A Systems Engineering Approach to IEEE-CIS Fraud Detection

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Leonardo Rodríguez Salas
Davidson Esfleider Sánchez Gordillo
Santiago Marín Paez
Luis Mario Ramírez

Eng. Carlos Andrés Sierra Virguez
Systems Analysis & Design

Diciembre 2025

Abstract

This report provides a detailed analysis of a fraud detection system in online transactions, integrating approaches from *machine learning (ML)* and *event-based simulations* such as *cellular automata*. Through an approach incorporating *chaos theory* and *sensitivity analysis*, the system's responses to data perturbations were explored. Using 590,540 real transactions, a Random Forest model was trained, achieving a *ROC-AUC* of 0.9244 and *accuracy* of 97.03%. The sensitivity analysis revealed that the system is highly sensitive to small variations in data, with performance degradation occurring when noise is injected into features, especially when the noise amplitude exceeds 0.10. Cellular automata revealed emergent fraud patterns, providing spatial insights into how fraud propagates through transactions. This analysis highlights the importance of robust and adaptive systems that can handle model sensitivity to perturbations and continuously monitor changes in fraud patterns.

Contents

1 Introduction

Fraud detection in online transactions is a critical challenge for e-commerce platforms, as fraud represents only a small fraction of all transactions. This creates a class imbalance, complicating the precise identification of fraud. This report describes a fraud detection system that uses Random Forest to classify transactions as fraudulent or legitimate and employs event-based simulations using cellular automata to model how fraud propagates through the transaction space. Additionally, sensitivity analysis and chaos theory are applied to assess how small variations in the data can affect model performance and lead to unpredictable deviations.

The goal is to provide a deep understanding of how machine learning models can be sensitive to initial conditions and how event-based simulations can help model the emergent behavior of fraud, providing a more robust and adaptive system.

2 Literature Review

In the literature on fraud detection, common approaches include the use of machine learning algorithms such as Random Forest, SVM, and LightGBM, which have proven effective in identifying fraud patterns in large datasets. However, these approaches face several challenges, such as class imbalance, high dimensionality of the data, and non-linear relationships between transaction features.

Furthermore, chaos theory has been applied to study complex, non-linear systems, such as fraud detection, where small changes in data can lead to large differences in prediction outcomes. In this context, the Random Forest model was chosen for its ability to handle noisy, high-dimensional data, while cellular automata provide a spatial simulation of fraud, allowing us to observe how fraudulent transactions group and propagate through the system.

3 Background

Fraud detection faces multiple challenges, including class imbalance (since only a small fraction of transactions is fraudulent), inconsistent identity data, and the variability of fraud patterns. To address these issues, the system described in this report follows a modular approach, with modules for data ingestion, preprocessing, modeling, evaluation, and real-time monitoring. The integration of chaos theory and sensitivity analysis enables the detection of drastic changes in model performance due to small data perturbations.

Additionally, the use of event-based simulations, such as cellular automata, helps model the spatial behavior of fraud, providing insights into high-probability fraud areas and understanding how fraud propagates through the system.

4 Objectives

The main objectives of this report are to:

- Describe the components of the fraud detection system, including preprocessing, modeling, and evaluation modules.
- Evaluate the system's sensitivity to variations in input data, particularly how small changes affect model performance.
- Explore the emergent behavior of fraud through cellular automata simulations, to observe how fraud clusters and propagates in the transaction space.
- Propose improvements to make the system more robust and adaptable to new fraud tactics and changing conditions.

5 Scope

This report focuses on fraud detection using two complementary approaches:

- Machine Learning Model (Random Forest): A Random Forest model is used to classify transactions as fraudulent or legitimate.
- Event-Based Simulation (Cellular Automata): Cellular automata model fraud propagation within the transaction space, with the goal of capturing emergent patterns and spatial behaviors of fraud.

The report does not address issues related to cybersecurity or real-time infrastructure optimization.

