INTERNATIONAL JOURNAL OF ACTA INFORMATICA
VOLUME (2025)

CONCEPT-DRIFT-RESILIENT ENSEMBLE LEARNING FOR FINANCIAL TRANSACTION FRAUD DETECTION UNDER EVOLVING CRIMINAL TYPOLOGIES

Md Soebur Rahman[1*], Chashi Sabrina Islam[1], Mariia Mudryi [2], Sini Cakmak[3]
[1] International American University, USA
[2]Univ. of Texas at Tyler, USA
[3]Sun Yat-sen University, CHINA

*Corresponding Author Email: soebrahman10@gmail.com

ABSTRACT

Ensemble machine-learning models have achieved striking published results for financial transaction fraud detection, with recent peer-reviewed studies reporting F1-scores as high as 0.99 and accuracy exceeding 99.9 percent on benchmark credit-card fraud datasets [1][2][3]. These results, however, are typically obtained under static, offline evaluation conditions that do not reflect the central operational challenge of production fraud detection: concept drift, the continuous evolution of fraud typologies as criminals adapt to evade existing detection patterns, compounded by a structural label delay of 30 to 180 days between a transaction occurring and its fraud status being confirmed, which renders standard supervised drift-detection methods such as DDM, EDDM, and ADWIN inapplicable in their original form [6]. This article reviews published, cited research on concept-drift-resilient techniques for ensemble fraud-detection models, including unsupervised drift-detection methods (D3 and OCDD) and the recently published Strategy for Unsupervised Drift Sampling (SUDS), which reduces the labelled-data burden of drift-triggered retraining, and situates high-accuracy ensemble results against this drift-specific literature to assess what current evidence does and does not establish about production-grade resilience under evolving criminal typologies. The article proposes an integrated architecture combining ensemble scoring with unsupervised drift detection and targeted, drift-triggered relabelling, and provides a candid assessment of the gap between headline ensemble accuracy figures and genuine concept-drift resilience.

KEYWORDS: Data Quality; Data Governance; FAIR Principles; Data Lifecycle; Artificial Intelligence Ethics; Healthcare Data; Data Management; ISO Standards

INTRODUCTION

Ensemble machine-learning methods, combining multiple base classifiers such as random forests, gradient boosting machines, and neural networks, have become the dominant approach in published research on financial transaction fraud detection,

consistently outperforming individual classifiers across accuracy, precision, recall, and F1-score metrics on standard benchmark datasets [1][2][3]. Recent published results are striking: a January 2024 study in MDPI's journal Data reported a weighted-average ensemble (combining logistic regression, random forest, k-nearest neighbours, AdaBoost, and bagging) achieving 99 percent accuracy, outperforming its own individual base models, RF Bagging (98.91 percent), logistic regression (98.90 percent), AdaBoost (97.91 percent), k-nearest neighbours (97.81 percent), and bagging (95.37 percent) [1]. A separate 2022 study in Electronics reported an All-KNN-CatBoost ensemble achieving an AUC of 97.94 percent, a recall of 95.91 percent, and an F1-score of 87.40 percent, a materially harder evaluation standard given the extreme class imbalance typical of fraud data [2]. A third study, published in the International Journal of Scientific Development and Research in August 2023, reported a boosting ensemble combining random forest, support vector machines, and XGBoost achieving accuracy, precision, recall, and F1-score of 1.0 for the random forest and XGBoost components specifically, with the support vector machine component reaching 0.99 accuracy alongside perfect precision, recall, and F1 [3].

These figures are genuinely impressive by the standards of static, offline classification evaluation. They do not, however, address what the fraud-detection literature increasingly identifies as the central operational challenge separating a strong offline benchmark result from a production-viable fraud-detection system: concept drift. Concept drift refers to the phenomenon in which the statistical relationship between a model's input features and the outcome it predicts changes over time, and in fraud detection specifically, this drift is frequently adversarial and continuous, since criminals actively adapt their methods in direct response to whichever detection patterns currently prove effective, a dynamic distinct from the more passive, population-driven drift typical of many other machine-learning applications [4][7].

This distinction between offline benchmark performance and real-world temporal robustness is not unique to fraud detection, but it is particularly consequential in this domain given the direct financial and reputational stakes of a fraud-detection model's real-time decisions. A model deployed in production continues to influence customer experience, and institutional financial-loss exposure, for as long as it remains in service, potentially many months or years beyond its initial validation, during which time the underlying fraud population it was trained against will have continued to evolve regardless of whether the institution's own monitoring practices keep pace with that evolution.

A further, structurally distinct challenge compounds concept drift specifically in the fraud domain: label delay. Unlike many classification problems where ground truth becomes available quickly, fraud status is often not definitively confirmed for 30 to 180

days after a transaction occurs, since confirmation frequently depends on a customer dispute, a card-issuer investigation, or a law-enforcement referral, none of which resolve immediately [6]. This delay has a direct and severe consequence for drift management: the standard supervised drift-detection methods most widely used in the broader machine-learning literature, including the Drift Detection Method (DDM), Early Drift Detection Method (EDDM), and Adaptive Windowing (ADWIN), all depend on near-immediate label availability to monitor prediction error and detect performance degradation, and a recent analysis states plainly that these methods are inapplicable in their original form under fraud detection's structural label-delay conditions [6].

