Development of an efficient machine learning algorithm for reliable credit card fraud identification and protection systems

Dr. K. Maithili¹*, Dr. T. Satish Kumar², Dr. R. Subha³, P. L. Srinivasa Murthy⁴, Dr. M. N. Sharath⁵, Kopparavuri Gurnadh Gupta⁶, and Praseeda Ravuri⁷, TNP madhuri⁸, Vikas Verma⁹

¹ Associate Professor, Department of CSE – AIML, KG Reddy College of Engineering & Technology, Moinablad, Hyderabad, Telangana-501504
² Associate Professor, Department of Computer Science and Engineering, Hyderabad Institute of Technology and Management, Hyderabad
³ Assistant Professor, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore 21
⁴ Professor, Department of Computer Science and Engineering, IARE
⁵ Associate Professor, Rajeev Institute of Technology, Hassan, Karnataka
⁶ Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur Dist., Andhra Pradesh - 522302, India
⁷ Computer Science Engineer, Oregon State University, Corvallis, Oregon, USA 97331
⁸Department of IT, GRIET, Hyderabad, Telangana, India
⁹Lovely Professional University, Phagwara, Punjab, India.

Abstract. Recent developments in e-commerce and e-payment systems have led to a rise in financial fraud incidents, particularly credit card fraud. Software tools to identify credit card theft are essential. Critical characteristics of credit card fraud are crucial in utilizing Machine Learning (ML) for credit card fraud identification and must be selected carefully. This study suggests a An Efficient Machine Learning Algorithm for Reliable Credit Card Fraud Identification (EMLA-RCCFI) was constructed using ML, which utilizes the Genetic Algorithm (GA) to select features. Once the optimum characteristics are determined, the suggested detecting module utilizes the subsequent ML-based classifications. The proposed EMLA-RCCFI system is assessed using a dataset produced by European cardholders to confirm its efficacy. Based on the results, the suggested EMLA-RCCFI method surpassed existing systems regarding accuracy, precision, and F score.

1 Introduction to credit card frauds

Within the fourth industrial revolution, the e-commerce network has evolved into financial enterprises’ most comprehensive management system [1]. People tend to choose services offered via e-commerce and the Internet to improve their productivity and reduce jobs that
need a lot of time. The use of the Internet for financial transactions is gaining popularity. As a result, schemes have been considerably rising across a variety of businesses all over the globe, notably in financial services. Regarding financial institutions, credit card fraud is regarded as the most troublesome kind of fraud, and there is an urgent need to create solutions that can detect it as quickly as feasible [2]. Considerable growth has occurred in the overall number of Credit Card Frauds, which has climbed from around $17.5 billion in 2018 to roughly $42.7 billion in 2028 worldwide. A forecast is that credit card fraud will suffer a loss of $37.1 billion in 2023, equivalent to 8.2 cents for every $100 of the total volume. Examining various fraud detection systems is necessary to manage credit card fraud's repercussions rigorously. This will allow for a significant reduction in the adverse effects. Credit card fraud detection systems are often educated using prior transactions to identify whether transactions will occur in the future.

The study of Artificial Intelligence (AI) is divided into many subfields, one of which is Machine Learning (ML), which enables computers to gain insight from their past experiences and increase their ability to anticipate outcomes without having been tailored to do so [3]. The purpose of this study is to build ML algorithms for the identification of fraudulent credit card activity. In financial transactions, "credit card fraud" refers to any forged transaction (payments) by an unauthorized person utilizing a credit or debit card. The currently available ML models for identifying credit card fraud have poor detection accuracy. They cannot overcome the problem of the severely skewed nature of the datasets used for credit card fraud. For this reason, it is of the utmost importance to design ML models that can perform at their highest possible level and identify instances of credit card fraud with high accuracy.

A credit card fraud dataset that was developed from European credit cardholders is employed in this study. It is very uncommon for these datasets to have a multitude of variables that have the potential to adversely affect the efficacy of the classification algorithms while they are participating in the training phase. The research developed a feature selection based on the Genetic Algorithm (GA). It uses the Random Forest (RF) approach in its fitness function to handle the problem of a high-featured dimension field. The RF technique can take a high number of input elements. It can immediately accommodate missing values and is not influenced by noisy data, which is why it is employed in the GA fitness algorithm.

The rest of the sections are listed below: section 2 overviews the literature survey on Credit Card Fraud identification. Section 3 proposed a GA-based Credit Card Fraud identification and protection model named Efficient Machine Learning Algorithm for Reliable Credit Card Fraud Identification (EMLA-RCCFI). Section 4 analyses the different ML models for credit card fraud identification, and the results are compared with the EMLA-RCCFI models in terms of accuracy, precision, recall, and sensitivity. The conclusion and future scope are shown in Section 5.

