Annals of R.S.C.B., ISSN:1583-6258, Vol. 24, Issue 2, 2020, Pages. 261 - 269
Received 24 October 2020; Accepted 15 December 2020.

Implementation and Design onFraud Detection and Prediction of Mobile Money Transaction Using ML Techniques

Ms. NishaBalani
Asst. Professor, Dept. of Computer Science and Engineering Jhulelal Institute of Technology
Nagpur, India

Ms. MeherBhawnani
Asst. Professor, Dept. of Computer Science and Engineering Jhulelal Institute of Technology
Nagpur, India

Ms. AnkitaKamle
Dept. of Computer Science and Engineering Jhulelal Institute of Technology
Nagpur, India

Abstract— Search Money Laundering is increasing considerably with the development of modern technology and the global superhighways of communication. Money Laundering and other frauds cost consumers and financial companies billions of dollars annually, and fraudsters continuously try to find new rules and tactics to commit illegal actions. Thus, fraud detection systems became essential for banks and financial organizations, to attenuate their losses. However, there is a lack of published literature on money laundering through mobile transaction techniques, due to the unavailable dataset on financial services and especially in the emerging mobile money transaction domain for researchers. Along with the great increase in mobile money transactions, fraud has become increasingly rampant in recent years. This study investigates the efficacy of applying different classification models to mobile money transaction fraud detection problems. Three different classification methods, i.e., Random Forest Classifier, KNNeighbors classifier, and logistic regression are tested for their applicability in fraud detections. The performance evaluation is performed on a synthetic mobile money transaction dataset to demonstrate the benefit of the different models.

***Index Terms* — Money transfer fraud, Fraud Detection, Random Forest Classifier, KN-Neighbor's classifier and logistic regression.**

I. INTRODUCTION

In recent years, credit card users have suffered a huge amount of loss because of fraud. Many researchers are working on the early detection of credit card frauds. The main tools used by researchers for credit card fraud detection include ML Algorithms, Neural networks, Classification, and Clustering Methods.Machine learning algorithms are AI techniques that are used in various disciplines to solve problems mainly deals with a large amount of data. Many researchers applied machine learning and deep learning techniques to detect frauds in credit cards. However, there is still a need to analyze and apply the power of ML algorithms to detect frauds in credit card transactions. The areas in which machine learning algorithms are in use are as follows:

Classification finds some conclusions from a huge amount of data. When given some input values from the data, the classification algorithms attempt to select one or more outputs on the basis of the input data. Machine learning algorithms are very useful in classification.

Regression is a supervised learning technique. It is used to predict output values from given input values. It is mostly used to predict continuous data. Regression techniques are machine learning techniques that are very useful in prediction.

Clustering It refers to dividing the problem space into groups on the basis of the similarities between the data. The items in one cluster are very similar to each other. Items in different clusters are different from each other in their properties. Machine learning algorithms are very useful in clustering.

The use of machine learning algorithms is not limited to these areas only. Even many researchers are working on areas in which machine learning algorithms are applicable and will give better results. In this paper, machine learning algorithms are applied to the detection of frauds using credit card transactions. The next section is discussing the proposed work.

II. PROPOSED WORK

In this paper, Machine learning algorithms employed in prediction and detection specifically Random Forest, KNeighbors, logistic Regression, and XGBOOST square measure applied on a true information set having

Annals of R.S.C.B., ISSN:1583-6258, Vol. 24, Issue 2, 2020, Pages. 261 - 269
Received 24 October 2020; Accepted 15 December 2020.

information of over one lacks credit cards. The operating of those machine learning algorithms is as follows:

A. Random Forests (RF)

Random Forest (RF) could also be a really helpful machine learning formula. it's principally used in areas like classification, statistical procedure, etc. At the coaching time, the RF formula creates several call trees.

RF may be a supervised learning approach that wants to check information for the model for coaching. It creates random forests for the matter set then finds the answer to mistreatment of these random forests.

Ensemble learning is through the gathering of hypotheses and combines their predictions to urge a far better prediction than one hypothesis prediction. For instance, generating a hundred completely different call trees for constant information or the subsets of the info have them vote on the simplest classification. in an exceedingly random Forest, the key motivation is to cut back the error rate and also the hope is that it'll become way more unlikely that the ensemble will misclassify an associate example. once hairdressing multiple freelance and various choices every of which may be a minimum of additional correct than random guess, random errors cancel each other out, and correct choices square measure strengthened.

