International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 12 | Dec - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Fraud Detection in Internet Banking Using Machine Learning

Malepu Shrishanth[1], L Nitish Reddy[2], K Harshavardhan Reddy[3], G Mahesh[4], Mrs P Venkata Pratima[5]

[1,2,3,4] *UG Scholars,* [5]*Assistant Professor*
[1,2,3,4,5] *Department of CSE [Artificial Intelligence & Machine Learning],*
[1,2,3,4,5] *Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India*

Abstract - *Banking fraud transactions refer to unauthorized or deceptive activities involving bank accounts or financial transactions. Various machine-learning algorithms can be employed to detect such fraudulent activities. This study examines several algorithms suitable for classifying transactions as either fraudulent or legitimate. The research utilizes the Banking Fraud Transactions dataset, often characterized by high imbalance. To address this issue, we are implementing multiple machine learning algorithms like Random Forest, K-Nearest Neighbour, and Decision Tree.*

Additionally, feature selection techniques are employed, and the dataset is divided into training and test sets. The algorithms evaluated in the study include Random Forest and KNN. The findings indicate that each algorithm demonstrates high accuracy in detecting banking fraud transactions. The proposed model holds promise for detecting other irregularities within financial transactions.

Key Words: *Banking Fraud Transactions*, *Machine Learning Algorithms*, *Imbalanced Dataset*

1 INTRODUCTION

The financial industry has seen a rise in fraudulent activity in recent years, including identity theft, money laundering, and credit card fraud. In addition to causing significant financial losses for consumers and financial institutions alike, these dishonest practices erode public confidence in the banking industry. It is becoming more and more necessary to implement advanced technical solutions that can identify and stop fraudulent transactions in real time in order to counter this growing threat.[3]

Machine learning has become a potent weapon in the fight against banking fraud because of its capacity to evaluate enormous volumes of data and spot intricate patterns. Machine learning algorithms can identify fraudulent and valid transactions by using past transaction data. This allows them to flag suspect activity for additional examination.[1]

The quality and representativeness of the training data, the algorithms selected, and the assessment metrics used to gauge model performance are some of the variables that affect how well machine learning models detect financial fraud. The imbalance present in banking fraud datasets—where instances of fraudulent transactions are greatly outnumbered by genuine ones—is one of the fundamental issues facing researchers and practitioners in this field. Biassed models that put accuracy ahead of accurately detecting fraudulent transactions—which are frequently the minority class—can result from this class imbalance.[14]

This project uses a comprehensive strategy to address this difficulty. First, to create a strong fraud detection model, a variety of machine learning methods are used, such as Random Forest, K-Nearest Neighbours (KNN), and Logistic Regression. These algorithms are well-suited to managing various facets of the fraud detection task since they each have unique benefits and trade-offs.[2][11]

Moreover, feature selection methods are used to find the most pertinent characteristics or traits that help differentiate between authentic and fraudulent transactions. The model's generalisation performance and predictive accuracy can be enhanced by concentrating on the most informative elements.[2][3]

To enable objective model evaluation, the dataset is also meticulously divided into training and test sets. The performance of the constructed models is thoroughly evaluated using performance metrics such area under the

receiver operating characteristic curve (AUC-ROC), precision, recall, and F1-score.[11][12]

In addition to creating an efficient fraud detection system for banking transactions, this project seeks to advance knowledge of how machine learning can be used to tackle intricate problems in the financial industry by methodically investigating and assessing these approaches. The study's conclusions could ultimately guide the creation of more resilient and flexible fraud detection systems, preserving the integrity of financial systems and shielding all parties involved from the constant danger of banking fraud.[10][13][14]

