FRAUD DETECTION

ABSTRACT

Many individuals and organizations today rely heavily on financial transactions, leading to an exponential increase in the volume and complexity of these transactions. Consequently, the risk of fraudulent activities—such as credit card fraud, identity theft, and money laundering—has risen significantly. Traditional rule-based fraud detection systems struggle to adapt to the rapidly evolving patterns of fraudulent behaviour, thereby underscoring the necessity for machine learning techniques in advanced fraud analysis. This system develops a robust model designed to predict and prevent fraud by meticulously analysing historical transaction data. Utilizing supervised learning methods such as Decision Trees and Random Forests, the model classifies transactions as either legitimate or suspicious. Feature engineering plays a pivotal role in this process, enabling the identification of critical trends related to time, location, transaction type, and user behaviour. The system evaluates sophisticated algorithms, including Random Forest and Gradient Boosting, based on comprehensive performance metrics such as accuracy, precision, and recall. By addressing the dynamic challenges of fraud detection, this approach illustrates how machine learning can significantly enhance the security and efficiency of financial institutions. The implementation effectively identifies anomalies and prevents fraudulent transactions, providing a vital tool in safeguarding financial systems against the threats posed by fraudulent activities.

KEYWORDS: Fraud in Financial Transaction, Random Forest, XGBoost, KNN, Decision-tree

1. INTRODUCTION

The surge in digital transactions has increased the risk of financial fraud. Detecting fraudulent transactions is crucial for safeguarding both financial institutions and customers. Machine learning algorithms can play a pivotal role in identifying unusual patterns and anomalies. Our main goal is to integrate ML into the manual fraud detection and convert the detection into an efficiently automated one.

We have taken various algorithms available such as Random Forest and Decision Trees are commonly used for their ability to handle complex data and interactions. Histogram Gradient Boost and XGBoost excel in accuracy and performance, especially with imbalanced datasets. K-Nearest Neighbors (KNN), a distance-based method, is effective for anomaly detection.

By leveraging these algorithms, institutions can enhance fraud detection, ensuring timely responses and improving the overall security of financial transactions. The automation of financial transactions will enhance the security and detect frauds among the transactions if any.

2. LITERATURE SURVEY

Fraud detection in financial transactions is very important nowadays, earlier methods have human error, high costs, time consuming and most importantly no real time detection of the fraud [1]. This paper aims to reduce frauds in financial transactions by using ML algorithms like Bayesian Networks, Recurrent Neural Networks (RNN) and Support vector machines

(SVM) to improve the detection accuracy and efficiency. This paper evaluates these techniques and proposes an effective solution.

Security of the financial institution is very crucial for the economic growth and stability. Thus proposed a ML approach to predict financial fraud using transaction level features enhanced with synthetic data via CTGAN [2]. Among all classifiers the XGBoost achieved the highest 99% accuracy. This paper proposes that it can help the financial institutions aiming to detect fraud and eliminate and compromise in money laundering.

Financial institutions play a key role in growth. And the money thus, should be secured. This paper reviews research on financial fraud from 2009 to 2019, categorizing it into various frauds [3]. 34 data mining techniques were used, SVM being most (23%) naive bayes and random forest (15%). The most frauds are bank and insurance fraud i.e 81% .the reviews serves as both academic and industry purpose.

Financial fraud involves unauthorized mobile transactions through identity stealing / credit card theft is a growing issue with the rise of online payment services [4]. This paper has surveyed financial fraud detection methods using ML and DL includes feature selection, sampling, and applying algorithms. This approach has been validated by actual financial data from Korea demonstrating the effectiveness compared to the traditional neural networking.

This paper categorizes, compares and summarizes nearly all fraud detection from the past 10 years [5]. It gives us information about the professional fraudster, outlines main fraud types, and examines the nature of the data collected. The paper discusses methods and challenges for fraud detection while data mining.

Traditional methods of fraud detection are not able to detect and analyse fraud on a large scale and also they are not sufficient due to the increasingly sophisticated tactics of fraudsters [6]. Models like Multiplayer Perceptron (MLP) and Artificial Neural Network (ANN) are designed to improve the accuracy of fraud detection. Performance of these models are measured in aspects such as Accuracy, Precision and Sensitivity. Paper found that the ANN model significantly enhances the ability to detect fraud transactions and help to reduce the financial losses caused by fraud.

