A Hyperparameters Tunned ML Algorithm for Fraud Identification in Banking and Financial Transactions

Srinivasa Rao Bogireddy
Senior Software Engineer
Horizon Systems Inc
Phoenix,Arizona , USA

Haritha Murari
Senior Software Engineer
Spark Infotech Inc
Phoenix, Arizona , USA

Abstract:- Banking, a pioneering industry, is experiencing rapid growth, leading to the adoption of cashless transactions. Digital banking offers better service quality but has faced challenges from fraudulent activities. Since the banking industry is expanding quickly throughout the globe, using cash for payments is becoming less common. Instead, people are using cashless transactions. Digital banking customers receive higher-quality services in money transfers, cashless payments, credit cards, and prepaid cards. Nonetheless, the fraudulent activities of scammers have drawn attention to the security of digital banking, as a lack of adequate protection has discouraged many users from using the service up to this point. Even though fraud is not a new problem, its associated actions always cause billions of dollars' worth of annual losses to the world economy. Fraudulent actions carry a wide range of severe financial hazards that might jeopardize an economy's profitability and reputation. The study aims to introduce an efficient hyperparameter-tuned machine learning approach to detect fraud in banking and financial transaction systems. Proper preprocessing and application of feature engineering, such as outlier rejection, null value handling, standardization, and parameter tuning, have been incorporated with the approach. Later, the Extreme gradient boosting model was trained with tunned parameters and evaluated with test data. The model demonstrated praiseworthy performance, having 99.63% accuracy. Extensive analysis using feature selection, confusion matrix, roc, and tunning evaluation graph was conducted to detect fraud in financial transactions.

Keywords:- *XGB, Financial Baking, Fraud, Grid Search, Tuning.*

I. INTRODUCTION

Fraud detection in banking and any financial dealings is one of today's most essential and challenging activities because fraud has become increasingly sophisticated. Financial institutions suffer from fraudulent transactions through financial losses, reputation impairment, and erosion of customer trust [1]. Volumes of transactions and modern fraud techniques make conventional fraud-identification methods, such as rule-based approach and manual reviews, increasingly ineffective [2]. Therefore, there is a pressing need for an advanced, efficient, and accurate automated system capable of detecting and preventing fraudulent activities. Machine learning approaches that use historical transaction data to identify trending patterns indicative of fraud can be up-and-coming solutions to this problem [3-4]. However, their performance strongly depends on their hyper-parameters and how proper their choice and setting are. This means setting optimal settings for a learning algorithm to improve its performance and make it more accurate and robust in distinguishing between legitimate and fraudulent transactions.

In our study, we have proposed an efficient machine learning-based automatic fraud detection of financial transactions. We applied modern preprocessing techniques such as null values imputation, outlier reduction, proper transformation by standardization, and duplicate values dropping to enhance the data quality. Later, the extreme gradient boosting model used hyperparameters to get the best-optimized parameters. Finally, the model was evaluated using well-known metrics. Our approach achieved superior performance.

➢ *The Study's Contributions May be Distilled into the Following:*

- Outlier reduction
- Null values imputations
- Duplicate values dropping
- Grid search optimization
- Proposing an efficient XGB model for fraud detection of financial transactions.

The methods and literature review are presented in Sections 2 and 3. Results analysis and its discussion are presented in Section 4. The conclusion and upcoming projects are covered in Section 5.

II. LITERATURE SURVEY

AI and ML services are used in many operational areas, including the commercial world, agriculture, and the medical field [5–6]. Similarly, machine learning assists in classifying fraud in financial services.

The study of [7] aims to improve fraud detection and prevention using machine learning-based predictive modelling. The proposed approach uses the Hidden Markov Model (HMM) to observe hidden states of financial transactions, followed by the Gradient Boosting Classifier (GBC) for fraud classification. A hybrid method combines HMM and GBC with experiments to ensure their effectiveness.

A new Fraud Detection System [8] combines graph mining and Machine Learning algorithms (MLA) to detect fraud involving money in digital ledgers and banking systems. Motivated by Benford's Law, the AntiBenford subgraph helps identify irregularities in financial transactions but can produce false positives. The system adapts over time and offers a 94.83% fraud detection accuracy rate, reinforcing confidence and safety in the financial industry.

