

## CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING AND DEEP LEARNING TECHNIQUES

Hariom Sharma\*<sup>1</sup>

\*<sup>1</sup>UG, Department Of Information Technology B.K Birla College Kalyan,  
(Empowered Autonomous Status), India.

DOI : <https://www.doi.org/10.56726/IRJMETS45225>

### ABSTRACT

In today's socio-economic scenario, people rely heavily on credit cards. Moreover, credit cards are a requisite financial tool that enables its holders to make assets. It is true that credit cards, as a new method of payment, have become socially amenable to the masses. But nowadays, improvements in technology lead to growth in illegal activities. During credit card transactions many fraudsters can breach security and make fraudulent transactions to withdraw or transfer funds from one's account or e-wallets. The relevant literature presents many machine learning-based approaches for credit card detection, such as Logistic Regression, Decision Tree, CatBoost, Random Forest, Support Vector Machine, KNN, RNN, and CNN. The main focus has been to apply the recent development of deep learning algorithms for this purpose. Comparative analysis of both machine learning and deep learning algorithms was performed to find efficient outcomes. A machine learning algorithm was first applied to the dataset, which improved the accuracy of detection of the fraud to some extent. The proposed approaches can be implemented effectively for the real-world detection of credit card fraud.

**Keywords:** Fraud, Machine And Deep Learning Techniques, Analysis.

### I. INTRODUCTION

Now these days digital, statistics are very easily available throughout the world because of digital online availability. All the information that also has a large volume, wide range, frequency, as well as importance is stored from small to large organizations over the cloud. The whole information is available from massive amounts of sources such as followers on social media, customer order behavior, likes, and shares. White-collar crime is an ever-increasing problem with far-reaching consequences for the finance sector, business institutions as well as governments.

Enhanced card transactions had already appreciated a heavy emphasis on communication technology. When credit card transactions are by far the most prevalent form of transaction for offline and online payments, the rate of card fraud accelerates as well. There are over 389,000 cases reported to the FTC in 2021 (Federal Trade Commission).

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 including random forest, tree classifiers, supporting vector machine, and logistic regression including CatBoost classifier approaches, on even a highly skewed credit card fraud database. This complete research paper is divided into different sections including; the introduction portion, related activities, credit card fraud obfuscation techniques for machine and deep learning, Result analysis, Future Work, and Conclusion.

## II. RELATED WORK

In this paper, some determined algorithms based on machine learning and neural networks are suggested and implemented to forecast credit card fraud detection. Machine learning approaches play a crucial role throughout numerous efficient areas for data processing. The allocation of datasets used in detection is balanced. Detection of card fraud is focused on an interpretation of the card actions in purchases. This article [1] evaluates several innovative methods of machine learning; supporting vector machines including random forests together with logistical regression as part of an attempt to better detect fraud when applying neural network and logistic regression to identity fraud detection issues. There are several different types of credit card fraud. One of these is stealing a physical card while the other is the stealing of confidential credit card information like account number, CVV key, card type, and many others. Businesses are using different techniques for machine learning that identify increasing transactions that constitute illegitimate and those that are not.

## III. 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. The dataset contains This dataset contains transactions made by European 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.

**TABLE 1.** Dataset 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 transaction is 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:

**TABLE 2.** Steps to perform experiments

SN.	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 hyperplane that best separates data points of different classes in feature space while maximizing the margin between the hyperplane 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 spill. 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 dataset 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 hyperparameter 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 graph that summarizes the performance of a classification algorithm by comparing the actual and predicted class labels for the dataset. The matrix provides insights into how many predictions were correct and how many were incorrect, breaking them down into different categories.

**TABLE 3.** 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. DEEP LEARNING 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.

