit card payment industry has expanded rapidly over the course of time with the rise & ubiquitous Internet. Many companies and industries have converted their firms into electronic platforms that give their customers easy to use e-commerce, connectivity and collaboration. This progress is very significant as it credits increased efficiency & profitability, but it still has drawbacks on its own. In relation to this growth, there is greater risk. One of the most significant obstacles for companies on the internet is that the payment method involves no involvement of the card or the cardholder. It stops the retailer from verifying whether the individual making an transaction is the actual cardholder or not. It helps a fraudster to perform anonymously a fraudulent transaction [1].

For financial transactions today, payment systems including online transactions, CCTs & digital payments are becoming more common. The number of fraudulent transactions is also growing as a consequence of these cashless transactions. The expense actions of consumers (users) may be differentiated from prior transaction data by an investigation of the fraud. When any deviation from accessible trends in the spending is found, the transaction can be fraudulent [2]. The stealing or compromising of banking data that result in fraud via e - mail, telecommunications, malware, non-secure security information , social networking & shoulder surfing. Fraudulent transactions may be identified either by classification or by detecting the outlying transactions. The first model is learned from training data for the classification method. Properties are extracted or transformed from raw data when the model is being used [3].

Credit card fraud (CCF) may be described as' an unauthorized person's fraudulent activity for their own benefit as well as the approved cardholder as well as the card issuer, when the transaction is done, are totally unaware of it'[4]. The problem is costly because it is capital or credibility, and so financial firms are searching at a variety of possible methods of stopping fraud. Nevertheless, these security mechanisms may often be created and compromised by fraudsters that utilize technologies. A successful performance of fraud detection was shown through Machine Learning (ML) Techniques. Some new methods for fraud prevention are also being used and used in other countries. The financial sector often uses ML techniques, due to the large number of frauds which impact it each year.[5] To find out how the frauds are being carried out, banks & CC companies use a variety of data mining approaches, such as decision-making tree, rule-based mining, neural network, a dynamic clustering method, a hidden markov model or hybrid approach. All of these approaches are used to assess consumers' usual consumption trend based on their previous behaviors.

Lack of real world evidence because of the importance to data and privacy concerns is one of the main problems shared with researchers on fraud identification. Most authorities investigated bank records through agreements for real-life details. Many tools may be used to produce synthetic data in order to solve this issue. Second concern is dealing with inequality or uneven distribution as there are far less fraudulent transactions relative to legal transactions.Throughout the end, a amount of low incidence data in the data set that produces the synthetic fraudulence of

transactions connected to original data sets are increased using over sampling approaches from the synthetic minority. In [6], cost-based experiments are applied to generate synthetic fraudulent data balance transactions. Data overlapping is another issue, because certain transactions tend to be fraudulent when in reality they are legal transactions. Fraudulent transactions can also appear to be normal transactions.

Objectives of Study

1) The aim is to decrease the financial damages suffered by both merchants ' or issuing banks due to payment fraud.
2) By utilizing predictive analytics to identify the fraud in a real-world CCFD.
3) Our key aim is to find the right approaches and integrate them into our effective CCFD classification model, and distinguish the instances in the dataset.
4) Forecast the successful identification of CCF & process the high dimension data.

Section II includes the required background or related work. The paper is structured as go after. Section III deals with our proposed system concept or algo. Section IV discusses outcomes of the simulation and analyzes. In Segment V, the paper is concluded.

II. LITERATURE SURVEY

A. A. Taha and S. J. Malebary(2020) This paper proposes a smart approach to CCFD with an OLightGBM (LightGBM) engineered light gradient booster. In the proposed solution, an optimization algorithm for the Bayesian-based hyperparameter is intelligently implemented to change te LightGBM parameters. Experiments have been performed using two public CCTs data collection in real world comprising of both permanent and temporary transactions in order to see how effective the OLightGBM plan is for detecting fraud in CCFs. Based on similarities with alternative techniques, the proposed approach exceeded other methods or reached maximum precision (98.40%), recipient region (AUC) (92.88%), accuracy (97.34%), or F1 (56.95%), respectively. The findings are focused on the contrast with certain methods utilizing both data sets.)[7].