The author of [9] conducted another study to enhance security in online credit card transactions by utilizing machine learning techniques. It develops a hybrid machine-learning model that distinguishes between legitimate and illegitimate transactions. The model uses AdaBoost, logistic regression, and random forest techniques to predict output values, enabling real-time identification of potentially fraudulent transactions. This approach significantly contributes to the overall security of online credit card transactions.

The research [10] aims to identify fraudulent transactions in branchless banking using Leiden community detection algorithms. It found that 25% of agents have done fraudulent transactions, with transactions 185% above the average in terms of transaction value and 90% above the average frequency. These transactions act as outliers in their respective communities, making detecting and preventing fraudulent activities difficult.

Abdulla et al. [11] used the methods Naïve Bayes, Decision Tree, Isolation Forest, Random Forest, and Cat Boost to study credit card datasets. They discovered Random Forest performed better than the rest.

To identify and reduce financial fraud in banking systems, Rathnakaret al. [12] suggested a machine learning-based method. The artificial intelligence model will lessen damage, expedite check verification, and stop counterfeits. The study resamples the dataset for improved accuracy and examines clever algorithms trained on a public dataset to find connections with fraudulence.

Asymmetric datasets were utilized by Anishuses et al. [13] to identify fraudulent credit card transactions using logistic regression. At 94% accuracy, the system finds important characteristics such as transaction amount, country of origin, and time of day as strong predictors. A more reliable and secure financial system for all parties involved can result from the results, which can aid in the development of advanced fraud identification algorithms for financial institutions.

V. Backiyalakshmi et al. [14] discussed fraud detection techniques using machines and deep learning in the banking sector. The paper introduces fraud detection, reviews existing literature, and provides a chronological assessment. It also discusses simulation tools, performance indices, and research gaps for future studies in the banking sector.

The literature review shows several works have been accomplished in fraud detection and financial and banking tracing. However, the approach's performance has scope to improve. No one uses parameter tuning, feature engineering, or advanced preprocessing. This study addresses these. We proposed a parameter-tuned XGB model incorporating modern preprocessing and tuning strategies and feature engineering strategies that outperform the previous ones.

