International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-9, Issue-3, May 2021
https://doi.org/10.21276/ijircst.2021.9.3.5
Article ID IRP1151, Pages 31-36
www.ijircst.org

Analysing Auto ML Model for Credit Card Fraud Detection

Vaishali Garg[1], Sarika Chaudhary[2] and Anil Mishra[3]

[1] Student, Department of Computer Science & Engineering, Amity University, Gurugram, India
[2] Assistant Professor, Department of Computer Science & Engineering, Amity University, Gurugram, India
[3] Assistant Professor, Department of Computer Science & Engineering, Amity University, Gurugram, India

Correspondence should be addressed to Vaishali Garg; vishugarg290@gmail.com

Copyright © 2021 Made Vaishali Garg et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT- Fraud Detection is a major concern these days because of digitalization. We are totally dependent on online transactions these days for even very small needs. There is no doubt that online transactions have made our life very easy but it has increased risk on other hand. And this risk can be very harmful one day. Confidential data is being stolen by the different apps and it is sold in international market. Which later on comes to us in totally different and very harmful way. So why not to use technology again to stop these risks and flaws. Various ML techniques has been observed by researchers but Auto ML is yet not discovered on a wider platform. Therefore, this paper at first aims to explore the trending technology Auto ML. Then a model for evaluating Auto ML is suggested and analysed with different classification algorithms. The experimental results ascertained the accuracy of Auto ML followed by a comparative analysis of ML and Auto ML.

KEYWORDS- Auto ML, Classification, Credit card, Fraud detection, Machine learning

I. INTRODUCTION

For years, fraud has been a serious issue in sectors like banking, medical, insurance, and lots of others. Due to the rise in online transactions through different payment options, like credit/debit cards, PhonePe, Gpay, Paytm, etc., fraudulent activities have also increased. Moreover, fraudsters or criminals became very skilled find escapes in order that they will loot more. Since no system is perfect and there is always a loophole them, it has become a challenging task to make a secure system for authentication and preventing customers from fraud [1]. So, Fraud detection algorithms are very useful for preventing frauds. The rapid growth in E-Commerce industry has led to an exponential increase in the use of credit cards for online purchases and consequently they has been surge in the fraud related to it. Machine learning plays an important role for detecting the master-card fraud within the transactions. Credit card fraud is the most common form of identity theft, affecting more than 10.7 million people annually [2]. It occurs when someone steals a card or snatches personal information to perform so-called card-not-present (CNP) transactions.

Automated machine learning provides methods and processes that enable machine learning professionals to access machine learning without machine learning, in order to improve machine learning efficiency and accelerate machine learning research. In recent years, machine learning (ML) has made great progress, and more and more disciplines are relying on it[3]. However, the key to this achievement rest on the grade to which machine learning experts accomplish the following tasks:

- Pre-process and clean the data.
- Select and construct appropriate features.
- Select an appropriate model family.
- Optimize model hyper-parameters.
- Post-process machine learning models.
- Critically analyse the results obtained.

As the difficulty of these tasks is a lot beyond non-ML-experts, the quick evolution of machine learning applications has formed a demand for off-the-shelf machine learning methods that can be used easily and without expert knowledge.

II. LITERATURE REVIEW

Fraud act because the unlawful or criminal deception intended to end in financial or personal benefit. It's a deliberate act that's against the law, rule or policy with an aim to achieve unauthorized financial benefit [4]. Numerous literatures concerning anomaly or fraud detection during this domain are published already and are available for public usage. A comprehensive survey conducted by Guedlek et al.[8] and his associates have revealed that techniques employed during this domain include data processing applications, automated fraud detection, adversarial detection. Albeit these methods and algorithms fetched an unexpected success in some areas, they did not provide a permanent and consistent solution to fraud detection. An identical research domain was presented by Quah et al.[6] where they used Outlier mining, Outlier detection mining and Distance sum algorithms to accurately predict fraudulent transaction in an emulation experiment of MasterCard transaction data set of 1 certain full service bank . Outlier mining may be a field of knowledge mining which is essentially utilized in monetary and internet fields. It deals with detecting objects that are detached from the most system i.e. the transactions that aren't genuine [7][16]. They need taken attributes of customer's behaviour and supported the worth of these attributes they've calculated that distance between the observed value of that attribute and its predetermined value. There have also been