D. Prusti and S. K. Rath (2019) Throughout this analysis, specific classification models are recommended throughout order to assess the precision and other output parameters of the fraudulent method by applying ML techniques. To order to objectively determine their efficiency and to test their performance, classification algorithms include K-NN, Extreme Learning Machine (ELM). Selection algos like as Random Forest (RF). We also proposed an ensemble of five algos as a predictive classification model because it provides stronger predictive efficiency [8].

A. Thennakoon et al.(2019) This paper focuses on four big real-world transaction fraud occasions. Each fraud is dealt with using a range of learning models as well as an evaluation selects the best form. This evaluation offers a detailed roadmap for the identification of an ideal fraud-type algo as well as the test is illustrated by a suitable output indicator. The significant area of our project is the identification of fraud via CC of real time. To accomplish so, the predictive analytics of ML models applied and an API framework are used to evaluate if the transaction is genuine or fraudulent. They frequently test a new approach that tackles biased data delivery effectively. The details used in our studies was given in confidential disclosure agreement by a financial institution [9].

G. Goyet al. (2019) A public data collection is included in this article. Using mixed sampling approaches combined, the imbalance issue of the data collection was overcome. Comparative performance evaluations were performed on this data collection. Unlike other research , in addition to normal output metrics the region under curve (AUC), which represents the quality of these data sets, was also used. As it is often crucial that fraud transactions via CC are quickly identified, it also demonstrates how well the various approaches operate. [10].

C. Jiang et al. (2018) Propose a modern four-stage fraud detection system. In order to optimize the actions of a cardholder, we first use the history transaction data of cardholders to separate all cardholder classes into such that participants of same cardholder community have similar transaction behaviors. And we suggest a window sliding technique for aggregating each group's transactions. Last, we derive a series of activities focused on aggregate transactions as well as historical transactions of the cardholder for each cardholder. We then create a set of classifiers focused on all behavioral characteristics within each party. Eventually, we use the on-line classification system to identify scams and a feedback framework is implemented to address the issue of idea drift if a new transaction becomes fraudulent. Our analysis findings indicate that our method is different than other studies. [11].

M. Jeragh and M. AlSulaimi (2018) The lack of fraudulent transaction data or distorted training data in the models for detecting fraud by credit card, as well as the selection of appropriate metric for measuring a model 's efficiency, among others. This paper proposes modern unsupervised learning paradigm focused on mixture of automobile encoder or one class SVM (OSVM). It offers entry to an automotive encoder & produces an entry reconstruction error. This implemented model is contrasted to other methods including the independent usage of OSVM, auto-encoders and a new type focused on a separate design from the mix of OSVM and the Auto encoder. The newly developed model provides a similar correlation with OSVM and better efficiency over all

other models if calculated using the geometric mean (GMean) and F1 results. [12].

S. Dhankhad et al. (2018) We are implementing different supervised algos for learning software to identify fraudulent CCFs using a real-world dataset. In fact, we use such algos to use ensemble training methods to introduce a super-classification. They define the key factors that may contribute to greater precision in the identification of fraudulent CCTs. We also compare and discuss the efficiency of various supervised ML algos against the super classifier that we have introduced in this paper in literature [13].

III. Proposed Methodology

Every year, fraudulent credit card transactions cause large financial damages for firms and consumers, and fraudsters are continually attempting to discover different techniques and methods of transactions fraud. The prevention of fraudulent transactions has been an important factor in the growing usage of online payments. Lack of real world data owing to data sensitivity & privacy problems is one of the main challenges connected with fraud detection researchers. Simple and effective approaches in CCTs for detect fraud are therefore required.

A. Methodology

To overcome the above given problems, we have used XGB classifier with Randomized Search CV. Firstly check the null values in default credit card fraud classification dataset. This dataset is freely available online at UCI machine learning repositories. If there is any null value then remove it. After it scale the data by standard scalar. Data is categorized into 80% for training & 20% for testing. Now these scaled data need to classify. For this purpose, we used XGB classifier with randomized Search CV and done hyper parameter selection. Finally get optimal parameters and train the model.