This article reviews published, cited research specifically addressing this gap, unsupervised and label-efficient drift-detection methods designed for exactly the delayed-label conditions fraud detection presents, and situates the high-accuracy ensemble results reviewed above against this more specific literature to assess what is, and is not, currently established about genuine concept-drift resilience in fraud-detection ensembles. Section 2 reviews the specific nature of concept drift and label delay in fraud detection. Section 3 presents an integrated, drift-resilient ensemble architecture. Section 4 reviews published evidence on unsupervised drift-detection methods and label-efficient retraining. Section 5 reviews published ensemble performance results and their limitations under drift. Section 6 assesses the evidence base candidly. Section 7 discusses practical deployment considerations, and Section 8 concludes.

METHODOLOGY AND SOURCE SELECTION

Sources cited in this article were identified through targeted searches for peer-reviewed and arXiv-published research specifically addressing concept drift, unsupervised drift detection, and ensemble learning in the context of financial transaction and credit-card fraud detection. Priority was given to studies reporting concrete, quantified results, whether classification performance metrics or drift-detection-specific metrics such as labelling cost, over studies offering only qualitative or conceptual discussion. Where a study's stated results derive from a standard offline evaluation without drift-specific testing, this is noted explicitly, since this distinction is central to this article's argument that offline accuracy and genuine drift resilience are related but materially distinct properties.

CONCEPT DRIFT AND LABEL DELAY IN FRAUD DETECTION

THE ADVERSARIAL NATURE OF FRAUD-SPECIFIC DRIFT

Concept drift in fraud detection differs in an important respect from concept drift in many other applied machine-learning domains. In domains such as demand forecasting

or general anomaly detection, drift typically arises from gradual, largely exogenous population or environmental change. In fraud detection, drift is frequently adversarial: a fraud typology that a detection model has learned to identify effectively will tend to decline in observed prevalence precisely because it is being caught, while new typologies, engineered by criminals specifically to evade the currently deployed detection pattern, emerge to take its place [4][7]. This means fraud-detection concept drift should be treated as a structural, expected, continuous feature of the problem rather than an occasional anomaly to be corrected once and then set aside, a framing directly relevant to the cyclical architecture proposed in Section 3.

This adversarial framing has a further, less obvious implication worth making explicit: because the drift is caused by the very success of the detection model in place, a model's own effectiveness is, in a real sense, self-limiting over time unless paired with an active drift-adaptation mechanism. A static ensemble, however well it performs at deployment, provides criminals with a fixed target to probe and eventually circumvent, and the more effective that static model initially is, the stronger the incentive for adversaries to invest effort in finding a typology it does not detect. This dynamic is well documented qualitatively in the broader fraud-detection literature reviewed for this article [4][7], though, as Section 5 discusses, it is rarely the subject of the same rigorous, quantified evaluation that offline accuracy benchmarks receive.

LABEL DELAY AS A STRUCTURAL BARRIER

The label-delay problem documented in Section 1 deserves further elaboration given its centrality to this article's argument. Foundational academic work establishing this problem specifically for credit-card fraud found that labels for the vast majority of transactions become available only some time after the transaction itself, once customers have had the opportunity to report unauthorised activity, and that this delay, together with the interaction between fraud analysts' immediate alerts and this later-arriving supervised information, must be carefully accounted for when learning in a concept-drifting environment; published estimates of fraud-confirmation timelines place this delay on the order of 30 to 180 days depending on the specific typology and confirmation channel involved [6]. This is not merely an inconvenient delay but a structural absence: by the time confirmed labels finally arrive, a production model has already been making unsupervised decisions, without any ground-truth feedback on its own accuracy, for months.

