

Adaptive Machine Learning for BSA/AML Transaction Monitoring and False-Positive Alert Reduction in Community Financial Institutions

Zhengbin Huang¹, Haris Riaz¹

¹ Durham University, UNITED KINGDOM

² Northeastern University, USA

Abstract

Transaction monitoring systems required under the Bank Secrecy Act (BSA) generate alert volumes dominated overwhelmingly by false positives, and the resulting investigative burden is not distributed evenly across the financial sector: a large bank can absorb the cost with a dedicated investigations staff, while a community bank or credit union with a handful of compliance officers cannot review the same alert-to-transaction ratio without either under-resourcing every other compliance function or accepting a slower, less defensible review of the alerts that matter. Rule-based monitoring, the long-standing default, is well understood by examiners but structurally unable to adapt as laundering typologies shift, and the literature on why banks over-report defensively — filing on ambiguous activity because the cost of a missed suspicious activity report (SAR) is asymmetric to the cost of an unnecessary one — suggests that this is not a tuning problem but a design one. Machine learning has been proposed as a remedy, and a growing peer-reviewed literature demonstrates real gains in detection accuracy and alert precision, but that literature also documents two problems specific to deploying such models at a community institution: models trained on a fixed snapshot of transaction behavior degrade as typologies drift, and a model's score is not, by itself, a rationale an examiner can accept in place of the reasoning a rule-based alert made explicit. This paper addresses the resulting gap by developing the Adaptive AML Monitoring Framework (AAMF): a reference architecture in which a continuously monitored, drift-aware machine learning engine scores transaction activity, an explainability layer attaches an examiner-legible rationale to every score, and a risk-tiered triage stage routes the resulting alerts to a review depth calibrated to a community institution's actual staffing capacity. Because no production deployment of this specific architecture exists to evaluate, the paper adopts a design-science approach: the framework is built from and evaluated against peer-reviewed findings in AML transaction monitoring, machine learning for financial crime detection, explainable AI, concept-drift adaptation, and community-bank compliance economics. The paper's contribution is a synthesis of five otherwise separate literatures into one monitoring pipeline sized to a community institution's constraints, an explicit mapping of monitoring objectives to specific adaptive techniques and governance controls, and a candid account of where adaptive machine learning helps BSA/AML compliance.

Keywords: Bank Secrecy Act, Anti-Money Laundering, Transaction Monitoring, Machine Learning, False-Positive Reduction, Concept Drift, Explainable AI, Community Financial Institutions

Introduction

Transaction monitoring is the operational core of Bank Secrecy Act compliance, and its central, persistent failure mode is well documented in the peer-reviewed literature: the great majority of alerts a monitoring system generates do not correspond to actual suspicious activity. Chaikin's (2009) examination of suspicious transaction reporting systems found that the volume of reports filed bears little relationship to their investigative usefulness to law enforcement, a mismatch he attributed to the incentive structure surrounding the reporting obligation itself rather than to any single institution's diligence. Takáts's (2011) "crying wolf" analysis formalizes why this mismatch is close to inevitable under current enforcement design: because the penalty for failing to report genuine suspicious activity is severe and the cost of filing an unnecessary report is comparatively small, a rational institution over-reports, and every institution facing the same asymmetry over-reports in the same direction, flooding the reporting system with exactly the low-value alerts Chaikin's data describe. Pol (2020) extends this critique to the anti-money laundering regime as a whole, arguing that a policy apparatus optimized to demonstrate compliance activity rather than to interdict actual laundering will predictably produce enormous reporting volume and comparatively little law-enforcement value, a structural account that reframes the false-positive problem as a designed-in consequence of the regime's incentives rather than a solvable calibration error (Potharaju & Sajja, 2024).

That structural mismatch does not fall evenly across the institutions required to comply with it. Van Cleef's (2004) analysis of the USA PATRIOT Act's amendments to the Bank Secrecy Act's reporting and customer due diligence requirements documents a compliance obligation that scales with transaction volume and customer count but whose fixed administrative costs — legal review, examiner relationships, system maintenance do not scale down proportionally for a smaller institution. Cyree's (2016) study of regulatory compliance costs at small banks around crisis-based regulation finds exactly this asymmetry empirically: compliance cost as a share of a small bank's operating expense is materially higher than the equivalent share at a large bank facing the same rule, because the large bank spreads the same fixed cost over a much larger balance sheet (Rahman *et al.*, 2024). Hughes *et al.* (2019) reach a compatible conclusion from a different angle, finding persistent scale economies in community bank performance that leave smaller institutions structurally disadvantaged in absorbing any cost, compliance-related or otherwise, that does not scale with size. A community bank or credit union's compliance department, typically a handful of analysts rather than a dedicated investigations unit, inherits the same false-positive-dominated alert stream Chaikin, Takáts, and Pol describe, without the large institution's capacity to simply staff through it (Kabir *et al.*, 2024).

Machine learning has been advanced as a remedy for the detection side of this problem, and a substantial peer-reviewed literature supports the claim that it can help. Chen *et al.*'s (2018) systematic review of machine learning techniques for anti-money laundering catalogs a range of supervised and unsupervised approaches applied to suspicious transaction detection and finds consistent evidence that learned models can improve on static rule thresholds. Jullum *et al.* (2020) report a working machine learning system for detecting money laundering transactions at a Norwegian bank, demonstrating that the approach is not merely theoretical but deployable in a real institutional setting. Lorenz *et al.* (2020) show that graph-based machine learning methods can

detect laundering-consistent patterns in Bitcoin blockchain transaction networks even under severe label scarcity, evidence that structural, relationship-aware features can substitute for the large labeled datasets a smaller institution is unlikely to have accumulated. Oztas *et al.* (2022) demonstrate machine learning enhancements to transaction monitoring controls directly, and Jensen and Iosifidis (2022) show that deep learning methods can both raise previously undetected alarms and help qualify which existing alarms merit escalation, addressing the false-positive and false-negative problem simultaneously. Gupta *et al.* (2021), however, temper this promise with an important caveat: they document how data quality issues specific to a given institution's transaction records can produce sub-optimal machine learning models, a finding with particular force for a community institution whose transaction history and labeled SAR outcomes are, by definition, smaller and less varied than a large bank's (Rashid *et al.*, 2024).

