Intelligent Credit Card Fraud Detection System

Gujjula Srisahajananda
Computer Science and Engineering
Dhanalakshmi Srinivasan University
Samayapuram,Trichy
sahajanandareddy251@gmail.com

Madamshetty Vamshi
Computer Science and Engineering
Dhanalakshmi Srinivasan University
Samayapuram,Trichy
madamshettyvamshivamshi@gmail.com

Madala Rosaiaha Chowdary
Computer Science and Engineering
Dhanalakshmi Srinivasan University
Samayapuram,Trichy
rasaiahachowdary7@gmail.com

Ms.R Santhana Lakshmi
Assistant Professor,SET
Dhanalakshmi Srinivasan University
Samayapuram,Trichy
santhanalakshmi.cse@dsuniversity.ac.in

Abstract— Credit card fraud has emerged as a key issue for the financial sector due to the increase in the number of cashless transactions. Traditional rule-based anti-fraud techniques are not found to perform well in the detection of complex credit card fraud patterns. [3]This paper introduces a machine learning approach for credit card fraud detection, inspired by earlier researchers in the field of data mining for credit card fraud detection. The proposed system follows a supervised learning approach to identify fraud transactions in a cashless environment. An appropriate preprocessing approach is adopted for handling class imbalance. The performance of these models is quantified through accuracy measures such as precision, recall, accuracy, and F1 score[6]; these are more appropriate for fraud detection problems compared to accuracy. The experiment results show that machine learning can improve fraud detection performance significantly compared to traditional approaches. This paper also emphasizes data-driven fraud detection approaches and their significant role in developing credit card fraud detection models that are trustworthy and scalable.

Keywords--Credit Card Fraud Detection, Machine Learning, Data Mining, Imbalanced Dataset, Classification Algorithms, Financial Security

I Introduction

The rapid rise of digital payment systems and online transactions has led to a rise in the risk of credit card fraud[3]. Credit card fraud involves the unauthorized use of credit card information to conduct financial transactions without the consent of the credit card owner or customer, resulting in huge financial losses. With millions of transactions happening every day, detection of fraudulent activities cannot be achieved based solely on traditional methods[7].

Traditional fraud detection systems use rulebased systems where rules and thresholds are set in advance based on domain expertise[9]. These systems, even though simple to implement, are unable to adapt to new and developing patterns of fraud. The reason for this is that fraudsters change their patterns of operation in order to avoid traditional systems securing the online environment. In recent years, machine learning has been identified as a promising approach in credit card fraud detection[8]. Here, historical transactions can be analyzed, and the patterns can be identified between genuine and fraudulent transactions using machine learning models.

In order to overcome this challenge, data preprocessing and imbalance handling play a vital role. In this context, data preprocessing and imbalance handling play a vital role[10]. Data preprocessing helps to find the appropriate model that can detect the maximum number of fraud transactions without raising false alarms. For instance, the use of precision, recall, and F1 score measures is more suitable than the accuracy of the model to evaluate a fraud detection system.

This paper aims to develop a machine learningbased credit card fraud detection system that deals effectively with the problem of imbalanced data to improve the performance of fraud detection. Various classification algorithms are compared based on their standard evaluation metrics[10]. The aim of this study is to illustrate how a data-driven approach can help in enhancing fraud-detection efficiency and simultaneously providing a reliable decision-support system for financial institutions.

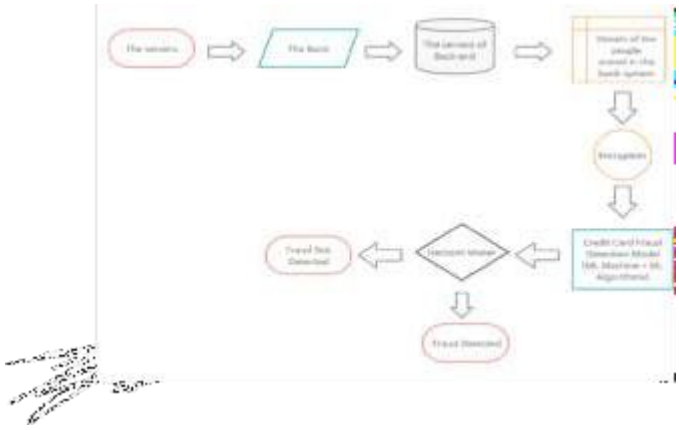

FIG. 1 Credit Card Fraud Detection Process Using Machine Learning

II Related Work

Various studies have also been done on using different data mining and machine learning approaches for solving credit card fraud detection problems. Bhattacharyya et a[5]l. carried out an extensive comparative study on the usage and evaluation of differrent classification methods to detect fraudulent credit card transaction instances.

