

Credit Card Fraud Detection using Machine Learning

Harsh

*Assistant Professor, Department of Computer Applications,
Panipat Institute of Engineering & Technology, Samalkha*

Abstract: Credit card fraud detection is presently the most frequently occurring problem in the present world. This is due to the rise in both online transactions and e-commerce platforms. Credit card fraud generally happens when the card was stolen for any of the unauthorized purposes or even when the fraudster uses the credit card information for his use. In the present world, we are facing a lot of credit card problems. To detect the fraudulent activities the credit card fraud detection system was introduced. In this paper, machine learning algorithms are used to detect credit card fraud. Standard models are firstly used. Then, hybrid methods which use Ada Boost and majority voting methods are applied. To evaluate the model efficacy, a publicly available credit card data set is used. Then, a real-world credit card data set from a financial institution is analyzed. In addition, noise is added to the data samples to further assess the robustness of the algorithms. The experimental results positively indicate that the majority voting method achieves good accuracy rates in detecting fraud cases in credit cards.

Keywords: Data Security, Credit card fraud detection, Network Security.

INTRODUCTION

The number of E-Commerce users has been steadily increasing in recent years, as has the size of online transactions. Fraudsters frequently employ a variety of channels to steal card information and transfer big sums of money in a short period of time, resulting in significant property losses for both customers and banks. As a result, machine learning and data mining can be used to create fraud detection systems. Techniques used for this purpose are primarily classification-based. Data mining is applied to the dataset in question, then, classification algorithms are implemented to detect fraudulent transactions.

Credit Card fraud detection is a heavily researched problem. Due to that, we were able to look at popularly used datasets and algorithms for the same purpose as well as devise a way that would prove to be better than them. SVM, Naïve Bayes, Logistic Regression, Artificial Neural Networks, Decision Trees, and K-Nearest Neighbours were initially widely used for classification in this domain. They provided a moderate accuracy between 80-90% but improved with the incorporation of data mining techniques and hybrid models the scores went up further. As we moved up the years, Random Forest was observed to be the preferred choice to classify fraudulent data. It was better at overcoming the errors caused by highly imbalanced data in the fraud detection dataset. This occurs because each tree is generated by a random vector and each tree votes for the most popular category to classify the inputs. Random Forest's generalization performance is superior. Although it did extremely well, it had a high training time and didn't provide excellent results while working with a huge dataset. XGBoost, a step above Random Forest, significantly reduced training times and increased the efficiency of memory usage. However, in the year 2017, a faster algorithm based on decision trees was released by Microsoft

by the name of Light Gradient Boosting Machine or LightBGM. It is lighter and faster than XGBoost hence, we selected that as our classifier.

Machine learning is the innovation of this century that eliminates conventional strategies and can function on huge datasets that humans can't immediately access. Strategies of machine learning break into two important categories; supervised learning versus unsupervised learning; Tracking of fraud can also be achieved in any form and may only be determined how to use as per the datasets. Supervised training includes anomalies to always be identified as before. Many supervised methods have been used over the last few decades to identify credit card fraud. The major obstacle in implementing ML for detecting fraud seems to be the presence of extremely imbalanced databases. To overcome this obstacle, we have used a balanced dataset. Due to this, it is really helpful to perform experiments easily.

Throughout this study, we introduce an effective credit card fraud identification system with a feedback system, centered on machine learning techniques. That feedback approach contributes to boosting the classifier's detection rate and performance. Also, analysis of the performance of different classification methods includes random forest, tree classifiers, supporting vector machine, and logistic regression including Cat-Boost classifier approaches, on even a highly skewed credit card fraud database.

LITERATURE REVIEW

There are many different approaches are available for the fraud detection. Different authors use different type of approaches. Here some of the used methods are listed below for skin cancer with different datasets and different approaches (Table1).