6 Assumptions

During this analysis, the following assumptions were made:

- The data used is representative of real online transactions and is sufficiently clean for model construction.
- The Random Forest model is capable of handling class imbalance through the use of class weights.
- The system is capable of adapting to new fraud tactics through periodic retraining and continuous monitoring.
- The perturbations applied in the sensitivity analysis reflect potential real-world fluctuations in transaction data.

7 Limitations

The analysis presents the following limitations that should be considered:

- Incomplete identity records: A significant portion of the transactions had incomplete or missing identity data, which limits the model's ability to detect fraud based on user identity.
- Model sensitivity: The system is highly sensitive to small perturbations in the data, which can lead to significant variations in the model's predictions. This sensitivity can affect the stability of the system in real-world applications.
- Lack of real-time security infrastructure: The study does not address the impact of cybersecurity threats or vulnerabilities outside the context of transaction data processing.
- Data representativeness: The data used may not fully reflect the diversity of fraudulent activities in various industries, potentially limiting the generalizability of the model.

8 Methodology

8.1 Data Preprocessing

Data preprocessing was crucial to ensure the features were properly prepared for modeling. Missing values were handled (imputation by median) and numerical features were normalized. Categorical variables were encoded using one-hot encoding.

8.2 Modeling and Evaluation

The Random Forest model was trained with 1000 trees (to reduce variance) and a max_depth of 30 to avoid overfitting. Class weights were used to compensate for class imbalance, and cross-validation was employed to ensure the model was evaluated fairly and without bias.

8.3 Cellular Automata Simulation

A cellular automaton was implemented to model fraud propagation in the feature space. Each transaction was represented as a cell, which could be in one of four states: normal, suspicious, fraudulent, or flagged. Transition rules were based on the number of fraudulent neighbors, and a random mutation rate was applied to simulate the appearance of new fraud cases.

Methodology Diagrams

Data Flow Diagram for Fraud Detection System: This diagram illustrates the process flow from the initial data ingestion to preprocessing, model training, evaluation, and real-time monitoring. It provides a high-level overview of how data is processed and how the system responds to new transactions.

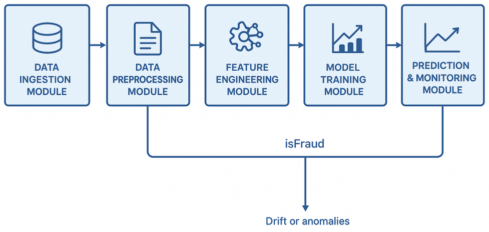
DATA INGESTION MODULE
DATA PREPROCESSING MODULE
FEATURE ENGINEERING MODULE
MODEL TRAINING MODULE
PREDICTION & MONITORING MODULE
isFraud
Drift or anomalies

Figure 1: Data flow from ingestion to monitoring.

Business Architecture Diagram for Fraud Detection System: This diagram shows the architecture of the fraud detection system, including the different modules (data ingestion, modeling, etc.), the connections between them, and how the system handles incoming transaction data. It provides a detailed view of how different components interact in the system.

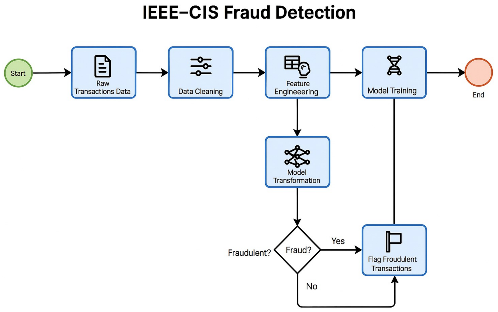

Figure 2: Business architecture and module interactions.

System Workflow for Fraud Detection Process: This diagram outlines the specific workflow followed by the fraud detection system. It details how the system processes each transaction, including steps such as data preprocessing, model classification, and fraud detection decision-making.