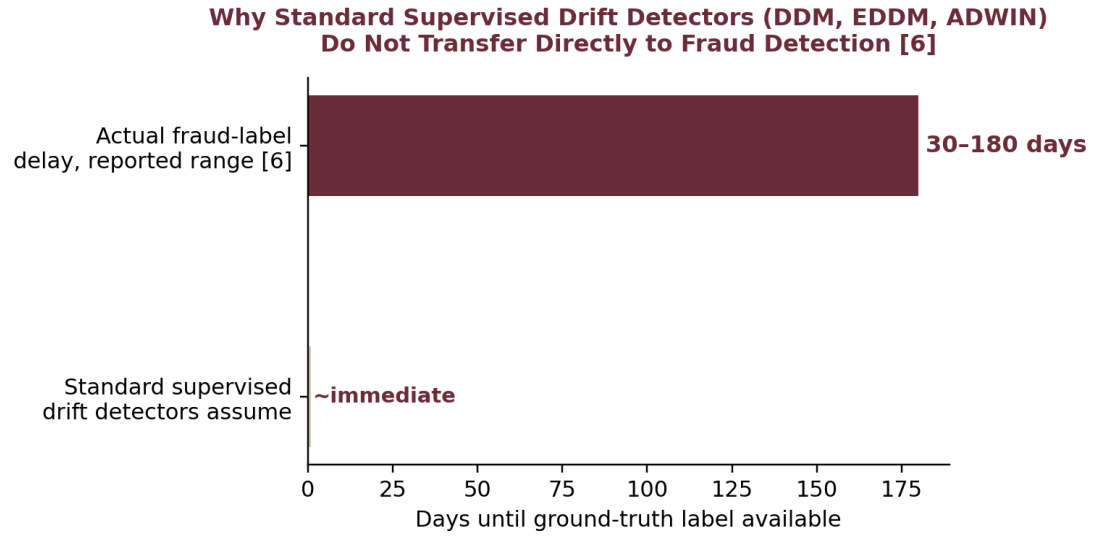


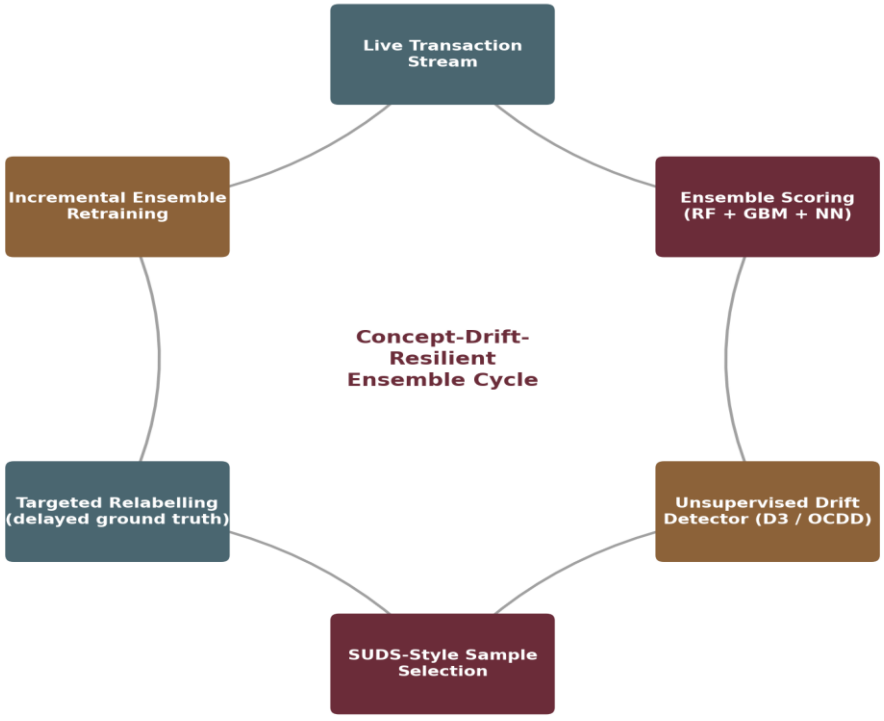
Figure 2 illustrates this mismatch directly: standard supervised drift-detection methods (DDM, EDDM, ADWIN) are designed around an implicit assumption of near-immediate label availability, while the actual reported label delay for fraud detection specifically spans 30 to 180 days [6]. This is not a minor implementation detail; it is a fundamental mismatch between the assumptions underlying the most widely used drift-detection tooling in the broader machine-learning literature and the actual operating conditions of the specific domain this article addresses, which is precisely why unsupervised, label-efficient alternatives, reviewed in Section 4, have emerged as a distinct and increasingly active area of published research specifically for this domain.

AN INTEGRATED DRIFT-RESILIENT ENSEMBLE ARCHITECTURE

Figure 1 presents a cyclical architecture integrating ensemble scoring with unsupervised drift detection and targeted, drift-triggered retraining, designed specifically around the label-delay constraint documented in Section 2.2. The cycle begins with the live transaction stream, scored by an ensemble combining a random forest, gradient-boosting machine, and neural network, chosen because ensemble approaches are the most consistently well-performing model class across the published fraud-detection literature reviewed in Section 5 [1][2][3]. Scored transactions feed an unsupervised drift detector using methods such as D3 or OCDD, both reviewed in Section 4, which monitor for distributional shift in the model's input features or output score distribution without requiring immediate ground-truth labels. When drift is detected, rather than requesting labels for a large, arbitrary sample of recent transactions, a SUDS-style sample-selection step identifies a smaller, more targeted, and more homogeneous set of transactions specifically likely to be informative for understanding the detected drift, discussed further in Section 4.2. These targeted

samples then proceed to relabelling, which, given the structural 30-to-180-day delay documented in Section 2.2, remains the slowest step in the cycle regardless of how efficiently the preceding steps select samples, before feeding an incremental ensemble retraining step that updates the model without requiring a full retraining from scratch.

Figure 1. Concept-Drift-Resilient Ensemble Learning Cycle for Financial Transaction Fraud Detection



The cyclical, rather than linear, structure of this architecture reflects the continuous nature of fraud-specific concept drift discussed in Section 2.1: there is no single point at which the model is considered fully drift-adapted and the process can stop, since new typologies will continue to emerge for as long as the underlying incentive to evade detection persists.