B. Randomized Search CV

In our proposed, RandomizedSearchCV is used for parameter tunning of XGB classifier. RandomizedSearchCV [14] Implement the fit approach and the forecasting approach like any classifier except that the classifier parameters often used forecast are cross-validated. Not all parameter values are used like compared to GridSearchCV, but a limited number of parameter settings from the defined distributions are sampled. N iter determines the amount of tried function settings.
When all parameters are shown as a number, sampling is carried out without substitution. When a sampling with such a replacement is being used, if at least one parameter is provided as a distribution. Constant distributions with constant parameters are highly recommended.

C. XGB Classifier

XGBoost is a widely used ML algo for organized or tabular results. XGBoost is a gradient-boosted implementation of quick & efficient decision trees designed. XGBoost offers a wrapping class to handle models in the scikit-learn context including classifiers or regressors. This implies that we will use the entire XGBoost platform scikit-learning collection. The classification XGBoost software is called the XGBClassifier. It is a common and powerful open-source implementation of gradually boosting algos of trees [15] &it can be generated and adapted to our training data collection. Gradient improvement is supervised learning algo that aims to predict target variable correctly by integrating predictions of a series of simpler, weaker models..

D. Proposed Algorithm

Input : Default credit card fraud classification dataset
Output : Classification accuracy.

Strategy:
Step:1 Start
Step:2 Collect the large volume of default credit card fraud classification datasetfrom UCI ML repository.
Step:3 Check the null value for all dataset
Step:4 Scale the data by standard scalar
Step:5 Categorize the data 80% for training and 20% for testing
Step:6 RandomizedSearchCV for Hyper Parameter tunning of XGB classifier
Step:7 Achieved optimal parameters
Step:8 Train the model using XGB classifier for training dataset
Step:9 Perform testing for test dataset
Step:10 Obtain confusion matrix and various performance parameters
Step:11 Stop

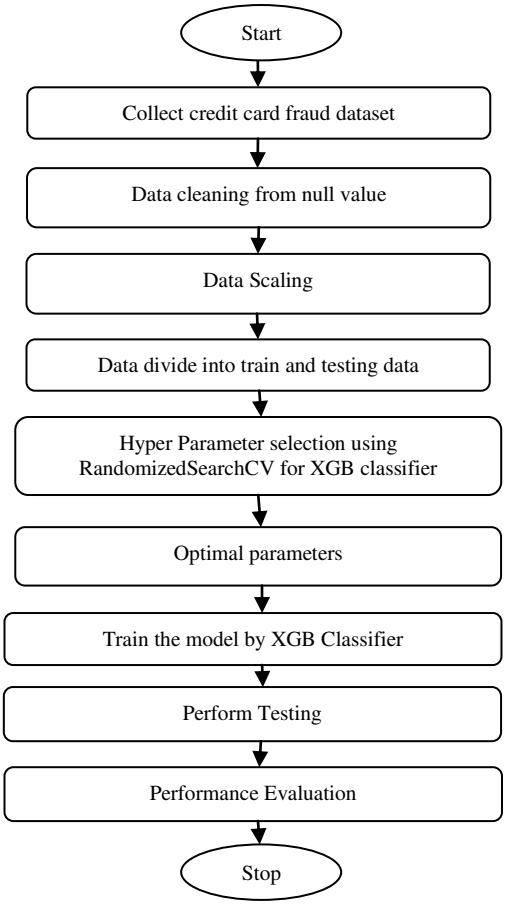

Fig. 1. Flowchart of Proposed Work

IV. Result and Discussion

In the analytic results, Jupyter Notebook version 5.5.0 experiment with Python 3.6 is carried out in this study.

A. Dataset used

An essential criteria for the application of the classification system is efficient data set use. The data set aspect will affect training or model testing. In our proposed classification model (https:/archive.ics.uci.edu / ml / machinelearningdatabases 00350/), vast numbers of default CCF-classification data were given. It has a minimum of 690000 field- and row-size details of 23 columns. The data base utilizes 80% of the product samples and 20% for processing. Since 80% is used for the preparation and 20% for research, the exactness ratio is improved..

