Fraud Detection in Banking Using Machine Learning Algorithm

Saptaparni Chatterjee
Faculty in Sister Nivedita University
Computer Science and Engineering

***Abstract:** The number of financial transactions has grown dramatically in the last few years due to the growth of financial institutions as well as the acceptance of web-based e-commerce. In internet banking, fraudulent transactions are becoming an increasingly common issue, and detecting fraudulent activity has never been easy. In banking, detection of fraud is observed as a binary machine learning problem where input is either categorized as fraud or not. In this proposed methodology, we have proposed a banking fraud detection using three steps: Pre-processing, feature extraction as well as classification. Pre-processing has done for data cleaning process, and then the feature extraction used for detecting transaction time and finally classification hybrid (Decision Tree-Random Forest) for detecting the fraud in banking. Results showed that the hybrid algorithm gives 95.8% accuracy and least error compared with single machine learning algorithms.*

***Keywords:** Fraud detection, Hybrid ML, Banking system, Decision Tree and Random forest.*

I. Introduction

In the modern era, nearly everyone has to interact with a bank, either in person or virtually, making the banking industry a very significant industry [1]. Both consumers and banks run the risk of falling victim to scammers when interacting with them. The following are a few instances of fraud: accountancy, credit card, insurance, and so on [2]. Thus, identifying fraudulent behavior is essential to keeping these expenses under control. Debit as well as credit card fraud, report fraud, assurance fraud, money laundering fraud, etc. are some of the most prevalent forms of bank fraud. The act of obtaining financial advantages through dishonest and unlawful means is known as financial fraud [3]. Financial fraud can happen in many different settings, such as the banking, insurance, corporate, and tax domains. Businesses and industries are facing an increasing number of financial fraud cases, including money laundering including financial transaction fraud recently [4, 5]. Since a lot of money is lost to fraud daily, although numerous attempts to curb it, financial fraud continues to hurt society and the economy [6]. Many years ago, many methods for detecting fraud were introduced. Most conventional methods need human work, which is not only impractical but also costly, time-consuming, and inaccurate. [7]. although more research is being done, it is ineffective in reducing losses brought on by deception [8]. As machine learning algorithms have advanced, they have been used to identify fraudulent activity in the financial sector. To forecast fraudulent activity, both supervised and unsupervised techniques were used [9]. This article uses machine learning approaches, such as hybrid machine learning (DT-RF), to detect bank fraud. Pre-processing, feature extraction, and classification made up the dataset used in the suggested method for detecting financial fraud [10].

II. Literature survey

On European, Small, as well as Tall cards in three distinct financial datasets, Nguyen et al. [11] examined the effectiveness of CNN also LSTM deep learning method-based approaches with ANN, RF, and SVM machine learning approaches. Sampling approaches were used in this work to solve the issue of class imbalance, which resulted in increased performance on examples that were already available but significantly decreased performance on newly discovered data. According to experimental results, the suggested deep learning techniques perform better at identifying credit card fraud than conventional machine learning models, suggesting that the suggested strategies may be applied to identify credit card fraud in practical settings. Compared with all other algorithms, LSTM by 50 blocks achieved an F1 score of 84.85%.

A goal of this research, according to Sinap, V. et al. (2024), is to compare machine learning algorithms based on a range of performance indicators after assessing each one's effectiveness in the detection of credit card fraud. Logistic Regression, Decision Trees, Random Forest, XGBoost, Naive Bayes, K-Nearest Neighbors, and Support Vector Machine were the seven supervised classification techniques that were employed. Thus, with 97% accuracy rates, the "Random Forest" as well as "K-Nearest Neighbors" algorithms attained best performance levels in this study.

Aslam, A. et al. conducted a comparative analysis of state-of-the-art machine-learning algorithms designed toward recognize credit card fraud in 2024. The study examines the effectiveness of these ML algorithms with a publicly accessible dataset of approximately 550,000 transactions of credit card made by European cardholders in 2023. With stated accuracy, recall, as well as F1 score of 1.00 for both classes, ML models such as LGBM, random forest, additional trees, along with logistic regression can obtain great precision and accuracy in the credit card fraud investigation dataset.

The depiction of several machine learning techniques, including Decision Tree, Random Forest, linear regression, along with Gradient Boosting approach, are compared meant for the detection and prediction of fraud instances utilizing loan fraudulent manifestations in 2023 by Maashi, M. et al. [13]. Additional model accuracy metrics, including the Receiver Operating Characteristic (ROC) curve and calculations of accuracy, precision, recall, and F-1 score, have been carried out using a confusion matrix.

