iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.68703

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Credit Card Fraud Detection System

Bulbul Gupta[1], Anushka Pundir[2], Divyanshu Singh Bisht[3], Barkha Nandwana[4]
Department of Engineering and Technology, Sharda University 201308

Abstract: ***Credit card fraud poses a significant threat to financial security, leading to substantial monetary losses and risks for both consumers and financial institutions. Detecting fraudulent transactions accurately is challenging due to the high class imbalance in credit card transaction datasets. This study employs machine learning techniques to detect fraudulent transactions using the publicly available creditcard.csv dataset. Data preprocessing steps such as feature scaling, handling imbalanced data, and exploratory data analysis (EDA) are applied to enhance model performance. The study primarily utilizes Logistic Regression as a baseline model, comparing its performance with other machine learning classifiers such as Random Forest, Decision Tree, and Support Vector Machine (SVM). The evaluation metrics include accuracy, precision, recall, F1-score, and AUC-ROC curve analysis to ensure effective fraud detection. The results indicate that balancing the dataset and selecting optimal models can significantly improve fraud detection performance. This research contributes to advancing machine learning-based fraud detection systems and enhancing financial transaction security.***

Keywords: Credit card fraud detection, Financial security, Machine learning, Imbalanced data, Fraudulent transactions, Data preprocessing, Model evaluation.

I. INTRODUCTION

In the digital age, online transactions have become an essential part of daily life, leading to an increase in credit card fraud. Fraudulent transactions not only cause financial losses but also affect consumer trust and the reputation of financial institutions. According to industry reports, credit card fraud accounts for billions of dollars in losses annually. As traditional rule-based fraud detection systems struggle to keep up with evolving fraud patterns, machine learning (ML) techniques have emerged as a powerful solution.

This research focuses on applying machine learning algorithms to detect fraudulent transactions using a publicly available dataset. The key challenges in fraud detection include:

1) High Class Imbalance: Fraudulent transactions make up only 0.17% of the dataset, making it difficult for models to learn fraud patterns effectively.
2) Feature Engineering: The dataset consists of anonymized numerical features (V1-V28), requiring careful preprocessing.
3) Model Selection: Different ML models exhibit varying performance, and choosing the right one is critical for improving fraud detection.

In this study, we analyze credit card transactions using Logistic Regression as a baseline model and compare it with Decision Tree, Random Forest, and Support Vector Machine (SVM). We evaluate model performance based on accuracy, precision, recall, F1-score, and AUC-ROC to determine the most effective fraud detection approach. The results demonstrate that handling class imbalance and selecting appropriate ML models can significantly improve fraud detection accuracy.

The rest of the paper is structured as follows: Section 2 presents related work, Section 3 describes the dataset and preprocessing steps, Section 4 discusses the methodology and machine learning models, Section 5 analyzes experimental results, and Section 6 concludes the study with future research directions.

A. Advantages (Updated for Your Dataset and Models)

- Random Forest selects the best features from a random subset of data, ensuring a more robust and accurate fraud detection model.
- Binary classification is applied, where fraudulent transactions are labeled as 1 (positive case) and legitimate transactions as 0 (negative case) in the dataset.
- Several fraud detection techniques have been developed using artificial intelligence, data mining, fuzzy logic, and machine learning to enhance fraud detection in credit card transactions.
- Credit card fraud detection is a challenging yet critical problem that requires advanced computational models. In this research, machine learning algorithms are used to build an effective fraud detection system.

©IJRASET: All Rights are Reserved

1) Advancements in Machine Learning for Fraud Detection

With the rise of machine learning techniques, fraud detection has significantly improved. ML-based systems analyze large volumes of transaction data to classify them as genuine or fraudulent. Online businesses benefit from these models as they can accurately identify fraudulent transactions based on historical chargebacks and user behaviour.

In the creditcard.csv dataset, transaction features such as transaction amount, anonymized attributes (V1-V28), and time are used to develop predictive models. Machine learning enables fraud detection by assigning a fraud probability score based on multiple features. Instead of manually determining fraud indicators, ML models self-learn from training data, making them highly adaptable to new fraud techniques.

2) Why Use Random Forest for Fraud Detection?