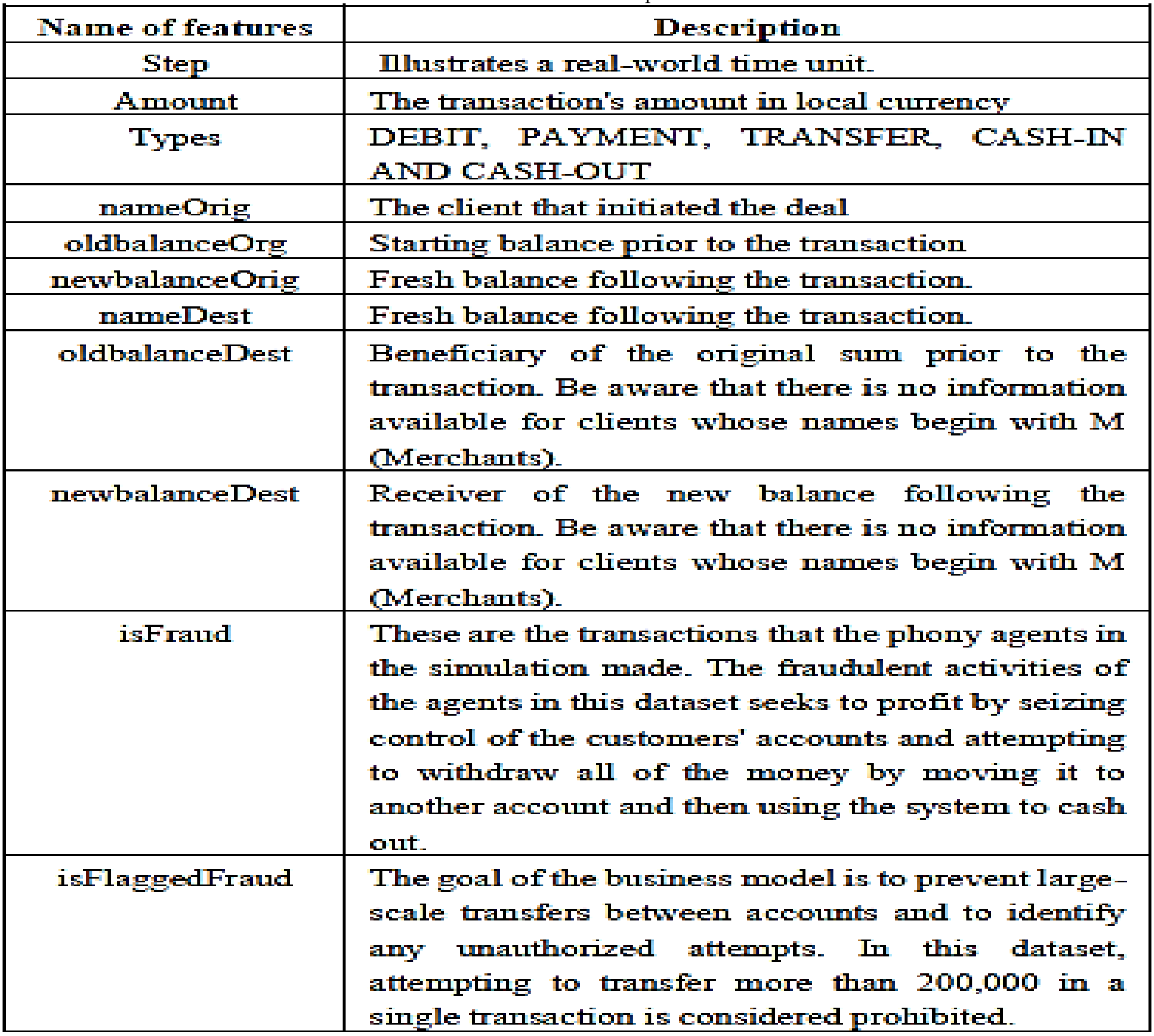
III. MATERIALS AND METHODOLOY

Various methods were used to manage the data collection, preprocessing, training, and testing processes for the suggested advanced machine-learning strategy for obesity classification. Fig. 4 shows the procedures that were followed.

A. Data Description

The dataset used in the study is publicly available in a trustworthy Kaggle repository [15]. It consists of 11 features, summarized in Table 1.

Table 1: Data Descriptions



Name of features	Description
Step	Illustrates a real-world time unit.
Amount	The transaction's amount in local currency
Types	DEBIT, PAYMENT, TRANSFER, CASH-IN AND CASH-OUT
nameOrig	The client that initiated the deal
oldbalanceOrg	Starting balance prior to the transaction
newbalanceOrig	Fresh balance following the transaction.
nameDest	Fresh balance following the transaction.
oldbalanceDest	Beneficiary of the original sum prior to the transaction. Be aware that there is no information available for clients whose names begin with M (Merchants).
newbalanceDest	Receiver of the new balance following the transaction. Be aware that there is no information available for clients whose names begin with M (Merchants).
isFraud	These are the transactions that the phony agents in the simulation made. The fraudulent activities of the agents in this dataset seeks to profit by seizing control of the customers' accounts and attempting to withdraw all of the money by moving it to another account and then using the system to cash out.
isFlaggedFraud	The goal of the business model is to prevent large-scale transfers between accounts and to identify any unauthorized attempts. In this dataset, attempting to transfer more than 200,000 in a single transaction is considered prohibited.

B. Exloratory Data Analysis and Preprocessing

Exploratory data analysis is a method for displaying the dataset's significant findings. We have statistically examined the fraud data frame, Boxplot, and fraudulent vs non-fraudulent transactions in our study, as seen in Figures 1, 2, and 3. There are five float datatype columns, three integer datatype columns, and three object datatype columns in the dataset's 11 columns. There are 6,362,620 rows of data in the dataset. Initially, we saw that all the columns in our dataset were devoid of null values. This indicates that either the missing values were already addressed, or our dataset was clean. Secondly, we observed that there are no duplicate values in our dataset. For our convenience and better comprehension, we then renamed a few of our columns and rearranged their locations. We use the TRANSFER and CASH_OUT transaction types for our fraud activities. 4,097 fraud transactions occurred in TRANSFER, and 4,116 in CASH_OUT. Most fraudulent transactions included customers trading with other customers. In fraudulent transactions, different transaction accounts were utilized for sending and receiving. We started by doing some feature engineering and introducing a new type of column that deals with transactions for both customers and merchants. Next, we removed a few superfluous columns and changed the column layout. 8,213 fraudulent transactions from one customer to another were recorded overall. 4,202,912 were the total number of valid transactions between customers. 2,151,495 transactions were made legitimately between the customer and the merchant.