Thirteen arithmetic as well as machine learning models for the detection of payment card fraud were examined in 2024 by Seera, M. et al. [14] utilizing both real and publicly accessible transaction information. Analyses and comparisons are made between the original feature and aggregated feature results. The results confirm in a good way how well-aggregated characteristics work when applied to actual payment card fraud detection tasks.

A number of actions have been conducted to study customer behavior in 2024, according to Hamidi, H.O.J.A.T. *et al.* [15]. This is in the detection of bank frauds. Following the application of intelligent tools to test various algorithms, the two algorithms—XGBoost and LightGBM—that produced highest ROC in the models were gradually chosen. Simultaneously, it has been employed in final examinations with a decrease in fictitious samples classified as fraudulent (FP). This model provides very good results in detecting card-to-card fraud and was constructed using real development data. This approach can be applied as a tool to lower financial crimes and greatly enhance the security of the banking system.

A comparative analysis of four Quantum Machine Learning (QML) models was carried out in 2024 by Innan, N. et al. [16] with the purpose of detecting financial fraud. We demonstrated that the model with the highest performance, Quantum Support Vector Classifier, had F1 scores of 0.98 for both frauds as well as non-fraud classes. The promise of QML financial applications classification is boosted by other models that show promising results, such as the Variational Quantum Classifier, Estimator Quantum Neural Network (QNN), as well as Sampler QNN.

Various approaches as well as patterns while enhancing elasticity along with accuracy in detection. In 2024 Huang, Z *et al.,* [17] proposed a machine learning-based K-means clustering way toward improve the exactness as well as effectiveness of detection of financial fraud. Furthermore, via allowing targeted checking plus avoidance efforts in high-risk areas, K-means clustering also helps financial institutions optimize the allocation of their resources, so successfully reducing the impact of fraud on the financial system as a whole. In summary, as it works to create a safer and dependable the K-means clustering method, which is based on machine learning, has great potential for use in the financial identification of fraud field owing to its transaction environment for the banking industry.

2.1 Research gap

The study aims to apply hybrid ML methods for the purpose of detecting financial fraud in relation to transaction volume and time. The suggested method of pre-processing, feature extraction, and classification was taken into consideration for the detection of financial fraud.

Initially, pre-processing has done to remove unwanted handling data and it can reduce complexity, correcting errors, inconsistencies, and improve the model's overall performance. And then the feature extraction based on time features can be extracted from the pre-processed data.

Hybrid machine learning algorithms have been utilized to solve classification problems and provide the highest level of accuracy. ML models such as Random Forest and Decision Tree were used for both regression and classification issues. When compared to individual trees, RF is known to reduce over fitting, making it useful for complicated fraud detection. It processes different kinds of data and delivers answers that are comprehensible, just like decision trees. Its ensemble technique, which combines many trees to capture complex patterns and relationships in data, improves accuracy and robustness. The ensemble technique offers much increased performance, which justifies the trade-off despite its computing needs.

2.2 Problem statement

On training datasets, binary machine learning methods like HMM, Naïve Bayes, Decision Tree, and others produce overfitting and classification issues. Additionally, making judgments based on the feature set requires additional time. Certain machine learning algorithms do not produce more accurate results for the input datasets. Large dataset training is challenging, and computational time is higher.

III. Methodology

The unlawful and unauthorized use of a bank account by someone who is not the account owner is known as fraud. When someone uses another person for their own grow at another person's bank that is known as banking fraud. Financial fraud is on the rise due to the lack of an appropriate method for safeguarding all accounts. The research needed to examine the scam is lacking. Several binary machine learning techniques are used to identify instances of banking fraud. In this proposed research, hybrid based ML techniques have used for detecting fraud in banking. Initially, pre-processing has done to remove the unwanted handling data and then the feature extraction step is carried out to extracting the transaction hour feature and then the classification has done for detecting fraud in banking. Result demonstrated that the hybrid algorithm gives the high accuracy for fraud detection in banking compared to binary ML algorithms.

