National College of Ireland

The Hidden Cost of Real-Time: Quantifying Performance and Economic Trade-offs Between Streaming and Batch Machine Learning Across Domains

MSc Research Project
Data Analytics

Shamsa Halima Kasozi Nantale
Student ID: X23344083

School of Computing
National College of Ireland

Supervisor: Christian Horn

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Shamsa Halima Kasozi Nantale
Student ID:	X23344083
Programme:	Data Analytics
Year:	2025
Module:	MSc Research Project
Supervisor:	Christian Horn
Submission Due Date:	11/08/2025
Project Title:	The Hidden Cost of Real-Time: Quantifying Performance and Economic Trade-offs Between Streaming and Batch Machine Learning Across Domains
Word Count:	6903
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	14th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

The Hidden Cost of Real-Time: Quantifying Performance and Economic Trade-offs Between Streaming and Batch Machine Learning Across Domains

Shamsa Halima Kasozi Nantale
x23344083

Abstract

Streaming machine learning offers real-time insights, but how much performance do we sacrifice compared to traditional batch processing? This research quantifies this trade-off for the first time across two contrasting domains: credit card fraud detection with 0.167% fraud rate and electricity price prediction with 42.5% minority class. The study compared batch algorithms against streaming alternatives, testing three enhancement strategies to improve streaming performance. Results show streaming completely fails on fraud detection achieving 0% F1 score without enhancements. Even with improvements, streaming reaches only 79.2% of batch accuracy while costing 63.2% more due to excessive false alarms. On balanced electricity data, streaming works reasonably well with just 7.7% accuracy loss. The stark contrast reveals that data imbalance doesn't hurt streaming gradually but exponentially, as a 440-fold difference in imbalance creates an 8-fold difference in performance loss. For businesses handling rare events like fraud, batch processing remains the better choice despite sacrificing real-time capability. This work helps practitioners choose between streaming and batch based on their data characteristics and cost constraints.

1 Introduction

1.1 Background and Problem Statement

The rise of real-time applications has led to an urgent need for systems that can learn and adapt on the fly, handling ongoing streams of data and making split-second predictions. Modern data stream analysis addresses the challenges of high-velocity, unbounded data sequences from smart devices, social networks, and evolving technological domains (Bahri et al. 2021). Organisations are increasingly relying on streaming algorithms for high-stakes decisions, such as evaluating transactions for fraud in real time or generating immediate forecasts in energy markets.

The fundamental architectural constraints of streaming algorithms create severe vulnerabilities in imbalanced scenarios where single pass learning prevents retrospective class balancing while strict memory limitations prohibit storing minority class examples for future replay. As our experiments demonstrate, these constraints can lead to complete failure when handling extreme class imbalance, with baseline streaming algorithms achieving

0% F1 score on fraud detection tasks. In cases like fraud detection where fraud makes up 0.167% of transactions, these systems might go through hundreds of samples before seeing a single fraud, leading them to favor common patterns. Recent studies show that local challenges in the data, such as how rare cases are spread out or appear in different forms, can have a bigger impact than just the overall imbalance. Standard real-time models often can't recover once they miss these rare patterns (Brzezinski et al. 2021) .

This transition from traditional batch processing represents a fundamental paradigm shift with largely unquantified performance implications. Whilst batch machine learning benefits from complete dataset access and multiple optimisation passes, streaming algorithms must make decisions with limited information and single pass learning constraints (Domingos and Hulten 2000). Recent advances demonstrate continued algorithmic evolution. The Extremely Fast Hoeffding Adaptive Tree (EFHAT) achieves statistically more efficient learning mechanisms (Manapragada et al. 2022), whilst the Multi-Label Hoeffding Adaptive Tree (MLHAT) addresses complex multi-output scenarios (Esteban et al. 2024). The River framework has standardised streaming algorithm implementation (Montiel et al. 2021), yet systematic performance quantification against batch equivalents remains limited across diverse application domains.

1.2 Literature Gap and Justification

Current studies on streaming machine learning largely emphasise developing algorithms and handling concept drift, with in-depth analysis showing how detection becomes more complex depending on the size and type of drift (Aguiar and Cano 2024). Recent work addresses concept drift in fraud detection using XGBoost (Shahapurkar and Patil 2023), yet these studies focus on single algorithm improvements rather than systematic cross-paradigm comparison. Existing comparative studies evaluate algorithms within single domains (Esteban et al. 2024), failing to capture real-world performance variations across different problem characteristics.

This gap in understanding creates major challenges for decision makers deciding between streaming and batch approaches. Without clear data on drops in performance, organisations struggle to balance real-time processing and predictive accuracy, risking significant financial losses. False positives cost $18.93 per declined transaction (Wallny 2022), while false negatives lead to direct losses equal to transaction amounts. Although cost modeling exists in some areas (Elkan 2001), cost-benefit analysis comparing streaming and batch methods are lacking. Crucially, no study quantifies performance drops when shifting from batch to streaming, leaving decision makers without clear guidance.

1.3 Research Questions

This research addresses the fundamental question: **What are the quantifiable trade-offs in performance and business value when choosing streaming over batch machine learning paradigms?** This is explored through four specific sub-questions:

- RQ1: What is the magnitude of systematic performance loss when transitioning from batch to streaming machine learning algorithms?

- RQ2: How does systematic performance loss vary across domains with different class imbalance characteristics?

- RQ3: To what extent can enhancement strategies mitigate systematic performance loss in streaming algorithms?

- RQ4: What is the business cost impact of systematic performance loss in fraud detection applications?

1.4 Contribution and Organisation

This research provides the first systematic quantification of streaming versus batch performance loss across domains with contrasting characteristics. By translating technical performance metrics into business costs, this work enables evidence based decision-making for streaming adoption in real-world applications.

The remainder of this report presents a comprehensive related work examining current streaming capabilities and limitations (Section 2), detailed cross-domain methodology for fair comparison (Section 3), Design Specification (Section 4), Implementation (Section 5) extensive evaluation results with statistical validation (Section 6) and Conclusion and Future Work (Section 7).

2 Related Work

2.1 Evolution and Limitations of Streaming Machine Learning

The development of streaming machine learning was marked by the foundational research of Domingos and Hulten (2000), which demonstrated that decision trees could be derived from data streams. Recent advancements encompass the Extremely Fast Hoeffding Adaptive Tree (EFHAT), which demonstrates enhanced concept drift recovery (Manapragada et al. 2022), and the River framework which has emerged as the standard implementation platform (Montiel et al. 2021). Streaming algorithms are great for real-time applications because they process data in small chunks, only learning from each one once, and let you keep getting updates without having to store old data.

However, a fundamental limitation persists where algorithmic developments are evaluated only against other streaming approaches, not against batch baselines. This within-paradigm comparison obscures the critical industry question: How much performance do I sacrifice for real-time capability? Without systematic cross-paradigm evaluation, organizations lack empirical guidance for deployment decisions.

2.2 The Compound Challenge of Streaming with Class Imbalance

Streaming's single-pass learning and strict memory constraints create severe vulnerabilities with imbalanced data, while batch algorithms can look at the same data many times. In addition to uneven class distribution, there are problems with scattered minority instances where borderline samples and rare cases that make classification harder (Brzezinski et al. 2021). The Dynamic Updated Ensemble (DUE) deals with class role reversal (Li et al. 2020), but it doesn't measure how much accuracy is lost compared to batch learning.