efforts to progress from a totally new aspect. Attempts are made to enhance the alert-feedback interaction just in case of fraudulent transactions [9]. Just in case of fraudulent transaction, the authorised system would be alerted and a feedback would be sent to deny the continued transaction. Table I shows the comparison among the various existing techniques for the detection of frauds [10]. Advantages and drawbacks of the techniques are discussed below.

Table 1: Summarized Fraud Detection Techniques

Fraud Detection Techniques	Observations	Limitations
K-nearest Neighbour Algorithm	Define anomalies in the target instance and is easy to implement.	Appropriate for detecting frauds with the limitations of memory.
Hidden Markov Model (HMM)[5]	Identify the fraudulent activity during transaction.	Unable to detect fraud with a less transactions.
Neural Network	Detect real-time credit card frauds.	Have many sub-techniques. So, if they pick-up this which is not suitable for credit card fraud detection, the performance of the method will decline.
Decision Tree	Handle non-linear credit card transaction as well.	DT cannot detect fraud at the real time of transaction.
Outlier Detection Method	Lesser memory and computation requirements. Works fast and well for large online datasets.	Cannot find anomalies accurately like other methods.
Deep Learning[14]	It can extract complex patterns	Only used in image recognition. No information to explain the other domains is available. The library of deep learning does not cover all algorithms.

After analysing the literature following gaps are identified:

- The major issue which comes into play is growing technology and with growing technology every day comes a new method of fraud especially in online transaction. Many companies do not reveal these frauds so as to protect their reputation, so most of the frauds remain unreported which leads to another harmful frauds.
- Another important issue is the maintenance of huge amount of database.
- It is very difficult to handle such a huge amount of data. Pre-processing of data takes so much time.
- Real time working problems as the incoming transactions are excessive and behavior of card holders and fraudsters change in rapid way.

III. PROPOSED MODEL DESIGN

This section describe the proposed model based upon the gaps identified. To detect the credit card fraud detection there are ample models that are available. But then the question arises which model to choose, which the best model is? There are many types of Machine Learning models specific to different use cases. As we work with datasets, a machine learning algorithm works in two stages. We usually split the data around 20%-80% between testing and training stages. Under supervised learning, we split a dataset into a training data and test data in Python ML. Fig 1. Depicts the workflow of the model.

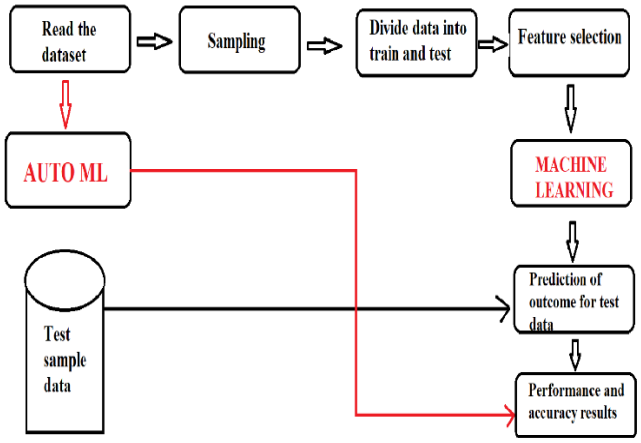

Fig. 1: Workflow of proposed model

A. Pre-processing

First of all the data is read using the panda's library. Data Pre-processing is that step in which the data gets transformed, or Encoded, to bring it to such a state that now the machine can easily parse it. In other words, the features of the data can now be easily interpreted by the algorithm.