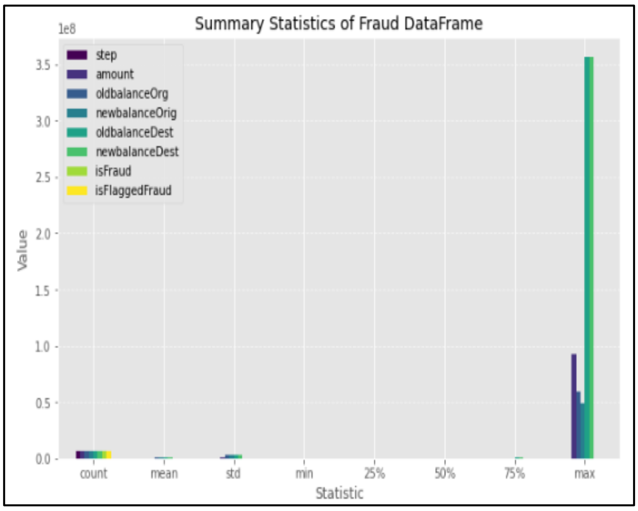


Fig 1: Summary Statistics of Fraud Data Frame

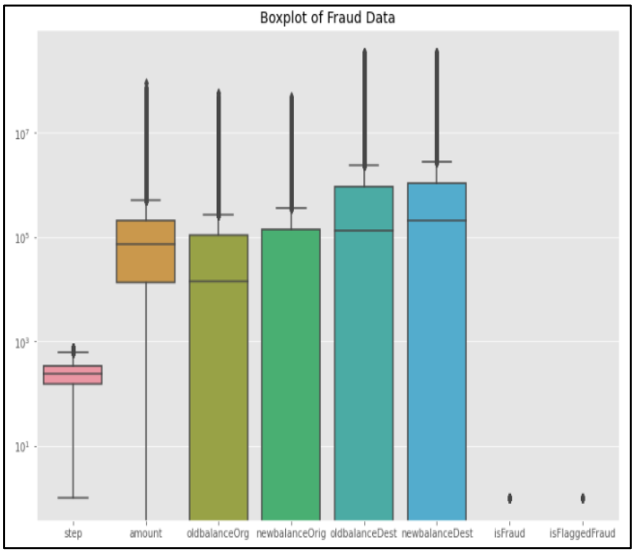


Fig 2: Boxplot of Fraud Data

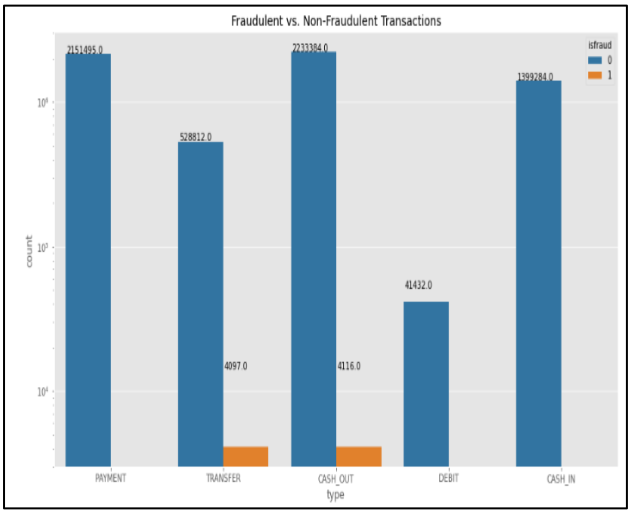


Fig 3: Fraudulent vs Non Fraudulent Transactions

C. Proposed Approach

Our study entailed data normalization, deletion of duplicate values in the dataset, cleaning the data, and replacing missing fields with the mean value as the data preparation step. Outliers were eliminated by applying the quartile approach. Some of the critical findings brought out the descriptively univariate and multivariate analysis. The dataset was then split 80:20 when it was determined that the ratio was optimum for testing and training. The Extreme Gradient Boosting was assessed using confusion matrices, accuracy, recall, F1 score and precision. Several models were then fine-tuned using grid search for hyperparameter tuning, which improved each model's performance tremendously. The models were trained again based on the new parameter, returning much better results while testing. Hence, the random search optimization was favorable since it efficiently gave the optimal values and reduced computational time. In the study, an estimate of the profile of tunning on AUC-ROC was also made. The result retrieved the tunned XGB model with a notable performance of 99.63%.

