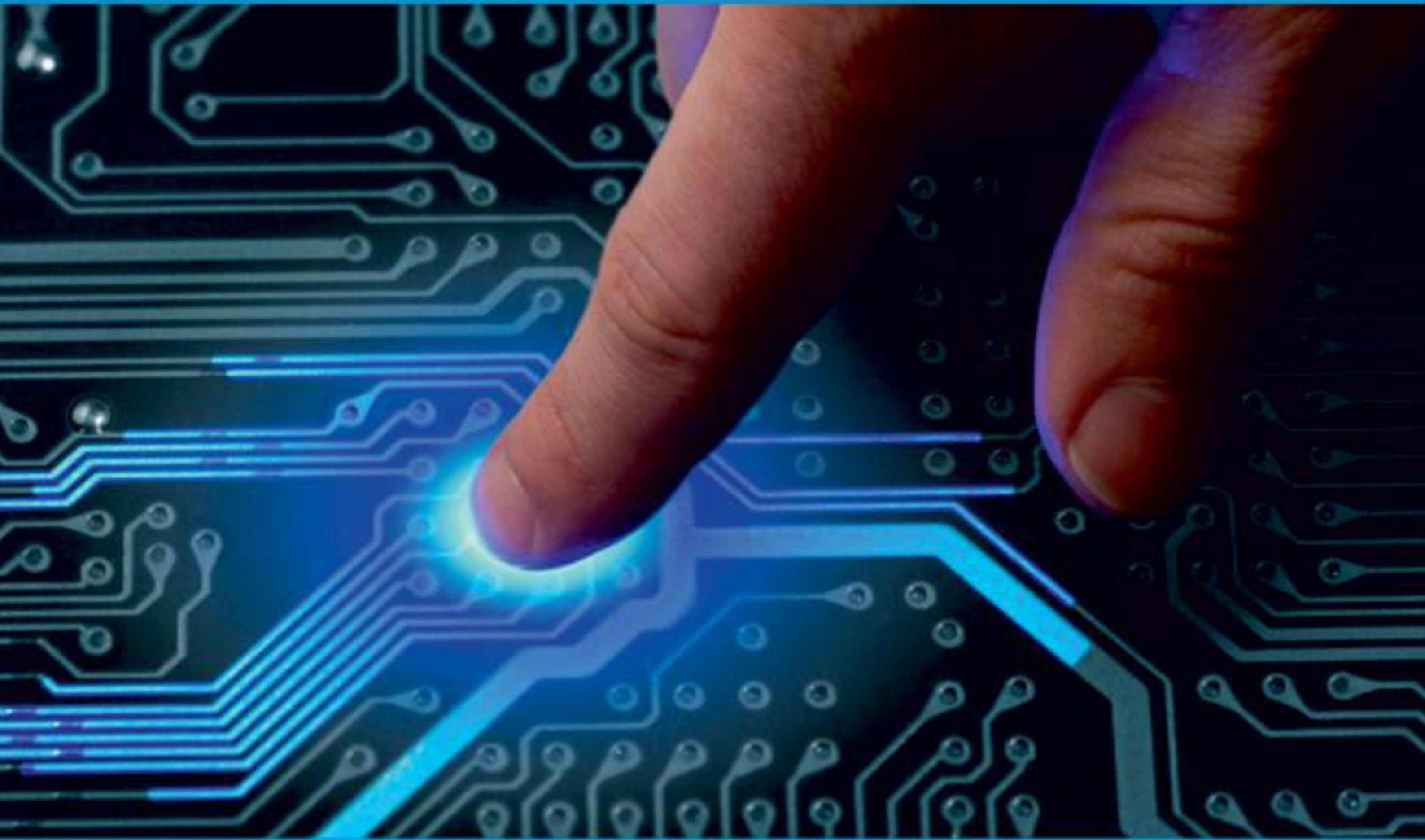
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 6, June 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006238|

Credit Card Fraud Detection using Machine Learning

Lasya TLV, Nagakodeeswari M, Mrs. Chitra P

U.G Scholar, Department of CSE, Velammal Institute of Technology, Chennai, Tamil Nadu, India

U.G Scholar, Department of CSE, Velammal Institute of Technology, Chennai, Tamil Nadu, India

Assistant Professor, Department of CSE, Velammal Institute of Technology, Chennai, Tamil Nadu, India