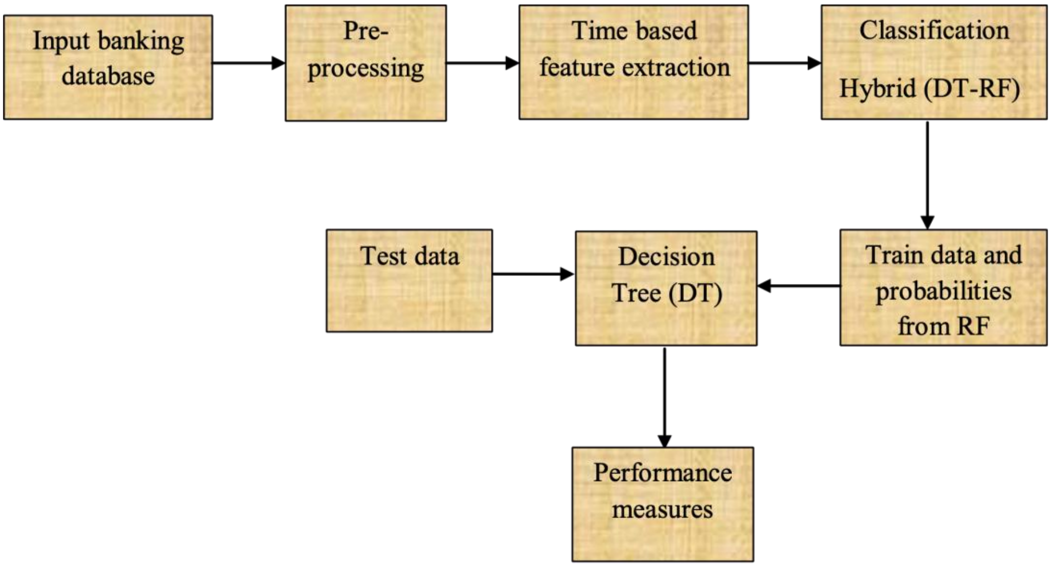

Fig 1: Overview of proposed method

3.1 Input Dataset

In the banking dataset contains individual deposits, withdrawals as well as transfers. Including, they contain credit card transactions, individual account details etc.

3.2 Pre-processing

It involves recognizing as well as error correcting, contradiction, and inaccuracies in data, thereby improving the value as well as reliability of information. Also, it removes data redundancy, remove unwanted handling data.

3.3 Feature Extraction

The time (in seconds) that passes connecting every transaction as well as the initial transaction is included in "time" feature. We extend the feature to take out the transaction hour feature, which provides us with additional information than the time feature alone, in order to maximize its usefulness.

3.4 Classification

In the classification hybrid ML utilized to detect the fraud activities and classify as fraud or non-fraud in banking system.

3.4.1 Decision tree

Decision Tree is a powerful algorithm that divides the feature space into segments for classification in order to detect fraud in online payments. It is well-known for being easily understood and simple to visualize. Decision trees are still a useful tool because of their ability to handle different kinds of data, even though they have a tendency to overfit [18]. It is appropriate for challenging fraud detection jobs due to its capacity to capture nonlinear correlations and interactions between features. Furthermore, Decision Tree models are useful in decision-making processes since they are simple enough for non-technical stakeholders to understand.

The decision tree model for the regression problem is constructed using the gini and gini index. When dealing with classification problems, the knowledge gained from the selection process is used to determine which root node has low entropy and high information once more. When solving regression problems, the feature with the lowest gini value is chosen as the root, allowing the depth of the tree to be calculated.

Predicting dependent variables on independent .

factors Y is done with a DT. $X = X_1, X_2 ......X_n$ DT, several nodes are created that connect input and output data samples. The frequency with which an input sample selected at random has an erroneous label is ascertained using the Gini impurity measure.

The Gini index for a data set by means of k classes can be calculated utilizing

$$F_j = 1 - \sum_{i=1}^{j} P_k^{\,2}$$

Where $i \in \{1, 2.....j\}$ and P_k represents the proportions of samples in class k. Each DT node encodes a rule that separates the input characteristics. Subject to a stopping requirement, new nodes can be formed to form a tree structure. The majority of data samples that originated from a leaf node of the tree are detected when an input sample [19] is given, resulting in a predicted target class.

3.4.2 Random Forest

Building on the foundation of decision trees, Random Forest is a powerful machine learning technique for fraud detection in banking systems. It creates a several decision trees are used in the training process, and the class mode is output for classification. Random Forest is useful for complicated [20] fraud detection because it is known to reduce overfitting as compared to individual trees. It processes different kinds of data and delivers answers that are comprehensible, just like decision trees. Its ensemble technique, which combines many trees to capture complex patterns and relationships in data, improves accuracy and robustness. The ensemble technique offers much increased performance, which justifies the trade-off despite its computing needs.

$$\theta_k = \theta_{k1}, \theta_{k2}, ...\theta_{kp}$$

It can be written as $h_k(x) = h(x|\theta_k)$. For classification function $f(x)$ connects every classifier outputs, whereby each DT outputs a vote for the majority likely class given input x.