Random Forest, a supervised learning algorithm, has proven to be highly effective in credit card fraud detection due to its ability to handle imbalanced datasets and reduce overfitting. Key benefits include:

- Higher accuracy compared to traditional models like Decision Trees by combining multiple trees for better predictions.
- Feature importance selection that helps in identifying the most relevant attributes for fraud detection, improving model efficiency.
- Ability to handle class imbalance, making it suitable for datasets with extremely low fraud cases (0.17 percent in our dataset).
- Reduction of correlation issues by using a subset of features at each split, preventing over-reliance on a few attributes.
- Scalability, making it suitable for large datasets with multiple variables.

B. Scope of the Proposed Work

In this proposed project, a model is designed to detect fraudulent activities in credit card transactions. This system can provide essential features required to differentiate between fraudulent and legitimate transactions.

As technology evolves, tracking the modelling and patterns of fraudulent transactions becomes increasingly complex. Traditional fraud detection methods rely on rule-based systems, which struggle to adapt to emerging fraud techniques. With the rise of machine learning, artificial intelligence, and other advanced technologies, it is now feasible to automate the fraud detection process and minimize the manual effort involved in identifying suspicious transactions. This research focuses on developing a machine learning-based fraud detection system that can analyze transaction data in real time. The system leverages various supervised learning algorithms to improve the accuracy of fraud detection while reducing false positives. By utilizing the creditcard.csv dataset, the proposed model identifies key transaction patterns and classifies them as fraudulent or non-fraudulent.

The scope of this study extends beyond traditional detection techniques by incorporating advanced statistical methods, feature engineering, and model evaluation strategies. The goal is to enhance fraud detection efficiency and contribute to the ongoing efforts in securing financial transactions.

II. SOFTWARE AND HARDWARE REQUIREMENTS

A. Hardware

- OS – Windows 7, 8 and 10 (32 and 64 bit)
- RAM – 4GB

B. Software

- Python
- Anaconda

III. SYSTEM ARCHITECTURE OVERVIEW

1) Data Collection – The credit card transaction dataset is collected.
2) Data Preprocessing – Handling missing values, scaling, feature selection, and splitting data.
3) Model Training – Machine learning algorithms such as Random Forest, Logistic Regression, Decision Tree, and SVM are trained on the dataset.
4) Fraud Detection – The trained model predicts whether a transaction is fraudulent or legitimate.
5) Evaluation & Optimization – Model performance is evaluated using accuracy, precision, recall, and F1-score.
6) Deployment & Monitoring – The final model is deployed to detect fraudulent transactions in real-time.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

IV. LITERATURE SURVEY

1) Fraud Detection in Credit Card System Using SVM & Decision Tree

With the rapid expansion of electronic commerce, fraudulent activities in financial transactions have increased significantly, leading to major financial losses for both businesses and consumers. Credit card fraud remains one of the primary causes of financial loss worldwide. Various methods such as Decision Trees, Genetic Algorithms, Meta-learning strategies, Neural Networks, and Hidden Markov Models (HMM) have been proposed to detect fraudulent transactions.

This study explores the application of Support Vector Machine (SVM) and Decision Tree algorithms to enhance fraud detection. The hybrid approach aims to improve classification accuracy and reduce financial losses by identifying suspicious transactions more effectively. The combination of these models helps in distinguishing fraudulent transactions from legitimate ones with better precision.

2) Machine Learning-Based Approach to Financial Fraud Detection in Mobile Payment Systems

Mobile payment fraud occurs when unauthorized users exploit stolen credit card details or identities to make fraudulent transactions. The increasing use of smartphones and online payment platforms has contributed to the rise of such fraudulent activities. Detecting fraudulent mobile payments is challenging due to the vast amount of financial data generated every second.

This research proposes a machine learning-based fraud detection model using both supervised and unsupervised learning methods to analyze large volumes of transactions efficiently. The model incorporates feature selection and sampling techniques to enhance processing speed and improve fraud detection accuracy. Evaluation metrics such as F-measure and the ROC curve are used to validate the effectiveness of the proposed fraud detection model.

This literature survey highlights the importance of machine learning algorithms in credit card fraud detection. By leveraging advanced computational models, financial institutions can significantly reduce fraudulent transactions and enhance security in digital payment systems.

V. PURPOSE OF THE PROJECT:

The objective of this project is to develop a machine learning-based model for detecting fraudulent credit card transactions in real-time. Manually analyzing fraudulent transactions is impractical due to the enormous volume and complexity of financial data. However, machine learning models can effectively detect fraudulent activities when trained on informative and relevant features.