2 LITERATURE SURVEY

In order to prioritise which financial transactions should be manually examined for possible money laundering, this study looks at the goal of developing, characterising, and validating a machine learning model. A sizable data set from DNB, the biggest bank in Norway, is used to test the model. Design, methodology, and strategy Three categories of historical data are used to train a supervised machine learning model: "normal" legal transactions, transactions that the bank's internal warning system flags as suspicious, and possible money laundering cases that have been reported to the authorities. Using data like the sender/receiver's background, past actions, and transaction history, the model is trained to forecast the likelihood that a new transaction should be reported. Results The study shows that the popular strategy of avoiding incorporating non-reported alerts—that is, transactions that are looked into but not reported—in the model's training process might produce less-than-ideal outcomes. The utilisation of routine (un-investigated) transactions is no different. In terms of a fair measure of performance, our devised method performs better than the bank's current strategy. Value and originality One of the few documented anti-money laundering (AML) models for suspicious transactions that has been used on a data set of a realistic scale is this research study. Additionally, a novel performance metric designed especially to contrast the suggested approach with the bank's current AML system is included in the article.[5][6]

In order to address this issue, this study suggests using the cross-validation approach to choose the best SVM classifier parameters. It also examines recent advancements in the selection of SVM model parameters that will impact the identification effect of suspicious financial transactions. By using grid search to choose the best parameters based on the highest classification accuracy rate, the cross-validation method significantly enhances the classifier's overall performance and successfully prevents over-learning and under-learning.[7][15]

In order to prevent antimony laundering, financial institutions must be able to identify suspected money laundering transactional behavioural patterns (SMLTBPs), which is covered in length in this study. This study develops a novel cluster-based local outlier factor (CBLOF) algorithm to discover SMLTBPs by combining distance-based unsupervised clustering and local outlier detection. Its applicability and efficacy are empirically tested using both synthetic and real data.[8]

The fraud detection techniques presented in this work are designed to process large amounts of data quickly. We present a chunk-based incremental classification method based on a neural network (MLP) and a memory model to overcome the stability-plasticity conundrum and data scalability concerns. With the incremental technique, previous data chunks are kept somewhat longer while the fraud model is modified successively with incoming data chunks. For training and testing, we use a sizable dataset on credit card fraud, which we divide into initial and incremental portions. We resolve the data skew problem, a crucial difficulty in fraud detection, by using data sampling. We use the testing chunk to assess the adjusted MLP classifier's performance following each incremental phase. The experimental findings show that our incremental approach is more successful and efficient than the non-incremental MLP.[9]

3 PROBLEM STATEMENT

The identification and detection of fraudulent activity in banking transactions, which constitute a serious danger to financial security in the digital age, is the issue this study attempts to solve. It can be difficult to correctly identify transactions as fraudulent or lawful because of the extremely unbalanced nature of the dataset that is usually linked to financial fraud. The purpose of this study is to investigate how well several machine learning algorithms—such as Random Forest, K-Nearest Neighbours, and Decision Tree—address this problem. The study aims to improve the detection accuracy of

banking fraud transactions and aid in the creation of reliable models for spotting financial irregularities by utilising feature selection techniques and assessing the performance of these algorithms on different training and test datasets.

4 PROPOSED METHODOLOGY

By using machine learning techniques, the suggested method aims to improve the identification of fraudulent activity in banking transactions. It starts by gathering transaction data, which includes crucial information like quantities, timestamps, merchant details, and client specifics. In order to train machine learning algorithms such as Random Forest, Decision Tree, and Logistic Regression to identify patterns suggestive of fraudulent activity, the pre-processed data must be separated into training and testing sets.

Dataset: The starting point, where data is collected and prepared.

Data Preprocessing & Feature Selection: Data is cleaned and normalized, and relevant features are selected.

Data Splitting: The dataset is divided into training and testing sets.

RF Algorithm & Evaluation: The Random Forest algorithm is applied to the training data, and its performance is evaluated on the testing data.

Prediction & Accuracy: The trained model makes predictions on new data, and its accuracy is assessed.

Deployment using FLASK: The model is deployed as a web service using the Flask framework for user interaction.