3.4.3 Hybrid (DT-RF)

The hybrid model in this case makes use of the random forest and decision tree methods. Running on random forest probabilities is the combined model. The decision tree method receives input from the random forest's probability and train data. To find and load test data with decision tree probabilities, same procedures are followed. At the end, values are forecast.

IV. Result and discussion

The categorization of the model depends on the dataset. The official website of the competent data science organization Kaggle provided the dataset. This dataset contains details on millions of transactions, some of which were fraudulent. The system is evolving more smoothly and steadily as a result. This dataset illustrates the difficulties in gathering information about the rising threat of financial fraud. There are 294,907 records in the banking dataset, of which 80% are used for training and 20% are tested to identify potential fraud activity.

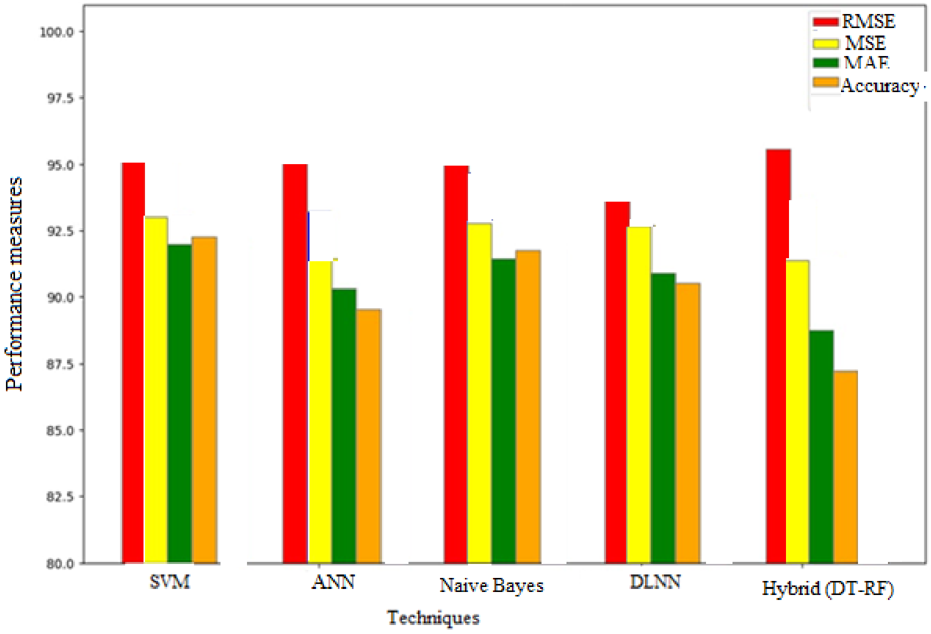

Fig 2: Analysis of performance measures

In the fig 2 explains the analysis of performance measures compared with various ML algorithms. Compared algorithms are SVM, ANN, Naïve Bayes, Hidden Markov Model and proposed hybrid (DT-RF). The fraud detection prediction of proposed hybrid ML gives the better accuracy and lessens the error rate.

V. Conclusion

The most prevalent issue that causes people to lose money is banking fraud, which also causes losses for certain banks and credit card companies. This paper, investigated the utilization of ML to detect banking fraud. For the purpose of enhancing fraud detection performance, we suggested a hybrid (DT-RF). Analysis of pre-processing feature extraction and classification is used to detect banking fraud. Pre-processing was done to clean the data; feature extraction was then carried out to extract the transaction time; and classification was the last step in identifying banking fraud. Hybrid algorithms have lower error performance and improved accuracy of 95.8% when compared to existing ML methods.

Reference

[1]. H. Feng, ‘‘Ensemble learning in credit card fraud detection using boosting methods,’’ in Proc. 2nd Int. Conf. Comput. Data Sci. (CDS), Jan. 2021, pp. 7–11.

[2]. E. F. Malik, K. W. Khaw, B. Belaton, W. P. Wong, and X. Chew, ‘‘Credit card fraud detection using a new hybrid machine learning architecture,’’ Mathematics, vol. 10, no. 9, p. 1480, Apr. 2022.

[3]. Matloob, S. A. Khan, R. Rukaiya, M. A. K. Khattak, and A. Munir, ‘‘A sequence mining-based novel architecture for detecting fraudulent transactions in healthcare systems,’’ IEEE Access, vol. 10, pp. 48447–48463, 2022.

