

---

## Adaptive and Explainable Machine Learning for Temporal Drift-Aware Financial Fraud Detection and Risk Monitoring: A Review

Yamil Osei <sup>1</sup>, Zofiana Akosua <sup>2</sup>

<sup>1</sup> Vistula Institute of Computer Science, POLAND

<sup>2</sup> Makerere University, UGANDA

---

### ABSTRACT

---

*Financial fraud detection is a moving target: fraud tactics, legitimate consumer behavior, and the statistical relationship between transaction features and fraud risk all shift over time, so a model that is accurate on the data it was trained on can quietly degrade in production without triggering any alarm built around static accuracy alone. This review synthesizes six sources to examine how the literature is responding to that problem along three fronts named in this review's title: adaptivity, explainability, and explicit drift-awareness. A systematic survey anchors the review's broadest claim, identifying class imbalance, concept drift, adversarial behavior, and explainability as the field's persistent, interlinked deployment barriers. Two applied studies evaluate models on the same public benchmark (284,807 European credit-card transactions, 0.17% fraud rate) and are therefore directly comparable: a soft-voting ensemble combined with explicit drift quantification (via the Population Stability Index, Kolmogorov-Smirnov testing, and Jensen-Shannon divergence) reaches an AUC of 0.9703, while a SHAP-explained Random Forest on the same data reaches AUC = 0.9766. A third applied study, built and validated on five years of real transaction data from Gulf Cooperation Council banks, combines drift detection, adversarial defense, and human-in-the-loop explainability in a single deployed framework, reporting a fraud-recall improvement from 35% to 85% and a reduction in adversarial attack success from 35% to 5%, with drift recovery within 24 hours and operational latency under 150 milliseconds. A fourth study addresses adaptivity most directly, comparing instance-incremental and batch-incremental streaming learning strategies and translating technical feature importance into investigator-facing concept explanations, though without a single reported benchmark accuracy figure this review could independently verify. A documented, non-peer-reviewed industry account of fraud-model failure during the COVID-19 pandemic's abrupt shift in consumer spending is included for real-world grounding, disclosed explicitly as not being an academic source. Read together, these sources suggest that adaptivity, explainability, and drift-awareness are not competing design goals but three views of the same underlying requirement: a fraud detection system whose statistical foundations are continuously checked against the population it is currently scoring, not just the population it was trained on.*

---

**Keywords:** Financial Fraud Detection, Concept Drift, Adaptive Learning, Explainable AI, Risk Monitoring, SHAP, Class Imbalance, Credit Card Fraud

---

### Introduction

#### Why a Static Fraud Model is a Decaying Asset

A fraud detection model is typically evaluated once, at deployment, against a held-out test set drawn from the same period as its training data, and that single evaluation is often treated as a durable measure of the model's quality. But the statistical relationship a model has learned between transaction features and fraud risk is a property of a particular population at a particular time; both the legitimate and fraudulent populations change continuously as consumer behavior shifts, payment technologies evolve, and fraud tactics adapt specifically to evade whatever detection method is currently deployed against them. A model with no mechanism for detecting

or adapting to that shift does not fail loudly — it fails quietly, as a slow rise in missed fraud or false positives that a static accuracy figure computed once at launch will never reveal [1].

### **Scope: three pillars this review treats as one problem**

This review's title names three properties — adaptive, explainable, and temporal drift-aware and treats them as three interdependent requirements of one system rather than three separate research agendas. A model cannot adapt responsibly to a population it is not monitoring for drift; a drift-monitoring system that flags a shift but cannot explain which features or transaction types are driving it gives an analyst little to act on; and an explanation method applied once, at training time, says nothing about whether the same explanation still holds for the population the model is scoring six months later [2-4].

### **Real-world grounding: when drift broke deployed fraud systems**

The clearest documented illustration of this review's central concern is not from the academic literature but from industry reporting during the COVID-19 pandemic. Nguyen (2020), writing in *American Banker*, documents how the abrupt, large-scale shift in consumer payment behavior in early 2020 — a third of consumers aged 65 and older reported planning to increase online spending specifically because of the pandemic, alongside a surge in bulk purchases, unusual first-time purchase categories, new government-issued payment sources, and rapid adoption of peer-to-peer payment technologies — destabilized fraud models built on pre-pandemic behavioral baselines. New, legitimate customers were flagged as fraudulent at elevated rates while some fraud went undetected in the surge of unusual-but-legitimate activity, precisely because the established risk thresholds and machine-learning baselines those models relied on had not been built to detect, let alone adapt to, a shift of that scale. This account is industry commentary rather than a peer-reviewed study, and this review cites it only for that real-world grounding, not as an evidentiary source for any quantitative claim elsewhere in this review.