B. Oversampling

After the pre-processing step then comes the oversampling. As the data is highly imbalanced. We need to do oversampling to bring the data to the balanced state.

C. Splitting the dataset into test and train data

Entire dataset is divided into two parts. The train data set and test data set. 80% of the data is feed into training and rest 20% is feed into testing.

D. Feature Selection

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model

construction. After the feature selection the results are analyzed and accuracy is measured.

E. AUTO ML

This is the main part of the entire model. With few commands only auto ml compares different models and extra trees classifier model is being built for further prediction.

IV. RESULTS AND DISCUSSIONS

A. Dataset

The Data-set used in this work as depicted in fig. 2 contains the transactions made in two days by European cards in September 2012, gathered and analyzed during a research collaboration of Worldline and the Machine Learning Group of ULB on big data mining and fraud detection. It is freely available on Kaggle. The data contains only numerical values. Due to confidentiality the values where changed by PCA transformation. The features time and amount have not been transformed and all other features are represented by V0, V1……V26 values.

Table 2: Dataset Description

Variable name	Description	Type
V0,V1------ V26	Transaction features after transformation	Integer
Time	Time elapsed between each and the first transaction	Integer
Amount	Amount of transaction	Integer
Class	Non fraud or Fraud	0 or 1

B. Performance Metrics

This Data-set classifies transactions by being fraudulent or not. We have 492 frauds out of 284807, which is highly unbalanced 0.173%. To solve this class unbalance, Random over-Sampling is used. Over Sampling shows the distribution of the Data-set. After Over-Sampling dataset is spliced into training and test sets. For a Pre-trained model performance check, we split the data into two separate training sets and one independent test set for final model comparison. Table III shows the instances.

Table 3: Instances of the dataset

Number of instances	284807
Split ratio for pre – training	0.2
Split ratio for training	0.4
Independent test set	0.4

C. Evaluation Metrics

The evaluation of the model was carried out using the various evaluation metrics such as Accuracy, Precision, F1-score, Recall.

Accuracy: is defined as the number of correct predictions made by the model. It is the proportion of the total number of correct predictions [11].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: defines the results classified as positive by the model, how many were actually positive. It is the number of items correctly identified as positive out of total true positives [12].

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positives}}$$

Recall: It is the number of items correctly identified as positive out of the total items classified as positive[13][15].

$$\text{Recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negatives}}$$

F1-Score: is the weighted average of the precision and the recall, it takes both false negatives and positives into the account and gives a better outlook especially in an uneven class distribution it is given as:

$$\text{F1 Score} = 2\left(\frac{\text{Precision} * \text{recall}}{\text{Precision} + \text{recall}}\right)$$

Where True positive (TP) represents data detected as fraudulent, True negative (TN) represents data detected as legitimate, False positive (FP) represents normal data detected as fraudulent, and False Negative (FN) is denoted as fraud data detected as normal[13].

D. Experimental results

The described model is evaluated using thirteen algorithms as described in the figure below. Python and Jupyter Notebook is utilized for implementation. Fig.2 shows the comparison of various existing models using auto ml on different parameters like accuracy, f-1 score, recall, time and precision.