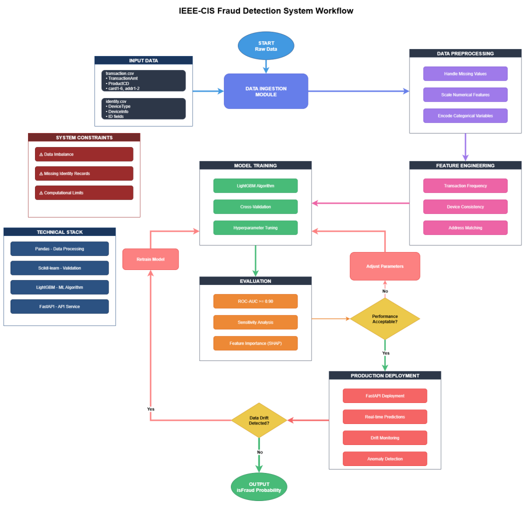

Figure 3: Workflow for transaction processing and classification.

9 Results

9.1 Overall Model Performance

The Random Forest model achieved a ROC-AUC of 0.9244 and an accuracy of 97.03%. This overall performance suggests that the model is quite effective at correctly classifying legitimate transactions. However, the model showed low precision for fraudulent transactions (0.58), indicating a high rate of false positives.

9.2 Sensitivity to Perturbations

The model was subjected to perturbation experiments, where Gaussian noise was injected with amplitudes ranging from 0.01 to 0.20. The results showed a chaotic behavior in model performance:

- With small perturbations (0.01-0.05), the model showed minor fluctuations in metrics like ROC-AUC, precision, and recall, maintaining stable performance. - However, with larger perturbations (above 0.10), the model experienced a significant degradation in performance: - AUC dropped by up to 15.6%. - Precision and recall for fraud became unstable, confirming the sensitivity of the model to data changes.

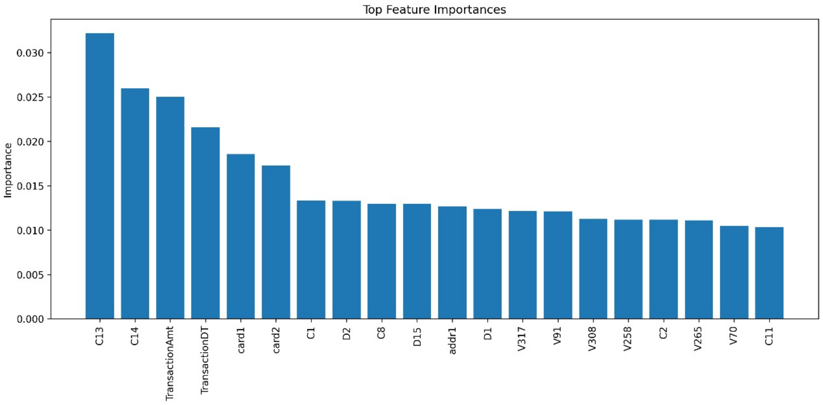

Figure 4: Resultados de la sensibilidad: variación de métricas con ruido.

Machine Learning Results Graphs

Sensitivity to Perturbations and Feature Importances: This graph presents the impact of Gaussian noise on the model's performance, showing how slight changes to the data can lead to drastic changes in metrics like AUC and precision. It also highlights the feature importance scores, which reveal the most influential features in fraud detection.

