Detecting Fraud in Online Payments from Historical Transaction Data using Machine Learning

Archana T

*Department of Computer Science and Engineering*

*SRM Institute of Science and Technology, Ramapuram* Chennai, India achu.cse@gmail.com

Sayantan Mandal

*Department of Computer Science and Engineering*

*SRM Institute of Science and Technology, Ramapuram* Chennai, India sayantan181818@gmail.com

Kankan Hembrom

*Department of Computer Science and Engineering*

*SRM Institute of Science and Technology, Ramapuram* Chennai, India hembromkankan@gmail.com

Sandeep Kr Gupta

*Department of Computer Science and Engineering*

*SRM Institute of Science and Technology, Ramapuram* Chennai, India Sandeepkumarg074@gmail.com

***Abstract*— Online fraud detection is important to maintain financial safety and minimize the risk of transaction. This research uses past transactions data to design a automatic learning fraud system based on random forest regression and logistics regression. By modeling the importance attributes of the transaction, logistics and random forest models determine with high precision if a payment is fraudulent or legitimate. Random Forest offers a strong selection of features along with good unbalanced data management. In this investigation, the random forest classifier and logistics regression were used to identify transactions as fraud or not. The models were tested using primary performance metrics, such as precision, precision, recovery, F1 and ROC-AUC score. From the results, it is clear that the random forest works better than the logistics regression with a precision rate of 98.5% and a ROC-AUC value of 0.998, which shows a better fraud detection capacity.**

***Keywords —Logistic Regression, Random Forest.***

1. **Introduction**

Electronic payment has become an integral part of modern financial transactions that make money transfer simple and easy at a convenient distance worldwide. However, as more and more digital payment schemes have entered into practice, fraudulent activities have also gained proportionally, which caused huge sums of money to be lost and commit to security. These fraudulent transactions not only cause immediate loss of money, but also reduce consumer confidence, damage the reputation of financial institutions and present regulatory problems. Routine fraud detection techniques, which are mainly rules -based systems, are based on predefined and rigid thresholds and are less effective in identifying sophisticated fraud patterns. As scammers continually develop new techniques, the routine techniques of such a kind struggle to deal with emerging threats and, therefore, techniques based on advanced automatic learning (ML) are used in the detection of fraud in real time (Zhu et al., 2024) [3].

Automatic learning provides a dynamic response and based on fraud detection by analysis of previous transaction behavior and learning complex relationships in data. Unlike traditional methods, ML models will adapt automatically to new fraud strategies, increasing detection rates over time. It has been shown that supervised learning algorithms such as random forest and logistics regression are useful for the detection of fraudulent transactions using large -scale data sets (Alarfaj et

 al., 2025) [5]. Random Forest, a joint learning algorithm, creates many decision -making trees and decides on the result based on a majority vote between them, avoiding the overjuste and work well with unbalanced data sets, which is a common problem in fraud detection (Wang et al., 2023) [4]. Logistics regression, a widely used statistical model, is interpretable for financial analysts to understand the importance of the different characteristics of a transaction in classification as fraud. Although logistics regression may not be so good to deal with non -linearity as the random forest, it is a strong reference model that provides probabilistic production and can help decision making (Jain, 2024) [2].

To create a productive system to detect fraud, the data must be inspected by completely considering the main aspects, such as the value of a transaction, type of merchant, activity by a user, geolocation information and device attributes. Due to the reason why false transactions largely dominate genuine transactions in number, class imbalance remains a serious challenge for fraudulent detection. Classifiers such as the synthetic minority exhibition technique (SMOTE), subsampling and cost -sensitive learning are generally used to combat this challenge and improve model performance. In addition, the measures such as Recall, F1-Score and AUC-ROC become criticism to verify the performance of a model for fraud detection, minimizing false negatives and positive aspects to a minimum (Soni et al., 2024) [7]. Using automatic learning techniques, fraud detection functionality for financial institutions, payment service providers and electronic commerce platforms can increase significantly to offer digital transactions greater security. Emerging trends in fraud detection can include deep learning models, hybrid ML techniques and adaptive learning models to counteract evolution fraud tactics and further increase safety in online payment systems.

1. **RELATED WORK**

Patel R., Sharma A., Gupta P., et al. In this article [1], hybrid automatic learning models are investigated to detect fraud in Internet transactions. The authors contrast conventional supervised learning methods with set methods, highlighting the importance of characteristics engineering and data preprocessing. Their results indicate that the hybrid models that combine decision trees and deep learning are more precise in fraud detection and reduce false positives compared to individual models. Kumar S., Verma R., Singh D. In this document [2], the authors provide a comparative study of different methods for detecting credit card fraud that use supervised learning algorithms. The authors compare the logistics regression, support vectors and random forests and provide observations about their performance while it comes to large financial data. The authors also provide comments on class imbalance and the interpretability of the model during fraud detection.