**TABLE 4.** Layers of CNN

Layers	Description
Input	It takes one-dimensional data as an input
Hidden	Convolutional: 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**

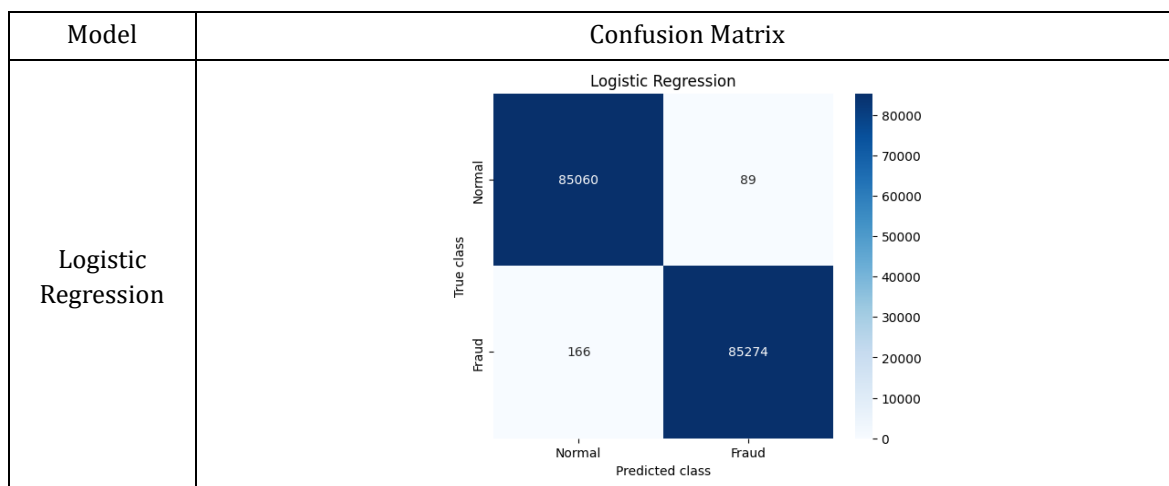
It is a type of artificial neural network architecture for processing sequences of data. RNNs are characterized by their ability to operate on variable-length sequences and are particularly suitable for tasks involving temporal dependencies and sequential patterns.

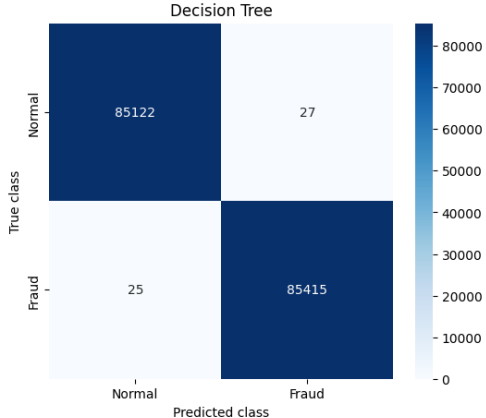
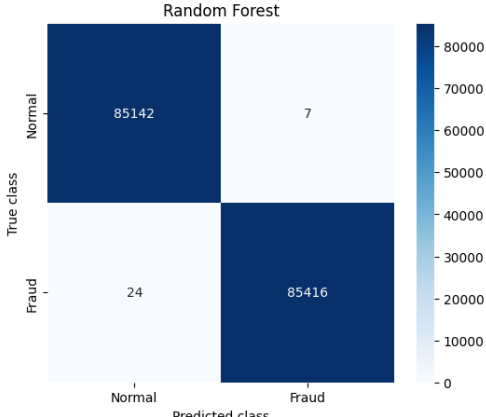
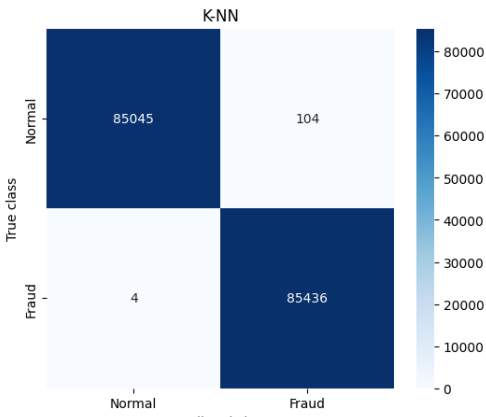
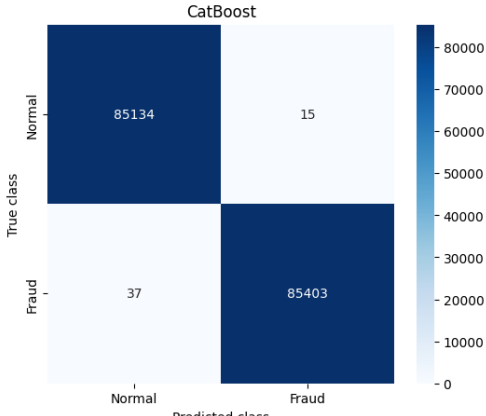
Model building of CNN and RNN is a crucial step in the entire experiment as it defines the number of layers along with their parameter used. A larger number of layers has a greater capacity to represent complex functions, which can lead to improved performance on challenging tasks as compared to a model having less number of layers. Training and Testing of the model have been performed using the layers with certainly required hyperparameters 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.