Random forest or random call for classification, regression, and different tasks, forest square measure associate ensemble learning technique that operates by constructing a mess of call trees at coaching time and outputting the class that's the mode of the categories (classification) or mean prediction (regression) of the individual trees. Random call forests are correct for call trees' habit of overfitting to their coaching set.

B. logistic Regression

Logistic Regression may be a classification formula. It's accustomed to predict a binary outcome (1 / zero, Yes / No, True / False) given a gaggle of freelance variables. In easy words, it predicts the likelihood of prevalence of an incident by fitting information to a logit performance. Hence, it's conjointly referred to as logistical regression. Since it predicts the likelihood, its output values lie between zero and one. logistic regression models a separate target variable as a performance of many feature variables. The target variable is that the sentiments column.

C. kNN

In the K-Neighbors classification, the output could also be a category membership. the associate object is assessed by a majority vote of its neighbors, with the issue being assigned to the class commonest among its K nearest neighbors. Here K is sometimes a positive number

This is a form of instance-based learning or lazy learning wherever they perform is just approximated domestically and each one the computation is delayed till the classification

The principle behind nearest neighbor strategies is to hunt out a predefined variety of employment samples nearest in distance to the new purpose and predict the label from these. The number of samples usually measure a user-defined constant (k-nearest neighbor learning) or vary supported by the native density of points (radius-based neighbor learning). The space will, generally, be any metric measure: customary Euclidean distance is the most common selection. Neighbors-based strategies square measure observed as non-generalizing machine learning strategies since they merely "remember" all of their coaching information (possibly reworked into a fast assortment structure.

D. XGBoost

XGBoost is an optimized distributed gradient boosting library designed to be extremely economical, flexible, and transportable. It implements machine learning algorithms underneath the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also mentioned as GBDT, GBM) that solves several information science issues throughout a fast and correct means.

This is an associate ensemble technique that seeks to create a sturdy classifier (model) supported "weak" classifiers. During this context, a weak and robust respect to a life of however correlated measures the learners to the particular target variable. By adding models on high of each different iteratively, the errors of the previous model square measure corrected by the next predictors, till the coaching information is accurately foretold or reproduced by the model. If you'd prefer to probe boosting slightly additional, examine info a couple of few fashionable implementations referred to as AdaBoost (Adaptive Boosting).

Now, gradient boosting conjointly includes associate ensemble technique that adds predictors and corrects previous models. However, rather than distributing completely different weights to the classifiers when each iteration, this technique fits the new model to new residuals of the previous prediction then minimizes the loss once adding the most recent prediction. So, at intervals the highest, you are changing your model mistreatment gradient descent and thence the name, gradient boosting. This is often supported for each regression and classification issues.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 24, Issue 2, 2020, Pages. 261 - 269
Received 24 October 2020; Accepted 15 December 2020.

III. DATA DESCRIPTION

A.Data Acquisition -

The dataset has been generated employing a machine referred to as PaySim. PaySim uses collective knowledge of 1 month of economic log from a mobile cash service enforced in AN African country. the info was created obtainable to kaggle (https://www.kaggle.com/ntnutestimon/paysim1/downloads/paysim1.zip/2) by a international company UN agency could be a mobile financial service supplier in additional than fifteen countries.

This dataset had 8213 frauds out of 6354407 transactions. The dataset was extremely unbalanced, the positive category (frauds) account for 0.2% of all transactions.

Column Name	Column Description
Step	Maps are a unit of time in the real world.
Type	Type of transaction.
Amount	Amount of transactions in local currency.
NameOrg	Customer who started the transaction.
oldbalanceOrig	Initial balance before the transaction.
newbalanceOrig	Customer's balance after transaction
nameDest	Recipient ID of the transaction.
oldbalanceDest	Initial recipient balance before the transactions.
newbalanceDest	Recipient's balance after transaction
isFraud	Identifies a fraudulent transaction (1) and non-fraudulent (0)
isFlaggegFraud	Flags illegal attempts to transfer more than 200,000 monetary unit in a single transaction

Table 1. Column Description

The dataset contains numerical, alphabetical, and categorical input variables containing 630000 rows. Dataset was given eleven features and every one containing labels. Feature's step contains a unit of your time wherever one signifies one hour of your time, nameOrig and nameDest incorporates a client or merchandiser IDs. Balance is the offered quantity in every customer's account. The amount is the group action amount. From these variables, the variables like step, amount, oldBalanceOrg, newBalanceOrg, oldbalanceDest, and newbalanceDest area unit mapped to numeric identifiers. The variable sort is mapped to a categorical variable.

