

Cost-Aware Adaptive Credit Card Fraud Detection Under Concept Drift: A Comparative Evaluation of Model-Maintenance Policies

Sailee Abhale

`sailee.abhale_24uds@sanjivani.edu.in`

Snjivani University

Om Bhadange

Sanjivani University

Samarth Gayake

Sanjivani University

Jagdish Malkar

Sanjivani University

Research Article

Keywords: credit card fraud detection, concept drift, adaptive re-training, ADWIN, Random Forest, cost-aware learning, imbalanced classification, streaming evaluation

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11398009/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Cost-Aware Adaptive Credit Card Fraud Detection

Under Concept Drift: A Comparative Evaluation of Model-Maintenance Policies

Sailee Abhale, Om Bhadange, Samarth Gayake, and Jagdish Malkar

Department of Artificial Intelligence and Data Science

Sanjivani University, Kopergaon, Maharashtra, India

sailee.abhale_24uds@sanjivani.edu.in, om.bhadange_24uds@sanjivani.edu.in, samarth.gayake_24uds@sanjivani.edu.in, jagdish.malkar_24uds@sanjivani.edu.in

Abstract— Credit card fraud detection combines severe class imbalance with transaction patterns that change over time, so a model trained on historical data can lose effectiveness. Retraining restores performance but costs computation, and fixed schedules need not coincide with actual change. This paper presents a comparative experimental evaluation of five model-maintenance policies for a Random Forest fraud classifier under controlled synthetic sudden concept drift: a static model, periodic retraining, ADWIN-triggered retraining, frequent retraining, and a multi-signal, cost-aware Adaptive Decision Engine (ADE). The ADE combines ADWIN error monitoring, prediction-confidence degradation and false-positive-rate deterioration with an economic gate and challenger validation. The experiment uses the European Credit Card Fraud Detection benchmark (284,807 transactions) with a chronological 70:30 split into 199,364 historical and 85,443 streaming transactions. We report precision, recall, F1-score, PR-AUC, false-positive rate, retraining count, training time, detection delay, sustained recovery time and a parameterised adaptation cost. In a single deterministic run, frequent retraining attained the highest F1-score (0.8173) with 17 retrainings; the ADE reached 0.8019 with four retrainings, and its first alert came 2,323 transactions after drift onset. Under the configured cost weights, the static and ADWIN-only policies had the lowest modelled cost (2,690 units) and the ADE did not minimise cost (2,780 units). The results illustrate a trade-off among detection quality, retraining frequency and assumed costs; they do not establish the superiority of any policy. Differences between policies correspond to only a few fraud transactions, and the study is limited by a single run, synthetic drift, immediate labels and assumed costs.

Index Terms— credit card fraud detection, concept drift, adaptive retraining, ADWIN, Random Forest, cost-aware learning, imbalanced classification, streaming evaluation

I INTRODUCTION

Credit card fraud detection is a highly imbalanced classification problem: fraudulent transactions are a tiny fraction of all transactions, so precision–recall-based evaluation is more informative than accuracy or ROC-based summaries [1]. Fraud behaviour also changes as fraudsters adapt and customer behaviour shifts, a phenomenon known as concept drift [2]. A classifier fitted to historical data may therefore degrade after de-

ployment [3].

For batch learners such as Random Forests [5], which cannot be updated one sample at a time, adaptation requires explicit retraining. Retraining has a cost, and the question of *when* to retrain has no obvious answer. Retraining on a fixed schedule may waste computation when the model is still adequate; triggering on a single statistical drift detector may miss changes that matter operationally, or fire when degradation is immaterial.

This paper compares five maintenance policies for one Random Forest fraud classifier under the same controlled streaming scenario. One of them, the Adaptive Decision Engine (ADE), combines three runtime signals with an economic gate and challenger validation. We do not claim that retraining-policy selection or ADWIN-based adaptation is unexplored, nor that the ADE is superior. The contribution is an integrated, reproducible comparison that measures classification performance, retraining effort, detection delay, sustained recovery and a parameterised cost together, and that reports where the proposed policy does *not* win.