Machine Learning techniques offer promising solutions as it analyses the historic data to identify fraud patterns [7]. The Personalized PageRank (PPR) algorithm to capture the

social dynamics of fraud by analysing relations between financial frauds. The PPR feature provides unique and valuable information, evidenced by its high feature importance score.

With the rapid growth of E-commerce, financial fraud cases also increase, leading to huge losses [8]. Detecting fraud involves analysing user behaviour to identify and prevent swindles and account evasion. Traditional pattern -matching methods fall short in accurately detecting fraud. This paper classifies fraudulent behaviours, identifies key data sources, and reviews various techniques for detecting different types of financial fraud, including credit card fraud, etc.

Online banking and e-commerce organizing are facing an increase in online transactions and credit-card transactions [9]. Credit card companies are experiencing a lot of fraud complaints causing financial loss of company and customers. This paper has used a Simulated Annealing algorithm used to train the Neural Networks for Credit Card fraud detection in real-time scenarios. This algorithm is beneficial for the organizations and for individual users in terms of cost and time efficiency.

Credit card frauds are also one of the major financial frauds we have seen [10]. By the application of Neural Network, the fraud detection has increased and these techniques prevented a number of frauds.

3. PROPOSED METHODOLOGY

Fraud Detection in Financial Transactions combines finance and machine learning, leveraging algorithms like Random Forest, XGBoost, and KNN to analyze large datasets for real-time fraud detection and financial security. Machine learning algorithms such as Decision Tree (86% accuracy), Random Forest, XGBoost (83% accuracy), and KNN are used to detect fraud, with the dataset divided into training and testing.Machine learning identifies anomalies in transactions categorized into 5 labels, predicting fraud based on tran_label and acc_bal after training.

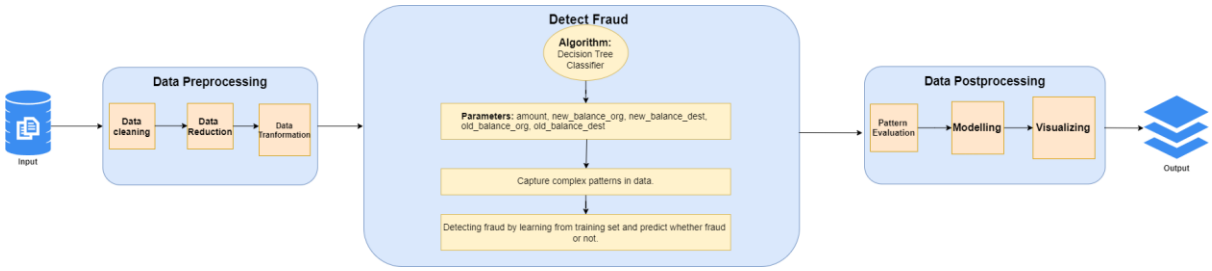


Fig 1 - System architecture – Detecting Fraud

4. MODULES

Detecting Frauds in Financial Transaction

4.1. Importing Modules

Import the necessary modules such as numpy, pandas, sklearn, seaborn, matplotlib, train_model, Decisiontreeclassifier, logistic regression, etc.

4.2 Data Preprocessing

The data preprocessing phase commences with the loading of the dataset, facilitating an initial examination of its structure and contents. Summary statistics are generated to provide insights into the distribution of values and to identify potential outliers. The assessment of missing values is conducted to highlight data quality issues, with visualizations employed to reveal patterns of absence. An analysis of the categorical variable representing transaction types yields the frequency of each category.

Subsequently, an interactive pie chart is created to depict the distribution of transaction types, offering a clear overview of the dataset's composition. This comprehensive preprocessing ensures that the data is clean and well-structured for subsequent analysis and modeling as shown below in Fig 2 .