ABSTRACT: A credit card is issued by a bank or financial services company that allows cardholders to borrow funds with which to pay for goods and services with merchants that accept cards for payment. Nowadays as everything is made cyber so there is a chance of misuse of cards and the account holder can lose the money so it is vital that credit card companies are able to identify fraudulent credit card transactions so that customers are not charged for items that they did not purchase. This type of problems can be solved through data science by applying machine learning techniques. It deals with modelling of the dataset using machine learning with Credit Card Fraud Detection. In machine learning the main key is the data so modelling the past credit card transactions with the data of the ones that turned out to be fraud. The built model is then used to recognize whether a new transaction is fraudulent or not. The objective is to classify whether the fraud had happened or not. The first step involves analyzing and pre-processing data and then applying machine learning algorithm on the credit card dataset and find the parameters of the algorithm and calculate their performance metrics.

I. INTRODUCTION

Credit card fraud detection is the process of identifying purchase attempts that are fraudulent and rejecting them rather than processing the order. There are a variety of tools and techniques available for detecting fraud, with most merchants employing a combination of several of them. Credit companies have developed extremely sophisticated tools for detecting fraud. They monitor every transaction on every card. Then, credit card issuers are complicated computer algorithms to look for unusual transactions. It's possible to detect credit card fraud early by routinely checking for signs ofshady activity on your credit accounts. Review your card statement monthly, whether you get them online or in hard-copy form, looking carefully for unexpected purchases or cashadvances. Overhere credit card companies can track where yourstolencredit cardwas last used, in most cases, only once the card is used by the person who took it. The credit card authorization process helps bank's trace this. However, by the time law enforcement arrives, the person may be gone long gone. Enormous Data is processed every day and the model build must be fast enough to respond to the scam in time. Imbalanced Data i.e most of the transaction(99.8%) are not fraudulent which makes it really hard for detecting the fraudulent ones. Data availability as the data is mostly private. Misclassified Data can be another major issue, as not every fraudulent transaction is caught and reported. Adaptive techniques used against the model by the scammers. The model used must be simple and fast enough to detect the anomaly and classify it as a fraudulent transaction as quickly as possible. For protecting the privacy of the user the dimensionality of the data can be reduced. Unsupervised Machine Learning methods use unlabeled data to find patterns or dependencies in the credit card fraud detection, making it possible to group data samples by similarities without manual labeling. Today, We have many machine learning algorithms that ca help us classify abnormal transactions. The only requirement is the past data and suitable algorithm that can fit our data in better form. In this article, It will help you in the complete end-to-end model training process. Finally, we will get the best model that can classify the transaction into fraud or not fraud types. The technique is proposed for the development of an automatic fraud detection system (a) Data Pre-processing (b) Data Validation/ Cleaning/ Preparing process (c) Comparing algorithm with prediction in the form of best accuracy result (d) Deployment.

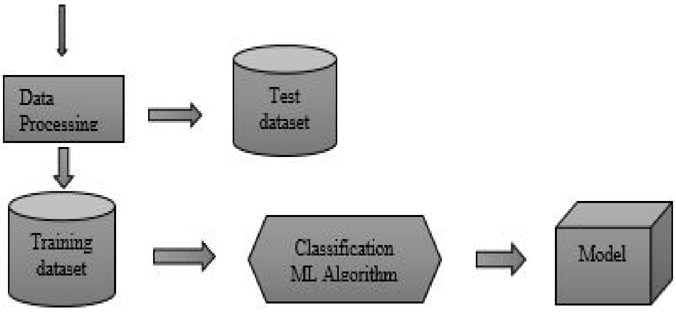
II. LITERATURE SURVEY

An Efficient Techniques for Fraudulent detection in Credit Card Dataset: Now a day, credit card transaction is one the famous mode for financial transaction. Increasing trends of financial transactions through credit cards also invite fraud activities that involve the loss of billions of dollars globally. It is also been observed that fraudulent transactions have increased by 35% from 2018. A huge amount of transaction data is available to analyze the fraud detection activities that require analysis of behavior/abnormalities in the transaction dataset to detect and ignore the undesirable action of the suspected person. The proposed paper lists a compressive summary of various techniques for the classification of fraud transactions from the various datasets to alert the user for such transactions. In the last decades, online transactions are growing rapidly and the most common tool for financial transactions. The increasing growth of online transactions also increases threats. Therefore, in keeping in mind the security issue, nature, an anomaly in the credit card transaction, the proposed work represents the summary of various strategies applied to identify the abnormal transaction in the dataset of credit card transaction datasets. This dataset contains a mix of normal and fraud transactions; this proposed work classifies and summarizes the various classification methods to classify the transactions using various Machine Learning-based classifiers. The efficiency of the method depends on the dataset and classifier used. The proposed summary will be beneficial to the banker, credit card user, and researcher to analyze to prevent credit card frauds. The future scope of this credit card fraud detection is to explore the things in each and every associations and banks to live safe and happily life. The data must be balanced in each place and we are getting the best results.