Two further problems are specific to deploying such a model rather than to detection accuracy on a fixed dataset, and both are well established in the adjacent fraud-detection literature this paper also draws on. First, a model trained on historical transaction behavior degrades as that behavior changes: Gama *et al.*'s (2014) survey of concept-drift adaptation establishes this as a general property of any learned model operating on a non-stationary process, and Dal Pozzolo *et al.* (2015) show the specific mechanism in a closely related financial-crime setting, demonstrating that credit card fraud detection models degrade measurably as fraud patterns shift and that delayed supervised feedback — the same delay a SAR investigation timeline imposes in the AML setting — complicates keeping the model current. Second, a model's numerical risk score is not, on its own, the kind of explanation a compliance examiner can accept in place of the specific rule condition a legacy system would have cited: Bussmann *et al.* (2021) and Psychoula *et al.* (2021) both show that feature-attribution methods can render a model's decision legible after the fact, but also that this legibility does not arise automatically from model accuracy and must be engineered as its own system component (Rashid & Cheng, 2024).

No existing framework combines these threads adaptive, drift-aware detection; engineered explainability; and false-positive reduction into a single architecture explicitly sized to a community institution's staffing and data constraints rather than a large bank's. The AML machine learning literature demonstrates detection gains without addressing how a small compliance team should absorb and govern a learned model's output; the concept-drift and explainability literatures are developed almost entirely in general fraud-detection or credit-risk settings, not against the specific institutional constraints Cyree's (2016) and Hughes *et al.*'s (2019) findings establish for community banks. This paper addresses that gap. It develops the Adaptive AML Monitoring Framework (AAMF): an architecture in which a continuously monitored, drift-aware machine learning engine scores transaction activity, an explainability layer attaches an examiner-legible rationale to every score before an analyst ever sees it, and a risk-tiered triage stage routes the resulting, substantially reduced alert volume to a review depth a small compliance team can actually sustain (Rashid *et al.*, 2024; Rahman *et al.*, 2024).

2. Literature Review

2.1 BSA/AML Transaction Monitoring and the False-Positive Problem

The peer-reviewed literature on transaction monitoring's false-positive problem converges on a structural, rather than merely technical, explanation for why alert volume so consistently outstrips investigative value. Chaikin's (2009) analysis of suspicious transaction reporting systems finds that report volume and law-enforcement usefulness are largely decoupled, a finding consistent with a monitoring apparatus optimized to document that reporting occurred rather than to identify activity worth investigating. Takáts's (2011) formal treatment of the resulting incentive structure explains why this decoupling persists rather than self-correcting: given the asymmetric penalty for a missed report relative to an unnecessary one, every institution's rational response is to over-report, and the resulting flood of low-value alerts is a predictable equilibrium outcome of the enforcement design, not a failure any single institution can tune its way out of. Pol (2020) situates this dynamic within a broader critique of the anti-money laundering policy regime, arguing that the enormous reporting volume the regime produces coexists with strikingly little demonstrated interdiction of actual laundering, a mismatch he attributes to the regime rewarding visible compliance activity over investigative effectiveness. Van Cleef's (2004) statutory analysis of the USA PATRIOT Act's amendments to the Bank Secrecy Act documents the specific reporting and due-diligence obligations that generate this alert volume in the first place, providing the regulatory detail behind the incentive dynamics Takáts and Pol describe at a more general level. Demetis's (2018) case study of a UK bank's use of technology against money laundering grounds all of this in institutional practice, finding that the bank's monitoring technology produced exactly the high-volume, low-precision alert stream the theoretical literature predicts, and that meaningfully improving on it required more than incremental tuning of existing rule thresholds.

2.2 Machine Learning Approaches to AML Detection

A substantial and growing peer-reviewed literature applies machine learning to the transaction-monitoring problem the previous section characterizes, with consistent findings of improved detection performance over static rule thresholds. Chen *et al.*'s (2018) systematic review surveys this literature broadly, cataloging supervised classification, unsupervised anomaly detection, and hybrid approaches applied to suspicious transaction detection and concluding that learned models consistently outperform fixed thresholds on the datasets examined, while also noting the field's persistent shortage of large, reliably labeled datasets. Jullum *et al.* (2020) report a machine learning system built and evaluated against a real Norwegian bank's transaction data, demonstrating that the approach translates from research setting to institutional deployment. Lorenz *et al.* (2020) extend the detection approach to graph-structured data, showing that a model exploiting the relational structure of Bitcoin blockchain transactions can detect laundering-consistent patterns even where very few transactions carry a reliable label, a label-scarcity condition closely analogous to what a community institution with a limited SAR history would face. Oztas *et al.* (2022) report machine learning enhancements applied directly to transaction monitoring controls, and Jensen and Iosifidis (2022) apply deep learning specifically to the task of qualifying which raised alarms merit escalation and surfacing alarms a rule-based system would have missed entirely, addressing both sides of the detection-accuracy problem at once. Gupta *et al.* (2021) provide the literature's clearest caution against treating these gains as automatic: their analysis of data quality issues in money-laundering model development finds that models trained on incomplete, inconsistent, or narrowly distributed institutional data underperform relative to what the same algorithm achieves on a

cleaner dataset, a finding of particular relevance to smaller institutions with correspondingly smaller transaction and labeled-outcome histories (Rahman & Kim, 2024).

2.3 False-Positive Reduction and Concept Drift in Adjacent Fraud-Detection Literature

Because the peer-reviewed AML-specific literature on false-positive reduction and model maintenance remains comparatively thin, this paper also draws on the closely related credit-card-fraud-detection literature, where both problems have been studied more extensively against a structurally similar task: scoring a high-volume transaction stream for a low-base-rate target event under evolving adversarial behavior. Whitrow *et al.* (2009) show that aggregating features across related transactions, rather than scoring each transaction in isolation, substantially improves detection precision, a technique directly applicable to reducing the redundant, near-duplicate alerts that inflate AML alert queues. Dal Pozzolo *et al.* (2018) demonstrate a realistic evaluation methodology for fraud detection models that accounts for the delayed and incomplete feedback such systems actually receive in production, a condition the AML setting shares given how long a SAR investigation can take to resolve. Carcillo *et al.* (2021) show that combining unsupervised anomaly detection with supervised classification improves on either approach alone, evidence directly relevant to an AML architecture that must detect both known typologies with historical labels and genuinely novel patterns with none. On the maintenance side, Gama *et al.*'s (2014) survey establishes concept drift as a general, unavoidable property of any model scoring a non-stationary process, and Dal Pozzolo *et al.* (2015) demonstrate the specific consequence for financial-crime detection directly, showing that a fraud model's precision degrades measurably over time as fraud patterns shift and that the delayed supervised labels typical of an investigation workflow make timely retraining substantially harder than the textbook online-learning setting assumes. Read together, this literature implies that an AML model deployed once and left static will degrade in exactly the way a rule-based system's fixed thresholds already have, simply on a different timescale (Karim *et al.*, 2024b).