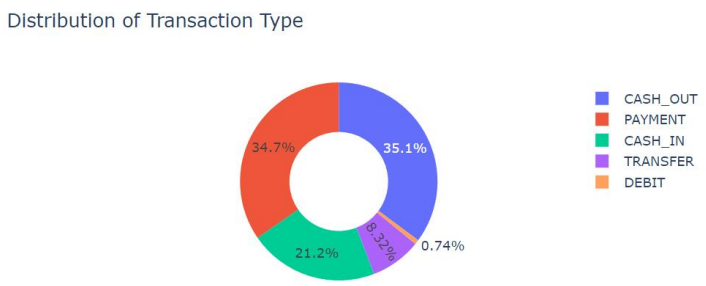
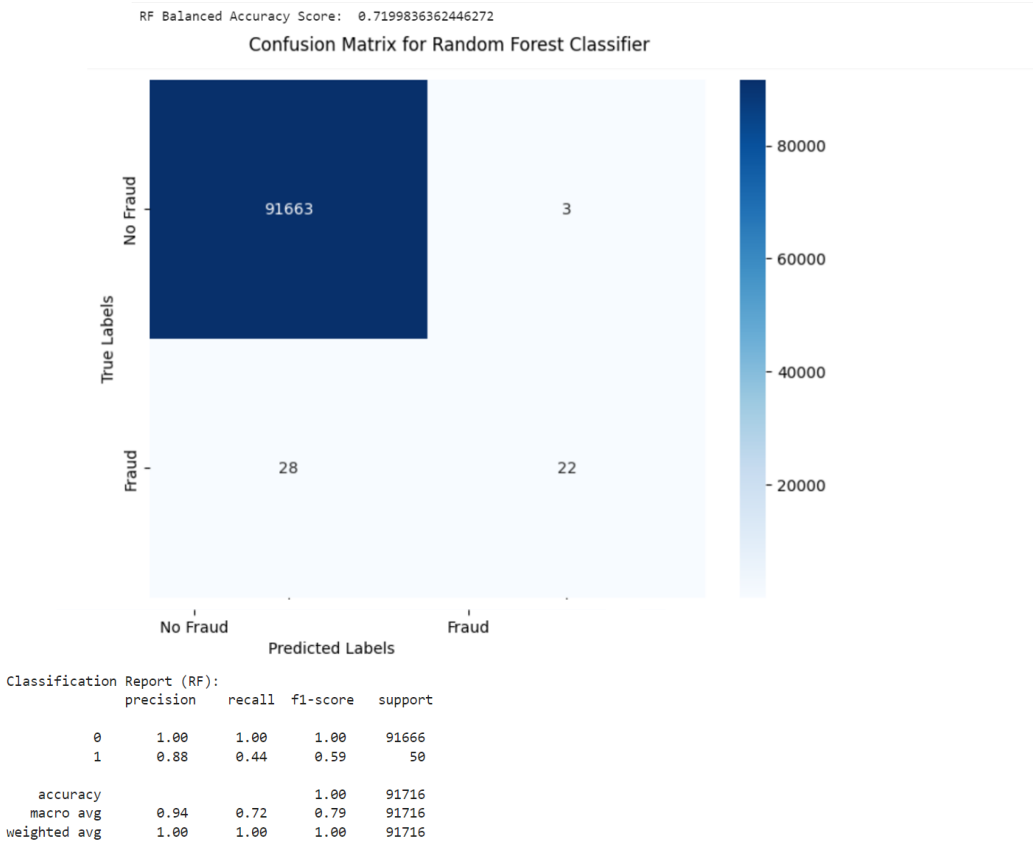


Fig 2-Transaction type representation

4.3 Algorithm

4.3.1 Random Forest

The implemented approach detects fraudulent financial transactions using a Random Forest Classifier. The process starts by loading the dataset, followed by encoding the categorical feature type using LabelEncoder. Key features selected for the model include step, type_encoded, amount, oldbalanceOrg, newbalanceOrig, oldbalanceDest, and newbalanceDest. The dataset is split into training and testing sets (70% training, 30% testing). Performance is evaluated through metrics such as balanced accuracy, confusion matrix (TP, TN, FP, FN), and classification report is represented in fig 3 below. Results, visualized using a heatmap, demonstrate effective fraud detection by classifying transactions as legitimate or fraudulent.



Classification Report (RF):

	precision	recall	f1-score	support
0	1.00	1.00	1.00	91666
1	0.88	0.44	0.59	50
accuracy			1.00	91716
macro avg	0.94	0.72	0.79	91716
weighted avg	1.00	1.00	1.00	91716

Fig 3-Confusion matrix for RF

4.3.2. XGBoost

The implemented approach detects fraudulent financial transactions using an XGBoost Classifier. The process begins by loading the dataset and encoding the categorical feature type using LabelEncoder. Key features selected for the model include step, type_encoded, amount, oldbalanceOrg, newbalanceOrig, oldbalanceDest, and newbalanceDest. The dataset is split into training (70%) and testing (30%) sets. The XGBoost model is trained on the training data, and performance is evaluated using metrics such as balanced accuracy, confusion matrix (TP, TN, FP, FN), and classification report is represented in fig 4 below.. The results, visualized using a heatmap, show the model's effectiveness in classifying transactions as legitimate or fraudulent.