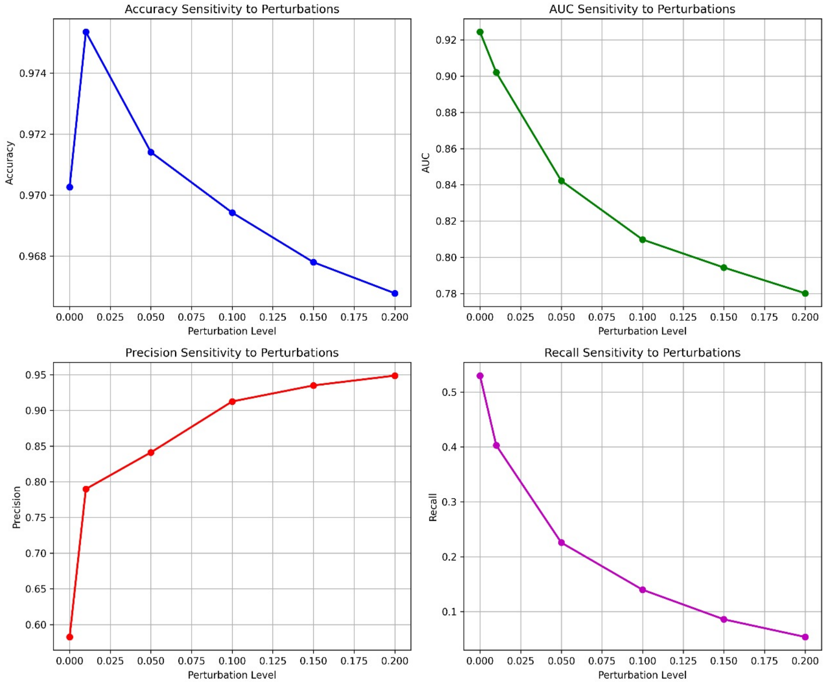

Figure 5: Importancia de características y efecto del ruido en las métricas.

9.3 Feature Importance

The Random Forest model revealed that the following features were most influential in classification:

- TransactionAmt_log: The log transformation of the transaction amount was crucial for reducing heavy-tailed variability in the data and improving the model's stability.
- Card usage frequency: This feature revealed patterns of unusual usage,

indicating that fraud often involves cards with irregular or repeated use.

- Email domain: The presence of inconsistent or suspicious email domains was a key indicator of fraud. Invalid or inconsistent emails increase the likelihood of fraud.
- Device metadata: Inconsistent device patterns, such as using different devices for consecutive transactions, were also a determining factor in classification.

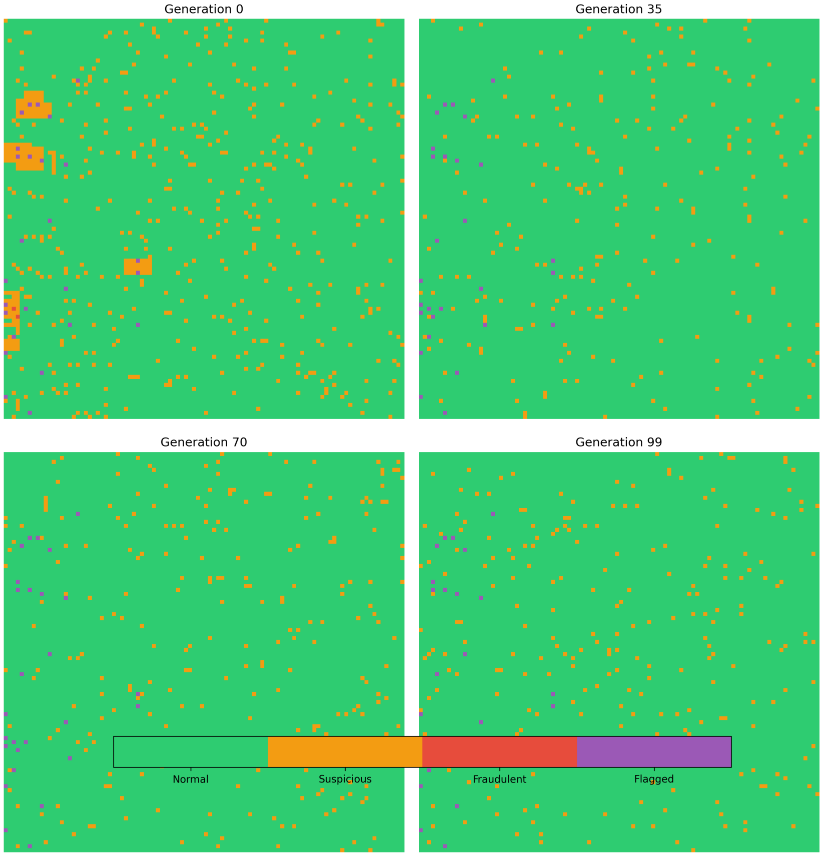

Figure 6: Evolución espacial del autómata celular (fraud hotspots).