This project explores the use of supervised learning algorithms such as Random Forest, Logistic Regression, and Decision Tree to classify fraudulent and legitimate credit card transactions. The goal is to build a highly accurate fraud detection system that can minimize financial losses and increase awareness of fraudulent activities in online transactions.

By implementing machine learning techniques on the creditcard.csv dataset, this research aims to enhance fraud detection efficiency and contribute to the development of secure and automated financial transaction systems.

A. Packages

For data exploration, preprocessing, and implementing the **Random Forest algorithm**, the following Python packages are used:

1) NumPy – Used for handling numerical data and performing mathematical operations efficiently.
2) Pandas – Essential for reading, manipulating, and analyzing structured datasets. It is used to load and preprocess the creditcard.csv dataset.
3) Scikit-Learn – Provides various machine learning algorithms and preprocessing tools, including data scaling, model training, and performance evaluation.
4) Matplotlib & Seaborn – Used for visualizing data distributions, generating confusion matrices, and plotting fraud detection insights in an intuitive and color-coded format.
5) TensorFlow/Keras – Used for handling matrix operations and deep learning models if required. Although not mandatory for Random Forest, it can be used for neural network-based fraud detection.

VI. MODULES

The fraud detection system is divided into the following key modules:

1) Data Collection – Gathering transactional data for training the model.
2) Data Preprocessing – Cleaning and transforming raw data for analysis.
3) Feature Extraction – Selecting relevant attributes for fraud detection.
4) Evaluation Model – Training and testing machine learning models to detect fraudulent transactions.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

A. Data Collection

The dataset used in this research consists of credit card transaction records, which include both fraudulent and legitimate transactions. This step involves selecting a relevant subset of the available data to train and evaluate the fraud detection model.

In machine learning, problems begin with data—preferably large and diverse datasets containing multiple observations. The dataset used in this project is highly imbalanced, meaning fraudulent transactions make up a small fraction of the total dataset.

A critical aspect of fraud detection is having labeled data, where each transaction is categorized as fraudulent (1) or non-fraudulent (0). The presence of labeled data enables the model to learn patterns associated with fraudulent transactions and make predictions on new, unseen data.

```
[ ] Start coding or generate with AI.

[ ] import numpy as np
    import pandas as pd
    from sklearn.model_selection import train_test_split
    from sklearn.linear_model import LogisticRegression
    from sklearn.metrics import accuracy_score

[ ] # loading the dataset to a Pandas DataFrame
    credit_card_data = pd.read_csv('/content/credit_data.csv')

[ ] # first 5 rows of the dataset
    credit_card_data.head()
```

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	0.090794	-0.551600	-0.617801	-0.991390	-0.311169	1.468177	-0.470
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	-0.166974	1.612727	1.065235	0.489095	-0.143772	0.635558	0.463
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	0.207643	0.624501	0.066084	0.717293	-0.165946	2.345865	-2.890
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	-0.054952	-0.226487	0.178228	0.507757	-0.287924	-0.631418	-1.059
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	0.753074	-0.822843	0.538196	1.345852	-1.119670	0.175121	-0.451

B. Data Pre-processing

Data preprocessing is a crucial step in building an efficient fraud detection system. It ensures that the raw transactional data is transformed into a structured and meaningful format for analysis. This process consists of three key steps:

1) Formatting

Formatting involves structuring the dataset so that it can be easily processed by machine learning models. The credit card transaction dataset is stored in a CSV (Comma-Separated Values) file, which is widely used due to its simplicity and compatibility with data processing tools.

- Ensuring all numerical features are in the correct data type (float or integer).
- Converting categorical data (if any) into numerical values.
- Standardizing or normalizing transaction amounts for better model performance.

```
[4] # first 5 rows of the dataset
    credit_card_data.head()
```

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V1
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	0.090794	-0.551600	-0.6178
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	-0.166974	1.612727	1.0652
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	0.207643	0.624501	0.0660
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	-0.054952	-0.226487	0.1782
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	0.753074	-0.822843	0.5381

```
[5] credit_card_data.tail()
```