The research emphasized the value of understanding transaction patterns rather than relying on rule-based systems.

In their research, various models of machine learning,[10] i.e., Logistic Regression, Decision Tree, Random Forest, and Neural Networks, were tested using practical data related to credit card transactions. Various points, as mentioned in the research, clarified that no particular model is fully efficient in any situation, i.e., the efficiency of models is still dependent upon the data applied as well as evaluation metrics. The research indicated better results of ensemble models, particularly Random Forest, in relation to the detection of fraudulent transactions.

This work provided a valuable contribution to the discussion on class imbalance, as it is a critical challenge in credit card fraud detection[9], given the very small portion of fraudulent transactions over the whole dataset. Traditional accuracy-based evaluation may mislead the model selection. In this work, performance measures like precision and recall and cost-sensitive evaluation were presented as suitable measures for fraud detection systems.

The results obtained in Bhattacharyya et al. give a concrete base for further studies in this area. Their comparative analysis really encourages the use of machine learning-based approaches, keeping appropriate data preprocessing and evaluation strategies in mind for an effective and reliable credit card fraud detection system[9]

Another important contribution of this work was its focus on evaluation methodology.
Bhattacharyya et al. pointed out that fraud detection is inherently a cost-sensitive and imbalanced classification problem[7]. Fraudulent transactions form only a tiny percentage of the transaction volume, and models learned without any class balancing mechanism tend to overgeneralize to the class representing normal transactions, leading to very poor fraud detection results[5]. The use of accuracy alone can be misleading under such circumstances, and the authors suggested precision, recall, and cost-based measures instead.

III Problem Statement and Objectives

Credit card usage has, therefore, rapidly increased online and offline transactions, leading

to a corresponding rise in fraudulent activities[4].

With financial institutions processing millions of transactions every day, fraud detection cannot be an efficient and practical method[7]. Most fraud detection systems are rule-based and call for the use of predefined thresholds or expert knowledge. Such systems lack adaptability and mostly fail to detect fraud patterns that are new and constantly changing.

Another big challenge to detecting credit card fraud lies in the seriously imbalanced nature of the transaction data[2]. The fraudulent cases come only as a very small portion of the overall dataset, and this causes most machine learning models to be biased toward legitimate transactions. Consequently, it may misclassify fraud cases and increase financial loss and a decline in customer

Trust Moreover, an improper fraud detection system might produce many false positive results, which could wrongly classify honest transactions as fraudulent ones.

This is detrimental to customer service and adds extra costs to the corresponding financial organizations[12]. Thus, it is necessary to develop an intelligent, adaptive, and reliable fraud detection system to successfully detect fraud transactions while avoiding false positive results The main objective of the research lies in the design and development of a viable machine learning-based credit card fraud detection system that enables the differentiation between fraudulent and nonfraudulent transactions with the use of historical transactional data.

This specific research aims to address the class imbalance problem as well. Data preprocessing and resampling techniques are used to achieve the best results in fraud detection[7]. Another specific objective of the proposed research lies in comparing a number of different classification algorithms that assist in selecting the best algorithm in the detection of fraudulent activities. In order to assess the proposed system effectively, a number of metrics are used to measure the performance. It is hoped

IV. Proposed Methodology

The proposed methodology aims to design a machine learning-based technique that can be effectively followed to conduct credit card fraud detection schemes[9]. The overall process of credit card fraud detection involves data collection, preprocessing, training, testing, and performance evaluation. The proposed system's workflow has been depicted in Fig. 2.

First, a historical dataset for credit card transactions is given, and this dataset has features associated with the transaction and corresponding class labels for the transaction, indicating whether the transaction is a fraud or not[9]. If the dataset for the transaction is too large, as is normally the case for a real-world scenario, some preprocessing techniques are applied.concentrating the model on the face features, eliminating noise variables or other irrelevant information found in the background. The collected facial images are split into training and testing datasets.