### **Review scope and methodology**

This review synthesizes six sources spanning 2020 to 2026, selected on relevance and evidentiary strength rather than a fixed recency cutoff. One is a systematic survey that anchors this review's broadest cross-cutting claims; three are individually verified applied studies with directly checkable datasets and quantitative results, two of which share a public benchmark and are therefore directly comparable to each other; one is an applied study whose contribution is architectural rather than benchmark-driven, and is discussed with that distinction made explicit; and one is a non-academic industry account included solely for real-world grounding, with its non-peer-reviewed status disclosed wherever it is cited.

### **The Scale of the Problem: A Systematic View**

#### **What the field looks like at scale**

Thivaivos, Kostopoulos, Stefani, and Kotsiantis (2026) provide this review's broadest evidentiary base: a PRISMA-guided systematic survey of machine learning and deep learning approaches to financial fraud detection, spanning classical supervised learning, anomaly detection, graph-based methods, deep neural architectures, multimodal approaches, and cost-sensitive learning frameworks, with particular attention to emerging techniques such as graph neural networks and federated learning. Their analysis of real-world deployment challenges converges on exactly the set of interlinked barriers this review's title addresses: extreme class imbalance, concept drift (explicitly discussed as "drift-aware learning frameworks" needed to detect distributional change as fraud tactics continuously evolve), adversarial behavior, data-privacy constraints, and real-time deployment requirements, alongside a significant emphasis on explainable AI techniques as necessary for regulatory compliance and decision traceability. That a survey scoped to the entire field, rather than to any one technique, independently arrives at the same three pillars this review treats as one problem is itself evidence that they are not an arbitrary framing choice [5-9].

### **Where this review's five applied and grounding sources sit within that broader taxonomy**

Thivaivos *et al.*'s (2026) taxonomy spans architectures well beyond what this review's five other sources individually cover — convolutional and recurrent networks, autoencoders, generative adversarial networks, transformers, and graph neural networks applied to relational transaction networks, alongside federated learning approaches that train across multiple institutions without centralizing raw data. None of Sections 3 or 4's applied studies uses a graph-based or federated approach; Zuberi *et al.* (2026) and Awad *et al.* (2026) work with classical and ensemble classifiers plus one custom deep network, while Al-Daoud and Abu-AlSondos (2025) and Alessi and Fugini (2026) focus on hybrid pipelines and streaming architectures rather than graph or federated methods. This is a genuine scope limitation on what this review can conclude: the applied evidence reviewed in Sections 3 and 4 speaks specifically to tree-based, ensemble, and streaming approaches, and this review makes no claim, one way or the other, about how graph-based or federated methods would perform against the same drift, explainability, and adaptivity criteria.

### **Quantifying and Detecting Temporal Drift**

#### **A direct, multi-method measurement of drift on a public benchmark**

Zuberi *et al.* (2026) address temporal drift not as a secondary concern but as one of the study's three explicit research questions, alongside extreme class imbalance (0.173% fraud prevalence) and the cost asymmetry between missed fraud and false alarms. Working with the well-known Kaggle European credit-card dataset (284,807 transactions over 48 hours), they enforce chronological, leakage-free time-based splitting (70/15/15) and fit all preprocessing, SHAP-based feature selection, and oversampling exclusively to the training partition — a methodological discipline that matters directly to this review's concerns, since evaluating a drift-monitoring approach on a dataset contaminated by future information would understate exactly the kind of degradation a real deployment would eventually face. Across 85 model

combinations (12 classical classifiers, 3 multilayer perceptron architectures, a custom deep network the authors call FraudNet, and 7 ensemble methods, each tested against 6 oversampling strategies), a Soft Voting ensemble achieves the best overall result: MCC = 0.8060, AUC = 0.9703. Applying the Population Stability Index, Kolmogorov-Smirnov testing, and Jensen-Shannon divergence directly to quantify drift across their 48-hour window, they find only modest distributional shift consistent with ordinary intra-day variation rather than the long-horizon concept drift this review's broader discussion (Section 1.1, Section 4) is centrally concerned with — an important, honestly reported negative finding: drift is not always present just because a study sets out to measure it, and a 48-hour window is a genuinely different, shorter time horizon than the multi-year drift Al-Daoud and Abu-AlSondos (2025) study in Section 4.1 [10].