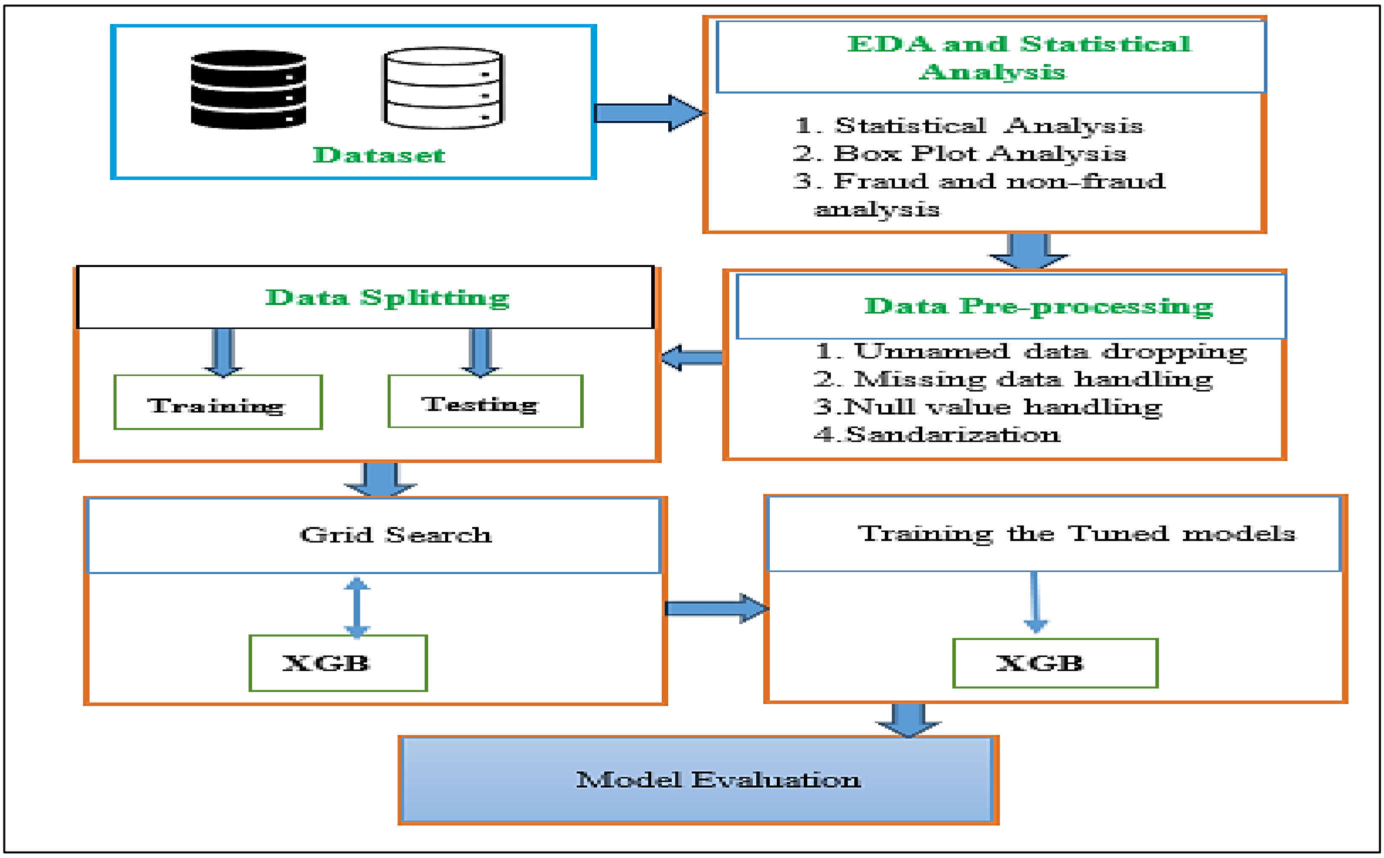


Fig 4: Proposed Architecture

D. Model Evalution

The proposed approach was assessed using the following parameters from equation (1-4).