Year	Title	Dataset	Model	Accuracy
2020	Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison	http://www.ulb.ac.be/di/map/adalpozz/imb_alancedatasets.z	Decision Tree, KNN, Random forest, Logistic Regression, Naïve Bayes	They take sensitivity and precision. Didn't mention accuracy
2019	Credit Card Fraud Detection using Machine Learning and Data Science	Kaggle	Local outlier factor, Isolation forest algorithm	99.67%, 99.77%
2021	Credit Card Fraud Detection Using Machine Learning	Kaggle	Decision tree, Random Forest, logitstic regression, naïve bayes	91.12%, 96.77% 95.16%
2019	Real-time Credit Card Fraud Detection Using Machine Learning	Real time data	Svm, Knn, Logistic regression, naïve bayes	91%, 72%, 74%, 83%
2020	Credit Card Fraud Detection Using Machine Learning	Kaggle	Random forest, Adaboost algorithm	Random forest has highest than adaboost algorithm
2019	Credit Card Fraud Detection - Machine Learning methods	Kaggle	Logistic regression, Naïve bayes, Ranodm forest, Multilayer percepton	97.46%, 99.23%, 99.96%, 99.93%
2021	Credit Card Fraud Detection using Machine Learning	Kaggle	Adaboost algorithm, Random Forest, LightGBM	96.13%, 95.95%, 95.78%
2021	Prediction of credit card defaults through data analysis and machine learning techniques	Uci library	KNN, random forest, Logistic regression, Svm, Naïve bayes	79%, 80%, 81%, 82%, 76%

Samidha Khatri et al. (2020) Credit card information about a specific person might be fraudulently acquired and used for fraudulent transactions. To solve this problem, certain Machine Learning Algorithms can be used to collect data. This study compares three well-

known supervised learning techniques for distinguishing between legitimate and fraudulent transactions. They provide precision and sensitivity.

S P Maniraj et al. (2019) On the PCA converted Credit Card Transaction data, we focused on evaluating and preprocessing data sets, as well as applying different anomaly detection techniques such as the Local Outlier Factor and Isolation Forest algorithm.

V.Sellam et al. (2021) For handling the highly imbalanced dataset, this research provides various machine learningbased classification techniques such as logistic regression, random forest, and Naive Bayes. Finally, the accuracy, precision, recall, f1 score, confusion matrix, and Roc-AUC score will be evaluated in this study.

Anuruddha Thennakoon et al. (2019) In this paper, Authors are using real time data and predictive analytics which is performed by ML models and an API module to detect the transaction is fraud or not. We also look at a new technique for dealing with data distribution that is skewed. According to a private disclosure agreement, the data used in our research came from a financial institution.

Ruttala Sailusha et al. (2020) In this research they focused on the random forest algorithm and the Adaboost algorithm are the algorithms employed. The accuracy, precision, recall, and F1-score of the two algorithms are used to compare their result. The confusion matrix is used to plot the ROC curve.

Dejan Varmedja et al. (2019) The Credit Card Fraud Detection dataset was used in this study. Because the dataset was highly imbalanced, the SMOTE technique was used to oversample it. The dataset was divided into two sections: training data and test data. The authors used Logistic Regression, Random Forest, Naive Bayes, and Multilayer Perceptron in this research. The findings show that each algorithm is capable of accurately detecting credit card fraud.

D. Tanouz et al. (2021) They make a graph, often known as a plot, and they analyze it. The model's recall, precision, and accuracy are then determined using three machine learning algorithms: light GBM, Adaboost, and random forest classifier. There's also a function for calculating the time it takes to run various algorithms. Finally, the value producedby these three algorithms is compared to determine which one produces the best result.

Saurabh Arora et al. (2021) They assess the dataset in this study, then do feature selection and apply various machine learning methods.

Hasan I and Rizvi S (2022) In this paper, the authors reviewed some Artificial intelligence and machine learning techniques to reduce fraud detection. They analyzed some techniques for the research challenge and provide the advantages and disadvantages of the techniques. From that, they provide the best techniques for credit card fraud detection.

Hussein, Ameer Saleh et al. (2021) In this paper, the authors used the fuzzy-rough nearest neighbor and sequential minimal optimization as base classifiers. They represent a combination of multiple classifiers through ensemble classifiers. They consider logistic classifiers as an outcome of the predictive model.

