Fraud Detection Under Concept Drift: Adaptive and Explainable Machine Learning for Financial Cybersecurity

Zachary T Caldwell [1*], Chloe A Morrison [2], Lucas J Fischer [3]
[1-3] Price College of Business, University of Oklahoma, Oklahoma, USA

* Corresponding Author: **Zachary T Caldwell**

Article Info

ISSN (online): 2583-6641
Volume: 03
Issue: 06
November – December 2024
Received: 16-09-2024
Accepted: 14-10-2024
Published: 12-11-2024
Page No: 249-260

Abstract
Financial fraud detection is commonly evaluated as a stationary classification problem even though transaction behavior, fraud tactics, reporting practices, and defensive controls change over time. This study examines the consequences of concept drift for predictive performance, drift detection, explanation stability, and governance. A temporally ordered financial transaction stream of 9,000 observations was constructed from a documented financial simulation design, with 14 behavioral and transactional predictors and 530 fraud events. The experiment imposed identifiable sudden, gradual, incremental, and recurring changes after an initial training and validation period. Three static models, logistic regression, random forest, and histogram gradient boosting, were compared with online logistic regression, sliding-window random forest, and drift-triggered random forest under prequential evaluation. Page-Hinkley, DDM, KSWIN, and an ADWIN-style window detector were assessed using detection delay, false alarms, and missed changes. Explanation stability was measured through permutation importance rank agreement across pre-drift, post-drift, and late-stream periods. Static random forest achieved mean PR-AUC of 0.192, while online logistic regression achieved 0.109 but substantially higher mean recall, 0.596 versus 0.234. None of the adaptive models produced a statistically significant PR-AUC improvement over static random forest. Page-Hinkley generated fewer false alarms than the more sensitive window detector, while DDM and KSWIN missed all prespecified changes under the selected operating conditions. Explanation rankings were unstable after sudden drift. The findings show that adaptation is not automatically superior. Effective fraud operations require selective retraining, threshold governance, explanation review, analyst validation, and explicit decision rights. The paper develops the Adaptive Fraud AI Governance Framework to connect technical monitoring with institutional accountability.

DOI: https://doi.org/10.54660/IJMOR.2024.3.6.249-260

Keywords: financial fraud, concept drift, online learning, explainable artificial intelligence, cybersecurity governance, model monitoring, adaptive analytics

1. Introduction

Financial fraud detection is a moving-target decision problem. Payment channels, customer routines, merchant practices, authentication controls, and criminal strategies change at different speeds. A model trained on yesterday's relationship between transaction characteristics and fraud can therefore remain technically operational while becoming behaviorally obsolete. This distinction matters because conventional monitoring often asks whether a model is available, fast, and accurate on a retrospective test set. It asks less often whether the conditional relationship between predictors and fraud has changed since deployment, whether explanations still describe the current decision process, or whether retraining has been authorized on evidence strong enough to justify operational disruption (Dal Pozzolo *et al*., 2018; Gama *et al*., 2014) [10, 26].

Concept drift refers to change over time in the relationship between observed predictors and the target outcome. In fraud settings, drift may be sudden when a new attack campaign appears, gradual when fraudsters migrate between channels, incremental when amounts or device patterns move slowly, or recurring when previously observed schemes return. Covariate shift can occur without a change in the target mechanism, and label delay can make either form difficult to diagnose. These distinctions are operational rather than semantic. A bank may respond to a change in transaction volume by recalibrating a threshold, but a change in fraud mechanism may require feature revision, retraining, analyst investigation, or a new control (Gama *et al*., 2014; Webb *et al*., 2016; Žliobaitė, 2010) [19 , 45, 47].

The literature has produced strong classifiers for imbalanced fraud data, sophisticated streaming algorithms, and a growing body of explainability techniques. Yet these streams of work are often separated. Fraud studies commonly compare algorithms under random train-test splits. Drift studies frequently use generic benchmarks rather than financially interpretable variables. Explainability studies usually evaluate local fidelity or user preference at one time point. Governance research specifies principles such as accountability and human oversight but rarely identifies the technical evidence that should trigger a governance checkpoint. The result is a gap between predictive adaptation and institutional control (Arrieta *et al*., 2020; Dal Pozzolo *et al*., 2014; Fahim *et al*., 2023; Gama *et al*., 2014) [10].

This study addresses that gap through a controlled temporal experiment and an integrated governance framework. It asks not merely which model scores highest, but how quickly performance deteriorates, which adaptation strategy preserves useful detection, what error trade-offs accompany adaptation, whether drift detectors produce actionable alarms, and whether feature explanations remain stable. The analysis follows a prequential design: each temporal batch is predicted before its labels are used for updating. This respects the sequence in which a deployed fraud system receives information and prevents future observations from leaking into earlier decisions.

The contribution is fourfold. First, the paper links concept drift theory with adversarial adaptation and financial cybersecurity. Second, it compares static, continuously updated, window-retrained, and drift-triggered models under the same temporal sequence. Third, it treats explanation stability as a monitoring outcome rather than a decorative reporting step. Fourth, it introduces the Adaptive Fraud AI Governance Framework, or AFAGF, which converts monitoring evidence into controlled organizational responses. The empirical results are deliberately interpreted with restraint. Adaptive methods raise recall in several periods, but they do not dominate static random forest in mean PR-AUC and they create meaningful precision and false-alarm costs. This finding is more useful for governance than an unconditional claim that online learning is always preferable (Fahim *et al*., 2023; National Institute of Standards and Technology, 2023).