$$Accuracy = \frac{True\ Positive + False\ Negative}{all} \quad (1)$$

$$F1\ score = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{Recall}} \quad (2)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ positive} \quad (3)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ negative} \quad (4)$$

IV. RESULTS AND DISCUSSION

The grid search technique employed to tune the extreme gradient boosting model was performed on the four parameters with a grid space. The parameters and grid space are shown in Table II. The parameters need Subsample, colsample_bytree, max_depth, and learning_rate, which are prominent in the XGB model. The outcome of the tunned model has been listed in Table III. The Subsample was found to be 0.8, and colsample_bytree, max_depth, and learning_rate is 0.6, 8, and 0.02, respectively, for the best performance. The changes in the AUC-ROC values due to grid search are dipped in Fig.5. It presents the best fitting with tunned parameters. The preprocessed feature was analyzed for the importance of predominance in fraud detection. The importance of the feature is shown in Fig.6. The amount of money transferred is the most critical parameter to classify. Later, the model was evaluated using the confusion matrix and Roc curve, as shown in Figs. 7 and 8. The confusion matrix displays only 4722missclassifcation out of 1272524. The Auc values are one. Similarly, the Recall and Accuracy are calculated, shown in Table 4. The recall and accuracy are 0.9958 and 0.9963 respectively. The models achieved the highest accuracy over the previous study. The parameter tuning, feature engineering, preprocessing, sense, and able approach of XGB play a significant role in getting high performance.

Our study has been compared with the previous work on the detection of fraud in financial transactions. Table V displays the summary. The recent papers referenced [8] and [13] illustrate different approaches to detect fraud detection from financial transcription. The accuracy achieved was 94.83 and 94%, respectively. But our proposed model improved to 99.63%. Our approach is the best in terms of performance in fraud detection in the banking system

Table 2: Gird Search Spaces

The Parameters	Space of Tunning
Subsample	0.6,0.8,1
Col_sample bytree	0.6,0.8,1
Max depth	8, 12, 16
Learning rate	2e-2, 3e-1, 1e-1

Table 3: Tunned Parameters

Parameters Name	Tunned Values
Subsample	0.8
Col_sample bytree	0.6
Max depth	8
Learning rate	0.02