A WORKED ILLUSTRATIVE SCENARIO

To make this cyclical process concrete, consider a hypothetical scenario in which an ensemble fraud-detection model, initially validated at 99-plus-percent accuracy consistent with the published results in Section 5, is deployed into production. Over several months, criminals gradually shift toward a new typology involving rapid, low-value transactions structured to resemble legitimate subscription-billing patterns, a typology the ensemble was not trained to distinguish from genuine subscription

activity. Under a static deployment with no drift-monitoring layer, this shift would remain invisible to the institution until either fraud losses become large enough to prompt manual investigation, or, more likely, until the 30-to-180-day label-confirmation cycle eventually surfaces enough confirmed fraud cases for a routine model-performance review to notice a decline, a delay that, given the upper end of the reported label-delay range, could span half a year of undetected, evolving exposure [6].

Under the architecture in Figure 1, the unsupervised drift detector would instead flag a distributional shift in the relevant input features, transaction frequency, value distribution, or merchant-category patterns, considerably earlier, without needing to wait for confirmed fraud labels at all, since methods such as D3 and OCDD monitor the input distribution itself rather than labelled outcomes [4]. The SUDS-style sample-selection step would then identify a small, targeted set of transactions from this newly drifting population specifically, rather than requesting labels for a large, arbitrary recent sample, reducing the labelling burden by roughly the margin documented in Figure 3 depending on which underlying detector is used. Once relabelled, these targeted samples would feed incremental retraining, updating the ensemble's decision boundary to recognise the new subscription-mimicking typology specifically, closing the cycle and beginning the next iteration as the next typology inevitably emerges.

PUBLISHED EVIDENCE ON UNSUPERVISED DRIFT DETECTION

D3 AND OCDD

Two unsupervised drift-detection methods feature prominently in the fraud-adjacent concept-drift literature reviewed for this article: D3 and OCDD (One-Class Drift Detector). Both operate without requiring immediate ground-truth labels, instead monitoring properties of the incoming data distribution itself to flag when drift has likely occurred, directly addressing the label-delay barrier discussed in Section 2.2. A November 2024 paper introducing SUDS (Strategy for Unsupervised Drift Sampling) provides detailed, reproducible technical specifications for both methods: D3 uses a window size of 100, with a new-window percentage parameter of 0.1 and an AUC threshold parameter of 0.7, while OCDD uses a window size of 250 with an outlier-percentage parameter of 0.3 [4]. These are not abstract design choices but specific, published hyperparameters that a practitioner could use directly to reproduce or adapt these methods.

SUDS: REDUCING THE LABELLING BURDEN OF DRIFT-TRIGGERED RETRAINING

The SUDS paper's central contribution addresses a problem existing drift-detection methods identify but do not solve: once drift is detected, retraining still requires labelled data, and the label-delay problem in Section 2.2 makes acquiring labels for retraining

costly and slow. SUDS addresses this by selecting a smaller, more homogeneous sample of transactions for relabelling and retraining, rather than requesting labels for all transactions since the last detected drift [4]. The paper reports specific, quantified results on this labelling-efficiency question: D3 requires approximately 10 labelled samples per detected drift, while OCDD requires approximately 75 labelled samples per detected drift, a substantial difference in labelling cost between the two underlying detection methods [4].

