IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, October 2026
Impact Factor: 8.2

Real-Time Credit Card Fraud Detection Using Machine Learning, Multi-Agent Systems and Generative AI: A Research Review

Prof. Dr. Santosh Baraku Pawar, Prof. Punam Popat Deshmukh, Rajendra Rohidas Hadawale
Vidya Niketan Institute of Management & Computer Application, Bota

Abstract: *Credit card payments are now part of everyday life, and so is credit card fraud. Banks must decide within a fraction of a second whether a payment is genuine, and a wrong decision either lets a thief through or blocks an honest customer. This paper reviews how machine learning, multi-agent systems and generative AI are used to detect credit card fraud in real time. It explains the basic ideas in simple terms, compares the common techniques, discusses the practical challenges such as rare fraud cases, changing fraud behavior and the need for explainable decisions, and points out future research directions including federated learning, graph-based models and large language models that explain alerts to human analysts.*

Keywords: Credit Card Fraud, Machine Learning, Random Forest, Multi-Agent System, Explainable AI, Generative AI, Real-Time Detection

I. INTRODUCTION

Card payments have grown quickly with online shopping, mobile wallets and contactless terminals. The same growth has given criminals more chances to misuse stolen card details. Fraud is a small part of all transactions, but the money lost and the damage to customer trust are large. Because a payment must be approved or declined while the customer is still standing at the counter or waiting on the checkout page, detection has to be both accurate and fast [6][5].

Early systems relied on rules written by experts, for example "decline if the amount is more than ten times the customer's average". Rules are easy to understand but fraudsters learn them and work around them. Machine learning models learn patterns from past transactions and can notice combinations that no person would write as a rule [2][1]. More recently, two ideas have joined this field: multi-agent design, where several small specialized programs look at the same payment from different angles [9], and generative AI, where a language model turns the technical output into a short explanation that a human analyst can read [12].

This paper reviews these ideas and the challenges around them. It is meant as a clear overview for a student or a beginner in the field, not as an exhaustive survey.

II. KEY CONCEPTS

Credit Card Fraud and Its Common Types

Fraud is usually grouped by how the card is misused. The patterns below are the ones most often described in the literature [5][6]:

Card-not-present fraud: stolen card details are used for online purchases, often for goods that are easy to resell such as electronics or gift cards.

Card testing: the thief makes many tiny payments in a short time to check whether the stolen card still works.

Location anomalies: the card is used in a city or country far away from where the owner normally shops.

Account takeover: the thief gets into a real account, so the behavior looks close to normal and is hard to catch.

Machine Learning for Fraud Detection

Most fraud systems use supervised learning: the model is trained on past transactions that are labelled as genuine or fraud, and then predicts a fraud probability for a new payment. Common choices are logistic regression, decision trees, support vector machines, neural networks and ensembles of trees. Comparative studies often find that ensembles such as the random forest [3] give strong results because they combine many decision trees and are not easily fooled by noise [2][7]. Unsupervised methods such as the isolation forest [14] do not need labels and look for payments that are very different from the rest. Carcillo et al. show that combining supervised and unsupervised outputs can improve detection [8].

Rare Fraud and How Models Are Measured

Fraud is rare, often well below five percent of all payments. A model that says "genuine" every time would be 95 percent accurate and completely useless, so accuracy alone is a poor measure. Researchers prefer precision (how many alerts were really fraud), recall (how many frauds were caught), the F1-score and the area under the ROC curve [1]. To help models learn from so few examples, techniques such as SMOTE create extra synthetic fraud examples [4], and class weights can make a mistake on a fraud case count more than a mistake on a genuine case.

Multi-Agent Systems

A multi-agent system is a group of independent programs, called agents, each with its own job, that work together to solve a problem [9][10]. For fraud detection this means one agent can apply fixed rules, another can compare the payment with the customer's own history, and another can run a machine learning model. A coordinator then combines their opinions. The benefit is that each agent is simple, easy to test and easy to replace, and the final decision is easier to explain than one large black-box model.