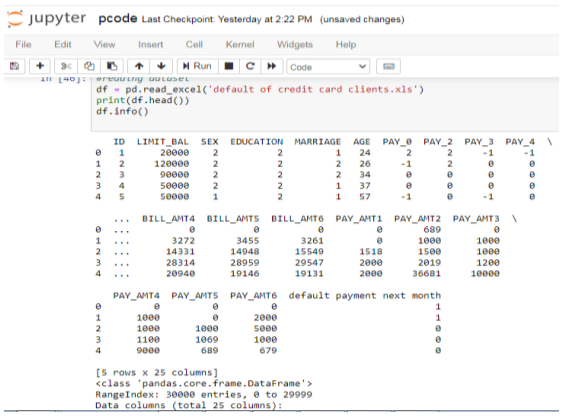

Fig. 2.Dataset Information

B. Screenshot of Results

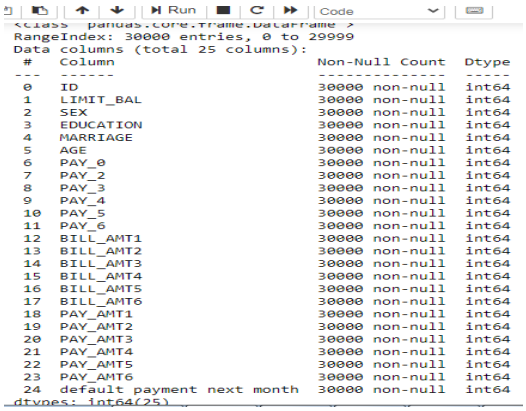

Fig.3.Null Values

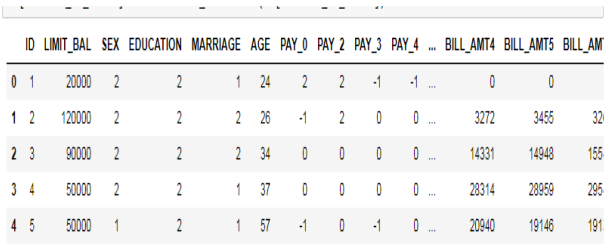

Fig. 4. Scaling the data by standard scaler

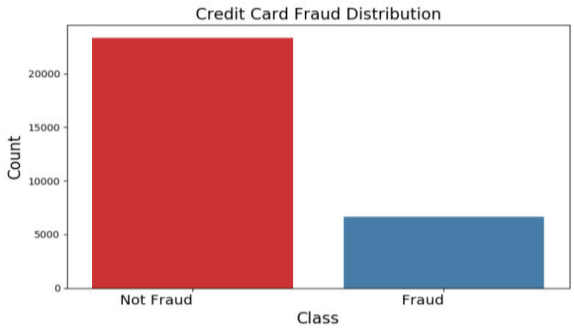

Fig. 5. Credit card fraud distribution

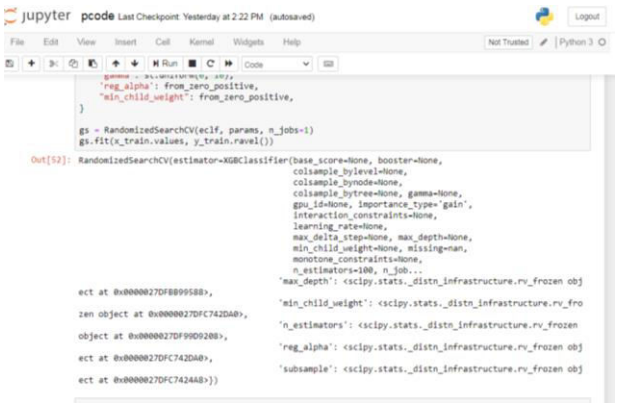

Fig. 6. XGBClassifier with Randomized Search CV

```
{'colsample_bytree': 0.8685022112302376,
 'gamma': 8.086312823088146,
 'learning_rate': 0.319852373027533,
 'max_depth': 34,
 'min_child_weight': 35.825156436937064,
 'n_estimators': 80,
 'reg_alpha': 27.09787131293087,
 'subsample': 0.969374420902256}
```

Fig. 7. Parameter Selected for XGBClassifier

Table I. Test Confusion Matrix for the Proposed Model

	Actual class	
Predicted class	True Positive 4496	False positive 207
	False Negative 812	True Negative 485

The confusion matrix parameters of our proposal to label the samples in table I as true positive class, false positive class or false negative class.

