Streaming Data Credit Card Fraud Detection

By: Gunjan Mukeshbhai Gangwani

Student No.: 501345559

Supervisor: Dr. Tamer Abdou

Date of Submission: October 20, 2025

Toronto Metropolitan University

Table of Contents

Abstract

This study investigates adaptive machine learning techniques for credit card fraud detection using streaming data. The analysis is based on the publicly available Kaggle Credit Card Fraud dataset, which comprises 284,807 transactions, including 492 fraudulent cases. The dataset's high class imbalance and anonymized PCA-transformed features present key challenges for model interpretability and detection accuracy. Data preprocessing involved robust scaling of Amount and Time variables, removal of duplicates, and the creation of sliding-window aggregates to simulate real-time transaction flows. Several supervised learning algorithms like Logistic Regression, Random Forest, Support Vector Machine, k-Nearest Neighbors, and XGBoost will be implemented to assess comparative performance under batch and incremental learning scenarios. The adaptive framework, developed in Python using PySpark and MLlib, employs micro-batch standardization and passive learning to address concept drift in streaming data. Evaluation metrics emphasizes the Matthews Correlation Coefficient (MCC) and Area Under the Precision–Recall Curve (AUCPR) to account for the dataset's extreme imbalance. Results highlight the potential of adaptive ensemble methods and incremental updates for maintaining fraud detection accuracy over time. The study contributes to scalable, real-world fraud detection research by combining reproducible machine learning practices with adaptive learning strategies for evolving financial data streams.

Past Research

Introduction

Financial fraud remains one of the most pressing challenges in today's digital economy, with global losses estimated at over $32 billion annually (Deboran, 2017). The rapid expansion of digital payments and e-commerce has created new opportunities for increasingly sophisticated fraud schemes. Traditional rule-based detection systems struggle to adapt to evolving patterns and often generate high false positive rates, creating unnecessary costs and friction for legitimate customers.

Machine learning offers more dynamic approaches to fraud detection, with real-time performance becoming critical as detection windows are measured in milliseconds (Jeyachandran et al., 2024). However, the challenge extends beyond simple classification. Fraud detection systems must handle severe class imbalance fraudulent transactions typically represent less than 1% of total volume while adapting to concept drift as fraudsters continuously evolve their tactics (Lucas et al., 2020). Success requires balancing fraud detection rates, false positives, real-time performance at scale, and continuous adaptation to emerging patterns.

This literature review examines machine learning-based fraud detection across three key dimensions: algorithmic approaches and performance optimization, adaptive learning strategies for evolving fraud patterns, and real-time deployment considerations. The synthesis reveals that effective fraud detection requires moving beyond static batch models toward adaptive streaming architectures. While systems like SCARFF (Carcillo et al., 2018) demonstrate scalability to 100 million daily transactions with performance comparable to batch learning, significant challenges remain in optimizing trade-offs between accuracy, efficiency, and complexity.

Literature Review Table

Paper	Methodology	Results	Limitations
Adaptive machine learning for credit card fraud detection (**Dal Pozzolo, 2015)**	Comprehensive PhD thesis examining adaptive learning for nonstationary environments. Developed mathematical derivation of undersampling effectiveness, concept drift taxonomy, and verification latency formalization. Evaluated static, sliding window, propagate-and-forget, and ensemble strategies on 75+ million transactions.	Sliding window improved CPk by 18-25% over static approaches. Propagate-and-forget outperformed sliding window by 8-12% for rapid fraud drift. Ensemble of 7 classifiers with exponential decay achieved best overall performance. Established optimal window size of 30 days for gradual drift.	Requires empirical determination of window sizes and ensemble parameters. Computational overhead for multiple ensemble members. Framework complexity may limit adoption in resource-constrained environments.
Credit card fraud detection: A realistic modeling and a novel learning strategy (Dal Pozzolo et al., 2018)	Formalized fraud detection with five-layer control system. Developed separated learning strategy training distinct classifiers on feedback (F_t) and delayed samples (D_t), aggregating via weighted combination (α=0.5). Used Random Forest with balanced bootstrap undersampling on 75+ million transactions.	Aggregated classifiers (AW and AE) achieved 15-20% higher Card Precision at k=100 compared to pooled training. AW achieved CPk=0.75 vs. W=0.58 on 2013 dataset. Demonstrated undersampling's theoretical foundation and feedback informativeness despite sample selection bias.	Requires careful tuning of aggregation weights. Sample selection bias from investigator feedback. Computational overhead of maintaining multiple ensemble members. Static importance weighting techniques failed to improve performance.