B. Sample review -

Annals of R.S.C.B., ISSN:1583-6258, Vol. 24, Issue 2, 2020, Pages. 261 - 269
Received 24 October 2020; Accepted 15 December 2020.

step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlag
1	PAYMENT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.0	0.0	0	0
1	PAYMENT	1864.28	C1666544295	21249.0	19384.72	M2044282225	0.0	0.0	0	0
1	TRANSFER	181.00	C1305486145	181.0	0.00	C553264065	0.0	0.0	1	0
1	CASH_OUT	181.00	C840083671	181.0	0.00	C38997010	21182.0	0.0	1	0
1	PAYMENT	11668.14	C2048537720	41554.0	29885.86	M1230701703	0.0	0.0	0	0

Table 2. Review of the Sample data

IV. LITERATURE SURVEY

YEAR	NAME OF PAPER	ADV ANT AGES	DIS-ADVANTAGES	ALOGO-RITHM	FEATURES
2018	Using Genetic Algorithm to Improve Classification of Imbalanced Datasets for credit card fraud detection	to enhance fraud detection in e-banking by using K-Means clustering and genetic algorithm as an oversampling strategy.	Class distribution is extremely unbalanced in credit card transactions, since frauds are typically less than 1% of the overall transactions.	K-means clustering and the genetic algorithm.	1.To enhance classified performance of the minority of credit card fraud instances in the imbalanced data set, for that we propose a sampling method based on the K-means clustering and the genetic algorithm. 2. K-means algorithm to cluster and group the minority kind of sample, and in each cluster, we use the genetic algorithm to gain the new samples and construct an accurate fraud detection classifier.
2018	Bank Fraud Detection Using Support Vector Machine	SVM-S is more reliable and accurate than BPN. With time complexity, SVM-S uses a few times for predicting anomalies as compared with other algorithms such as BPN.	For the method, the slight improvement on credit scoring databases was because of the difficulty of obtaining real databases. The results can be improved by studying the influence of various parameters used by the SVM-S architecture.	Supervised learning methods Support Vector Machines with Spark (SVM-S)	1.The results obtained from databases of credit card transactions show that these techniques are effective in the fight against banking fraud in big data. 2.Experiment results from the study show that SVM-S have better prediction performance than Back Propagation Networks (BPN). 3.Besides the average prediction, accuracy reaches a maximum when training the data ratio arrives at 0.8.

2019	Comparative Evaluation of Credit Card Fraud Detection Using Machine Learning Techniques.	The method of logistic regression showed the greatest accuracy of results across the assessment metrics used.	This exploration on distinguishing charge card extortion has extraordinary potential for future ramifications. The skewed data handling techniques need to be simplified.	Machine Learning, Naive Bayes, Logistic Regression, Support Vector Machine, K-Nearest Neighbor.	1.The execution of these techniques is assessed dependent on accuracy, sensitivity, precision, specificity. 2.The outcomes show an ideal accuracy for logistic regression, Naive Bayes, k-Nearest neighbor and Support vector machine classifiers are 99.07%, 95.98%, 96.91%, and 97.53% respectively. 3.The relative outcomes demonstrate that logistic regression performs superior to other algorithms.
2019	Building a robust mobile payment fraud detection system with adversarial examples	Compared some of the most popular fraud detection methods in an adversarial attack condition. Then, oversampled these adversarial examples to construct a more robust mobile payment fraud detector.	There is no guarantee that SMOTE synthesized points will help the model to reach the task decision boundary. As for adversarial oversampling, no strong assumption is made. Above all, it helps to push back the decision boundary towards the task decision boundary (i.e. the theoretical decision boundary for the task) by anticipating fraudsters next moves.	Adversarial machine learning, oversampling.	1.To build a robust mobile fraud detection system using adversarial examples.
2020	Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison	used an imbalanced dataset to check the suitability of different supervised machine learning models to predict the chances of occurrence of a fraudulent transaction.	Accuracy as a parameter was not used as it is not sensitive to imbalanced data and does not give a conclusive answer.	Machine Learning, Supervised Learning.	1.to evaluate an imbalanced dataset with the help of various supervised machine learning models. 2.To determine which one of those is the best suited for detecting credit card frauds. 3. To evaluate a dataset on the basis of various predefined criteria.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 24, Issue 2, 2020, Pages. 261 - 269
Received 24 October 2020; Accepted 15 December 2020.