[4]. R. Almutairi, A. Godavarthi, A. R. Kotha, and E. Ceesay, ‘‘Analyzing credit card fraud detection based on machine learning models,’’ in Proc. IEEE Int. IoT, Electron. Mechatronics Conf. (IEMTRONICS), Jun. 2022, pp. 1–8.

[5]. T. Vairam, S. Sarathambekai, S. Bhavadharani, A. K. Dharshini, N. N. Sri, and T. Sen, ‘‘Evaluation of Naïve Bayes and voting classifier algorithm for credit card fraud detection,’’ in Proc. 8th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS), Mar. 2022, pp. 602–608.

[6]. D. Almhaithawi, A. Jafar, and M. Aljnidi, ‘‘Example-dependent costsensitive credit cards fraud detection using SMOTE and Bayes minimum risk,’’ Social Netw. Appl. Sci., vol. 2, no. 9, pp. 1–12, Sep. 2020.

[7]. Y. Chen and X. Han, ‘‘CatBoost for fraud detection in financial transactions,’’ in Proc. IEEE Int. Conf. Consum. Electron. Comput. Eng. (ICCECE), Jan. 2021, pp. 176–179.

[8]. Zahid, S.Z.S., Muhammad, H.M.U.H.H., Hafeez, U., Iqbal, M.J.I.M.J., Asif, A.A.A., Yaqoob, S.Y.S. and Mehboob, F.M.F., 2024. Credit Card Fraud Detection using Deep Learning and Machine Learning Algorithms. Journal of Innovative Computing and Emerging Technologies, 4(1).

[9]. Gorte, A.S., Mohod, S.W., Keole, R.R., Mahore, T.R. and Pande, S., 2022, November. Credit Card Fraud Detection Using Various Machine Learning and Deep Learning Approaches. In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2022, Volume 3 (pp. 621-628). Singapore: Springer Nature Singapore.

[10]. Prajapati, D., Tripathi, A., Mehta, J., Jhaveri, K. and Kelkar, V., 2021, December. Credit Card Fraud Detection Using Machine Learning. In 2021 International Conference on Advances in Computing, Communication, and Control (ICAC3) (pp. 1-6). IEEE.

[11]. T. T. Nguyen, H. Tahir, M. Abdelrazek, and A. Babar, “Deep learning methods for credit card fraud detection,” 2020 Dec.

[12]. Sinap, V., 2024. Comparative analysis of machine learning techniques for credit card fraud detection: Dealing with imbalanced datasets. Turkish Journal of Engineering, 8(2), pp.196-208.

[13]. Aslam, A. and Hussain, A., 2024. A Performance Analysis of Machine Learning Techniques for Credit Card Fraud Detection. Journal of Artificial Intelligence (2579-0021), 6.

[14]. Maashi, M., Alabduallah, B. and Kouki, F., 2023. Sustainable financial fraud detection using garra rufa fish optimization algorithm with ensemble deep learning. Sustainability, 15(18), p.13301.

[15]. Seera, M., Lim, C.P., Kumar, A., Dhamotharan, L. and Tan, K.H., 2024. An intelligent payment card fraud detection system. Annals of operations research, 334(1), pp.445-467.

[16]. Hamidi, H.O.J.A.T. and Karbasiyan, M., 2024. Presenting a Model to Detect the Fraud in Banking using Smart Enabling Tools. International Journal of Engineering, 37(3), pp.529-537.

[17]. Innan, N., Khan, M.A.Z. and Bennai, M., 2024. Financial fraud detection: a comparative study of quantum machine learning models. International Journal of Quantum Information, 22(02), p.2350044.

[18]. Huang, Z., Zheng, H., Li, C. and Che, C., 2024. Application of Machine Learning-Based K-Means Clustering for Financial Fraud Detection. *Academic Journal of Science and Technology*, *10*(1), pp.33-39.

[19]. Pradhan, S.K., Rao, N.K., Deepika, N.M., Harish, P., Kumar, M.P. and Kumar, P.S.,

2021, December. Credit Card Fraud Detection Using Artificial Neural Networks and Random Forest Algorithms. In 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1471-1476). IEEE.

[20]. Chauhan, A., Mogha, S. and Verma, D., Banking Fraud Detection using Python and Machine Learning, International Journal of Innovations in Management, Science and Engineering (IJIMSE), ISSN: 2582-6778, Volume-03, Issue-01, September 2022.