Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6079582, 14 pages
https://doi.org/10.1155/2021/6079582

Research Article

An Enhanced Secure Deep Learning Algorithm for Fraud Detection in Wireless Communication

Sumaya Sanober,[1] **Izhar Alam**,[2] **Sagar Pande**,[2] **Farrukh Arslan**,[3] **Kantilal Pitambar Rane**,[4] **Bhupesh Kumar Singh**,[5] **Aditya Khamparia**,[6] **and Mohammad Shabaz**[5,7]

[1]*Computer Science and Engineering, Prince Sattam Bin Abdul Aziz University, Saudi Arabia Wadi Aldwassir*
[2]*Computer Science and Engineering, Lovely Professional University, Punjab, India*
[3]*School of Electrical and Computer Engineering, Purdue University, USA*
[4]*KCEs COEM JALGAON, India*
[5]*Arba Minch University, Ethiopia*
[6]*Babasaheb Bhimrao Ambedkar University, Lucknow, India*
[7]*Department of Computer Science Engineering, Chitkara University, India*

Correspondence should be addressed to Sumaya Sanober; s.sanober@psau.edu.sa, Sagar Pande; sagarpande30@gmail.com, and Mohammad Shabaz; mohammad.shabaz@amu.edu.et

Received 9 July 2021; Revised 20 July 2021; Accepted 26 July 2021; Published 15 August 2021

Academic Editor: VIMAL SHANMUGANATHAN

In today's era of technology, especially in the Internet commerce and banking, the transactions done by the Mastercards have been increasing rapidly. The card becomes the highly useable equipment for Internet shopping. Such demanding and inflation rate causes a considerable damage and enhancement in fraud cases also. It is very much necessary to stop the fraud transactions because it impacts on financial conditions over time the anomaly detection is having some important application to detect the fraud detection. A novel framework which integrates Spark with a deep learning approach is proposed in this work. This work also implements different machine learning techniques for detection of fraudulent like random forest, SVM, logistic regression, decision tree, and KNN. Comparative analysis is done by using various parameters. More than 96% accuracy was obtained for both training and testing datasets. The existing system like Cardwatch, web service-based fraud detection, needs labelled data for both genuine and fraudulent transactions. New frauds cannot be found in these existing techniques. The dataset which is used contains transaction made by credit cards in September 2013 by cardholders of Europe. The dataset contains the transactions occurred in 2 days, in which there are 492 fraud transactions out of 284,807 which is 0.172% of all transaction.

1. Introduction

Credit card fraud might be a significant issue which requires payment card as Mastercard as illegal supply of money in transactions. Fraud is illegal because of getting funds and goods. The objective of such unlawful transaction might be urging items without paying and also obtain an unauthorized fund from an account. Identifying such fraud might be a troublesome and must risk the company as well as business organizations. Within the world of Fraud Detection System (FDS) [1], investigators are not prepared to examine each transactions. Here, the Fraud Detection System monitors all of the authorized transactions and alerts the foremost distrustful one. Investigator verifies these alerts and also provides FDS with responses in case the transaction was authorized and fraudulent. Verifying all of the alerts each day might be a time intensive and dear process. Hence, investigator is in a place to confirm just a number of alerts each day. The rest of the transactions stay unchecked until client identifies them and reports them to be a fraud. Also, the techniques employed for fraud, and consequently, the cardholder paying behavior changes over time. This particular alteration

in Mastercard transaction is called as idea drift [1, 2]. Thus, usually, it is hard to notice the Mastercard fraud. Machine learning is taken into consideration collectively of the foremost profitable method for fraud identification. Classification is used by it and also regression strategy for knowing fraud in Mastercard. The machine learning algorithms are split into 2 kinds, supervised [3] along with unsupervised [4] learning algorithm. Supervised learning algorithm uses labeled transactions for instructing the classifier whereas unsupervised mastering algorithm uses coeval's analysis that groups customers in line with the profile of theirs and identifies fraud supported clients spending behavior.

Many learning algorithms are offered for fraud detection in Mastercard that features neural networks, logistic regression (LR), Naive Bayes (NB), Support Vector Machines (SVM), decision tree (DT), and K-nearest neighbors (KNN) as well as random forest (RF). This paper examines the functionality of above algorithms supported the ability of theirs to classify whether the transaction was authorized, and fraudulent next compares them. The comparison is created utilizing performance measure accuracy, precision, and specificity. The end result proved that random forest algorithm showed improved precision and accuracy than some other methods. Further the obtained accuracy was improved by using deep Autoencoder.

The following are the main contributions in this paper:

(i) Novel deep learning framework is implemented using Spark for financial fraudulent detections

(ii) Comparative analysis is performed with proposed deep architecture using various machine learning algorithms

(iii) Performance factors like accuracy, specificity, and precision are used for comparing their performance measures

(iv) The importance of feature selection techniques is discussed and explored with five different techniques

(v) A stacked-based novel approach for feature selection is proposed

(vi) Comparative analysis is performed with proposed deep architecture using various machine learning algorithms

(vii) Novel deep learning framework is implemented using Spark for financial fraudulent detections

(viii) Performance factors like accuracy, specificity, and precision are used for comparing their performance measures

The paper is organized as follows: review of the related papers has been done in the literature review section, and next section proposed the methodology where discussion on the dataset is provided along with its description. Further section is of result analysis where comparison of all the algorithms is done by using the performance factor. The experiment is performed on a system having the configuration of 8 gigabytes of RAM, Intel i5 8th generation quad-core processor with 1.6 GHz clock speed. In the last section, the conclusion and future scope are explained.

2. Literature Survey

Various papers were reviewed and are discussed as follows.

Altiti [1] states that the fast evolution of technology all around the world is more often using cards as compared to cash in their day to day life. The Mastercard becomes the highly useable equipment for Internet shopping. This increase in use causes a considerable damage and fraud cases also. It is very much necessary to stop the fraud transactions because it impacts on financial conditions over time the anomaly detection is having some important application to detect the fraud detection. This paper is mainly focused on checking if the transaction is legal or fraud. They present models like "Bidirectional Long short-term memory (BiLSTM)" and "Bidirectional Gated recurrent unit (BiGRU)." They also apply deep learning and Machine Learning algorithms. But their model shows much better results than the machine learning classifiers which is 91.37% score.

Makki et al. [2] describe that in today's era of technology especially in the Internet commerce and banking, the transactions by the Mastercards have been increasing rapidly. The Mastercard becomes the highly useable equipment for Internet shopping. This increase in use causes a considerable damage and fraud cases also. It is very much necessary to stop the fraud transactions because it impacts on financial conditions over time the anomaly detection is having some important application to detect the fraud detection. The paper mainly focused on the solution that tackles the imbalance problem of classification they explore the solution for fraud detection using machine learning algorithms. They also find the summarized results and weakness that they get using credit card fraud labeled dataset. They give us the conclusion that the imbalanced classification is ineffective when the data are highly imbalanced. In this paper, the authors found that the existing methods were costlier and show many false alarms.

Ounacer et al. [3] state that in today's era of technology especially in the Internet commerce and banking, the transactions by the Mastercards have been increasing rapidly. The Mastercard becomes the highly useable equipment for Internet shopping. This increase in use causes a considerable damage and fraud cases also. It is very much necessary to stop the fraud transactions because it impacts on financial conditions over time the anomaly detection is having some important application to detect the fraud detection. Logistic regression, decision tree, SVM, and so on are some approaches to detect anomalies. But these methods are limited because they are supervised algorithms which are trained by the labels to know whether the transactions are legitimate or not.

Benchaji et al. [4] state that in today's era of technology especially in the Internet commerce and banking, the

transactions by the Mastercards have been increasing rapidly. The Mastercard becomes the highly useable equipment for Internet shopping. This increase in use causes a considerable damage and fraud cases also. It is very much necessary to stop the fraud transactions because it impacts on financial conditions over time the anomaly detection is having some important application to detect the fraud detection. The purpose of this paper is to enhance the performance of the classified instances in the imbalanced dataset for which they proposed the unsupervised sampling method based on the genetic algorithm and K-means clustering.

Dal Pozzolo et al. [5] describe that in today's era of technology especially in the Internet commerce and banking, the transactions by the Mastercards have been increasing rapidly. The Mastercard becomes the highly useable equipment for Internet shopping. This increase in use causes a considerable damage and fraud cases also. It is very much necessary to stop the fraud transactions because it impacts on financial conditions over time the anomaly detection is having some important application to detect the fraud detection.

Zheng et al. [6] describe that with the increase of e-commerce, transactions are also increasing in which some of them were fraud. To detect the fraud transaction, it is important to extract historical transaction records on the behavior profile of the users. To represent the BPs of the user, the Markov chain model is popular. Whose transaction behaviors are stable, this will affect them. The Mastercard becomes the highly useable equipment for Internet shopping. This increase in use causes a considerable damage and fraud cases also.

Venkata Suryanarayana et al. [7] address states in today's era of technology especially in the Internet commerce and banking, the transactions by the Mastercards have been increasing rapidly. The Mastercard becomes the highly useable equipment for Internet shopping. This increase in use causes a considerable damage and fraud cases also. It is very much necessary to stop the fraud transactions because it impacts on financial conditions over time the anomaly detection is having some important application to detect the fraud detection. This paper states the overall performance of LR, RF, and DT for charge card fraud detection. The 3 methods are used for the dataset, and function is applied in the R language. The functionality of the methods is actually evaluated for diverse variables grounded on awareness, specificity, and reliability as well as error rate. The end result displays of reliability for LR, RF, and DT classifier are actually 90.0, 95.5.3, and 94.3, respectively. The comparative results indicate that the random forest does much better compared to the logistic regression as well as decision tree techniques.

Thennakoon et al. [8] state that in today's era of technology especially in the Internet commerce and banking, the transactions by the Mastercards have been increasing rapidly. The Mastercard becomes the highly useable equipment for Internet shopping. This increase in use causes a considerable damage and fraud cases also. It is very much necessary to stop the fraud transactions because it impacts on financial conditions over time the anomaly detection is having some important application to detect the fraud detection. Fraud transactions are one of the major financial issues in the banks. There are 10 million transactions that are fraudulent out of 12 billion which can cause a huge loss. So to analyze these, they have predicted the fraud transaction using isolation forest and local outlier factor. They also calculated the no. of error and accuracy of both algorithms.

Shukur and Kurnaz [9] project that in today's era of technology especially in the Internet commerce and banking, the transactions by the Mastercards have been increasing rapidly. The Mastercard becomes the highly useable equipment for Internet shopping. This increase in use causes a considerable damage and fraud cases also. It is very much necessary to stop the fraud transactions because it impacts on financial conditions over time the anomaly detection is having some important application to detect the fraud detection. Such issues can also be tackling with the help of data science with the combination of machine learning. The main objective here is to find all the fraud transactions while increasing the accuracy. Mastercard fraudulent detection is actually a sample of classification. With this procedure, centering on preprocessing datasets and analyzing in addition to the deployment of several anomaly detection algorithms like isolation forest algorithm as well as local outlier factor on the PCA changed Mastercard transaction information.

John and Naaz [10] describe that in today's era of technology especially in the Internet commerce and banking, the transactions by the Mastercards have been increasing rapidly. The Mastercard becomes the highly useable equipment for Internet shopping. This increase in use causes a considerable damage and fraud cases also. It is very much necessary to stop the fraud transactions because it impacts on financial conditions over time the anomaly detection is having some important application to detect the fraud detection. Online transaction fraud detection may be the vast majority of challenging issue for financial businesses and banks. So it is much crucial for financial businesses and also banks to have highly effective fraud detection techniques to reduce the losses of theirs as an outcome of these fee card fraud transactions. Different techniques are found by many researchers till morning to be able to recognize these frauds at the same time as to take down them. After the analysis of the dataset, the reliability is ninety-seven % by LOF and seventy-six % by IF.

Yu et al. [11] state that that in today's era of technology especially in the Internet commerce and banking, the transactions by the Mastercards have been increasing rapidly. The Mastercard becomes the highly useable equipment for Internet shopping. This increase in use causes a considerable damage and fraud cases also. It is very much necessary to stop the fraud transactions because it impacts on financial conditions over time the anomaly detection is having some important application to detect the fraud detection. This algorithm detects the frauds very quickly resulting in the reduction of loss and risks.

3. Methodology

3.1. Dataset Description. The datasets consist of card purchases made by European cardholders in September 2013. This dataset describes transactions that happened in 2 days, specifically where 492 frauds beyond 284,807 transactions. The dataset is highly unbalanced; most transactions account for 0.172 per cent of the beneficial group (frauds). Figure 1 depicts the neural network architecture. The proposed framework is represented in Figure 2. Generalized block diagram is represented in Figure 3.

3.2. Autoencoder. AE is used to reduce input sizes to a smaller representation. They will recreate it from the compressed data if someone wants the original data. Having a similar algorithm in machine learning, i.e., PCA, performs the same task. AE is a class of unmonitored networks consisting of two main networks: Encoders and Decoders. The standard Autoencoder working can be seen in Figure 4. An AE neural network is an unsupervised learning algorithm which applies back propagation and sets target values equal to the inputs; i.e., they are using $B(i) = A(i)$. Simply put, an AE is made up of two parts, an encoder and a decoder. Taking into account, a data model A with samples and f attributes, the encoder output B represents a reduced representation of A, and the decoder is optimized to recreate the original dataset A from the representation B of the encoder by minimizing the gap between A and $A0$. The encoder is simply a function f, which maps an input A to hidden representation B. The method is set out as [12].

$$B = f(A) = a_f(W_m A + b_x), \tag{1}$$

where a_f is a nonlinear activation function and the AE must do linear projection if it is an identity function. The encoder is parameterized by a W_m matrix of weight and a bias vector by $b \epsilon R^n$.

The decoder function d maps hidden representation B back to a reconstruction A' as follows:

$$A' = d(B) = a_d\left(W'_m B + b_y\right), \tag{2}$$

where a_d is the activation function of the decoder, either the identity (rendering linear reconstruction) or a sigmoid is usually used. Parameters of the decoder are by and matrix W'_m a bias vector. In this paper, we explore only the case of bound weights where $W'_m = W_m^T$. Training an AE involves finding parameters like $\Theta = (W_m, b_x, b_y)$ which minimize the loss of reconstruction on the given dataset X and the objective

$$\Theta = \min_\Theta L\left(A, A'\right) = \min_\Theta L(A, d(f(A))). \tag{3}$$

For linear reconstruction, the reconstruction loss (L1) is generally from the squared error as follows:

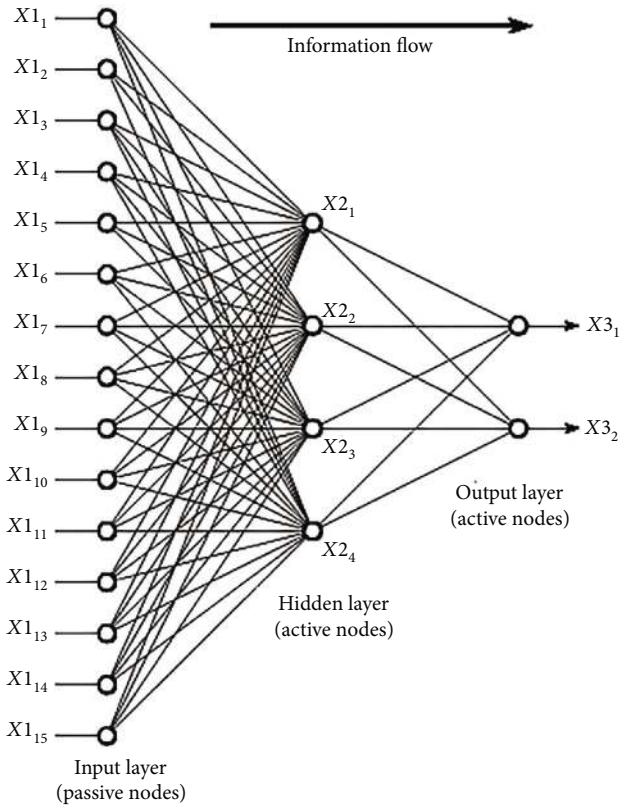

Figure 1: Neural network architecture [12].

$$L1(\Theta) = \sum_{i=1}^{n} \left\|a_i - a'_i\right\|^2 = \sum_{i=1}^{n} \|a_i - d(f(a_i))\|^2. \tag{4}$$

For nonlinear reconstruction, the reconstruction loss (L2) is generally from cross-entropy as follows:

$$L2(\Theta) = -\sum_{i=1}^{n} [a_i \log(b_i) + (1 - a_i) \log(1 - b_i)], \tag{5}$$

where $a_i \epsilon A$, $a'_i \epsilon A'$, and $b_i \epsilon B$.

Apache Spark3 is a streaming-enabled Map-Reduce implementation that distributes the computation automatically among the allocated resources and aggregates the results on a distributed file system. Spark offers both a deep and machine learning database and a streaming database. A strong point for Spark is its ability in the same framework to enable batch and stream analyses. The proposed framework is focused on Spark Streaming which processes data streams in minilots that trail the order of the latency of the second. Although this may be considered a downside in some streaming contexts, it is harmless in quasi-real-time setting. The Spark module of the system is written in Scala, a language that blends functional programming with object-oriented one. Scala runs atop Java VM and is fully compliant with the Java libraries. Overall, in the process, Spark fulfills two tasks: aggregating

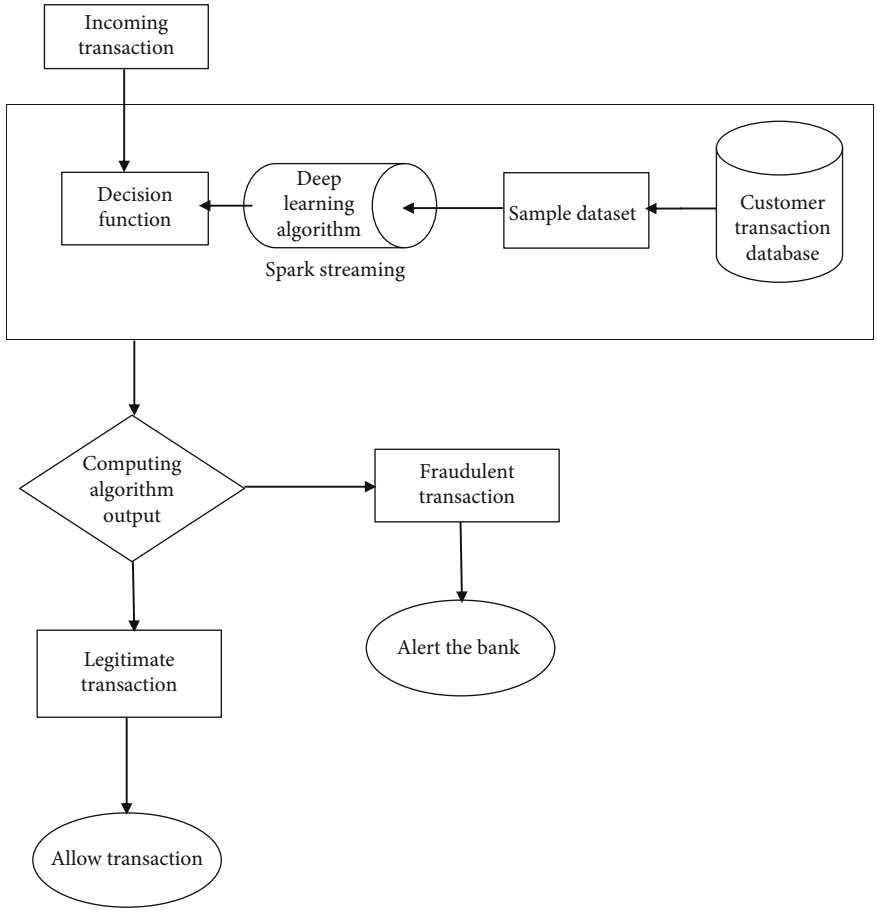

Figure 2: System flow.

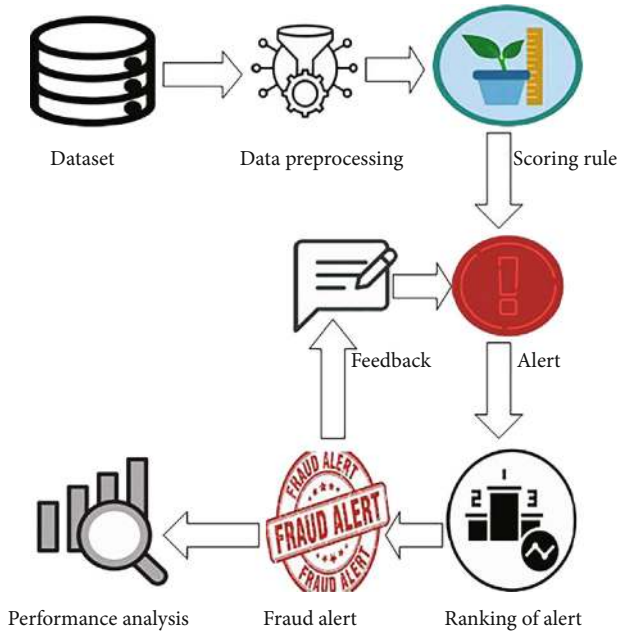

Figure 3: Proposed system block diagram.

historical transactions to perform design engineering and classifying transactions online that return the estimated risk of fraud.

3.3. Random Forest (RF) Algorithm. RF is a supervised learning algorithm, which can be used in addition to regression for both groups. But it is mainly used for classification issues. Because a forest is made up of more plants and leaves, it means a much better forest. Likewise, the RF algorithm selects trees on knowledge samples and then collects the prediction from all of them and eventually selects the optimal alternative by voting. It is an ensemble strategy that is much better than an individual choice tree because by averaging the end result it reduces the over fit.

The following is the implementation of random forest in scikit learn:

(i) node_j = importance of node j

(ii) weight_j = weightage no.of sample reaching node_j

(iii) C_j = impurity of node_j

(iv) left_j = child node from left split on node_j

(v) right_j = child node from right split on node_j

1: *procedure* $AE(T_E, B_S, a, D_S, L_R, \Theta)$:
2: $a = [a_1, a_2, \cdots . a_i] \in R^{i+j}$ is the *input matrix*, in *which* $a_n \in [0, 1]^j (1 \le n \le j)$ is *a single input data*
3: $T_E = training\_epochs = 10$
4: $B_S = batch\_size = 256$
5: $D_S = display\_step = 1$
6: $L_R = learning_{rate} = 0.01$
7: $\Theta = (W_m, b_x, b_y)$, where Θ is the *parameter* of *AE*
8: for 0 to T_E *do*
9: for 0 to B_S *do*
10: $f(A) = a_f(W_m A + b_x)$
11: $d(B) = a_d(W'_m B + b_y)$
12: $L1(\Theta) = \sum_{i=1}^{n} \|a_i - d(f(a_i))\|^2$
13: $L2(\Theta) = -\sum_{i=1}^{n} [a_i \log(b_i) + (1 - a_i) \log(1 - b_i)]$
14: $\Theta = \min_{\Theta} L(A, A')$
15: $C = compute$ the *cost with respect* to Θ
16: for Θ_n, C_n in (Θ, C) *do*
17: $\Theta_n = \Theta_n - L_R * C_n$
18: *end* for
19: *end* for
20: *end for*
21: *end procedure*

Algorithm 1

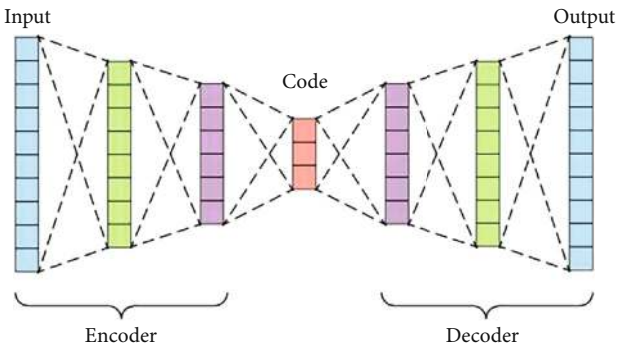

Figure 4: AE architecture [13].

$$\text{node}_j = \text{weight}_j C_j - \text{weight}_{\text{left}(j)} C_{\text{left}(j)} - \text{weight}_{\text{right}(j)} C_{\text{right}(j)}, \tag{6}$$

(vi) fi_i = importance of feature i

(vii) node_k = importance of node j

$$\text{fi}_i = \frac{\sum_{j:\text{node } j \text{ split on feature } i} \text{node}_j}{\sum_{k \in \text{all nodes}} \text{node}_k}, \tag{7}$$

(viii) RFfi_i = importance of feature I calculated from all trees in the random forest model

(ix) Normfi_{ij} = normalized feature importance for I in tree j

$$\text{normfi}_i = \frac{\text{fi}_i}{\sum_{j \in \text{all features}} \text{fi}_j}, \tag{8}$$

(x) T_{all} = total no.of trees

$$\text{RFfi}_i = \frac{\sum_{j \in \text{all trees}} \text{normfi}_{ij}}{T_{\text{all}}}. \tag{9}$$

3.4. K-Nearest Neighbor

(i) K-nearest neighbor (KNN) algorithm is a kind of supervised ML algorithm that can be used for predictive problems in both categories and regression. Nevertheless, it is mainly used in industry for predictive classification issues. The next 2 attributes could well decide KNN [13]

(ii) *Lazy Mastering Algorithm*. K-nearest neighbor is a sluggish learning algorithm since it does not possess a special education phase and also requires all of the information for education while classification

(iii) *Nonparametric Mastering Algorithm*. KNN is additionally a nonparametric learning algorithm since it does not believe anything about the main information

Implementation of K-nearest neighbor (KNN) Algorithm.

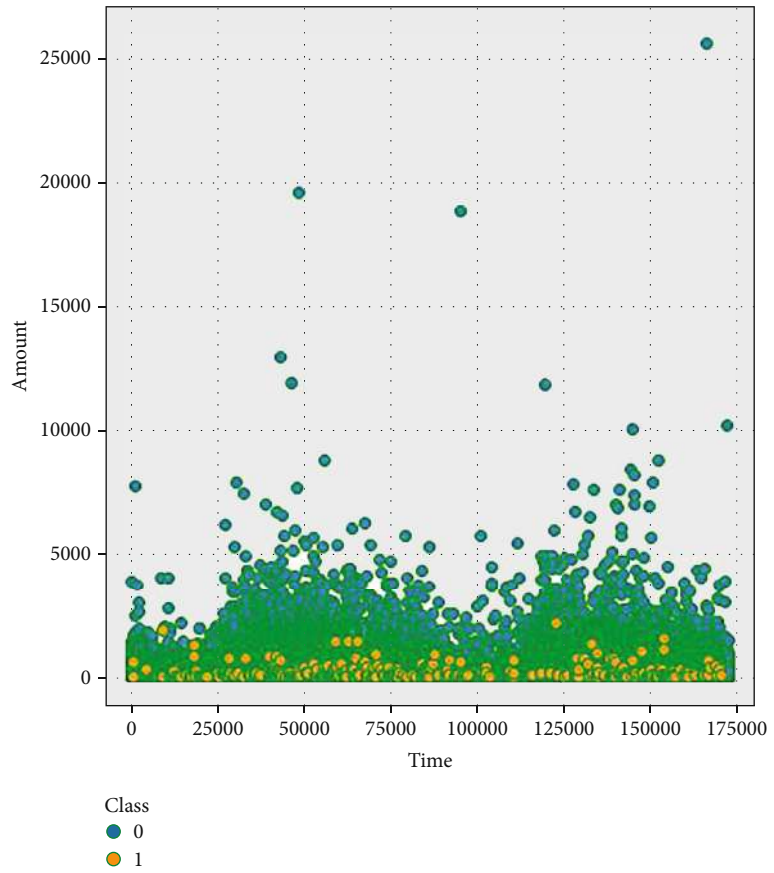

Figure 5: 2D scatter plot of distribution of all the cases.

We are able to know its working with the aid of pursuing steps:

Step 1. For applying some algorithm, we require dataset. So, throughout the initial stage of KNN, we should load the instruction and evaluation data.

Step 2. Next, we have to select the importance of K, i.e., probably the nearest data points. K could be any integer.

Step 3. For every stage within the test information do the following:

(a) Measure the gap between the training specifics of each row and check with the help of the strategy: Euclidean, Manhattan, or even Hamming distance. The most often used method to compute distance is Euclidean

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}, \tag{10}$$

(b) Now, dependent on the distance worth, sort them in ascending order

(c) Next, the high K rows from the sorted array are to be chosen

(d) Now, it is going to assign a course to the test stage based on many regular categories of these rows

Step 4. End.

3.5. Decision Tree Algorithm. The supervised learning algorithm contains decision tree. The general purpose of utilizing decision tree is creating a training type that will utilize to predict value or class of goal variables by mastering choice regulations inferred from prior data (training data). The comprehension amount of decision tree algorithm is very simple in contrast to some other group algorithms [14].

Implementation of Decision Tree Algorithm:

3.5.1. Gini Index (GI). It is the title of the price feature which is utilized to assess the binary splits in the dataset and also works together with the categorical target variable "Failure" or "Success." The higher the importance of GI, the higher will be the homogeneity. A great GI value is zero, and worst is 0.5 (for two class problem). Gini list for a split may be estimated with the aid of the following steps [14]:

Figure 6: Plot of all instances.

(i) For starters, compute Gini index for subnodes by utilizing the system $p^2 + q^2$ and that is the amount of the square of likelihood for failure and success

(ii) Then, compute Gini list for split using weighted Gini rating of every node of that particular split

Classification and Regression Tree (CART) algorithm employs Gini technique to produce binary splits.

3.5.2. Split Index. A split is simply incorporating a characteristic in a value and the dataset. We are able to develop a split in dataset with the assistance of the following 3 parts:

(i) *Calculating Gini Score*. We have simply talked about this particular component in the prior section

(ii) *Splitting a Dataset*. It might be described as separating a dataset into 2 lists of rows keeping index of a characteristic along with a split worth of that feature. After getting the 2 groups, right and also remaining, from the dataset, we are able to compute the importance of split by utilizing Gini score calculated in original part. Split value is going to decide where the team the attribute will reside

(iii) *Evaluating Almost All Splits*. Next component after finding Gini score as well as splitting dataset is

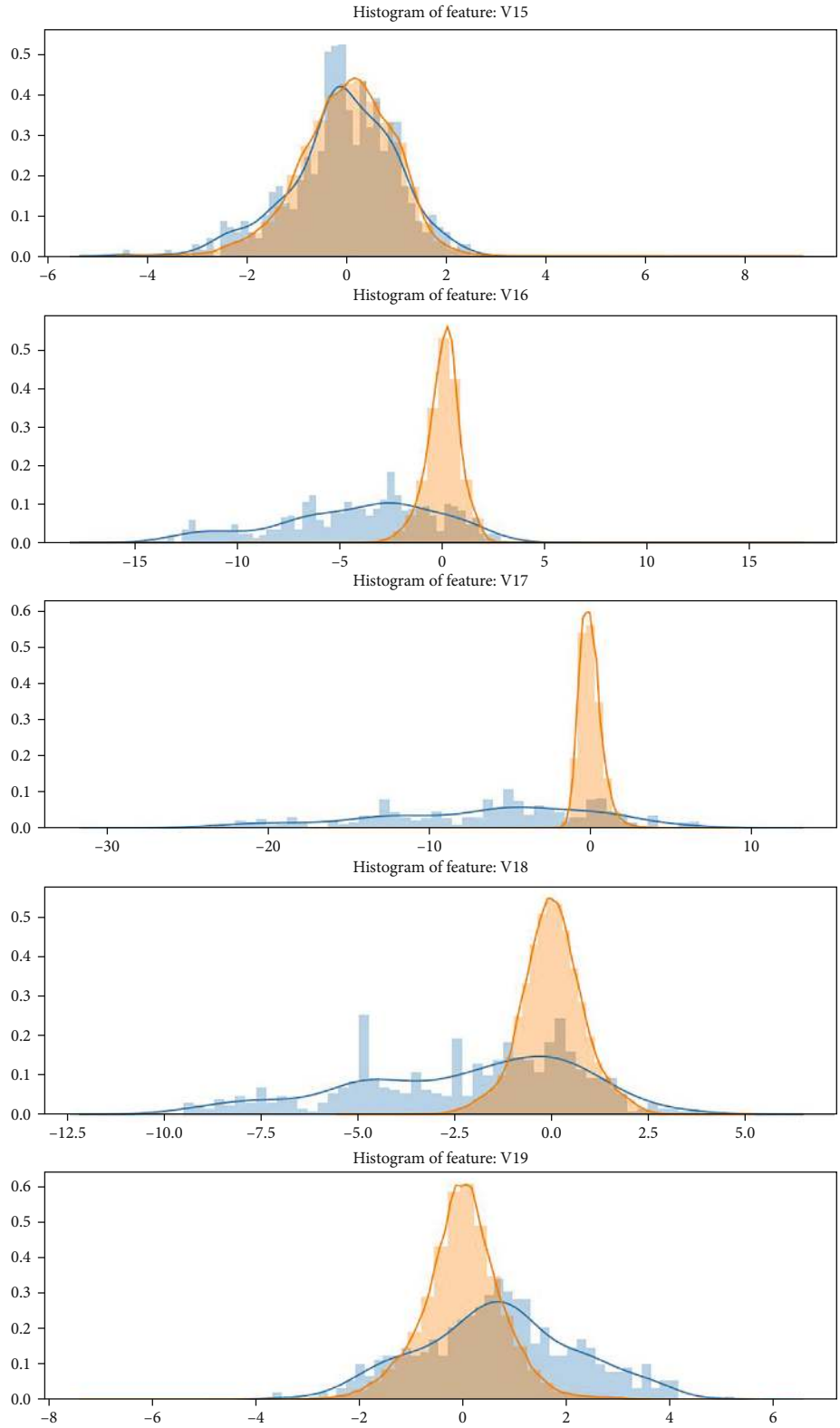

Figure 7: V15-V19 feature histogram.

definitely the analysis of all splits. For this particular purpose, for starters, we should examine each value connected to each feature as being a candidate split.

Next, we have to discover the absolute best split by analyzing the price of the split. The most effective split would be used as a node in the decision tree

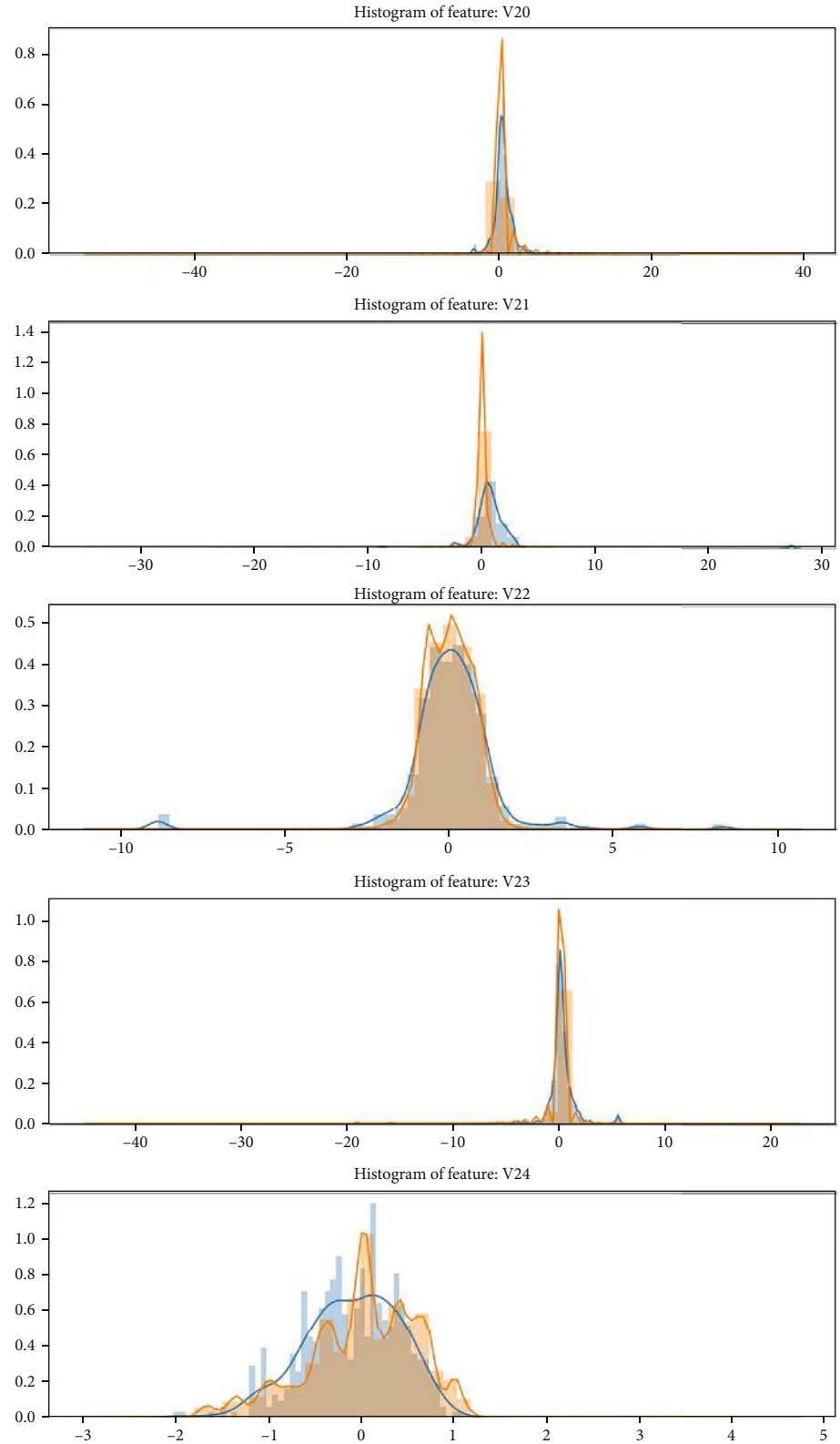

Figure 8: V20-V24 feature histogram.

3.6. Logistic Regression (LR). LR is a supervised learning category algorithm used to predict the likelihood of an adjustable goal. The target dynamics, or maybe dependent component, are dichotomous; meaning, there will be only 2 possible courses. In simple words, the dependent element is binary in nature to get knowledge written as

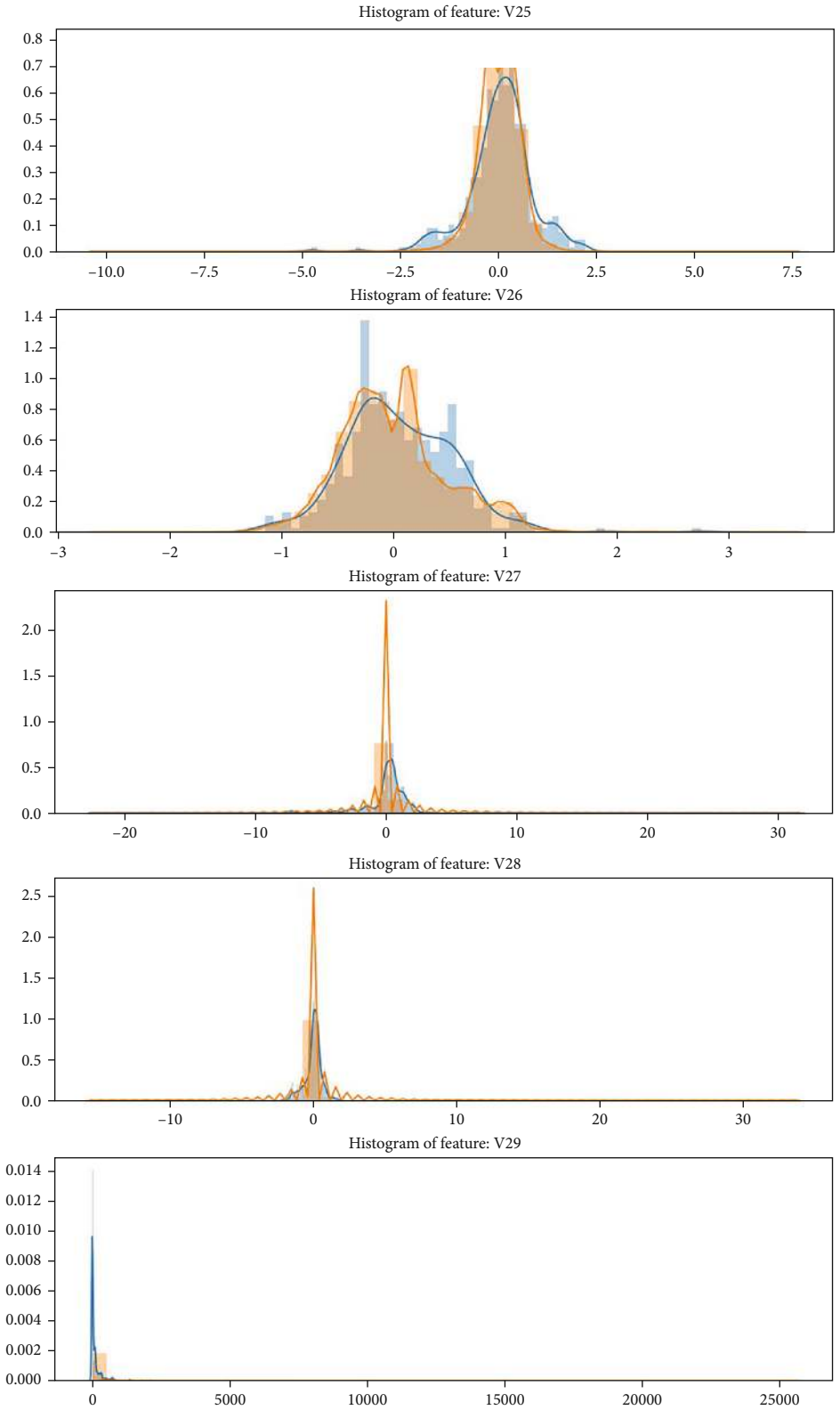

Figure 9: Amount and V25-V28 feature histogram.

theoretically one or even zero. Mathematically, $P(Y = 1)$ is predicted as a characteristic of X by an LR algorithm. It is among the simplest ML algorithms that could be used to detect spam, cancer detection, diabetes prediction, etc., for various classification complications [14].

3.6.1. Types of LR. In general, LR suggests binary LR owning binary goal variables, but there could be 2 more types of target variables which may be predicted by it. Based upon those numbers of types, LR is split into the following types:

(i) *Binomial*. In such a type of classification, a reliant variable is going to have just 2 possible kinds both one and zero

(ii) *Multinomial*. In such a type of category, dependent variable should have three or maybe more potential unordered styles or even the kinds getting no quantitative significance

(iii) *Ordinal*. In such a type of category, dependent variable should have three or maybe more feasible ordered styles or even the kinds with a quantitative significance

3.6.2. LR Assumptions

(i) Before diving into the implementation of LR, we should be conscious of the coming assumptions about the same

(ii) In case of binary logistic regression, the goal variables should be binary constantly, and the desired outcome is represented by the aspect level one

(iii) Right now, there should not be some multicolinearity within the product; this means the independent variables should be outside of one another

(iv) We need to have significant variables in the model of ours

(v) We must select a big sample size for LR

3.7. Support Vector Machine. In 1960s, SVMs were first released, but eventually, they have enhanced in 1990. SVMs have the unique way of theirs of setup as compared to various other machine learning algorithms. Recently, they are incredibly well known due to their capability to deal with a couple of continuous and categorical variables [15].

3.7.1. Working of SVM. An SVM unit is simply a representation of various courses in a hyperplane in space that is multidimensional. The hyperplane would be created within an iterative fashion by Support Vector Machine; therefore, the mistake could be lessened. The objective of Support Vector Machine is dividing the datasets into martial arts classes to locate an optimum marginal hyperplane.

The following are important ideas in SVM:

Table 1: Machine learning comparison table on the basis of performance measures.

Classifiers	Accuracy %	Specificity %	Precision %	F1-score %
Random forest	96.2	98.7	99.7	92
Logistic regression	94.7	97.9	99.6	91.7
SVM	93.8	98.4	78.2	80.2
Decision tree	90.8	91.2	91	86
KNN	94.2	97.1	41.0	50.6

Table 2: Deep learning evaluation for training dataset.

Train data	Test data	Train instances	Train fraud cases	% of train fraud cases	Training accuracy	Time elapsed (sec)
90	10	256326	470	0.00183	0.959921	18.35
80	20	227845	417	0.00183	0.954843	17.36
70	30	199364	384	0.00193	0.957749	14.63
60	40	170884	360	0.00211	0.963300	12.54
50	50	142403	269	0.00189	0.953874	10.63

(i) *Support Vectors*. Data points what are nearest to the hyperplane is called SVs. Separating line would be identified with the aid of these data points

(ii) *Hyperplane*. It is a choice plane or maybe room that is split between a pair of items having various classes

(iii) *Margin*. It might be described as the gap between 2 lines on the closet information points of various courses. It may be estimated as the perpendicular distance out of the series on the assistance vectors. Huge margin is viewed as an excellent margin, and tiny margin is as a terrible margin

The primary objective of SVM is dividing the datasets into classes which can be achieved inside the next 2 steps as follows:

(i) First, SVM is going to generate hyperplanes iteratively that segregates the classes in most effective way

(ii) Next, it is going to choose the hyperplane which separates the classes properly

4. Experiment and Result Analysis

Several machine learning algorithms are analyzed for the performance measures in the credit card fraud detection dataset. Along with this, deep Autoencoder is implemented using various training and testing split ratio. Majorly five core machine learning algorithms, namely, RF, LR, KNN, DT, and SVM algorithm, are implemented. From Figure 5, it is clearly visible that there are frauds only on the transactions which have transaction amount approximately less than 2500. Transactions which have transaction amount approximately above 2500 have no fraud. As per with the time, the

Table 3: Deep learning evaluation for testing dataset.

Train data	Test data	Test instances	Test fraud cases	% of test fraud cases	Testing accuracy	Time elapsed (sec)
90	10	28481	22	0.00077	0.923378	00.07
80	20	56962	75	0.00132	0.949008	00.13
70	30	85443	108	0.00126	0.939815	00.23
60	40	113923	132	0.00116	0.944843	00.32
50	50	142404	223	0.00157	0.965315	00.38

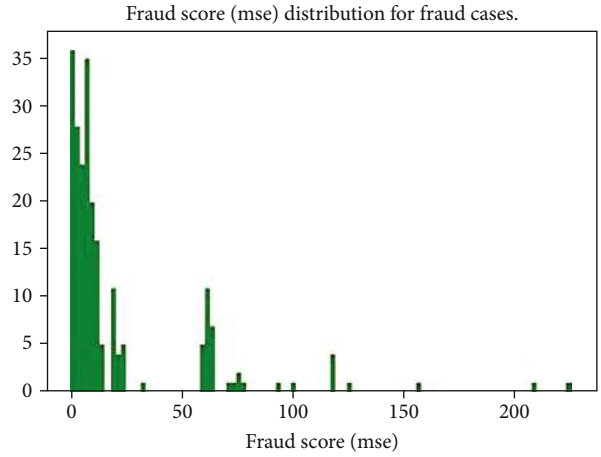

Figure 10: Fraud score distribution for 50-50 split ratio.

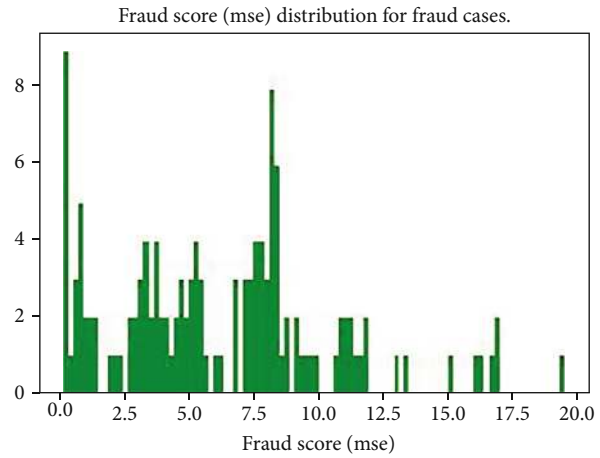

Figure 11: Fraud score distribution for 60-40 split ratio.

frauds in the transactions are evenly distributed throughout time. Amount and time distribution can be seen in Figure 6. Feature distribution from 15 to 30 is depicted in Figures 7–9, respectively. Table 2 narrates the obtained machine learning results. Tables 2 and 3 provide the result obtained by using Deep AE with various training and testing split ratio. Figures 10 and 11 depict fraud score distribution for 50-50 and 60-40 split ratio, respectively. Accuracy, precision, and specificity for all machine learning algorithms are calculated as follows which is shown in Table 1:

$$\begin{aligned} &\text{TP} = X, \text{TN} = Y, \text{FP} = P, \text{FN} = Q, \\ &\text{Accuracy} = \frac{X + Y}{X + Y + P + Q}, \\ &\text{Precision} = \frac{X}{X + P}, \\ &\text{Specificity} = \frac{Y}{Y + P}, \\ &F1\ \text{Score} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}}. \end{aligned} \tag{11}$$

5. Conclusion

As in today's era of technology, especially in the Internet commerce and banking, the transactions by the Mastercards have been increasing rapidly. The Mastercard becomes the highly useable equipment for Internet shopping. This increase in use causes a considerable damage and enhances inflation rate of fraud cases also. It is very much necessary to stop the fraud transactions because it impacts on financial conditions over time the anomaly detection is having some important application to detect the fraud detection. This paper has reviewed several algorithms to identify fraud in card transaction. Autoencoder is used to classify the alert as fraudulent or even authorized in spark environment. Next, it will aggregate every probability to discover alerts. Further, proposed model utilizes ranking approach where alert is positioned based on priority. The model is able to resolve the class imbalance. In today's era, we just detect the fraudulent transaction, but we are not able to prevent it. Preventing fraud transaction dynamically is not easy, but it is possible. The system which proposed is design to detect fraud transaction, but in future by some advancement, it can became fraud prevention system.

Data Availability

The data shall be made available on request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] O. Altiti, "Credit card fraud detection based on machine and deep learning," in *2020 11th International Conference on Information and Communication Systems (ICICS)*, pp. 204–208, Irbid, Jordan, 2020.

[2] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M. S. Hacid, and H. Zeineddine, "An experimental study with imbalanced classification approaches for credit card fraud detection," *IEEE Access*, vol. 7, pp. 93010–93022, 2019.

[3] S. Ounacer, H. A. El Bour, Y. Oubrahim, M. Y. Ghoumari, and M. Azzouazi, "Using Isolation Forest in anomaly detection: the case of credit card transactions," *Periodicals of Engineering & Natural Sciences*, vol. 6, no. 2, pp. 394–400, 2018.

[4] I. Benchaji, S. Douzi, and B. El Ouahidi, "Using genetic algorithm to improve classification of imbalanced datasets for credit card fraud detection," *Lecture Notes in Networks & Systems*, vol. 66, pp. 220–229, 2019.

[5] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, "Credit card fraud detection: a realistic modeling & a novel learning strategy," *IEEE Transactions on Neural Networks & Learning Systems*, vol. 29, no. 8, pp. 3784–3797, 2018.

[6] L. Zheng, G. Liu, C. Yan, and C. Jiang, "Transaction fraud detection based on total order relation & behavior diversity," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 3, pp. 796–806, 2018.

[7] S. Venkata Suryanarayana, G. N. Balaji, and G. Venkateswara Rao, "Machine learning approaches for credit card fraud detection," *International Journal of Engineering & Technology (UAE)*, vol. 7, no. 2, pp. 917–920, 2018.

[8] A. Thennakoon, C. Bhagyani, S. Premadasa, S. Mihiranga, and N. Kuruwitaarachchi, "Real-time credit card fraud detection using machine learning," in *Proceedings of the 9th International Conference on Cloud Computing, Data Science & Engineering*, vol. 7no. 10, pp. 488–493, Noida, India, 2019.

[9] H. A. Shukur and S. Kurnaz, "Credit card fraud detection using machine learning methodology," *International Journal of Computer Science and Mobile Computing*, vol. 8, no. 3, pp. 257–260, 2019.

[10] H. John and S. Naaz, "Credit card fraud detection using local outlier factor & isolation forest," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 4, pp. 1060–1064, 2019.

[11] W. F. Yu and N. Wang, "Research on credit card fraud detection model based on distance sum," in *IJCAI international joint conference on artificial intelligence*, pp. 353–356, Hainan, China, 2009.

[12] https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc.

[13] "Autoencoders tutorial : a beginner's guide to autoencoders," https://www.edureka.co/blog/autoencoders-tutorial/.

[14] Q. Meng, D. Catchpoole, D. Skillicom, and P. J. Kennedy, "Relational autoencoder for feature extraction," in *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 364–371, Anchorage, AK, 2017.

[15] F. Carcillo, A. D. Pozzolo, Y. L. Borgne, O. Caelen, Y. Mazzer, and G. Bontempi, "SCARFF: a scalable framework for streaming credit card fraud detection with spark," *Information Fusion*, vol. 41, pp. 182–194, 2018.