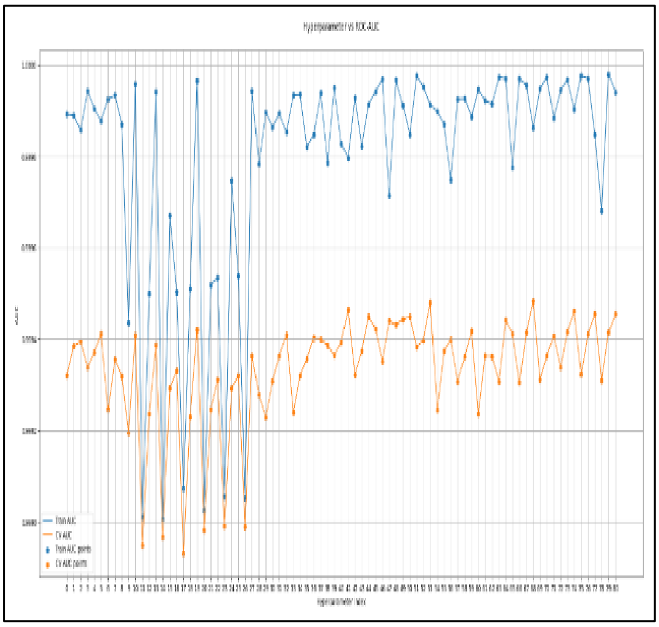


Fig 5: Hyper Parameters vs AUC-ROC

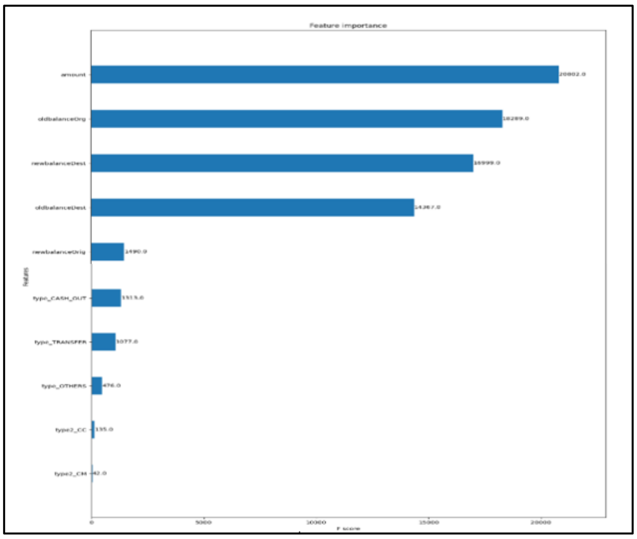


Fig 6: Feature Selection

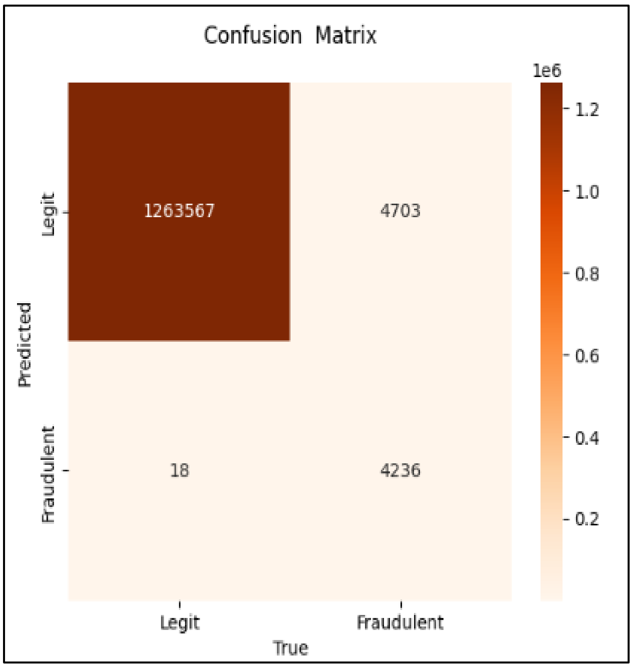


Fig 7: Confusion Matrix

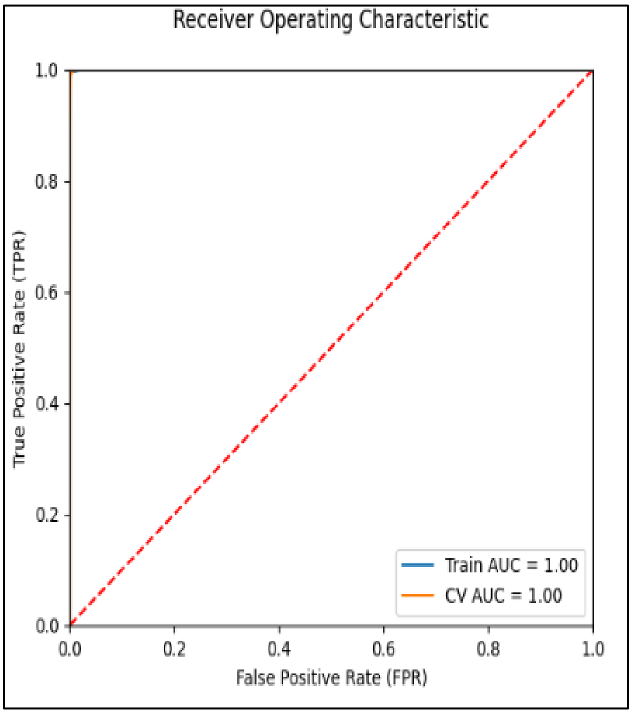


Fig 7: ROC Curve

Table 4: Performance Metrics.

Matrices	Values
Recall	0.9958
Accuracy	0.9963

Table 5: Comparison with Previous Studies

Ref.	Published Year	Performance
[13]	2023	Accuracy= 94%
[8]	2023	Accuracy=94.83%
Our proposed	-	Accuracy=99.63%

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24AUG458

V. CONCLUSION AND FUTURE WORK

Using grid search optimization significantly enhanced the performance of Extreme Gradient boosting machine learning models for Fraud Identification in Banking and Financial Transactions. Tuning the models' parameters resulted in substantial improvement, recall, and accuracy metrics. On the other hand, pre-processing, outlier rejection, null value, duplicated value management, and feature engineering also impact the models' performance. The approach achieved 99.63% accuracy. These enhancements were visually validated through confusion matrices, which showed reduced misclassification. The feature selection techniques also add the importance of features that explain the model predictability. The tuning evaluation also shows the fruitfulness of the approach.