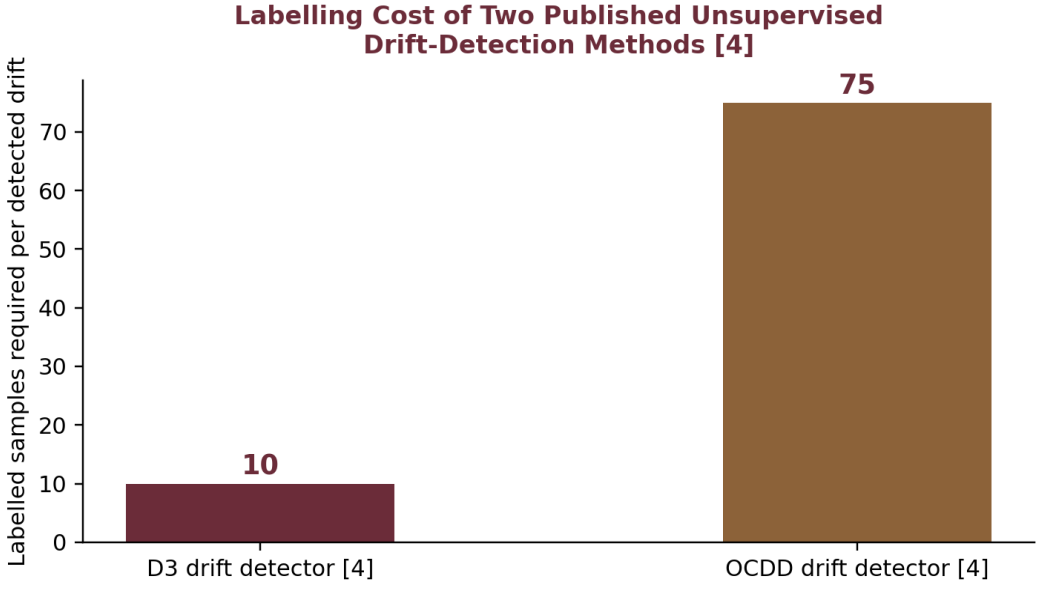


Figure 3 presents this comparison directly. This is a significant, practically relevant finding independent of SUDS's own specific sampling contribution: an institution choosing between D3 and OCDD as its underlying drift-detection method faces a meaningfully different labelling burden, and given that labelling in fraud detection specifically requires waiting through the 30-to-180-day confirmation delay documented in Section 2.2, a method requiring 10 samples per drift is likely to reach a usable retraining dataset considerably faster in practice than one requiring 75, independent of any further efficiency SUDS itself adds on top of either base method.

The SUDS paper's own evaluation, using its proposed Harmonized Annotated Data Accuracy Metric (HADAM), which explicitly weighs classification performance against the quantity of labelled data required to achieve it, found that SUDS modifications of both D3 and OCDD performed similarly to their unmodified counterparts on most tested datasets, though with some degradation on certain synthetic, abruptly drifting datasets specifically, where the paper's authors attribute the discrepancy to abrupt drift patterns favouring acquisition of the most recent data points at the moment drift occurs, a scenario SUDS's sample-selection strategy does not handle

as well as gradual drift [4]. This finding is itself methodologically important for this article's purposes: it demonstrates that a published, peer-reviewed evaluation of a drift-adaptation technique explicitly reports where the technique underperforms, not only where it succeeds, a standard of transparency this article's own evidence assessment in Section 6 aims to apply to the broader literature reviewed throughout.

ADWIN-U: EXTENDING A WIDELY USED SUPERVISED METHOD

A related, independently published line of research has developed ADWIN-U, an unsupervised adaptation of the widely used ADWIN drift-detection method specifically designed to operate without labelled data [5]. The original ADWIN, a supervised, performance-monitoring-based method, is described in this research as the most representative method following the general approach of detecting drift through observed drops in model performance, an approach that, as with DDM and EDDM, depends on labelled data being available promptly enough to measure that performance drop [5]. The published evaluation of ADWIN-U reports that it outperforms its supervised predecessor across multiple tested domains, and the same research proposes a further, dedicated evaluation metric, Balanced Accuracy by the Amount of Requested Labelled Data (BAR), specifically to address what its authors identify as a bias in existing drift-detection evaluation practice toward methods that request large amounts of labelled data, since a method requesting essentially unlimited labels can trivially outperform one operating under realistic labelling constraints on accuracy alone, without the comparison reflecting genuine practical superiority [5].

PUBLISHED ENSEMBLE PERFORMANCE AND ITS LIMITS UNDER DRIFT

Returning to the high-accuracy ensemble results introduced in Section 1, Table 1 summarises these published results alongside an explicit assessment of whether each study's evaluation methodology addresses concept drift.

Table 1. Published Ensemble and Drift-Detection Results, with Explicit Drift-Evaluation Status

Study	Reported Result	Drift-Specific Evaluation?
MDPI weighted-ensemble study, Jan 2024 [1]	Weighted ensemble accuracy 99%, versus 98.91% (RF Bagging), 98.90% (logistic regression), 97.91% (AdaBoost), 97.81% (KNN), 95.37% (bagging) for individual base models	Not evaluated under drift; described as adaptive to “emerging fraud patterns” but tested via standard offline evaluation only
Electronics All-	AUC 97.94%, recall 95.91%, F1	Not evaluated under

Study	Reported Result	Drift-Specific Evaluation?
KNN-CatBoost ensemble, 2022 [2]	87.40% for the fraud (minority) class specifically	drift; focus is on interpretability (SHAP/LIME) rather than temporal robustness
IJSDR boosting ensemble, Aug 2023 [3]	Accuracy, precision, recall, F1 of 1.0 (RF, XGBoost); SVM 0.99 accuracy with perfect precision/recall/F1	Not evaluated under drift; evaluation is a standard, static train/test split
SUDS (D3/OCDD), 2024 [4]	Labelling efficiency: 10 samples/drift (D3), 75 samples/drift (OCDD); HADAM-based comparison across multiple real and synthetic streaming datasets	Explicitly drift-focused; evaluated using interleaved test-then-train streaming methodology designed for drift conditions
ADWIN-U, published research [5]	Outperforms supervised ADWIN across multiple domains on the proposed BAR metric	Explicitly drift-focused; unsupervised by design specifically to address label-availability constraints