Credit Card Fraud Detection and Prevention using Machine Learning: This research focused mainly on detecting credit card fraud in real world. We must collect the credit card data sets initially for qualified data set. Then provide queries on the user's credit card to test the data set. After random forest algorithm classification method using the already evaluated data set and providing current data set[1]. Finally, the accuracy of the results data is optimised. Then the processing of a number of attributes will be implemented, so that affecting fraud detection can be found in viewing the representation of the graphical model. The techniques efficiency is measured based on accuracy, flexibility, and specificity, precision. The results obtained with the use of the Random Forest Algorithm have proved much more effective.

III. PROPOSED SYSTEM

The proposed model is to build a classification model to classify whether its fraud or not. The dataset of previous credit card cases are collected where it is used to make the machine to learn about the problem. The first step for involves the analysis of data where each and every column is analyzed and the necessary measurements are taken for missing values and other forms of data. Outliers and other values which are not much impact is dealt. Then preprocessed data is used to build the classification model where the data will be split into two parts one is for training and remaining data for testing purpose. Machine learning algorithms are applied on the training data where the model learns the pattern from the data and the model will deal with test data or new data and classify whether its fraud or not .The algorithms are compared and the performance metric of the algorithms are calculated.



Proposed System Design

Advantages:
→Performance and accuracy of the algorithms can be calculated and compared.
→Class imbalance can be dealt with machine learning approaches.

Data Pre-processing:

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters.

Data Validation/ Cleaning/Preparing Process:

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

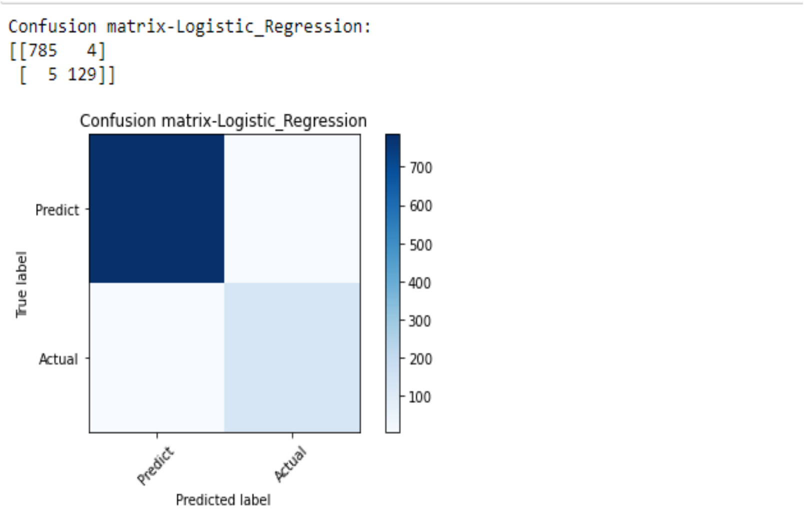
ALGORITHM AND TECHNIQUES

Algorithm Explanation:

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too.

a)Logistic Regression:

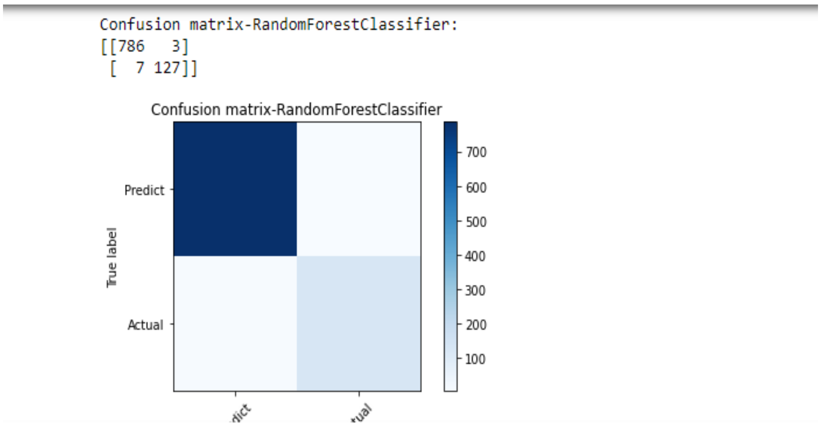
It is a statistical method for analysing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is independent (predictor or explanatory) variables. Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of



```
Confusion matrix-Logistic_Regression:
[[785   4]
 [  5 129]]
```

b)Random Forest Classifier:

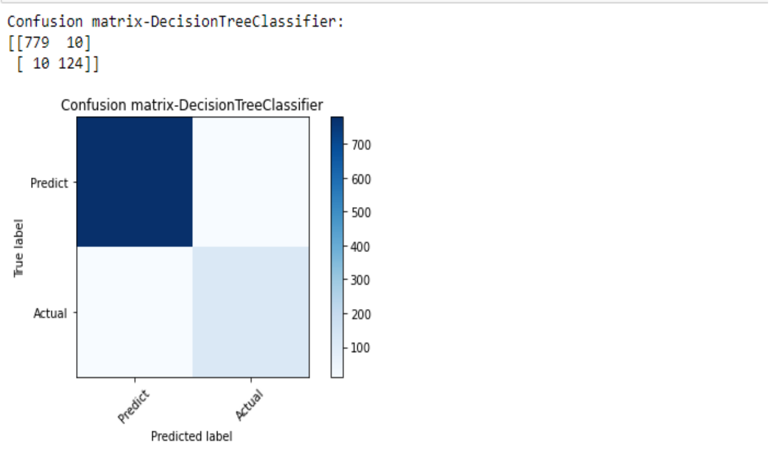
Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision *trees,* resulting in a *forest of trees,* hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.



```
Confusion matrix-RandomForestClassifier:
[[786   3]
 [  7 127]]
```

c)Decision Tree Classifier:

It is one of the most powerful and popular algorithm. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

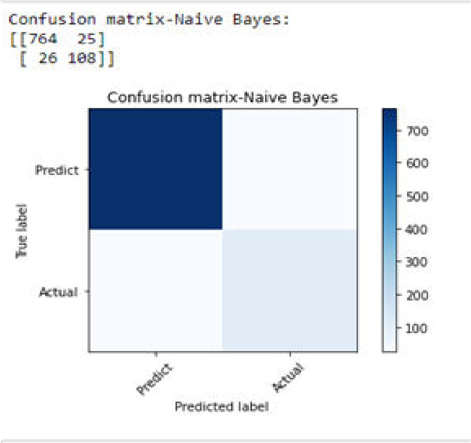


```
Confusion matrix-DecisionTreeClassifier:
[[779  10]
 [ 10 124]]
```

d)Naive Bayes Algorithm:

The Naive Bayes algorithm is an intuitive method that uses the probabilities of each attribute belonging to each class to make a prediction. It is the supervised learning approach you would come up with if you wanted to model a predictive modeling problem probabilistically. Naive bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method. The probability of a class value given a value of an attribute is called the conditional probability. By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class. To make a prediction we can calculate probabilities of the instance belonging to each classand select the class value with thehighest probability. Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms.Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets.

Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. For example, a loan applicant is desirable or not depending on his/her income, previous loan and transaction history, age, and location.



```
Confusion matrix-Naive Bayes:
[[764  25]
 [ 26 108]]
```

Deployment:
Flask (Web FrameWork) :

Flask is a micro web framework written in Python. It is classified as a micro-framework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

IV. CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score will be find out. This application can help to find the Prediction of credit card fraud or not.

V. FUTURE WORK

Credit card fraud prediction to connect with cloud model.
To optimize the work to implement in Artificial Intelligence environment.

REFERENCES

1. J. Wang, Y. Sun, Z. Zhang, and S. Gao, “Solving multitrip pickup and delivery problem with time windows and manpower planning using multiobjective algorithms,” IEEE/CAA J. AutomaticaSinica, vol. 7, no. 4, pp. 1134–1153, Jul. 2020.
2. Y. Liu, H. Ishibuchi, G. G. Yen, Y. Nojima, and N. Masuyama, “Handling imbalance between convergence and diversity in the decision space in evolutionary multi-modal multi-objective optimization,” IEEE Trans. Evol. Comput., vol. 24, no. 3, pp. 551–565, Jun. 2020.
3. X. Zhang, K. Zhou, H. Pan, L. Zhang, X. Zeng, and Y. Jin, “A network reduction-based multiobjective evolutionary algorithm for community detection in large-scale complex networks,” IEEE Trans. Cybern., vol. 50, no. 2, pp. 703–716, Feb. 2020.
4. Q. Wu, M. Zhou, Q. Zhu, Y. Xia, and J. Wen, “MOELS: Multiobjective evolutionary list scheduling for cloud workflows,” IEEE Trans. Autom. Sci. Eng., vol. 17, no. 1, pp. 166–176, Jan. 2020.
5. L. Huang, M. Zhou, and K. Hao, “Non-dominated immune-endocrine short feedback algorithm for multi-robot maritime patrolling,” IEEE Trans. Intell. Transp. Syst., vol. 21, no. 1, pp. 362–373, Jan. 2020.