Research questions. **RQ1:** How do the five policies differ in fraud-detection performance under the same sudden-drift scenario? **RQ2:** How do retraining frequency and training time vary? **RQ3:** How do the ADE’s detection and recovery behaviour compare with those of the baselines? **RQ4:** How does the cost ranking change if the assumed cost weights change?

Contributions. (i) A chronological streaming protocol with a controlled sudden-drift intervention; (ii) a side-by-side evaluation of five policies, including a three-signal, cost-gated ADE; (iii) metric definitions that separate drift alert, model update and sustained recovery; (iv) a candid account of the results, including that static and ADWIN-only policies have the lowest modelled cost under the chosen weights.

II RELATED WORK

Fraud detection under imbalance and drift. Dal Pozzolo *et al.* [3] model realistic fraud-detection conditions, including class imbalance, delayed supervision and concept drift, and motivate adaptive learning. The benchmark data used here was

introduced alongside work on probability calibration under undersampling [4]. Lebichot *et al.* [7] study incremental learning strategies for card fraud detection; our study instead concerns a non-incremental learner for which adaptation requires explicit retraining.

Drift detection. ADWIN [6] maintains an adaptive window over a stream and signals change when two sub-windows differ significantly, with a confidence parameter δ governing false-alarm tolerance. Surveys of concept-drift adaptation [2] describe detector-triggered and windowed strategies. In fraud detection, ROSFD [8] uses DDM, EDDM and ADWIN to trigger updates in a “train-only-when-required” scheme and reports that ADWIN performed best among the detectors it evaluated; it is a drift-only trigger and does not combine confidence or false-positive signals.

When to retrain. Regol *et al.* [9] cast retraining as a decision problem that trades off retraining cost against the cost of degraded performance, and show that drift detection alone cannot account for that trade-off. Dasari [10] compares periodic, error-threshold and ADWIN-triggered policies against a no-retrain baseline under budget and latency constraints; for learners without incremental updates, that study reports that policy choice strongly separates outcomes and that simple periodic retraining outperformed the reactive policies under abrupt and gradual drift, while reactive policies retained an advantage under recurring drift. Fujiwara *et al.* [11] observe that detecting drift does not by itself determine whether enough post-drift data exists for reliable retraining. Cost-sensitive evaluation more broadly is discussed by Elkan [12].

Positioning. The studies above treat retraining policy generally or use drift-only triggers. This paper evaluates a specific combination, ADWIN error evidence plus confidence-margin degradation plus false-positive-rate deterioration, with an economic gate, for a non-incremental Random Forest on the card-fraud benchmark, against static, periodic, ADWIN-only and frequent baselines. It is a small, controlled comparison rather than a general theory of retraining.

III DATASET AND EXPERIMENTAL PROTOCOL

A Data

We use the European Credit Card Fraud Detection benchmark: 284,807 transactions, of which 492 are fraudulent and 284,315 legitimate. Features are the anonymised principal components V1–V28, the transaction `Amount`, `Time`, and the binary `Class` label. The data cover a short historical period and do not form a naturally evolving long-term fraud stream; drift is therefore injected synthetically (Section E).

B Chronological partitioning

The first 70% of transactions form the historical training partition (199,364 transactions, 384 frauds) and the remaining 30% form the streaming evaluation partition (85,443 transactions, 108 frauds). The historical partition is split chronologically for model selection: the first 80% (159,491 transactions, 356 frauds) fits candidate classifiers and the final 20% (39,873 transactions, 28 frauds) validates them (Table 1).

Table 1: Data partitions

Partition	Transactions	Frauds
Model-selection training	159,491	356
Chronological validation	39,873	28
Historical training (full)	199,364	384
Streaming evaluation	85,443	108
Complete dataset	284,807	492

C Preprocessing

The raw `Time` feature is removed from the model inputs. `Amount` is standardised into `Scaled_Amount` with a `StandardScaler` fitted on historical training data only and then applied to the stream. The class imbalance of the stream is preserved; no oversampling is applied to it. The stream is processed sequentially in batches of 500 transactions, with performance monitored over a rolling window of 5,000 transactions.