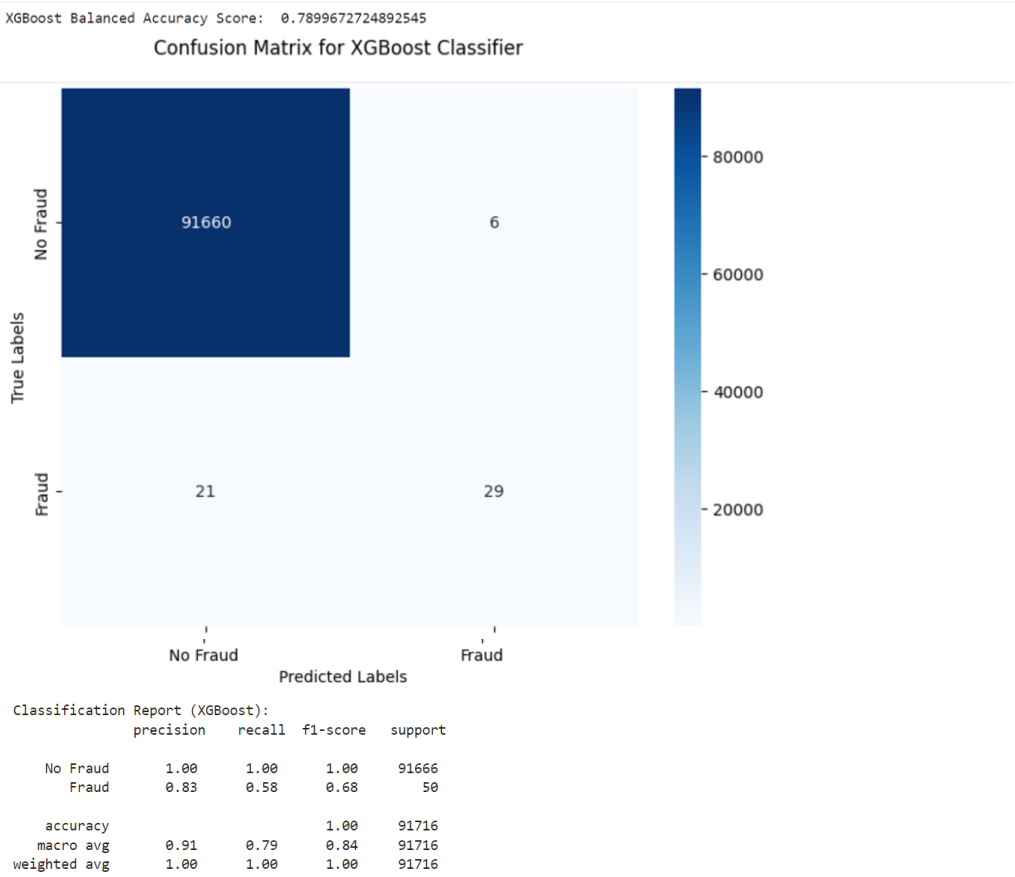


Fig 4-Confusion matrix for XGBoost

4.3.3. KNN

The implemented approach detects fraudulent financial transactions using a K-Nearest Neighbors (KNN) Classifier. The process begins by loading the dataset, followed by encoding the categorical feature type using LabelEncoder. Key features selected for the model include step, type_encoded, amount, oldbalanceOrg, newbalanceOrig, oldbalanceDest, and newbalanceDest. The dataset is split into training (70%) and testing (30%) sets. The KNN model is trained with 'n_neighbors=5', and performance is evaluated using metrics such as balanced accuracy, confusion matrix (TP, TN, FP, FN), and classification report is represented in fig 5 below.. The results, visualized using a heatmap, demonstrate the model's ability to classify transactions as legitimate or fraudulent.