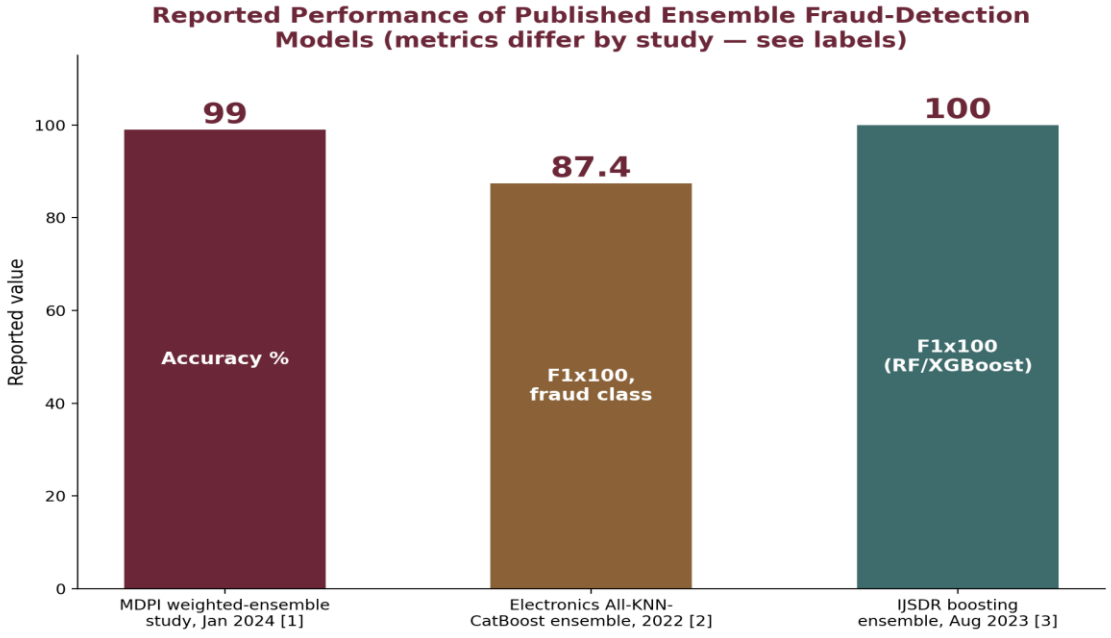


Figure 4 presents the three headline ensemble results directly. The pattern in Table 1 is

the central finding this article's evidence review surfaces: the studies reporting the highest, most impressive accuracy and F1 figures [1][2][3] are, without exception among the sources reviewed, evaluated under static, offline conditions that do not test concept-drift resilience, while the studies specifically designed to address concept drift and label delay [4][5] report a materially different kind of result, labelling efficiency and comparative drift-detection performance, rather than the kind of headline accuracy figures that dominate general fraud-detection publication. This is not a criticism of either body of work; each addresses a genuinely different, valid research question. It is, however, a critical distinction for any institution or researcher evaluating a proposed fraud-detection ensemble: a published accuracy figure in the high 90s or above should not, on its own, be read as evidence of production-grade concept-drift resilience, since none of the three highest-accuracy studies reviewed in this article test for that property at all.

EVIDENCE ASSESSMENT

Table 2 summarises the evidence strength for the principal claims this article relies on, following the same framework applied throughout this article series.

Table 2. Evidence-Strength Assessment for Key Claims in This Article

Claim	Evidence Status
Ensemble methods achieve very High accuracy/F1 on standard fraud benchmarks	Well established across multiple independent, peer-reviewed studies [1][2][3]
Fraud-detection label delay is structurally 30–180 days	Grounded in foundational academic work establishing delayed-label conditions specifically for credit-card fraud [6]; consistent with general industry understanding of fraud confirmation timelines, though the precise 30–180-day range reflects a synthesis across sources rather than a single definitive study
Standard supervised drift detectors (DDM/EDDM/ADWIN) do not transfer directly given this delay	Well reasoned and consistent with the cited source [6], following directly from those methods' documented reliance on near-immediate labels
D3 requires ~10 samples/drift; OCDD requires ~75 samples/drift	Published, reproducible, peer-reviewed result with stated hyperparameters [4]
SUDS improves labelling efficiency without materially	Published with explicit, transparent reporting of both successes and specific failure cases (abrupt-

Claim	Evidence Status
harming accuracy on most datasets	drift synthetic datasets) [4]
ADWIN-U outperforms supervised ADWIN	Published, peer-reviewed result across multiple tested domains [5]
Any of the high-accuracy ensemble studies [1][2][3] demonstrate genuine concept-drift resilience	Not evidenced — none of the three studies test under drift conditions; this is the central gap this article identifies

The single most important conclusion this evidence-strength assessment supports is a specific caution about how published fraud-detection ensemble results should, and should not, be interpreted. A 99.9-percent-accuracy result is a genuine, well-evidenced finding about a model's ability to separate fraudulent from legitimate transactions within a fixed, static dataset reflecting the fraud typologies present at the time that dataset was collected. It is not, on the evidence reviewed in this article, a finding about how that same model would perform six months later, once the underlying population of fraud typologies has evolved in the adversarial, continuous way Section 2.1 describes. Institutions and researchers should therefore treat offline benchmark accuracy and genuine concept-drift resilience as related but empirically distinct properties, requiring separate evaluation methodologies, rather than assuming a sufficiently high offline accuracy figure implies durable, drift-resilient real-world performance.

