# Credit Card Fraud Detection Using Machine Learning

Gokul G (22BDS026)

IIIT Dharwad, Karnataka, India
22bds026@iiitdwd.ac.in

**Abstract.** Credit card fraud is a growing concern in the financial sector. Detecting fraudulent transactions in real-time requires robust machine learning models capable of handling imbalanced datasets. This paper presents a hybrid machine learning approach combining supervised (XGBoost, Random Forest) and unsupervised (Isolation Forest) models for fraud detection. A Streamlit dashboard is also developed for real-time prediction and visualization. Experimental results demonstrate the effectiveness of the hybrid model in achieving high recall and accuracy.

**Keywords:** Credit Card Fraud Detection · Machine Learning · Hybrid Model · Streamlit · Anomaly Detection

## 1 Introduction

The rise of digital payments has increased the risk of fraudulent transactions. Credit card fraud detection aims to identify rare fraudulent activities within large transactional datasets. Due to class imbalance and evolving fraud patterns, traditional detection methods often fail. This study proposes a hybrid model that leverages both supervised and unsupervised learning techniques for effective fraud detection.

## 2 Related Work

Existing studies have applied Logistic Regression, Random Forest, and XGBoost to detect fraud. While Logistic Regression provides interpretability, it often fails on nonlinear patterns. Tree-based methods like Random Forest and boosting methods like XGBoost provide better performance on tabular data. Unsupervised methods such as Isolation Forest detect anomalies without labeled data. Hybrid models combine these strengths to improve detection robustness.

## 3 Dataset and Preprocessing

### 3.1 Dataset Description

The dataset consists of 284,807 credit card transactions, including 492 fraudulent cases. Key features include:

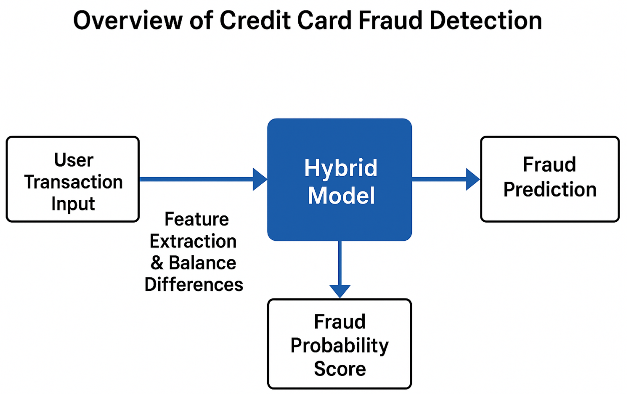

Overview of Credit Card Fraud Detection
User Transaction Input
Feature Extraction & Balance Differences
Hybrid Model
Fraud Prediction
Fraud Probability Score

**Fig. 1.** Overview of Credit Card Fraud Detection

- step, type, amount
- nameOrig, oldbalanceOrg, newbalanceOrig
- nameDest, oldbalanceDest, newbalanceDest
- isFraud, isFlaggedFraud

### 3.2 Data Preprocessing

Missing values were handled and new features were computed:

$$\text{balanceDiffOrig} = \text{oldbalanceOrg} - \text{newbalanceOrig}$$
$$\text{balanceDiffDest} = \text{newbalanceDest} - \text{oldbalanceDest}$$

Top senders and receivers were analyzed, and zero balance anomalies in TRANSFER and CASH_OUT transactions were noted.

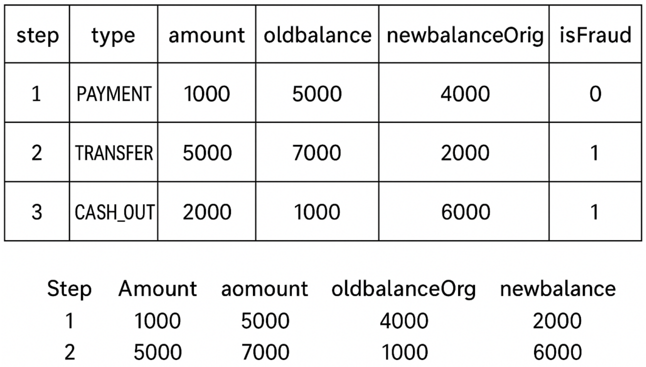
## 4 Exploratory Data Analysis

- Distribution of transaction types
- Fraud occurrence by type (TRANSFER and CASH_OUT dominate)
- Log-scaled amount histogram
- Boxplot: amount vs fraud for transactions ¡50k

## 5 Methodology

### 5.1 Algorithm Selection

Selected algorithms:



| step | type | amount | oldbalance | newbalanceOrig | isFraud |
|---|---|---|---|---|---|
| 1 | PAYMENT | 1000 | 5000 | 4000 | 0 |
| 2 | TRANSFER | 5000 | 7000 | 2000 | 1 |
| 3 | CASH_OUT | 2000 | 1000 | 6000 | 1 |

| Step | Amount | aomount | oldbalanceOrg | newbalance |
|---|---|---|---|---|
| 1 | 1000 | 5000 | 4000 | 2000 |
| 2 | 5000 | 7000 | 1000 | 6000 |

**Fig. 2.** Example of Transaction Features and Fraud Labels

- Logistic Regression – baseline linear model
- Random Forest – ensemble tree-based
- XGBoost – gradient boosting
- Isolation Forest – unsupervised anomaly detection
- Hybrid Model – combines Random Forest, XGBoost, and Isolation Forest for robustness

### 5.2 Model Training

- 80% training, 20% testing split
- Standard scaling for numeric features
- One-hot encoding for categorical features
- Metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC

## 6 Results and Discussion

**Table 1.** Model Performance Comparison

| Model | Accuracy | Precision | Recall | F1-score | ROC-AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.963 | 0.974 | 0.0304 | 0.874 | 0.588 |
| Random Forest | 0.999 | 0.998 | 0.998 | 0.996 | 0.997 |
| XGBoost | 0.999 | 0.999 | 0.783 | 0.986 | 0.873 |
| Isolation Forest | 0.997 | 0.845 | 0.008 | 0.008 | 0.008 |
| Hybrid Model | **0.999** | **0.999** | **0.794** | **0.982** | **0.878** |

Observations:

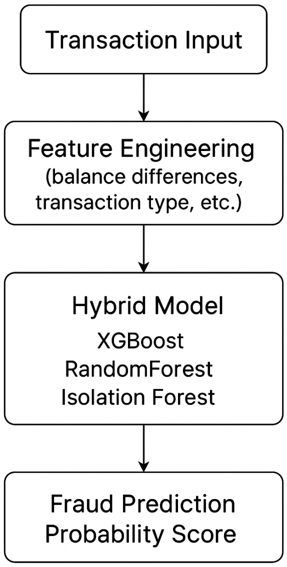


**Fig. 3.** Hybrid Model Architecture for Fraud Detection

- Hybrid model achieves best balance of recall and precision.
- Supervised models capture patterns, while Isolation Forest detects anomalies.
- Fraudulent transactions are mostly TRANSFER and CASH_OUT types.

## 7 Streamlit Dashboard

An interactive dashboard was developed to:

- Input transaction details
- Predict fraud probability
- Display transaction summary and visualizations

## 8 Conclusion

The hybrid model successfully integrates supervised and unsupervised learning for credit card fraud detection. It demonstrates high recall and robustness against unseen fraud patterns. Future work includes batch predictions and advanced visualizations.

## Acknowledgments

We thank the dataset providers on Kaggle and the Streamlit community for supporting deployment.

## References

1. Kaggle Credit Card Fraud Dataset: `https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud`
2. Scikit-learn documentation: `https://scikit-learn.org/stable/`
3. Streamlit documentation: `https://docs.streamlit.io/`
4. XGBoost documentation: `https://xgboost.readthedocs.io/`