International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-9, Issue-3, May 2021
https://doi.org/10.21276/ijircst.2021.9.3.5
Article ID IRP1151, Pages 31-36
www.ijircst.org

```
best_model = compare_models()
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
et	Extra Trees Classifier	0.9996	0.9456	0.7959	0.9460	0.8635	0.8633	0.8670	17.059
rf	Random Forest Classifier	0.9995	0.9474	0.7807	0.9406	0.8528	0.8525	0.8565	125.848
lda	Linear Discriminant Analysis	0.9993	0.9009	0.7368	0.8514	0.7875	0.7871	0.7904	1.081
ada	Ada Boost Classifier	0.9992	0.9701	0.7003	0.8235	0.7556	0.7552	0.7583	39.809
lr	Logistic Regression	0.9991	0.9459	0.6045	0.8139	0.6900	0.6896	0.6991	7.421
dt	Decision Tree Classifier	0.9991	0.8763	0.7530	0.7449	0.7461	0.7456	0.7471	10.843
ridge	Ridge Classifier	0.9989	0.0000	0.4257	0.8214	0.5525	0.5520	0.5859	0.177
gbc	Gradient Boosting Classifier	0.9989	0.5691	0.4188	0.7796	0.5054	0.5050	0.5441	221.055
knn	K Neighbors Classifier	0.9984	0.6034	0.0586	0.8167	0.1083	0.1081	0.2137	2.501
svm	SVM - Linear Kernel	0.9982	0.0000	0.0000	0.0000	0.0000	-0.0001	-0.0002	5.727
lightgbm	Light Gradient Boosting Machine	0.9951	0.6923	0.5381	0.2131	0.2999	0.2982	0.3328	3.414
nb	Naive Bayes	0.9926	0.9662	0.6259	0.1370	0.2246	0.2224	0.2903	0.167
qda	Quadratic Discriminant Analysis	0.9758	0.9678	0.8667	0.0584	0.1093	0.1065	0.2212	0.597

Fig. 2: Comparison of various ML algorithms using auto ML

As analyzed from the above figure, extra tree classifier comes out to be the best algorithm with Auto ML. Fig.3 shows the Precision-Recall curve for extra trees classifier. Average precision comes out to be 0.73 and fig. 4 illustrate confusion matrix for the same.

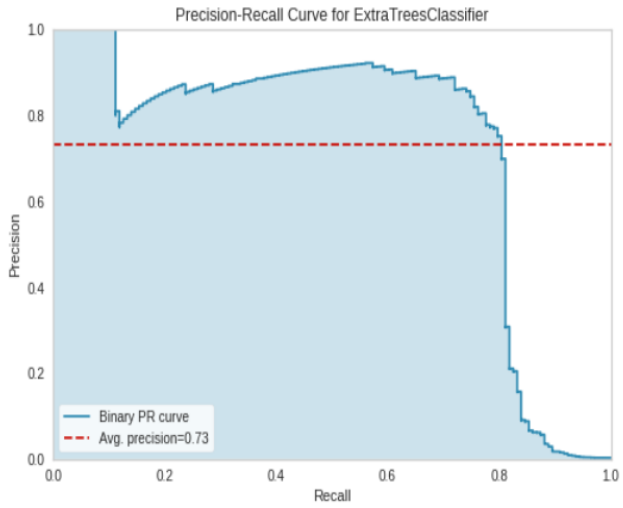

Fig.3: Precision-Recall curve for extra tree classifier

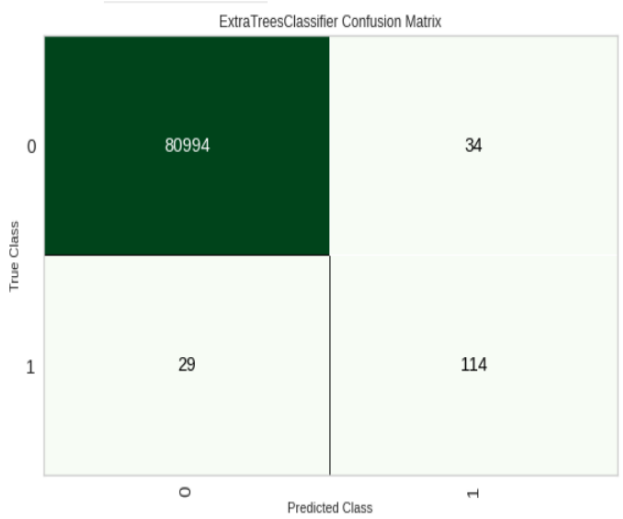

Fig.4: Confusion matrix of extra trees classifier

Fig.5 (a-d) shows the comparison of existing models using Auto ML on the basis of accuracy.

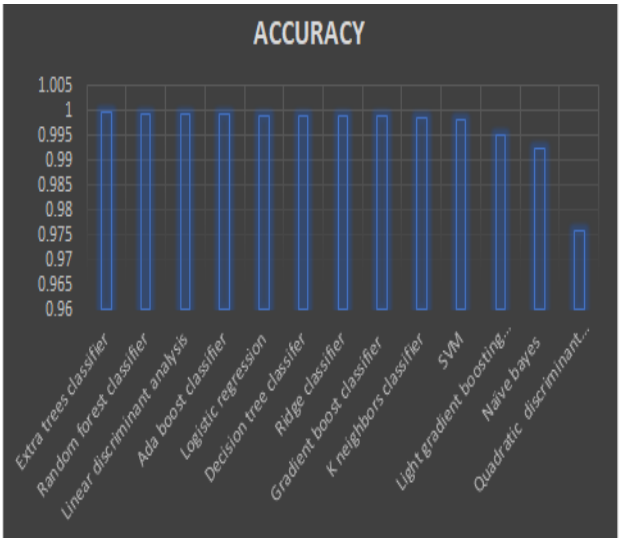