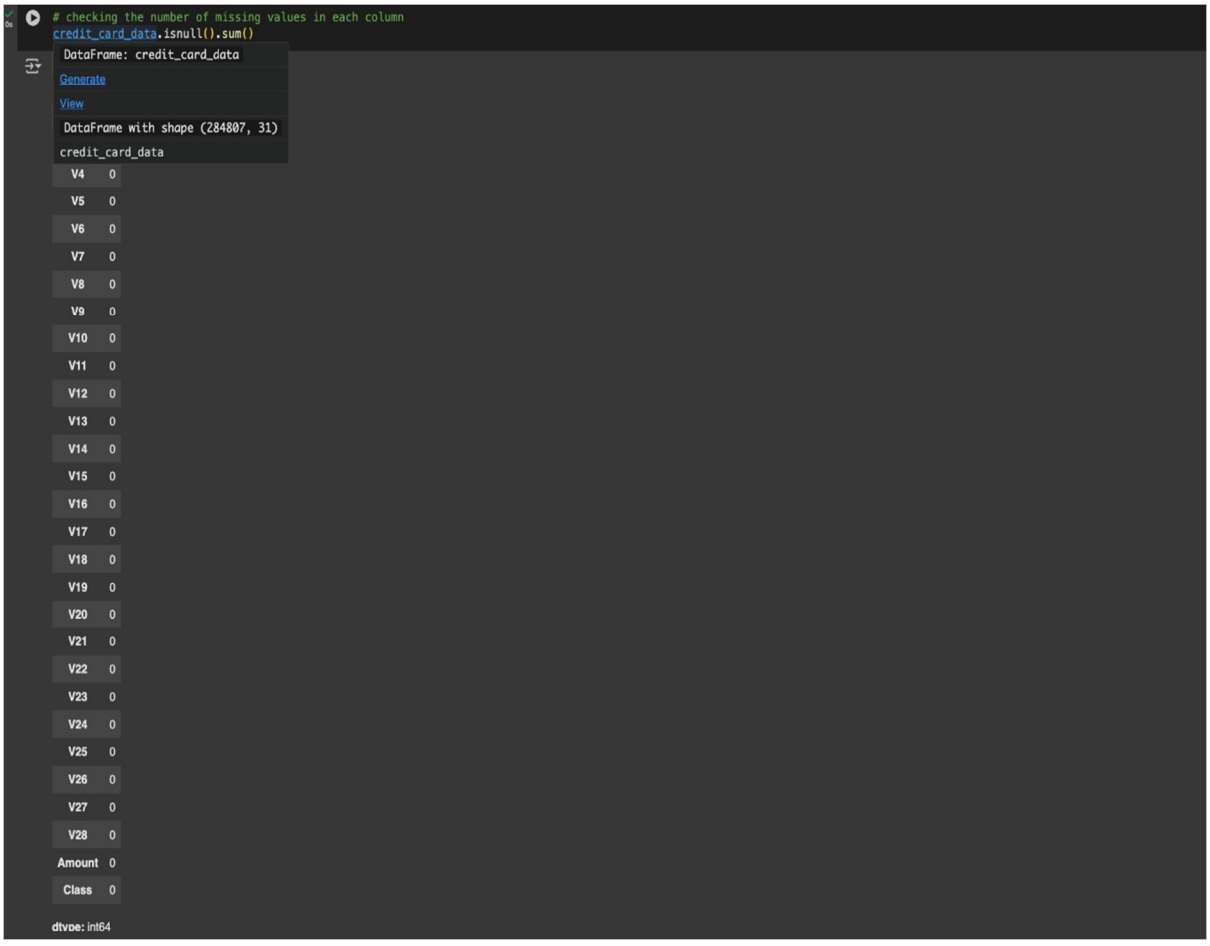
	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V1
284802	172786.0	-11.881118	10.071785	-9.834783	-2.066656	-5.364473	-2.606837	-4.918215	7.305334	1.914428	4.356170	-1.5931
284803	172787.0	-0.732789	-0.055080	2.035030	-0.738589	0.868229	1.058415	0.024330	0.294869	0.584800	-0.975926	-0.1501
284804	172788.0	1.919565	-0.301254	-3.249640	-0.557828	2.630515	3.031260	-0.296827	0.708417	0.432454	-0.484782	0.4116
284805	172788.0	-0.240440	0.530483	0.702510	0.689799	-0.377961	0.623708	-0.686180	0.679145	0.392087	-0.399126	-1.9338
284806	172792.0	-0.533413	-0.189733	0.703337	-0.506271	-0.012546	-0.649617	1.577006	-0.414650	0.486180	-0.915427	-1.0404

2) Cleaning

Data cleaning is essential for removing inconsistencies and handling missing values. In fraud detection, data quality is critical since inaccurate or incomplete information can lead to poor model predictions. The following steps are performed:

- Handling Missing Values: Transactions with missing or incomplete information are either filled with appropriate values or removed.
- Removing Duplicates: Duplicate transactions are identified and removed to prevent bias in model training.
- Encoding Categorical Data: If there are categorical variables (e.g., transaction types), they are converted into numerical representations.