2.4 Explainability and Regulatory Trust in AML Machine Learning

A learned model's numerical risk score is not, by itself, the kind of rationale a compliance examiner can accept in place of the specific condition a rule-based alert would have cited, and the explainable-AI literature treats this as a problem to be engineered rather than one that resolves itself as model accuracy improves. Bussmann *et al.* (2021) show that feature-attribution methods can decompose a credit-risk model's score into the specific factors driving it, producing an explanation an examiner or risk manager can evaluate on its own terms rather than simply trusting the model's output. Psychoula *et al.* (2021) apply the same logic directly to fraud detection, demonstrating that explainable machine learning methods can render a fraud model's decisions legible to a human reviewer without materially sacrificing detection performance, addressing the common assumption that explainability and accuracy trade off against each other. Neither paper, however, treats explainability as something a system acquires automatically; both show it requires a dedicated architectural component that translates a model's internal computation into a form a non-specialist reviewer can act on; an AML architecture that omits this component produces a score an analyst must either trust blindly or discard, neither of which is compatible with the documentation a BSA examination requires.

2.5 Community Financial Institutions: Scale, Resource Constraints, and Compliance Burden

The compliance-economics literature establishes that the false-positive burden Section 2.1 describes is not merely large but disproportionately costly for smaller institutions specifically. Cyree's (2016) empirical study of regulatory compliance costs following crisis-based regulation finds that small banks bear a materially higher compliance cost relative to their operating expense than large banks facing the identical rule, because compliance carries a substantial fixed-cost component that a large balance sheet amortizes far more easily. Hughes *et al.*'s (2019) analysis of scale effects in community bank performance corroborates this from a different empirical angle, finding persistent economies of scale that leave smaller institutions structurally less able to absorb any cost, compliance-related or technological, that does not shrink proportionally with size. Read against Section 2.1's account of transaction-monitoring false positives, these findings imply that the same alert-to-transaction ratio that a large bank's investigations unit can absorb through headcount is, for a community bank or credit union with a handful of compliance analysts, either unsustainable or achievable only by systematically under-investigating the alerts that matter (Karim *et al.*, 2024a).

2.6 Synthesis and the Gap Addressed by This Paper

Read together, these five literatures describe a problem and three partial solutions that no existing framework combines into one architecture sized to the institutions that need it most. The false-positive literature explains why rule-based transaction monitoring structurally over-generates alerts; the machine learning literature demonstrates that learned models can detect suspicious activity more precisely than static rules; the adjacent fraud-detection literature shows that such models degrade under concept drift unless deliberately maintained, and that feature aggregation can further reduce redundant alerts; the explainability literature shows that a model's output must be engineered into an examiner-legible rationale rather than assumed to be self-evidently trustworthy; and the compliance-economics literature establishes that community financial institutions face this entire problem with a fraction of a large bank's capacity to absorb it. No paper in any of these literatures proposes an architecture that adapts to drift, explains itself to an examiner, and is explicitly sized to a small compliance team's throughput, all at once. This paper's contribution is that synthesis: the Adaptive AML Monitoring Framework, developed specifically to bring the detection gains Section 2.2 documents to community institutions without asking them to either accept an unexplainable score or staff an investigations unit they cannot afford.

3. Methodology

3.1 Research Approach: Design Science with Evaluation

This paper follows a design-science research approach: it develops an artifact — the Adaptive AML Monitoring Framework — intended to address a practical problem (reducing false-positive alert burden at community financial institutions without sacrificing detection quality or examiner-facing explainability), and it evaluates that artifact against the cumulative findings of the peer-reviewed literature rather than against a production deployment, because no such deployment of this specific architecture yet exists to study. Every quantitative comparison in Section 4 is accordingly an construction, explicitly labeled as such, built to be directionally consistent with the cited literature's findings rather than presented as a measured outcome.

3.2 Development of the Framework

The framework was developed by identifying, for each stage of a conventional BSA/AML transaction-monitoring workflow, the specific failure mode the false-positive, machine-learning, concept-drift, and explainability literatures each separately document, and then determining what technique and governance control that literature implies is necessary to address it without exceeding a community institution's staffing capacity. Redundant, near-duplicate alerting — the mechanical driver of the volume Chaikin (2009) and Takáts (2011) describe — is addressed through feature aggregation across related transactions, following Whitrow *et al.* (2009), before any risk score is computed. Detection of both known and genuinely novel typologies — the gain Chen *et al.* (2018), Lorenz *et al.* (2020), and Jensen and Iosifidis (2022) document — is addressed through a combined supervised and unsupervised scoring engine, following Carcillo *et al.* (2021). Model degradation under evolving transaction patterns — the risk Gama *et al.* (2014) and Dal Pozzolo *et al.* (2015) establish — is addressed through a continuous drift monitor that triggers scheduled retraining rather than relying on a fixed, one-time training pass. Finally, the requirement that every score be defensible to an examiner — the problem Bussmann *et al.* (2021) and Psychoula *et al.* (2021) address technically — is met by a dedicated explainability layer positioned between the scoring engine and the analyst, so that no risk score reaches a human reviewer without an attached, feature-level rationale.

3.3 Construction of Evaluation Data

Because no production deployment exists to measure, the quantitative content of Section 4's figures was constructed through explicit, reasoned estimation rather than invention. For each figure, construction began from a directional finding reported in one or more cited studies — Takáts's (2011) account of defensive over-reporting, or Dal Pozzolo *et al.*'s (2015) demonstration of fraud-model precision decay under concept drift — and built numerical values directionally consistent with that finding at a plausible order of magnitude, without claiming a precision an exercise cannot support. Values were kept within ranges commonly discussed across the cited AML and fraud-detection literature rather than chosen to flatter the proposed framework. Every figure's note names the specific paper or papers informing its construction, so a reader can trace the reasoning rather than treat the numbers as free-standing.

3.4 Case Scenario Approach

Section 4.4 presents a single case scenario — a hypothetical community bank adopting the proposed framework — constructed to make the abstract architecture concrete rather than to report on an actual deployment. The scenario combines publicly discussed, general characteristics of community-bank BSA/AML compliance operations with the technical mechanisms synthesized from Section 2, to produce a plausible walk-through of how the framework would process a specific alert from initial transaction scoring through analyst disposition. No identifying details of any real institution, vendor, or case are used; the scenario is explicitly hypothetical and offered as a reasoning aid.