Paper	Methodology	Results	Limitations
Reproducible machine learning for credit card fraud detection: Practical handbook. (Le Borgne et al., (2022)	Developed comprehensive reproducible framework with eight chapters covering problem formulation, performance metrics, model selection, imbalanced learning, deep learning, feature engineering, and interpretability. Implemented prequential validation and transaction simulator for synthetic data generation.	Established best practices: temporal splitting essential, multiple baselines required, transaction aggregation features outperform raw attributes, ensemble methods improve robustness. All methods implemented in executable Jupyter notebooks ensuring reproducibility.	Synthetic simulator may not capture all real-world fraud complexities. Framework focuses primarily on batch processing with limited streaming implementation guidance.
Credit card fraud detection using machine learning: A survey (Lucas et al., 2020)	Comprehensive survey synthesizing data-driven credit card fraud detection approaches. Characterized typical fraud detection tasks including dataset attributes, metric selection, and class imbalance handling methods. Focused extensively on dataset shift (concept drift) caused by evolving fraudster tactics and changing cardholder behaviors. Reviewed machine learning methods including traditional algorithms (RF, SVM), sequential models (HMM, LSTM), and graphical models.	Survey identified key challenges: severe class imbalance, concept drift, sequential transaction patterns, importance of transaction aggregation features, need for proper evaluation metrics beyond accuracy. Highlighted that LSTM-based discriminative approaches outperform traditional models by modeling succession patterns in cardholder transaction sequences.	Limited discussion of production deployment challenges, computational costs, and real-time streaming implementation details. Does not extensively cover recent advances in graph neural networks and federated learning for fraud detection.

Paper	Methodology	Results	Limitations
Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative survey (Rahman, 2021)	Six algorithms (LR, KNN, SVM, DT, RF, XGBoost), European dataset, temporal splitting, with/without SMOTE, multiple metrics including computation time. PCA-transformed features limited interpretability	XGBoost best F1=0.87, LR competitive with 60% faster speed, KNN improved with SMOTE, SVM high precision = 0.91 and low recall = 0.68, SMOTE improved recall 12-18%, training vs inference time trade-offs	PCA-transformed features limited interpretability, no concept drift evaluation, deployment considerations not addressed, hyperparameter specificity limits generalizability
Incremental learning strategies for credit cards fraud detection (Lebichot et al., 2021)	Compared five learning paradigms: static batch, periodic retraining, incremental with fixed windows, adaptive incremental with variable windows, and ensemble incremental. Used Logistic Regression, Random Forest, and XGBoost on 150 days of e-commerce data.	Incremental learning matched/exceeded batch performance while reducing computational costs by 60-70%. Ensemble incremental achieved highest AUCPR (0.76 vs. 0.69 static). XGBoost maintained performance better than RF in incremental mode. Transfer learning showed promise for recurring patterns.	Window size management critical too small loses history, too large fails adaptation. Ensemble increases memory and latency. Lacks principled guidance for discarding old models.
SCARFF: A scalable framework for streaming credit card fraud detection with Spark (Carcillo et al., 2018)	Presented SCARFF architecture integrating Apache Kafka for ingestion, Spark Streaming for processing (1-5 second micro-batches), and Cassandra for storage. Implemented Balanced Random Forest with 100 trees, daily updates using 30-day sliding windows, exponential decay weighting (λ=0.05).	Demonstrated linear scalability to 100M transactions/day. Sub-100ms inference latency (95th percentile: 180ms). Achieved AUCPR=0.71-0.73 comparable to batch learning. Maintained stable performance over 6-month deployment with 70% less storage than batch approaches.	Performance depends on cluster configuration requiring DevOps expertise. Window size and update frequency involve empirical trade-offs. Does not handle late-arriving out-of-order transactions. Rare fraud patterns may be lost in aggregation.