In scenarios with less than 1% minority class, streaming algorithms may process hundreds of instances before encountering a single minority case. This delay results in persistent bias that the model cannot overcome. Although previous studies have described these failures in general terms, Li et al. (2020) highlight the inability of models to adapt when distributions shift, and Chrysakis and Moens (2020) note severe degradation during continual learning. The experiments presented here offer concrete evidence, showing that both VFDT and HAT achieve an F1 score of 0% on the fraud detection task when no enhancements are applied, revealing the vulnerability of baseline streaming algorithms in real-world imbalanced environments.

2.3 Algorithmic Responses to Streaming Imbalance

Researchers have come up with a number of ways to make streaming work better in difficult situations. Bernardo and Della Valle (2021) presented VFC-SMOTE, which preserves logarithmic time complexity while yielding superior outcomes in over 50% of the assessed scenarios. Cano and Krawczyk (2022) showed that ROSE, with its adjustable parameters and balanced buffers for each class, worked better than 30 other ensemble methods. In network traffic classification, Liu et al. (2023) highlighted the need for streaming systems to manage shifting imbalance ratios, concept drift and limited access to labeled data. Their MicFoal framework involves human expert input, pointing to real-world costs that go beyond algorithmic complexity.

More importantly, Amekoe et al. (2024) raised a critical point with one of the few direct comparisons between batch and streaming methods. Their findings show that batch models achieve 22% higher AUCPR on fraud detection tasks, confirming what many practitioners have suspected as streaming often trades off a substantial amount of performance to meet real-time demands.

2.4 Concept Drift and Environmental Dynamics

The interaction between concept drift and class imbalance creates compound problems. Aguiar and Cano (2024) analysed drift locality across 2760 scenarios, revealing how drift characteristics influence performance. Azeem et al. (2022) demonstrate 5-15% accuracy loss when new parameters emerge in electrical load forecasting, as models built for static feature spaces cannot adapt to evolving parameters. Chrysakis and Moens (2020) achieve 40% improvement against catastrophic forgetting using Class-Balancing Reservoir Sampling (CBRS).

In highly imbalanced settings, traditional drift detection often fails because changes in the minority class barely affect overall accuracy, making it difficult for models to respond. While some advanced solutions have been proposed, their performance is rarely compared directly with batch models that are periodically retrained, thus leaving an important gap in understanding.

2.5 The Missing Economic Perspective

Business cost modeling remains largely absent in streaming literature. Wallny (2022) estimates false positives at $18.93 per rejected transaction, but this model hasn't been used yet for streaming versus batch comparisons. Recent empirical evidence from Amekoe et al. (2024) demonstrate that batch incremental models achieve up to 22% relative

improvement in AUCPR over streaming approaches in fraud detection yet the economic implications of this performance gap remain unquantified. This is a serious gap in fraud detection where slight performance decrease corresponds to severe financial losses. The literature's focus on algorithmic innovation over economic impact disconnects academic research from practical deployment needs.

2.6 Enhancement Strategies and Their Contextual Limitations

The literature presents several strategies aimed at improving streaming performance on imbalanced data. Bernardo and Della Valle (2021) propose VFC-SMOTE, a synthetic oversampling technique that continuously generates minority class examples to maintain their presence in the data stream. Cano and Krawczyk (2022) introduce ROSE, which combines self-adjusting weighted learning with ensemble methods to address imbalance in streaming contexts.

These strategies show clear improvements within the streaming paradigm. However, their evaluations focus solely on comparing enhanced streaming models to baseline streaming, without considering how they stack up against batch learning. While VFC-SMOTE boosts minority class performance and ROSE outperforms 30 state-of-the-art streaming methods, neither study includes batch baselines. This narrow focus makes it impossible to tell whether such enhancements actually make streaming competitive with batch processing, or if they simply close part of the gap. Without cross-paradigm benchmarks, the literature overlooks a crucial question, leaving practitioners without the evidence needed to decide when streaming is truly a viable alternative.

2.7 Synthesis and Critical Research Gaps

The review shows three major gaps in the current literature on streaming machine learning.

First, evaluations are conducted strictly within the same paradigm. Streaming methods are compared only to other streaming methods, and batch methods are compared only to batch methods. One of the few exceptions, Amekoe et al. (2024) directly compares both and confirms that batch learning outperforms streaming in fraud detection. This separation limits our ability to understand the actual trade-offs between real-time processing and predictive accuracy.

Second, economic considerations are largely missing. While Wallny (2022) provides cost models for fraud detection that quantify false positives at $18.93 per declined transaction, no study uses them to quantify the financial implications of choosing streaming over batch paradigms. Recent empirical comparisons like Amekoe et al. (2024) demonstrate batch superiority in predictive performance but do not translate these differences into business costs. Without this economic analysis, practitioners are left without the insights needed to make informed decisions about model deployment.

Third, research tends to focus on a single domain at a time. Whether it is fraud detection, network traffic analysis, or electricity forecasting, studies often stay within one context. This makes it difficult to evaluate how streaming performance holds up across different types of problems, especially when class imbalance, cost sensitivity, or time-based patterns vary.

Together, these gaps show that the field emphasises algorithmic innovation over practical deployment guidance. Techniques like VFC-SMOTE, ROSE, and CBRS are prom-

ising, but they are typically tested only against other streaming methods, not against the broader range of available options.

This study responds to these gaps by directly comparing batch and streaming performance using both technical and business metrics across very different domains. By translating predictive performance into financial terms and testing under contrasting conditions, we aim to support data-driven decisions about when the speed of real-time learning justifies the performance trade-off.

This leads to our guiding research question: **What are the quantifiable trade-offs in performance and business value when choosing streaming over batch machine learning paradigms?**

3 Methodology

3.1 Research Design and Approach

This study takes a practical data-driven approach to compare how machine learning performs in batch versus streaming setups, using proven methods from recent research (Montiel et al. 2021, Aguiar and Cano 2024, Cano and Krawczyk 2022). This study focuses on the measurable drop in performance that occurs when shifting from batch to streaming processing and tests how effective specific strategies are at reducing that loss. It employs a systematic comparison across two different application areas, allowing the results to be tested for consistency across domains as recommended by recent streaming evaluation guidelines (Esteban et al. 2024).

This study adopts the Knowledge Discovery in Databases (KDD) process (Fayyad et al. 1996) to guide the experimental workflow. KDD provides a structured sequence of steps for extracting insights from existing datasets, making it well-suited to both batch and streaming contexts. Within this study, the phases comprise: selecting datasets that represent distinct streaming challenges; preprocessing to clean data, remove duplicates, and address class imbalance; transforming features for model compatibility; applying and optimising both batch and streaming algorithms; and interpreting results through statistical validation and business cost analysis.

The approach extends concept drift evaluation frameworks and goes beyond what has been done so far by including systematic loss quantification. The business cost integration is based on cost-sensitive learning (Elkan 2001) and is applied to streaming using previously proposed frameworks (Wallny 2022). This technique guarantees that the findings are generalisable principles to improve streaming algorithms and not domain-specific artifacts.