D Base classifier

All five policies share one base learner, a Random Forest [5] with 100 trees, maximum depth 10, balanced class weights and random state 42. It was selected on the chronological validation split (validation PR-AUC 0.6560) and then refitted on all 199,364 historical transactions before streaming evaluation.

E Synthetic sudden drift

A sudden drift is injected into the stream over the stream-relative interval 34,177–51,265 (17,088 transactions; global coordinates 233,541–250,629). Within the interval, V14 is shifted by -2.0 , V4 by $+1.5$, and the scaled amount is multiplied by 2.0 . This is a controlled stress test; it is not evidence of naturally occurring or production-validated drift.

IV MAINTENANCE POLICIES

A Baselines

Static: the initial model is never retrained. *Periodic*: retrain every 10,000 transactions. *Frequent*: retrain every 5,000 transactions (a high-frequency scheduled baseline, not retraining after every batch). *ADWIN-only*: ADWIN with $\delta = 0.002$ monitors the prediction-error stream and triggers retraining on detected change.

B Adaptive Decision Engine

The ADE monitors three signals.

1. **Error and missed-fraud monitoring.** ADWIN monitors the streaming error signal and missed-fraud events.
2. **Confidence-margin degradation.** With $P(\text{fraud})$ the model’s predicted probability, the certainty is

$$c_t = 2 P(\text{fraud}) - 0.5, \quad c_t \in [0, 1]. \quad (1)$$

A decline in mean certainty beyond a configured threshold signals that predictions are moving towards the decision boundary.

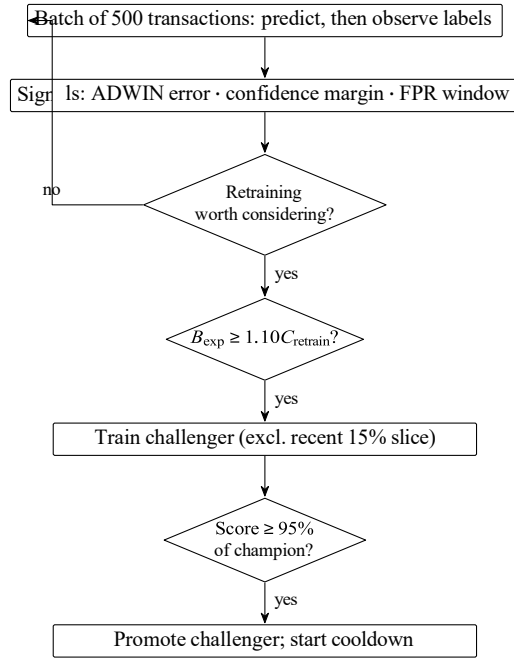


Figure 1: Decision flow of the ADE for each evaluated batch. A negative outcome at the economic gate or at challenger validation keeps the current model.

3. **False-positive-rate deterioration.** A recent 5,000-transaction window monitors changes in false alarms as an operational-friction signal.

The policy also uses a 5,000-transaction warm-up and a 5,000-transaction retraining cooldown. When the signals indicate that retraining is worth considering, an *economic gate* estimates the expected benefit

$$B_{\text{exp}} = 0.70 N_{\text{FN}} C_{\text{FN}} + 0.50 N_{\text{FP}} C_{\text{FP}}, \quad (2)$$

and retraining is approved only if $B_{\text{exp}} \geq 1.10 C_{\text{retrain}}$. A *challenger* model is then trained on historical data excluding a recent 15% validation slice, and it replaces the active (champion) model only if its validation score is at least 95% of the champion's. These are configured experimental rules, not thresholds claimed to be optimal for any banking environment.

V EVALUATION METRICS

With TP , FP , FN , TN the confusion-matrix counts over the streaming partition: precision = $TP/(TP + FP)$; recall = $TP/(TP + FN)$; $F1 = 2PR/(P + R)$; false-positive rate (FPR) = $FP/(FP + TN)$; PR-AUC is the area under the precision-recall curve. Retraining count and measured training time describe maintenance effort.