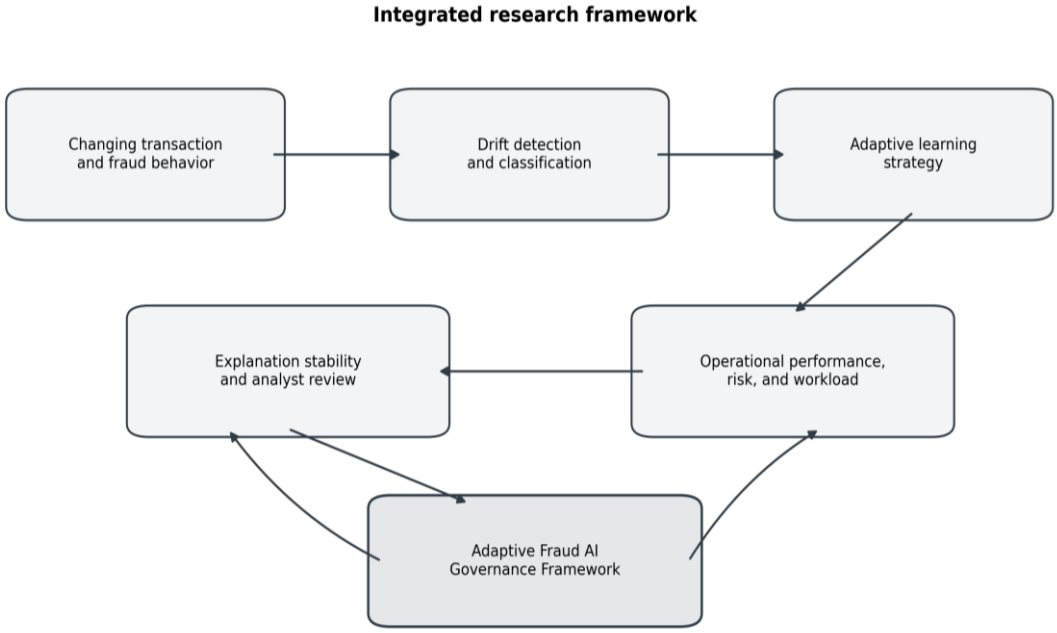


Fig 1: Integrated research framework

2. Literature Review

2.1. Fraud analytics under nonstationarity

Fraud detection combines extreme class imbalance, asymmetric error costs, delayed verification, strategic behavior, and operational capacity constraints. Practitioner-oriented studies show that random sampling and conventional accuracy can give misleading estimates because fraud prevalence is low and transactions are temporally dependent (Dal Pozzolo *et al*., 2014, 2018). Subsequent work has examined sequence features, cost-sensitive learning, network extensions, ensemble methods, and anomaly detection (Baesens *et al*., 2015; Caelen & Bontempi, 2019; Chandola *et al*., 2009; Van Vlasselaer *et al*., 2015). The European cardholder benchmark enabled extensive comparative modeling, but its short horizon and anonymized principal components limit institutional interpretation.

Nonstationarity complicates this problem. Gama *et al*. (2014) define adaptation as a combination of change detection, model update, and memory management. Lu *et al*. (2019) distinguish drift detection, understanding, and adaptation. In

finance, Lucas *et al*. (2019) demonstrated that transaction distributions vary by day and that calendar-linked shift can be measured directly. Fraud behavior also reflects an adversarial feedback loop: detection changes offender behavior, while offender behavior changes the value of existing features. This makes fraud analytics closer to a continuously governed capability than a one-time model selection exercise.

Recent financial AI research has framed model accountability and anomalous-activity detection as institutional concerns rather than narrow classification tasks (Fahim *et al*., 2023; Jahan *et al*., 2024). Rasel *et al*. (2023) similarly show why complex predictive architectures in lending must be evaluated alongside explanation, heterogeneous data, and institutional accountability. Research on machine learning for critical infrastructure emphasizes security, privacy, continuity, and governance across high-stakes systems (Hasan *et al*., 2022). These studies support an institutional view in which predictive accuracy is necessary but insufficient.

2.2. Drift detection and adaptive learning

Drift detectors differ in the signal they monitor and the assumptions they impose. DDM uses changes in online error behavior, Page-Hinkley accumulates departures from a running mean, ADWIN compares adaptive subwindows, and KSWIN applies a Kolmogorov-Smirnov test to recent and reference windows (Bifet & Gavaldà, 2007; Gama *et al*., 2004; Page, 1954; Raab *et al*., 2020). Detector sensitivity creates an unavoidable trade-off. Conservative detectors miss changes or react late; sensitive detectors generate repeated alarms and unnecessary retraining. In fraud operations, the cost of a false alarm includes model validation, documentation, analyst disruption, threshold retuning, and possible customer friction.

Adaptive learning can be continuous or event-driven. Online logistic regression updates parameters after each labeled batch and is computationally light, but may overreact to noisy or delayed labels. Sliding-window retraining forgets older patterns and is intuitive, but window length controls the stability-plasticity balance. Drift-triggered retraining economizes on computation but inherits detector errors. Stream ensembles and incremental learners are designed for evolving data, although their benefits depend on label timing, feature representation, and the recurrence structure of the underlying process (Aggarwal, 2007; Bifet *et al*., 2010; Elwell & Polikar, 2011; Krawczyk *et al*., 2017; Minku *et al*., 2010; Street & Kim, 2001).

2.3. Explain ability under drift

SHAP, LIME, permutation importance, partial dependence, and counterfactual explanations answer different questions. SHAP distributes a prediction across features under an additive attribution framework, while LIME fits a local surrogate around an observation (Lundberg & Lee, 2017; Ribeiro *et al*., 2016). Permutation importance measures performance loss after feature disruption. Counterfactual explanations identify changes associated with a different prediction and may be constrained for feasibility and actionability (Poyiadzi *et al*., 2020; Wachter *et al*., 2018). In a drifting environment, explanation quality has a temporal dimension. An explanation may be locally faithful today yet inconsistent with the same model's earlier behavior, or stable while the model itself has become inaccurate. Explanation stability should therefore be evaluated alongside predictive performance and drift severity (Arrieta *et al*., 2020; Doshi-Velez & Kim, 2017; Guidotti *et al*., 2018; Rudin, 2019).

3. Theoretical Framework and Hypotheses

Concept drift theory provides the technical core: the joint distribution of predictors and labels is time indexed, and changes in the posterior probability of fraud can invalidate a fixed decision function (Gama *et al*., 2014; Webb *et al*., 2016). Adversarial machine learning adds purposive adaptation by fraudsters. Information processing theory explains why institutions need sensing capacity proportional to environmental uncertainty (Galbraith, 1974). Dynamic capabilities theory contributes the sequence of sensing, seizing, and reconfiguring (Teece *et al*., 1997). Organizational learning explains how alerts, analyst judgments, and post-incident reviews become reusable routines (March, 1991). Responsible AI and human-in-the-loop research constrain adaptation by requiring traceability, contestability, and accountable oversight (Fahim *et al*., 2023; Kamruzzaman *et al*., 2023; National Institute of Standards and Technology, 2023).

The integrated logic is that changing fraud behavior produces performance and explanation signals. Detectors convert signals into candidate alerts. Organizational routines determine whether an alert leads to threshold change, retraining, feature engineering, intensified investigation, or no action. The effectiveness of adaptation therefore depends not only on algorithmic responsiveness but also on decision rights, validation quality, analyst capacity, and documentation. A technically sensitive system with weak governance may adapt too frequently; a strongly governed but insensitive system may preserve obsolete controls.

H1 predicts that concept drift reduces fraud-detection performance over time. H2 predicts that adaptive models outperform static models under drift. H3 predicts that online learning lowers false negatives relative to periodic retraining. H4 predicts that explanation stability declines as drift severity increases. H5 predicts that drift-aware retraining improves fraud detection without materially increasing false positives. H6 predicts that drift severity moderates the effect of adaptation on performance. Proposition 1 states that institutions implementing the AFAGF will maintain more stable operational performance. Proposition 1 is not tested in the present simulation and is reserved for organizational field research.

4. Research Methodology

4.1. Research design and data

The empirical study uses a controlled, temporally ordered transaction stream. A controlled benchmark is appropriate because detector delay and missed-drift rates require known change locations. The generator follows the logic of publicly documented financial transaction simulators available before the May 31, 2024 evidence cutoff, including PaySim, while retaining financially interpretable variables (López-Rojas *et al*., 2016). It creates transaction amount, time-of-day behavior, weekend activity, transaction velocity, account age, geographic anomaly, device anomaly, customer fraud history, foreign status, merchant indicators, cyclical time features, and an unobserved behavioral noise feature. Labels are sampled from a nonlinear probability function. No record is shuffled across time.