Whatever, the task of identifying fraud in banking and financial transactions has been completed properly. The study's limitation is that the model has yet to be deployed. Future research can be deployed and done in real-time to check its effectiveness.

REFERENCES

[1]. Tomasic, R., & Akinbami, F., "The role of trust in maintaining the resilience of financial markets", Journal *of corporate law studies*, Vol. 11, pp. 369-394, 2011.

[2]. Bolton, R. J., & Hand, D. J.," Statistical fraud detection: A review", *Statistical science*, Vol. 17, pp. 235-255, 2002.

[3]. Stojanović, B., Božić, J., Hofer-Schmitz, K., Nahrgang, K., Weber, A., Badii, A., ... & Runevic, J., "Follow the trail: Machine learning for fraud detection in Fintech applications", Sensors, Vol. 21, pp.1594, 2021.

[4]. Ryman-Tubb, N. F., Krause, P., & Garn, W., "How Artificial Intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark", *Engineering Applications of Artificial Intelligence*, Vol. *76*, pp. 130-157, 2018.

[5]. M. S. H. Talukder, A. H. Nur, S. Zaman, M. R. Noor, M. A. U. Khan and F. Amir, "Optimizing Diabetes Prediction Accuracy: A Comprehensive Approach with Advanced Preprocessing and Diverse Machine Learning Classifiers," 2024 3rd International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE), pp. 1-6, 2024.

[6]. M. S. Hasan Talukder, A. Krishno Sarkar and M. Nuhi-Alamin, "An Improved Model for Nutrient Deficiency Diagnosis of Rice Plant by Ensemble Learning," 2022 4th International Conference on Sustainable Technologies for Industry 4.0 (STI), pp. 1-6, 2022.

[7]. K. W. Thar and T. T. Wai, "Machine Learning Based Predictive Modelling for Fraud Detection in Digital Banking," *2024 IEEE Conference on Computer Applications (ICCA)*, pp. 1-5, 2024.

[8]. R. K. Somkunwar, A. Pimpalkar, K. M. Katakdound, A. S. Bhide, S. P. Chinchalkar and Y. M. Patil, "A Fraud Detection System in Financial Networks Using AntiBenford Subgraphs and Machine Learning Algorithms," *2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE)*, pp. 1-6, 2023.

[9]. A. Al-Maari and M. Abdulnabi, "Credit Card Fraud Transaction Detection Using a Hybrid Machine Learning Model," *2023 IEEE 21st Student Conference on Research and Development (SCOReD)*, pp. 119-123, 2023.

[10]. M. Auleria, D. E. Saputra and Y. Yustiawan, "Data Driven Analysis of Fraudulent Transaction Characteristics in Branchless Banking," *2024 3rd International Conference on Digital Transformation and Applications (ICDXA)*, pp. 68-73,2024.

[11]. H. Ali Mohamed and S. Subramanian, "Fraud Classification In Financial Statements Using Machine Learning Techniques," *2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD)*, pp. 1-4, 2023.

[12]. R. Achary and C. J. Shelke, "Fraud Detection in Banking Transactions Using Machine Learning," 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), pp. 221-226, 2023.

[13]. Mahajan, V. S. Baghel and R. Jayaraman, "Credit Card Fraud Detection using Logistic Regression with Imbalanced Dataset," *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 339-342, 2023.

[14]. Backiyalakshmi and B. Umadevi, "A Systematic Short Review on Intelligent Fraud Detection Approaches in the Banking Sector using Deep Learning and Machine Learning with Future Trends," *2023 6th International Conference on Recent Trends in Advance Computing (ICRTAC)*, pp. 474-481, 2203.

[15]. "Synthetic Financial Datasets For Fraud Detection", Online Available: https://www.kaggle.com/datasets/ealaxi/paysim1/data . (assesed on 28 July, 2024)

[16]. P. Singla and V. Verma, "Towards Personalized Job Recommendations: A Natural Language Processing Perspective," *2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, pp. 768-773, 2023.