Fig. 5(a): Graph of existing models on the basis of accuracy

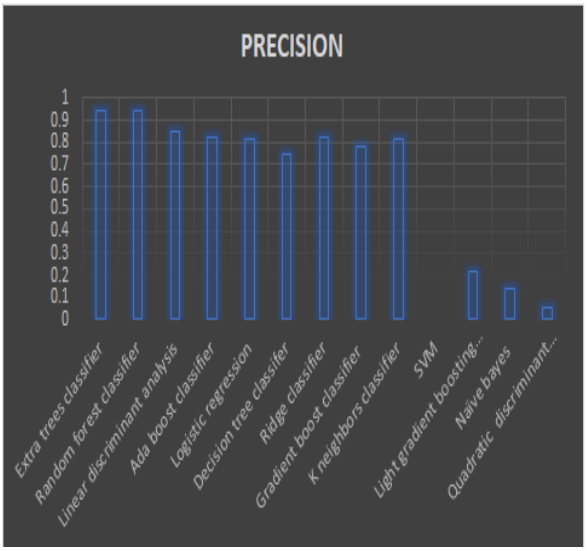

Fig. 5(d): Comparison of existing models on precision

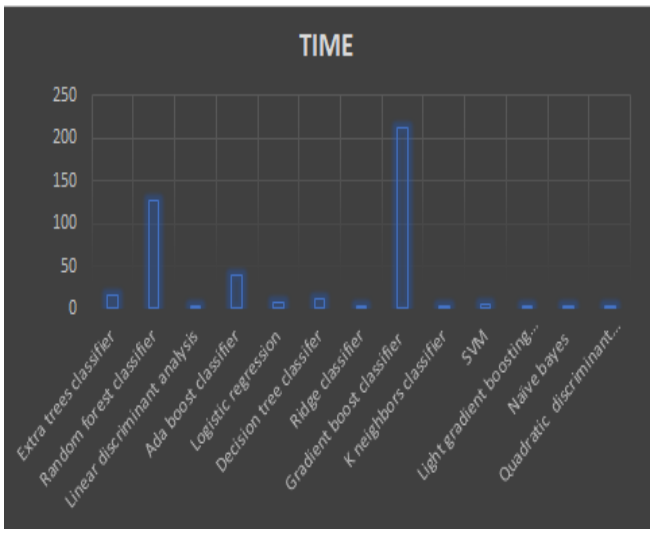

Fig.5(b): Graph of existing models on the basis of time.

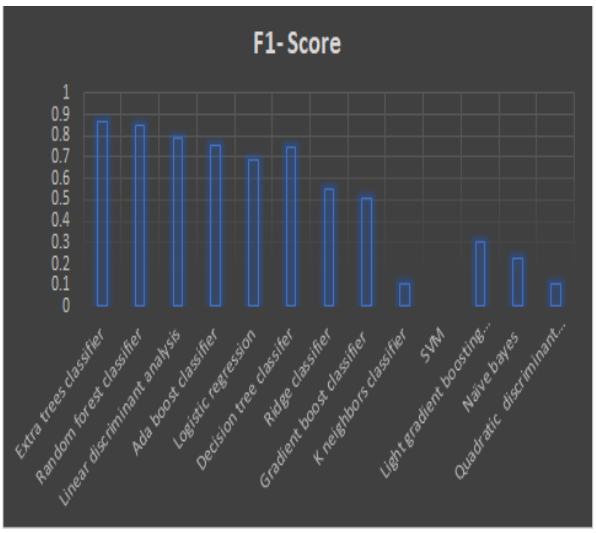

Fig. 5(e): Comparison of existing models on F1- score

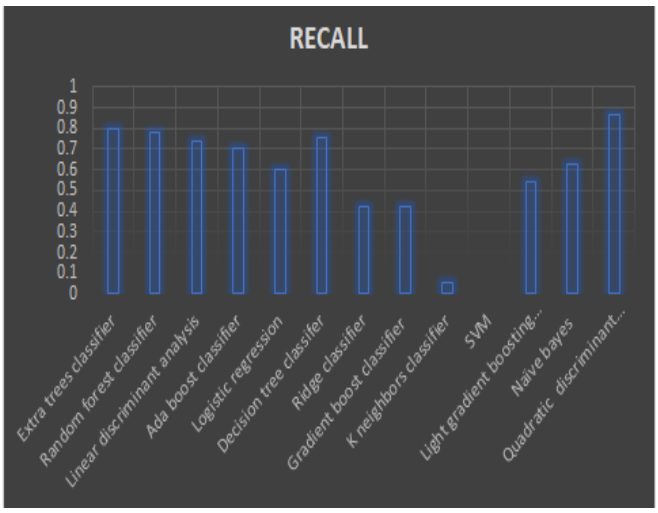

Fig. 5(c): Comparison of existing models on recall

Table 4: Comparative Analysis of Auto Ml and ML based on Experimental Evaluation

Feature	Auto ML	ML
Programming Complexity	Line of code is less	Line of code is more
Skill Requirement	Less skilled data scientists can also work	Skilled data scientists are needed to build models
Processing Time	Less	More
Hyper-Parameter Optimization	Present	Absent

V. CONCLUSION

In this paper different ML and Auto ML techniques are reviewed to discover the research gap. One of the significant observed approach in ML is classification. Mostly ML techniques have been utilized for credit card

fraud detection in past. Auto ML has still not discovered yet on a bigger platform for the same. Also, a novel auto ML based model is proposed. The model is capable to detect credit card fraud in comparison to various existing ML model in terms of processing time and easiness. Also, the quality of model is measured in terms of factors like accuracy, time, precision, recall. Then a comparative analysis of auto ml on existing models is being done and extra trees classifier comes out to be the best model on the factors like accuracy and time. The datasets examined for the analysis have been retrieved from online libraries, in future they can be directly collected from software industries to draw a fair and reasonable comparisons to measure the effectiveness of evaluation process .

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] Maniraj, S., Saini, A., Ahmed, S., & Sarkar, S., "Credit card fraud detection using machine learning and data science". International Journal of Engineering Research and, 8(09) 2019.