Detection delay is the first valid detection coordinate minus the drift onset. Scheduled policies do not detect drift, so no delay applies to them. **Sustained recovery time** is the start of the first qualifying streak minus drift onset, where a streak requires a post-drift drop below 95% of the pre-drift baseline F1 followed by three consecutive completed 500-transaction batches

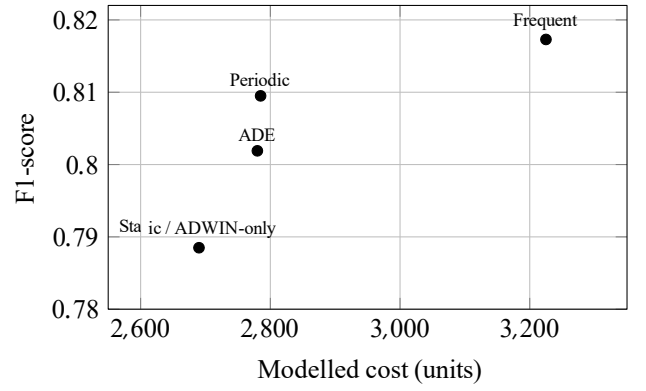


Figure 2: F1-score versus modelled adaptation cost. Static and ADWIN-only coincide.

at or above that threshold. A drift alert, a completed retraining and sustained recovery are distinct events and are reported separately.

Adaptation cost. Total modelled cost is

$$C_{\text{total}} = N_{\text{retrain}} C_{\text{retrain}} + N_{\text{FP}} C_{\text{FP}} + N_{\text{FN}} C_{\text{FN}}, \quad (3)$$

with $C_{\text{retrain}} = 50$, $C_{\text{FP}} = 5$ and $C_{\text{FN}} = 100$. These are relative experimental weights, not estimates of real monetary loss; costs are therefore reported in model units.

VI RESULTS

A Classification performance and cost

Table 2 reports all five policies on the same stream. All values are from a single deterministic run (random seed 42).

Frequent retraining attained the highest F1 (0.8173), precision (0.8500) and lowest FPR (0.000176), but also the largest number of retrains (17), the longest training time (175.44 s) and the highest modelled cost (3,225). Periodic retraining reached F1 0.8095 and the best PR-AUC (0.8105) with eight retrains (cost 2,785). The ADE reached F1 0.8019 and PR-AUC 0.8088 with four retrains (cost 2,780), which is 445 units below frequent retraining and 5 units below periodic retraining. The static and ADWIN-only policies produced identical classification results, never retrained, and had the *lowest* modelled cost (2,690), 90 units below the ADE.

Scale of the differences. The stream contains only 108 frauds. The ADE missed 25 of them against 23 for periodic and frequent retraining and 26 for the static model; the F1 differences among policies thus correspond to the handling of two or three fraud transactions. With one run and no confidence intervals, these differences should be read as descriptive, not statistically established. The 5-unit cost gap between the ADE and periodic retraining is immaterial.

Fig. 2 plots F1 against modelled cost; no policy is both best on F1 and lowest on cost.

B Drift detection and sustained recovery

The drift starts at stream coordinate 34,177. The ADE's first valid detection alert occurs at coordinate 36,500, a detection

Table 2: Five-policy comparison on the 85,443-transaction stream (108 frauds; single run). Cost in model units.

Policy	Precision	Recall	F1	PR-AUC	FPR	FP	FN	Retrains	Train time (s)	Cost
Static	0.8200	0.7593	0.7885	0.7998	0.000211	18	26	0	7.55	2,690
Periodic (10k)	0.8333	0.7870	0.8095	0.8105	0.000199	17	23	8	86.23	2,785
ADWIN-only	0.8200	0.7593	0.7885	0.7998	0.000211	18	26	0	7.50	2,690
Frequent (5k)	0.8500	0.7870	0.8173	0.8094	0.000176	15	23	17	175.44	3,225
ADE	0.8384	0.7685	0.8019	0.8088	0.000187	16	25	4	85.71	2,780

Table 3: Re-priced cost for alternative weights ($C_{\text{retrain}}, C_{\text{FP}}, C_{\text{FN}}$), observed counts fixed. Lowest per row in bold.