### **A directly comparable study, same benchmark, different explanatory lens**

Awad, Hamza, Sobh, and Bahaa-Eldin (2026) evaluate Logistic Regression, Random Forest, XGBoost, and LightGBM on the identical Kaggle credit-card dataset Zuberi *et al.* (2026) use in Section 3.1, alongside two additional, larger datasets (1.85 million and 24 million transactions respectively), deliberately excluding deep-learning architectures such as LSTMs and Transformers to preserve what they describe as consistent, interpretable explainability across all three datasets via SHAP applied at both global and local (individual-prediction) levels. On the shared Kaggle benchmark, their best model, Random Forest, reaches AUC = 0.9766 with a fraud-class F1-score of 0.8223 — a result directly comparable to, and marginally ahead of, Zuberi *et al.*'s AUC of 0.9703 on the same data, though the two studies use different model families, oversampling choices, and evaluation protocols, so the difference should be read as two independently strong results on the same benchmark rather than as a controlled head-to-head comparison. Their explicit choice to forgo deep learning in favor of tree-based ensembles specifically to preserve explainability is a direct empirical instance of the accuracy-interpretability trade-off this review's broader discussion (Section 6.2) returns to [11].

### **Benchmark scale and time horizon: what these numbers can and cannot tell us**

Sections 3.1 and 3.2's directly comparable AUC figures both come from a dataset spanning only 48 hours of transactions, and while Awad *et al.* (2026) additionally test their models on two much larger datasets (1.85 million and 24 million transactions respectively), neither of those larger datasets is described as spanning a longer real-world time horizon than the original Kaggle set — they are larger in transaction count and feature richness (geographic and merchant metadata; a synthetic 24-million-record benchmark), not necessarily in the number of days or months of drift they could expose. This matters directly for this review's argument: a model can achieve an excellent AUC on a large, feature-rich dataset while still being untested against the kind of multi-year behavioral and tactical drift that Section 1.3's real-world account describes and that only Al-Daoud and Abu-AlSondos's (2025) five-year study, reviewed next, actually spans. Readers should not infer from Sections 3.1 and 3.2's strong AUC figures alone that either study's model has been shown to resist long-horizon drift; both studies say so

themselves, directly or by omission, rather than this review imposing that caveat on their behalf.

## **Toward Adaptive, Drift-Resilient Systems**

### **A deployed hybrid framework combining all three pillars at once**

Al-Daoud and Abu-ALSondos (2025) provide the applied study in this review that most fully integrates adaptivity, explainability, and drift-awareness into a single system, evaluated not on a public benchmark but on five years (January 2020–December 2024) of real transaction data from Gulf Cooperation Council banking partners — a time span long enough to plausibly contain the kind of durable, multi-year drift Section 3.1's 48-hour window cannot speak to. Their hybrid machine-learning framework combines SMOTEBoost with cost-sensitive learning for class imbalance, adversarial training augmented with a purpose-built FraudGAN for synthetic adversarial simulation, the Drift Detection Method (DDM) and ADWIN for detecting concept drift specifically, and a three-part explainability layer combining SHAP, LIME, and structured human-in-the-loop review. Reported improvements are substantial and specific: fraud recall rises from a 35% baseline to 85%, adversarial attack success rate falls from 35% to 5%, drift recovery is achieved within 24 hours of a detected shift, operational latency stays below 150 milliseconds, and overall ROC-AUC reaches 0.93 using the hybrid imbalance-handling approach. The authors frame this explicitly as a regulation-compliant, auditable solution for high-risk financial contexts, directly addressing the accountability requirement Section 2.1's systematic survey identifies as a persistent barrier [12].