3.2 Cross-Domain Framework and Data Selection

Two datasets illustrate different scenarios that highlight key challenges in data stream analysis. The Credit Card Fraud Detection dataset from Kaggle originally contains 284,807 transactions with 31 columns including the binary target variable Class. Data preprocessing identified and removed 1,081 duplicate transactions resulting in 283,726 unique records with 30 input features. The cleaned dataset retains a fraud rate of 0.167% with 473 fraud cases, illustrating the severe class imbalance typical of financial transaction records. This dataset reflects a context in which errors can result in direct financial impact. The Electricity Price Prediction dataset from CapyMOA includes 45,312 records

with 9 features, including the class variable showing UP/DOWN. The class distribution is reasonably balanced with 42.5% UP and 57.5% DOWN, and the dataset shows temporal variation, reflecting changing conditions in real-time energy markets.

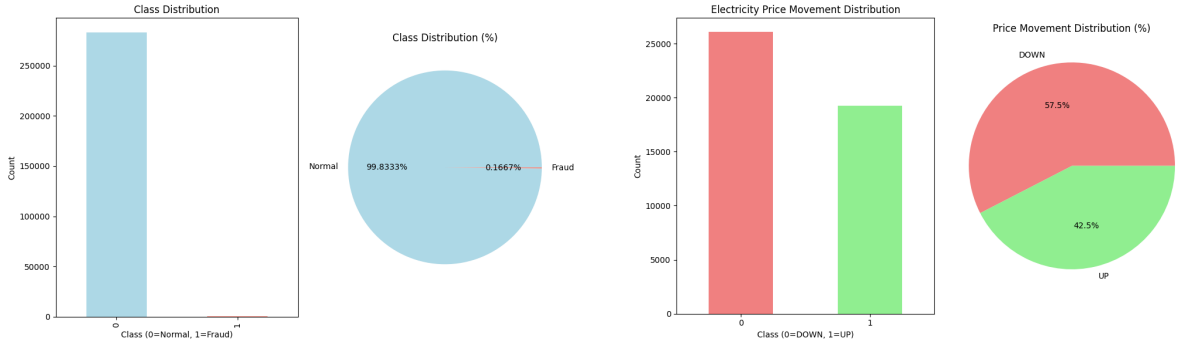


Figure 1: Dataset Class Distributions

This systematic domain variation enables identification of performance boundary conditions whilst testing hypotheses regarding class imbalance impact on systematic loss magnitude. The selection spans high-stakes decision environments for fraud detection and temporal forecasting scenarios in energy markets, providing complementary evaluation contexts that address different streaming algorithm challenges as identified in recent literature (Shahapurkar and Patil 2023, Bahri et al. 2021).

3.3 Algorithm Selection and Enhancement Strategies

Algorithm selection spans both paradigms for systematic comparison. Batch algorithms comprise Random Forest (ensemble baseline), Decision Tree (interpretable comparison), and Logistic Regression (with/without scaling). Streaming algorithms implemented via River framework, include VFDT (basic streaming trees), HAT (adaptive to concept drift) and Streaming Logistic Regression (incremental learning) following recent classifications (Manapragada et al. 2022). Three enhancement strategies target streaming limitations identified in literature:

Weighted Learning: Amplifies minority class influence through increased class weights in the 1-20x range, countering inherent majority bias in single-pass learning.

Pre-training: Initializes models using balanced historical data with 500-15,000 samples range, mitigating cold-start problems where initial imbalance causes irreversible bias.

Artificial Fraud Injection: Synthetically inserts minority examples at regular intervals in credit card data only, testing whether continuous fraud exposure helps maintain detection capability during extended periods of normal transactions.

Bayesian optimization determined optimal parameters for each combination of strategy algorithms in both domains. The parameter ranges were selected based on preliminary experiments and computational constraints. Notably, artificial fraud injection emerged as a valuable negative result, consistently underperforming weighted learning thus suggesting label quality outweighs exposure frequency.

3.4 Experimental Protocol and Statistical Validation

While F1 score serves as the primary metric for comparing machine learning model performance, systematic loss is measured differently in each domain. It uses business cost for fraud detection and the F1 score for electricity prediction. This approach allows each domain to be evaluated by its most relevant metric. Business cost analysis follows Wallny (2022). In this context, false positives cost $18.93 for each declined transaction, while false negatives cost the actual fraud amount. Batch algorithms calculate costs based on average fraud amounts, while streaming algorithms use actual transaction values. This reflects real-world deployment.

Domain-specific metric selection employing business cost for fraud detection versus F1 for electricity ensures systematic loss calculations reflect real-world priorities rather than imposing uniform metrics across disparate applications. Systematic Loss is calculated using these domain-specific metrics: percentage cost increase for fraud detection and percentage F1 reduction for electricity prediction.

Statistical validation employs pairwise t-tests ($\alpha = 0.05$) with multiple comparison corrections and Cohen's d for effect sizes. Enhancement parameters undergo Bayesian optimisation using Tree-structured Parzen Estimator (TPE) in Optuna, maximising F1 score through 3-fold cross-validation over 30 trials. All metrics are reported with 95% confidence intervals from multiple independent runs following current statistical validation standards.

3.5 Research Process and Workflow

The systematic research process follows a logical progression ensuring methodological rigor. First, baseline performance is established by evaluating both batch and streaming algorithms without enhancements on both datasets, identifying fundamental performance gaps.

Second, these baseline results inform the design of three enhancement strategies targeting identified weaknesses: weighted learning addresses class imbalance bias, pretraining tackles cold-start problems, and artificial fraud injection maintains minority class awareness.

Third, Bayesian optimization determines optimal parameters for each enhancement-algorithm combination through 30 trials with 3-fold cross-validation using the TPE algorithm, ensuring each strategy is evaluated at its best configuration.

Fourth, using these optimized parameters, enhanced algorithms are evaluated across both domains to quantify improvement and systematic loss, revealing the effectiveness of each enhancement strategy.

Finally, comprehensive evaluation includes business cost analysis using the Wallny (2022) framework and statistical validation through 10 independent experimental runs with fixed seeds. This enables t-tests to confirm statistical significance of findings and addresses variability in non-deterministic algorithms like HAT.

3.6 Limitations and Strengths

The methodology has several notable strengths. Cross-domain validation helps reduce domain-specific biases, improving the robustness and generalisability of the findings. Bayesian optimization facilitates objective and reproducible hyperparameter selection,

minimizing manual tuning bias. Furthermore, comprehensive statistical validation including the use of confidence intervals and effect sizes strengthens the reliability of performance comparisons. The integration of business cost considerations anchors the evaluation in practical real-world deployment contexts thus addressing critical gaps in the existing streaming evaluation literature.

Several limitations must be acknowledged. Testing only two domains may limit generalisability. HAT's inherent variability requires multiple runs, which increases computational overhead. The credit card dataset from 2013 may not reflect current fraud patterns. Pre-training strategies assume access to historical data, which differs from pure streaming scenarios. Artificial fraud injection showed limited improvement, indicating that data quality is more important than exposure frequency. These limitations are addressed through thorough statistical validation, cautious interpretation and clear documentation of boundary conditions for practitioners to assess applicability.

3.7 Research Questions Mapping and Implementation

The methodology systematically addresses each research question through targeted strategies. RQ1 quantifies systematic loss magnitude through controlled batch-streaming comparisons using domain-specific metrics where business cost for fraud and F1 for electricity. RQ2 examines how the 440-fold difference in class imbalance with 0.167% vs 42.5% affects performance degradation. RQ3 evaluates three enhancement strategies that are weighted learning, pre-training and artificial fraud injection with Bayesian-optimized parameters and statistical validation. RQ4 applies the Wallny (2022) cost framework of $18.93 per false positive to quantify economic impact beyond technical metrics. Reproducibility is ensured through fixed random seeds and 10 independent runs. Despite inherent stochasticity in Bayesian optimization and HAT with $\pm 5\%$ parameter variation, statistical validation confirms significant and generalisable findings.

4 Design Specification

The experimental framework uses a comparative structure that evaluates batch and streaming methods across different situations with varying levels of class imbalance. This new approach fills a significant gap in streaming research by creating direct performance baselines under extreme conditions.

Core Architecture: The framework has a three-phase evaluation process. In the first phase, it sets baseline performance using standard algorithms without changes. The second phase uses improvement strategies optimised through Bayesian search over 200 iterations. The third phase combines results through paired statistical comparison, highlighting consistent performance gaps between methods.

Technical Components: The system integrates multiple algorithmic approaches. For streaming, Hoeffding Trees (VFDT/HAT) provide incremental learning through single-pass processing. For batch processing, Random Forest, Decision Trees, and Logistic Regression establish performance ceilings using complete data access. This dual-paradigm architecture enables quantitative comparison previously unavailable in literature.

Enhancement Framework: Three improvement strategies tackle streaming challenges under extreme imbalance:

- Weighted Learning: Dynamically adjusts class weights during training, multiplying

the importance of the minority class by optimised factors ranging from 3x to 17x based on the severity of imbalance.

- Artificial Fraud Injection: Adds synthetic minority examples at regular intervals, ensuring class visibility in the data stream.
- Pre-training: Initializes models using balanced historical data before streaming begins.

Economic Evaluation Model: The framework includes the Wallny (2022) cost model, which assigns $18.93 for each false positive and transaction penalties for false negatives. This turns abstract performance measures into tangible business effects, revealing the hidden costs of streaming deployment.

Statistical Validation Requirements: All experiments need 10 independent runs with fixed seeds, paired t-tests for significance testing, and 95% confidence intervals for reliability checks. This ensures that observed differences indicate real gaps rather than random variation.

5 Implementation

The experimental framework was built in Python 3.10 using River 0.15 for streaming algorithms, scikit-learn 1.3 for batch models, and Optuna 3.1 for Bayesian optimisation.

The implementation produces four key outputs. First, it provides temporal performance visualisations that track F1 scores, cumulative costs, and weighted accuracy over 50,000 samples. These visualisations show how the algorithms adapt. Second, there are comparison tables that rank all algorithm enhancement combinations based on business cost and technical performance. Third, it includes statistical validation reports from 10 independent runs. These reports feature confidence intervals, pairwise t-tests, and systematic loss quantification between streaming and batch methods. Fourth, it offers one-sided t-tests that confirm enhancement effectiveness. The tests confirm significant improvements, with p value less than 0.0001 for both algorithms across the different domains.

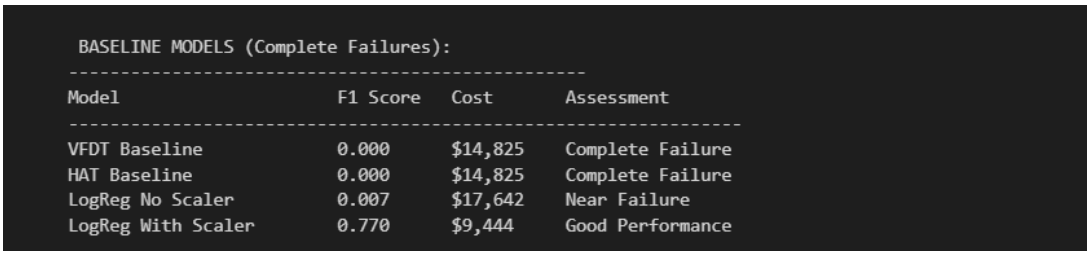
The system also creates Bayesian optimisation traces that record parameter convergence over 200 iterations for each enhancement strategy. All experiments can be reproduced because of fixed random seeding. The complete source code and datasets are available in the accompanying repository.

6 Evaluation

This evaluation presents comprehensive experimental results comparing batch and streaming machine learning paradigms across across two contrasting domains of credit card fraud detection with extreme imbalance of 0.167% minority and electricity price prediction with 42.5% minority. The stark differences in outcomes reveal that streaming algorithm viability depends exponentially on dataset characteristics, with systematic performance loss ranging from 7.7% to complete algorithmic failure.

6.1 Catastrophic Failure versus Functional Performance under Class Imbalance

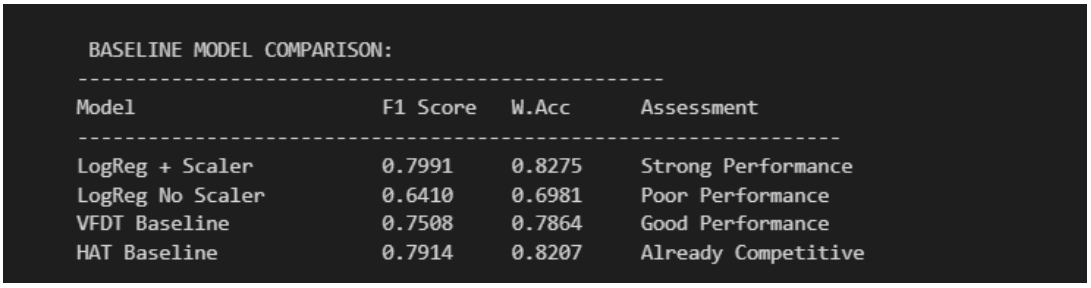
The most interesting finding comes from looking at the baseline streaming performance across different domains. Figure 2 and Figure 3 show the first results of testing the algorithm, revealing a fundamental divide in streaming viability.



BASELINE MODELS (Complete Failures):

Model	F1 Score	Cost	Assessment
VFDT Baseline	0.000	$14,825	Complete Failure
HAT Baseline	0.000	$14,825	Complete Failure
LogReg No Scaler	0.007	$17,642	Near Failure
LogReg With Scaler	0.770	$9,444	Good Performance

Figure 2: Baseline Models - Credit card fraud



BASELINE MODEL COMPARISON:

Model	F1 Score	W.Acc	Assessment
LogReg + Scaler	0.7991	0.8275	Strong Performance
LogReg No Scaler	0.6410	0.6981	Poor Performance
VFDT Baseline	0.7508	0.7864	Good Performance
HAT Baseline	0.7914	0.8207	Already Competitive

Figure 3: Baseline Model - Electricity

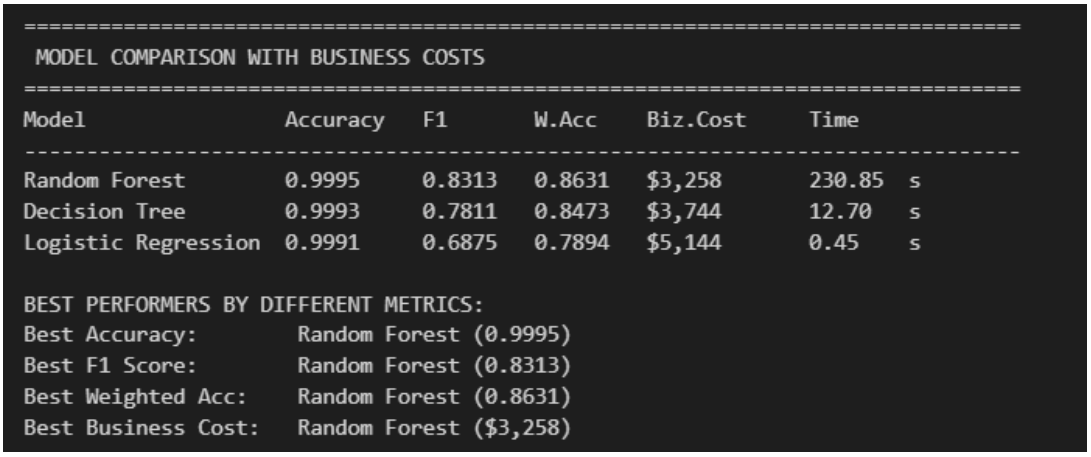
The difference is notable because streaming algorithms completely fail with a 0% F1 score on highly imbalanced fraud data but show varied performance on more balanced electricity data, with F1 scores ranging from 64.18% to 79.91% . This huge 440-fold gap in class balance of 0.167% vs 42.5% leads to dramatically different results. Fraud detection sees total failure versus reasonable to good performance on electricity data though feature scaling proves critical for some models.

For fraud detection, this catastrophic failure occurs because streaming algorithms process hundreds of normal transactions before encountering a single fraud case, causing irreversible bias toward the majority class. The single-pass learning constraint prevents recovery from this initial misclassification pattern.

This substantial difference confirms the theoretical concerns raised by Aguiar and Cano (2024) and Cano and Krawczyk (2022) about how concept drift interacts with class imbalance, while also quantifying to what Brzezinski et al. (2021)'s qualitative observation that streaming algorithms struggle when faced with extreme data challenges.

6.2 Batch Algorithm Performance Benchmarks

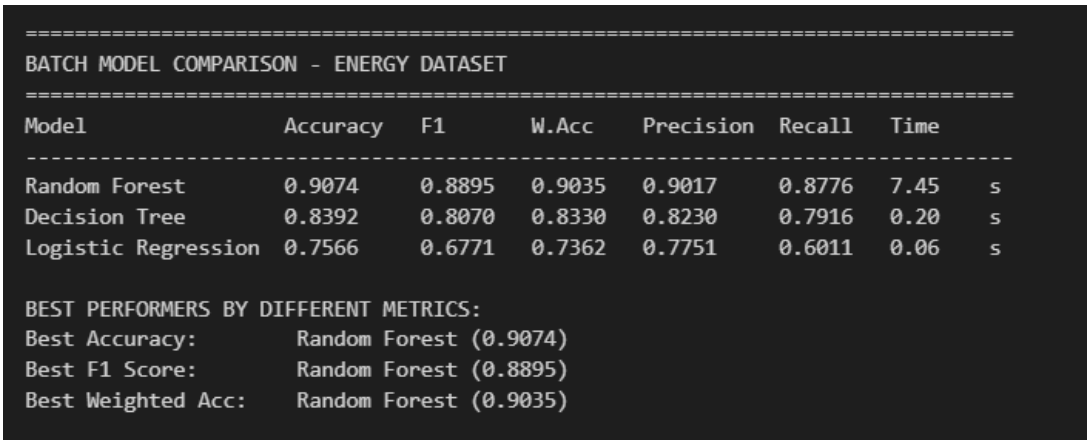
To establish performance benchmarks, three batch algorithms were evaluated over 10 independent runs using a fixed random seed of 42. Figures 4 and 5 present the batch results for fraud detection and electricity prediction respectively.



```
================================================================================
 MODEL COMPARISON WITH BUSINESS COSTS
================================================================================
Model               Accuracy   F1       W.Acc    Biz.Cost     Time
--------------------------------------------------------------------------------
Random Forest       0.9995     0.8313   0.8631   $3,258       230.85  s
Decision Tree       0.9993     0.7811   0.8473   $3,744       12.70   s
Logistic Regression 0.9991     0.6875   0.7894   $5,144       0.45    s

BEST PERFORMERS BY DIFFERENT METRICS:
Best Accuracy:        Random Forest (0.9995)
Best F1 Score:        Random Forest (0.8313)
Best Weighted Acc:    Random Forest (0.8631)
Best Business Cost:   Random Forest ($3,258)
```

Figure 4: Enhanced Batch Model Comparison with Business Costs - Credit Card Fraud



```
================================================================================
BATCH MODEL COMPARISON - ENERGY DATASET
================================================================================
Model               Accuracy   F1       W.Acc    Precision  Recall   Time
--------------------------------------------------------------------------------
Random Forest       0.9074     0.8895   0.9035   0.9017     0.8776   7.45    s
Decision Tree       0.8392     0.8070   0.8330   0.8230     0.7916   0.20    s
Logistic Regression 0.7566     0.6771   0.7362   0.7751     0.6011   0.06    s

BEST PERFORMERS BY DIFFERENT METRICS:
Best Accuracy:        Random Forest (0.9074)
Best F1 Score:        Random Forest (0.8895)
Best Weighted Acc:    Random Forest (0.9035)
```

Figure 5: Batch Model Comparison - Electricity Dataset

Random Forest is the best batch algorithm for both domains. For fraud detection, it achieves the best performance across all metrics with F1 of 0.8313 and Business Cost of $3,258. It also does very well at predicting electricity, with an F1 score of 0.8895 and a weighted accuracy of 0.9074. This better performance is because it is an ensemble which is better at dealing with class imbalance than single classifiers.

Statistical validation across 10 independent runs establishes the clear performance benchmarks used throughout this analysis. Random Forest achieves an F1 score of 0.8531 and a Business Cost of $2,763 for fraud detection. It also reaches an F1 score of 0.8874 and a weighted accuracy of 0.9017 for electricity prediction. Statistical tests confirm Random Forest's advantage with a p-value less than 0.001 for all pairwise comparisons in the energy dataset. Significant improvements appear in fraud detection, where F1 scores show a p-value less than 0.01. The narrow confidence intervals indicate reliable performance for production deployment. The fraud F1 scores range from 0.8338 to 0.8724, while the electricity F1 scores range from 0.8850 to 0.8898.

These validated metrics, rather than results from single runs form the basis for all streaming comparisons and set the performance ceiling for evaluating streaming algorithms.

6.3 Temporal Learning Dynamics

The illustrations below demonstrate the temporal evolution of algorithm performance, revealing fundamentally different learning dynamics between domains and paradigms.

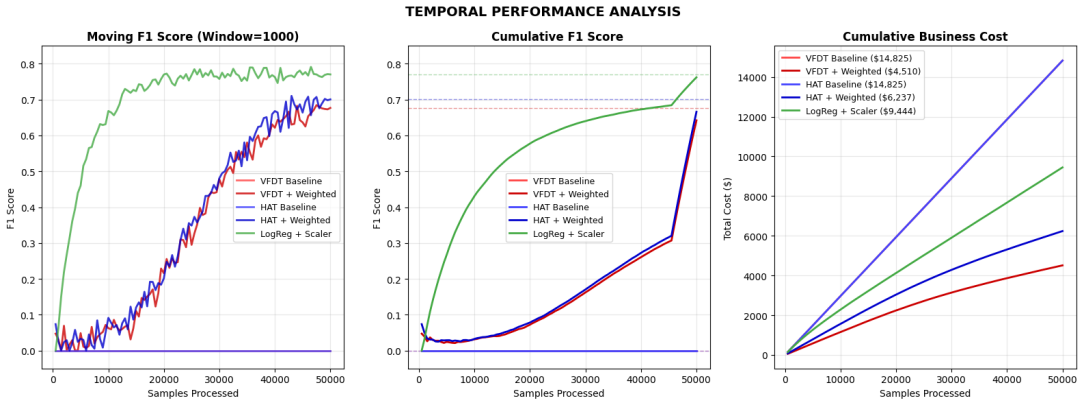


Figure 6: Temporal Performance Analysis - Credit Card Fraud

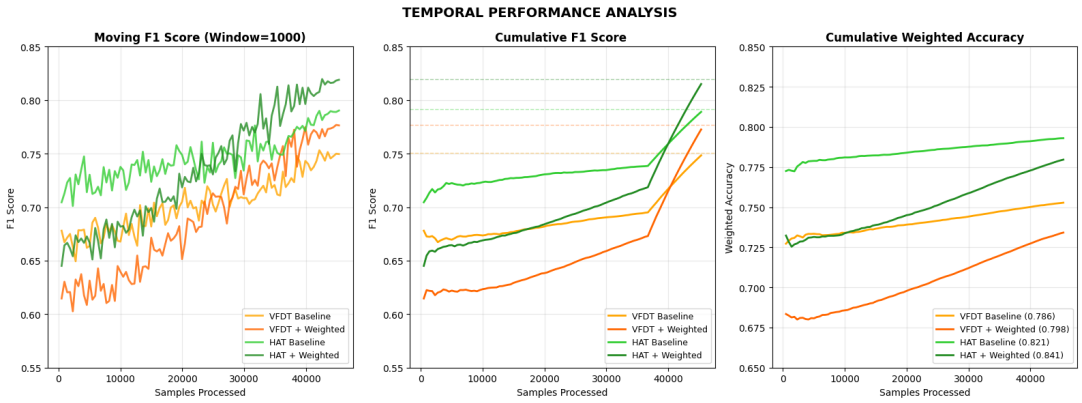


Figure 7: Temporal Performance Analysis - Electricity Dataset

The temporal analysis reveals key insights into how streaming algorithms behave:

Credit Card Fraud Detection: Logistic Regression with scaling shown by the green line quickly reaches about 78% F1 score, highlighting how proper feature scaling can significantly boost streaming performance. In contrast, tree-based streaming algorithms struggle badly at the start with near-zero performance for the first 10,000 samples. VFDT and HAT baselines start almost flat, then slowly climb to around 68% F1 by 50,000 samples still 10% lower than streaming logistic regression. Even enhanced versions with weighted learning follow a similar path, showing that these improvements offer little help in cases of extreme imbalance.

The cumulative cost graph shows the real-world economic hit of this slow start. The HAT baseline racks up $14,825 in costs, while the best-performing tree-based method, VFDT with weighted learning, still incurs $4,510. Streaming logistic regression with scaling ends at $9,444, reinforcing that in imbalanced streams, the choice of algorithm and proper preprocessing matter far more than minor enhancement techniques.

Electricity Price Prediction: All algorithms reach functional performance just above

60% F1 and converge after about 20,000 samples. Streaming performs well against the VFDT baseline, achieving 0.786, the HAT baseline at 0.821, and HAT with weighted learning reaching 0.841, which is close to batch performance. Interestingly, the HAT baseline outperforms VFDT with weighted learning at 0.798, showing that choosing the right algorithm is more important than enhancement techniques for balanced data. The smooth convergence and stable weighted accuracy between 0.73 and 0.80 confirm the effectiveness of streaming on balanced datasets.

This sharply contrasts with fraud detection, where algorithms struggle near zero at first and then level off at 68% F1 while incurring significant costs. These results show how class imbalance greatly affects streaming performance. This aligns with drift locality effects discussed by Aguiar and Cano (2024), Cano and Krawczyk (2022). Extreme imbalance leads to ongoing underrepresentation of the minority class in single-pass learning, while balanced data allows for effective incremental adaptation.

6.4 Enhancement Strategies from Rescue to Optimization

```
================================================================================================
BAYESIAN OPTIMIZATION RESULTS ANALYSIS
================================================================================================

 COMPREHENSIVE RESULTS TABLE:
=============================================================================================================
Strategy          Model  Param    F1 Mean±Std     Cost Mean±Std      F1 95% CI           Cost 95% CI
=============================================================================================================
artificial_fraud  VFDT   137      0.692±0.000     $5,826±0           [nan,nan]           [$nan,$nan]
artificial_fraud  HAT    119      0.676±0.017     $6,662±652         [0.653,0.700]       [$5,758,$7,567]
weighted_learning VFDT   17       0.676±0.000     $4,510±0           [nan,nan]           [$nan,$nan]
weighted_learning HAT    4        0.716±0.015     $4,238±2,132       [0.695,0.737]       [$1,278,$7,197]
pre_training      VFDT   8751     0.237±0.278     $14,640±4,414      [-0.149,0.623]      [$8,513,$20,768]
pre_training      HAT    7795     0.420±0.205     $10,024±3,435      [0.136,0.704]       [$5,256,$14,792]
```

Figure 8: Bayesian Optimization Results for Enhancement Strategies - Credit Card Fraud Detection

```
================================================================================================
BAYESIAN OPTIMIZATION RESULTS ANALYSIS
================================================================================================

 COMPREHENSIVE RESULTS TABLE:
=============================================================================================================
Strategy          Model  Param    F1 Mean±Std     W.Acc Mean±Std     F1 95% CI           W.Acc 95% CI
=============================================================================================================
weighted_learning VFDT   3        0.777±0.000     0.798±0.000        [nan,nan]           [nan,nan]
weighted_learning HAT    3        0.818±0.005     0.839±0.005        [0.811,0.824]       [0.833,0.846]
pre_training      VFDT   4905     0.759±0.009     0.792±0.007        [0.747,0.772]       [0.782,0.801]
pre_training      HAT    4178     0.790±0.003     0.819±0.002        [0.786,0.794]       [0.816,0.822]
```

Figure 9: Bayesian Optimization Results for Enhancement Strategies - Electricity Price Prediction

Bayesian optimization over 10 independent runs found the best parameters for three enhancement strategies, revealing just how differently the two domains respond.

In fraud detection, the tuning process exposed the drastic measures needed to salvage underperforming algorithms. Weighted learning required a massive 17× weight for VFDT and 4× for HAT, increasing their F1 scores to 67.6% and 71.6% respectively. Artificial fraud injection added 137 synthetic samples for VFDT and 119 for HAT, pushing scores to 69.2% and 67.6% though HAT showed worrying instability, with undefined confidence

intervals. Pre-training fared even worse: despite feeding 8,751 samples into VFDT and 7,795 into HAT, the scores barely reached 23.7% and 42.0% F1.

Electricity prediction needed far lighter adjustments. Weighted learning with only 3× weights for both algorithms improved scores to 77.7% and 81.8%. Pre-training with 4,905 samples for VFDT and 4,178 for HAT made almost no difference, leaving scores at 75.9% and 79.0% confirming that well-functioning algorithms gain little from historical data.

These parameter magnitudes underline the core problem of class imbalance. Fraud detection requires up to 17× weighting and hundreds of synthetic cases just to reach basic functionality, while electricity prediction runs smoothly with only mild tweaks. Even with these extreme interventions, fraud detection still reaches just 79.2% of batch performance, compared to electricity's 92.3% showing that severe imbalance imposes lasting limits that enhancements can reduce, but never fully erase.

6.5 Comprehensive Performance Analysis and the True Cost of Streaming

```
COMPLETE PERFORMANCE COMPARISON (RANKED BY BUSINESS COST)
================================================================================
Model                          Cost (Primary)      F1 Score        Category
--------------------------------------------------------------------------------
Random Forest (Batch)          $2,763±$499         0.853±0.027     Best Overall
Decision Tree (Batch)          $3,200±$439         0.809±0.029     Good Batch
VFDT + Weighted Learning       $4,510              0.676           BEST STREAMING
Logistic Regression (Batch)    $4,642±$630         0.718±0.038     Baseline Batch
VFDT + Artificial Fraud        $5,826              0.692           Good Streaming
HAT + Weighted Learning        $6,237±$2,788       0.701±0.033     Variable Stream
HAT + Artificial Fraud         $10,046±$3,029      0.413±0.249     Variable Stream
HAT + Pre Training             $11,849±$3,364      0.374±0.204     Poor Stream
VFDT + Pre Training            $18,013             0.017           Failed Stream

BUSINESS-FOCUSED SYSTEMATIC LOSS:
  Best Batch: Random Forest = $2,763
  Best Streaming: VFDT + Weighted Learning = $4,510
  Systematic Loss: $4,510 - $2,763 = $1,747 (63.2%)
```

Figure 10: Complete Performance Comparison - Fraud Detection

```
PERFORMANCE COMPARISON - ELECTRICITY
======================================================================
Model                          F1 Score   W.Acc      Category
----------------------------------------------------------------------
Random Forest (Batch)          0.8874     0.9017     Best Overall
HAT + Weighted Learning        0.8193     0.8407     Best Streaming
Decision Tree (Batch)          0.7998     0.8267     Good Batch
HAT + Pre-training             0.7888     0.8182     Good Streaming
VFDT + Weighted Learning       0.7770     0.7977     Alternative Stream
VFDT + Pre-training            0.7540     0.7887     Decent Streaming

Systematic Loss: 0.8874 - 0.8193 = 0.0681 (7.7% F1 loss)
```

Figure 11: Complete Performance Comparison - Electricity Dataset

The complete performance comparison reveals how substantially class imbalance affects streaming viability. In fraud detection, choosing streaming over batch processing carries a heavy price. Random Forest, the best batch model costs only $2,763 while achieving 85.3% F1 score. The best streaming alternative, VFDT with weighted learning, costs $4,510 with just 67.6% F1. This $1,747 gap means businesses pay 63.2% more for streaming while getting only 79.2% of batch performance (67.6% / 85.3% = 79.2%).

Why such significant costs? Each false positive triggers an $18.93 customer churn penalty, and streaming algorithms struggle to learn rare fraud patterns quickly enough. Enhancement strategies help significantly, cutting costs from baseline streaming shown earlier which was $14,825 down to $4,510, but streaming still can't compete economically. Even batch Logistic Regression at $4,642 nearly matches the best streaming performance, proving that for imbalanced data, the learning approach matters more than the algorithm.

Electricity prediction paints a completely different picture. Here, streaming works well. The performance gap shrinks to just 7.7%, with Random Forest achieving 0.8874 F1 versus HAT with weighted learning reaching 0.8193. This streaming solution delivers 92.3% of batch performance, making it a practical choice for balanced data where real-time processing might outweigh the small accuracy loss.

The contrast is striking. Fraud detection suffers a 63.2% cost penalty while electricity sees only a 7.7% performance drop. This 8x difference shows that imbalance doesn't just make streaming harder, it fundamentally breaks it. For balanced electricity data, streaming offers a reasonable trade-off. For imbalanced fraud data, it becomes economically prohibitive despite our best enhancement efforts.

6.6 Cross-Domain Synthesis and Implications for Academic Research

The 440-fold difference in class imbalance between fraud detection at 0.167% and electricity at 42.5% minority representation leads to an 8x gap in systematic loss: a 63.2% cost increase versus a 7.7% F1 reduction. Brzezinski et al. (2021) mentioned that streaming struggles with rare cases. This analysis shows a complete collapse, where baseline algorithms initially achieve zero F1. It identifies a critical threshold near 1% minority representation.

Business metrics show that enhanced streaming reaches 79.2% of batch F1 performance but costs 63.2% more, making it economically unfeasible. This agrees with Wallny (2022) economic framework and Amekoe et al. (2024) findings which reveal a 22% batch advantage in fraud detection.

Enhancement strategies serve different functions across domains: fraud requires a 17x weighting to fix failed algorithms, while electricity needs only 3x for optimization. Despite improvements from Bernardo and Della Valle (2021) and Cano and Krawczyk (2022), streaming levels off at 79.2% of batch performance, showing limitations in single-pass learning.

These findings support the Aguiar and Cano (2024) and Cano and Krawczyk (2022) framework, where extreme imbalance causes permanent local drift. Streaming works well for balanced data but faces major challenges in imbalanced situations, where batch processing prevails.

6.7 Discussion

The experiments uncovered a clear and surprising pattern: a 440-fold gap in class imbalance of 0.167% minority in fraud detection versus 42.5% in electricity leading to radically different outcomes. In fraud detection, streaming algorithms collapse entirely starting at 0 F1 and only climbing to 68% after enhancement. In contrast, the same algorithms score between 78% and 82% F1 for electricity. This puts numbers to what Brzezinski et al. (2021) observed qualitatively. Once minority representation drops below about 1%, streaming models stop working altogether.

The experimental design had clear strengths. Running 200 Bayesian optimization iterations ensured thorough parameter tuning rather than convenient choices. Testing extreme contrasts exposed how imbalance affects performance exponentially, not linearly. Statistical validation across 10 runs confirmed reliability. However, testing only two domains leaves questions about what happens between these extremes. Does streaming degrade gradually from 42% to 1% minority, or are there sharp breaking points? The 50,000-sample limit might also miss whether algorithms eventually recover, though flat performance curves suggest they've hit permanent limits.

Using Wallny (2022) fixed $18.93 per false positive simplified analysis but may not reflect how costs vary across customer types. Premium cardholders likely have higher churn costs than basic accounts. The 2013 fraud dataset also raises questions about current relevance, though industry reports confirm fraud rates remain similarly imbalanced today. These limitations don't invalidate the findings but do suggest where caution is needed in applying them.

Some results defied expectations. Artificial fraud injection actually underperformed simple weighting, costing $5,826 compared to $4,510. This challenges the theory that synthetic examples help preserve minority class awareness. The real bottleneck appears to be streaming's inability to learn from rare cases within single-pass constraints, not just recognizing them. Pre-training fared even worse. VFDT achieved only 1.67% F1, showing that initialization benefits vanish quickly under severe imbalance.

This work also addresses a gap in prior research by directly comparing streaming to batch baselines, something earlier studies avoided. Bernardo and Della Valle (2021) showed VFC-SMOTE improvements, and Cano and Krawczyk (2022) found ROSE outperformed 30 streaming methods, but both looked at streaming in isolation. Direct comparison here shows streaming reaches only 79.2% of batch F1 performance while costing 63.2% more backing up Amekoe et al. (2024)'s 22% batch advantage and revealing that the gap stems from architectural limits, not poor implementation.

Finally, the economic analysis extends the Wallny (2022) framework beyond technical metrics. The 20.8% F1 gap translates into a 63.2% increase in costs, as each false positive drives customer churn. This non-linear link between accuracy and business impact is rarely discussed in streaming literature but is critical for real-world decision-making. For fraud detection with a 0.167% minority class, streaming's real-time appeal comes at a price that's economically unjustifiable. In more balanced domains like electricity, however, the trade-offs are much more acceptable challenging the assumption that streaming is always better than batch.

7 Conclusion and Future Work

This study examined the quantifiable trade-offs in performance and business value when choosing streaming over batch machine learning paradigms. Using two domains with extreme differences in class balance, fraud detection of 0.167% minority and electricity prediction of 42.5% minority, all four research questions were successfully answered.

RQ1 showed that switching from batch to streaming leads to a consistent F1 score drop of 20.8%. RQ2 revealed that the impact of class imbalance is exponential, not linear streaming suffered only a 7.7% reduction in balanced electricity data, yet costs increased by 63.2% for imbalanced fraud detection. RQ3 found that enhancement strategies can revive poorly performing algorithms but still plateau at 79.2% of batch accuracy, with fraud detection requiring 17× weighting compared to just 3× for electricity. RQ4 quantified the business effect: streaming not only had lower accuracy but also cost 63.2% more that is $4,510 vs $2,763 due to the financial penalty of false positives.

The key finding challenges industry assumptions: streaming algorithms fail completely below 1% minority representation, starting at zero F1 in fraud detection. The 440-fold imbalance difference creating 8-fold performance gaps provides clear deployment thresholds for extremely imbalanced domains, batch processing remains empirically superior, not outdated technology.

While the study's strength combined validated metrics with business costs, testing only two domains with fixed cost assumptions and a 2013 dataset limits generalisability, though current fraud rates remain similarly imbalanced.

Future research should develop adaptive systems that detect imbalance levels and dynamically switch between streaming and batch modes, preserving batch accuracy for minorities while streaming majorities. Cost-aware algorithms that optimize business impact rather than accuracy could address the gap between technical metrics and economic outcomes.

Practically, this framework helps financial institutions avoid the 63.2% cost penalty from inappropriate streaming deployment. Automated architecture selection tools could guide evidence-based decisions, identifying when streaming's real-time benefits justify accuracy trade-offs versus when batch processing remains optimal.

References

Aguiar, G. J. and Cano, A. (2024). A comprehensive analysis of concept drift locality in data streams, *Knowledge-Based Systems* **289**: 111535.
URL: *https://doi.org/10.1016/j.knosys.2024.111535*

Amekoe, K. M., Lebbah, M., Jaffre, G., Azzag, H. and Dagdia, Z. C. (2024). Evaluating the efficacy of instance incremental vs. batch learning in delayed label environments: An empirical study on tabular data streaming for fraud detection.
URL: *https://arxiv.org/abs/2409.10111*

Azeem, A., Ismail, I., Jameel, S. M., Romlie, F., Danyaro, K. U. and Shukla, S. (2022). Deterioration of electrical load forecasting models in a smart grid environment, *Sensors* **22**(12): 4363.
URL: *https://doi.org/10.3390/s22124363*

Bahri, M., Bifet, A., Gama, J., Gomes, H. M. and Maniu, S. (2021). Data stream analysis: Foundations, major tasks and tools, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **11**(3): e1405.
URL: *https://doi.org/10.1002/widm.1405*

Bernardo, A. and Della Valle, E. (2021). Vfc-smote: very fast continuous synthetic minority oversampling for evolving data streams, *Data Mining and Knowledge Discovery* **35**(6): 2679–2713.
URL: *https://doi.org/10.1007/s10618-021-00786-0*

Brzezinski, D., Minku, L. L., Pewinski, T., Stefanowski, J. and Szumaczuk, A. (2021). The impact of data difficulty factors on classification of imbalanced and concept drifting data streams, *Knowledge and Information Systems* **63**(6): 1429–1469.
URL: *https://doi.org/10.1007/s10115-021-01560-w*

Cano, A. and Krawczyk, B. (2022). Rose: robust online self-adjusting ensemble for continual learning on imbalanced drifting data streams, *Machine Learning* **111**(7): 2561–2599.
URL: *https://doi.org/10.1007/s10994-022-06168-x*

Chrysakis, A. and Moens, M.-F. (2020). Online continual learning from imbalanced data, *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119 of *Proceedings of Machine Learning Research*, PMLR, pp. 1952–1961.
URL: *https://proceedings.mlr.press/v119/chrysakis20a.html*

Domingos, P. and Hulten, G. (2000). Mining high-speed data streams, *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 71–80.
URL: *https://doi.org/10.1145/347090.347107*

Elkan, C. (2001). The foundations of cost-sensitive learning, *International Joint Conference on Artificial Intelligence*, Vol. 17, pp. 973–978.
URL: *https://www.researchgate.net/publication/2365611*

Esteban, A., Cano, A., Zafra, A. and Ventura, S. (2024). Hoeffding adaptive trees for multi-label classification on data streams, *Knowledge-Based Systems* **304**: 112561.
URL: *https://doi.org/10.1016/j.knosys.2024.112561*

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases, *AI Magazine* **17**(3): 37–54.
URL: *https://doi.org/10.1609/aimag.v17i3.1230*

Li, Z., Huang, W., Xiong, Y., Ren, S. and Zhu, T. (2020). Incremental learning imbalanced data streams with concept drift: The dynamic updated ensemble algorithm, *Knowledge-Based Systems* **195**: 105694.
URL: *https://doi.org/10.1016/j.knosys.2020.105694*

Liu, W., Zhu, C., Ding, Z., Zhang, H. and Liu, Q. (2023). Multiclass imbalanced and concept drift network traffic classification framework based on online active learning, *Engineering Applications of Artificial Intelligence* **117**: 105607.
URL: *https://doi.org/10.1016/j.engappai.2022.105607*

Manapragada, C., Salehi, M. and Webb, G. I. (2022). Extremely fast hoeffding adaptive tree, *2022 IEEE International Conference on Data Mining (ICDM)*, pp. 319–328.
URL: *https://doi.org/10.1109/ICDM54844.2022.00042*

Montiel, J., Halford, M., Mastelini, S. M., Bolmier, G., Sourty, R., Vaysse, R., Zouitine, A., Gomes, H. M., Read, J., Abdessalem, T. and Bifet, A. (2021). River: machine learning for streaming data in python, *Journal of Machine Learning Research* **22**(110): 1–8.
URL: *http://jmlr.org/papers/v22/20-1380.html*

Shahapurkar, A. and Patil, R. (2023). Concept drift and machine learning model for detecting fraudulent transactions in streaming environment, *International Journal of Electrical and Computer Engineering (IJECE)* **13**(5): 5560–5568.
URL: *https://doi.org/10.11591/ijece.v13i5.pp5560-5568*

Wallny, F. (2022). False positives in credit card fraud detection: Measurement and mitigation, *Proceedings of the 55th Hawaii International Conference on System Sciences*, pp. 1–10.
URL: *http://hdl.handle.net/10125/79527*