V. RESULT AND ANALYSIS

1. Distribution of Type field:

The Bar chart below shows the distribution of transactions in percentage with respect to the transaction types.

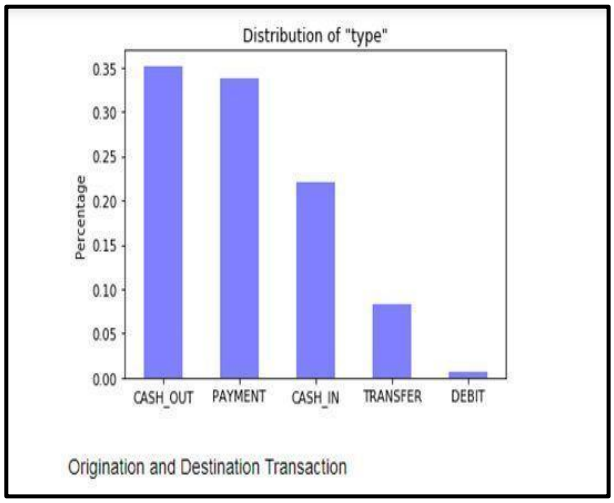

Fig.2. Bar chart for No of transaction

2. Transaction Amount and the step:

A scatter plot could be a style of plot or mathematical diagram victimization philosopher coordinates to show values for generally 2 variables for a collection of information. the info or displays as a group of points, every having the worth of 1 variable deciding the position on the horizontal axis and also the value of the opposite variable deciding the position on the vertical axis.

The below scatter plot shows the distribution of transactions with relevancy hours (steps). Here we have a tendency to see that there's no relation between the hours (steps) and deceitful transactions. Thus we are going to not be a victimization of this field for our predictions.

The scatter plot below describes the dispersion of the dealing quantity with fraud transactions with relevancy time (step within the column) and that we here conclude that the pattern is non-linear the fraudsters don't follow any specific time or day to perform any fraud.

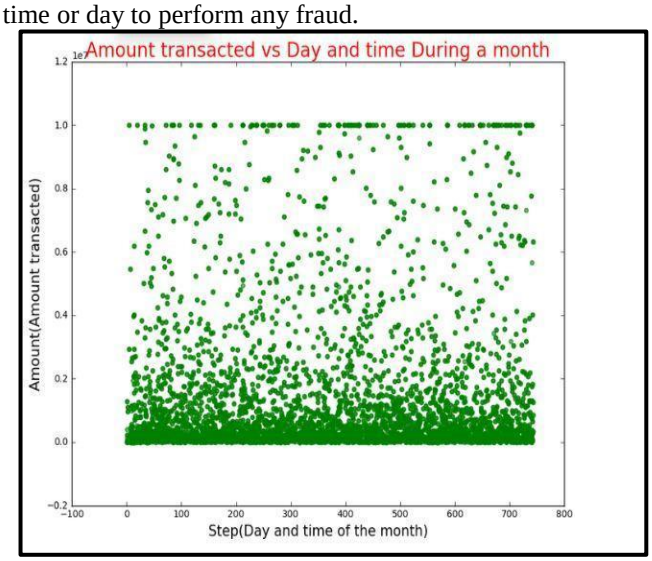

Fig.3. Transaction Amount and the step (Day and time of the Month – Scatter Plot

Annals of R.S.C.B., ISSN:1583-6258, Vol. 24, Issue 2, 2020, Pages. 261 - 269
Received 24 October 2020; Accepted 15 December 2020.

3. Correlation:

A great thanks to exploring new information is to use a pairwise matrix. This can live the correlation between each combination of your variables. It doesn't matter if you have got an AN outcome (or response) variable at now, it'll compare everything against everything else.