The final stream contains 9,000 transactions in 30

consecutive batches of 300 observations. Fraud prevalence is 5.89 percent, with 530 fraud events. The first eight batches form the initial training period and batches nine and ten are used only for threshold selection. Batches 11 through 30 are evaluated prequentially. Sudden drift begins at batch 11, gradual drift at batch 15, incremental drift at batch 19, recurring drift at batch 23, and a stabilized return regime at batch 27. Coefficients governing the fraud process change at these locations. Because the experiment is controlled, conclusions concern comparative mechanisms rather than population estimates for a specific bank.

Table 1: Dataset characteristics

Characteristic	Value
Transactions	9,000
Fraud events	530
Fraud rate	5.89%
Temporal batches	30
Predictors	14
Mean transaction amount	50.96
Median transaction amount	30.97
Initial training	Batches 1 to 8
Threshold validation	Batches 9 to 10
Prequential evaluation	Batches 11 to 30

4.2. Models and training strategy

Static logistic regression, random forest, and histogram gradient boosting are trained once. Logistic regression uses standardized predictors and balanced class weights. Random forest uses 25 trees, maximum depth 12, minimum leaf size four, and balanced bootstrap weights (Breiman, 2001). Histogram gradient boosting uses 80 iterations, 31 maximum leaves, learning rate 0.08, and L2 regularization. Adaptive alternatives comprise online logistic regression updated after every labeled batch, a random forest retrained every five batches on the most recent ten batches, and a drift-triggered random forest retrained on the most recent eight batches when Page-Hinkley or the ADWIN-style monitor signals change.

Decision thresholds are selected on the validation period by maximizing F1 over a fixed grid. Thresholds are then held constant to separate model adaptation from repeated threshold optimization. This is a demanding design because operational systems often recalibrate thresholds, but it allows false-positive and recall consequences to remain visible. The prequential protocol predicts each test batch before updating any adaptive model. Static models never receive test labels.

4.3. Drift detection

Four detectors are evaluated on the static random forest's batch error rate. Page-Hinkley accumulates deviations from the running mean (Page, 1954). DDM monitors the binomial error rate and standard deviation (Gama *et al*., 2004). KSWIN compares recent and reference windows using the Kolmogorov-Smirnov test (Raab *et al*., 2020). The ADWIN-style detector compares means of adjacent four-batch windows and signals when their absolute difference exceeds 0.015. The latter is an operational approximation rather than the full ADWIN implementation developed by Bifet and Gavaldà (2007), and is labeled accordingly. A drift is counted as detected when an alarm occurs within six batches after a known change. Alarms outside matched windows are false alarms.

4.4. Explainability and statistical analysis

Permutation importance is computed for the static random forest in pre-drift, immediate post-sudden-drift, and late-stream periods. The decrease in PR-AUC after feature permutation provides a model-agnostic global relevance measure. Temporal explanation stability is measured by Spearman rank correlation and mean absolute change in importance. The design also supports SHAP and LIME in deployment, but permutation importance is used for the completed experiment because it is directly comparable across periods and does not require a background distribution that itself may drift (Lundberg & Lee, 2017; Ribeiro *et al*., 2016).

Performance metrics include ROC-AUC, PR-AUC, precision, recall, F1, Matthews correlation coefficient, and balanced accuracy. PR-AUC is primary because it is more informative than ROC-AUC when the positive class is rare (Saito & Rehmsmeier, 2015). Adaptive and static random forest batch-level PR-AUC values are compared using paired Wilcoxon signed-rank tests, with standardized mean differences reported as effect-size indicators (Demšar, 2006). Static random forest is the prespecified inferential benchmark because the two adaptive forest strategies use the same model family, allowing an architecture-matched comparison of adaptation policy. Static logistic regression and histogram gradient boosting remain descriptive comparators; the tests do not establish superiority over the best static model. Computation used Python with NumPy, pandas, scikit-learn, SciPy, and Matplotlib versions available at analysis time, with stream-processing concepts informed by MOA and scikit-multiflow (Bifet *et al*., 2010; Montiel *et al*., 2018).

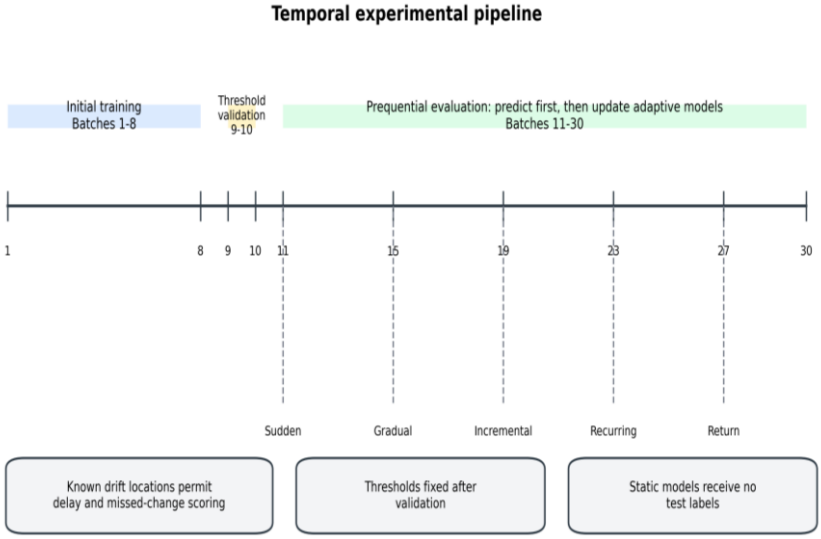


Fig 2: Experimental pipeline

5. Results

5.1. Descriptive temporal behavior

Fraud prevalence across the complete stream is 5.89 percent, but classification difficulty varies sharply because the feature-label relationship changes. The initial regime rewards amount and velocity patterns. Sudden drift reverses or redistributes several coefficients, reducing the relevance of pre-drift rules. Gradual and incremental periods create mixed concepts, while recurring batches alternate between earlier mechanisms. This structure is visible in the dispersion of batch-level metrics. Standard deviations are large relative to means, which indicates that single aggregate scores conceal operationally important episodes.