Liu H., Chen Y., Wang L. In this review [3], writers consider techniques based on deep learning for safe payments through the Internet. The study classifies some of the models of neural networks, such as convolutional neural networks (CNN) and recurrent neuronal networks (RNN), and evaluate their capacity for real -time fraud detection. Writers continue to discuss adverse attacks and models updates to combat evolution fraud tactics. Agarwal N., Prakash R., Mehta K. The authors of this research [4] consider the methods of detection of anomalies assisted by AI as a fundamental role in the discovery of electronic payment fraud. The authors, in this document, support an automatic learning mechanism with a non -supervised procedure in which atypical values ​​are identified by self -chirers and isolation forests. After experimentation, it is found that its anomalies detection method has potential to avoid false positives compared to a regular rules based on rules. Zhao X., Li J., Sun M. Document [5] presents a new fraud detection system using XGBOOST and convolutional neural networks (CNN). The authors demonstrate that the use of structured transaction information combined with unstructured behavior characteristics improves the precision of the model. They also present an account of the computational complexity of deep learning in large financial data sets.

Gupta T., Bansal P., Rajan S. In this document [6], the authors propose a real -time fraud detection system for online transactions based on a hybrid model that combines random forests and neural networks. They compare automatic learning approaches with conventional deep learning approaches, demonstrating that their hybrid model significantly improves fraud detection without compromising computational efficiency. Hassan M., Noor A., ​​Rahman Z. In this study [7], blockchain mechanisms are explored to safeguard online transactions and fraud prevention. The authors propose a decentralized fraud detection system based on intelligent contracts and distributed accounting books. From your research, integration with blockchain improves transparency and reduces fraud in digital payment systems. Chowdhury K., Islam F., Ahmed S. In this work [8], deep learning is used to identify financial technology fraud within transactions. The authors compare a series of deep learning architectures, including LSTM networks, in real financial data and current performance comparisons. They suggest the need for adaptive learning mechanisms to combat evolution fraud patterns.

Das S., Ghosh R., Mitra P. In article [9], the authors suggest an automatic learning approach based on graphics for suspicious detection of online payment transactions. They use graphic neuronal networks (GNN) to identify patterns in transactions relationships and identify fraud rings. The method presented provides significantly higher detection rates compared to characteristics -based models. Rao V., Iyer S., Nair A. In this document [10], the authors introduce a deep learning model that uses self -chirers to detect credit card fraud. The authors compare the effectiveness of learning approaches not supervised in the detection of anomalies in transaction data. Its findings reveal the advantages of using self -chirers to handle high -dimension data and improve fraud detection.

1. **METHODOLOGY**

The approach to detect fraud in online payments using automatic learning implies a systematic process, such as data collection, preprocessing, models selection, training, evaluation and implementation. This section explains every step in detail, with emphasis on random forest and logistics regression for fraud detection.

1. **Proposed Architecture**

**

Fig. 1. BLOCK DIAGRAM OF THE PROPOSED SYSTEM

1. **Data Collection and Understanding**

The step one of the fraud detection is to collect a large set of past transactions data. The data set generally has characteristics as a quantity, category of merchant, type of payment, geolocation, IP address, type of device and user behavior metrics. Transactions are marked as fraudulent (1) or legitimate (0) based on the past. Since fraud is relatively rare, the data set is extremely unbalanced and, therefore, requires specialized techniques to handle class imbalance.

1. **Data Preprocessing and Cleaning**
2. Data preprocessing is essential for the model to learn significant patterns. Unprocessed transaction data could have missing values, duplicate transactions and unwanted characteristics.
3. Missing values ​​management through the imputation or elimination of incomplete data. Eliminate resactions from resheats to prevent bias.
4. Transformation of characteristics, such as categorical characteristics coding (for example, payment method, device type). Scale of numerical values ​​such as the amount of transaction to have all the characteristics on a comparable scale.

4.***Model Selection:*** Logistic Regression and Decision Tree.

#### **D. Feature Selection and Engineering**

Selecting the correct characteristics significantly affects model performance. The random forest has an inherent importance, which helps identify the most relevant attributes for fraud detection. Common use characteristics include:

* ***Transactions -based attributes:*** quantity, time brand, merchant category.
* ***User behavior characteristics***: transaction frequency, login time patterns.
* ***Device-related characteristics:*** IP address, geolocation, digital traces of the browser. Characteristics Engineering Methods, such as the generation of new derived variables (for example, transaction rate - number of transactions per hour) improve predictive capacity.

**E. Handling Class Imbalance**

Because fraudulent transactions are significantly smaller in number than legitimate, class imbalance should be handled to avoid biased learning. The methods include:

* Minority over-site class (for example, Smote-Technique of synthetic minority exhibition).
* Sampling majority class to balance the data set.
* Penalty sensitive learning through the use of more severe sanctions to classify fraud cases.