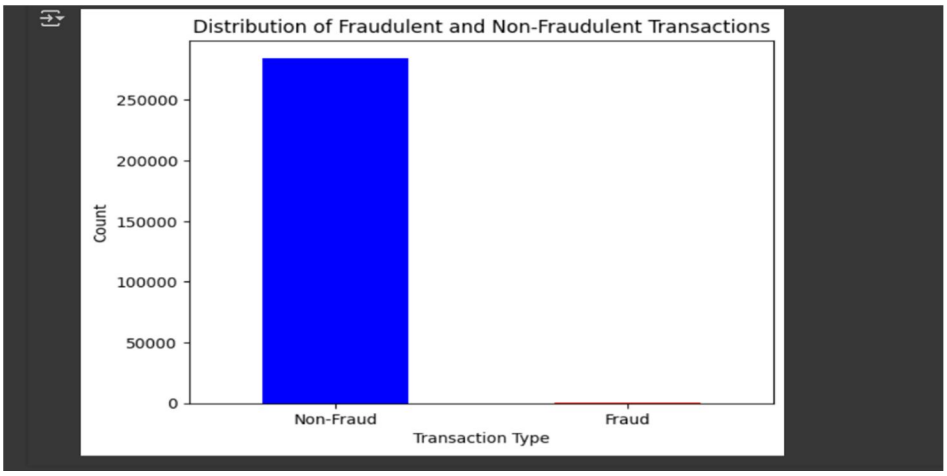
Since credit card fraud datasets are often highly imbalanced, special techniques such as oversampling (adding more fraudulent cases) or undersampling (removing excess legitimate cases) may be used to balance the dataset.



3) Sampling

Due to the large volume of financial transactions, sampling techniques help in reducing computational complexity while preserving key data patterns. In this research, Stratified Sampling is used, ensuring that the proportion of fraudulent and non-fraudulent transactions remains consistent within the training and testing datasets.

By implementing effective data preprocessing techniques, the system ensures that the machine learning model receives high-quality input data, improving the accuracy of fraud detection.



C. Data Exploration

Data Exploration

```
# first 5 rows of the dataset
credit_card_data.head()
```

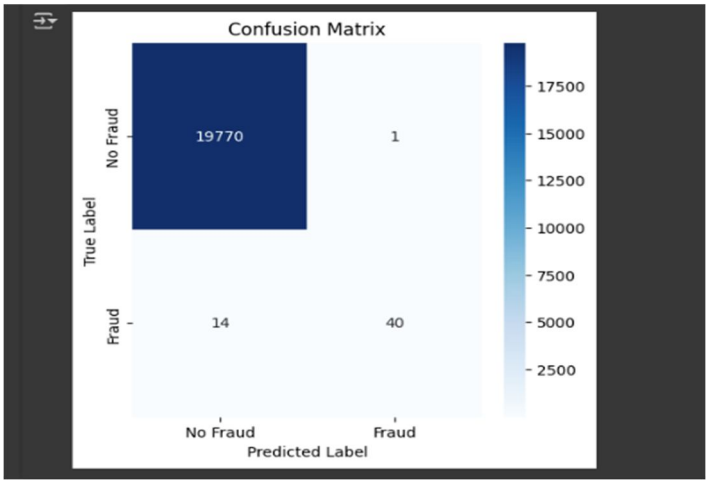
	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	0.090794	-0.551600	-0.617801	-0.991390	-0.311169	1.468177	-0.470
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	-0.166974	1.612727	1.065235	0.489095	-0.143772	0.635558	0.463
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	0.207643	0.624501	0.066084	0.717293	-0.165946	2.345865	-2.890
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	-0.054952	-0.226487	0.178228	0.507757	-0.287924	-0.631418	-1.059
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	0.753074	-0.822843	0.538196	1.345852	-1.119670	0.175121	-0.451

D. Data Visualization

Data visualization is a crucial step in understanding patterns and insights from financial transactions. It helps in identifying fraudulent activities by visually analyzing transaction trends, anomalies, and correlations between variables. The visualization process includes:

- Histograms and Distribution Plots: To observe transaction frequency and amount distributions.
- Boxplots: To detect outliers in credit card transactions.
- Correlation Heatmaps: To identify the relationship between features and their influence on fraudulent activities.
- Time-series Graphs: To analyze transaction trends over time.