Weights	Static/ADWIN	Periodic	Frequent	ADE
(50, 5, 100) base	2,690	2,785	3,225	2,780
(5, 5, 100)	2,690	2,425	2,460	2,600
(200, 5, 100)	2,690	3,985	5,775	3,380
(50, 5, 20)	610	945	1,385	780
(50, 5, 500)	13,090	11,985	12,425	12,780
(50, 50, 100)	3,500	3,550	3,900	3,500

delay of $36,500 - 34,177 = 2,323$ transactions. The periodic and frequent policies are scheduled and have no detection delay. The ADE’s sustained-recovery onset is also at coordinate 36,500 (2,323 transactions after onset), coinciding with challenger promotion. Periodic and frequent retraining reached sustained recovery 10,323 transactions after onset under the same criterion.

The static and ADWIN-only policies showed a rolling-window threshold crossing at coordinate 35,500 (1,323 transactions after onset), attributed to a local cluster of four fraud transactions that the existing model detected. Because neither policy retrained and both degraded afterwards, this is a transient threshold crossing and not adaptation. The temporal coincidence between the ADE’s promotion and its recovery does not prove that the update caused the recovery; a single run cannot separate the effect of retraining from changes in sample difficulty or fraud clustering.

C Sensitivity of the cost ranking (RQ4)

Because C_{total} is a linear function of retrain, FP and FN counts, the ranking can be re-priced for other weights *holding the observed counts fixed*. This is a post-hoc re-pricing, not a new experiment: a policy such as the ADE would itself behave differently under other weights through its economic gate. Table 3 shows that the lowest-cost policy depends on the weights: cheap retraining or expensive missed fraud favours periodic retraining, while the base weights, cheap missed fraud and expensive retraining favour the static policies. The ADE is never strictly cheapest in these scenarios (it ties the static policies when false positives are weighted at 50).

VII DISCUSSION

Performance versus maintenance effort (RQ1, RQ2). Frequent retraining gave the best F1 but the highest effort and mod-

elled cost, so the highest classification score is not automatically the most economical policy. Retraining count and training time rose together: 0, 4, 8 and 17 retrainings corresponded to roughly 7.5, 85.7, 86.2 and 175.4 s. The ADE and periodic policies had similar training time in this run despite different retraining counts, so retraining count alone should not be read as a proxy for compute.

The ADE trade-off (RQ3). The ADE retrained half as often as periodic retraining and about a quarter as often as frequent retraining, at slightly lower F1, and it detected drift 2,323 transactions after onset. It did not minimise modelled cost, and the static policies were cheaper under the base weights. A plausible reading is that, in a single sudden-drift scenario where static performance degrades only modestly in absolute counts, the fixed retraining cost outweighs the avoided errors; this reading is not tested here.

Relation to prior work. Dasari [10] reports that, for non-incremental learners, simple periodic retraining outperformed reactive policies under abrupt drift and that ADWIN can be blind to some drift types. Our ADWIN-only policy never retrained and matched the static model, which is consistent with that observation, but we do not conclude that ADWIN is ineffective in general: only one δ and one scenario were run. Regel *et al.* [9] emphasise that the right retraining decision depends on the cost ratio, and Table 3 illustrates that dependence on this stream.

Detection is not recovery. The ADE’s alert, promotion and recovery onset coincided, but causality cannot be inferred from one run. Likewise, a threshold crossing without retraining (static, ADWIN-only) is not adaptation.

VIII LIMITATIONS AND THREATS TO VALIDITY

1. **Single deterministic run** (seed 42): no multi-seed means, deviations or confidence intervals.
2. **Synthetic sudden drift**: the injected perturbations are interventions on a historical benchmark and do not represent all real drift processes. Gradual, recurring and natural drift were not evaluated here.
3. **Immediate labels**: the simulation uses zero label lag; real fraud labels (investigations, chargebacks) arrive late, which affects any supervised monitoring signal.
4. **Assumed costs**: retraining, FP and FN weights are assumptions, not measured financial losses.