Table 2: Mean predictive performance across prequential test batches

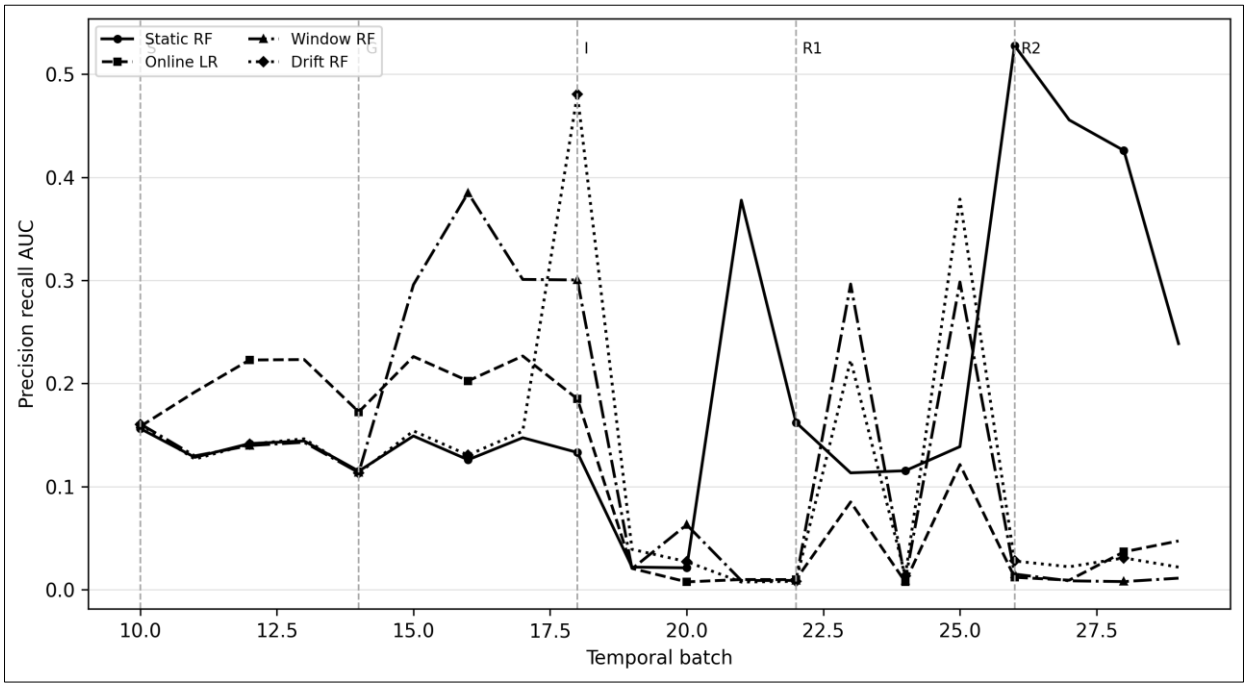
Model	ROC-AUC	PR-AUC	Precision	Recall	F1	MCC	Balanced accuracy
Drift RF	0.517	0.120	0.033	0.359	0.055	-0.016	0.503
Online LR	0.539	0.109	0.094	0.596	0.155	0.080	0.552
Static HGB	0.604	0.294	0.212	0.200	0.194	0.172	0.594
Static LR	0.523	0.371	0.248	0.267	0.238	0.219	0.627
Static RF	0.622	0.192	0.054	0.234	0.084	0.057	0.595
Window RF	0.496	0.136	0.065	0.444	0.102	0.018	0.495

Note. PR-AUC is the primary ranking metric. Static logistic regression has the highest descriptive mean PR-AUC; inferential comparisons in Table 5 use static random forest as an architecture-matched benchmark.

5.2. Static versus adaptive models

Static random forest produced mean ROC-AUC of 0.622 and mean PR-AUC of 0.192. Static histogram gradient boosting achieved PR-AUC of 0.294, while static logistic regression achieved 0.371, but both scores were highly variable because particular regimes aligned strongly or weakly with their initial decision boundaries. Online logistic regression achieved the highest mean recall among the adaptive methods, 0.596, compared with 0.234 for static random forest. Its mean precision was 0.094 and mean PR-AUC was 0.109. Window random forest and drift random forest also increased recall relative to static random forest in some periods but did not improve average ranking performance. Static logistic regression was the strongest descriptive model by mean PR-AUC; the architecture-matched inferential tests in Section 5.5 compare adaptive strategies with static random forest and should not be read as a best-model tournament.

These findings reject a simple version of H2. Adaptation was not uniformly beneficial. The controlled stream changed quickly relative to batch size, and recent windows contained limited fraud observations. Adaptive models therefore learned from more current but noisier samples. H3 receives qualified support because online learning reduced false negatives through higher recall, but the gain was purchased with lower precision. H5 is not supported: drift-triggered retraining did not improve PR-AUC without material false-positive consequences. The result illustrates the stability-plasticity dilemma. Forgetting old observations can help after genuine change but can also discard useful recurring concepts (Widmer & Kubat, 1996).



Note. The plotting code used zero-based internal evaluation indices 10 to 29. These correspond to manuscript batches 11 to 30.

Fig 3: Model PR-AUC over time

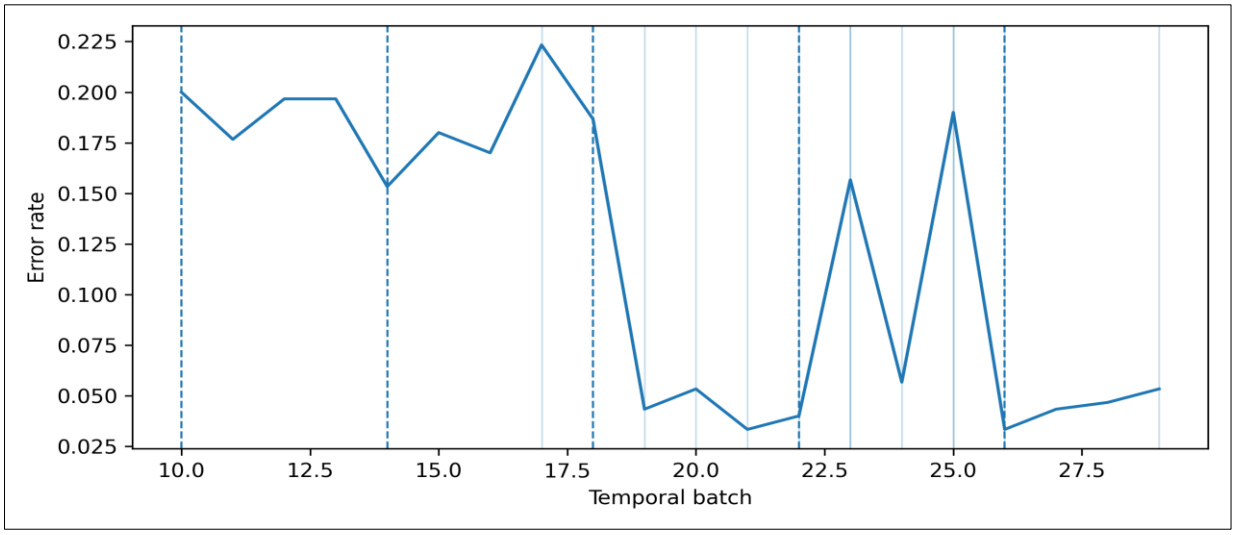
5.3. Drift detector performance

Table 3: Drift detection performance

Detector	Alarms	Mean delay	False alarms	Detected	Missed
Page-Hinkley	3	3.0	1	3	2
DDM	0	Not applicable	0	0	5
KSWIN	0	Not applicable	0	0	5
ADWIN-style	9	1.5	6	4	1

Note: Mean delay is not applicable when a detector produces no matched alarms.