Generative AI and Large Language Models

Large language models such as GPT-3 and LLaMA are trained on huge amounts of text and can write fluent answers to instructions [12][13]. They are not good fraud classifiers on their own, but they are very good at explaining. Given the facts about an alert (amount, city, time, which agent raised the score), a language model can write two or three plain sentences for an analyst and suggest the next step. If the prompt tells the model to use only the supplied facts, the risk of invented details is reduced.

III. CHALLENGES AND LIMITATIONS

Class Imbalance

With very few fraud cases, models tend to favour the majority class. Sampling methods and class weights help, but they must be tested carefully because they can also make the model too eager to raise alarms [4][1].

Changing Fraud Behavior

Fraudsters change their methods when they are caught. A model trained on last year's data slowly becomes less accurate, an effect known as concept drift. Systems need regular retraining and monitoring [1][6].

False Alarms

Every wrongly blocked payment annoys an honest customer. Many banks therefore use a middle option, sending doubtful payments for manual review instead of blocking them straight away.

Speed

A decision must be returned in milliseconds. Heavy models, slow databases or repeated work make this difficult, which is why caching and efficient data pipelines matter as much as the model.

Explainability and Trust

A bank must be able to tell a customer, an auditor or a regulator why a payment was stopped. Methods such as SHAP assign a share of the decision to each input feature [11], and simpler what-if comparisons can do something similar.

Privacy and Data Availability

Real card data is highly sensitive and banks cannot share it freely because of data protection rules. This makes public datasets rare and pushes many researchers to use simulated data, which is easier to share but may not capture every real-world behaviour [1].

IV. FUTURE TRENDS

Federated Learning

In federated learning the model is trained at each bank on its own data and only the learned parameters are shared, not the transactions. This allows banks to learn from each other without exposing customer data [15].

Graph-Based Detection

Fraud often involves networks of cards, merchants and devices. Graph neural networks [16] can look at these links and find rings of connected accounts that single-transaction models miss.

Explainable and Responsible AI

Regulators increasingly ask for clear reasons behind automated decisions. Systems that show which factors raised the risk, and that keep a human in the loop for doubtful cases, are likely to become standard.

Language Models as Analyst Assistants

Language models can summarise alerts, answer questions about dashboards and help investigators work faster. Open models such as LLaMA [13], served through fast hosted APIs, make this affordable even for small teams.

Real-Time Streaming Architectures

Payments naturally arrive as a stream. Web technologies such as Server-Sent Events and message queues let a dashboard show scored payments live, while caches keep repeated lookups fast.

V. CONCLUSION

Credit card fraud detection has moved from fixed rules to learning systems, and is now moving towards teams of cooperating agents with a language model that explains the result. Each approach has strengths: rules are transparent, behaviour checks are personal, machine learning is accurate on complex patterns, and generative AI makes the output understandable. Challenges such as rare fraud, drifting behaviour, false alarms and privacy remain open. Future work on federated learning, graph models and explainable AI will shape the next generation of fraud detection systems.

REFERENCES

1. Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., &Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems, 29(8), 3784-3797.
2. Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision Support Systems, 50(3), 602-613.
3. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
4. Chawla, N. V., Bowyer, K. W., Hall, L. O., &Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.
5. Phua, C., Lee, V., Smith, K., &Gayler, R. (2010). A comprehensive survey of data miningbased fraud detection research. arXiv:1009.6119.
6. Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. Journal of Network and Computer Applications, 68, 90-113.
7. Awoyemi, J. O., Adetunmbi, A. O., &Oluwadare, S. A. (2017). Credit card fraud detection using machine learning techniques: A comparative analysis. 2017 International Conference on Computing Networking and Informatics (ICCNI), IEEE.

8. Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oble, F., &Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. Information Sciences, 557, 317-331.
9. Wooldridge, M. (2009). An Introduction to MultiAgent Systems (2nd ed.). John Wiley & Sons.
10. Russell, S., &Norvig, P. (2020). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.
11. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 30, 4765-4774.
12. Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems 33, 1877-1901.
13. Touvron, H., Lavril, T., Izacard, G., et al. (2023). LLaMA: Open and efficient foundation language models. arXiv:2302.13971.
14. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. 2008 Eighth IEEE International Conference on Data Mining, 413-422.
15. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), Article 12.
16. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations (ICLR).