Kumar S et al. (2022) In this research, they tried support vector machine to overcome the drawbacks and gave result to detect the fraud using SVM.

METHODOLOGY

To proceed with our research we require a balanced dataset. The dataset used to perform experiments has been taken from Kaggle which was updated in 2023. The dataset is perfectly balanced with its class of fraudulent and Normal. Because of the balanced dataset, we do not require long pre-processing steps to balance the imbalance data. This dataset contains transactions made by cardholders in the year 2023. It comprises over 555,000 records, and the data has been anonymized to protect cardholders' identities. The primary objective of this dataset

is to facilitate the development of fraud detection algorithms and models to identify potentially fraudulent transactions.

TABLE1. DATA SET FEATURES AND DESCRIPTION.

Features	Description
id	Unique identifier for each transaction.
V1-V28	Anonymized features representing various transaction attributes (e.g., time, location, etc.).
Amount	Amount of transaction.
Class	A Binary label indicating whether the transactions fraudulent (1) or not (0).

There have been many approaches in Machine learning and Deep learning Methods to detect credit card fraud, but in this research, we focus on some Machine learning classifiers and neural networks. Experiments are nothing but hands-on experience in designing, conducting, and analyzing the research. The experiments are divided into the following four steps:

TABLE2. STEPS TO PERFORM EXPERIMENTS

S. No.	Steps
1.	Importing necessary libraries and loading the dataset.
2.	Pre-processing and EDA on a dataset.
3.	Building and Training of developed model to make predictions.
4.	Evaluation and Conclusion.

1. MACHINE LEARNING APPROACH

A. LOGISTIC REGRESSION

It is used for binary classification problems. The outcome is measured with a dichotomous variable. It is a type of generalized linear model (GLM).

B. DECISION TREE

It is primarily used for classification and regression tasks. It recursively partitions data into subsets based on feature value and each split is chosen to maximize information gain or minimize impurity, depending on the specific model used.

C. RANDOM FOREST

It is an ensemble of multiple decision trees, where each tree is trained on a random subset of the data and features. The name “Random Forest” stems from the idea of creating a forest of decision trees, and randomness is introduced in the construction of each tree to improve overall model performance.

D. K-NN

It is a supervised machine learning algorithm that makes predictions based on the similarity between query data points and their k-nearest neighbors in a labeled training dataset. The algorithm assigns the class label for classification or computes the weighted average for regression of the k-nearest neighbors to make predictions for the query data point.

E. CATBOOST

It is a gradient-boosting framework for supervised machine learning tasks that excels at handling categorical features, while also providing high predictive accuracy for classification and regression. It is an open-source library that uses a combination of gradient boosting and decision trees with several innovative techniques for efficient training.

F. SUPPORT VECTOR MACHINE

It aims to find the optimal hyper plane that best separates data points of different classes in

feature space while maximizing the margin between the hyper plane and the nearest data points. They are known for their ability to handle high-dimensional data and have applications in various fields.

G. ISOLATION FOREST

This algorithm works by randomly selecting a feature and then choosing a random value within the range of that feature's value to create a split. This process is repeated recursively until the anomalies are isolated into short partitions, while normal data points require more splits to isolate. The experiments start according to TABLE 2, importing the necessary libraries that are required for data visualization, model building, training, testing, and lastly result evaluation. The balanced data set is loaded. After performing Exploratory Data Analysis, the data is split into features and target labels. Where features represent the whole data except the column Class which is going to be used during the training of specific models and the target represents the data of column Class which is used to make predictions or classifications based on target labels. After this, the data is split into training and testing sets for training each model and to predict unseen data. Model training requires 80% and 20% of the dataset for training and testing respectively. The Standard Scalar function has been used to make data in a particular structure for ease of analysis and training. After this model has been developed and trained based on training sets of data. The prediction has been made by testing sets of data using the trained model. The unique factor in these algorithms is the hyper parameter used during the model building of each algorithm. They are parameters that are not learned from the data but are set before training and remain constant during training. It controls various aspects of the model's behavior and performance. Lastly, results have been evaluated using the Confusion Matrix. It is a table or graphs that summarize the performance of a classification algorithm by comparing the actual and predicted class labels for the dataset.