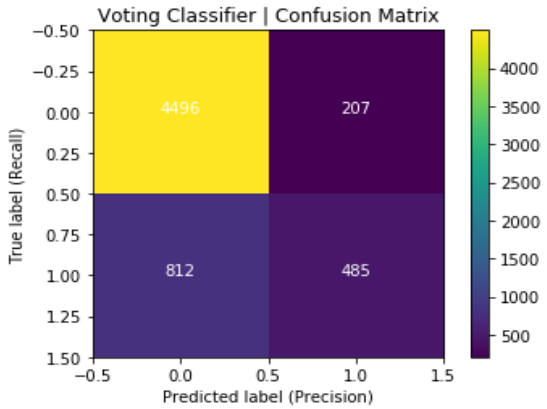

Fig. 8. Confusion matrix

C. Performance Parameters

Diverse output parameters like precision , recall or accurate results, F1-score, Matthews correlation coefficient (MCC), etc. are measured by means of a uncertainty matrix, as well as the value of these parameters is described for classification model.

Table II. Performance Results for Proposed Model

ClassifiersParameters	Existing	Proposed
Accuracy	82.20%	83.0167%
Precision	71.8608%	70.0867%
Recall	35.0037%	37.3940%
F1-Score	47.0763%	48.7682%
MCC	0.4142	0.4252
ROC AUC	0.7634	0.7849

Table II displays the output parameters.
They are calculated with 20% of the test results. The results metrics for precision, recall, accuracy, F1-Score or MCC are evaluated. There has been a statistical performance of 83.02 percent for the new model, substantially higher than for the existing classification model..

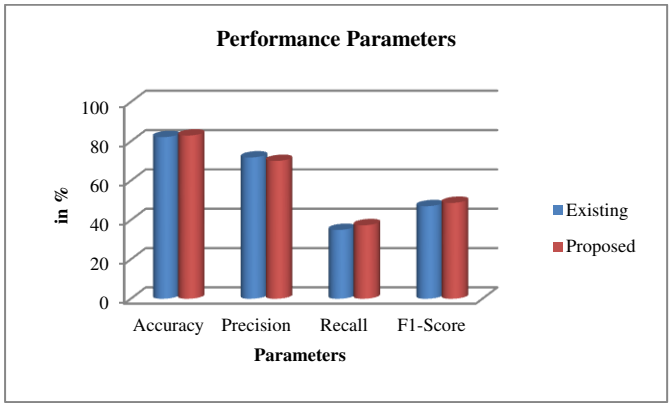

Fig. 9. Comparison Graph of Performance Parameters

Figure 9 displays the output parameters including accuracy, In comparison to the existing classification models, precision or F1-Score are seen. We found, comparison with the present classification scheme, that the exactitude of the proposed model is highest.