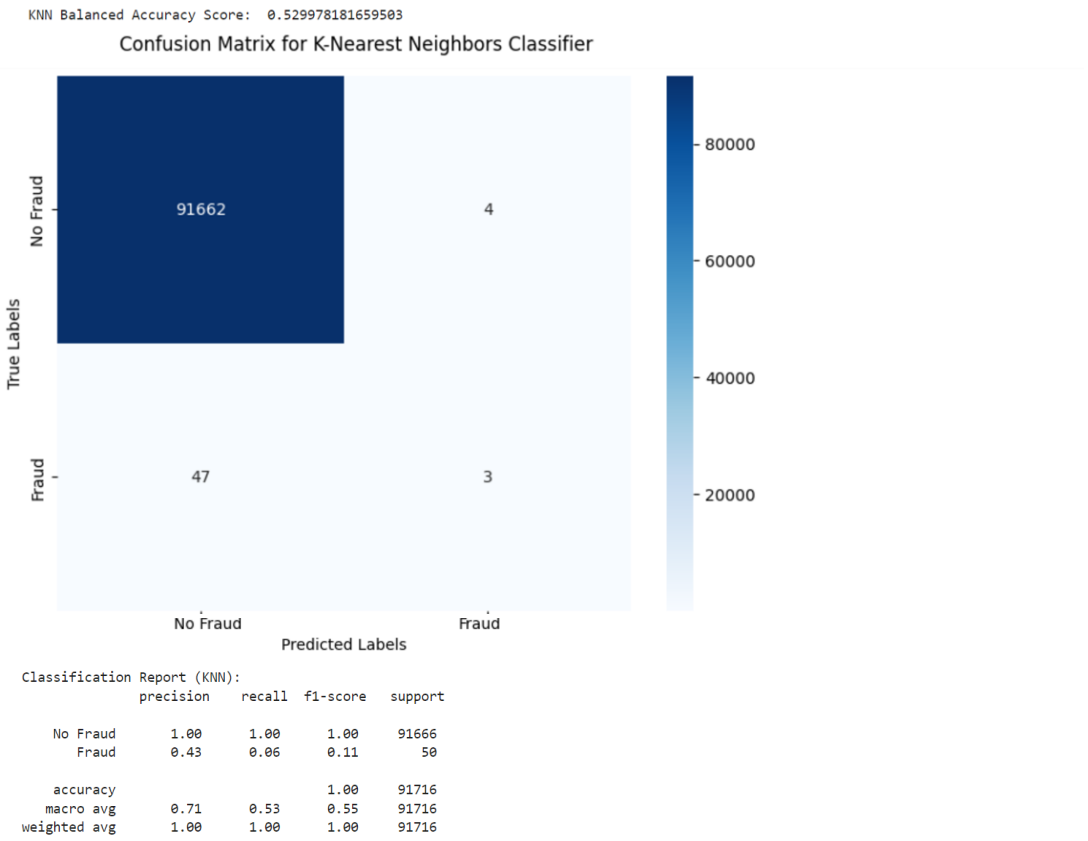


Fig 5-Confusion matrix for KNN

4.3.4. Histogram Gradient Boosting Classifier

The implemented approach detects fraudulent financial transactions using a HistGradientBoostingClassifier. The process starts by loading the dataset, followed by encoding the categorical feature type using LabelEncoder. Key features selected for the model include step, type_encoded, amount, oldbalanceOrg, newbalanceOrig, oldbalanceDest, and newbalanceDest. The dataset is split into training (70%) and testing (30%) sets. The HistGradientBoostingClassifier is trained and its performance is evaluated using metrics such as balanced accuracy, confusion matrix (TP, TN, FP, FN), and classification report is represented in fig 6 below. The results, visualized through a heatmap, demonstrate effective fraud detection by classifying transactions as either legitimate or fraudulent.

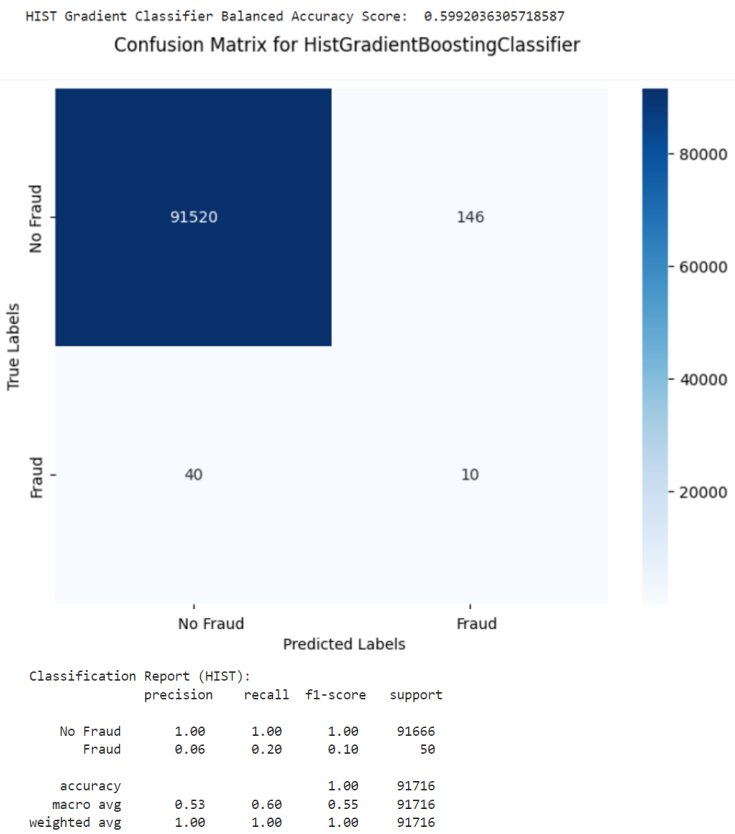
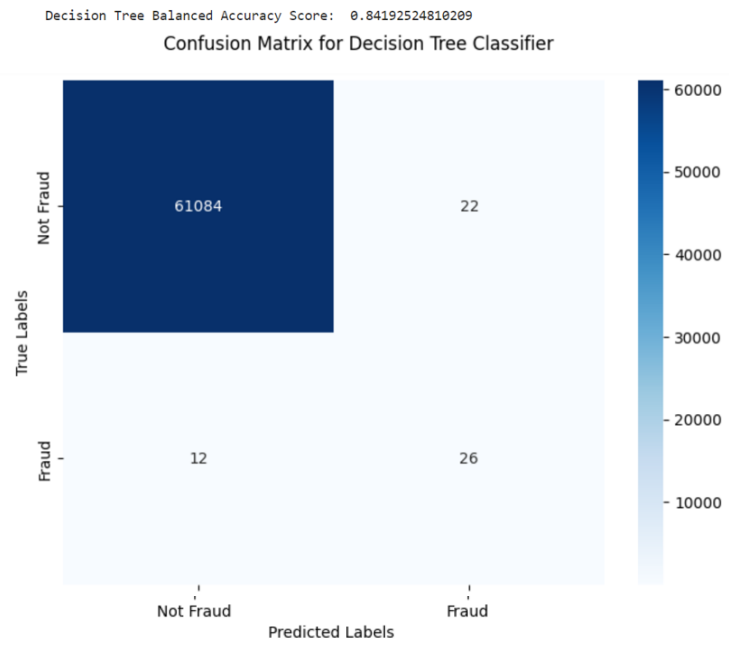


Fig 6 -Confusion matrix for HistgradientBoosting

4.3.5. Decision Tree Classifier

The implemented approach detects fraudulent financial transactions using a Decision Tree Classifier. The process begins with loading the dataset, focusing on key features such as `amount`, `oldbalanceOrg`, `newbalanceOrig`, `oldbalanceDest`, and `newbalanceDest`. The independent variables (features) are separated from the dependent variable (`isFraud`). The dataset is then split into training (80%) and testing (20%) sets. The Decision Tree Classifier is trained on the training data, and predictions are made on the test data. Performance is evaluated using metrics such as balanced accuracy, confusion matrix (TP, TN, FP, FN), and classification report is represented in fig 7 below. The results, visualized with a heatmap, illustrate the model's effectiveness in classifying transactions as either fraudulent or not fraudulent.



