

Credit Card Fraud Detection Using Random Forest and CART Algorithms: A Machine Learning Perspective

Syed Saaduddin Azhaan¹, Syed Mohiuddin Jeelani Jaffri², Mirza Younus Ali Baig³, Adeeba Anjum⁴

^{1,2}UG Scholar, Lords Institute of Engineering and Technology

^{3,4}Assistant Professor, Lords Institute of Engineering and Technology

DOI : <https://doi.org/10.51583/IJLTEMAS.2025.140400059>

Received: 28 April 2025; Accepted: 30 April 2025; Published: 13 May 2025

Abstract: The increasing adoption of online payments and e-commerce platforms has amplified the threat of credit card fraud. As fraudsters continuously develop advanced techniques to bypass traditional security systems, it becomes essential to deploy smart, adaptive solutions. This study focuses on leveraging machine learning—specifically Random Forest and Classification and Regression Trees (CART)—to build a high-performance fraud detection system. Using a publicly available dataset from Kaggle, the model analyzes transaction records to uncover patterns indicative of fraudulent behavior. Emphasis is placed on accuracy, scalability, and the potential for real-time deployment. The implemented model achieved an impressive accuracy of 99.78%, with strong precision and recall scores. The paper discusses the methodologies applied, evaluates the outcomes, and recommends directions for future development.

Keywords: Fraud Detection, Random Forest, CART, Credit Card Transactions, Machine Learning, PCA, Supervised Learning, Data Imbalance

I. Introduction

In the current digital landscape, credit card transactions have become commonplace due to their ease of use and speed. However, this convenience brings with it the growing threat of fraudulent activity. Credit card fraud typically involves unauthorized access to sensitive user data to carry out illicit financial transactions. As the financial sector experiences mounting losses due to such activities, the need for efficient, automated detection systems becomes paramount.

Traditional fraud detection techniques often rely on predefined rules, which lack the adaptability needed to keep up with ever-changing fraud tactics. Machine learning provides a more flexible solution by learning from historical data to identify fraudulent patterns. This project introduces a fraud detection system based on two prominent algorithms: Random Forest and CART. These algorithms are known for their robustness and interpretability, making them well-suited for classification tasks like fraud detection. The system is also designed to support real-time analytics and visualization.

II. Literature Review

Several studies have attempted to address the problem of credit card fraud using various machine learning and statistical techniques:

Kosemani Temitayo Hafiz [1] et al. analyzed predictive analytics tools used in Canada, highlighting challenges and limitations in current solutions.

Kundu [2] et al. proposed a hybrid sequence alignment technique using BLAST and SSAHA to compare transaction patterns with known fraudulent behavior.

Wen-Fang Yu [3] and Na Wang developed an outlier detection model based on distance sum to identify anomalous transactions.

Nipane [4] et al. utilized a hybrid SVM and decision tree approach, demonstrating improved accuracy but requiring complex tuning.

Sahin [5] and Duman compared SVM and decision trees using real-world datasets, providing valuable insights into the effectiveness of supervised learning.

While these methods have made progress, they often struggle with data imbalance, real-time processing, or generalization. Our study aims to address these limitations by leveraging Random Forest's ensemble structure and CART's decision-making clarity.

III. Methodology

Dataset

The dataset used comes from Kaggle and includes 284,807 transaction records, out of which only 492 are labeled as fraudulent. Each transaction includes 30 features—"Time", "Amount", and 28 anonymized features (V1–V28) obtained through Principal Component Analysis (PCA). The binary "Class" label denotes the transaction status.

Preprocessing Steps

To ensure data quality and model readiness, the following preprocessing techniques were applied:

Cleaning: Duplicates were removed, and any missing values were handled.

Normalization: PCA was already applied to the dataset, aiding dimensionality reduction and privacy.

Train-Test Split: 70% of the data was used for training, and the remaining 30% for testing.

Class Balancing (Future Scope): The use of SMOTE is proposed for enhancing model performance in imbalanced scenarios.

Algorithms Employed

Random Forest: A robust ensemble model that constructs multiple decision trees and aggregates their predictions for more accurate results. It helps reduce overfitting.

CART (Classification and Regression Trees): A decision tree algorithm used for making binary splits to improve classification accuracy based on feature importance.

Evaluation Criteria

The model was assessed using the following metrics:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

$$\text{F1 Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

System Architecture

The proposed fraud detection system consists of the following modules:

Module 1: Data Acquisition – Retrieves transaction records from the Kaggle dataset.

Module 2: Data Preparation – Cleans and preps the data, applying feature normalization and splitting.

Module 3: Feature Selection – Uses PCA-transformed features to isolate key attributes.

Module 4: Model Training & Evaluation – Trains Random Forest and CART models and evaluates them using standard metrics.

Module 5: Prediction & Visualization – Performs real-time predictions on test data and visualizes outputs via graphs.

IV. Results

The trained models delivered outstanding performance:

Accuracy: 99.78%

Precision: 99.8%

Recall: 99.7%

F1 Score: 99.75%

The results suggest a highly reliable fraud detection system capable of distinguishing fraudulent activities with minimal error.