9.4 Emergent Behavior of Fraud (Cellular Automata Simulation Results)

The analysis of cellular automata simulations showed that fraud tends to cluster in specific areas of the transaction space, forming fraud hotspots. This supports the hypothesis that fraud is an emergent phenomenon, not a random event, and follows spatial patterns that can be better modeled and detected through spatial techniques.

Cellular Automata Simulation Results: This graph shows the spatial distribution of fraudulent transactions as they propagate across the feature space. The emergent patterns illustrate how fraud clusters and spreads, revealing potential fraud hotspots. This spatial model provides insights into the propagation dynamics of fraud within the transaction data.

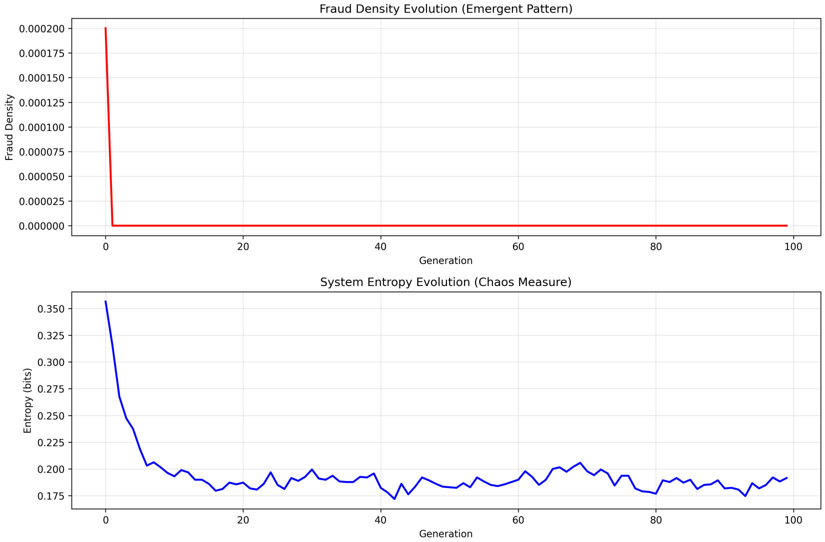

Figure 7: Métricas del autómata: densidad y entropía de fraude.

10 Discussion

10.1 Sensitivity to Initial Conditions and Non-Linear Behavior

The current analysis demonstrates the utility of the Random Forest model in fraud detection, particularly in online transaction systems. However, as evidenced in the sensitivity analysis, the model is highly susceptible to small perturbations in the data. The results indicate that even small variations, such as noise injections between 0.01 and 0.05, caused minor fluctuations in performance. These fluctuations, although minimal, suggest that the model exhibits some level of stability. However, as noise levels increased above 0.10, the performance of the model degraded significantly. This is a critical finding as it underscores the challenges in applying this model in real-world applications where data can fluctuate unpredictably.

The use of cellular automata to simulate the propagation of fraud adds another layer of insight into the behavior of fraud in transaction data. It was observed that fraud tends to cluster in specific areas within the feature space, and these fraud hotspots emerge over time, revealing the spatial dynamics of fraudulent behavior. This spatial behavior supports the notion that fraud detection models should not only focus on the features of individual transactions but also account for the evolving patterns in transaction flows, much like how fraud propagates in real-world systems.

The Random Forest model's success in classifying legitimate transactions with high accuracy indicates its potential for operational use, but its limitations in detecting fraud with the same precision highlight areas for improvement. The relatively low precision for fraudulent transactions emphasizes the need for further refinement in the system. Additionally, the model's sensitivity to perturbations indicates the importance of developing more robust preprocessing techniques, which will help in stabilizing the model's performance in the face of noisy data.