TABLE3. COMPONENTS OF THE CONFUSION MATRIX

Component	Description
True Positive	The model correctly predicts a positive When the true class is indeed positive.
True Negative	The model correctly predicts a negative When the true class is indeed negative.
False Positive	The model incorrectly predicts a positive When the true class is negative.
False Negative	The model incorrectly predicts a negative When the true class is negative.

2. DEEPLARNING APPROACH

A. CONVOLUTIONAL NEURAL NETWORK

CNN is used to perform experiments on datasets containing images. However, the one-dimensional feature of CNN allows us to perform experiments on the Comma Separated Value (i.e. CSV) data format. It contains three layers namely the input, hidden, and output layers.

TABLE4. LAYERS OF CNN

Layers	Description
Input	It takes one-dimensional data as an input
Hidden	Convolution: It operates on input data to extract local patterns and features. Pooling: It is used to reduce spatial dimensions of feature maps. Fully Connected (Dense): It includes dropout, batch normalization, and the activation function to improve model training and generalization.
Output	It produces the final prediction based on the features learned by the preceding layers.

B. RECURRENT NEURAL NETWORK

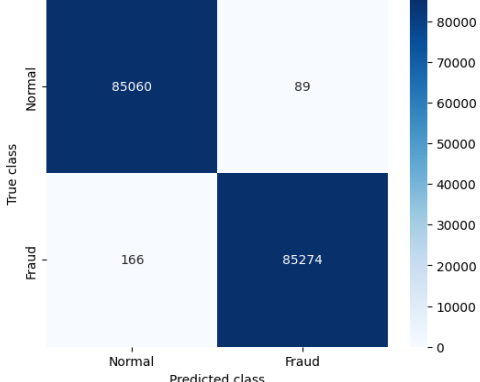
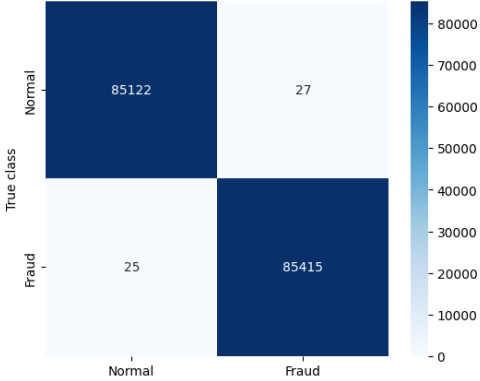
It is a type of artificial neural network architecture for processing sequences of data. RNNs are characterized by their ability to operate on variable-length sequences and are particularly suitable for tasks involving temporal dependencies and sequential patterns.