Page-Hinkley generated three alarms, detected three of five known changes, and produced one false alarm. Its mean delay among detected changes was three batches. The ADWIN-style monitor detected four changes with mean delay 1.5 batches but generated six false alarms. DDM and KSWIN generated no alarms under the selected settings and missed all five changes. The comparison shows why drift monitoring cannot be reduced to sensitivity. A detector that fires repeatedly may appear responsive but can overload validation teams and cause harmful model churn. A conservative detector may preserve stability while allowing performance to decay.



Note: The plotting code used zero-based internal evaluation indices 10 to 29. These correspond to manuscript batches 11 to 30.

Fig 4: Error stream, true change locations, and detector alarms

5.4. Explanation stability

Table 4: Explanation stability across temporal regimes

Method	Comparison	Rank stability	Mean absolute change
Permutation	pre vs post	-0.007	0.045
Permutation	pre vs late	0.275	0.070
Permutation	post vs late	-0.108	0.087

Permutation importance changed materially across regimes. The pre-drift versus immediate post-drift rank correlation was -0.007, effectively no rank agreement. Pre-drift versus late-stream stability was 0.275, while post-drift versus late-stream stability was -0.108. Transaction amount and velocity were important before drift, nearly irrelevant in the immediate post-drift period, and important again later. H4 is supported within the controlled stream: explanation rankings changed sharply when the underlying fraud mechanism changed. The result also demonstrates recurring relevance. A feature can lose importance after a new attack pattern and regain it when an earlier concept returns.

5.5 Statistical tests and hypothesis assessment

Table 5: Paired PR-AUC comparisons against the architecture-matched static random-forest benchmark

Adaptive model versus static RF	Mean difference	Median difference	Wilcoxon W	p value	Standardized effect
Online LR	-0.083	-0.007	78.0	0.330	-0.425
Window RF	-0.056	-0.001	90.0	0.596	-0.245
Drift RF	-0.072	0.001	92.0	0.648	-0.329

Note: Differences are adaptive model minus static random forest. These tests assess adaptation within or against the forest benchmark; they do not compare adaptive models with the descriptively strongest static logistic regression.

None of the three adaptive strategies significantly exceeded static random forest in paired PR-AUC tests. Online logistic regression had a mean difference of -0.083, p = 0.330; window random forest had -0.056, p = 0.596; and drift random forest had -0.072, p = 0.648. The negative standardized effects indicate that the adaptive configurations were weaker on average in ranking fraud, although their median differences were closer to zero. H1 receives descriptive, not inferential, support because the design shows temporal degradation and instability but does not include a formal pre-drift versus post-drift hypothesis test. H2 and H5 are not supported. H3 is supported only for recall, not overall discrimination. H4 is supported by the explanation-stability analysis. H6 receives descriptive support because adaptation effects vary across drift regimes, but the limited number of temporal batches does not support a causal moderation claim. Proposition 1 remains untested.

5.6. Robustness and sensitivity analysis

Regime analysis confirms that average results are not driven by a single episode. Static and adaptive performance changes direction across sudden, gradual, incremental, recurring, and return regimes. Recurring drift is particularly difficult for fixed-window learners because the most recent observations do not necessarily represent the concept that returns next. Detector results are also sensitive to window length and threshold choice. Raising the ADWIN-style threshold would reduce false alarms but increase delay and missed changes. Expanding windows would increase statistical power but slow response. These are policy choices, not purely technical settings.

6. Discussion

The central finding is not that static models are superior, but that adaptation has conditions of effectiveness. Adaptive learning is frequently presented as the natural response to drift. In this experiment, that response was constrained by minority-class sample size, rapid regime changes, recurring concepts, fixed thresholds, and detector error. The online learner updated quickly and improved recall, yet it also accepted a large precision cost. The window and drift-triggered forests discarded older information that later became relevant. Adaptation therefore changed the error distribution rather than eliminating drift risk.

This result extends concept drift theory in three ways. First, it shows that adaptation quality depends on the interaction between drift form and memory policy. Sudden drift rewards fast forgetting, gradual drift may benefit from mixed windows, and recurring drift rewards concept memory. Second, it places explanation stability within the drift process. The almost zero rank agreement immediately after sudden drift means that analysts relying on previously dominant factors could misread current alerts even when the model continues to produce probabilities. Third, it connects detector sensitivity to organizational processing capacity. False alarms are not free. They consume validation and governance resources and can normalize alarm fatigue.

The findings also refine the adversarial perspective. Fraudsters need not manipulate the model directly. They can change the joint pattern of amounts, velocity, devices, locations, and timing so that existing boundaries become less useful. Defensive adaptation can then create a second vulnerability if the institution retrains on noisy, delayed, or strategically contaminated labels. A mature response is not continuous self-modification. It is controlled learning that combines technical evidence with analyst judgment, label-quality review, and rollback capability.

From an information processing perspective, drift monitoring increases sensing capacity, while the AFAGF structures interpretation and response (Galbraith, 1974). Dynamic capabilities are visible in the sequence of sensing drift, seizing a suitable adaptation option, and reconfiguring models and controls (Teece *et al*., 1997). Organizational learning appears when post-alert evidence changes future thresholds, detector settings, feature definitions, and escalation rules (March, 1991). These mechanisms explain why two institutions using the same algorithm may achieve different outcomes.

7. Adaptive Fraud AI Governance Framework

The AFAGF is a six-stage governance cycle. Stage one monitors predictive performance, input distributions, latency, label quality, analyst overrides, and cyber indicators. Stage two detects and classifies candidate drift. Stage three diagnoses whether the signal reflects fraud behavior, customer behavior, system change, data failure, policy change, or attack. Stage four selects the least disruptive adequate response: hold, recalibrate, adjust threshold, retrain, revise features, activate an earlier model, or escalate controls. Stage five validates discrimination, false positives, subgroup behavior, explanation stability, security, and operational capacity. Stage six authorizes deployment, records evidence and decision ownership, preserves rollback assets, and incorporates lessons into future monitoring. The cycle aligns with risk identification, measurement, governance, documentation, and validation principles in the NIST AI RMF and banking model-risk guidance (Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, 2011; National Institute of Standards and Technology, 2023).