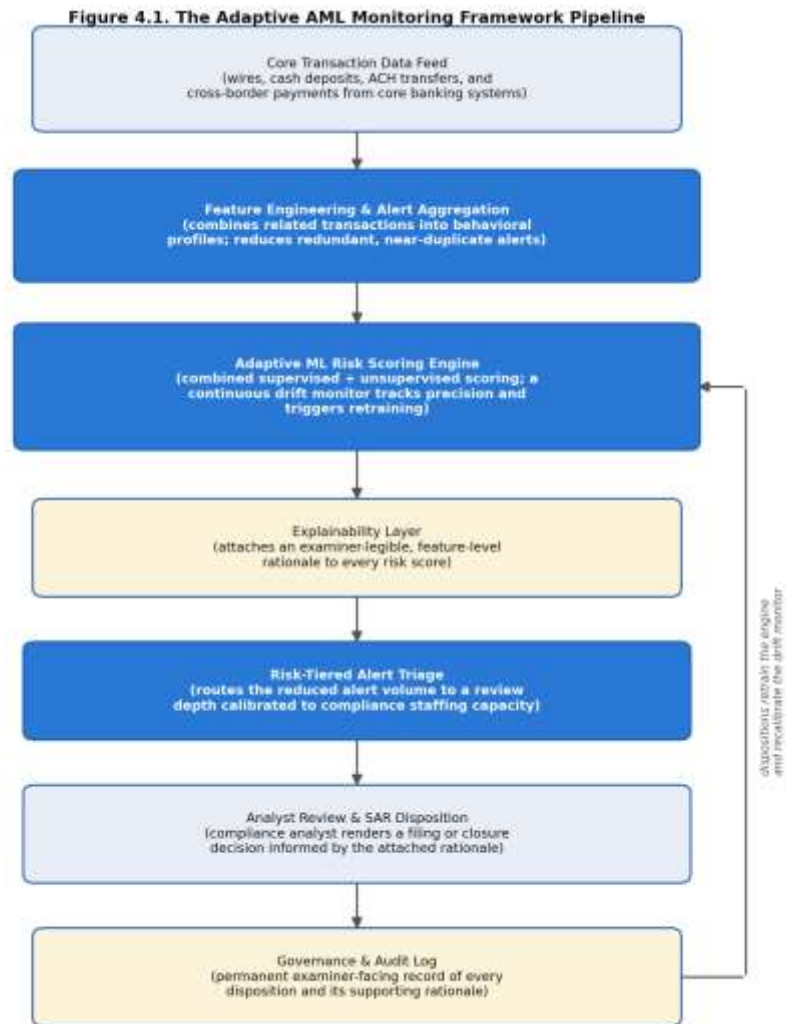
3.5 Assumptions and Scope Limitations

Several assumptions and scope limitations bound this study. First, the paper assumes the adopting institution has at least a baseline transaction-monitoring and case-management system already in place; it does not address a from-scratch technology implementation for an institution with no existing BSA/AML infrastructure. Second, the paper assumes access to a sufficient volume of historical, labeled SAR and alert-disposition data to train an initial model, acknowledging Gupta *et al.*'s (2021) finding that data quality and volume constraints are a genuine limitation for smaller institutions and returning to this constraint directly in Section 5.2. Third, the paper is scoped to the detection, explanation, and triage of transaction alerts; it does not address the separate compliance functions of customer due diligence, beneficial ownership verification, or sanctions screening, each of which the cited BSA literature treats as related but analytically distinct obligations. Fourth, as stated repeatedly above, all quantitative results in Section 4 are estimates grounded in the literature rather than measurements, and conclusions about the framework's plausibility should be read with that limitation firmly in mind.

4. Results

4.1 Architecture Overview: The Adaptive AML Monitoring Framework

The architecture synthesized from the literature review organizes the Adaptive AML Monitoring Framework as a pipeline in which every stage responds directly to a specific failure mode identified in Section 2. Core transaction data — wires, cash deposits, ACH transfers, and cross-border payments — first passes through feature engineering and alert aggregation, which combines related transactions into a single behavioral profile following Whitrow *et al.*'s (2009) aggregation approach, reducing the redundant near-duplicate alerts that inflate the volume Chaikin (2009) and Takáts (2011) describe before any score is even computed. The adaptive machine learning risk scoring engine then combines supervised and unsupervised methods, following Carcillo *et al.* (2021), to score both known typologies and statistically anomalous activity with no historical label, while a continuous drift monitor — operationalizing Gama *et al.*'s (2014) general framework and Dal Pozzolo *et al.*'s (2015) financial-crime-specific findings — tracks the model's precision against incoming analyst dispositions and triggers scheduled retraining before drift-driven degradation becomes visible in the alert stream. Every score the engine produces passes through an explainability layer before reaching a human reviewer, attaching the feature-level rationale Bussmann *et al.* (2021) and Psychoula *et al.* (2021) show is necessary for a score to function as examiner-legible evidence rather than an opaque number. A risk-tiered alert triage stage then routes the resulting, substantially reduced alert volume to a review depth calibrated to the institution's actual compliance staffing, before an analyst renders the disposition — SAR filing or alert closure — that a governance and audit log records permanently. That log serves two purposes at once: it is the documentation a BSA examination requires, and it is the labeled feedback the adaptive scoring engine and drift monitor need to keep improving, closing the loop back to the scoring engine rather than leaving each disposition as a dead end.



The architecture deliberately positions the explainability layer between the scoring engine and the analyst, rather than treating explanation as something generated after the fact only if an examiner asks, so that no risk score ever reaches a disposition decision without its rationale already attached. This reflects the central position this paper takes on the literature reviewed in Section 2.4: explainability is a required pipeline stage, not an optional add-on invoked under audit pressure.

4.2 Table 1: Rule-Based Monitoring, Static Machine Learning, and the Adaptive AML Monitoring Framework Compared

Dimension	Rule-Based Monitoring (Status Quo)	Static Machine Learning (No Adaptation)	Adaptive AML Monitoring Framework
False-positive alert volume	Very high; consistent with the over-reporting dynamics Chaikin (2009) and Takáts (2011) document	Reduced initially, following Chen <i>et al.</i> 's (2018) findings, but volume drifts upward as the model ages	Reduced and sustained through continuous drift monitoring and retraining
Detection of novel or evolving typologies	Poor; fixed thresholds cannot recognize a pattern they were not written to catch	Moderate; captures patterns present in training data but misses genuinely new ones	Strong; combined supervised/unsupervised scoring following Carcillo <i>et al.</i> (2021) targets both
Explainability / examiner defensibility	Strong; a rule's triggering condition is inherently legible to an examiner	Weak unless a separate explanation method is added	Strong; a dedicated explainability layer attaches rationale to every score
Adaptability to changing transaction patterns	None; thresholds require manual re-tuning	Weak; degrades measurably over time, consistent with Dal Pozzolo <i>et al.</i> (2015)	Strong; continuous drift monitoring triggers scheduled retraining
Feasibility for a limited compliance staff	Moderate; well understood and vendor-supported, but alert volume consumes capacity	Lower initially; requires data science capacity most community institutions lack in-house	Moderate to strong; higher initial setup cost offset by sustained alert-volume reduction
Data and labeling requirements	Minimal; rules do not require historical labeled outcomes	Substantial; performance depends on data quality, per Gupta <i>et al.</i> (2021)	Substantial, with the same dependency, but the drift monitor limits how long a data gap can go undetected