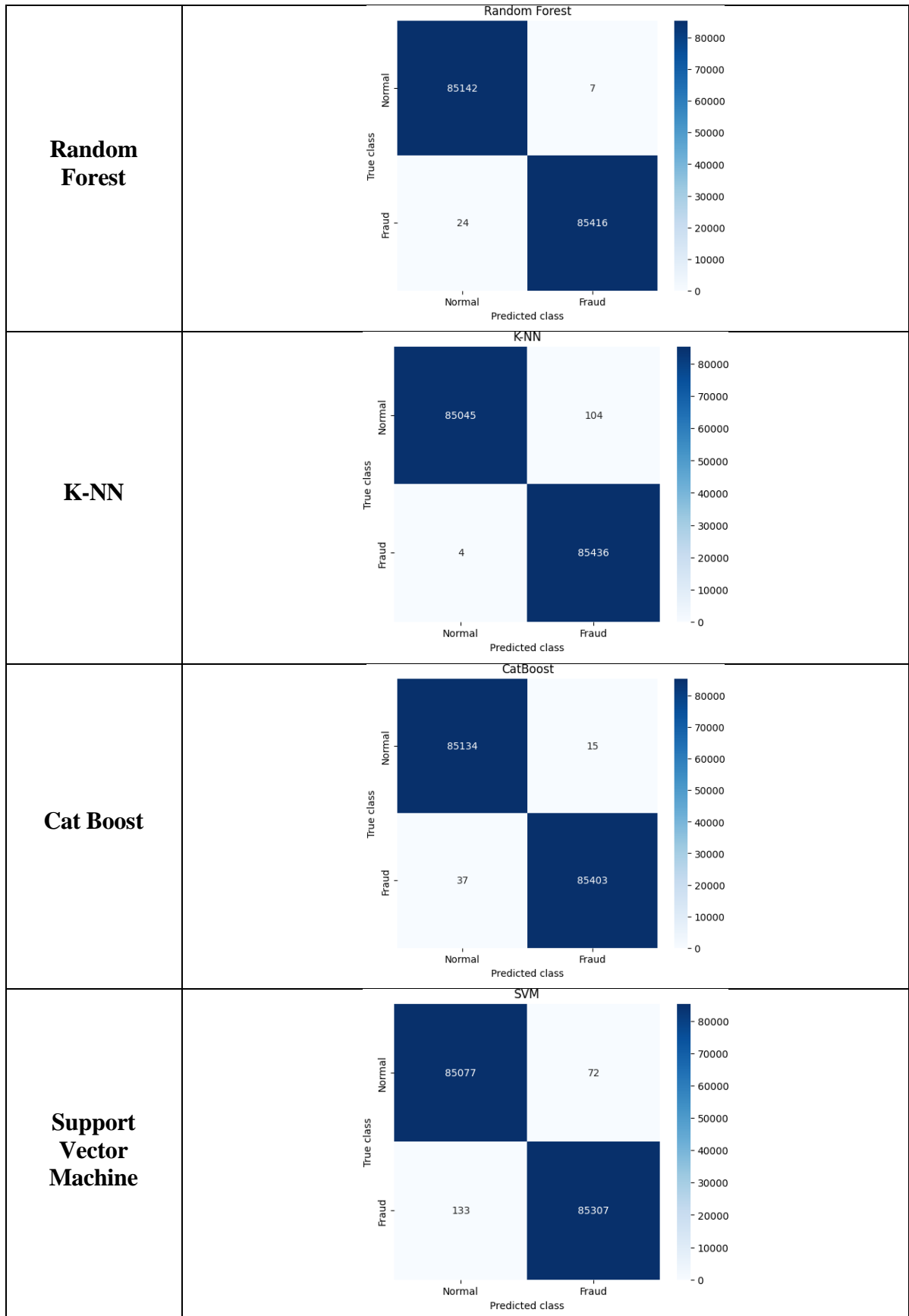
Model building of CNN and RNN is a crucial step in the entire experiment as it defines the number of layers along with their parameter used. A larger number of layers have a greater capacity to represent complex functions, which can lead to improved performance on challenging tasks as compared to a model having less number of layers. Training and Testing of the model have been performed using the layers with certainly required hyper parameters to evaluate the model performance with specific features. The model has been trained on an epoch size of 30 iterations. Testing is performed to showcase model accuracy and ability to make predictions.

RESULT EVALUATION

The balanced dataset is a key factor in our research experiments and evaluation process. It provides ease of access and understanding of data in performing operations of different algorithms on it. TABLE 5 classifies all the necessary factors like accuracy, precision, Recall, and F1-score for a better understanding of results and confusion matrix of performed experiments are mentioned below.