The dataset is then split into two subsets, one for training and one for testing. It is observed that 80% of the dataset is used for training the model, whereas 20% is used for testing[12]. Such distribution is used to train the model with regard to

transaction patterns

To overcome the problem of class imbalance, appropriate resampling techniques are included as part of the training process. Such techniques will ensure that a significant number of fraudulent transactions are represented, helping the algorithm to accurately identify clear characteristics of fraudulent cases[15].

A. Data Collection

As mentioned earlier, the first step of the proposed methodology includes the collection of historical credit card transactions[11]. The dataset includes various features related to credit card transactions along with corresponding class labels for fraud and non-fraud transaction**s.**

B. Data Preprocessing

The process of data preprocessing is a vital step to ensure the quality and accuracy of the dataset. Data preprocessing is defined as the step where

irrelevant features are removed[11], missing values are dealt with, and feature scaling is performed. Due to the extreme class imbalance, resampling is performed to ensure the training of the model considers the fraudulent values adequately.

C. Splitting of the Dataset

After preprocessing, the dataset is divided into two subsets: a training set and a testing set[15]. About 80% of the data is used to train the model, with the remaining 20% reserved for testing. This split evaluates how well the model generalizes on unseen transaction data.

C. Splitting of the Dataset

The data is then divided into two sets, which include a training set and a testing set[12]. It is estimated that 80% of the data is used in training, while 20% is used as a test for the model. The test is a measure of the generality of the model, focusing particularly on transactions.

E. Model Testing and Evaluation

Once the training is complete, the model is subjected to a test using the testing data to produce predictions on the data provided[15]. The predictions produced by the system are compared to the actual labels to determine the effectiveness of the system. Accuracy, precision, recall, F1 score, etc., are used to determine the effectiveness of the fraud detection system, particularly in identifying frauds.

Algorithms:

Logistic Regression:

It is a classification technique that can be used to predict fraud[13]. It is a supervised learning algorithm that predicts the probability of fraud (0 = genuine, 1 = fraud).

It is suitable for linearly separable data and is fast, interpretable, and scalable to large financial datasets.

However, it may not be able to handle complex, nonlinear patterns of fraud.

K-Nearest Neighbors (KNN):

KNN relies on classifying a transaction according to its similarity to its nearest neighbors.[10]

KNN performs well when fraudulent transactions are grouped together in the feature space.

It is easy to implement but computationally expensive during the prediction phase.[9]

It is sensitive to the choice of the optimal value of K and the distance measure.

Decision Tree:

Decision Tree divides the transaction data based on the conditions of the features.[8]

Decision Tree is effective in handling nonlinear fraud patterns and feature interactions.

Decision Tree is easy to interpret and visualize for financial decision analysis.

Decision Tree has the potential to overfit if not pruned properly.

Random Forest:

Random Forest is a combination of several Decision Trees.[10]

It handles the problem of overfitting by taking the average of the predictions of many Decision Trees.

It is very accurate and robust in fraud analysis problems.

It performs well on imbalanced datasets when used along with resampling methods.

XGBoost :

XGBoost is a sophisticated boosting algorithm[16].

Improves weak models one by one to minimize errors in fraud prediction.

Very accurate and efficient for handling large financial fraud datasets.

Usually the best-performing algorithm for credit card fraud detection

VI. Performance Evaluation Metrics

Several standard performance metrics are employed to find the effectiveness of the proposed credit card fraud detection system. Since credit card fraud detection deals with highly imbalanced data, relying on accuracy may be misleading; therefore, multiple evaluation metrics have been considered to give a comprehensive view of model performance[14].

A. Accuracy

The accuracy measure mainly focuses on the overall accuracy of the classification model by calculating the ratio of correctly classified

transactions to the total number of transactions. However[10], the accuracy measure would not be sufficient for understanding the performance of the model for the fraud detection problem, as the frequency of fraudulent transactions is much lower compared to legitimate transactions.

Table I: Accuracy Comparison of Machine Learning Models

Algorithm	Accuracy (%)
Logistic Regression	94.88
Decision Tree	93.68
K-Nearest Neighbors (KNN)	97.72
Random Forest Classifier	98.12

B. Precision

Precision is the ratio of correctly classified fraudulent transactions to the overall number of transactions the model predicted as fraudulent. A high precision value means that the model generates fewer false positives, which again is very important not to flag genuine transactions and distress customers.[12].

C. Recall

Recall, also known as sensitivity, is defined as the proportion of actual fraudulent transactions that have been correctly identified by the model. Recall is regarded as an important measure in fraud detection systems because failing to identify fraudulent activity results in heavy losses[15].

