Journal of Contemporary Issues in Business and Government Vol. 27, No. 3,2021
https://cibg.org.au/

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.220

CREDIT CARD FRAUD IDENTIFICATION USING MACHINE LEARNING ALGORITHM

[1]Vinutha H, [2]Amutharaj Joyson, [3]Apoorva J, [4]Ashitha G R, [5]B Tejashwini

[1,2,3,4,5]Department of Information Science and Engineering

RajaRajeswari College of Engineering, Bangalore, India
[1]vinuthamadhusudhana@gmail.com,[2]amutharajj@yahoo.com,
[3]apoorvajanardhan1998@gmail.com,[4]ashitha.gr3720@gmail.com,
[5]tejashwinibellad@gmail.com

ABSTRACT: Credit card extortion occasions happen often time and then bring about gigantic monetary misfortunes. Culprits can take advantage of some advancement, for example, phishing or Trojan to take another individual's credit card data. Furthermore, strong fraud identification technique is important to recognize the fraud in time when criminals use stolen cards. One strategy is to utilize historical transaction data to obtain normal and fraudulent transactions under the behavioural features of machine learning strategies, and the use of this feature is to verify whether the transaction is valid transaction or invalid transaction. This paper considers four different machine learning algorithms that are Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbor (KNN), and Random Forest to discipline the behavior of ordinary transactions and fraud features. Confusion matrix is utilized for estimating the performance analysis of the algorithms. Results obtained from the processing of datasets provide an accuracy of about 99-100%.

Keywords: Accuracy, Credit Card, Fraud Identification, Transaction, Prediction .hyperplane.

1. Introduction

Credit cards are widely utilized for e-business advancement and improvement of mobile smart devices. Cardless exchanges are increasingly popular, especially for all of the tasks performed by the credit card payment gateway web, for example, Alipay and PayPal. Credit cards have made online payments simple and more profitable. In any case, there is a growing pattern of fraudulent transactions

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.220

causing tremendous misfortune of cash annually. It has been estimated that misfortunes will be extended at binary digit figures in future. As a credit card is not physically required in the web-based payment conditions, the owner's card information is sufficient to settle any transaction making it simpler to steer misrepresentation than ever before. Credit card extortion has become a barrier to progress on a web based business and have a dramatic impact on the economy [2]. Therefore, identification of fraud is necessary and important.

Identification of fraud is the method used for observing the behaviour of exchange of a customer to identify whether an on-going transaction is completed by the customer or other. Eventually, there are two types of techniques for identifying fraud, anomaly detection and misuse detection. Anomaly detection builds a profile of ordinary exchanges of customers based on their past exchange information, and determine that the recent exchange is a fraudulent transaction if it varies from the typical exchange behavior. In Misuse detection, the gathered data is compared and analyzed with the huge dataset. Overall, this strategy should take into account current varieties of extortion for building frameworks by studying several misrepresentations patterns.

Machine learning techniques are used to analyze the official exchanges and unusual transactions are reported back. The professionals investigate this report and contact the cardholder to make confirmation if the transaction is valid or fraudulent. Based on this, the researchers provide feedback to the automated systems used for updating and training algorithms for the identification of fraud by ultimately improving their performance over time. This process can be categorized into unsupervised learning and supervised learning [1]. The purpose of the paper is to identify invalid transaction accurately by applying machine learning algorithms such as KNN, SVM, Naive Bayes, and Random Forest algorithm..

2. Literature review

A broad understanding of fraud identification technology might be useful for solving the problem of card-not-present fraud. Various literature works pertaining to fraud detection have been published already and are accessible for open use. The authors in [3] proposed the random forest classification algorithm to analyze the original data set and the current user dataset, and evaluate the accuracy of the resulting data.

The comprehensive research conducted by S P Maniraj and his associates has revealed that card-not-present misrepresentation identification problem includes modeling of historical master card exchanges with transaction information of the people who turn out as frauds. They focus on deploying some anomaly detection techniques namely transformed credit card data transactions, Isolation Forest

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.220

algorithm and Local Outlier Factor on Principal Component Analysis (PCA) and also perform the analysis and pre-processing of data sets[4].

Similar work was carried out in [5] where an intelligent fraud identification model was proposed to identify fraudulent card-not-present exchanges from huge dataset which was anonymous and highly imbalanced. The problem of Class imbalance was described by looking for patterns in fraudulent as well as non-fraudulent transactions for each cardholder by making use of frequent item set mining. A convenient method was presented for identifying to which arrangement(fraudulent/non-fraudulent) the arriving exchange of a cardholder was nearer to and based on this the decision would be made.

A completely new architecture for online and offline card-not-present transactions was proposed in [6]. This architecture was based on trust-building and mutual authentication between a merchant and a customer for web based transactions by integrating a digital signature and one-time-password method using public key infrastructure system.

In [7], the authors focused on fraud detection in real time and presented a new way to understand the spending patterns to decipher the possible cases of fraud. It used a self-organizing map to decipher, filter, and analyze the conduct of the client to detect the occurrence of misrepresentation. In a fraud identification solution, the key purpose is to restrict the number of incorrectly classified exchanges. However, incorrect classification of each exchange does not have the same effect as in if the fraudsters have the card then they will use its entire available limit. For the method of solution, a new collaboration of two meta-heuristic methods, scatter search and genetic algorithms can be suggested. Ultimately the application of this method on original data yielded successful results [8].

Andrea Dal Pozzolo et al. [9] proposed a conception for the problem of fraud identification which described the working prerequisites of FDSs that analyzed huge streams of card-not-present exchanges every day. The overall performance measure to be used for the fraud identification purpose was also illustrated. The experiments demonstrated the affect of the concept drift and class unbalance on a real-world information stream consisting of about seventy five million exchanges.

According to Sahil Dhankhad et al.[10], various supervised machine learning algorithms can be utilized for identifying invalid and valid card transactions using original datasets. Moreover, a super classifier ensemble learning method can be implemented by employing these machine learning algorithms. It makes it easier to identify the highest accuracy providing variables and compare the performance of each algorithm.

The existing research work carried by different researchers on the credit card misrepresentation identification case studies, in which normalization of raw data

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.220

input is done using unsupervised learning algorithms such as Cluster Analysis and the results of these techniques on detection of fraud can be shown with attribute grouping. Input neurons can be reduced and the outcomes of each can be calculated using the normalized raw input. Similarly, researches were done using various supervised machine learning algorithms such as Naïve Bayes classifier, Random forest, KNN, and SVM along with Neural Network for regression and classification of dataset and the dataset is classified using a confusion matrix. An analysis is done over these categorized data and the results on fraud detection were obtained. The significance of this paper was to develop a new method that combines different supervised algorithms for misrepresentation identification and to compare the accuracy of algorithms with each other.

3. Machine Learning Algorithms for Credit Card Fraud Detection

This research study considers four different machine learning algorithms namely Support Vector Machine (SVM), Random Forest, K-Nearest Neighbour (KNN), and Naïve Bayes to train the behavioral features of fraudulent as well as normal exchanges. Support Vector Machine, Random forest, and Naive Bayes are used for the process of regression and classification whereas the K-Nearest Neighbour algorithm is used for clustering. First step is to collect the dataset of the customer card. The dataset used in this study is based on real-time transaction information shared by a China-based company. The analysis will be performed on a dataset collected followed by cleaning the dataset. Generally, it is required to clean the information to eliminate all the null and duplicate values existing in the dataset. Once the cleaning process is completed, the dataset is divided into two different categories namely Test dataset and Training dataset. After categorizing the dataset into two types, the machine learning algorithms are applied for classifying the dataset. This classification provides performance analysis which is used to provide accuracy about the fraudulent transactions. Each algorithm provides different accuracy and this accuracy will be depicted in the form of graphical representation. Figure 1 depicts the proposed Machine Learning algorithm based credit card fraud detection.

3.1. Random Forest Algorithm

Random Forest is a technique that is fit for performing both regression and classification tasks. They consist of a huge number of individual decision trees that collectively operate as an ensemble. One of the well-known examples of supervised learning is random forest algorithm and it is a classification algorithm that uses bagging and feature randomness techniques. The key advantage of this technique is that it can be utilized for Regression as well as Classification. The random forest algorithm used in this experiment can identify a fraudulent

Journal of Contemporary Issues in Business and Government Vol. 27, No. 3,2021
https://cibg.org.au/

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.220

transaction with an accuracy of about 99%. Figure 2 depicts the Random Forest Algorithm.

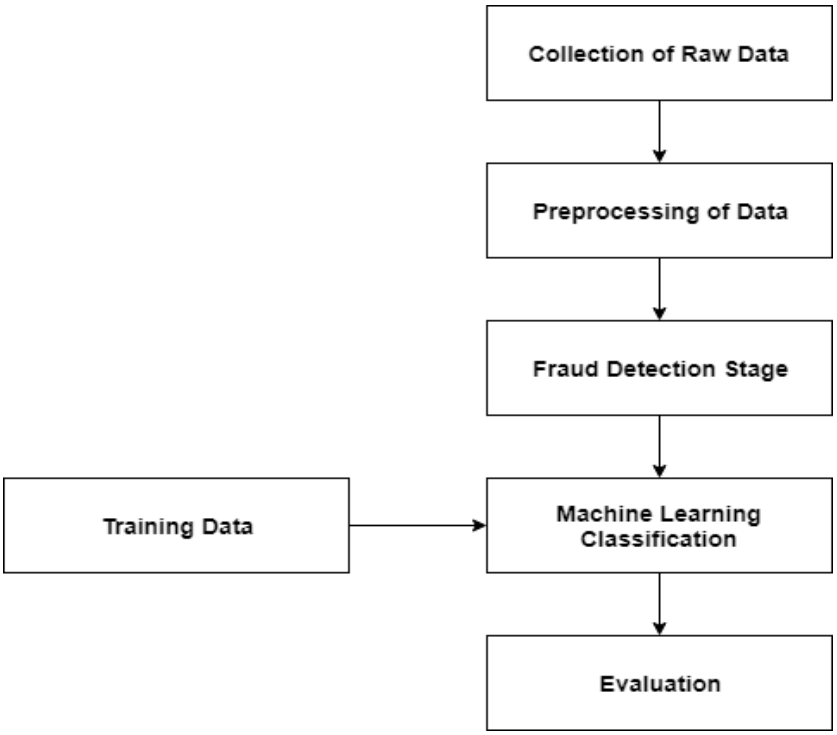

Figure 1: Machine Learning algorithm based credit card fraud detection

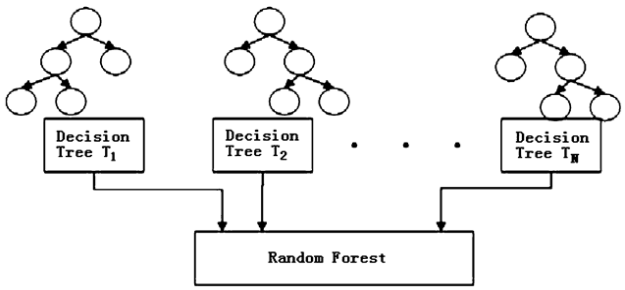

Figure 2: Random Forest

Random Forest Algorithm follows the following steps:

1. Load the dataset.
2. Label the data as training and testing set.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.220

3. Train the classifier using Random Forest algorithm.
4. Then fit the Random Forest model with the data.
5. Total count of fraudulent cases and valid transactions is done.
6. Finally percentage of accuracy is computed.

Table 1 illustrates the results obtained by using Random Forest algorithm for credit card fraud detection.

Table 1: Results obtained by using the Random Forest Algorithm

Number of fraudulent transactions	492
Number of valid transactions	284315
Precision score	0.9238095238095239
Recall score	0.8016528925619835
Accuracy score	0.9995505744220669

3.2. K-Nearest Neighbor (KNN) Algorithm

The K-Nearest Neighbor (KNN) method classifies the new instances on the measure of similarity. It belongs to instance-based learning. The new instances are classified to the respective class by using Euclidian formula. The new instances are grouped with the neighbor instance which has less distance compared to other neighbors. In this paper, k value is used has 5 for the classification of new instances and by varying k value, the results obtained using this algorithm may also vary respectively. Results obtained using KNN algorithms are presented in Table 2.

Table 2: Results obtained for KNN algorithm

Number of fraudulent transactions	492
Number of valid transactions	284315
Precision score	0.7777777777777778
Recall score	0.06140350877192982
Accuracy score	0.9984691441251651

3.3. Naïve Bayes Classifier

Naïve Bayes (NB) is one of the supervised learning methods which use the labeled data. In this technique, the system is trained using both the input and respected output values. Later the model is tested by giving input as the test data. Based on the correct predictions made by the model using Naïve Bayes classifier

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.220

we can determine the efficiency of the system. It mainly uses the Bayes theorem for classification and further uses the concept of conditional probability. The independence of the attributes of training examples is used as the major feature in Naïve Bayes method. Results obtained by using Naïve Bayes Classifier are depicted in Table 3.

Table 3: Results obtained by using Naïve Bayes Algorithm as Classifier

Number of fraudulent transactions	492
Number of valid transactions	284315
Precision score	1.0
Recall score	1.0
Accuracy score	1.0

3.4. Support Vector Machine (SVM)

The Support Vector Machines (SVM) are the statistical techniques of machine learning which was first introduced in 1995 by Cortes and Vapnik and were very useful in the classification tasks. The SVM mainly includes concepts of decision planes that define the decision boundaries in a graph. The set of different classes are separated using this decision planes. Later the decision plane segregates the sample data into negative and positive classes, thus the SVM algorithm constructs a hyperplane for the classification of samples instead of the decision plane. Representation of the kernel and optimization of the margin is the two major features of this method. Hyperplane is used to separate the different classes in the training examples. To provide the better separation between the different classes, maximum margin hyperplane is used. The nearest instance to the maximum margin hyperplane is known as support vectors. There can be one or more support vectors associated to the class. This model has the greatest accuracy and efficient classification of instances compared to other machine learning algorithms. A large training dataset is required to achieve maximum prediction accuracy in the case of SVM. Even for multidimensional data and continuous featured data, SVMs methods are one of the first choices made for the classification and regression process of this dataset.

Table 4: Results obtained for Support Vector Machine Algorithm

Number of fraudulent transactions	492
Number of valid transactions	284315
Precision score	1.0
Recall score	1.0
Accuracy score	1.0

Journal of Contemporary Issues in Business and Government Vol. 27, No. 3,2021
https://cibg.org.au/

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.220

4. Experimental Set up and Analysis

This study proposed four different machine learning algorithms namely Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN), and Naïve Bayes. Real time dataset of credit card transactions from Chinese Company is collected which contains all essential information regarding the credit card transactions. This study does not consider the parameters such as amount and time for the further analysis and pre-processing of the dataset. Then followed by the methods such as scaling, normalization, binarization, standardization, and data labeling of the dataset values. The dataset records are divided into Test dataset which is of 20% of the actual dataset and trained dataset which is of 80% of the actual dataset. The next step is applying the machine learning algorithms for classifying the dataset. Later the accuracies of these algorithms in detecting the fraudulent transactions are compared along with the bar graph which provides the clear representation of the valid and fraudulent transactions present in the dataset. The System Architecture is depicted in Figure 3.

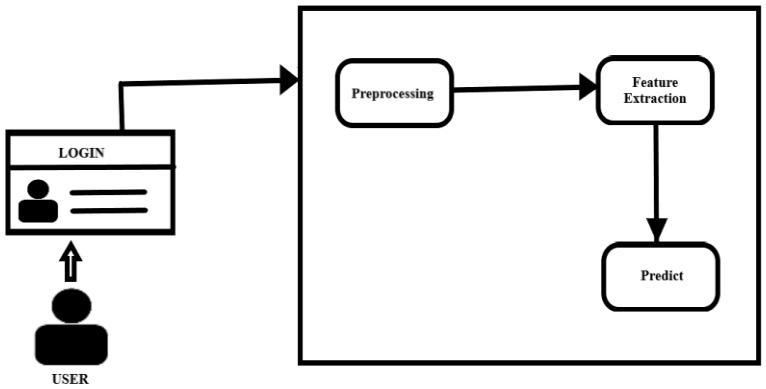

Figure 3: System Architecture

4.1. Dataset Collection

The dataset used in this paper has been taken from an e-commerce based company from China. It consists of a total of 31 columns. The sensitive user information is stored from 1 to 28 columns which are represented as V1 to V28. The time, class and amount are the other columns of the dataset. The time gap between the successive credit card transactions is tabulated in Time column of the dataset. The last column represents the class in the dataset and the class 0 is referred as valid transactions and class 1 is referred as the in-valid or fraudulent transactions.

4.2. Data Pre-processing

Data pre-processing consists mainly of three basic steps: Formatting, Cleaning, and Sampling. Formatting involves organizing the data according to the required specification. Cleaning is the method of removing null values or fixing of missing

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.220

data. Sampling involves taking a smaller representative sample of data and working with it individually. During the Data pre-processing, the dataset will be cleaned from raw data by following 5 steps:-

1. Scaling: The raw data is scaled from 0 to 1.
2. Normalization: Re-scaling the data to have the desired length.
3. Binarization: The process of converting raw data into binary values.
4. Standardization: The dataset values are transformed into numerical values.
5. Data Labeling: It is used to label the type of data.

4.3. Feature Extraction

Feature Extraction is an attribute reduction method. It is a process in which the initial dataset is reduced to manageable groups of data for processing. The transformed attributes are considered to be the linear combination of the original attributes. Significantly, the model is trained to classify the pre-processed values using the four different machine learning algorithms mentioned in this paper. The feature extraction mainly includes the extraction of numerical values from the cardholder information and there are a total of 28 extracted features from the dataset.

4.4. Prediction

The final step in the implementation part is to predict the fraudulent and valid transactions along with the bar graph representation for better visualization. Class 1 contains fraudulent transactions and class 0 contains valid transactions. By implementing the four different algorithms used in this paper, a prediction is made using the test data in the dataset and the accuracy of each algorithm is obtained.

4.5. Sequence diagram

Sequence diagram for the Credit Card Fraud Identification is given in Figure 4. There are three main modules involved in the identification of credit card fraud/valid transactions. They are:-

4.5.1. Registration Module

Registration module includes the signing up of the user by entering the details such as username, password, e-mail, and phone number. These details will be stored in a database in MySQL software.

4.5.2. Login Module

Once the registration process is successful, the user can log in to the system by entering the username and password. If the details entered are correct then the user can successfully login or else the login session will fail.

Journal of Contemporary Issues in Business and Government Vol. 27, No. 3,2021
https://cibg.org.au/

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.220

4.5.3. Prediction Module

This is the third module where the dataset is being imported using python code and all the four algorithms are applied on the dataset to analyze the fraud and valid transactions. The graphical representation of the fraud and valid transactions along with the accuracy of each algorithm is predicted which the final outcome of this study.

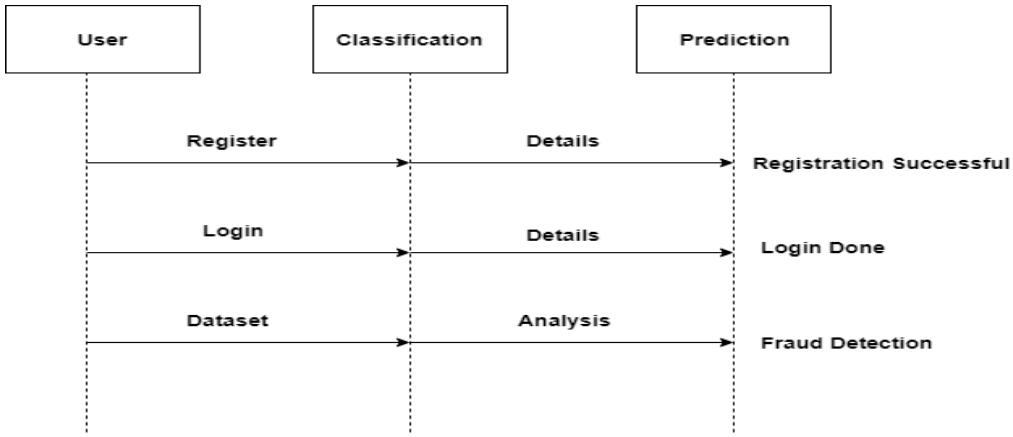

Figure 4: Sequence Diagram for modules of Credit Card fraud Identification

5. Result and Discussions

This research study analyzed the results obtained from four different machine learning algorithms namely Random Forest algorithm, SVM algorithm, KNN algorithm, Naïve Bayes algorithm. These machine learning algorithms classify the transactions in to two classes such as fraudulent transactions (class 1) and non-fraudulent transactions (class 0).

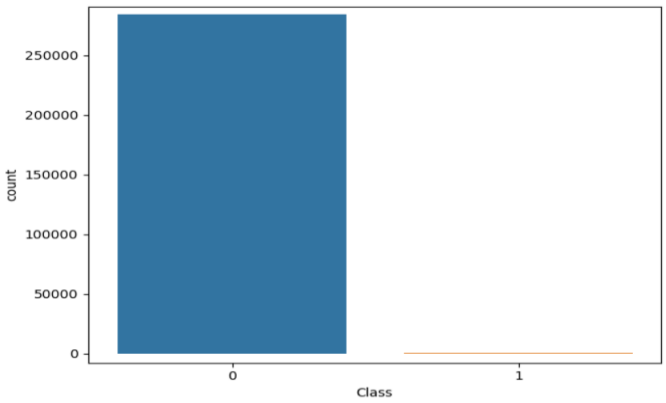

Figure 5: Graphical Representation of valid and fraud transaction

The experimental results collected by implementing four different algorithms were collected and classified transactions are plotted in the bar graph which represents class (class 0 and class 1) transactions on the x-axis and the number of records on the y-axis.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.220

6. Conclusion

The Credit card Fraud Identification system helps us to find out the number of fraudulent and non-fraudulent credit card transactions by making use of machine learning techniques such as Random Forest, SVM, KNN, and Naïve Bayes Classifier. The results of these algorithms are compared to obtain the highest accuracy. SVM and Naïve Bayes algorithms are proved to provide the highest accuracy for the chosen dataset compared to the other algorithms. The graphical representation of the fraud and valid transactions is provided for better visualization and easy identification of the fraud transactions.

References

[1] M. Suresh Kumar, V. Soundarya, S. Kavitha, E. S. Keerthika, and E. Aswini. "Credit Card Fraud Detection Using Random Forest Algorithm", IEEE International Conference on Communication Systems and Network Technologies IEEExplore, **(2019)**, pp.149-153.

[2] ShiyangXuan, Guanjun Liu, Zhenchuan Li, LutaoZheng, Shuo Wang, and Changjun Jiang. "Random Forest For Credit Card Fraud Detection", International Conference on Networking, Sensing, and Control IEEE, **(2018).**

[3] Devi Meenakshi B, Jeanne B, Gayathri S, Indira N. "Credit Card Fraud Detection Using Random Forest", International Research Journal of Engineering and Technology, (2019), pp.6662-6666.

[4] AdityaSaini, Swarna Deep Sarkar, and Shadab Ahmed. "Credit Card Fraud Detection Using Machine Learning And Data Science", International Journal of Engineering Research & Technology (IJERT), **(2019)**, pp.110-115.

[5] K. R. Seeja and Masoumeh Zareapoor. "Fraud Miner: A Novel Credit Card Fraud Detection Model Based on Frequent Itemset Mining." , The Scientific World Journal, **(2014)**..

[6] Shalini Gupta and Rahul Johari. "A New Framework for Credit Card Transactions involving Mutual Authentication between Cardholder and Merchant", International Conference on Communication Systems and Network Technologies, **(2011),** pp.22-26.

[7] Jon T.S. Quah and M. Sriganesh. "Real-time Credit card Fraud Detection using Computational Intelligence" , Expert Systems with Applications, Science Digest, **(2008),** pp.1721–1732.

[8] EkremDuman and M. HamdiOzcelik. "Detecting credit card fraud by genetic algorithm and scatter search." Expert Systems with Applications 38 **(2011),** pp.13057–13063

[9] Andrea Dal Pozzolo, GiacomoBoracchi, Olivier Caelen, CesareAlippi, and GianlucaBontempi. "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy." International Conference On Neural Networks and

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.220

Learning Systems IEEE, **(2018).**

[10] SahilDhankhad, Emad Mohammed, and BehrouzFar."Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study." International Conference on Information Reuse and Integration(IRI) IEEE**, (2018).**

Authors

Mrs.Vinutha H received her Bachelor of Engineering degree from VTU University, Belgaum, India in 2010 and Master of Technology degree from VTU University, Belgaum India in 2016. She is currently working as an Asst. Professor in the Department of Information Science and Engineering, RajaRajeswari College of Engineering, Bangalore, India. She has ten years of experience in teaching. Her research interests include Image Processing, Data mining, Internet of Things, Network security.

Dr. Amutharaj Joyson received his Bachelor of Engineering degree from Manonmaniam Sundaranar University, Tirunelveli, India in 1999 and Master of Engineering degree from Madurai Kamaraj University, Madurai in 2002, and Ph.D from Anna University, Chennai, India in 2012. He is currently working as a Professor and Head in the Department of Information Science and Engineering, RajaRajeswari College of Engineering, Bangalore, India. His research interests include Internet of Things, Healthcare Applications, Network security, Cellular Automata and Cloud Computing. Dr. Amutharaj Joyson is a reviewer of International Journal of Network and Computer Applications (JNCA), Computer and Communication, Elsevier Publications. He has served as a Technical Programme Committee member and reviewer at various Conferences including the IEEE International Conference on Electrical, Electronics, Communication, Computer Technologies & Optimization Techniques (ICEECCOT-2018), IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing(INCOS-'19). He is a Fellow of the Institution of Engineers (IE), Kolkatta, India, Fellow of Institution of Electronics and Telecommunication Engineers, New Delhi, lifetime member of Computer Society of India (CSI), Mumbai and life member of ISTE, New Delhi.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.220

Apoorva. J completed her Bachelor of Engineering in Information Science and Engineering from Visveswayara Technological University, Belagavi in 2020. She has undergone industry Internship at ADE, branch of DRDO organization and at LIVEWIRE in 2019. She is one of the organizing committee members of International Women's Day Celebration on 8th March 2019 organized by Rajarajeswari Groups of institutions, Bangalore. It is recorded in the Limca Book of Records for the largest human image of Karnataka map in 2 colours, most women taking Oath on women Empowerment and Most women displaying messages on Women Empowerment.

Ashitha G R completed her Bachelor of Engineering in Information Science and Engineering from Visveswaraya Technological University, Belagavi in 2020. She has undergone industry Internship at BESCOM in 2019. She is one of the organizing committee members of International Women's Day Celebration on 8th March 2019 organized by Rajarajeswari Groups of institutions, Bangalore. It is recorded in the Limca Book of Records for the largest human image of Karnataka map in 2 colours, most women taking Oath on women Empowerment and Most women displaying messages on Women Empowerment.

B Tejashwini completed her Bachelor of Engineering in Information Science and Engineering from Visveswayara Technological University, Belagavi in 2020. She has undergone industry Internship at LIVEWIRE in 2019. She is one of the organizing committee members of International Women's Day Celebration on 8th March 2019 organized by Rajarajeswari Groups of institutions, Bangalore. It is recorded in the Limca Book of Records for the largest human image of Karnataka map in 2 colours, most women taking Oath on women Empowerment and Most women displaying messages on Women Empowerment.