### **Adaptivity examined directly: streaming learning strategies and investigator-facing explanation**

Alessi and Fugini (2026) address the "adaptive" pillar of this review's title most directly of any source reviewed here, building a streaming machine-learning pipeline on a lightweight data-stream management system for real-time transaction monitoring, and explicitly comparing adaptive learning strategies along a spectrum from instance-incremental (updating the model after every single transaction) to batch-incremental (updating after accumulated batches) approaches under realistic operational constraints, including delayed feedback verification — the practical reality that a transaction's true fraud label is often confirmed only after a delay, complicating any adaptive-learning approach that assumes immediate ground truth. Their detection engine combines rule-based logic with adaptive models and dynamic threshold management, and their interpretability contribution is distinctive among this review's sources: rather than presenting SHAP or LIME output directly, they translate technical feature-importance metrics into what they describe as investigator-friendly concept-importance measures, aimed at a fraud analyst rather than a data scientist. This review was able to confirm the paper's authorship, venue, and abstract content directly, but the publisher blocked automated retrieval of the full text, so no single benchmark accuracy, precision, or AUC figure from this study is reported here; its contribution to this review is architectural and

methodological rather than a verified quantitative result, and it is discussed accordingly rather than placed alongside Sections 3.1, 3.2, and 4.1's directly comparable metrics.

### Delayed feedback: a practical constraint the benchmark studies do not face

Alessi and Fugini's (2026) explicit treatment of delayed feedback verification surfaces a constraint that Sections 3.1 and 3.2's benchmark evaluations, by construction, do not have to confront: on a fixed, already-labeled historical dataset, every transaction's true fraud status is known at evaluation time, but in a live deployment a transaction's true label is often confirmed only after a customer dispute, a chargeback, or a manual investigation concludes, sometimes days or weeks later. An adaptive model that updates itself between a prediction and the eventual arrival of that transaction's true label is, for a time, learning from an incomplete or provisional picture of its own recent accuracy — a genuine complication for the instance-incremental updating strategies Alessi and Fugini (2026) evaluate, and one that none of Sections 3.1, 3.2, or 4.1's studies discusses directly, since none of them describes updating its model incrementally in production against live, progressively-arriving labels [13-5].

### Comparative Synthesis

Table 1. All six sources reviewed

Reading this table alongside Sections 2 through 4, three columns matter more than they might first appear: the dataset column separates the two directly comparable public-benchmark studies (Zuberi *et al.*, 2026; Awad *et al.*, 2026) from the two studies evaluated on private, real institutional or operational data (Al-Daoud & Abu-AlSondos, 2025; Alessi & Fugini, 2026), which cannot be benchmarked against each other on the same numbers even though both address this review's "adaptive" and "drift-aware" pillars directly. The asterisk on Nguyen (2020) marks the one row in this table that is not a peer-reviewed study at all, included only for the real-world grounding Section 1.3 describes. Table 1 is therefore best read as five methodologically distinct pieces of evidence bearing on the same underlying question, not as five entries in a single ranked leaderboard.

| Study                          | Domain dataset /                                      | Approach                                                                                           | Key reported result                                                                                                                            |
|--------------------------------|-------------------------------------------------------|----------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------|
| Thivaivos <i>et al.</i> (2026) | Financial fraud detection — broadly systematic survey | PRISMA-guided review of ML/DL architectures, data modalities, and real-world deployment challenges | Identifies class imbalance, concept drift, adversarial behavior, and explainability as the field's persistent, interlinked deployment barriers |
| Zuberi <i>et al.</i> (2026)    | Credit-card fraud — Kaggle                            | 85 model combinations (12 classifiers, 3 MLPs,                                                     | Soft Voting ensemble: MCC = 0.8060, AUC = 0.9703;                                                                                              |