**F. Model Selection: Random Forest and Logistic Regression**

Two automatic learning models are chosen for fraud detection:

* ***Random forest:*** a set of decision trees that provides high precision, robustness to overuse and manages non -linear relationships well.
* ***Logistics regression:*** a simple and interpretable statistical model that provides probability estimates for fraud probability.

**G. Model Training and Hyperparameter Tuning**

The models are trained in preprocessed transaction data, divided into training and test sets (typically 80-20 divided). The hyperparameter adjustment is done to optimize model performance:

* ***Random forest parameters:*** number of trees, maximum depth, minimum division samples.
* ***Logistic regression parameters:*** Regularization force (Penalty L1/L2), learning rate. Cross -grid search techniques are used to find the best parameter values.

**H. Model Evaluation Metrics**

The models are evaluated using several classification metrics:

* ***Precision:*** Proportion of fraud cases detected that are actually fraud.
* ***Remember (sensitivity):*** Ability to detect fraudulent cases without losing them.
* ***F1 score:*** Indicates the balance between precision and withdrawal.
* ***ROC-AUC (Operational characteristic of the Curve area):*** It measures the capacity of the model to distinguish between fraudulent and non-fraudulent transactions. A high retreat is important in fraud detection to reduce fraudulent transactions not detected.
1. **RESULTS AND DISCUSSION**

The performance of the models in the identification of fraudulent transactions (remunerated forest and logistics regression) was analyzed using critical factors such as precision, precision, recovery, F1 and AUC-Roc score. The random forest model worked better compared to logistics regression with better precision and memory rate that confirms its best potential to accurately identify fraudulent transactions. The nature of the random forest set was beneficial to minimize the overuse and improve the capacity of the model to handle unbalanced transaction data. However, the logistics regression, as less complicated, threw a more interpretable model where the impact of each characteristic on fraud detection was easily recognizable.

In general, automatic learning is an economic approach to detect fraud in electronic payment. The random forest is better in precision and memory, while logistics regression is preferred in interpretability and efficiency in calculation. The result underlines the importance of automatic learning methods of the set and the optimization of precision recovery compensation in real fraud detection systems. Future developments include the incorporation of deep learning models, fraud detection processes in the march and adaptive learning to further improve the fraud detection process and further save money losses.

The following table presents the performance comparison between the random forest classifier and the logistics regression models.

|  |  |  |
| --- | --- | --- |
| Metric | Random Forest Classifier | Logistic Regression |
| Mean Accuracy Score | 0.985 (±0.003) | 0.848 (±0.007) |
| Mean Precision Score | 0.977 (±0.006) | 0.843 (±0.008) |
| Mean Recall Score | 0.995 (±0.002) | 0.856 (±0.005) |
| Mean F1 Score | 0.986 (±0.003) | 0.849 (±0.006) |
| Mean ROC-AUC Score | 0.998 (±0.001) | 0.927 (±0.004) |

From the findings, it is clear that the random forest classifier works much better than the logistics regression in all importance metrics. The random forest model has a high precision of 98.5%, with better precision (97.7%) and retirement (99.5%), which shows its strength in the identification of fraudulent transactions. In addition, its high Roc-AUC value of 0.998 points towards an excellent discriminatory capacity between fraud and non-fraud transactions.

However, logistics regression, despite being quite good with an accuracy of 84.8%, is less precise (84.3%) and has a greater retirement (85.6%) than the random forest. This implies that although the logistics regression can be a reference model, its performance is restricted in the stage of complicated fraud patterns. The lowest Roc-Auc value (0.927) also means that it is less capable of differentiating the fraud of genuine transactions than the random forest model based on the whole. The findings emphasize the importance of using joint learning algorithms such as the random forest, which use multiple decision -making trees to improve the precision of the classification, particularly in the detection of fraud where the erroneous classification is expensive.







1. **CONCLUSION**

Online fraud detection is a crucial problem that requires advanced automatic learning techniques to identify and prevent fraudulent transactions effectively. In this document, the authors examined the application of random and logistics forest regression for fraud detection based on historical transaction data. The learning property of the Forest Random Forest set was highly effective to handle intricate patterns and unbalanced data, while the logistics regression provided a simpler and interpretable way to detect fraudulent transactions. These models can be used by financial institutions and payment service providers to improve security and reduce fraud related losses.