Paper	Methodology	Results	Limitations
Streaming active learning strategies for real-life credit card fraud detection: Assessment and visualization (Carcillo et al., 2018).	Evaluated active learning strategies: uncertainty sampling, query-by-committee, expected model change. Simulated investigator feedback with prequential evaluation on 10 million transactions. Compared active strategies against passive learning baseline.	High-risk querying (HR) outperformed exploration strategies (CPk=0.68 vs. 0.43 for uncertainty sampling). HR combined with semi-supervised learning improved to CPk=0.73. Active learning reduced labeling costs by 30-40% but benefits diminished with rapid drift.	Active learning adds deployment complexity requiring real-time uncertainty estimation. Sample selection bias can harm performance if misaligned with actual patterns. Computational overhead for uncertainty adds latency to scoring pipeline.
Leveraging Machine Learning for Real-Time Fraud Detection in Digital Payments (Jeyachandran et al., 2024)	Multiple supervised/unsupervised algorithms, real-time constraints- sub-100ms latency, 10K+ transactions/second, synthetic + real fraud patterns, feature engineering for behavioral/network features	Random Forest performance with 94% precision, 82% recall, 45ms latency, neural network trade-offs, isolation forests for novel fraud about 23% detection of unseen patterns, feature selection impact with 60% latency reduction, concept drift degradation of 15-20% over 3 months	Synthetic data limitations, limited infrastructure discussion, label acquisition challenges not addressed, privacy-preserving techniques not evaluated
SMOTE: Synthetic minority over-sampling technique (Chawla et al., 2002)	Original SMOTE paper introducing synthetic minority oversampling via interpolation between existing minority samples. Algorithm selects k-nearest neighbors (typically k=5) and generates synthetic examples along line segments connecting neighbors.	SMOTE improved minority class recognition across multiple datasets. Combining SMOTE with undersampling of majority class achieved better classifier performance than either technique alone. Particularly effective for C4.5 decision trees.	Generated synthetic samples may not represent realistic patterns, potentially introducing noise. Requires specifying k parameter with no theoretical guidance. Linear interpolation assumption may not hold for complex non-linear patterns.

Summary

The reviewed literature provides important insights into modern fraud detection challenges and solutions. Dal Pozzolo's (2015) work showed that adaptive approaches particularly sliding windows and ensemble strategies consistently outperform static methods by 18-25%, with 30-day windows proving optimal for gradual drift.

Addressing class imbalance requires careful handling. While SMOTE improves recall by 12-18% (Rahman, 2021), Dal Pozzolo et al. (2018) demonstrated that balanced bootstrap undersampling within Random Forest ensembles achieves 15-20% improvements in Card Precision. Their separated learning strategy training distinct classifiers on immediate feedback versus delayed samples effectively addresses verification latency challenges. Real-time performance at scale is achievable. Ensemble methods like Random Forest and XGBoost consistently perform well, with XGBoost achieving F1 scores of 0.87 (Rahman, 2021). However, Logistic Regression remains competitive with 60% faster inference times, highlighting important computational trade-offs.

Le Borgne et al.'s (2022) framework established critical best practices: temporal splitting is essential, transaction aggregation features outperform raw attributes, and multiple baselines ensure fair evaluation. Despite these advances, challenges persist around hyperparameter selection, label acquisition given verification delays, and the interpretability-performance trade-off that complicates regulatory compliance and investigator trust.

Research Questions

RQ1: How do different classification algorithms (e.g. logistic regression, random forest, XGBoost, SVM, KNN) perform on streaming credit-card transaction data?

RQ2: How does incremental or online learning, especially with sliding window technique compared to traditional batch learning in terms of fraud detection accuracy and robustness to imbalance?

RQ3: Which sampling or ensemble strategies (e.g. SMOTE oversampling, undersampling, boosting) most effectively mitigate class imbalance?

RQ4: How sensitive are the models to concept drift and delayed labels in terms of the impact of window size or retraining frequency on performance?

Dataset Statistics and Explanation

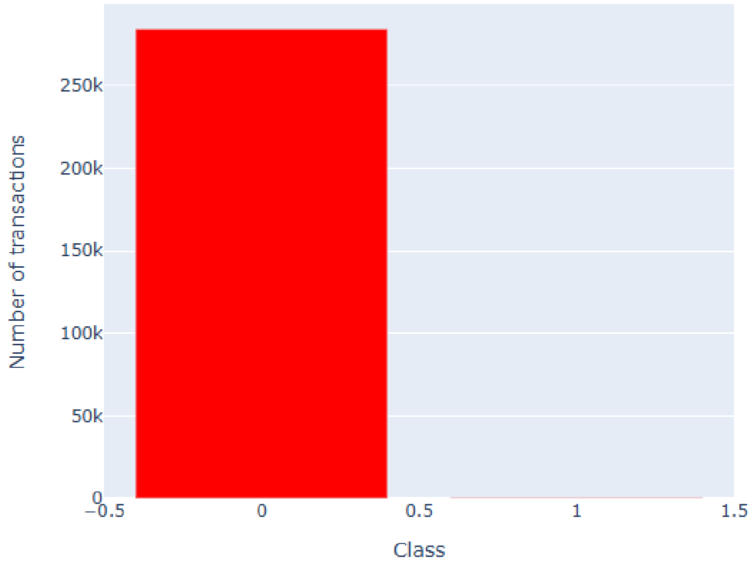
Credit Card Fraud Detection Dataset

Source: Kaggle (*Credit Card Fraud Detection* 2018)
URL: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
Size: 284,807 transactions, including 492 fraud cases
Features: 28 anonymized PCA features + time, amount, and class label

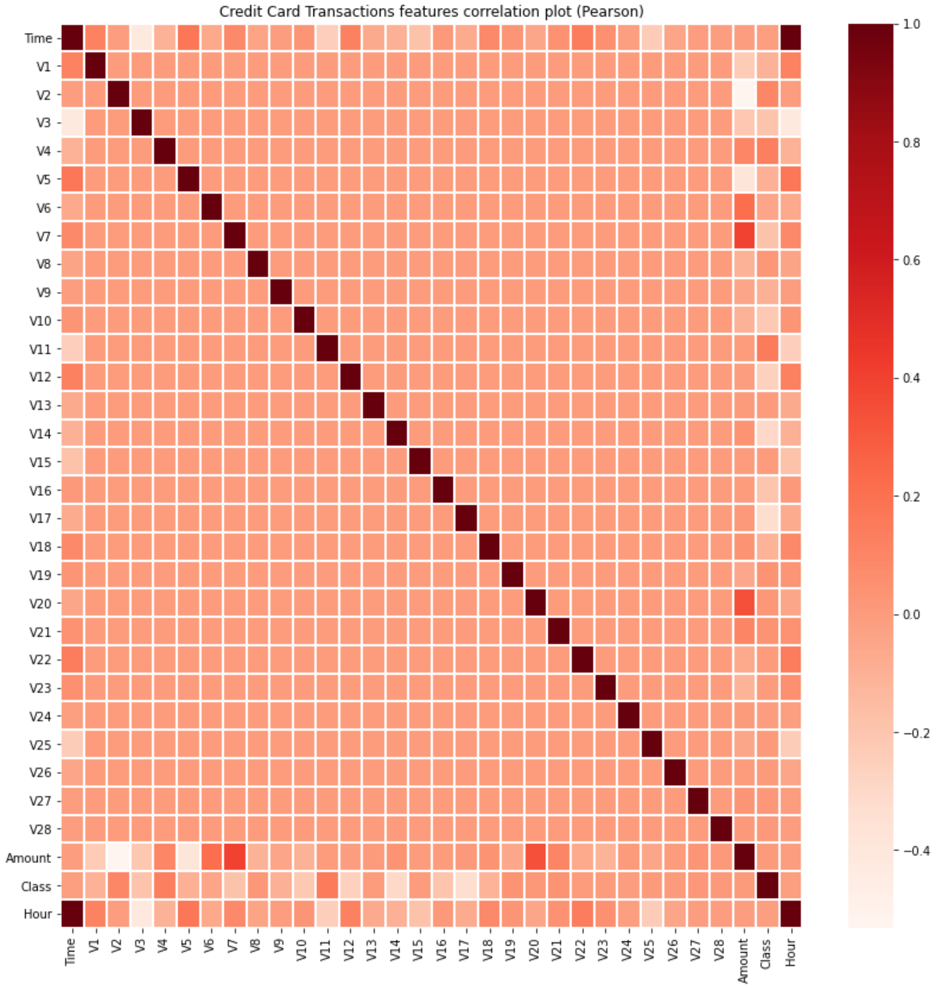
Descriptive statistics

Class balance: Extremely imbalanced – 492 frauds vs 284,315 non-frauds.
Time: spans 0 → 172,792 seconds (seconds elapsed between each transaction in the dataset).
Amount: strongly right skewed; most transactions are small, with a long tail.
V1-V28: centered around 0 with unit-scale behavior. Most components show tight distributions for legitimate transactions and broader, skewed distributions for fraudulent transactions.

Class Imbalance Plot



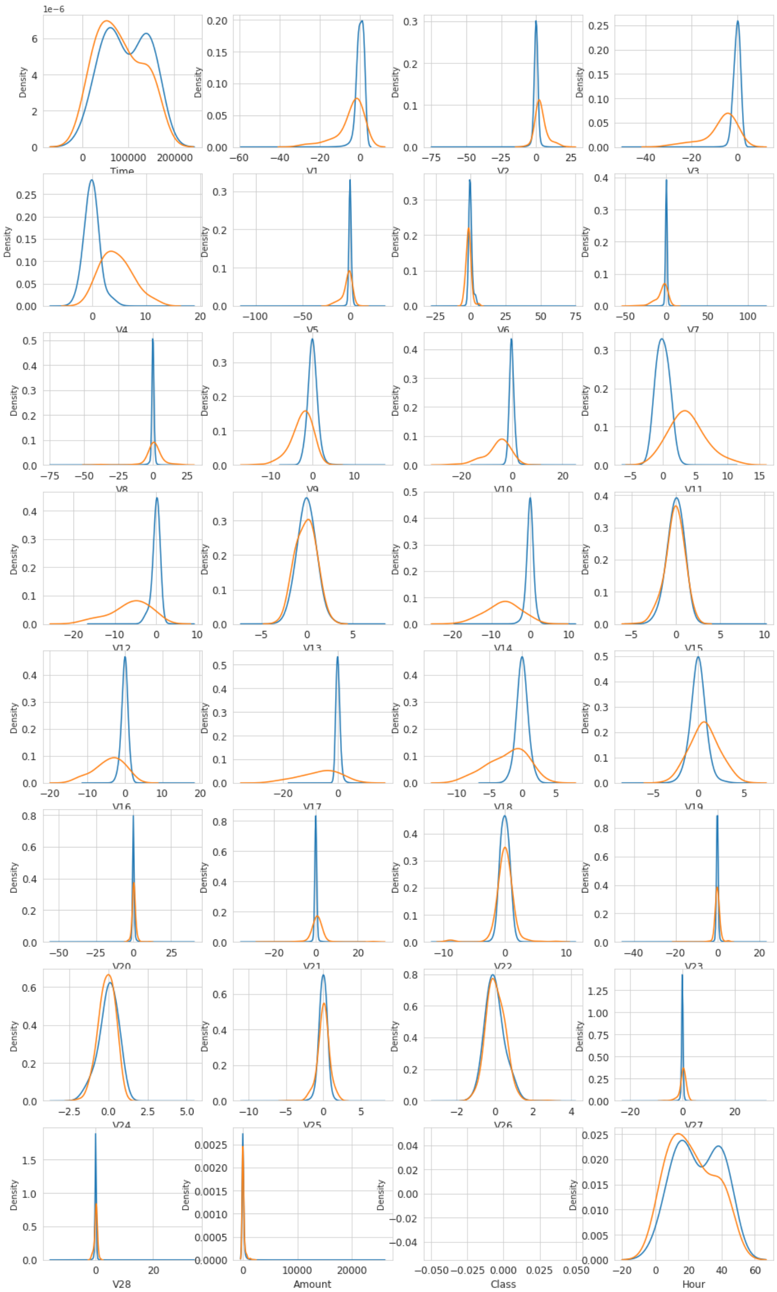
Correlation Plot



PCA components by design are approximately orthogonal, so pairwise correlations between V features are small, however, several components still show meaningful association with the target (Class) and should be examined jointly.

- Top component signals: V4 and V11 show the strongest univariate separation with the label; V12 / V14 / V18 also correlate with Class more than most other components.
- Amount vs Class: low direct Pearson correlation — amount alone is not a reliable fraud indicator; combine with PCA features and temporal context.

Feature Distribution



Based on the density plots (legitimate transactions in blue, frauds in orange) the following patterns were observed:

- V4, V11 — distributions for are markedly different and show strong discriminatory potential.
- V12, V14, V18 — partial separation; these components show differing shapes between classes.
- V1, V2, V3, V10 — show a noticeably different profile or skew for frauds vs non-frauds.

- V25, V26, V28 — near-identical or highly overlapping distributions for the two classes.

Time and Amount show different characteristics relative to the PCA components:

- Amount has a heavy right tail and is only weakly correlated with Class alone.
- Time reveals temporal clustering of some fraud events which can be useful for sliding-window features.

Data Cleaning

- No missing values or negative amounts in the dataset.
- No exact duplicates were present in the working dataset.
- PCA features are already scaled; Amount and Time feature were scaled using Robust Scaling.

Data Preprocessing

- Sliding-window aggregates (transactions per minute/hour).
- For streaming: use incremental or micro-batch standardization to avoid lookahead data leakage.
- Class imbalance handling: focused sampling like SMOTE and oversampling is used.

Constraints

- PCA anonymization prevents domain-level feature interpretation, limiting feature explainability.

The dataset is clean, realistically simulation and well-suited for prototyping streaming fraud-detection systems. The main technical challenges are extreme class imbalance and anonymized features (PCA).

Methodology

Research Design

This study adopts an experimental quantitative research design aimed at evaluating the performance of adaptive machine learning models for credit card fraud detection in streaming data environments. The design follows an iterative approach comprising data preprocessing, feature engineering, model training, incremental updating, and performance evaluation. The central objective is to assess the extent to which adaptive learning mechanisms, particularly incremental and ensemble-based models, improve detection performance under evolving data distributions.

Grounded in prior research by Dal Pozzolo (2017) and Lebichot et al. (2021), this work extends traditional batch learning paradigms into streaming contexts, where models must continuously learn from incoming transactions with minimal latency. The methodology therefore integrates PySpark Streaming and MLlib within Python environment, providing a distributed computing framework that simulates real-time data ingestion and model adaptation.

Data Cleaning and Preprocessing

An exploratory data analysis was conducted to assess variable distributions and identify distinguishing features between legitimate and fraudulent transactions. Several PCA components exhibited noticeable discriminatory potential such as different distributions between classes, near-identical distributions, etc. Data quality assessment confirmed the absence of missing values, negative transaction amounts, and duplicate records. PCA features were already standardized; however, Amount and Time were normalized using Robust Scaling to reduce sensitivity to extreme values. This ensured that all features maintained comparable magnitudes during model training.

To simulate real-world fraud detection, a sliding-window mechanism will be applied to create temporal aggregates. These aggregates facilitated the development of streaming-compatible models by allowing micro-batch feature updates without data leakage. For the incremental learning pipeline, incremental or micro-batch standardization will be adopted. Given the extreme class imbalance, resampling strategies is employed to improve model sensitivity. Techniques such as Synthetic Minority Oversampling Technique (SMOTE) and focused oversampling will be applied to the training folds. The evaluation strategy emphasized avoiding oversampling in the test set to maintain realistic performance estimates.

A major constraint of the dataset is the anonymization of features through PCA transformation. This limits interpretability, as domain-specific insights about transaction types or merchant categories cannot be inferred. Nevertheless, this structure enables fair comparison of algorithmic performance across studies while focusing on methodological innovation rather than feature semantics.

Modeling Framework

Multiple machine learning approaches will be examined, spanning both supervised and unsupervised paradigms. Algorithms such as Logistic Regression, k-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest, and Extreme Gradient Boosting (XGBoost) will be implemented to establish baselines. These models will be evaluated under both batch and incremental learning regimes to assess adaptability in evolving data streams.

For adaptive and streaming learning, the study focused on passive adaptation mechanisms, in which models update incrementally with new data rather than employing retraining strategies. A sliding-window ensemble will be implemented using PySpark Streaming and MLlib, allowing continuous ingestion of transactions in micro-batches. Model performance will be periodically re-evaluated to capture the impact of concept drift, reflecting changing fraud patterns over time while testing the frequency of retraining to get the best performance.

Evaluation Metrics

Performance evaluation emphasizes the use of metrics suitable for highly imbalanced data. In addition to standard measures such as precision, recall, and F1-score, the Matthews Correlation Coefficient (MCC) and Area Under the Precision–Recall Curve (AUCPR) will be prioritized. These metrics provide a more balanced assessment of classifier quality under severe imbalance. Model comparisons will be performed across multiple time windows to assess stability and adaptability in streaming contexts.