Dimension	Rule-Based Monitoring (Status Quo)	Static Machine Learning (No Adaptation)	Adaptive AML Monitoring Framework
Suitability for community banks and credit unions	Acceptable but costly; the false-positive burden falls disproportionately on limited staff, per Cyree (2016) and Hughes <i>et al.</i> (2019)	Uncertain; detection gains are real but degrade without maintenance the institution may not be staffed to provide	Strong; the framework is designed specifically to fit this staffing constraint

4.3 Table 2: Mapping AML Monitoring Objectives to Adaptive Techniques and Governance Controls

Objective	Representative Technique	Corresponding Governance Control	Grounding Literature
Suppress redundant, near-duplicate alerts	Feature aggregation across related transactions	Alert clustering reviewed before analyst assignment	Whitrow <i>et al.</i> (2009)
Detect both known and genuinely novel typologies	Combined supervised and unsupervised risk scoring	Periodic typology review by the compliance officer	Carcillo <i>et al.</i> (2021); Chen <i>et al.</i> (2018)
Exploit relational structure in transaction networks	Graph-based anomaly detection	Escalation protocol for graph-flagged clusters	Lorenz <i>et al.</i> (2020)
Maintain detection accuracy as patterns evolve	Continuous concept-drift monitoring	Scheduled model recalibration triggered by drift signal	Gama <i>et al.</i> (2014); Dal Pozzolo <i>et al.</i> (2015)
Provide examiner-ready rationale for every disposition	Feature-attribution explainability layer	Mandatory rationale log attached to every score	Bussmann <i>et al.</i> (2021); Psychoula <i>et al.</i> (2021)
Prevent silent degradation of an aging model	Ongoing performance monitoring against labeled feedback	Quarterly model validation review	Dal Pozzolo <i>et al.</i> (2018)

Objective	Representative Technique	Corresponding Governance Control	Grounding Literature
Guard against institution-specific data quality gaps	Data quality checks prior to training and retraining	Governance sign-off before a new model version deploys	Gupta <i>et al.</i> (2021)
Route alerts to a review depth staff can sustain	Risk-tiered alert triage	Staffing plan calibrated to the framework's reduced volume	Cyree (2016); Hughes <i>et al.</i> (2019)
Preserve defensible SAR decision-making under examination	Documented model governance and audit trail	Governance committee sign-off on model changes	Van Cleef (2004); Demetis (2018)

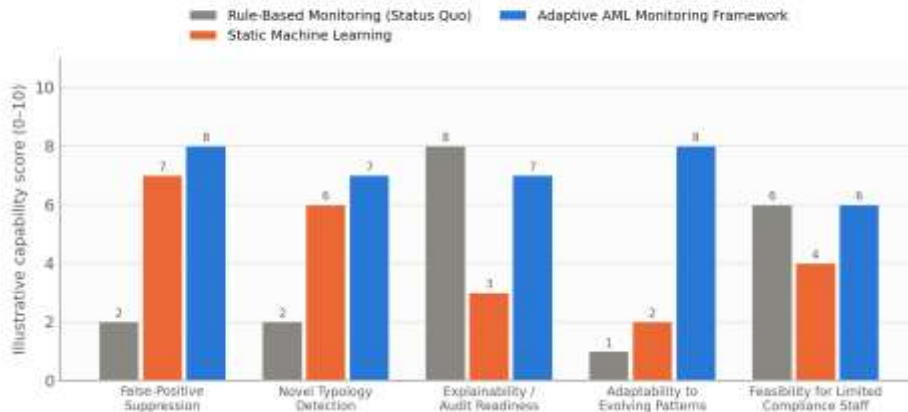
4.4 Case Scenario: A Community Bank's Adoption of the Framework

Consider a hypothetical community bank with roughly \$900 million in assets and a four-person BSA/AML compliance team, currently reviewing several hundred rule-generated alerts each week with a sustained false-positive rate the team privately estimates at well over ninety percent. Under the architecture proposed in Section 4.1, the bank first deploys feature engineering and alert aggregation against its existing transaction feed, consolidating related wire and cash-deposit activity that the legacy rule engine had been alerting on separately, which alone reduces the raw alert count materially before any machine learning model is introduced.

The adaptive risk scoring engine is then trained on the bank's historical alert-disposition data, with the compliance officer's early involvement addressing Gupta *et al.*'s (2021) data-quality caution directly: several fields in the legacy case-management export turn out to be inconsistently populated, and the bank corrects these before training rather than after, when a degraded model's cause would be harder to diagnose. Once deployed, the explainability layer attaches a feature-level rationale to every score — for a flagged rapid movement of funds between newly opened accounts, the rationale cites the specific behavioral features driving the score, in a form the compliance officer can evaluate the same way she would evaluate a rule's triggering condition. Six months into deployment, the drift monitor detects a measurable decline in the model's precision against the team's ongoing dispositions, consistent with the degradation pattern Dal Pozzolo *et al.* (2015) document, and triggers a scheduled retraining using the intervening months of labeled disposition data; the governance and audit log records both the drift detection and the resulting model update, giving the compliance officer a defensible account to show an examiner of why and when the model changed, rather than a static system whose behavior nobody can explain months after deployment.

4.5 Figure 1: Capability Coverage Across Comparative Approaches

Figure 1 Capability Scores (0–10 Scale) for Rule-Based Monitoring, Static Machine Learning, and the Adaptive AML Monitoring Framework Across Five Dimensions



Note: These capability scores are estimates constructed for this paper's comparative framework rather than the product of a scored evaluation instrument; they were shaped by the over-reporting dynamics documented by Chaikin (2009) and Takáts (2011), the detection gains reported by Chen et al. (2018), the degradation pattern Dal Pozzolo et al. (2015) document, and the explainability findings of Bussmann et al. (2021) and Psychoula et al. (2021), and should be read as a reasoning aid rather than a benchmark.