| Study                          | Domain dataset /                                                                                   | Approach                                                                                                                                                               | Key reported result                                                                                                                                                                                  |
|--------------------------------|----------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                | European dataset (284,807 transactions, 0.173% fraud)                                              | FraudNet, 7 ensembles × 6 oversampling methods); drift measured via PSI, KS test, Jensen–Shannon divergence                                                            | drift found modest (intra-day), not long-horizon concept drift                                                                                                                                       |
| Awad <i>et al.</i> (2026)      | Credit-card fraud — same Kaggle dataset, plus two additional datasets (1.85M and 24M transactions) | Logistic Regression, Random Forest, XGBoost, LightGBM compared; SHAP applied at global and local (per-prediction) levels                                               | Random Forest on Kaggle data: AUC = 0.9766, F1 = 0.8223 — directly comparable to Zuberi <i>et al.</i> 's AUC on the same benchmark                                                                   |
| Al-Daoud & Abu-AlSondos (2025) | GCC bank transactions, Jan 2020–Dec 2024 (private, real institutional data)                        | Hybrid framework: SMOTEBoost + cost-sensitive learning, adversarial training + FraudGAN, DDM/ADWIN drift detection, SHAP + LIME + human-in-the-loop review             | Fraud recall 35%→85%; adversarial attack success 35%→5%; drift recovery <24h; latency <150ms; AUC = 0.93                                                                                             |
| Alessi & Fugini (2026)         | Financial transaction streams — real-time operational setting                                      | Streaming ML pipeline; instance-incremental vs. batch-incremental adaptive learning strategies compared; investigator-facing concept-importance interpretability layer | No single benchmark accuracy figure reported in the abstract reviewed; contribution is architectural (adaptive learning strategy comparison + interpretability translation layer), disclosed as such |
| Nguyen (2020)*                 | Real-world grounding case — U.S. consumer payments                                                 | Industry commentary (American Banker), not a peer-reviewed study                                                                                                       | Documents deployed fraud models becoming unreliable as consumer spending patterns shifted abruptly and durably; cited for real-world grounding only (Section 1.3)                                    |

| Study | Domain dataset / | Approach | Key reported result |
|-------|------------------|----------|---------------------|
|       | during COVID-19  |          |                     |

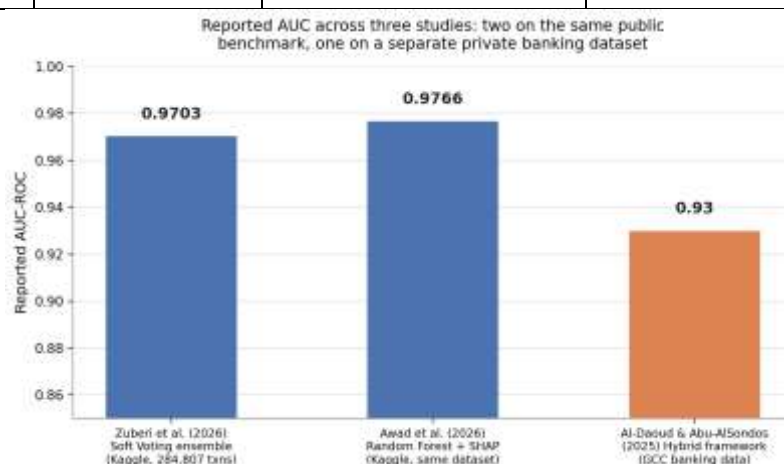


Figure 1. Reported AUC across the three studies in this review with a single, individually verified figure. Zuberi et al. (2026) and Awad et al. (2026) share the same public benchmark; Al-Daoud and Abu-AlSondos (2025) evaluate a separate private dataset, shown in a different color. Thivaos et al. (2026) is a survey without one figure of its own, and Alessi and Fugini's (2026) full text could not be retrieved (Section 4.2).

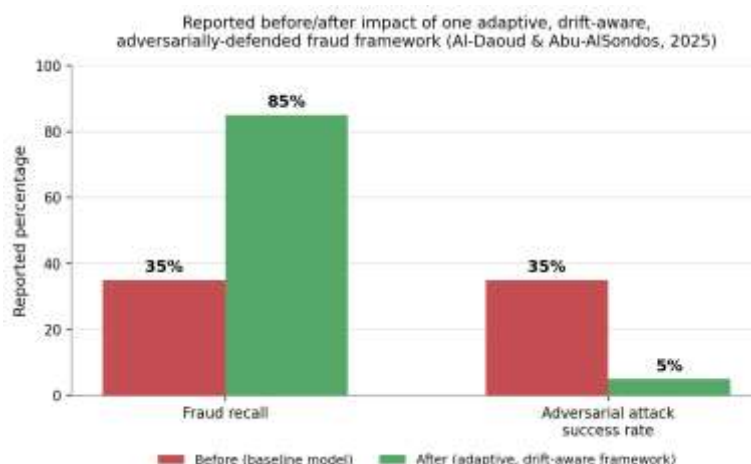


Figure 2. Before/after impact of Al-Daoud and Abu-AlSondos's (2025) hybrid adaptive, drift-aware, adversarially-defended framework, reported directly in the paper from five years of real GCC banking transaction data [16].