TABLE5. CONFUSION MATRIX

MODEL	CONFUSION MATRIX									
<p style="text-align: center;">Logistic Regression</p>	<p style="text-align: center;">Logistic Regression</p>  <table border="1" style="margin-left: auto; margin-right: auto;"> <tr> <td></td> <td>Normal</td> <td>Fraud</td> </tr> <tr> <td>Normal</td> <td>85060</td> <td>89</td> </tr> <tr> <td>Fraud</td> <td>166</td> <td>85274</td> </tr> </table>		Normal	Fraud	Normal	85060	89	Fraud	166	85274
	Normal	Fraud								
Normal	85060	89								
Fraud	166	85274								
<p style="text-align: center;">Decision Tree</p>	<p style="text-align: center;">Decision Tree</p>  <table border="1" style="margin-left: auto; margin-right: auto;"> <tr> <td></td> <td>Normal</td> <td>Fraud</td> </tr> <tr> <td>Normal</td> <td>85122</td> <td>27</td> </tr> <tr> <td>Fraud</td> <td>25</td> <td>85415</td> </tr> </table>		Normal	Fraud	Normal	85122	27	Fraud	25	85415
	Normal	Fraud								
Normal	85122	27								
Fraud	25	85415								



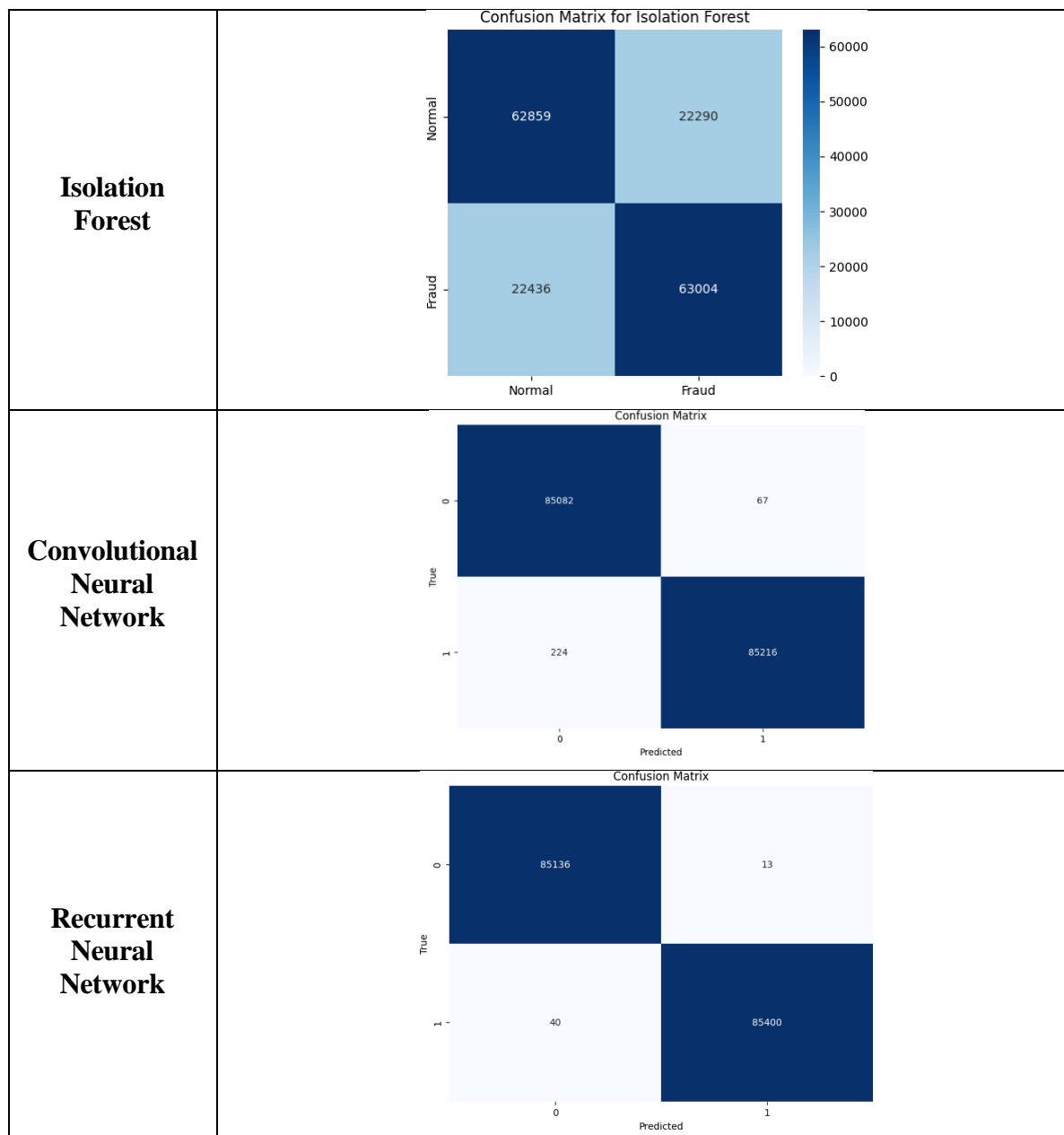


TABLE6. CLASSIFICATION REPORT

Models	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.9985	0.9989	0.9980	0.9985
Decision Tree	0.9996	0.9996	0.9997	0.9996
Random forest	0.9998	0.9999	0.9997	0.9998
K-NN	0.9993	0.9987	0.9999	0.9993
Cat Boost	0.9996	0.9998	0.9995	0.9996
SVM	0.9987	0.9991	0.9984	0.9987
Isolation Forest	0.7378	0.7386	0.7374	0.7321
CNN	0.9982	0.9973	0.9991	0.9985
RNN	0.9996	0.9998	0.9995	0.9996

CONCLUSION

Our exploration of the topic of credit card fraud detection by using machine and deep learning approaches unfolded as a journey full of understanding new concepts of ML and DL. Coming to the performance of the algorithm concludes that the Isolation factor performs poorer as compared to the rest algorithms. While other algorithms have performed exceptionally well in