Classification Report (Decision Tree):

	precision	recall	f1-score	support
0	1.00	1.00	1.00	61106
1	0.54	0.68	0.60	38
accuracy			1.00	61144
macro avg	0.77	0.84	0.80	61144
weighted avg	1.00	1.00	1.00	61144

Fig 7 -Confusion matrix for Decision Tree

4.4. POST PROCESSING

In the post-processing phase, the model's performance is evaluated by calculating the balanced accuracy scores for different classifiers, including Random Forest, XGBoost, KNN, HistGradient Boosting, and Decision Tree, with the Decision Tree classifier emerging as the most accurate as shown in fig 8. The transactional data is prepped by mapping categorical transaction types (such as "CASH_OUT," "PAYMENT") to numerical values, while the target variable (isFraud) is transformed to binary labels of "Fraud" or "No Fraud." Key features like transaction type, amount, and balances are selected, and the dataset is split into training and testing sets. A Decision Tree classifier is trained using these features, and after successful training, it is capable of predicting whether a given transaction is fraudulent or legitimate. By using the trained model, a sample transaction can be evaluated for potential fraud.

```
RF Balanced Accuracy Score:  71.99836362446273

XGBoost Balanced Accuracy Score:  78.99672724892545

KNN Balanced Accuracy Score:  52.99781816595029

HIST Gradient Classifier Balanced Accuracy Score:  59.92036305718587

Decision Tree Balanced Accuracy Score:  84.19252481020901


 As seen above decision tree matrix has the best accuracy
```

Fig 8 -Model Selection

Here, after training the model we have selected the algorithm as a decision tree algorithm as it has the best accuracy score of 84.19%. We then test the model by creating a function and passing the input parameters and the trained model predicts whether the transaction is fraud or not.

```
features = np.array([[4, 9000.60, 9000.60, 0.0]])

prediction = model.predict(features)[0]  # Access the first (and only) element of the prediction

if prediction == "Fraud":
    print("Fraud")
elif prediction == "No Fraud":
    print("No Fraud")
```

```
Fraud
```

Fig 9 - Prediction Output

As depicted above in Fig 9 the outcome/prediction for the parameters given was given “fraud” which tells us that the transaction happened was a fraud one. As depicted in the below fig 10 the outcome /prediction for the parameters given was “no fraud” which tells no fraud had happened in the transaction. The predictions given were 100% accurate, thus concluding that the model has been trained correctly.

```
features = np.array([[4, 900, 900.60, 0.0]])

prediction = model.predict(features)[0]  # Access the first (and only) element of the prediction

if prediction == "Fraud":
    print("Fraud")
elif prediction == "No Fraud":
    print("No Fraud")
```

```
No Fraud
```

Fig 10 -Prediction O/P

5. COMPARISON TABLE

Algorithm / Model Name	F1 -score	Accuracy Score
Random Forest	0.79	71.99
XGBoost	0.84	78.99
KNN	0.55	52.99
Histogram Gradient Boost	0.55	59.92
Decision Tree Classifier	0.80	84.19

Table 1 Comparison with our models

As mentioned in the comparison table i.e Table 1 the Random Forest model had an average f1-score of 0.79 and an accuracy score of 71.99%, the XGBoost model had an average f1-score of 0.84 and an accuracy score of 78.99%, the KNNmodel had an average f1-score of 0.55 and an accuracy score of 52.99%, the Histogram Gradient Boost model had an average f1-score of 0.55 and an accuracy score of 59.92% and the Decision tree classifier algorithm has an average f1-score of 0.80 and accuracy score of 84.19%. The parameters we have taken are amount, old balance origin, new balance origin, old balance destination and new balance destination.

6. EVALUATION METRICS

The model performance is accessed using the following metrics:

- **Precision** is calculated by dividing the actual true prediction by the model's total number of predictions as shown in eqn 1.

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{------- (1)}$$

- **Recall** is determined in a classification problem with two classes by dividing the total number of true positives by the sum of true positives and false negatives as shown in eqn 2.

$$\text{Recall} = \frac{TP}{TP + FN} \qquad \text{------- (2)}$$

- **F1 score:** Weighted average of recall and precision as shown in eqn 3.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad \text{------- (3)}$$

7. COMPARISON WITH EXISTING WORK

Metric / Comparison model	**ACCURACY**	**F1 SCORE**
Proposed System	84.19	0.80
Existing Model Catayoun Azarm [7]	82.45	0.82

Table 2 Comparison with existing work

As depicted in the above table i.e table 2 the Existing Model based on XGBoost has an accuracy score of 82.45 and F1-score of 0.82 while the work done by us gives an accuracy score of 84.19% and f1-score of 0.80 .During the test , the model trained gave 100% accurate results .