2 Literature survey and analysis

When it comes to data processing and mining in most real-world study fields, particularly online banking operations, the unbalanced data distribution provides several problem opportunities. The fraudulent use of credit cards has caused a significant loss for users as well as for financial institutions all around the globe. As a result, experts are looking for technologies tailored to identify and prevent fraud of this kind.

The unique strategy was extreme outlier removal, employing k Reverse Nearest Neighbors (k-RNNs) with mixed sampling techniques [4]. This method aimed to develop trustworthy anticipations for situations that were not only fraudulent but also for situations that were not fraudulent.
The study presents a complete analysis of numerous ensemble classifications for credit card fraud identification [5]. This analysis was carried out by regression and voting. After that, these classifications were compared to several successful single classifications, such as KNN, Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), and Multilayer Perceptron (MLP). The performance of these methods was examined using three distinct datasets that were processed using SMOTE.

The SMOTE-Edited Nearest Neighbor (E-NN) methodology was the best at identifying credit card fraud compared to other distinct classifications amongst a set of oversampling methods [6]. The SMOTE-Tomek's Links (TL) approach demonstrated excellent results compared to the under-sampling methods.

An identifiable database of credit card activities was used to facilitate the integrated probabilistic and neuro-adaptive technique [7]. This particular combo exemplified a high level of fraudulent activity. A Hidden Markov Model (HMM) was applied to develop a model of the processes and sequences involved in handling credit card transactions [8]. The study also demonstrated how the framework was utilized to identify fraudulent transactions. The outcomes of the experiments revealed that the system based on HMM is both successful and beneficial in understanding the spending profile of the cardholders they are associated with. Anomalies in Fintech systems were identified by deploying ML techniques [9]. It was determined to focus on the suspicious behavior in the financial database, and algorithms were developed to forecast future thefts. The strategies now accessible for identifying anomalies were also discussed, and they showed satisfactory performance in credit card fraud identification.

An innovative hybrid strategy using Dynamic Weighted Entropy (DWE) was developed to address the issue of category imbalance with overlapping in credit card fraud identification [10]. This approach relied on the concept of divide and rule to tackle the problem. In this investigation, a novel measurement given the DWE was established to assist with choosing one of the model's hyper-parameters.

Several contemporary methods based on ML, data mining, sequence position, fuzzy computation, and genetic algorithms were investigated to determine how to identify fraudulent credit card activity [11]. However, the strategies discussed in this research have had varying degrees of success and failure, depending on the particular circumstances of each instance.

A study of unbalanced categorization systems was examined to analyze fraudulent credit-related actions [12]. Based on the findings of this research, it was determined that the use of unbalanced categorization methodologies could have been more efficient and adequate in addressing the issue of data imbalances.

3 Proposed GA-based credit card fraud identification

The system design offers a comprehensive design of the presented framework. The structure of the system that has been developed is explained. Figure 1 illustrates the architecture of the proposed framework. The design shows that the user must first register to use the platform. Then, they must log in by inputting their credentials, which include their username, password, email address, and telephone number. Following that, it presents the method selection openings, which are the places where the consumer must choose the method. Immediately after selecting the process, the data undergoes pre-processing, which includes cleaning the information and normalizing the database before it is fed into the actual model, followed by the information partitioning. At this point, the system is being trained with the help of the ML method, and it is eventually being tested and forecasted whether or not the purchase would be considered fraudulent.
Feature selection is an essential stage in adopting ML techniques. The dataset has a huge feature space, which might adversely affect the models' efficiency during development and validation. The feature selection approach relies on the specific topic that a researcher aims to address.

### 3.1 GA for selecting features

The GA is a kind of Evolutionary Algorithm (EA) frequently employed for optimizing jobs with lower processing requirements. GAs often have the following characteristics:

- Demographics GA keeps a sample of potential solutions known as a population.
- A person inside a population is a solution. A genetic profile and an indicator of physical capability define every person.
- A person undergoes evolution via mutations influenced by natural gene development.

The research utilizes the Random Forest technique as the fitness strategy inside the GA. The RF approach addresses the issue of over-fitting, which is often seen with standard DTRF. RF has high performance with continuous as well as elements and variables and is particularly effective on datasets with category imbalance issues. The RF algorithm operates based on rules, eliminating the need for data normalization. The fitness technique is a function that evaluates a candidate solution (a characteristic vector) to assess its suitability. Fitness level is defined by the correctness of a particular characteristic vector during the testing phase of the GA. Figure 2 illustrates the critical phases of the GA tailored to credit card fraud identification.
3.2 Fraud detection model

Figure 3 illustrates the framework of the suggested credit card fraud identification technique. The first step is utilizing the min-max scaling approach to normalize the training database in the normalizing inputs. Scaling is performed to guarantee that all input values fall inside a predetermined range. The GA is used in the feature-selecting phase, employing the normalized information from the phase of the normalizing input. During every round of the GA feature selecting a unit, the potential characteristic vector \( v_n \) is generated by the GA. This vector is subsequently utilized to train the algorithms in the training, which consists of the training information and the model phases. The identical vector evaluates the training models employing the testing information. Testing is carried out using the trained modeling block with the testing information. The testing method is carried out on every \( v_n \) in the selected model until the intended outcomes are achieved.