Summary

Overall, the methodology integrates both classical and adaptive machine learning paradigms within a streaming framework. The cleaned and preprocessed Kaggle dataset provides a robust foundation for prototyping real-time fraud detection systems. Despite the constraints introduced by PCA anonymization, the pipeline design allows for comprehensive experimentation with incremental learning, ensemble adaptation, and the interplay between supervised and unsupervised detection methods under realistic data conditions.

References

Carcillo, F., Le Borgne, Y. A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2018). Streaming active learning strategies for real-life credit card fraud detection: Assessment and visualization. *International Journal of Data Science and Analytics*, *5*, 285-300. https://arxiv.org/pdf/1804.07481

Carcillo, F., Dal Pozzolo, A., Le Borgne, Y. A., Caelen, O., Mazzer, Y., & Bontempi, G. (2018). SCARFF: A scalable framework for streaming credit card fraud detection with Spark. *Information Fusion*, *41*, 182-194. https://arxiv.org/pdf/1709.08920

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. https://doi.org/10.1613/jair.953

Dal Pozzolo, A. (2015). Adaptive machine learning for credit card fraud detection [Doctoral dissertation, Université Libre de Bruxelles] https://dalpozz.github.io/static/pdf/Dalpozzolo2015PhD.pdf

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.

Deboran, A.-M. (2017, June 20). Credit Card Fraud in Canada. https://mfacc.utoronto.ca/management/media/727/download?inline

Ghosh Dastidar , K., Caelen , O., & Granitzer, A. (2025). *Machine Learning Methods for Credit Card Fraud Detection: A Survey*. IEEE Explore. https://opus4.kobv.de/opus4-uni-passau/frontdoor/deliver/index/docId/1579/file/goshdastidar_granitzer_CCFraudDetection.pdf

Jeyachandran, P., Akisetty, A. S. V. V., Subramani, P., Goel, O., Singh, D. S. P., & Shrivastav, Er. A. (2024). Leveraging Machine Learning for Real-Time Fraud Detection in Digital Payments. Integrated Journal for Research in Arts and Humanities, 4(6), 70–94. https://doi.org/10.55544/ijrah.4.6.10

Kumar Tambi, V. (2022). AI-Powered Fraud Detection in Real-Time Financial Transactions. https://philarchive.org/archive/VARAFD-2

Le Borgne, Y. A., Bontempi, G., Caelen, O., Lebichot, B., & Paldino, G. M. (2022). Reproducible machine learning for credit card fraud detection: Practical handbook. Université Libre de Bruxelles. https://fraud-detection-handbook.github.io/fraud-detection-handbook/Foreword.html

Lebichot, B., Paldino, G. M., Siblini, W., He-Guelton, L., Oblé, F., & Bontempi, G. (2021). Incremental learning strategies for credit cards fraud detection. International Journal of Data Science and Analytics, 12, 165-174. https://publications.uni.lu/bitstream/10993/50241/1/Lebichot2021_Article_IncrementalLearningStrategiesF.pdf

Liu, C., Tang, H., & Yang, Z. (n.d.). Big Data-Driven Fraud Detection Using Machine Learning and Real-Time Stream Processing. https://www.arxiv.org/pdf/2506.02008

Lucas, Y., & Jurgovsky, J. (2020). Credit card fraud detection using machine learning: A survey. https://arxiv.org/abs/2010.06479

Machine Learning Group - ULB (2018) Credit Card Fraud Detection, Kaggle. Available at: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud (Accessed: 20 September 2025).

Naik, S., & Pise, N. (2022). Credit Card Fraud Detection Using Machine Learning. https://ieomsociety.org/proceedings/2022india/170.pdf

Rahman, R. (2021). Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative survey. University of Victoria. https://dspace.library.uvic.ca/server/api/core/bitstreams/e25f1e1b-176f-4d4e-ae82-324e805277ab/content

Sarker, A., Must. Asma Yasmin, Md. Atikur Rahman, Harun, M., & Bristi Rani Roy. (2024). Credit Card Fraud Detection Using Machine Learning Techniques. *Journal of Computer and Communications*, *12*(06), 1–11. https://doi.org/10.4236/jcc.2024.126001

Appendix

GitHub Repository for the project

https://github.com/GunjanGangwani/CIND820-Credit-Card-Fraud-Detection