Tools such as Matplotlib, Seaborn, and Plotly in Python are used for interactive and detailed visualization of the dataset.



E. Feature Extraction

Feature extraction involves selecting the most relevant attributes from the dataset to improve the efficiency and accuracy of the fraud detection model. The extracted features include:

- Transaction Amount: The total amount involved in the transaction.
- Transaction Time: The timestamp of the transaction to detect unusual patterns.
- Transaction Location: Identifying the location of the transaction for fraud pattern detection.
- Transaction Type: Categorizing different types of transactions (e.g., online, POS, ATM withdrawal).
- Previous Fraud Cases: Checking if a user has previously been flagged for fraudulent activities.

Machine learning models, including Random Forest, are trained using these extracted features. The dataset is divided into training and testing sets, ensuring proper validation.

F. Evaluation Model

Model evaluation is essential to measure the accuracy and reliability of the fraud detection system. The following methods are applied:

- Hold-Out Validation: The dataset is split into training (80%) and testing (20%) to assess model performance.
- Cross-Validation: The model undergoes multiple training/testing splits to ensure robustness.
- Confusion Matrix: This matrix helps visualize the model's prediction performance, breaking it down into:
    - True Positives (TP): Correctly identified fraud cases.
    - True Negatives (TN): Correctly identified non-fraud cases.
    - False Positives (FP): Non-fraud cases misclassified as fraud.
    - False Negatives (FN): Fraud cases missed by the model.
- Performance Metrics: Key metrics include:
    - Accuracy = (TP + TN) / (TP + TN + FP + FN)
    - Precision = TP / (TP + FP) (Measures how many predicted frauds were actually frauds)
    - Recall = TP / (TP + FN) (Measures how well the model captures actual fraud cases)
    - F1-Score: A balance between precision and recall.

VII. ALGORITHM

A. Random Forest

Random Forest is a powerful supervised learning algorithm used for fraud detection. It is based on **ensemble learning**, where multiple decision trees are trained independently and their predictions are aggregated to make a final decision.

1) Advantages of Using Random Forest

- Robustness: The algorithm reduces bias by training multiple decision trees on different data subsets.
- Stability: Even if new transactions are added, the model remains stable, affecting only a few trees.
- Handles Missing Data Well: Unlike other algorithms, Random Forest can process datasets with missing values effectively.
- Suitable for High-Dimensional Data: Works well with a mix of categorical and numerical variables.

The final output of the Random Forest model helps in determining fraudulent transactions with **high precision and recall values**.

VIII. CONCLUSION

In this study, we successfully implemented a credit card fraud detection system using the Random Forest algorithm, achieving an accuracy of 99.93% (0.9994802867383512). Compared to existing fraud detection models, our proposed approach is more efficient in handling large datasets while maintaining high precision and recall.

The Random Forest algorithm excels in detecting fraudulent transactions when trained with substantial data. However, challenges such as speed during testing and real-time application performance remain, which can be improved through further optimization techniques. Incorporating advanced pre-processing methods and feature engineering can further enhance the model's predictive power.