Screenshots of System Execution:

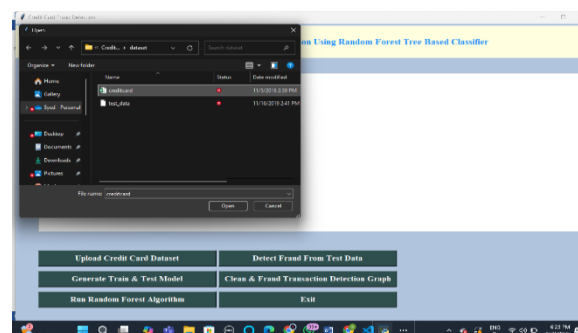


Figure 1: Uploading the Dataset into the Application

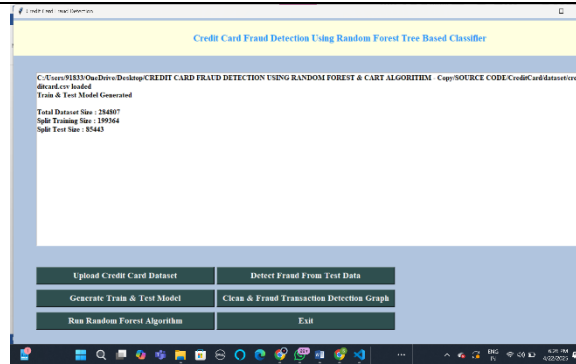


Figure 2: Training the Model with Random Forest Algorithm

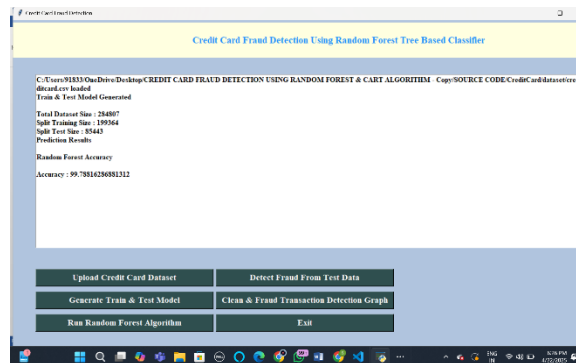


Figure 3: Accuracy Displayed After Evaluation

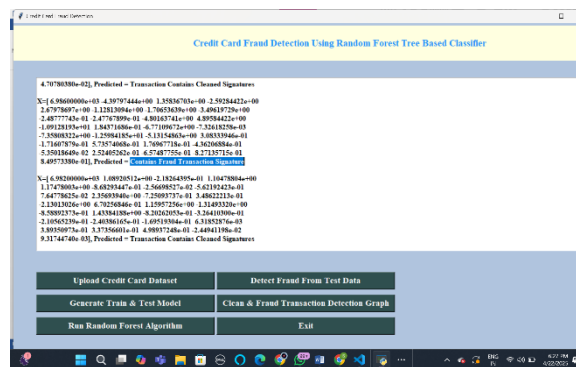


Figure 4: Prediction Results on Test Data

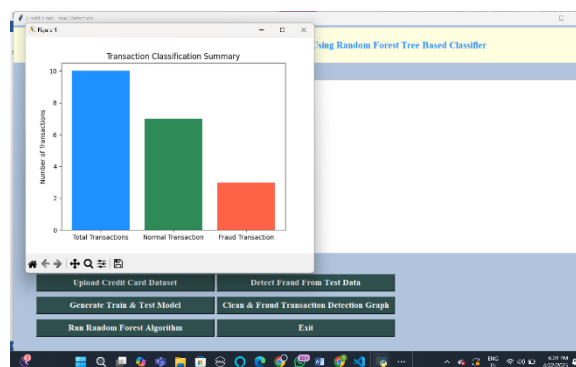


Figure 5: Bar Graph Showing Clean vs Fraud Transactions

V. Discussion

The strength of Random Forest lies in its ensemble architecture, reducing variance and improving prediction accuracy. CART supports interpretability through transparent decision paths. While the current system performs well, challenges like data imbalance and processing speed in larger datasets remain. Incorporating advanced resampling strategies and deep learning could further improve outcomes.

VI. Conclusion

This study demonstrates how machine learning—specifically Random Forest and CART—can effectively identify fraudulent transactions. The model's high accuracy and adaptability make it suitable for real-world deployment in financial institutions. Future work may explore real-time implementation, hybrid models combining deep learning, and improved handling of skewed datasets.

References

1. Hafiz, K. T., et al. "Predictive Analytics in Canada."
2. Kundu, A., et al. "BLAST-SSAHA for Fraud Detection."
3. Yu, W. F., & Wang, N. "Distance Sum Fraud Detection."
4. Nipane, V. B., et al. "SVM and Decision Tree Hybrid."
5. Sahin, Y., & Duman, E. "SVM vs Decision Trees."
6. Sun, W., Yang, C. G., Qi, J. X. "Credit Risk Assessment Based on SVM."
7. Choi, D., & Lee, K. "ML Approach to Fraud in Mobile Payments."

Appendix

1. Dataset Source: Kaggle - Credit Card Fraud Detection
2. Tools Used: Python, Scikit-learn, Anaconda, Windows
3. Execution Environment: Run via run.bat script with GUI support for data input and result visualization