[2] Patil, S., Nemade, V., & Soni, P. K. (2018). Predictive modelling for credit card fraud detection using data analytics. Procedia computer science, 132, 385-395.

[3] Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2019, March). Credit card fraud detection-machine learning methods. In 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-5). IEEE.

[4] Carcillo, F., Le Borgne, Y. A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2019). Combining unsupervised and supervised learning in credit card fraud detection. Information Sciences.

[5] Srivastava, A., Kundu, A., Sural, S., & Majumdar, A. (2008). Credit card fraud detection using hidden Markov model. IEEE Transactions on dependable and secure computing, 5(1), 37-48.

[6] Quah, J. T., & Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. Expert systems with applications, 35(4), 1721-1732.

[7] S. Akila and U. Srinivasulu Reddy, "Cost-sensitive Risk Induced Bayesian Inference Bagging (RIBIB) for credit card fraud detection," Journal of Computational Science, vol. 27, pp. 247–254, Jul. 2018, doi: 10.1016/j.jocs.2018.06.009.

[8] M. Ozbayoglu, M. U. Gudelek, and O. B. Sezer, "Deep learning for financial applications : A survey," Applied Soft Computing, vol. 93, p. 106384, Aug. 2020, doi: 10.1016/j.asoc.2020.106384.

[9] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," Decision Support Systems, vol. 50, no. 3, pp. 602–613, Feb. 2011, doi: 10.1016/j.dss.2010.08.008.

[10] G. C. de Sá, A. C. M. Pereira, and G. L. Pappa, "A customized classification algorithm for credit card fraud detection," Engineering Applications of Artificial Intelligence, vol. 72, pp. 21–29, Jun. 2018, doi: 10.1016/j.engappai.2018.03.011.

[11] Carcillo, Y.-A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi, "Combining unsupervised and supervised learning in credit card fraud detection," Information Sciences, May 2019, doi: 10.1016/j.ins.2019.05.042.