The integration of event-based simulations, specifically the use of cellular automata, proved valuable in understanding the spatial and emergent properties of fraud. It provides a novel approach to capture how fraud propagates and clusters, which could lead to more targeted and efficient detection strategies. This feature could be further optimized by incorporating real-time data and feedback loops to monitor and adjust the fraud detection mechanisms based on the evolving nature of fraud.

11 Conclusion

This study confirms the effectiveness of the Random Forest model in detecting fraudulent transactions, achieving strong overall performance metrics such as ROC-AUC and accuracy. However, the results also reveal critical weaknesses in the model's ability to detect fraud with sufficient precision, particularly in the presence of noisy data. The low precision score for fraudulent transactions and the significant sensitivity to data perturbations highlight the need for a more robust and adaptive system capable of handling real-world data fluctuations.

The cellular automata simulations have successfully illustrated the emergent behavior of fraud, demonstrating that fraud is not a random occurrence but rather an emergent phenomenon that follows spatial patterns within the transaction data. These insights open up avenues for developing more sophisticated detection systems that leverage spatial dynamics and the evolving nature of fraud.

Based on these findings, several recommendations are proposed:

Continuous monitoring and drift detection: Implementing mechanisms for continuous monitoring of the model's performance will help identify shifts in fraud patterns in real-time, allowing for timely interventions.

Periodic retraining: Given the dynamic nature of fraud tactics, periodic retraining of the model using the latest data will ensure that the system remains effective as new fraud patterns emerge.

Enhanced preprocessing: Addressing the system's sensitivity to small data perturbations through more robust preprocessing methods will improve the model's stability and reliability, even in noisy environments.

These recommendations, along with the insights gained from this analysis, lay the groundwork for further improving the fraud detection system to make it more reliable, adaptive, and capable of handling the complexities of online transaction fraud.

References

1 IEEE-CIS Fraud Detection Competition. Kaggle, 2025.

2 Vesta Corporation, "Fraud Prevention Technologies in E-Commerce."

3 Breiman, L. "Random Forests." Machine Learning, 2001.

4 Saltelli, A. et al., "Sensitivity Analysis in Practice," Wiley, 2004.

5 Gleick, J., Chaos: Making a New Science. Penguin Books, 1987.

Glossary

Machine Learning (ML) A subfield of artificial intelligence that allows systems to learn from data without being explicitly programmed. In this report, ML is used to train models that predict fraudulent transactions based on historical data.

Random Forest A machine learning algorithm that builds multiple decision trees and merges their outputs to improve prediction accuracy. It is used in this study to classify transactions as fraudulent or legitimate.

AUC (Area Under the Curve) A metric used to evaluate the performance of classification models. It represents the ability of the model to distinguish between classes (fraud/non-fraud) on a ROC curve. A higher AUC indicates better model performance.

Precision In fraud detection, precision measures the proportion of transactions identified as fraudulent that are actually fraudulent. Low precision indicates a high rate of false positives.

Recall This metric indicates the model's ability to correctly identify fraudulent transactions. A high recall means the model detects most fraudulent transactions but may also include more false positives.

Cellular Automata A mathematical model where a grid of cells represents transaction states (normal, suspicious, fraudulent). Cells interact with their neighbors based on rules that simulate how fraud propagates through the system.

Entropy A measure of disorder or unpredictability in the system. In cellular automata, a decrease in entropy over time suggests that fraudulent transactions are clustering and stabilizing in specific areas of the transaction space.

Mutation Rate In cellular automata, the rate at which random changes are introduced into the system to simulate the appearance of new fraud patterns, adding an element of chaos to the simulation.

Fraud Density A measure of how concentrated fraudulent transactions are in certain areas of the transaction space, observed through cellular automata simulations.

Sensitivity to Perturbations The model's responsiveness to small changes or noise in the data. High sensitivity can lead to large changes in model performance, highlighting the importance of robust data preprocessing.

List of Figures