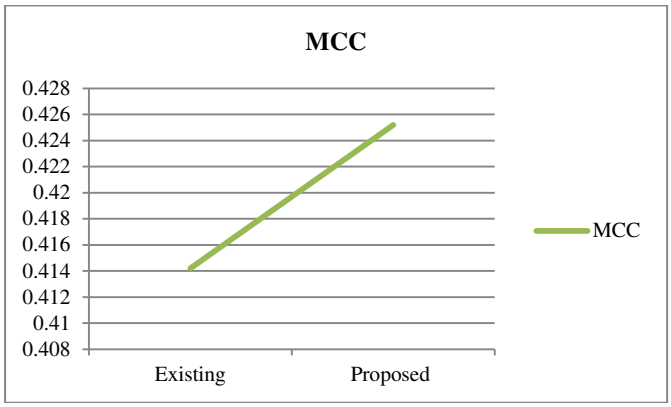

Fig. 10. Comparison Graph of MCC

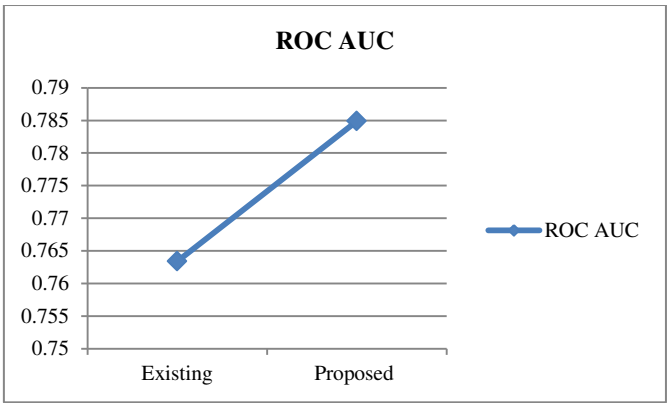

Fig. 11. Comparison Graph of ROC

V. CONCLUSION

Recently, the amount of CCTs , especially in respect of online sales, has risen significantly. The increased usage of CCs for individuals significantly enhances CCTs. Despite the large amount of CCFs it is challenging to identify fraudulent transactions. The requirement for all CC issuing banks is therefore to incorporate successful fraud prevention mechanisms to reduce their losses. Fraudulent transactions in real life are interspersed with genuine transactions but easy pattern matching is not always enough to reliably distinguish them.. In this work, we have used XGB classifier with Randomized Search CV. Randomized Search CV is used to select the parameters for XGB classifier. The findings in terms of precision, recall, accuracy, F1-Score or MCC were obtained via the test results study. There has been a statistical performance of 83.02 per cent for the new model, substantially higher than for the current classification model..

References

[1] A. Srivastava, M. Yadav, S. Basu, S. Salunkhe and M. Shabad, "Credit card fraud detection at merchant side using neural networks," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 667-670.

[2] K. Modi and R. Dayma, "Review on fraud detection methods in credit card transactions," 2017 International Conference on Intelligent Computing and Control (I2C2), Coimbatore, 2017, pp. 1-5, doi: 10.1109/I2C2.2017.8321781.

[3] Alejandro Correa Bahnsen, Djamila Aouada, Aleksandar Stojanovic, Björn Ottersten, "Feature engineering strategies for credit card fraud detection", 0957-4174/ 2016 Elsevier.

[4] S. Askari, A. Hussain, "Credit Card Fraud Detection Using Fuzzy ID3", Proc. of International Conf.On Computing, Communication and Automation (ICCCA),Greater Noida, India, pp.446-452, 2017.

[5] I. SADGALI, N. SAEL and F. BENABBOU, "Fraud detection in credit card transaction using machine learning techniques," 2019 1st International Conference on Smart Systems and Data Science (ICSSD), Rabat, Morocco, 2019, pp. 1-4, doi: 10.1109/ICSSD47982.2019.9002674.

[6] Kang Fu, Dawei Cheng, Yi Tu, and Liqing Zhang, "Credit Card Fraud Detection Using Convolutional Neural Networks", Springer International Publishing AG 2016.

[7] A. A. Taha and S. J. Malebary, "An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine," in IEEE Access, vol. 8, pp. 25579-25587, 2020, doi: 10.1109/ACCESS.2020.2971354.

[8] D. Prusti and S. K. Rath, "Fraudulent Transaction Detection in Credit Card by Applying Ensemble Machine Learning techniques," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 2019, pp. 1-6, doi: 10.1109/ICCCNT45670.2019.8944867.

[9] 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.

[10] G. Goy, C. Gezer and V. C. Gungor, "Credit Card Fraud Detection with Machine Learning Methods," 2019 4th International Conference on Computer Science and Engineering (UBMK), Samsun, Turkey, 2019, pp. 350-354, doi: 10.1109/UBMK.2019.8906995.

[11] C. Jiang, J. Song, G. Liu, L. Zheng and W. Luan, "Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism," in IEEE Internet of Things Journal, vol. 5, no. 5, pp. 3637-3647, Oct. 2018, doi: 10.1109/JIOT.2018.2816007.

[12] M. Jeragh and M. AlSulaimi, "Combining Auto Encoders and One Class Support Vectors Machine for Fraudulant Credit Card Transactions Detection," 2018 Second World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), London, 2018, pp. 178-184, doi: 10.1109/WorldS4.2018.8611624.

[13] S. Dhankhad, E. Mohammed and B. Far, "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study," 2018 IEEE International Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, 2018, pp. 122-125, doi: 10.1109/IRI.2018.00025.

[14] https://scikit-learn.org/0.16/modules/generated/sklearn.grid_search.RandomizedSearchCV.html.

[15] Tianqi Chen and Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System",ACM, KDD '16, San Francisco, CA, USA, August 13-17, 2016. DOI: http://dx.doi.org/10.1145/2939672.2939785.