6. Q. Kang, X. Song, M. Zhou, and L. Li, “A collaborative resource allocation strategy for decomposition-based multiobjective evolutionary algorithms,” IEEE Trans. Syst., Man, Cybern. Syst., vol. 49, no. 12, pp. 2416–2423, Dec. 2019.
7. Y. Feng et al., “Target disassembly sequencing and scheme evaluation for CNC machine tools using improved multiobjective ant colony algorithm and fuzzy integral,” IEEE Trans. Syst., Man, Cybern. Syst., vol. 49, no. 12, pp. 2438–2451, Dec. 2019.
8. Y. Cao, H. Zhang, W. Li, M. Zhou, Y. Zhang, and W. A. Chaovalitwongse, “Comprehensive learning particle swarm optimization algorithm with local search for multimodal functions,” IEEE Trans. Evol. Comput., vol. 23, no. 4, pp. 718–731, Aug. 2019.
9. X. Wang, K. Xing, C.-B. Yan, and M. Zhou, “A novel MOEA/D for multiobjective scheduling of flexible manufacturing systems,” Complexity, vol. 2019, pp. 1–14, Jun. 2019.
10. L. Ma, X. Wang, M. Huang, Z. Lin, L. Tian, and H. Chen, “Two-level master–slave RFID networks planning via hybrid multiobjective artificial bee colony optimizer,” IEEE Trans. Syst., Man, Cybern. Syst., vol. 49, no. 5, pp. 861–880, May 2019.
11. C. Yue, B. Qu, and J. Liang, “A multiobjective particle swarm optimizer using ring topology for solving multimodal multiobjective problems,” IEEE Trans. Evol. Comput., vol. 22, no. 5, pp. 805–817, Oct. 2018.
12. Credit Card Fraud Detection: A Realistic Modeling and a Novel
NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 8, AUGUST 2018
14. J. Liang, Q. Guo, C. Yue, B. Qu, and K. Yu, “A self-organizing multiobjective particle swarm optimization algorithm for multimodal multiobjective problems,” in Proc. 9th Int. Conf. Advances Swarm Intell. (ICSI), Shanghai, China, Jun. 2018, pp. 550–560.
15. “Credit Card Fraud Detection through Parenclitic Network Analysis-By Massimiliano Zanin, Miguel Romance, ReginoCriado, and
Article ID 5764370, 9 pages
17. Machine learning group — ULB, credit card fraud detection (2018), Kaggle
18. W. Dong and M. Zhou, “A supervised learning and control method to improve particle swarm optimization algorithms,” IEEE Trans. Syst., Man, Cybern. Syst., vol. 47, no. 7, pp. 1135–1148, Jul. 2017.
19. Credit card fraud detection based on transaction behavior -by john Richard D. Kho, larry A. Vea published by proc. Of the 2017 IEEE region 10 conference (TENCON), malaysia, november 5-8, 2017
20. G. Tian, M. Zhou, P. Li, C. Zhang, and H. Jia, “Multiobjective optimization models for locating vehicle inspection stations subject to stochastic demand, varying velocity and regional constraints,” IEEE Trans. Intell. Transp. Syst., vol. 17, no. 7, pp. 1978–1987, Jul. 2016.
21. “Credit Card Fraud Detection-by Ishu Trivedi, Monika, Mrigya, Mridushi” published by International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016
22. “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
23. L.J.P. Van der maaten and G.E. Hinton, visualizing high-dimensional Data using t-sne (2014), journal of machine learning research
24. J. J. Liang, S. T. Ma, B. Y. Qu, and B. Niu, “Strategy adaptative memetic crowding differential evolution for multimodal optimization,” in Proc. IEEE Congr. Evol. Comput., Jun. 2012, pp. 1–7.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING