The Performance Analysis of Machine Learning Algorithms for Credit Card Fraud Detection

https://doi.org/10.3991/ijoe.v19i03.35331

Muhammad Zohaib Khan, Sarmad Ahmed Shaikh(✉), Muneer Ahmed Shaikh,
Kamlesh Kumar Khatri, Mahira Abdul Rauf, Ayesha Kalhoro, Muhammad Adnan
Department of Computer Science, Sindh Madressatul Islam University, Karachi, Pakistan
sarmad@smiu.edu.pk

Abstract—This paper studies the performance analysis of machine learning (ML) and data mining techniques for anomaly detection in credit cards. As the usage of digital money or plastic money grows in developing nations, so does the risk of fraud. To counter these scams, we need a sophisticated fraud detection method that not only identifies the fraud but also detects it before it occurs efficiently. We have introduced the notion of credit card fraud and its many variants in this research. Numerous ML fraud detection approaches are studied in this paper including Principal Component Analysis (PCA) data mining and the Fuzzy C-Means methodologies, as well as the Logistic Regression (LR), Decision Tree (DT), and Naive Bayes (NB) algorithms. The existing and proposed models for credit card fraud detection have been thoroughly reviewed, and these strategies have been compared using quantitative metrics including accuracy rate and characteristics curves. This paper discusses the shortcomings of existing models and proposes an efficient technique to analyze the fraud detection.

Keywords—PCA, Fuzzy C-Means, Logistic Regression, Decision Tree, Naive Bayes algorithms

1 Introduction

There have always been those who would develop new ways to illegally access someone's funds since the inception of e-commerce payment platforms. This has become a significant issue in the present day, as all transactions can be readily accomplished online by simply inputting your credit card information. Because the numbers are projected to rise in the future, many academics in this subject are focusing on detecting fraudulent behavior early by using powerful machine learning (ML) methods [1]. Several ML approaches are actively in use, especially in industry area. Some of the approaches require supervised learning techniques while others require un-supervised. Supervised learning technique works on structured or labeled data. After that, supervised learning takes what it has learned from previous data and applies it to updated data (new data). It provides model training data, and the machine predicts whether it is a circle or a square, for instance. This supervised learning method is used for subcategory categorization and regression. Unsupervised learning uses unstructured or concealed data and requires the machine to be fed with a variety of inputs. Unsupervised learning examines information before classifying it; as a result, after classification, we obtain different data in each group, and each group is distinct from the others [2].

The systematic approach of grouping mechanisms into distinct groups and subcategories based on their commonalities is known as classification. In ML, there are several types of categorizations i.e., Linear Regression, Nave Bayes Classifier, and Linear Classifier. Structured and tagged data are included in classifications. Linear and nonlinear regression techniques are based on supervised and unsupervised learning approaches, respectively, since various regression models differ based on the kind of relationship between dependent and independent variables. It carries out regression tasks. The ML methods employ a variety of regression characteristics, including structured and unstructured data. Both linear and nonlinear regression techniques include regression model features one and two. Apart of this, clustering is the most frequent type of unsupervised learning which has a wide range of applications particularly in businesses. It is the process of separating and executing information on behalf of the information machine, which results in a collection of data that we label with a unique identifier (ID).

The credit card fraud (CCF) detection using ML technique is a method in which data science investigates the provided data and develops a model that will deliver the best outcomes in detecting and preventing fraudulent transactions. This is accomplished by aggregating all the relevant information of card users, transactions, such as date, user zone, product category, amount, provider, client's behavioral patterns, and so on. The data is then fed into a slightly trained model that looks for patterns and rules to determine if a transaction is fraudulent or lawful. To achieve this task, generally, two steps are catered i.e., data mining and patterns recognition. Data mining is used to categorize, aggregate, and segment data in order to scan millions of transactions for trends and detect fraud. On the other hand, patterns recognition entails spotting suspicious behavior classes, clusters, and patterns. In this context, ML refers to the selection of a model or combination of models that best fits a certain business challenge. The neural networks technique of ML, for example, assists in automatically identifying the traits most commonly presented in fraudulent transactions; this method is most successful if you have a large number of transaction samples [3–5].

2 Literature review

The CCF seems to be increased, though as a result, financial losses are growing dramatically. Every year, as a result of fraud, billions of dollars are lost. There is a dearth of studies to assess the scam. To identify real-world credit card fraud, a variety of machine learning techniques are used. Many researchers want to discover viruses, anomalies, and farads in IoT devices as early as feasible. Sumaya Sanober in [6] proposed Deep Learning (DL) and ML techniques, i.e., the random forest method, to classify CCF detection. The rate of success was discovered to be 96.2%. Other study included in [7] anticipated an improved technique such as Decision Tree (DT) to predict the credit card fraud. The experimental success rate was 97.93%. John O. Awoyemi in [8], proposed the advanced fraud detection of credit card using the model and Logistic Regression algorithm. The performance achieved was 98% as success rate. The author, Hassan Najadat in [9], employed the classification prediction model with logistic regression algorithm and improved the credit card fraud detection based on DL and ML techniques. The rate of success was 80% in this study. In another study in [10], the authors used the DL technique i.e., ODAE for the fraud detection in credit card transactions. They achieved 84.1% success rate. Moreover, the authors in [11], gained utmost highest outcome of 91.48% by using the random forest algorithm of ML.

Furthermore, the paper in [12] used the ML prediction model and implementation approach through the K-Nearest Neighbors (KNN) algorithm and achieved 91.11% accuracy results. Additionally, the study in [13] anticipated the focus on the classifier model i.e., DT in addition to other ML algorithms. The success rate reported in this study with the implementation is 89.91%. Abhimanyu Roy, in [14], worked on the DL method as well as LSTM algorithm for the CCF detection and showed the outcome as 91.2% success rate. In [15], the authors analyze the different ML techniques on the European cardholders and employed the Naïve Bayes (NB) algorithm for detection/categorization and reached 97.92% accuracy. Manjeevan et al in [16] implemented an intelligent card fraud detection system using the genetic algorithm (GA) for feature selection and aggregation. The researchers implemented several ML algorithms to validate the effectiveness of their proposed method. The results demonstrated that their proposed method attained an accuracy of 81.97%.

Most of the reported techniques have achieved the performance of CCF detection with complex structures and the performance still needs to be improved. In this paper, we detect fraudulent transactions by examining both supervised and unsupervised learning methods. As described in detail in the following section, we use the classifying technique Principal Component Analysis (PCA) and Fuzzy C-Means (FCM), as well as prediction techniques including Logistic Regression (LR), Decision Tree (DT), and Naive Bayes (NB) algorithms.

3 Proposed methodology

This section describes the proposed method which begins with an algorithm, subsequently load data from a database, and finally prepare the information. Data is preprocessed, normalized, and PCA is implemented when the data preparation procedure is done. PCA analyses data in the same manner as a database does, then performs PCA analysis and determines the result. Afterwards, we establish supervised classification for the credit card fraud prediction model using the Random Forest Algorithm and the FCM Clustering algorithm for unsupervised clustering (because the FCM clustering algorithm cannot transform label data and clean the FCM cluster result). That's because combining diverse strategies yielded more impressive results. To design and construct the goal model, the features of PCA, FCM, LR, DT, and NB Supervised classification algorithm will be applied. The sample size for the test is 25%, whereas the sample size for the implementation training is 75%. Figure 1 depicts a flowchart that can assist in comprehending the research activity and outlining the job breakdown structure concisely. In this flowchart, initially we retrieve data form the database, then prepare data, start preprocessing the data in data preprocessing block, standardize data and then use the PCA techniques to obtain PCA stop and repeat PCA. Moreover, PCA's reduced 3-Dimensional data is used to reduce the dimension of the data and then clustering and classification algorithms are applied, implementing 75% processed trained data and 25% test data for the model validation.

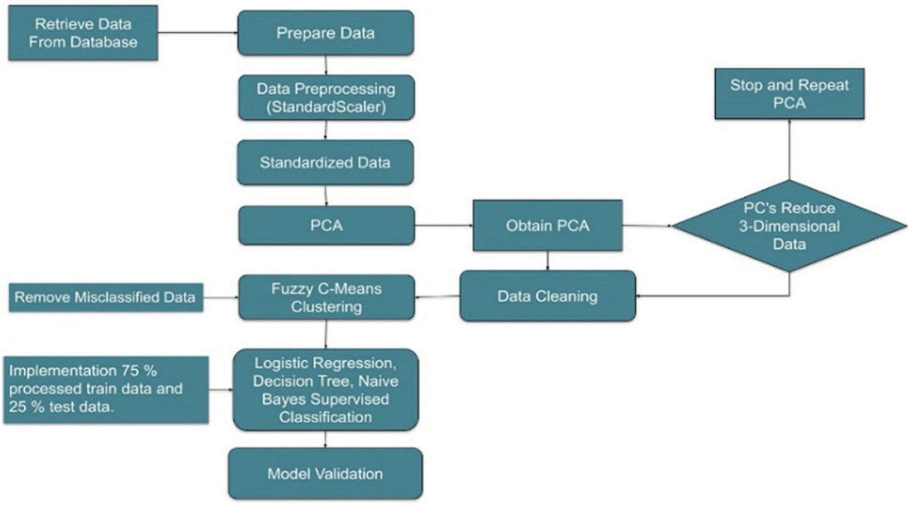
Retrieve Data From Database
Prepare Data
Data Preprocessing (StandardScaler)
Stop and Repeat PCA
Standardized Data
PCA
Obtain PCA
PC's Reduce 3-Dimensional Data
Remove Misclassified Data
Fuzzy C-Means Clustering
Data Cleaning
Implementation 75 % processed train data and 25 % test data.
Logistic Regression, Decision Tree, Naive Bayes Supervised Classification
Model Validation

Fig. 1. Proposed method for credit card fraud detection

The proposed method consists of the following two subsections: 1) Pre-processing of data, and 2) Classification.

3.1 Preprocessing

In preprocessing phase, we clean the data and use it for clustering in order to get useful data. For this, we investigate two well-known methods i.e., PCA and FCM clustering, described below, and then further process the data in the classification phase.

Principal Component Analysis (PCA). PCA is a dimension reduction approach that is extensively used to deal with large quantities of data predictions by splitting many variables into small pieces and combining numerous data from information machines into a massive collection of data. Attempting to limit the number of components in the test set affects accuracy obviously. Still, the issue with lowering dimensionality is that it sacrifices a little inflexibility for convenience. Because smaller bits of data are easy to inspect and reproduce, data assessment is likely less demanding and faster for ML models with free components to test. PCA commonly employs preprocessing, depreciating measurement reduction, covariance and correlation eigenvalues, and eigenvectors techniques. More detail about PCA can be found in [17] while the main mathematical steps are given below.

$$\text{Standardization} \quad z = \frac{x-\mu}{\sigma} \tag{1}$$

$$\begin{gathered} \textit{For Population} \\ \text{Covariance} \quad Cov(x,y) = \frac{\sum |x_i - x| * |y_i - y|}{N} \\ \textit{For Sample} \\ Cov(x,y) = \frac{\sum |x_i - x\ | * |y_i - y\ |}{(N-1)} \end{gathered} \tag{2}$$

$$\text{Eigenvalues \& Eigenvectors} \quad Av - \lambda v = 0; (A - \lambda I)\, v = 0 \tag{3}$$

The method has a significant contribution in reducing the redundant characteristics that are worthless for grouping. PCA provides better performance over the FCM (described below) because it minimizes the number of variables in the original data set making it easier to deal with ambiguous or mislabeled data. The primary benefit of PCA is that it becomes an essential process in calculating the number of clusters as well as providing a conceptual mathematical model to model the structure of the sets once we have identified these principal components from the data. Table 1 briefly describes the characteristics and variables included in the implementation dataset for this study, which was fraud detection in banking credit cards. Table 2 shows the cleaned and preprocessed data using PCA algorithm. The dataset considered in this work for fraud detection in credit cards has been taken from the UCI machine-learning source (OpenML), as specified in Table 1. The statistics include card transactions done by European cardholders in September 2013. This dataset consists of 492 frauds out of 284,807 transactions that happened over the course of two days. The dataset is significantly imbalanced, with positive transactions accounting for 0.172 percent of all trades [18]. The dataset for detecting fraud in credit card transactions was gathered from photos of real and fake banknote-like specimens. An industrial camera, typically used for print inspection, was utilized for digitalization. The images are 400×400 pixels in size. Grayscale images with a resolution of roughly 660 dpi were obtained as a result of the object lens and distance to the researched item. To extract features from these photos, a Wavelet Transform tool was employed. The key attributes of dataset are as follows: (1) **V1.** The variance of Wavelet Transformed image (continuous), (2) **V2.** The skewness of Wavelet Transformed image (continuous), (3) **V3.** Kurtosis of Wavelet Transformed image (continuous), and (4) **V4.** The entropy of the image (continuous).

Table 1. Original dataset for fraud detection in banking credit card

Time	**V1**	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**	**V9**	**V10**
0	–1.359807	–0.072781	2.536347	1.378155	–0.338321	0.462388	0.239599	0.098698	0.363787	0.090794
0	1.191857	0.266151	0.166480	0.448154	0.060018	–0.082361	–0.078803	0.085102	–0.255425	–0.166974
1	–1.358354	–1.340163	1.773209	0.379780	–0.503198	1.800499	0.791461	0.247676	–1.514654	0.207643
1	–0.966272	–0.185226	1.792993	–0.863291	–0.010309	1.247203	0.237609	0.377436	–1.387024	–0.054952
2	–1.158233	0.877737	1.548718	0.403034	–0.407193	0.095921	0.592941	–0.270533	0.817739	0.753074
V11	**V12**	**V13**	**V14**	**V15**	**V16**	**V17**	**V18**	**V19**	**V20**	**V21**
–0.55160	–0.617801	–0.991390	–0.311169	1.468177	–0.470401	0.207971	0.025791	0.403993	0.251412	–0.018307
1.612727	1.065235	0.489095	–0.143772	0.635558	0.463917	–0.114805	–0.183361	–0.145783	–0.069083	–0.225775
0.624501	0.066084	0.717293	–0.165946	2.345865	–2.890083	1.109969	–0.121359	–2.261857	0.524980	0.247998
–0.226487	0.178228	0.507757	–0.287924	–0.631418	–1.059647	–0.684093	1.965775	–1.232622	–0.208038	–0.108300
–0.822843	0.538196	1.345852	–1.119670	0.175121	–0.451449	–0.237033	–0.038195	0.803487	0.408542	–0.009431
V22	**V23**	**V24**	**V25**	**V26**	**V27**	**V28**	**Amount**	**Target Class**		
0.277838	–0.110474	0.066928	0.128539	–0.189115	0.133558	–0.021053	149.62	0		
–0.638672	0.101288	–0.339846	0.167170	0.125895	–0.008983	0.014724	2.69	0		
0.771679	0.909412	–0.689281	–0.327642	–0.139097	–0.055353	–0.059752	378.66	0		
0.005274	–0.190321	–1.175575	0.647376	–0.221929	0.062723	0.061458	123.50	0		
0.798278	–0.137458	0.141267	–0.206010	0.502292	0.219422	0.215153	69.99	0		

Table 2. Preprocessed dataset for fraud detection in banking credit cards

Time	**V1**	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**	**V9**	**V10**
–1.996583	–0.694242	–0.044075	1.672773	0.973366	–0.245117	0.347068	0.193679	0.082637	0.331128	0.083386
–1.996583	0.608496	0.161176	0.109797	0.316523	0.043483	–0.061820	–0.063700	0.071253	–0.232494	–0.153350
–1.996562	–0.693500	–0.811578	1.169468	0.268231	–0.364572	1.351454	0.639776	0.207373	–1.378675	0.190700
–1.996562	–0.493325	–0.112169	1.182516	–0.609727	–0.007469	0.936150	0.192071	0.316018	–1.262503	–0.050468
–1.996541	–0.591330	0.531541	1.021412	0.284655	–0.295015	0.071999	0.479302	–0.226510	0.744326	0.691625
V11	**V12**	**V13**	**V14**	**V15**	**V16**	**V17**	**V18**	**V19**	**V20**	**V21**
0.083386	–0.618296	–0.996099	–0.324610	1.604014	–0.536833	0.244863	0.030770	0.496282	0.326118	–0.024923
1.580003	1.066089	0.491418	–0.149982	0.694360	0.529434	–0.135170	–0.218763	–0.179086	–0.089611	–0.307377
0.611830	0.066137	0.720700	–0.173114	2.562906	–3.298235	1.306868	–0.144790	–2.778561	0.680975	0.337632
–0.221892	0.178371	0.510169	–0.300360	–0.689837	–1.209296	–0.805445	2.345305	–1.514205	–0.269855	–0.147443
–0.806147	0.538627	1.352244	–1.168034	0.191323	0.191323	–0.279081	–0.045569	0.987037	0.529939	–0.012839
V22	**V23**	**V24**	**V25**	**V26**	**V27**	**V28**	**Amount**	**Target Class**		
0.382854	–0.176911	0.110507	0.246585	–0.392170	0.330892	–0.063781	0.244964	–0.041599		
–0.880077	0.162201	–0.561131	0.320694	0.261069	–0.022256	0.044608	–0.342475	–0.041599		
1.063358	1.456320	–1.138092	–0.628537	–0.288447	–0.137137	–0.181021	1.160686	–0.041599		
0.007267	–0.304777	–1.941027	1.241904	–0.460217	0.155396	0.186189	0.140534	–0.041599		
1.100011	–0.220123	0.233250	–0.395202	1.041611	0.543620	0.651816	–0.073403	–0.041599		