**IV. 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.

**TABLE 5.** Confusion Matrix



Decision Tree	 <p>Decision Tree</p> <table border="1"> <tr> <td>True class \ Predicted class</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>	True class \ Predicted class	Normal	Fraud	Normal	85122	27	Fraud	25	85415
True class \ Predicted class	Normal	Fraud								
Normal	85122	27								
Fraud	25	85415								
Random Forest	 <p>Random Forest</p> <table border="1"> <tr> <td>True class \ Predicted class</td> <td>Normal</td> <td>Fraud</td> </tr> <tr> <td>Normal</td> <td>85142</td> <td>7</td> </tr> <tr> <td>Fraud</td> <td>24</td> <td>85416</td> </tr> </table>	True class \ Predicted class	Normal	Fraud	Normal	85142	7	Fraud	24	85416
True class \ Predicted class	Normal	Fraud								
Normal	85142	7								
Fraud	24	85416								
K-NN	 <p>K-NN</p> <table border="1"> <tr> <td>True class \ Predicted class</td> <td>Normal</td> <td>Fraud</td> </tr> <tr> <td>Normal</td> <td>85045</td> <td>104</td> </tr> <tr> <td>Fraud</td> <td>4</td> <td>85436</td> </tr> </table>	True class \ Predicted class	Normal	Fraud	Normal	85045	104	Fraud	4	85436
True class \ Predicted class	Normal	Fraud								
Normal	85045	104								
Fraud	4	85436								
CatBoost	 <p>CatBoost</p> <table border="1"> <tr> <td>True class \ Predicted class</td> <td>Normal</td> <td>Fraud</td> </tr> <tr> <td>Normal</td> <td>85134</td> <td>15</td> </tr> <tr> <td>Fraud</td> <td>37</td> <td>85403</td> </tr> </table>	True class \ Predicted class	Normal	Fraud	Normal	85134	15	Fraud	37	85403
True class \ Predicted class	Normal	Fraud								
Normal	85134	15								
Fraud	37	85403								