each manner, the Random Forest algorithm proves its ability to perform better results. The main purpose of our research is to find out the best way to detect credit card fraud and that could make predictions for unseen data. Our findings have direct implications for the financial industry, where credit card fraud is a significant concern. The models developed in this research can be deployed in financial institutions to strengthen security measures and reduce financial loss due to fraud.

REFERENCES

1. N. K. Trivedi, S. Simaiya, U. K. Lilhore, and S. K. Sharma, "An Efficient Credit Card Fraud Detection Model Based on Machine Learning Methods," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 5, 2020.
2. D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, "Credit Card Fraud Detection -Machine Learning methods," in 2019 18th International Symposium INFOTEH JAHORINA (INFOTEH), East Sarajevo, Bosnia and Herzegovina: IEEE, Mar. 2019, pp. 1–5. doi: 10.1109/INFOTEH.2019.8717766.
3. A. Rb and S. K. Kr, "Credit card fraud detection using artificial neural network," *Glob. Transit. Proc.*, vol. 2, no. 1, pp. 35–41, Jun. 2021, doi: 10.1016/j.gltp.2021.01.006.
4. A. Thakarke, S. Ugale, S. Nale, and D. M. Dixit, "Credit Card Fraud Detection Using Bagging and Boosting Algorithms," vol. 10, no. 7, 2020.
5. G. Mhatre, O. Almeida, D. Mhatre, and P. Joshi, "Credit Card Fraud Detection Using Hidden Markov Model," vol. 5, 2014.
6. F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan, and M. Ahmed, "Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms," *IEEE Access*, vol. 10, pp. 39700–39715, 2022, doi: 10.1109/ACCESS.2022.3166891.
7. Ghosh and Reilly, "Credit card fraud detection with a neural network," in *Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences HICSS-94*, Wailea, HI, USA: IEEE Comput. Soc. Press, 1994, pp. 621–630. doi: 10.1109/HICSS.1994.323314.
8. T. T. Nguyen, H. Tahir, M. Abdelrazek, and A. Babar, "Deep Learning Methods for Credit Card Fraud Detection".
9. P. Gamini, S. T. Yerramsetti, G. D. Darapu, V. K. Pentakoti, and V. P. Raju, "Detection of Credit Card Fraudulent Transactions using Boosting Algorithms," vol. 8, no. 2, 2021.