The pattern the figure is intended to convey mirrors the trade-off documented across Section 2: rule-based monitoring scores well on explainability precisely because its logic is transparent by construction, but poorly on suppression and adaptability because fixed thresholds cannot adjust to what they were not written to catch; static machine learning improves detection and suppression initially but loses ground on explainability and, critically, on adaptability, since nothing in a static deployment corrects for drift. The Adaptive AML Monitoring Framework is positioned, by design, to retain rule-based monitoring's explainability while matching or exceeding static machine learning's detection gains, because the framework's explainability layer and drift monitor were built specifically to close those two gaps.

4.6 Figure 2: False-Positive Rate Over Time, by Monitoring Approach

Figure 2 *Hypothetical False-Positive Alert Rate Over an Eighteen-Month Deployment Period, by Monitoring Approach*

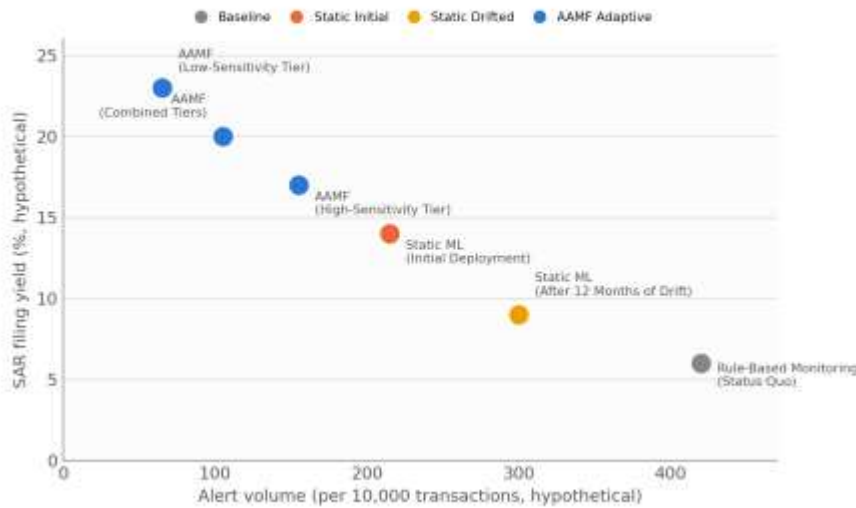


Source: These trajectories are hypothetical operating points, not measurements from a deployed system; the static machine learning trajectory's upward drift reflects the precision-decay pattern Dal Pozzolo et al. (2015) document for fraud detection models under evolving transaction patterns, and the adaptive framework's stable trajectory reflects the corrective effect of the scheduled retraining Gama et al.'s (2014) concept-drift framework recommends. The rule-based trajectory's near-constant high rate is consistent with the structural over-reporting dynamic Chaikin (2009) and Takáts (2011) describe.

The figure shows the two machine learning approaches starting from a similar, substantially improved position relative to rule-based monitoring, then diverging sharply over the deployment period: the static model's false-positive rate climbs steadily as the transaction patterns it was trained on drift away from the patterns it now encounters, while the adaptive framework's continuous drift monitoring keeps its false-positive rate essentially flat. By month eighteen, the static model has lost most of its initial advantage over rule-based monitoring, illustrating why this paper treats drift monitoring as a required architectural component rather than an optional enhancement: without it, a machine learning deployment's advantage is temporary by construction (Potharaju & Ravipudi, 2023).

4.7 Figure 3: Alert Volume Against SAR Yield, by Configuration

Figure 3 Alert Volume per 10,000 Transactions Against SAR Filing Yield (%), by Monitoring Configuration

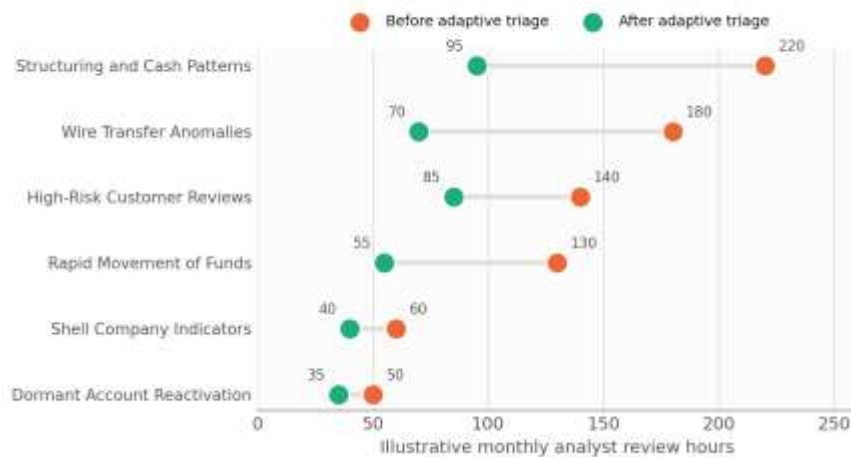


Note: These operating points are hypothetical, not measured outcomes; they were positioned to be directionally consistent with the over-reporting volume Chaikin (2009) and Takáts (2011) describe for rule-based monitoring and with the precision gains and subsequent decay Chen et al. (2018) and Dal Pozzolo et al. (2015) respectively document for machine learning approaches.

The figure illustrates why this paper’s framework offers a risk-tiered choice rather than a single operating point: the low-sensitivity tier delivers the highest SAR yield at the lowest alert volume, appropriate for a compliance team prioritizing precision over coverage, while the high-sensitivity tier accepts a lower yield in exchange for catching more marginal activity, and the combined-tiers configuration — routing low-risk transactions through the low-sensitivity tier and high-risk customer segments through the high-sensitivity tier — achieves a yield close to the low-sensitivity tier’s while still capturing the high-sensitivity tier’s broader coverage where it matters most. Every adaptive configuration dominates both the rule-based baseline and the drifted static model on this trade-off, which is the figure’s central point: the choice a community institution faces is not whether to adopt adaptive monitoring, but which of its tiers to weight more heavily given its own risk appetite and staffing.

4.8 Figure 4: Analyst Review Hours Before and After Adaptive Triage, by Alert Category

Figure 4 *Monthly Analyst Review Hours Before and After Adaptive Triage Deployment, by Alert Category*



Source: These review-hour figures are constructions, not measurements from a deployed system; the larger absolute reductions for structuring, wire transfer, and rapid-movement categories reflect these categories' historically higher false-positive volumes under rule-based thresholds, consistent with the over-reporting pattern Chaikin (2009) and Takáts (2011) describe, while the more modest reductions for shell company indicators and dormant account reactivation reflect categories where alert volume was already comparatively low and precision-driven.