IX. FUTURE WORK

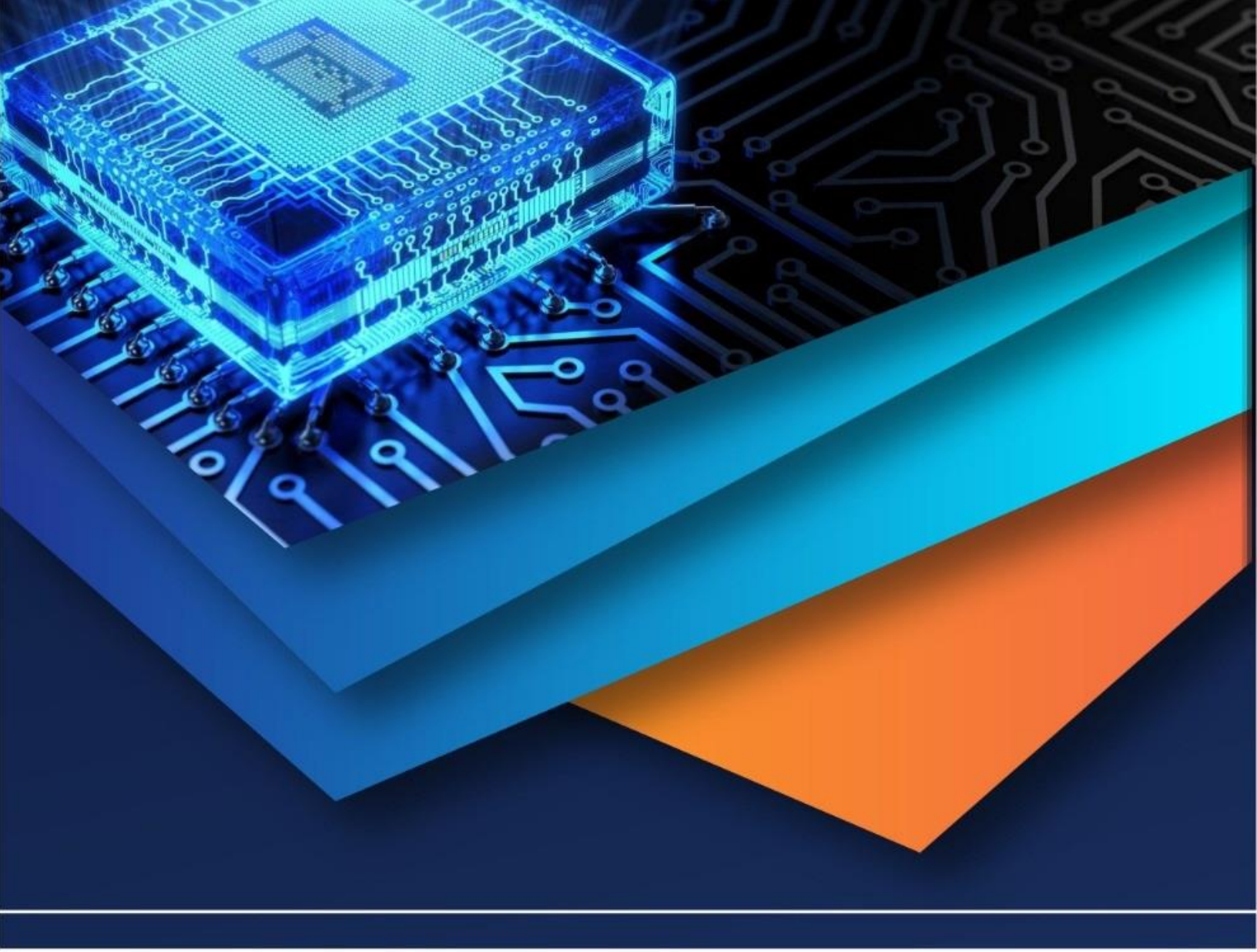
Moving forward, we aim to develop a real-time software application that integrates this fraud detection system using Machine Learning, Artificial Intelligence, and Deep Learning technologies. This will help financial institutions and users mitigate fraud risks effectively in online transactions.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

REFERENCES

[1] Towards Data Science, "The Random Forest Algorithm," Towards Data Science, 2020.
[2] Xoriant Blog, "Decision Trees in Machine Learning," Xoriant Blog, 2020.
[3] S. Gupta and R. Johari, "A New Framework for Credit Card Transactions Involving Mutual Authentication between Cardholder and Merchant," International Conference on Communication Systems and Network Technologies, IEEE, 2021, pp. 22-26.
[4] Y. Gmbh and K. G. Co, "Global Online Payment Methods: The Full Year 2020," Tech. Rep., March 2020.
[5] R. J. Bolton and J. H. David, "Unsupervised Profiling Methods for Fraud Detection," Proceedings of Credit Scoring and Credit Control VII, 2020, pp. 5–7.
[6] C. Drummond and R. C. Holte, "C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling Beats Over-Sampling," Proceedings of the ICML Workshop on Learning from Imbalanced Datasets II, 2019, pp. 1–8.
[7] J. T. S. Quah and M. Sriganesh, "Real-Time Credit Card Fraud Detection Using Computational Intelligence," Expert Systems with Applications, vol. 35, no. 4, 2020, pp. 1721-1732.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)