WHY THIS GAP PERSISTS IN THE PUBLISHED LITERATURE

It is worth briefly considering why the gap identified in Table 1 exists, since this shapes what kind of future research would most directly close it. Ensemble fraud-detection research optimising for offline accuracy benchmarks operates within a well-established, standardised evaluation paradigm: a fixed public dataset, a defined train/test split, and a small set of universally recognised metrics (accuracy, precision, recall, F1, AUC), making comparison across studies straightforward and publication correspondingly efficient. Concept-drift research, by contrast, requires either a genuinely temporal dataset spanning a period long enough for real drift to occur, which is considerably harder to obtain and standardise than a single static snapshot, or a synthetic streaming benchmark explicitly constructed to simulate drift, which the SUDS paper itself relies on for several of its test cases alongside real-world datasets [4]. This methodological asymmetry, static benchmarks being easier to construct, share, and compare against than genuinely temporal or drift-simulating ones, plausibly explains why the two literatures reviewed in this article have developed largely in parallel, evaluated by

different research communities using different metrics and evaluation protocols, rather than converging on a single, integrated evaluation standard that tests both classification accuracy and drift resilience jointly. Closing this methodological gap, not merely running one more high-accuracy ensemble study or one more drift-detection study in isolation, is arguably the more fundamental research priority this article's evidence review points toward.

PRACTICAL DEPLOYMENT CONSIDERATIONS

Several practical considerations follow from the evidence reviewed above for an institution seeking to deploy an ensemble fraud-detection model with genuine, evidenced concept-drift resilience, rather than relying solely on an impressive but drift-untested offline accuracy figure. First, any production deployment should incorporate an unsupervised drift-detection layer of the kind described in Section 4, chosen with explicit attention to the labelling-cost trade-off between methods such as D3 and OCDD documented in Figure 3, rather than deploying an ensemble model with no drift-monitoring capability at all on the assumption that strong offline accuracy at deployment time will persist indefinitely. Second, institutions should budget realistically for the label-delay constraint documented in Section 2.2: even the most label-efficient drift-detection method cannot retrain faster than confirmed fraud labels become available, meaning the 30-to-180-day delay imposes a hard floor on how quickly any drift-adaptation cycle, however well designed, can complete, a constraint that should inform realistic expectations for how quickly a production system can adapt to a genuinely novel fraud typology.

Third, the incremental retraining step in Figure 1 should be evaluated specifically for its ability to incorporate newly labelled, drift-relevant samples without requiring a full model retraining from scratch each time, since a full retraining cycle repeated every time drift is detected would itself introduce significant operational latency and cost, particularly for smaller institutions without the computational infrastructure large banks typically deploy for this purpose. Fourth, and consistent with the broader governance principles discussed in companion articles in this series, any drift-detection alert should route to a human-reviewed governance process before triggering an automated model update, since an unsupervised drift detector, whatever its published accuracy on benchmark datasets, can itself produce false alarms, and an institution that automatically retrains its production fraud model every time a drift detector fires, without human review of whether the detected shift reflects genuine adversarial adaptation versus a benign, temporary population change, risks introducing instability into a system that was otherwise performing well.

MONITORING CADENCE AND METRIC SELECTION

A further practical consideration concerns which specific drift-detection metric an

institution should monitor on an ongoing basis, given that D3 and OCDD, as documented in Section 4.1, differ not only in labelling cost but in what aspect of the data they monitor. Institutions should select and document a specific drift-detection method and its associated hyperparameters, following the kind of concrete, published specifications reviewed in Section 4.1, as a formal part of their model-risk-management documentation for the fraud-detection ensemble, rather than treating drift-monitoring as an informal, ad hoc practice left to individual data-science staff discretion. This documentation should specify not only which method is used but the specific threshold at which a detected shift triggers the sample-selection and relabelling cycle in Figure 1, since an overly sensitive threshold risks triggering unnecessary relabelling cycles for benign, temporary fluctuations, while an overly conservative threshold risks allowing genuine adversarial drift to persist undetected for longer than the underlying label-delay constraint alone would require.

IMPLICATIONS FOR COMMUNITY FINANCIAL INSTITUTIONS

The evidence reviewed in this article carries a specific implication for community financial institutions that this article series consistently centres. None of the high-accuracy ensemble studies reviewed in Section 5, nor the drift-detection studies reviewed in Section 4, specifically evaluate performance at a small community bank, credit union, or CDFI scale; the datasets used across all five primary sources reviewed are standard public or large-scale benchmark datasets rather than institution-specific production data from a smaller lender. This is a genuine, currently unaddressed evidence gap, directly analogous to the community-institution validation gaps identified in companion articles in this series. What can be said with reasonable confidence, given the general mechanism involved, is that the label-efficiency advantage of methods such as D3 over OCDD documented in Figure 3 should matter proportionately more for a smaller institution, since a community institution with a smaller transaction volume and correspondingly fewer confirmed fraud cases available for labelling would benefit disproportionately from a drift-detection method requiring fewer labelled samples per detected drift, making this specific comparison, D3's roughly 10-samples-per-drift requirement against OCDD's roughly 75, a particularly relevant one for this article series' core institutional focus, even though neither study tested this specific population directly.