5. **Causal attribution:** the link between challenger promotion and recovery is temporal only.
6. **Small fraud count:** 108 frauds in the stream make policy differences a matter of a few transactions.
7. **External validity:** results on one public benchmark do not establish production performance. No ablation of the three ADE signals, the economic gate or the champion–challenger step is reported, so the individual contribution of each signal is not isolated.

IX CONCLUSION AND FUTURE WORK

We compared five model-maintenance policies for a Random Forest fraud detector under controlled sudden drift. Frequent retraining had the highest F1-score (0.8173) at 17 retrains and the highest modelled cost; the ADE had F1 0.8019 with four retrains and cost 2,780 units, detecting drift 2,323 transactions after onset; and the static and ADWIN-only policies had the lowest modelled cost (2,690) under the chosen weights. The findings show that classification performance, retraining effort and assumed costs should be evaluated together, and that the preferred policy depends on the cost weights. They do not establish that the ADE or any other policy is generally superior.

Future work includes multi-seed runs with confidence intervals; gradual, recurring and natural-drift scenarios; delayed labels; ablations of the confidence, FPR, ADWIN, cost-gate and replay components; re-running the ADE under alternative cost weights; counterfactual replay to estimate how much recovery is attributable to retraining; and additional datasets.

REPRODUCIBILITY

Settings: seed 42; Random Forest (100 trees, depth 10, balanced weights); batch size 500; rolling window 5,000; periodic and frequent intervals 10,000 and 5,000; ADWIN $\delta = 0.002$; cost weights (50, 5, 100). [Add code repository URL / archive DOI before submission.]

ACKNOWLEDGMENT

The authors thank the project guide for guidance. AI tools were used to assist drafting and structuring the written content; the experimental design, implementation and analysis were performed and verified by the authors.

REFERENCES

- [1] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” *PLoS ONE*, vol. 10, no. 3, p. e0118432, 2015.
- [2] J. Gama, I. Žliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia, “A survey on concept drift adaptation,” *ACM Comput. Surv.*, vol. 46, no. 4, art. 44, 2014.
- [3] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, “Credit card fraud detection: A realistic modeling and a novel learning strategy,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3784–3797, 2018.
- [4] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, “Calibrating probability with undersampling for unbalanced classification,” in *Proc. IEEE Symp. Series on Computational Intelligence (SSCI)*, 2015.
- [5] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [6] A. Bifet and R. Gavaldà, “Learning from time-changing data with adaptive windowing,” in *Proc. SIAM Int. Conf. Data Mining (SDM)*, 2007, pp. 443–448.
- [7] B. Lebichot, F. Paldino, W. Siblini, L. He-Guelton, F. Oblé, and G. Bontempi, “Incremental learning strategies for credit card fraud detection,” *Int. J. Data Sci. Anal.*, vol. 12, pp. 165–174, 2021, doi: 10.1007/s41060-021-00258-0.
- [8] V. Yelleti, “ROSFD: Robust online streaming fraud detection with resilience to concept drift in data streams,” arXiv:2504.10229, 2025.
- [9] F. Regol, L. Schwinn, K. Sprague, M. Coates, and T. Markovich, “When to retrain a machine learning model,” in *Proc. 42nd Int. Conf. Mach. Learn.*, PMLR vol. 267, 2025, pp. 51369–51404.
- [10] S. Dasari, “When to retrain: An empirical study of retraining policies for streaming ML under concept drift, budget, and latency constraints,” arXiv:2608.19488, 2026.
- [11] R. Fujiwara, Y. Matsubara, and Y. Sakurai, “When to retrain after drift: A data-only test of post-drift data size sufficiency,” in *Int. Conf. Learn. Representations (ICLR)*, 2026, arXiv:2603.09024.
- [12] C. Elkan, “The foundations of cost-sensitive learning,” in *Proc. 17th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2001, pp. 973–978.