[12] S. M. S. Askari and M. A. Hussain, "IFDTC4.5: Intuitionistic fuzzy logic based decision tree for Etransactional fraud detection," Journal of Information Security and Applications, vol. 52, p. 102469, Jun. 2020, doi: 10.1016/j.jisa.2020.102469.

[13] C. S. Throckmorton, V. Mohan, J. M. William and C. Leslie, "Financial fraud detection using vocal, linguistic and financial cues," 2018.

[14] Y. Pandey, "Credit card fraud detection using deep learning" Int. J. Adv. Res. Comput. Sci., vol. 8, no. 5, May–Jun. 2017.

[15] Malik, Sanjay Kumar, and Sarika Chaudhary. "Comparative study of decision tree algorithms for data analysis." International Journal of research in Computer Engineering and Electronic. Page1 2 (2013).

[16] Kaur, Sonamdeep, Sarika Chaudhary, and Neha Bishnoi. "A Survey: Clustering Algorithms in Data Mining." International Journal of Computer Applications 975 (2015): 8887.

[17] Mandiratta, Sonam, Pooja Batra Nagpal, and Sarika Chaudhary. "A Perlustration of Various Image Segmentation Techniques." International Journal of Computer Applications 139.12 (2016).

[18]Sarika Chaudhary, Yojna Arora, Neelam Yadav (2020). Optimization of Random Forest Algorithm for Breast Cancer Detection IJIRCST Vol-8 Issue-3 Page No-63-66.

[19]Chaudhary,S., Nagpal.P (2019). "Live location tracker" , Global Research and Development Journal for Engineering, | Volume 4, Issue 10

ABOUT THE AUTHORS

Ms. Vaishali Garg is currently pursuing M.C.A from Amity University Haryana. She has published 4 research papers. Her research interest is Machine learning and deep learning.

Ms. Sarika Chaudhary is currently designated as Assistant Professor in CSE, Amity University, Haryana. She has published more than 34 research papers and 02 books. She is member of 16 Professional/Technical Committees and Editorial Board Member/Reviewer of 20 reputed Journals.

Dr. Anil Mishra is currently designated as Assistant Professor in CSE, Amity University, Haryana. He has published more than 16 research papers. He is an active member of ISTE, IEEE and CSI.