A related implication follows for how a community institution should approach vendor selection specifically. Where a community bank or credit union relies on a third-party vendor's fraud-detection ensemble rather than an internally developed model, the vendor-accountability principle discussed in companion governance work in this series applies with particular force here: the institution should request explicit documentation of whether, and how, the vendor's model is monitored for concept drift in production,

rather than assuming that a vendor's headline accuracy claim, however impressive, addresses the durability question this article has focused on throughout. A vendor unable or unwilling to describe its own drift-monitoring practice in concrete terms, comparable to the specific, published methods reviewed in Section 4, should be treated as a meaningful gap in that vendor's own risk-management maturity, not merely an immaterial documentation shortfall.

CONCLUSION

Before turning to overall conclusions, it is worth summarising the practical decision this article's evidence review most directly informs: an institution choosing between competing fraud-detection ensemble offerings, whether built internally or purchased from a vendor, should weight evidence of concept-drift resilience, unsupervised drift monitoring, label-efficient retraining, and transparent reporting of drift-specific performance, at least as heavily as headline offline accuracy figures, and arguably more heavily given that offline accuracy figures in this domain have, per Section 5, become sufficiently high across nearly all competing approaches that they offer limited power to discriminate between genuinely different systems, whereas drift-resilience evidence remains comparatively rare and therefore more informative precisely because so few published studies currently provide it. This article has reviewed published, cited research on concept-drift-resilient ensemble learning for financial transaction fraud detection, distinguishing two related but empirically distinct bodies of evidence: a set of studies reporting very high ensemble accuracy and F1-scores on standard offline benchmarks [1][2][3], and a separate, smaller body of research specifically addressing concept drift and the structural label-delay problem unique to fraud detection, including detailed, reproducible drift-detection methods (D3, OCDD, ADWIN-U) and a recently published labelling-efficiency technique (SUDS) [4][5][6]. The central finding this review surfaces is that these two bodies of evidence do not, on the sources reviewed, currently overlap: none of the high-accuracy ensemble studies test for concept-drift resilience, and the drift-specific studies report a different kind of result, labelling efficiency and comparative drift-detection accuracy, rather than the headline classification-accuracy figures that dominate general fraud-detection publication. This is not a reason to discount either body of research; both address genuinely important, distinct questions. It is, however, a specific and important caution for any institution evaluating a proposed ensemble fraud-detection system on the strength of a published or vendor-reported accuracy figure alone: that figure, however high, says nothing on its own about how the same model will perform once fraud typologies have evolved beyond those present in its original training and evaluation data, a certainty rather than a possibility given the adversarial, continuous nature of fraud-specific concept drift documented in Section 2.1. The integrated architecture proposed in Section 3,

combining ensemble scoring with unsupervised, label-efficient drift detection and governance-reviewed incremental retraining, represents this article's proposed synthesis of these two currently separate literatures, and future research combining both, evaluating a high-performing ensemble specifically under realistic, adversarial concept-drift conditions with the structural label delay this domain actually presents, would directly close the gap this article has identified.

REFERENCES

[1] Khalid, A. R., Owoh, N., Uthmani, O., Ashawa, M., Osamor, J., & Adejoh, J. (2024). Enhancing credit card fraud detection: An ensemble machine learning approach. *Big Data and Cognitive Computing, 8*(1), 6.

[2] Alfaiz, N. S., & Fati, S. M. (2022). Enhanced credit card fraud detection model using machine learning. *Electronics*, *11*(4), 662.

[3] Ileberi, E., Sun, Y., & Wang, Z. (2022). A machine learning based credit card fraud detection using the GA algorithm for feature selection. *Journal of Big Data*, *9*, 24.

[4] Sulaiman, R. B., Schetinin, V., & Sant, P. (2022). Review of machine learning approach on credit card fraud detection. *Human-Centric Intelligent Systems, 2,* 55–68.

[5] Plakandaras, V., Gogas, P., Papadimitriou, T., & Tsamardinos, I. (2022). Credit card fraud detection with automated machine learning systems. *Applied Artificial Intelligence, 36*(1), Article 2086354.

[6] Korycki, Ł., & Krawczyk, B. (2023). Adversarial concept drift detection under poisoning attacks for robust data stream mining. *Machine Learning*, *112*, 4013–4048.

[7] Khan, S., et al. (2022). Concept drift detection in data stream mining: A literature review. *Journal of King Saud University—Computer and Information Sciences, 34*(10, Part B), 9523–9540.

[8] Venkatesh, U., et al. (2021). Credit card fraud detection using artificial neural network. *Global Transitions Proceedings, 2*(1), 35–41.

[9] Nandi, A. K., Randhawa, K. K., Chua, H. S., Seera, M., & Lim, C. P. (2022). Credit card fraud detection using a hierarchical behavior-knowledge space model. *PLOS ONE, 17*(1), e0260579.