## Discussion

What "adaptive, explainable, drift-aware" actually requires, in combination

This review's title asks about all three properties at once, and Section 4.1 is the clearest evidence in the literature reviewed here that combining them is possible, not merely aspirational: Al-Daoud and Abu-AlSondos (2025) show a single deployed framework achieving drift detection, adversarial resilience, and multi-method explainability together, with each component's contribution reported separately. But their study is also the only one of the six that attempts all three at once on data spanning multiple years; Sections 3.1 and 3.2's two directly comparable, higher-AUC results are both measured on a 48-hour benchmark window too short to contain the durable drift Section 1.3's real-world account describes. The literature's strongest quantitative results and its strongest drift-handling demonstration currently come from different studies on different time horizons, which is itself a finding: a benchmark's AUC figure and a system's demonstrated resilience to multi-year drift are not the same claim, and this review's title's three pillars are not yet routinely validated together on the same long-horizon data.

### **Explainability is not one thing: three different audiences across these sources**

The six sources reviewed use explainability for at least three distinct purposes rather than one uniform practice. Zuberi *et al.* (2026) apply SHAP purely for feature selection, reducing 33 features to 15 before modeling — an internal, model-building use of an XAI technique rather than an explanation delivered to any end user. Awad *et al.* (2026) apply SHAP at both global and local levels explicitly to support what they frame as regulatory interpretability requirements — closer to a compliance-facing use. Alessi and Fugini (2026), by contrast, build an explicit translation layer converting technical feature metrics into concept-level explanations aimed at a human fraud investigator's workflow — a genuinely different design goal from either of the other two. Section 2.1's systematic survey treats "explainability" as one item on its list of deployment barriers, but Sections 3.1, 3.2, and 4.2 show that item resolving into at least three different technical answers depending on who the explanation is actually for, a distinction worth making explicit rather than treating "uses SHAP" as a single, comparable property across studies [17].

### **Regulatory context**

Financial institutions in major jurisdictions operate model risk under long-standing supervisory guidance — most notably SR 11-7, the joint Federal Reserve and Office of the Comptroller of the Currency guidance on model risk management, which requires ongoing monitoring for exactly the kind of performance degradation drift produces, not just pre-deployment validation. The GDPR's Article 22, where applicable, grants individuals a right to meaningful information about automated decisions that significantly affect them, which bears directly on Section 6.2's finding that explainability serves compliance-facing as well as internal purposes. None of the six sources reviewed here discusses SR 11-7 or GDPR Article 22 by name, which is a reasonable scope limitation for methods-focused technical research, but it leaves a real gap between what Section 2.1's survey identifies as a regulatory-compliance driver for explainability and any of these studies' own account of how their explanations would satisfy a specific regulatory requirement in practice.

## **Ethical considerations**

Section 1.3's real-world account of the COVID-19 disruption illustrates an equity dimension worth stating plainly: a fraud model destabilized by rapid behavioral change does not fail uniformly across a customer population. Nguyen (2020) specifically notes elevated false-positive risk for new and first-time customers and for behavior outside a model's established baseline — precisely the customers, including newly banked or newly online populations, least likely to have an existing track record a model can fall back on. Section 4.1's reported recall improvement from 35% to 85% is a substantial gain, but a 15-percentage-point residual miss rate, and whatever false-positive rate accompanies it, still falls unevenly on whichever customers' transaction patterns least resemble the training population — a distributional consequence none of the six sources reviewed quantifies directly.