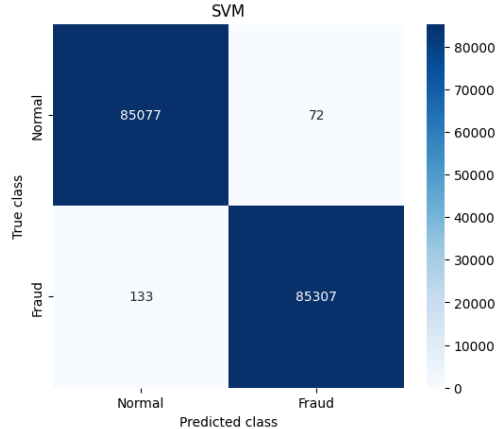
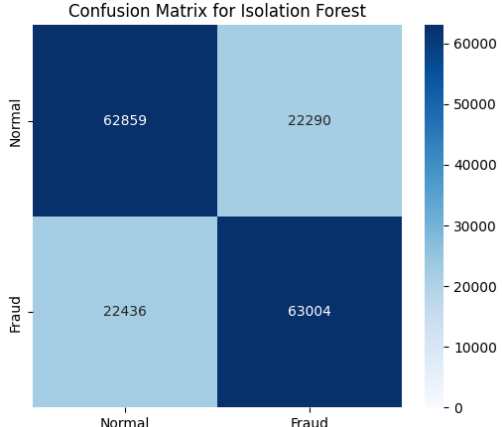
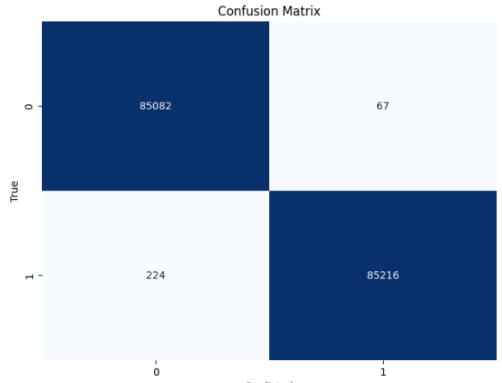
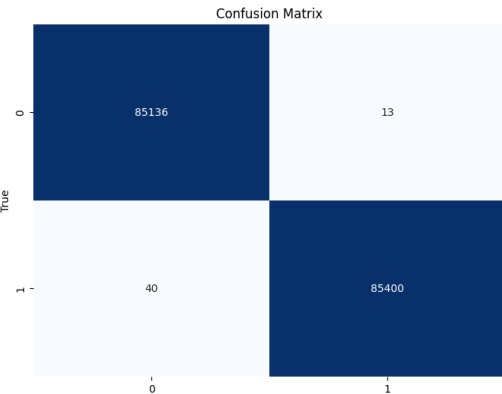
<p>Support Vector Machine</p>	 <p>SVM</p> <table border="1"> <tr> <td>True class \ Predicted class</td> <td>Normal</td> <td>Fraud</td> </tr> <tr> <td>Normal</td> <td>85077</td> <td>72</td> </tr> <tr> <td>Fraud</td> <td>133</td> <td>85307</td> </tr> </table>	True class \ Predicted class	Normal	Fraud	Normal	85077	72	Fraud	133	85307
True class \ Predicted class	Normal	Fraud								
Normal	85077	72								
Fraud	133	85307								
<p>Isolation Forest</p>	 <p>Confusion Matrix for Isolation Forest</p> <table border="1"> <tr> <td>True class \ Predicted class</td> <td>Normal</td> <td>Fraud</td> </tr> <tr> <td>Normal</td> <td>62859</td> <td>22290</td> </tr> <tr> <td>Fraud</td> <td>22436</td> <td>63004</td> </tr> </table>	True class \ Predicted class	Normal	Fraud	Normal	62859	22290	Fraud	22436	63004
True class \ Predicted class	Normal	Fraud								
Normal	62859	22290								
Fraud	22436	63004								
<p>Convolutional Neural Network</p>	 <p>Confusion Matrix</p> <table border="1"> <tr> <td>True \ Predicted</td> <td>0</td> <td>1</td> </tr> <tr> <td>0</td> <td>85082</td> <td>67</td> </tr> <tr> <td>1</td> <td>224</td> <td>85216</td> </tr> </table>	True \ Predicted	0	1	0	85082	67	1	224	85216
True \ Predicted	0	1								
0	85082	67								
1	224	85216								
<p>Recurrent Neural Network</p>	 <p>Confusion Matrix</p> <table border="1"> <tr> <td>True \ Predicted</td> <td>0</td> <td>1</td> </tr> <tr> <td>0</td> <td>85136</td> <td>13</td> </tr> <tr> <td>1</td> <td>40</td> <td>85400</td> </tr> </table>	True \ Predicted	0	1	0	85136	13	1	40	85400
True \ Predicted	0	1								
0	85136	13								
1	40	85400								

TABLE 6. 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
CatBoost	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

## V. FUTURE WORK

Trying out different Machine and Deep Learning Algorithms to find the best way of finding fraud in transactions. Further in this field of research one can develop a system for real-time detection of frauds.

## VI. 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.

## VII. 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.gltip.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.