4 Simulation analysis and outcomes

The experiments were carried out using Google Colab. The computer characteristics are outlined below: Intel processor running 2.4GHz with two cores. This study utilizes the Scikit-Learn ML framework. This research includes samples categorized as true positive \((T_+}\), false positive \((F_-}\), true negative \((T_-}\), and false negative \((F_+}\). In this research, samples containing fraud are classified as positive, whereas samples without fraud are recognized as negative. The confusion matrix categorizes accuracy, precision, and fraud as part of the negative category.

Classifying accuracy is often used to evaluate the efficacy of categorization. It is the ratio of correctly categorized specimens to the total number of specimens, as shown in Equation (1).

\[
A = \frac{T_+ + T_-}{T_+ + T_- + F_- + F_+}
\]

The positive prediction value, or precision, is the proportion of positive specimens correctly identified out of the overall amount of specimens projected as positive, as seen in Equation (2).

\[
P = \frac{T_+}{T_+ + F_-}
\]

Sensitivity, also known as recall, is the proportion of positive specimens correctly identified out of the total positive and negative specimens that were expected, as mentioned in Equation (3).

\[
R = \frac{T_+}{T_+ + T_-}
\]

Specificity is the proportion of true negative specimens identified adequately out of the overall amount of negative predicted specimens, as seen in Equation (4).
\[ S = \frac{T_-}{T_- + T_+} \]  

(Equation 4)

**Fig. 4.** Accuracy evaluation of different credit card fraud-detecting models

Figure 4 displays the accuracy assessment of several credit card fraud detection methods. The EMLA-RCCFI model has superior accuracy to other models due to using a genetic algorithm for feature selection. The proposed method results are compared with Logical Regression (LR), DT, RF, SVM, KNN, Naïve Bayes (NB), Principal Component Analysis (PCA), and Long Short Term Memory (LSTM). The \( T_+ \) and \( T_- \) values acquired using the suggested technique are much greater than the \( F_+ \) and \( F_- \) values. The accuracy achieved is 78.28%, which surpasses that of other machine learning models. This enhances credit card fraud detection and boosts user safety.

**Fig. 5.** Precision evaluation of different credit card fraud-detecting models

Figure 5 displays a detailed assessment of several credit card fraud detection methods. The accuracy of the suggested EMLA-RCCFI model is calculated using Equation (2) and is determined to be 86.31%, far surpassing the precision of other models. The findings demonstrate the effectiveness of the GA model in identifying and categorizing credit card fraud.
Fig. 6. Recall the evaluation of different credit card fraud-detecting models

Recall is calculated for several machine learning models for credit card fraud detection according to Equation (3), and the collective outcomes of all approaches are shown in Figure 6. The EMLA-RCCFI model has an accuracy of 85.32%, surpassing previous machine learning models. Maximizing the outcome improves the efficacy of credit card fraud detection in terms of detection and security performance.

Fig. 7. Sensitivity evaluation of different credit card fraud-detecting models

Figure 7 displays the sensitivity assessment of several credit card fraud detection methods. The performance of the suggested EMLA-RCCFI model surpasses that of other models when compared. The EMLA-RCCFI has an accuracy rate of 87.32%, making it the most effective model for detecting and preventing credit card fraud.

5 Conclusion and findings

This study suggests an Efficient Machine Learning Algorithm for Reliable Credit Card Fraud Identification (EMLA-RCCFI) was constructed using ML, which utilizes the Genetic Algorithm to select features. Once the optimum characteristics are selected, the suggested detecting module utilizes the subsequent ML-based classifications. The Genetic Algorithm was used on the dataset of credit card transactions made by European customers, generating five ideal feature vectors. The testing findings showed that the EMLA-RCCFI obtained higher results utilizing the GA-chosen qualities. The outcomes from this investigation surpassed those of other ML approaches. The research used the EMLA-RCCFI framework to synthesize a credit card fraud database to confirm the findings obtained from the European...
credit card fraud database. The study's results demonstrated that the EMLA-RCCFI achieved an accuracy of 78.28%, precision of 86.31%, recall of 85.32%, and sensitivity of 87.32%. The system will use more databases in future projects to verify the EMLA-RCCFI approach.

References