## **Future research directions**

Three directions follow from the gaps this review has identified. First, Section 6.1's finding that the strongest AUC results (Sections 3.1, 3.2) and the strongest demonstrated multi-year drift resilience (Section 4.1) come from studies on different time horizons suggests an open need for a single study validating high accuracy, explainability, and multi-year drift resilience together on the same dataset. Second, given Section 6.2's finding that "explainability" already resolves into at least three distinct technical purposes across just these six sources, future work would benefit from stating explicitly which audience — model builder, compliance reviewer, or front-line investigator — an explanation method is designed to serve, rather than treating XAI as a single checklist item. Third, Section 4.2's inability to verify Alessi and Fugini's (2026) quantitative claims points to a broader, practical need for adaptive-learning research under realistic delayed-feedback conditions to be published in venues this review's tools can access and verify directly, since architectural contributions this promising are difficult for a review like this one to weigh against directly verified alternatives.

## **Toward a converged adaptive, explainable, drift-aware framework**

Pulling Sections 1 through 5 together, a genuinely converged system matching this review's title would need to combine at least four elements this review has found only separately across its six sources. First, direct, ongoing drift quantification of the kind Zuberi *et al.* (2026) demonstrate on a short window and Al-Daoud and Abu-ALSondos (2025) demonstrate over five years — not a one-time measurement at deployment, but a continuously updated check against the population the model is currently scoring. Second, an explanation layer matched to its actual audience, per Section 6.2's finding that model-builder-facing, compliance-facing, and investigator-facing explanations are three different technical answers, not one checklist item. Third, adaptive updating that accounts for delayed feedback rather than assuming immediate ground truth, the specific gap Section 4.3 identifies in Alessi and Fugini's (2026) otherwise-relevant streaming approach. Fourth, validation on a time horizon long enough to contain the kind of durable, tactic-level drift Section 1.3's real-

world account describes, which only Al-Daoud and Abu-AlSondos's (2025) study currently spans among this review's sources. No single study reviewed here combines all four elements with fully verified, independently checkable results; closing that gap — rather than any further improvement to benchmark AUC alone — is what would most directly advance the unified system this review's title names.

### **Cost asymmetry: a research question Section 3.1 raises but this review cannot fully resolve**

Zuberi *et al.* (2026) name the cost asymmetry between a missed fraud case and a false alarm as one of their study's three explicit research questions, alongside class imbalance and temporal drift, and their use of the F2 score rather than plain accuracy or F1 (giving recall roughly twice the weight of precision) is a direct methodological response to that asymmetry: for most financial institutions, letting a fraudulent transaction through is typically far costlier than briefly delaying or reviewing a legitimate one. None of this review's other five sources states its own assumed cost ratio explicitly, which makes any cross-study comparison of reported precision-recall trade-offs, including Section 4.1's recall improvement from 35% to 85%, incomplete in one specific respect: a recall gain achieved at an unstated cost in false positives cannot be judged as an unambiguous improvement without knowing what that institution's actual cost asymmetry is, information this review's sources do not uniformly report.

### **Conclusion**

Financial fraud detection research increasingly recognizes that adaptivity, explainability, and drift-awareness are not separable design choices but interdependent requirements of any system meant to keep working after deployment, and the evidence reviewed here — a systematic survey converging on the same three barriers, two directly comparable applied studies on a shared public benchmark, one deployed framework integrating all three pillars on real multi-year banking data, and one architecturally focused study on adaptive streaming strategies — supports that view. The clearest concrete finding is that the pieces exist and, in at least one case (Al-Daoud & Abu-AlSondos, 2025), have been combined with substantial reported success, but that no single study reviewed here validates high benchmark accuracy, multiple distinct forms of explainability, and multi-year drift resilience together on identical data. Building that combined, long-horizon validation — the genuinely unified adaptive, explainable, drift-aware system this review's title names — remains the clearest concrete direction for future work.