For those not conversant in the parametric statistic, it's merely a live of similarity between 2 vectors of numbers. The measured worth will vary between one and -1, wherever one is dead correlative, -1 is dead reciprocally correlative, and zero isn't correlative in any respect. Thus, we have a tendency to cypher the correlation of the output with all input numerical options. Higher the correlation between the output and therefore the options higher area unit the possibilities of fitting AN correct model.

The following graph shows the correlation between completely different options in crucial whether or not the dealings were a real or a fallacious one.

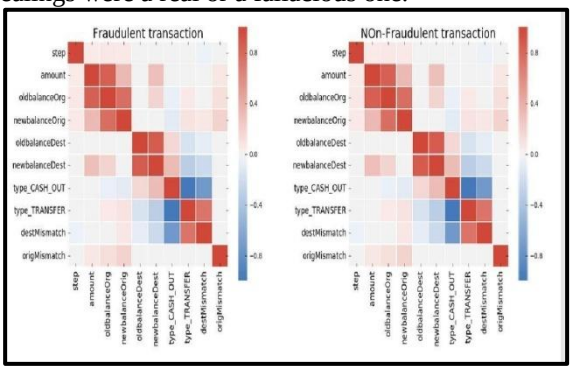

Fig.4. Pearson Correlation for all parameters

4. Transactions, Amount and Types

We more analyzed the dealings with relation to the dealing's quantity and whether or not it's Fraud or not. we tend to see that the dealings that square measure fraud aren't having extreme higher values for the transaction quantity. The red color plot below describes an equivalent.

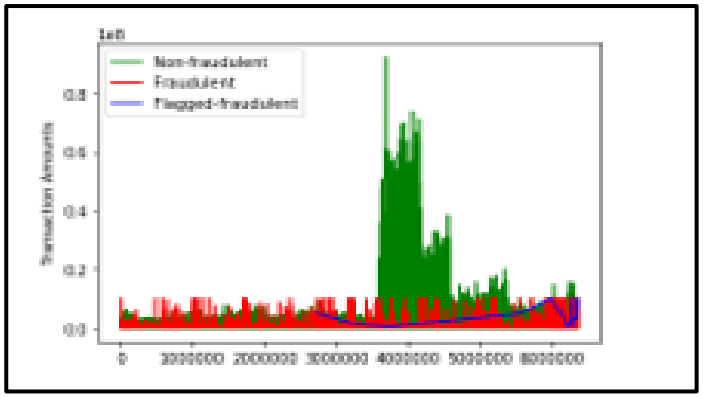

Fig5. Fraud and Non-Fraud with Transaction Amount

5. Results summary

Algorithm	Precision	Recall	F1-Score
Random Forest	**0.99**	**0.82**	**0.92**
KNeighbors	**0.90**	**0.76**	**0.82**

Annals of R.S.C.B., ISSN:1583-6258, Vol. 24, Issue 2, 2020, Pages. 261 - 269
Received 24 October 2020; Accepted 15 December 2020.

Logistics Regression	**0.90**	**0.74**	**0.80**
XGBoost	**1.00**	**0.99**	**0.99**

Result Summary

CONCLUSION AND FUTURE WORK

In this study, we tend to use Associate in Nursing unbalanced dataset to envision the suitableness of assorted supervised machine learning models to predict the possibilities of prevalence of a deceitful dealing. We tend to use sensitivity, precision, and time as a result of the deciding parameters to come back to a selected conclusion. Accuracy as a parameter wasn't used as a result of it's not sensitive to unbalanced knowledge and does not give a conclusive answer. We tend to analyze the kNN, XGBoost, logistic Regression, and Random Forest models throughout this study. We tend to use these models for predicting the possibilities of prevalence of a deceitful Mastercard dealings out of a given variety of transactions. Mastercard frauds are a contemporary issue which we tend to come to the conclusion that the only suited model for predicting such frauds is that the choice Tree model. The analysis shows that the sensitivity of the kNN model is larger than that of a choice tree, however because the time taken by kNN for testing the information is extraordinarily massive, we tend to elect call Tree over kNN. Simply just in case of fraud detection, we'd prefer to make certain that minimum time is taken for prediction, therefore, a choice Tree is the most popular model. Future researchers throughout this field might apply the resampling techniques to the various datasets obtaining used. This method helps to cut back the imbalance quantitative relation of datasets that in turn produces higher classification results.