8. CONCLUSION

Detecting fraud in financial transactions is essential to maintaining the integrity and security of financial systems in an increasingly digital world. Machine learning algorithm Decision Trees offer powerful tools for identifying fraudulent patterns within vast, complex datasets. By implementing the algorithm, institutions can move beyond traditional rule-based systems and enable real-time, accurate fraud detection with minimal false positives. This enhances their ability to proactively prevent financial losses and protect customers. As fraud techniques evolve, the continuous refinement of the model ensures that financial institutions remain ahead of potential threats, securing a safer transaction environment. The accuracy of the model is 86%.

FUTURE ENHANCEMENT

Currently we apply the decision tree model on the dataset and find frauds in the dataset. We plan to develop an end-to-end site wherein the user can just enter the credentials and check whether fraud or not in real time. This would require access to real time data and real time OS will also be required. Our future plans are to secure you from any type of financial fraud. Are u also ready to secure your money with us?

REFERENCES

[1] Thushara Amarasinghe, Achala Aponso, and Naomi Krishnarajah. 2018. Critical Analysis of Machine Learning Based Approaches for Fraud Detection in Financial Transactions. In Proceedings of the 2018 International Conference on Machine Learning Technologies (ICMLT '18). Association for Computing Machinery, New York, NY, USA, 12–17.

[2] Alwadain, A.; Ali, R.F.; Muneer, A. Estimating Financial Fraud through Transaction-Level Features and Machine Learning. *Mathematics* **2023**, *11*, 1184.

[3] Khaled Gubran Al-Hashedi, Pritheega Magalingam,Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019,Computer Science Review,Volume 40,2021,100402,ISSN 1574-0137

[4] Panigrahi S., Kundu A., Sural S., and Majumdar A. K., Credit card fraud detection: a fusion approach using dempster shafer theory and bayesian learning, *Information Fusion*. (2009) 10, no. 4, 354–363.

[5] Li, Shing-Han and Yen, David C. and Lu, Wen-Hui and Wang, Chiang, Identifying the signs of fraudulent accounts using data mining techniques,volume=28,ISSN=0747-5632

[6] Bassam Kasasbeh, Balqees Aldabaybah, Hadeel Ahmad.2022. Indonesian Journal of Electrical Engineering and Computer Science. Vol. 26, No. 1, April 2022, pp. 362~373

[7] Catayouan Azarm, Erman Aacar, Mickey van Zeeltv.2022. On the Potential and Feature for Fraud Detection

[8] Pankaj Richhariya,Prashant K Singh.A Survey on Financial Fraud Detection Methodologies.International Journal of Computer Applications (0975 – 8887).Volume 45– No.22.

[9] Azeem Ush Shan Khan, Nadeem Akhtar and Mohammad Naved Qureshi.Real-Time Credit-Card Fraud Detection using Artificial Neural Network Tuned by Simulated Annealing Algorithm Vol .35,no - 13-17.

[10] Imane Sadgali, Nawal Sael, and Faouzia Benabbou. 2019. Fraud detection in credit card transactions using Neural Networks. In Proceedings of the 4th International Conference on Smart City Applications (SCA ’19). Association for Computing Machinery, New York, NY, USA, Article 95, 1 – 4

[11] S. Maes, K. Tuyls, B. Vanschoenwinkel, and B. Manderick, "Credit Card Fraud Detection. Applying Bayesian and Neural networks," Oct. 2017.

[12] B. Wiese and C. Omlin, "Credit Card Transactions, Fraud Detection, and Machine Learning: Modelling Time with LSTM Recurrent Neural Networks," vol. 247, 1970, pp. 231--268.

[13] T. Razooqi, P. Khurana, K. Raahemifar, and A. Abhari, "Credit Card Fraud Detection Using Fuzzy Logic and Neural Network," in Proceedings of the 19th Communications & Networking Symposium, San Diego, CA, USA, 2016, p. 7:1--7:5.

[14] B. Baesens, V. V. Vlasselaer, and W. Verbeke, Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection, 1st ed. Wiley Publishing, 2015.

[15] K. R. Sungkono and R. Sarno, "Patterns of fraud detection using coupled Hidden Markov Model," in 2017 3rd International Conference on Science in Information Technology (ICSITech), 2017, pp. 235--240.