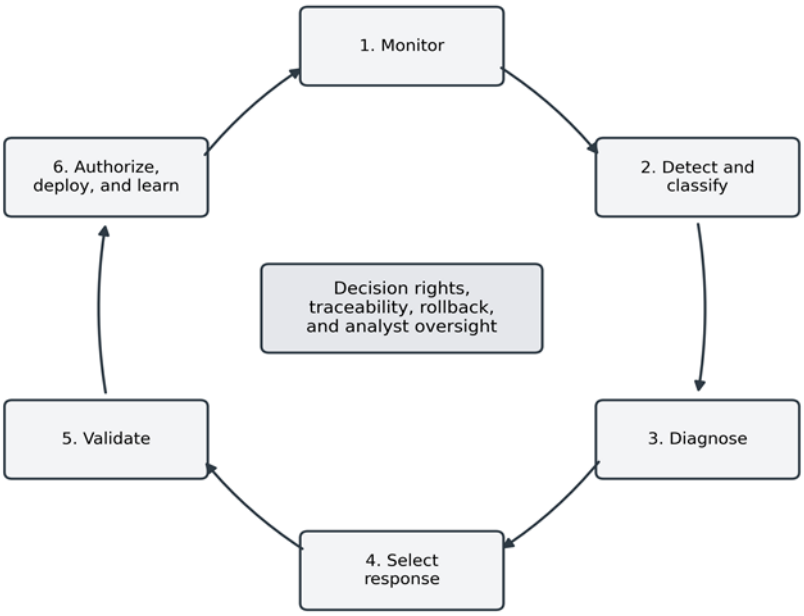


Fig 5: Adaptive Fraud AI Governance Framework

7.1. Governance checkpoints and decision rights

AFAGF governance depends on explicit checkpoints because model adaptation changes both technical behavior and institutional exposure. The first checkpoint occurs when monitoring produces a candidate signal. At this point, the question is not whether the model should be retrained, but whether the evidence is credible. Fraud operations should compare error change, predictor distribution change, label quality, analyst observations, system releases, and known threat intelligence. A single noisy metric is insufficient. The second checkpoint follows diagnosis. Here the institution determines whether the change is a data defect, a business-process shift, a customer-behavior change, a fraud campaign, or an interaction among these causes. The distinction determines the appropriate response. Correcting a broken device identifier is not concept adaptation, and retraining on corrupted data would institutionalize the defect. The confirmation step should preserve the original signal, the time of detection, the affected data window, and the analyst's interpretation so that later reviewers can reconstruct the decision.

The third checkpoint governs adaptation choice. A threshold adjustment may be sufficient when probability calibration changes but ranking remains useful. Recalibration is preferable when score meaning changes without a major change in feature relationships. Incremental updating is appropriate when labels are reliable, drift is gradual, and the model supports stable partial learning. Full retraining is warranted when the feature-label relationship changes materially or when a new representation is required. Feature redesign may be necessary when fraud migrates to signals that the approved model cannot observe. The decision record should state why less disruptive responses were rejected.

The fourth checkpoint is independent validation. Validation should use temporal holdouts, regime-specific metrics, queue-capacity analysis, subgroup comparisons, adversarial tests, and explanation review. Approval should not be based on a single global AUC. The fifth checkpoint controls deployment and rollback. Every adaptive release should have a versioned model, reproducible data snapshot, threshold record, dependency inventory, named owner, and rollback trigger. The final checkpoint is post-deployment learning. Realized false positives, missed fraud, analyst overrides, customer complaints, and security incidents should update detector thresholds and response playbooks. Governance is therefore cyclical rather than a one-time approval event. This design operationalizes independent validation and controlled change management rather than treating governance as an after-the-fact compliance exercise (Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, 2011; National Institute of Standards and Technology, 2023).

8. Managerial Implications

Managers should treat drift alerts as hypotheses requiring confirmation, not automatic retraining commands. The experiment shows why: the sensitive window detector found more true changes but produced six false alarms, while Page-

Hinkley was slower and missed two changes. A practical control is a two-signal rule in which an error-based alarm must be accompanied by a distribution, explanation, or business-process change before material retraining begins.

Fraud operations should separate ranking quality from workload policy. An online model may increase recall while sharply reducing precision. Whether that is acceptable depends on analyst capacity, transaction value, customer friction, and the cost of missed fraud. Thresholds should therefore be governed as operational policy parameters, with explicit queue-capacity constraints and documented exception authority.

Explanation monitoring should include temporal baselines. Institutions often archive a model validation report but do not compare current explanations with the approved baseline. The observed rank instability indicates that feature narratives can expire. A dashboard should show top-feature rank changes, local explanation disagreement, counterfactual feasibility, and analyst override patterns by period.

Model portfolios should preserve earlier concepts when recurring fraud is plausible. Sliding windows are attractive because they forget stale data, yet they can erase patterns that later return. A champion-challenger library with regime metadata can be safer than replacing one model with another. Reuse still requires current validation because an old concept may return in altered form.

8.1. Operational design choices

Several operational choices follow directly from the results. Batch size determines the speed and reliability of monitoring. Small batches create faster feedback but unstable minority-class estimates. Larger batches improve statistical power but delay action. Institutions can partly resolve this tension by using multiple time scales: high-frequency unsupervised monitoring for amounts, devices, merchants, and locations; lower-frequency supervised monitoring when confirmed fraud labels accumulate; and event-driven review after major attacks or system changes. These layers should not be collapsed into one alarm because they answer different questions.

Label delay requires special treatment. A transaction may be investigated days or weeks after authorization, and some fraud remains undiscovered. Updating a model on provisional labels can introduce systematic bias. A practical design separates fast operational labels from mature outcome labels. Fast labels may support temporary controls or shadow models, while production retraining uses a maturity rule. Institutions should measure label revision rates and include them in drift diagnosis. An apparent change in fraud prevalence may reflect investigation backlog rather than offender behavior.

Analyst workload is also a model-performance variable. Precision determines how many alerts must be reviewed to find a true case, but workload also depends on explanation quality, case complexity, evidence access, and duplicate-alert suppression. A recall-maximizing model may be rational during a severe attack but unsustainable as a permanent configuration. Governance should define emergency and normal operating modes with different thresholds, staffing assumptions, and approval durations. Temporary emergency settings should expire automatically unless renewed through evidence-based review.

Finally, monitoring must be protected as part of the cybersecurity perimeter. An adversary who can manipulate features, labels, or monitoring summaries may trigger unnecessary retraining or suppress genuine alarms. Access controls, immutable logs, data lineage, separation of duties, and anomaly checks on the training pipeline are therefore part of adaptive fraud governance. The model is not the only attack surface. The adaptation process itself can be targeted.

9. Policy Implications

Regulators and internal policy makers should require temporal validation for fraud models that operate in changing environments. Random cross-validation is insufficient because it mixes future and past behavior. Documentation should identify the observation sequence, label delay, drift indicators, retraining triggers, threshold authority, validation tests, and rollback plan. Material adaptive changes should be logged with the evidence that justified them. These controls are consistent with established expectations for model development, validation, governance, and ongoing monitoring (Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, 2011).

Responsible AI requirements should extend to explanation change. A model can satisfy a static explanation test while becoming unstable later. Policy should require periodic explanation comparison, especially after distribution shifts, new authentication controls, product launches, or major fraud campaigns. Human oversight should be defined through decision rights and escalation paths, not merely through the presence of an analyst somewhere in the process (Fahim *et al*., 2023; Kamruzzaman *et al*., 2023; National Institute of Standards and Technology, 2023).