D. F1-Score

The F1-score can be viewed as the harmonic mean of precision and recall[16]. It provides a regularized measure of the model's performance. It works well on problems related to imbalanced classification, since it considers both false positives and false negatives. A higher F1-score means a better tradeoff between precision and recall.

VII. Comparison with Existing Methods

Typically, traditional fraud detection systems based on credit card fraud employ rule-based and statistical approaches and focus on setting thresholds and defining rules based on expert opinions and knowledge[16]. Although the approaches are quite simple in nature and straightforward to execute, high false positive rates are generated by fraudsters due to dynamic patterns and methods used by fraudsters in fraud transactions as traditional fraud detection systems lack flexibility and dynamically update the rules on a frequent basis.

Though there are existing machine learning models like Logistic Regression, Decision Tree, and Support Vector[17] Machines, which have shown improved accuracy in fraud detection compared to traditional models, these models have been seen to perform poorly with imbalanced datasets and are also not effective in handling non-linear relationships. This is because the models are vulnerable to noise and overfitting

VIII. Conclusion

Thus, it can be concluded that the proposed system will provide an efficient solution for the detection of credit card fraud, which would not only reduce losses but also eliminate false alarms[14]. This article emphasizes the need for data-driven and dynamic approaches to machine learning for improving the security of financial transaction System

From the experiment performed, it is evident that the Random Forest classifier is better than other machine learning models, especially in terms of accuracy and detection capabilities[17]. It is apparent that employing multiple evaluation mechanisms, like precision, recall, and F1 score, is a suitable strategy for evaluating model competencies, particularly when dealing with class imbalance problems. Indeed, ensemble models are more consistent while detecting sophisticated fraudulent patterns.

IX. References

[1] Adversarial Drift Detection (IJCNN2014) https://ieeexplore.ieee.org/document/6889387

[2] Credit Card Fraud Detection Using Machine LearningTechniques(IJERT2020) https://www.ijert.org/credit-card-fraud- detectionusing-machine-learning-techniques

[3] Data Mining for Credit Card Fraud: A Comparative Study(Decision Support Systems) https://www.sciencedirect.com/science/article/pii /S 0167923610002604

[4] SCARFF: A Scalable Framework for Streaming CreditCardFraudDetection https://www.sciencedirect.com/science/article/pii/S 156625351730628X

[5] XGBoost: A Scalable Tree Boosting System https://dl.acm.org/doi/10.1145/2939672.2939785

[6] UCI Machine Learning Repository – Credit Card Dataset https://archive.ics.uci.edu/

[7] Kaggle – European Credit Card Fraud Dataset https://www.kaggle.com/mlg-ulb/creditcardfraud

[8] SMOTE: Synthetic Minority Over-sampling Technique https://jair.org/index.php/jair/article/view/10302

[9] **Chen & Guestrin (2016)** – Gradient Boosting for Fraud Detection Proceedings of ACM SIGKDD Conference https://dl.acm.org/doi/10.1145/2939672.2939785

[10] Anomaly Detection for CreditCard Fraud https://www.sciencedirect.com/science/article/pii/S1877050 91831097

[11] Handling Imbalanced Datasets in Machine Learning – ACM Computing Surveys https://dl.acm.org/doi/10.1145/3344996

[12] Random Forests – Leo Breiman (FoundationalPaper) https://www.stat.berkeley.edu/~breiman/rando mforest2001.pdf

[13] Logistic Regression in Credit Risk and Fraud Detection https://link.springer.com/chapter/10.1007/978-3-030-16841-4_6

[14]A Survey of Credit Card Fraud Detection Methods – IEEE Access https://ieeexplore.ieee.org/document/8254255

[15] Isolation Forest Algorithm for Anomaly Detection https://ieeexplore.ieee.org/document/4781136

[16] Credit Card Fraud Detection: A Realistic Modeling and New Public Dataset – Dal Pozzolo et al. https://ieeexplore.ieee.org/document/7475706

[17] Machine Learning for Credit Card Fraud Detection – A Comparative Study https://www.researchgate.net/publication/328078 670

[18] IEEE Xplore – Credit Card Fraud Detection Research Collection https://ieeexplore.ieee.org/search/searchresult.jsp ?queryText=credit%20card%20fraud%20detecti on