After the comparative analysis of the numerous supervised Learning models, we'll infer that the selection Tree Model is the most effective approach to be used for police work Mastercard fraud detection. But, the performance of the choice Tree Model should even be evaluated with the help of unsupervised machine learning models at intervals in the long run to provide an additional conclusive result. This tells America whether or not the model that is chosen is also a higher more robust an improved choice or the unsupervised machine learning techniques perform better

REFERENCES

1. [K. K. Tripathi and M. A. Pavaskar, “Survey on credit card fraud detection methods,” International Journal of Emerging Technology and Advanced Engineering, vol. 2, no. 11, pp. 721–726, 2012.
2. A. Srivastava, A. Kundu, S. Sural, and A. Majumdar, “Credit card fraud detection using hidden markov model,” IEEE Transactions on dependable and secure computing, vol. 5, no. 1, pp. 37–48, 2008.
3. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199, 2013.
4. N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in 2017 IEEE Symposium on Security and Privacy (SP), pp. 39–57, IEEE, 2017.
5. A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” arXiv preprint arXiv:1607.02533, 2016.
6. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” arXiv preprint arXiv:1706.06083, 2017.
7. K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel, “Adversarial examples for malware detection,” in European Symposium on Research in Computer Security, pp. 62–79, Springer, 2017.
8. N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, “Distillation as a defense to adversarial perturbations against deep neural networks,” in 2016 IEEE Symposium on Security and Privacy (SP), pp. 582–597, IEEE, 2016.
9. M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, “Can machine learning be secure?,” in Proceedings of the 2006 ACM Symposium on Information, computer and communications security, pp. 16–25, ACM, 2006.
10. F. N. Ogwueleka, “Data Mining Application in Credit Card Fraud Detection System,” vol. 6, no. 3, pp. 311-322, 2011.
11. O. S. Yee, S. Sagadevan, N. Hashimah, and A. Hassain, “Credit Card Fraud Detection Using Machine Learning As Data Mining Technique,” vol. 10, no. 1, pp. 23-27.

12. C. Phua, D. Alahakoon and V. Lee, "Minority report in fraud detection", ACMSIGKDD Explorations Newsletter, vol. 6, no. 1, p. 50, 2004.
13. N. Sethi and A. Gera, "A Revived Survey of Various Credit Card Fraud Detection Techniques", International Journal of Computer Science and Mobile Computing, vol. 3, no. 4, pp. 780-791, 2014.
14. J. Awoyemi, A. Adetunmbi and S. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis", 2017 International Conference on Computing Networking and Informatics (ICCNI), 2017.
15. D. Veeraiah and J. N. Rao, "An Efficient Data Duplication System based on Hadoop Distributed File System," 2020 International Conference on Inventive Computation Technologies (ICICT), 2020, pp. 197-200, doi: 10.1109/ICICT48043.2020.9112567.
16. Rao, J. Nageswara, and M. Ramesh. "A Review on Data Mining & Big Data." Machine Learning Techniques. Int. J. Recent Technol. Eng 7 (2019): 914-916.
17. Karthik, A., MazherIqbal, J.L. Efficient Speech Enhancement Using Recurrent Convolution Encoder and Decoder. Wireless PersCommun (2021). https://doi.org/10.1007/s11277-021-08313-6
18. S. Mittal and S. Tyagi, "Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection", 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2019.
19. S.Dutt, A.K.Das and S.Chandramouli, Machine Learning. Pearson Education India, 2018.
20. S. N. Ajani and S. Y. Amdani, "Probabilistic path planning using current obstacle position in static environment," 2nd International Conference on Data, Engineering and Applications (IDEA), 2020, pp. 1-6, doi: 10.1109/IDEA49133.2020.9170727.
21. S. Ajani and M. Wanjari, "An Efficient Approach for Clustering Uncertain Data Mining Based on Hash Indexing and Voronoi Clustering," 2013 5th International Conference and Computational Intelligence and Communication Networks, 2013, pp. 486-490, doi: 10.1109/CICN.2013.106.