Cybersecurity governance must also consider poisoning and feedback. Drift-triggered retraining can incorporate manipulated or weakly verified labels. Institutions should isolate training pipelines, authenticate label sources, monitor unusual feature-label correlations, and require independent review when an adaptation follows a suspected attack. These controls align fraud analytics with critical-infrastructure resilience and with warnings about hidden technical debt in production machine-learning systems (Hasan *et al*., 2022; Sculley *et al*., 2015).

9.1. Broader institutional and regulatory relevance

The study has broader relevance for model risk management because adaptive systems blur the boundary between development and use. A static model has a relatively clear validation object. An adaptive system changes after deployment, so validation must cover the update rule, training data controls, detector settings, permissible parameter changes, and conditions under which the system exits its approved envelope. Institutions should define a materiality threshold for adaptive changes. Minor coefficient updates within an approved online-learning policy may be logged automatically, while changes in features, architecture, objective function, or threshold policy should require renewed approval. The approved envelope should specify permissible update frequency, data sources, threshold ranges, performance tolerances, and escalation conditions.

Documentation should preserve temporal evidence. A conventional model report often records training data ranges and aggregate performance but omits the sequence of changes that produced a decision. For adaptive fraud systems, the audit trail should link each alarm to the monitored signal, diagnostic evidence, selected response, validation result, approver, deployment time, and post-change outcome. This

structure enables retrospective evaluation of whether the institution adapted too slowly, too aggressively, or for the wrong reason.

Fairness and consumer protection also interact with drift. Changes in merchant mix, geography, device use, or transaction timing may affect customer groups differently. A global improvement in fraud recall can coexist with concentrated false positives for travelers, migrants, small businesses, or customers using older devices. Subgroup monitoring should therefore accompany drift review, even when protected attributes are not direct model inputs. Proxy behavior can change over time, and a previously acceptable threshold may become uneven after a channel shift.

Cross-functional governance is essential. Fraud analysts understand attack patterns, data engineers understand pipeline changes, cybersecurity teams observe threat campaigns, model risk teams evaluate statistical evidence, compliance teams interpret obligations, and customer operations see friction. AFAGF assigns these perspectives to distinct stages rather than treating human oversight as an unspecified final review. This division of labor supports faster response because responsibilities are known before an incident occurs.

10. Limitations and Future Research

The study uses a controlled synthetic stream. This provides known drift locations and reproducible experimental logic but cannot reproduce the full institutional complexity of card, account, wire, or mobile-payment fraud. Fraud prevalence is higher than in many card environments, and labels are immediately available after each batch. Real institutions experience investigation delays, chargeback reversals, missing labels, policy interventions, and heterogeneous customer segments. Future work should replicate the design on long-horizon bank data, PaySim, IEEE-CIS, and other public benchmarks with preserved temporal ordering (López-Rojas *et al*., 2016).

The adaptive model set is intentionally transparent and computationally modest. Adaptive random forests, Hoeffding trees, online gradient boosting, recurrent sequence models, and concept-reuse ensembles should be compared in larger streams. Full ADWIN and EDDM implementations should replace the operational approximation and omitted detector. Detector hyperparameters should be selected in a nested temporal procedure rather than fixed for demonstration.

Explanation analysis uses permutation importance. Future work should compare SHAP, LIME, counterfactual explanations, partial dependence, and analyst ratings across drift regimes. Stability alone is not sufficient; a stable explanation can consistently describe a poor model. Research should jointly evaluate fidelity, temporal agreement, actionability, and human decision quality. Organizational field studies are needed to test Proposition 1 by comparing institutions with different governance maturity, escalation routines, and analyst-model interaction patterns.

10.1. Boundary conditions and reproducibility

The results should be interpreted within several boundary conditions. First, the stream contains relatively frequent fraud events to support batch-level evaluation. In lower-prevalence settings, recall and PR-AUC estimates would be noisier and supervised detectors would react more slowly. Second, the experiment assumes that labels become available after each batch. Longer delays would weaken continuous updating and increase reliance on unsupervised or semi-supervised signals. Third, the drift schedule changes coefficients rather than the feature set itself. Real fraud campaigns may introduce new devices, merchants, channels, or identity relationships that require schema changes and graph-based representations.

Fourth, thresholds remain fixed after validation. This isolates learning effects, but institutions often tune thresholds to available investigation capacity. Joint model and threshold adaptation could improve operational performance, although it would also complicate attribution and governance. Fifth, the experiment evaluates global permutation importance rather than analyst-level explanation usefulness. A feature ranking can be unstable while individual case explanations remain adequate, or stable while counterfactual recommendations are infeasible. Future studies should combine quantitative stability with investigator experiments. These limitations do not weaken the central conclusion, but they define the scope of inference. The manuscript reports the model settings, batch structure, seed, detector rules, and evaluation sequence. Exact replication of every numerical result also requires the original coefficient schedule, synthetic stream, selected thresholds, and analysis script. Those files should accompany the journal submission as supplementary material or be deposited in a persistent repository. The appropriate research direction is not a search for one universally adaptive classifier. It is the design of sociotechnical systems that choose among memory, updating, recalibration, human review, and control escalation based on the type and severity of change.

11. Conclusion

Fraud detection under concept drift is a coupled prediction and governance problem. The experiment shows that changing fraud mechanisms can destabilize performance and explanations, but it does not support the assumption that adaptation automatically improves minority-class ranking. Online learning raised recall while reducing precision, window retraining struggled with recurring concepts, and drift-triggered retraining inherited detector errors. Page-Hinkley offered a more conservative alarm profile, while the ADWIN-style detector was faster but substantially noisier. Explanation rankings changed sharply after sudden drift.

The practical implication is selective adaptation. Institutions need continuous monitoring, but model change should pass through diagnosis, validation, explanation review, human approval, deployment control, and post-change learning. The AFAGF provides that connection. Its purpose is not to slow defensive response. It is to ensure that speed is directed by reliable evidence and bounded by accountable decision rights. In financial cybersecurity, a model that learns continuously without institutional control can be as risky as a model that never learns at all.