Using the Banking Fraud Transactions dataset, the project seeks to create a reliable machine learning-based system for identifying banking fraud transactions. The implementation of several methods, such as Random Forest, K-Nearest Neighbours (KNN), and Logistic Regression, addresses the issue of the high-class imbalance in this dataset. The dataset is separated into training and test sets for assessment, and feature selection approaches are used to improve model performance. The study assesses how well each algorithm performs in correctly identifying transactions as either legitimate or fraudulent. To give a thorough evaluation, the project takes into account a number of performance indicators in addition to accuracy, including precision, recall, F1-score, and AUC-ROC. The results of thorough testing and analysis show that the suggested model may successfully identify banking fraud activities, with the possibility of wider uses in spotting anomalies in financial transactions.

4.1 METHODOLOGIES

4.1.1 Data Collection:

The process of gathering data is the first actual step in creating a learning model. This crucial stage will have a cascading effect on the model's performance; the more and better data we collect, the better the model will function. Data can be gathered using a variety of methods, including manual interventions, online scraping, and money laundering detection.

4.1.2 Dataset:

The dataset is made up of individual data; it has 11 columns and 1048576 rows, which are explained below.[16]

1)step
2)type
3)amount
4)nameOrig
5)oldbalanceOrg
6)newbalanceOrig
7)nameDest
8)oldbalanceDest
9)newbalanceDest
10)isFraud
11)isFlaggedFraud

4.1.3 Data Preparation:

Sort data and get it ready for training. Remove duplicates, fix mistakes, deal with missing numbers, normalise, convert data types, and clean up everything that could need it.Data should be randomised to eliminate the impact of the specific sequence in which they were gathered and/or otherwise prepared.

Use data visualisation to carry out exploratory research or find pertinent correlations between variables or class imbalances (beware of bias!). Divide the sets into training and evaluation.

4.1.4 Model Selection:

We developed our money laundering detection system using the **RANDOM FOREST** system. After achieving a **98.04%** accuracy rate on the test set, we put this method into practice.

4.1.5 Analyze and Prediction:

In the actual dataset, we chose only main 2 features:

➢ Amount transactions - detailed descriptions of the Amount transactions data.

➢ Fraud – it indicates whether the details of the transactions are fraudulent or not.

4.2 WORKFLOW

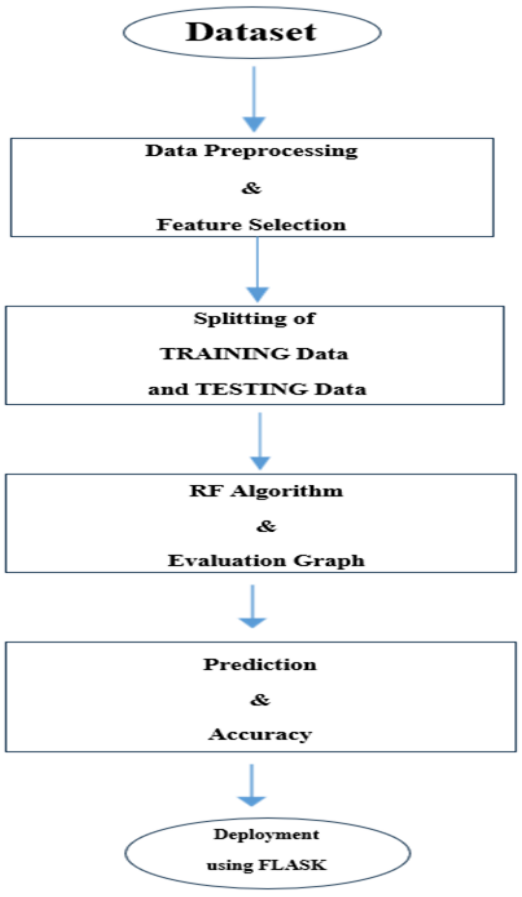

Fig1: Represents WorkFlow of the Model.