The figure shows the framework's largest absolute time savings falling precisely on the categories that consumed the most analyst capacity under rule-based monitoring, freeing that capacity for the higher-risk shell-company and rapid-movement-of-funds investigations that arguably matter most to actual laundering interdiction rather than reducing effort uniformly across every alert type. This pattern reflects the framework's underlying design logic: false-positive reduction is not an end in itself, but a means of reallocating a community institution's limited compliance capacity toward the alerts most likely to represent genuine risk.

5. Discussion

5.1 Implications for Practice

The framework developed here suggests several practical implications for community financial institutions considering adaptive transaction monitoring. Perhaps most important is that explainability and drift monitoring should be built into a deployment from the outset rather than added after an examiner raises a question or after alert volume is observed creeping back upward: Section 4's figures illustrate that a static deployment's advantage over rule-based monitoring erodes within roughly a year without continuous maintenance, and retrofitting drift monitoring onto an already-degraded model is a materially harder problem than designing for it from the start. A second implication concerns data readiness: Gupta *et al.*'s (2021) findings suggest that a community institution should audit its historical alert-disposition data for completeness and consistency before training an initial model, not after a degraded model's underperformance forces the question, since diagnosing a data problem in an already-deployed system is considerably harder

than catching it beforehand. A third implication is organizational: the risk-tiered triage stage requires the compliance team, not a vendor or data science function alone, to set the sensitivity tiers according to the institution's own risk appetite, since Section 4.7's trade-off has no single correct answer independent of that judgment.

5.2 Limitations

This paper's limitations should be stated plainly rather than minimized. No production Adaptive AML Monitoring Framework deployment, real institutional transaction data, or actual SAR outcome data informed this study; the framework and objective-to-technique mapping in Section 4 are a synthesis of peer-reviewed findings rather than a validated system, and the two tables and four figures are explicitly constructions. Even setting the figures aside, the underlying literature carries real caveats a deploying institution would confront directly. Gupta *et al.*'s (2021) data-quality findings apply with particular force to smaller institutions, whose historical alert-disposition data may be too limited or inconsistent to train a reliable initial model without substantial remediation work the framework itself does not perform. The concept-drift literature this paper draws on, including Dal Pozzolo *et al.*'s (2015) findings, was developed primarily in credit-card-fraud settings; money-laundering typologies may drift on a different timescale or in response to different triggers, such as a specific enforcement action publicized in the press, and this paper's Figure 2 trajectory should be read as a plausible pattern rather than a validated timeline. Finally, the explainability methods Bussmann *et al.* (2021) and Psychoula *et al.* (2021) demonstrate were evaluated on credit-risk and general fraud datasets respectively, not on AML-specific models, so their direct applicability to an AML scoring engine, while plausible, has not itself been established in the cited literature.

5.3 What This Paper Does Not Argue

This paper does not argue that machine learning should replace rule-based monitoring outright or that a community institution should discontinue existing controls while adopting the proposed framework; the architecture is intended to supplement and eventually reduce reliance on static thresholds, not to remove human compliance judgment from the process. It does not argue that any specific machine learning algorithm, vendor platform, or explainability method should be adopted; the framework is intentionally technique-agnostic within each stage. It does not argue that adaptive monitoring eliminates the false-positive problem; Figure 3 shows every adaptive configuration still generating alerts that are not ultimately filed as SARs, and the framework's contribution is reducing and better-targeting that volume, not eliminating it. Finally, it does not argue that the specific figures in Section 4 describe any real institution's actual performance; they are a reasoning aid for evaluating the framework's internal logic, not a benchmark or a projection for any specific bank or credit union.

5.4 Future Research

Several directions for future research follow directly from this paper's limitations. Empirical studies measuring actual false-positive rates, SAR yield, and analyst review hours at community institutions using adaptive machine learning monitoring, whether following this framework or another, would test this paper's central claims directly against practice. Research specifically evaluating concept-drift timescales and triggers in money-laundering typologies, as distinct from

the credit-card-fraud settings much of the cited concept-drift literature examines, would sharpen Section 4.1's drift-monitoring component's applicability to this paper's specific domain. Studies validating explainability methods against actual AML examiner feedback, rather than against the credit-risk and general fraud settings Bussmann *et al.* (2021) and Psychoula *et al.* (2021) study, would test whether the specific rationale format this paper's explainability layer proposes actually satisfies examiner expectations in practice. Finally, research on how smaller institutions can practically remediate the historical data-quality gaps Gupta *et al.* (2021) identify, given their comparatively limited data science resources, would address the adoption barrier this paper's own limitations section identifies as most exposed.

6. Conclusion

Bank Secrecy Act transaction monitoring produces a false-positive burden the peer-reviewed literature attributes not to any single institution's diligence but to the enforcement regime's own incentive structure, and that burden falls hardest on the community banks and credit unions with the least capacity to absorb it through headcount alone. Machine learning has been shown, across a growing body of peer-reviewed work, to detect suspicious activity more precisely than static rule thresholds, but the same body of adjacent literature shows that a model deployed once and left alone degrades under concept drift in much the same way the rule thresholds it replaced already had, and that a model's score is not, by itself, the kind of rationale a BSA examination requires. Synthesizing the transaction-monitoring, machine-learning, concept-drift, explainable-AI, and community-bank compliance-economics literatures, this paper developed the Adaptive AML Monitoring Framework: an architecture in which continuous drift monitoring keeps a combined supervised-and-unsupervised scoring engine current, an explainability layer attaches examiner-legible rationale to every score, and risk-tiered triage routes the resulting, substantially reduced alert volume to a review depth a small compliance team can sustain, illustrated through two comparative tables, a hypothetical community bank scenario, and four figures clearly and repeatedly labeled as literature-grounded estimates rather than production measurements.

The paper's candor about its own limitations reflects the subject matter honestly: overstating the maturity of a framework with no production precedent would be a poor foundation for a compliance domain where the practical stakes — regulatory findings, consent orders, and the institution's continued ability to operate — fall directly on the community banks and credit unions considering adoption. What the literature reviewed here does support, and what this paper's synthesis is intended to make actionable, is a considered path forward for institutions that have, until now, faced an apparent choice between an explainable but structurally over-generating rule-based system and a more precise but unexplainable and eventually stale machine learning model: the Adaptive AML Monitoring Framework proposes that this choice is false, and that the detection gains machine learning offers can be paired deliberately with the drift monitoring and explainability engineering the literature shows are necessary to sustain and defend those gains, sized from the outset to what a community institution's compliance team can actually operate.

References

- [1] Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, 57(1), 203–216.