### **References**

- [1] Al-Daoud, K. I., & Abu-AlSondos, I. A. (2025). Robust AI for financial fraud detection in the GCC: A hybrid framework for imbalance, drift, and adversarial threats. *Journal of Theoretical and Applied Electronic Commerce Research*, 20(2), Article 121. <https://doi.org/10.3390/jtaer20020121>

- [2] Alessi, G., & Fugini, M. (2026). Adaptive real-time financial fraud detection with explainable AI tools. *Digital Threats: Research and Practice*.
- [3] Shahiduzzaman, M., Sourav, M. S. A., Kabtia, M., Shammah, R. S., Islam, M. A., Chakraborty, P., ... & Adnan, M. (2026). Attention-guided quantum-inspired extreme learning machine with grey wolf optimization for rare-attack detection in IIoT networks. *Discover Artificial Intelligence*.
- [4] Awad, S. S., Hamza, A. A., Sobh, M. A., & Bahaa-Eldin, A. M. (2026). Applying explainable artificial intelligence to interpret supervised ensemble learning models for robust credit card fraud detection. *Scientific Reports*, 16, Article 15220. <https://doi.org/10.1038/s41598-026-49939-5>
- [5] Nguyen, N. (2020, August 13). COVID-19 has upended fraud detection. *American Banker*.
- [6] Potharaju, B., & Ravipudi, J. (2025). Automated Data Validation and Resilient Infrastructure Design for Enterprise Regulatory Reporting Platforms. *International Journal of Computer Science and Engineering Innovations*, 1(1), 46-64.
- [7] Rahman, M. S., Islam, C. S., & Mudryi, M. (2025). Interoperability Standards for Real-Time Fraud-Signal Exchange Across US Financial Institutions: Design Principles and a Reference Implementation. *The Computertech*, 47-60.
- [8] Ahmed, K. R., Shammah, R. S., Kowser, S., Faruq, O., Sufian, M. A., & Ahmmmed, M. R. (2025, July). Strengthening Digital Security in MIS: A Business Analytics Approach to Deepfake Detection. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)* (pp. 1-6). IEEE.
- [9] Karim, N. B., & Kovac, T. (2025). Cloud-Based Big-Data Architectures for AI-Driven Financial Planning and Analysis and Early Risk Signaling in Resource-Constrained Community Financial Institutions. *The Metascience*, 3(3), 29-40.
- [10] Thivaivos, S., Kostopoulos, G., Stefani, A., & Kotsiantis, S. (2026). A survey of machine learning and deep learning for financial fraud detection: Architectures, data modalities, and real-world deployment challenges. *Algorithms*, 19(5), Article 354.
- [11] Rashid, S. U., & Sattiraju, S. R. (2025b). Multi-Institution Validation of Drift-Resilient Early-Warning Models for SME Loan Distress and Transaction Fraud: Evidence from US Community Financial Institution Pilots. *International Journal of Modern Computing*, 8(1), 40-54.
- [12] Zuberi, N., Rahman, M. M., Rashid, S. U., Ahsan, M. N., Rahman, M. S., & Mollah, M. H. O. R. (2026). A robust machine learning framework for detecting temporal drift in financial fraud prevention. *Scientific Reports*, 16, Article 27776.

- [13] Potharaju, B., & Ravipudi, J. (2023). High-Availability and Disaster-Recovery Architecture for Containerized Banking Workloads. *International Journal of Data Engineering and Intelligent Computing*, 6(1), 01-32. Potharaju & Ravipudi, 2023).
- [14] Rahman, M. S., Kabir, A. A., Reon, M. A. H., & Rashid, S. U. (2025). From Financial Transactions to Clinical Records: A Unified Machine Learning Methodology for Real-Time Anomaly Detection and Institutional Decision Support In High-Stakes Domains. *International Journal of Information Technology and Management Information Systems (IJITMIS)*, 16(6), 23-42.
- [15] Siddiqui, M. I. H., Rahman, M. S., Kabir, A. A., Mahmud, F. U., Rashid, S. U., & Shammah, R. S. (2023). Comparative analysis of explainable machine learning models for cancer classification using cytological features. *Journal of Medical and Health Studies*, 4(5), 110-150.
- [16] Rahman, M. S., Islam, C. S., Mudryi, M., & Cakmak, S. (2025). Concept-Drift-Resilient Ensemble Learning for Financial Transaction Fraud Detection Under Evolving Criminal Typologies. *International Journal of Acta Informatica*, 4(1), 52-67.
- [17] Karim, N. B., Tolosana, D., & Simko, R. (2025b). Responsible and Legally Compliant Enterprise Artificial Intelligence in Small-Business Lending: Governance, Explainability, and Fair-Lending Safeguards for Early-Intervention Recommendation Engines. *International Journal of Modern Computing*, 8(1), 55-66.