5 PROPOSED TECHNIQUE AND ALGORITHM USED:

Random Forest and KNN: During training, Random Forest builds a large number of decision trees, making it a flexible and potent ensemble learning technique. Every tree in the forest learns to categorise occurrences on its own using a subset of features, and the sum of all the trees' predictions determines the final classification. A straightforward yet powerful approach for classification and regression applications is K-Nearest Neighbours (KNN). It works on the similarity principle, which states that a data point's classification is based on the feature space class labels of its closest neighbours. Because KNN saves all training data points and their associated labels, it eliminates the need for explicit training and is simple to use and understand.

$$\text{ACCURACY} = \frac{(\text{TruePositive} + \text{TrueNegative})}{\text{Total Sample Accuracy}}$$

TruePositive :- In this instance, both the actual output and our prediction were **YES**.

TrueNegative :- In this instance, we predicted **NO**, and the actual result was likewise **NO**.

Total Sample Accuracy :-The percentage of properly identified samples over the whole test dataset is known as total sample accuracy.

CONFUSION MATRIX :-

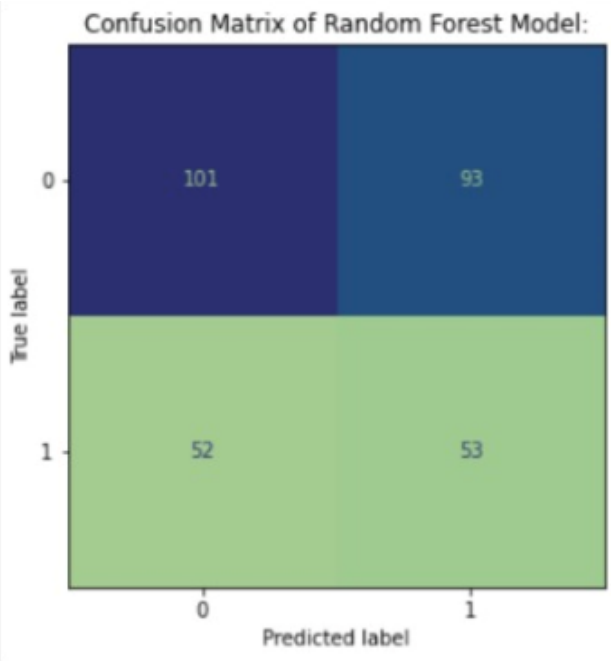

Fig2 **:** Represents confusion matrix of random forest model.[16]

Results:-

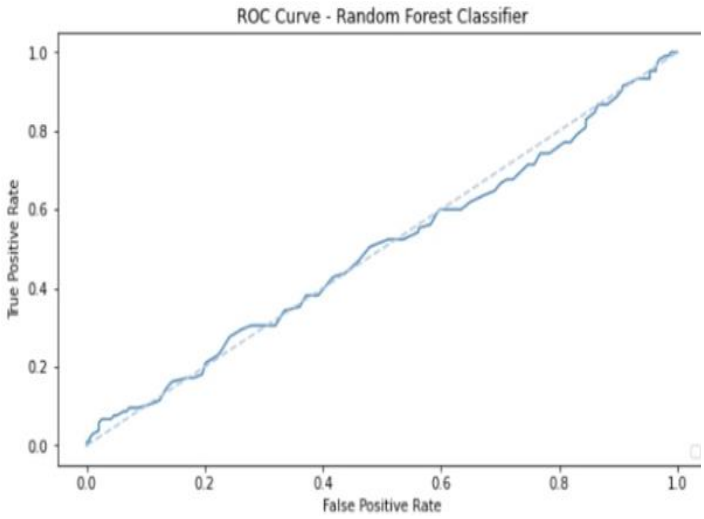

fig3 : Represents ROC Curve of TruePositive Rate and TrueNegative Rate.[16]