References

1. Aggarwal CC. *Data streams: Models and algorithms*. New York: Springer; 2007.
2. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, *et al*. Explainable artificial intelligence: Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82-115. doi:10.1016/j.inffus.2019.12.012.
3. Baesens B, Van Vlasselaer V, Verbeke W. *Fraud analytics using descriptive, predictive, and social*

network techniques. Hoboken (NJ): Wiley; 2015.
4. Bifet A, Gavaldà R. Learning from time-changing data with adaptive windowing. In: *Proceedings of the 2007 SIAM International Conference on Data Mining*; 2007. p. 443-448. doi:10.1137/1.9781611972771.42.
5. Bifet A, Holmes G, Kirkby R, Pfahringer B. MOA: Massive online analysis. J Mach Learn Res. 2010;11:1601-1604.
6. Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency. *Supervisory guidance on model risk management (SR Letter 11-7/OCC Bulletin 2011-12)*. Washington (DC); 2011.
7. Breiman L. Random forests. Mach Learn. 2001;45:5-32. doi:10.1023/A:1010933404324.
8. Caelen O, Bontempi G. Inclusive and active learning toward cost-sensitive fraud detection. J Big Data. 2019;6:64. doi:10.1186/s40537-019-0224-8.
9. Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. ACM Comput Surv. 2009;41(3):15. doi:10.1145/1541880.1541882.
10. Dal Pozzolo A, Boracchi G, Caelen O, Alippi C, Bontempi G. Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Trans Neural Netw Learn Syst. 2018;29(8):3784-3797. doi:10.1109/TNNLS.2017.2736643.
11. Dal Pozzolo A, Caelen O, Le Borgne YA, Waterschoot S, Bontempi G. Learned lessons in credit card fraud detection from a practitioner perspective. Expert Syst Appl. 2014;41(10):4915-4928. doi:10.1016/j.eswa.2014.02.026.
12. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006;7:1-30.
13. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv. 2017. doi:10.48550/arXiv.1702.08608.
14. Elwell R, Polikar R. Incremental learning of concept drift in nonstationary environments. IEEE Trans Neural Netw. 2011;22(10):1517-1531. doi:10.1109/TNN.2011.2160459.
15. Fahim ASM, Ibrahim M, Pritty AA, Tania TA. Algorithmic accountability in U.S. consumer FinTech: Governance mechanisms for credit risk, fair lending, and financial stability. J Econ Finance Account Stud. 2023;5(4):80-93. doi:10.32996/jefas.2023.5.4.8.
16. Galbraith JR. Organization design: An information processing view. Interfaces. 1974;4(3):28-36. doi:10.1287/inte.4.3.28.
17. Gama J. *Knowledge discovery from data streams*. Boca Raton (FL): Chapman & Hall/CRC; 2010.
18. Gama J, Medas P, Castillo G, Rodrigues P. Learning with drift detection. In: *Advances in Artificial Intelligence – SBIA 2004*. Berlin: Springer; 2004. p. 286-295. doi:10.1007/978-3-540-28645-5_29.
19. Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. ACM Comput Surv. 2014;46(4):44. doi:10.1145/2523813.
20. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv. 2018;51(5):93. doi:10.1145/3236009.
21. Hasan N, Rasel IH, Rahman M, Islam K, Arman M, Jahan N. Securing U.S. healthcare infrastructure with machine learning: Protecting patient data as a national security priority. Int J Comput Exp Sci Eng. 2022;8(3):85-93. doi:10.22399/ijcesen.3987.
22. Jahan N, Pritty AA, Ibrahim M, Zadid MU, Fahim ASM, Mahmud S. Machine learning-driven early warning analytics for identifying market manipulation, irregular trading activity, and suspicious market signals in U.S. stock markets. J Comput Sci Technol Stud. 2024;6(2):257-283. doi:10.32996/jcsts.2024.6.2.26.
23. Kamruzzaman M, Saha S, Islam MK. Augmented intelligence in healthcare: Bridging machine learning, human expertise, and business process automation. Int J Eng Technol Res Manag. 2023;7(6):250-264.
24. Krawczyk B, Minku LL, Gama J, Stefanowski J, Woźniak M. Ensemble learning for data stream analysis: A survey. Inf Fusion. 2017;37:132-156. doi:10.1016/j.inffus.2017.02.004.
25. López-Rojas EA, Elmir A, Axelsson S. PaySim: A financial mobile money simulator for fraud detection. In: *Proceedings of the 28th European Modeling and Simulation Symposium*; 2016. p. 249-255.
26. Lu J, Liu A, Dong F, Gu F, Gama J, Zhang G. Learning under concept drift: A review. IEEE Trans Knowl Data Eng. 2019;31(12):2346-2363. doi:10.1109/TKDE.2018.2876857.
27. Lucas Y, Portier PE, Laporte L, Calabretto S, He-Guelton L, Oblé F, *et al*. Dataset shift quantification for credit card fraud detection. In: *2019 IEEE Symposium Series on Computational Intelligence*; 2019. p. 163-170.
28. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*. 2017;30.
29. March JG. Exploration and exploitation in organizational learning. Organ Sci. 1991;2(1):71-87. doi:10.1287/orsc.2.1.71.
30. Minku LL, White AP, Yao X. The impact of diversity on online ensemble learning in the presence of concept drift. IEEE Trans Knowl Data Eng. 2010;22(5):730-742. doi:10.1109/TKDE.2009.156.
31. Montiel J, Read J, Bifet A, Abdessalem T. Scikit-multiflow: A multi-output streaming framework. J Mach Learn Res. 2018;19(72):1-5.
32. National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1)*. Gaithersburg (MD): NIST; 2023. doi:10.6028/NIST.AI.100-1.
33. Page ES. Continuous inspection schemes. Biometrika. 1954;41(1-2):100-115. doi:10.1093/biomet/41.1-2.100.
34. Poyiadzi R, Sokol K, Santos-Rodriguez R, De Bie T, Flach P. FACE: Feasible and actionable counterfactual explanations. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*; 2020. p. 344-350. doi:10.1145/3375627.3375850.
35. Raab C, Heusinger M, Schleif FM. Reactive soft prototype computing for concept drift streams. Neurocomputing. 2020;416:340-351. doi:10.1016/j.neucom.2019.11.111.
36. Rasel IH, Ibrahim M, Pritty AA, Fahim ASM, Jahan N. Beyond FICO: Enhancing mortgage default forecasting and inclusive lending via multimodal graph neural networks and urban mobility analytics. Front Comput Sci Artif Intell. 2023;2(2):62-81. doi:10.32996/fcsai.2023.2.2.5.
37. Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining; 2016. p. 1135-1144. doi:10.1145/2939672.2939778.
38. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206-215. doi:10.1038/s42256-019-0048-x.
39. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432.
40. Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, *et al*. Hidden technical debt in machine learning systems. In: *Advances in Neural Information Processing Systems*. 2015;28.
41. Street WN, Kim Y. A streaming ensemble algorithm for large-scale classification. In: *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2001. p. 377-382. doi:10.1145/502512.502568.
42. Teece DJ, Pisano G, Shuen A. Dynamic capabilities and strategic management. Strateg Manag J. 1997;18(7):509-533.
43. Van Vlasselaer V, Bravo C, Caelen O, Eliassi-Rad T, Akoglu L, Snoeck M, *et al*. APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. Decis Support Syst. 2015;75:38-48. doi:10.1016/j.dss.2015.04.013.
44. Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv J Law Technol. 2018;31(2):841-887.
45. Webb GI, Hyde R, Cao H, Nguyen HL, Petitjean F. Characterizing concept drift. Data Min Knowl Discov. 2016;30:964-994. doi:10.1007/s10618-015-0448-4.
46. Widmer G, Kubat M. Learning in the presence of concept drift and hidden contexts. Mach Learn. 1996;23:69-101. doi:10.1007/BF00116900.
47. Žliobaitė I. Learning under concept drift: An overview. arXiv. 2010. doi:10.48550/arXiv.1010.4784.