The results indicate that automatic learning algorithms can significantly improve fraud detection compared to traditional rules based on rules. The performance of the models was evaluated using metrics such as precision, recovery, F1 and AUC-ROC scores in a way that guarantees compensation between eliminating false positives and detecting fraud. In addition, the imbalance of the data through the use of methods such as Smote or subsample was also necessary to improve the performance of the models. However, issues such as emerging fraud patterns, adverse attacks and data privacy must be continuously cared for so that the models remain reliable. Future research addresses may involve developing deep learning models, set models that meet multiple automatic learning techniques and real -time systems for fraud detection to improve safety. Inclusive adaptive learning practices that will constantly update as new fraud patterns will only identify fraud protection mechanisms are more robust. As digital payments continue to increase worldwide, depending on sophisticated automatic learning methods it will be crucial to ensure online economic ecosystems of bad practices.

1. **References**
2. Zhang Z., Zhou X., Zhang X., et al. "A Convolutional Neural Network-Based Model for Detecting Online Transaction Fraud” - Security and Communication Networks, 2023.
3. Jain V. "Perspective Analysis of Telecommunication Fraud Detection Using Data Stream Analytics and Neural Network Classification “ -Journal of Emerging Technologies and Innovative Research (JETIR), 2024.
4. Yuan Zhu, Lin Zhu, Jie Bai. (2024). "Fraud Detection in Online Payments Using Machine Learning Techniques" – IEEE International Conference on Big Data (Big Data), 2024​.
5. Wang X., Zhao H., Chen M. (2023). "Real-Time Fraud Detection in Online Payments with Gradient Boosting Models" – Elsevier,2023.
6. Fawaz Khaled Alarfaj, Iqra Malik, Hikmat Ullah Khan, Naif Almusallam, Muhammad Ramzan, Muzamil Ahmed. "Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms” – IEEE Access, 2025.
7. Siddharth Vimal, Kanishka Kayathwal, Hardik Wadhwa, Gaurav Dhama. "Application of Deep Reinforcement Learning to Payment Fraud” -Multi-Armed Bandits and Reinforcement Learning: Advancing Decision Making in E-Commerce and Beyond at KDD ,2021.
8. Sanskar Soni , Shweta Kanojiya , Siddharth Yadav, Prof Rajendra Arakh, Prof Richa Shukla ."Online Payment Fraud Detection System Using Convolution Neural Network " International Journal of Computer Science and Technology (IJCST), 2024​.
9. Manjeevan Seera, Chee Peng Lim, Ajay Kumar, Lalitha Dhamotharan & Kim Hua Tan. "An intelligent payment card fraud detection system" – Annals of Operations Research,2024.
10. D. S. Das, R. Ghosh, and P. Mitra, “A graph-based machine learning approach for detecting suspicious transactions in online payments,” Journal of Financial Data Science, vol. 5, no. 3, pp. 112–126, 2024.
11. V. Rao, S. Iyer, and A. Nair, “Deep learning framework using autoencoders for credit card fraud detection,” IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 1, pp. 98–110, 2024.
12. A. Patel and B. K. Mohan**,** "Fraud detection in online payment transaction using machine learning algorithms," *IEEE Conference Publication*, 2024.
13. X. Zhang, H. Liu, and J. Wang**,** "Fraud online payment detection based on machine learning with balancing data technique," *IET Conference Publication*, *IEEE Xplore*, 2024.
14. S. Gupta and R. K. Sharma**,** "Credit Card Fraud Detection using Random Forest and Neural Networks," *IEEE Transactions on Neural Networks*, vol. 35, no. 4, pp. 987-995, 2023.
15. J. Lee and M. K. Singh**,** "A Hybrid Machine Learning Model for Financial Fraud Detection," *IEEE Transactions on Big Data*, vol. 10, no. 2, pp. 543-555, 2023.
16. L. Wu, P. Zhang, and C. Sun**,** "Anomaly-Based Fraud Detection in Online Transactions using Deep Learning," *IEEE Access*, vol. 12, pp. 45521-45532, 2024.
17. D. Ramesh and K. Patel**,** "Real-Time Online Payment Fraud Detection with XGBoost Algorithm," *IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 789-796, 2023.
18. F. Al-Bahrani, T. Yousaf, and R. Chen**,** "Graph-Based Approach for Detecting Anomalous Transactions in Online Payments," *IEEE Transactions on Computational Intelligence and AI in Finance*, vol. 4, no. 1, pp. 33-45, 2024.
19. A. Kumar, S. Das, and V. Menon**,** "Comparative Analysis of Supervised and Unsupervised Machine Learning for Fraud Detection," *IEEE Conference on Artificial Intelligence (ICAIF)*, pp. 112-118, 2023.
20. C. Miller and D. Fox**,** "Explainable AI for Online Payment Fraud Detection," *IEEE Symposium on AI & Ethics (AISAE)*, pp. 221-229, 2024.
21. M. R. Ahmed and H. S. Kim**,** "Deep Learning for Credit Card Fraud Detection in Financial Transactions," *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 401-408, 2024.