- ROC Curve of TruePositive Rate and TrueNegative Rate

----- - Total Sample Accuracy Rate

FUTURE ENHANCEMENT & CONCLUSION

Improving the calibre and applicability of the features used for model training is a future improvement in machine learning that will eventually result in improved generalisation and predictive performance. Feature transformation, feature engineering, and feature selection are some of the methods for improving features. Finding the most informative subset of features from the original feature space is the goal of feature selection, which lowers computational cost and dimensionality while maintaining or even increasing model accuracy. Finally, this work has shown that machine learning algorithms such as Random Forest, K-Nearest Neighbours (KNN), and Logistic Regression are successful in identifying transactions that involve banking fraud. We have created strong models that can reliably identify transactions as either fraudulent or legitimate by tackling the class imbalance issue with meticulous algorithm selection, feature enhancement strategies, and thorough assessment criteria. The results emphasise how crucial it is to use a variety of methods for detecting fraud because every algorithm has unique advantages.

REFERENCES

[1] M. Jullum, A. Løland, R. B. Huseby, G. A˚ non-sen, and J. Lorentzen, "Detecting money laundering transactions with machine learning," Journal of

Money Laundering Control, vol. 23, no. 1, pp. 173–186, Jan 2020.

[2] L. Keyan and Y. Tingting, "An improved support-vector network model for anti-money laundering," in 2011 Fifth International Conference on Management of e-Commerce and e-Government. IEEE, 2011, pp. 193– 196.

[3] R. Liu, X.-l. Qian, S. Mao, and S.-z. Zhu, "Research on anti-money laundering based on core decision tree algorithm," in 2011 Chinese Control and Decision Conference (CCDC). IEEE, 2011, pp. 4322– 4325.

[4] Z. Gao, "Application of cluster-based local outlier factor algorithm in anti-money laundering," in 2009 International Conference on Management and Service Science. IEEE, 2009, pp. 1–4.

[5] J. de Jes´us Rocha Salazar, M. Jes´us Segovia-Vargas, and M. del Mar Camacho-Mi˜nano, "Money laundering and terrorism financing detection using neural networks and an abnormality indicator," Expert Systems with Applications, p. 114470, Dec 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417420311209

[6] E. L. Paula, M. Ladeira, R. N. Carvalho, and T. Marzagao, "Deep learning anomaly detection as support fraud investigation in Brazilian exports and anti-money laundering," in 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2016, pp. 954–960.

[7] F. Anowar and S. Sadaoui, "Incremental Neural-Network Learning for Big Fraud Data," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 2020-Octob.
Institute of Electrical and Electronics Engineers Inc., Oct 2020, pp. 3551–3557.

[8] G. A. Carpenter and S. Grossberg, "A massively parallel architecture for a self-organizing neural pattern

recognition machine," Computer vision, graphics, and image processing, vol. 37, no. 1, pp. 54–115, 1987.

[9] G. Carpenter, "An adaptive resonance algorithm for rapid category learning and recognition," Neural Networks, vol. 4, pp. 439–505, 1991.

[10] T. Kohonen, "Self-organized formation of topologically correct feature maps," Biological cybernetics, vol. 43, no. 1, pp. 59–69, 1982.

[11] A. Ultsch, "Kohonen's self-organizing feature maps for exploratory data analysis," Proc. INNC90, pp. 305–308, 1990.

[12] Senator T E, Goldberg H G, Wooton J, et al. The Financial Crimes Enforcement Network AI System(FAIS)-identifying Potential Money Laundering from Reports of Large Cash Transactions[J]. AI Magazine, 1995, pp. 21-39.

[13] Zdanowicz John S. Detecting Money Laundering and Terrorist Financing Via Data Mining [J]. Communications of the ACM, 2004, pp.53-55.

[14] Bolton R J, Hand D J. Statistical Fraud Detection[J]. Statistical Science, 2002, pp. 235-254.

[15] Zhang Yan, Ouyang Yiming, Wang Hao, Wang Xidong, Application of Data Mining in the Financial Field Computer Engineering and Applications, vol.18, pp.208-211, 2004

[16] https://github.com/Nitish-37/Fraud-detection-in-Online-Banking/blob/main/fraud.csv