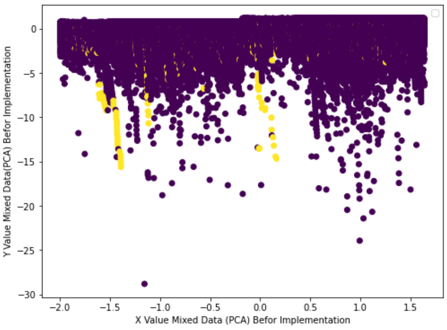

Fig. 2. The mixed data chart before applying PCA algorithm

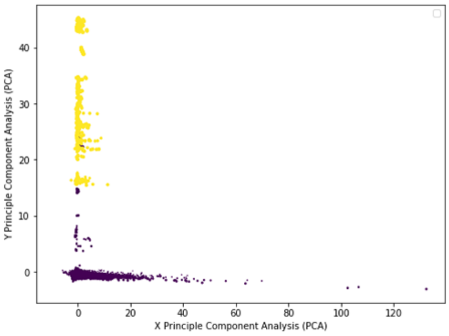

Fig. 3. The achieved data clusters after applying PCA algorithm

After training the machine algorithm over the designed module, the execution was catered as above. The Figures 2 and 3 illustrate the clustering data before and after applying the PCA algorithm, respectively, using the X-axis and Y-axis Component Analysis approach to obtain visualized data and determine the alternative outcomes.

Fuzzy C-Means (FCM) clustering algorithm. The Fuzzy C-means (FCM) clustering technique is most extensively used for data clustering purpose. FCM is a soft clustering approach in which each data point is assigned a probability or likelihood score indicating whether or not it belongs to that cluster. It does this by sequentially creating new uncorrelated variables that maximize variance. This logic is immediately applied to the data matrix to generate a membership matrix that displays the degree of the link between the samples and each cluster. In essence, this approach employs a clustering technique, as evidenced by the centroid clustering values, the FCM algorithm, and a three-dimensional dataset. This approach gives participation to each data point corresponding to each centroid based on the distance between the centroid and the data point. The closer the data is to the cluster centroid, the more it belongs to it. Each data point's actual number should clearly be one [19]. The following formula is used to adjust participation and cluster centroid after each cycle:

$$\mu_{ij} = 1/\sum\nolimits_{k=1}^{c}\left(d_{ij}/d_{ik}\right)^{(2/m\text{-}1)}, \tag{4}$$

$$v_j = \left(\sum_{i=1}^{n}(\mu_{ij})^m x_i\right) \Big/ \left(\sum_{i=1}^{n}(\mu_{ij})^m\right), \forall_j = 1, 2, 3, \ldots, c, \tag{5}$$

where, 'n', 'v_j', 'm', 'c', 'μ_{ij}', and 'd_{ij}' denote the number of data points, cluster center, the fluffiness directory, the number of cluster centers, the connection between the ith data and the jth cluster center, and the Euclidean distance between the ith data and the jth cluster center, respectively.

The main objective of the FCM strategy is to reduce the Euclidean distance given below:

$$J(\mathrm{U},\mathrm{V}) = \sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{c} (\mu_{ij})^m \left\| x_i - v_j \right\|^2 . \tag{6}$$

The Euclidean distance between both the *i*th data point and the *j*th cluster center is represented by '$||x_i - v_j||$'. There are some mathematical functions which are used for any FCM algorithm i.e., Euclidean, Manhattan, and Hamming, as given below.

$$\text{Euclidean} \quad \sqrt{\sum\nolimits_{i=1}^{k} (x_i - y_i)^2} \tag{7}$$

$$\text{Manhattan} \quad \sum\nolimits_{i=1}^{k} \left| x_i - y_i \right| \tag{8}$$

$$\text{Minkowski} \quad \left(\sum\nolimits_{i=1}^{k} (|\ x_i - y_i\ |)\ q \right) 1/q \tag{9}$$

In this way, let $X = \{x1, x2, x3\ldots, xn\}$ represent a set of data points and $V = \{v1, v2, v3\ldots, vc\}$ represent a set of centers, then the steps of the FCM are as follows.

1. Randomly select 'c' cluster centroids
2. O use the following formula to calculate the fuzzy participation 'μ_{ij}':

$$\mu_{ij} = 1/\sum\nolimits_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)} \tag{10}$$

3. Compute the fuzzy centroids 'v_j' as follows:

$$v_j = \left(\sum_{i=1}^{n} (\mu_{ij})^m x_i \right) / \left(\sum_{i=1}^{n} (\mu_{ij})^m \right), \forall_j = 1, 2, 3, \ldots, c \tag{11}$$

4. Repeat steps 2 and 3 until the smallest '*j*' value is attained or $\|U(k+1) - U(k)\| < \beta$ is achieved, whereas the repetition step is signified by the letter '*k*.', the completion criteria between [0, 1] is 'β', the fuzzy relationship matrix is well-defined as '$U = (ij)$ $n*c$', and the impartial function is represented by the letter '*J*.'

The fraud detection in banking credit card and clusters dataset is explicitly defined as a 3-dimensional dataset, with three features based on banking credit card fraud detection values and one feature target property cluster number (see Table 3). The FCM clustering centroid value briefly defined is given below, whereas Figure 4 shows the three clusters after the unstructured dataset has been converted to structured data. This graph depicts two clusters: red and green, with centroid values specified by the yellow star (*). Figure 5 shows FCM determined sum of squared error line chart.

FCM Clustering Centroid Value.

```
array [[–0.14096411, –0.36526464, 1.20874051],
       [–0.05878056, 0.27174926, –1.02626197]]
```

Table 3. FCM two clusters preprocessed fraud detection in banking credit card dataset

S/No.	1-Dimension	2-Dimension	3-Dimension	Clusters
0	0.417977	0.708503	–2.443670	1
1	–0.392663	0.567814	–1.996839	1
2	1.873758	0.727602	–2.461528	1
3	0.314127	0.463391	–1.710503	1
4	–0.004095	0.405264	–1.433439	1
…	**…**	**…**	**…**	**…**
284804	–0.194457	–0.554986	1.978335	0
284806	0.763326	–0.605939	1.814907	0
284807 rows × 4 columns				

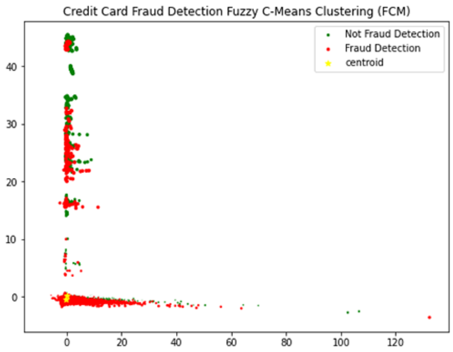

Fig. 4. The FCM two clusters

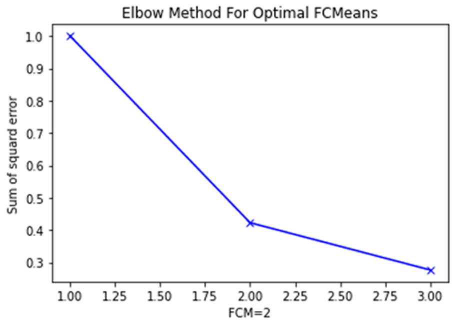

Fig. 5. The FCM sum of squared error line chart

Performance of CFD using both clustering techniques is compared using precision, Recall, and f-measure. A suspicion score is calculated according to the extent of deviation from the normal patterns and thereby, the transaction is classified as legitimate or suspicious or fraudulent.

3.2 Classification

The Classification algorithm is a supervised learning approach that uses training data to determine the category of observations. Classification is the process of learning from a dataset or observations and then classifying the observations into one of many classes or groups. We evaluate the following classification algorithms in our dataset and figure out which one is performing better.

Logistic regression algorithm. Logistic regression (LR) is a ML technique that is used for classification methods. It is a predictive analytic approach that is based on the probability hypothesis. A logistic regression model is comparable to a linear regression model. However, the LR utilizes a more sophisticated cost function, which may be characterized as the 'sigmoid function' or sometimes known as the 'logistic function' rather than a linear function. The LR hypothesis tends to restrict the differential equation between 0 and 1. Consequently, linear functions fail to describe it since they can have a value larger than one or even less than 0, which is not feasible according to the LR assumption [20]. These are some scientific approaches or functions that can be used for any algorithm. This approach, like the Sigmoid, Linear, Cost Linear, and Non-linear LR, is based on classification. Some of these approaches are as follows:

$$\text{Sigmoid} \quad S(z) = \frac{1}{(1+e^{-z})} \tag{12}$$

$$\text{Linear LR} \quad y = e \wedge (b0 + b1 * x) / (1 + e \wedge (b0 + b1 * x)) \tag{13}$$

$$\text{Cost Linear LR} \quad \begin{matrix} (Cost(h\theta\ (x), y)) = -log(h\theta\ (x)), if\ y = 1\, and \\ (Cost(h\theta\ (x), y)) = -log(1 - h\theta\ (x)), if\ y = 0 \end{matrix} \tag{14}$$

$$\text{Nonlinear LR} \quad \mathrm{Y} = \mathrm{f}(X, \beta) + \varepsilon \tag{15}$$

On a synthetic dataset, we can show this by plotting the Receiver Operating Characteristic (ROC) curve for a no-skill classifier and a LR model. We used the LR Classifier model accuracy and model loss concepts to assess our model in this study. This will increase the approximated prediction approach's accuracy while also ensuring that the patterns of credit-card fraud detection prediction are fulfilled on a frequent basis. Figure 6 shows the obtained ROC curve of LR. Moreover, we used the LR Algorithm model accuracy and model loss concepts to assess our model in this study. This will increase the approximated prediction approach's accuracy while also ensuring that the patterns of credit-card fraud detection prediction are fulfilled on a frequent basis. Figure 7 shows the model accuracy of trained data which is 0.999 and the tested was 0.998 in the LR algorithm. While the model loss of the trained data is 0.023 and the tested is 0.022 in the LR algorithm.

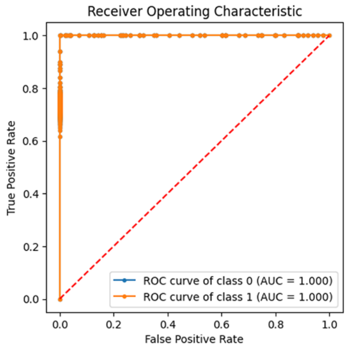

Fig. 6. LR Receiver Operating Characteristic (ROC) Curve

Note: The positive rate of the ROC curve = 1.000.

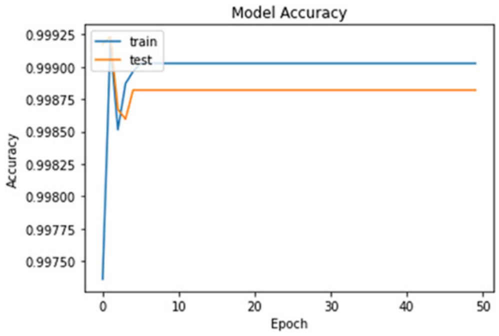

Fig. 7. Model accuracy LR algorithm

Decision tree algorithm. The Decision Tree (DT) algorithm is a branch of the supervised learning algorithm category. The DT approach can also be implemented to overcome regression and classification applications. The objective of implementing a choice tree is to build a training model that predicts the category or values of the target attribute by learning basic decision rules from previous results (training data). In DT, we begin at the root of the tree to forecast a class label for a record. The values of the root attribute are compared to the importance of the record's quality. Based on the comparison, we follow the branch corresponding to that value and proceed to the next node. These are some scientific methods or functions that can be applied to any algorithm [21]. This method, like the Information Gain and Entropy, is dependent on the creation of a DT and categorization. These approaches are mathematically given as follows.

Information Gain $$\text{Information Gain} = E(Y) - E(Y|X) \tag{16}$$

Entropy $$Entropy(s) = -P(+)log2\,P(+) - P(-)log2\,P(-) \tag{17}$$

The main steps of the DT algorithm can be found in [21]. We apply the DT algorithm on our dataset and label the old data values. It predicts the values of data, where we try to make the predictions fit the labels during preparation with the use of DT algorithm. In this way, Figure 8 demonstrates a binary classifier system's diagnostic performance when its discriminating threshold is modified. ROC analysis is inextricably linked to cost/benefit analysis in diagnostic decision-making. The achieved accuracy of the DT trained data, as shown in Figure 9, is 0.9989 and the tested data is 0.9999. Moreover, the model loss trained data is 0.025 and the tested data is 0.035.

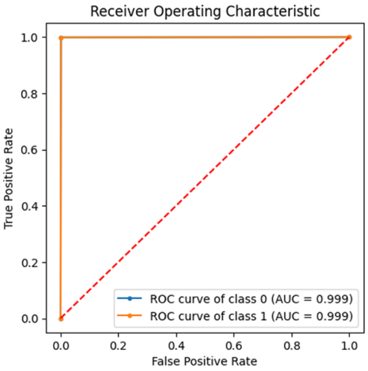

Fig. 8. Receiver Operating Characteristic (ROC) curve of DT algorithm

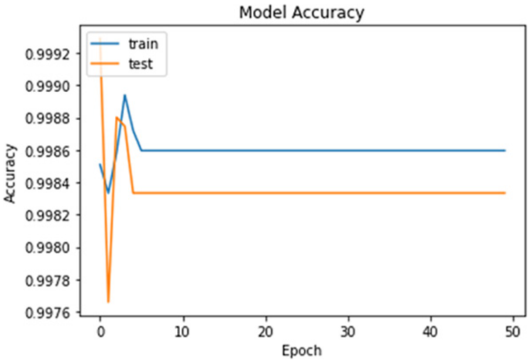

Fig. 9. Model accuracy of DT algorithm

Naïve Bayes Algorithm. The Naïve Bayes (NB) method is a supervised learning method that solves classification issues and is based on the Bayes theorem. It is primarily utilized in text categorization with a large training dataset, and it is a probabilistic and effective classification method that aids in the development of rapid machine learning models capable of making quick predictions. The following is the mathematical expression of the algorithm.

Bayes Theorem Formula $$P(A \mid B) = \frac{P(B \mid A).\,P(A)}{P(B)} \tag{18}$$

The probability of an incidence is calculated by the NB classifier in the steps given in [22]. We apply the NB algorithm on our above dataset and label the old data values. It predicts the value of data – where we tried to make our predictions fit the labels during preparation with the use of the NB algorithm. Furthermore, Figure 10 shows ROC analysis which is inextricably linked to cost/benefit research in diagnostic decision-making. This will gain the accuracy of the estimated prediction method while frequently meeting the patterns of card prediction. The achieved accuracy of the trained data using NB algorithm, shown in Figure 11, is 0.9979 and for the tested data is 0.9983. Moreover, the model loss of the NB algorithm for the trained data is 0.025 and the tested data is 0.026.

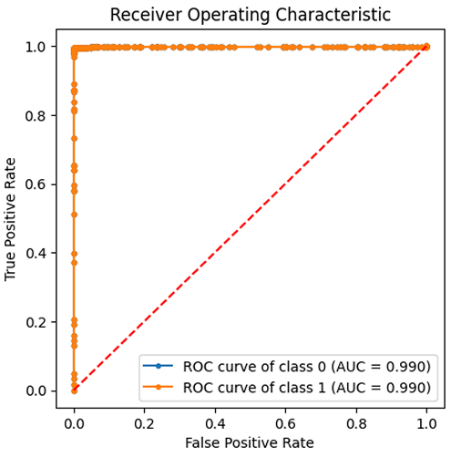

Fig. 10. NB – Receiver Operating Characteristic (ROC) curve

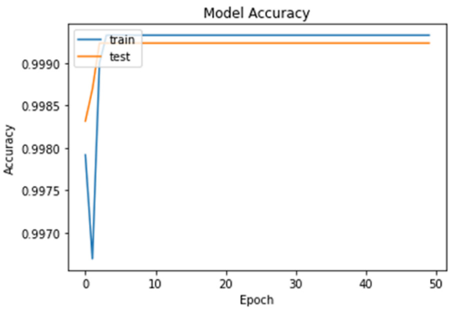

Fig. 11. Model accuracy of NB Algorithm

4 Experimental setup

Python is a high-level scripting language that is mainly used for general purpose programming and ML techniques, web development, and databases. We have used the Python tool's Anaconda Navigator ->Jupiter Notebook GUI framework for whole simulation work. We used the Python programming language to link datasets and perform PCA, FCM, Logistic Regression, Decision Tree, and Naive Bayes algorithms. Basically, the dataset is for detecting banking credit card fraud. On 284807, there are 31-dimensional characteristics (columns) and tuples (rows) in this dataset. We simulated three separate programs, the first of which used PCA, FCM, and LR techniques, the second of which used PCA, FCM, and DT method, and the third of which used PCA, FCM, and NB procedure, and all of which utilized the same dataset. On the personal computer, all of these apps are running. The following is the computer's configuration:

- Second Generation Intel (R) Core (TM) i5-2520M CPU @ 2.50 GHz.
- RAM of 4.00 GB.
- The system is a 64-bit operating system.
- Windows 10 (Home).
- 500 GB hard disk.

5 Results and discussions

The ML is a scientific technique where computers learn how to solve a problem without explicitly programming them. Deep learning is currently leading the ML race powered by better algorithms, computation power, and large data. Still, ML classical algorithms have a strong position in the field. This paper uses a new approach based on the integration between dimensionality reduction PCA with FCM. Then, a comparative study of the performance was performed, including other supervised ML algorithms, to reach the best classifier. These ML supervised techniques include LR, DT, and NB in this study. Here, we combine the algorithms to check the most accurate combination for the credit card fraud detection. The combinations of the algorithms are as follows: PCA – FCM – Logistic Regression; PCA – FCM – Decision Tree; and PCA – FCM – Naïve Bayes. The obtained accuracy and other performance parameters are provided in Tables 4 and 5, respectively.

Table 4. Model accuracy for the combination of algorithms for banking credit card fraud detection

Combination of Algorithms	Accuracy (%)
PCA – FCM – Logistic Regression	99.994382
PCA – FCM – Decision Tree	99.882025
PCA – FCM – Naïve Bayes	99.066037

Table 5. Parameter score for the combination of algorithms for banking credit card fraud detection

S/No.	Parameter Score (%)	PCA – FCM – Logistic Regression	PCA – FCM – Decision Tree	PCA – FCM – Naïve Bayes
1	Accuracy	0.99994382	0.99882025	0.99066037
2	Precision	0.99994382	0.99882025	0.99066037
3	Recall	0.99994382	0.99882025	0.99066037
4	Roc Auc	0.99993765	0.99880563	0.98965034
5	Sensitivity	1.0	0.99865868	0.99989772
6	Specificity	0.99987530	0.99895258	0.97940296
7	F1-Score	0.99994382	0.99882025	0.99066037

As shown in the observed results in Figures 12 and 13, it is clear that the combination PCA – FCM – Logistic Regression has reached at its maximum outcome in terms of accuracy. However, the combination PCA – FCM – Decision Tree is ranked second, followed by combination consisting of PCA – FCM – Naïve Bayes as third ranked. We may limit or alter the precision as per the requirement. For instance, the Parameter Score Accuracy, Precision, Recall, Roc Auc, Sensitivity, Specificity, and F1-Score are at their best correctness.

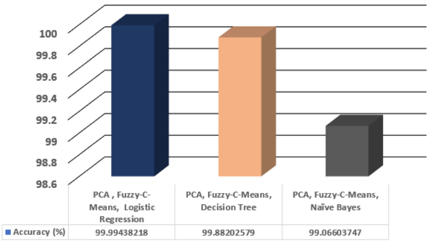

Fig. 12. Model accuracy for various combination of algorithms for banking credit card fraud detection

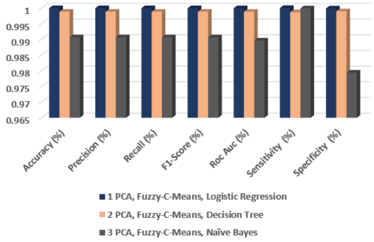

Fig. 13. Parameter score for various combinations of algorithms for banking credit card fraud detection

6 Conclusion

This study examined the performance of ML algorithms for credit card fraud detection. Credit card firms must be able to detect fraudulent transactions. The datasets containing the credit card transactions made by cardholders were analyzed using a variety of algorithms, including the FCM, PCA, LR, DT, and NB algorithms. The most efficient and accurate result is achieved by combining the algorithms PCA – FCM – Logistic Regression. We used fraud detection data sets for classification, regression, and clustering. To enhance the accuracy, we employed PCA data mining and the FCM technique. By employing the different algorithm combinations i.e., PCA – FCM – logistic regression; PCA – FCM – decision tree; and PCA – FCM – Nave Bayes, we achieved an accuracy of 99.99%, 99.88%, and 99.06%, respectively, as a consequence of each method being coupled with the PCA. As a result, it is concluded that the combination of PCA – FCM – logistic regression provides improved results.

7 References

[1] Stands with Ulkraine, Credit-Card-Fraud-Detection, https://spd.group/machine-learning/credit-card-fraud-detection/

[2] Machine Learning Mastery, Supervised-And-Unsupervised-Machine-Learning-Algorithms, https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

[3] Fatma Nasoz, Laxmi Gewali, Justin Zhan, "Application of Machine Learning Techniques in Credit Card Fraud Detection", https://digitalscholarship.unlv.edu/cgi/viewcontent.cgi?article=4457&context=thesesdissertations

[4] Datavedas, Classification-Problems, https://www.datavedas.com/wp-content/uploads/2018/05/3.1.1.2-CLASSIFICATION-PROBLEMS-1.png

[5] Researchgate, Document Clustering, https://www.researchgate.net/profile/Laith-Abualigah/publication/322455242/figure/fig1/AS:582970224644096@1516002340891/An-example-of-the-document-clustering.png

[6] Sumaya Sanober, Izhar Alam, Sagar Pande, Farrukh Arslan, Kantilal Pitambar Rane, Bhupesh Kumar Singh, Aditya Khamparia, Mohammad Shabaz, "An Enhanced Secure Deep Learning Algorithm for Fraud Detection in Wireless Communication", https://doi.org/10.1155/2021/6079582

[7] Yashvi Jain, NamrataTiwari, ShripriyaDubey, Sarika Jain, "A Comparative Analysis of Various Credit Card Fraud Detection Techniques", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277–3878, Volume-7 Issue-5S2, January 2019.

[8] John O. Awoyemi, Adebayo O. Adetunmbi, Samuel A. Oluwadare, "Credit Card Fraud Detection using Machine Learning Techniques: A Comparative Analysis", 978-1-5090-4642-3/17/$31.00 ©2017 IEEE.

[9] Hassan Najadat, Hassan Najadat, Ayah Abu Aqouleh, Mutaz Younes, "Credit Card Fraud Detection Based on Machine and Deep Learning", 2020 11th International Conference on Information and Communication Systems (ICICS). https://doi.org/10.1109/ICICS49469.2020.239524

[10] Zahra Kazemi, Houman Zarrabi, Using Deep Networks for Fraud Detection in the Credit Card Transactions, 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), 978-1-5386-2640-5/17/$31.00 ©2017 IEEE. https://doi.org/10.1109/KBEI.2017.8324876

[11] Maja Puh, Ljiljana Brkic, "Detecting Credit Card Fraud Using Selected Machine Learning Algorithms", MIPRO 2019, May 2019, p. 20–24, Opatija Croatia. https://doi.org/10.23919/MIPRO.2019.8757212

[12] Khatri S, Arora, Agrawal AP, "Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison", In: 10th International Conference on Cloud Computing, Data Science & Engineering (Confuence); 2020. p. 680–683. https://doi.org/10.1109/Confluence47617.2020.9057851

[13] Emmanuel Ileberi, Yanxia Sun, Zenghui Wang, "A Machine Learning Based Credit Card Fraud Detection Using the GA Algorithm for Feature Selection", Ileberi et al. Journal of Big Data (2022) 9:24. https://doi.org/10.1186/s40537-022-00573-8

[14] Awoyemi JO, Adetunmbi AO, Oluwadare SA, "Credit Card Fraud Detection Using Machine Learning Techniques: A Comparative Analysis". In: International Conference on Computer Networks and Information (ICCNI); 2017. p. 1–9. https://doi.org/10.1109/ICCNI.2017.8123782

[15] Abhimanyu Roy, Jingyi Sun, Robert Mahoney, Loreto Alonzi, Stephen Adams, Peter Beling, "Deep Learning Detecting Fraud in Credit Card Transactions", 978-1-5386-6343-1/18/$31.00 ©2018 IEEE.

[16] Seera M, Lim CP, Kumar A, Dhamotharan L, Tan KH, "An Intelligent Payment Card Fraud Detection System", Ann Oper Res 2021, p. 1–23. https://doi.org/10.1007/s10479-021-04149-2

[17] Analytics Vidhya, Principal Component Analysis (PCA), https://medium.com/analytics-vidhya/understanding-principle-component-analysis-pca-step-by-step-e7a4bb4031d9

[18] OpenML, Credit Card Fraud Detection Dataset, https://www.openml.org/search?type=data&sort=runs&id=42397

[19] Data Clustering Algorithms, fuzzy-c-means-clustering-algorithm, https://sites.google.com/site/dataclusteringalgorithms/fuzzy-c-means-clustering-algorithm

[20] Towards Data Science, Logistic Regression, https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

[21] KDnuggets, Decision Tree, https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

[22] DataCamp, Naive Bayes, https://www.datacamp.com/tutorial/naive-bayes-scikit-learn

8 Authors

Muhammad Zohaib Khan was born in Pakistan. He received a Master degree in Computer Science from the Sindh Madressatul Islam University, Karachi, Pakistan, in 2022, and a Bachelor degree in Computer Science from the University of Sindh, Jamshoro, Pakistan, in 2016. He worked as an IT Engineer in the Department of Information Technology, Sindh Public Procurement Regulatory Authority, from 2017 to 2019. His research interests include Data Science, Artificial Intelligence, Machine Learning, Deep Learning, and the Internet of Things.

Dr. Sarmad Ahmed Shaikh was born in Pakistan. He received Ph.D. degree in Information Technology from the University of Klagenfurt, Klagenfurt, Austria, in 2018. He is currently an Assistant Professor at the Faculty of Information Technology, Sindh Madressatul Islam University (SMIU), Karachi, Pakistan. His research interests include RF/microwave engineering, radio source localization, antenna array designing, wireless sensor networking (IoT), and artificial intelligence.

Muneer Ahmed Shaikh was born in Pakistan. He received the Bachelors degree in Computer Systems Engineering from Mehran University of Engineering & Technology Jamshoro, Pakistan, in 1994, the M.E. degree in Computer Systems Engineering from NED University, Karachi, Pakistan, in 2010, and the Ph.D. degree (Continued) in Computer Science from SMI University, Karachi, Pakistan. He is currently an Assistant Professor in Computer Science, College Education Department, Government of Sindh, Karachi, Pakistan. His research interests include radio source localization, antenna arrays designing, wireless communication, and artificial intelligence.

Dr. Kamlesh Kumar Khatri was born in Pakistan. He obtained his Bachelors degree in Computer Science from Shah Abdul Latif University, Khairpur, Pakistan, in 2006, the Master's degree in Computer Science from SALU Khairpur, Pakistan, in 2010, and the PhD degree in Computer Science & Technology from UESTC China in 2016. He is currently serving at Sindh Madressatul Islam University, Karachi, Pakistan, as an Assistant Professor in the Department of Software Engineering. His research interests include Image Processing, Computer Vision, Machine Learning, Knowledge Discovery and Data Mining, and Content Based Image Retrieval System.

Mahira Abdul Rauf was born in Pakistan. She received a Master degree in Computer Science from the Sindh Madressatul Islam University, Karachi, Pakistan, in 2022. Currently she is working as SAP BASIS Consultant at TallyMarks Consultancy, Pakistan. Her research interests include Machine Learning, Artificial Intelligence, Cloud Computing, the Internet of Things, and the Blockchain.

Ayesha Kalhoro was born in Pakistan. She received a Master degree in Computer Science from the Sindh Madressatul Islam University, Karachi, Pakistan, in 2022. Currently, she is working as a technical and research writer at BrandH20. Her research interests include Machine Learning, Artificial Intelligence, Cloud Computing, the Internet of Things, and Blockchain.

Muhammad Adnan was born in Pakistan. He received a Master degree in Computer Science from the Sindh Madressatul Islam University, Karachi, Pakistan, in 2022. Currently, he is working as a Freelance Data Science consultant on Kaggle. His research interests include Machine Learning, Artificial Intelligence, and Bioinformatics.

Article submitted 2022-09-14. Resubmitted 2023-01-26. Final acceptance 2023-01-26. Final version published as submitted by the authors.