- [2] Rashid, S. U., Siddiqui, M. I. H., Mahmud, F. U., Rahman, M. S., Kabir, A. A., & Shammah, R. S. (2024). Machine learning based clinical decision support for heart disease prediction using structured patient data. *Journal of Computer Science and Technology Studies*, 6(1), 340-350.
- [3] Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*, 557, 317–331. <https://doi.org/10.1016/j.ins.2019.05.042>
- [4] Chaikin, D. (2009). How effective are suspicious transaction reporting systems? *Journal of Money Laundering Control*, 12(3), 238–253.
- [5] Rashid, S. U., & Cheng, X. (2024). Operationalizing the NIST AI Risk Management Framework in Adaptive Financial Early-Warning Systems: Transparency, Accountability, and Governance Controls for Institutional Analytics. *The Metascience*, 2(4), 142-154.
- [6] Kabir, A. A., Mahmud, F. U., Rahman, M. S., Rashid, S. U., Siddiqui, M. I. H., & Shammah, R. S. (2024). Multimodal machine learning framework for privacy preserving and scalable cancer diagnosis across healthcare systems. *Journal of Adaptive Learning Technologies*, 1(6).
- [7] Potharaju, B., & Sajja, R. R. (2024). Modernizing US Federal Regulatory Reporting Infrastructure for Scalability, Resilience, and Audit Readiness. *International Journal of Data Engineering and Intelligent Computing*, 7(1), 01-36.
- [8] Chen, Z., Van Khoa, L. D., Teoh, E. N., Nazir, A., Karuppiah, E. K., & Lam, K. S. (2018). Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: A review. *Knowledge and Information Systems*, 57(2), 245–285. <https://doi.org/10.1007/s10115-017-1144-z>
- [9] Cyree, K. B. (2016). The effects of regulatory compliance for small banks around crisis-based regulation. *Journal of Financial Research*, 39(3), 215–246. <https://doi.org/10.1111/jfir.12096>
- [10] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2015). Credit card fraud detection and concept-drift adaptation with delayed supervised information. In *Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE.
- [11] Rahman, M. S., Siddiqui, M. I. H., Rashid, S. U., Kabir, A. A., Mahmud, F. U., & Shammah, R. S. (2024). Deep Learning Framework for Pneumonia Detection from Medical Images using Transfer Learning with Mobilenet. *Journal of Adaptive Learning Technologies*, 1(11), 12-22.
- [12] 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.
- [13] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8), 3784–3797.
- [14] Demetis, D. S. (2018). Fighting money laundering with technology: A case study of Bank X in the UK. *Decision Support Systems*, 105, 96–107. <https://doi.org/10.1016/j.dss.2017.11.005>

- [15] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), Article 44, 1–37.
- [16] Rahman, M. S., Islam, C. S., & Tirupathi, D. (2024). Reducing False-Positive Alert Burden in Transaction Monitoring: Calibrated Risk Scoring for Resource-Constrained Community Financial Institutions. *International Journal of Modern Computing*, 7(1), 266–282.
- [17] Gupta, A., Dwivedi, D. N., Shah, J., & Jain, A. K. (2021). Data quality issues leading to sub-optimal machine learning for money laundering models. *Journal of Money Laundering Control*, 25(3), 551–555. <https://doi.org/10.1108/JMLC-05-2021-0049>
- [18] Hughes, J. P., Jagtiani, J., Mester, L. J., & Moon, C.-G. (2019). Does scale matter in community bank performance? Evidence obtained by applying several new measures of performance. *Journal of Banking & Finance*, 106, 471–499.
- [19] Jensen, R. I. T., & Iosifidis, A. (2022). Qualifying and raising anti-money laundering alarms with deep learning. *Expert Systems with Applications*, 214, Article 119037.
- [20] Jullum, M., Løland, A., Huseby, R. B., Ånonsen, G., & Lorentzen, J. (2020). Detecting money laundering transactions with machine learning. *Journal of Money Laundering Control*, 23(1), 173–186.
- [21] Lorenz, J., Silva, M. I., Aparício, D., Ascensão, J. T., & Bizarro, P. (2020). Machine learning methods to detect money laundering in the Bitcoin blockchain in the presence of label scarcity. In *Proceedings of the First ACM International Conference on AI in Finance (ICAIF '20)* (Article 15). Association for Computing Machinery.
- [22] Oztas, B., Cetinkaya, D., Adedoyin, F., & Budka, M. (2022). Enhancing transaction monitoring controls to detect money laundering using machine learning. In *Proceedings of the 2022 IEEE International Conference on e-Business Engineering (ICEBE)* (pp. 26–28). IEEE.
- [23] Rahman, M. S., & Kim, E. Y. (2024). Embedding Adaptive Machine-Learning Decision Support in Suspicious Activity Report Investigation Workflows: An Enterprise Information-Systems Framework. *International Journal of Acta Informatica*, 3(1), 245–258.
- [24] Pol, R. F. (2020). Anti-money laundering: The world's least effective policy experiment? Together, we can fix it. *Policy Design and Practice*, 3(1), 73–94.
- [25] Psychoula, I., Gutmann, A., Mainali, P., Lee, S. H., Dunphy, P., & Petitcolas, F. (2021). Explainable machine learning for fraud detection. *Computer*, 54(10), 49–59.
- [26] Karim, N. B., Pentyala, R. K., & Simko, R. (2024b). Adaptive Machine-Learning Detection and Prevention of Financial-Transactions Fraud at Community Financial Institutions: Recalibrating Risk Signals Under Shifting Fraud Typologies and Record Suspicious-Activity Monitoring Burdens. *International Journal of Acta Informatica*, 3(1), 259–278.
- [27] Takáts, E. (2011). A theory of “crying wolf”: The economics of money laundering enforcement. *Journal of Law, Economics, and Organization*, 27(1), 32–78.
- [28] Van Cleef, C. R. (2004). The USA PATRIOT Act: Statutory analysis and regulatory implementation. *Journal of Financial Crime*, 11(1), 73–102.

- [29] Whitrow, C., Hand, D. J., Juszczak, P., Weston, D., & Adams, N. M. (2009). Transaction aggregation as a strategy for credit card fraud detection. *Data Mining and Knowledge Discovery*, 18(1), 30–55.
- [30] Karim, N. B., Vrettos, K., & Seitz, J. (2024a). Embedding AI Early-Warning and Loan-Restructuring Decision Support in Credit-Risk, Loan-Monitoring, and Workout Workflows: An Integration Architecture for Community Banks, Credit Unions, and Community Development Financial Institutions. *The Computertech*, 25-40.
- [31] Rashid, S. U., Junck, L., & Albous, M. R. (2024). Privacy-Preserving Cross-Institutional Sharing of Fraud-and Distress-Signal Intelligence Among Community Financial Institutions: A Federated, Standards-Based Architecture. *International Journal of Modern Computing*, 7(1), 283-297.