Developing A Conceptual Framework For AI-Assisted Compliance Monitoring and Anomaly Detection in High-Volume Transaction Environments

[1]Michael Ominyi, [2]Cyril Chimelie Anichukwueze, [3]Ngozi Samuel Uzougbo
[1]Ernst & Young (EY), EMEA & United States;
michael.c.ominyi@gmail.com
[2]The University of Texas at Dallas,
United States; cyrilanichukwueze@gmail.com
[3]Independent Researcher, USA ; Ngozi.uzougbo@gmail.com
Corresponding Author; michael.c.ominyi@gmail.com
DOI: 10.56201/wjimt.v8.no6.2024.pg204.228

Abstract
Financial institutions and payment platforms now process transaction volumes that routinely exceed the practical capacity of rule-based compliance monitoring. This paper develops a conceptual framework for embedding artificial intelligence into compliance monitoring and anomaly detection within high-volume transaction environments. Drawing on literature spanning anti-money laundering supervision, fraud detection, graph-based machine learning, federated learning, concept drift adaptation, and financial model risk governance, the framework organizes the compliance monitoring function into six interdependent layers: data ingestion and integration, feature engineering and entity resolution, a hybrid detection engine, explainability and risk scoring, human-in-the-loop review, and governance with continuous feedback. The framework treats detection accuracy and regulatory defensibility as joint design objectives rather than competing priorities, positioning explainability and human oversight as structural components rather than after-the-fact controls. The paper situates the framework against existing regulatory expectations, including the Federal Reserve's SR 11-7 model risk guidance and the European Union's Artificial Intelligence Act, discusses sector-specific extensions to digital assets, insurance, and cross-institutional consortia, illustrates the framework through a hypothetical high-volume payments scenario, and compares it against alternative approaches. It concludes by identifying the limitations of a purely conceptual contribution and outlining directions for empirical validation.

Keywords: *artificial intelligence, compliance monitoring, anomaly detection, anti-money laundering, transaction monitoring, model risk management, explainable AI, federated learning, concept drift*

1. Introduction

The volume, velocity, and channel diversity of financial transactions have expanded sharply with the growth of real-time payments, embedded finance, and cross-border digital commerce. Compliance functions built around static, rule-based monitoring were designed for a slower, lower-volume environment and are increasingly strained by this shift. Rule-based scenarios

operate on fixed thresholds and predefined logic; they provide structure and regulatory consistency, but their central weakness is rigidity, an inability to adapt to evolving tactics, which leaves them reactive rather than anticipatory. Because thresholds are fixed and, in practice, discoverable through repeated probing, actors seeking to evade detection can learn to operate just beneath them, eroding the value of static rules over time (FATF, 2021b).

The operational cost of this rigidity is substantial and has been recognized at the supervisory level. The Financial Action Task Force has observed that conventional rule-based monitoring generates large volumes of alerts that are ultimately determined not to be suspicious, consuming compliance capacity that would otherwise be directed toward genuine risk, and has identified this inefficiency as a principal motivation for exploring more advanced analytical techniques (FATF, 2021b). The underlying problem is not merely operational but effectiveness-related: long-standing estimates suggest that only a small fraction of illicit financial flows are ultimately intercepted, with the United Nations Office on Drugs and Crime estimating that less than one percent of such flows are seized or frozen (UNODC, 2011). A monitoring apparatus that is simultaneously expensive to operate and modest in interception rate creates pressure to improve detection without simply adding reviewers. This combination of rising volume, rising cost, and persistent risk has pushed compliance functions toward artificial intelligence as a complement to, rather than a wholesale replacement for, existing controls. Supervisory posture has shifted accordingly. In the United States, the federal banking agencies and the Financial Crimes Enforcement Network issued a joint statement encouraging banks to consider innovative approaches to Bank Secrecy Act and anti-money laundering compliance, including artificial intelligence and machine learning, and clarifying that pilot programs that do not succeed would not automatically result in supervisory criticism (FinCEN et al., 2018). The Financial Action Task Force has likewise catalogued the opportunities that new technologies present for improving the effectiveness of anti-money laundering and counter-terrorist-financing measures, alongside the challenges of explainability, data quality, and governance that accompany them (FATF, 2021b).

Despite this momentum, the literature connecting artificial intelligence to compliance monitoring remains fragmented across disciplines that rarely address one another directly. The technical literature on anomaly detection develops increasingly capable models, including deep and graph-based architectures, but typically evaluates them on detection metrics alone (Chandola et al., 2009; Pang et al., 2021; Pourhabibi et al., 2020). The explainability literature develops methods for interpreting model outputs but is generally domain-agnostic rather than tied to a specific regulatory obligation (Ribeiro et al., 2016; Lundberg & Lee, 2017; Ying et al., 2019). The model risk governance literature, anchored in supervisory guidance, specifies what institutions must document and validate but does not prescribe an analytical architecture (Federal Reserve Board of Governors, 2011). Each body of work is mature within its own boundaries.

What is comparatively scarce is an architecture that treats detection performance and regulatory defensibility as a single, integrated design problem, particularly for the high-volume, low-latency conditions typical of modern transaction environments. This gap matters operationally because a compliance system that optimizes detection accuracy while treating explainability, audit trail, and human oversight as downstream add-ons is likely to underperform on regulatory examination even when its underlying models perform well statistically. It matters legally as well: supervisory guidance on adverse action notices makes clear that the complexity or opacity of an algorithm does not excuse an institution from providing specific reasons for a decision, meaning that interpretability is a legal requirement attaching to the decision rather than an optional property of the model (CFPB, 2022).

This paper responds to that gap by developing a conceptual framework, rather than an empirically tested model, that integrates the technical and governance dimensions of AI-assisted compliance monitoring for high-volume transaction environments. In doing so, it asks what architectural components are needed to operationalize such monitoring at high transaction volume and low latency, how detection performance can be reconciled architecturally, rather than procedurally, with the explainability and documentation requirements regulators expect of financial models, and what role human oversight should play within an AI-augmented system, calibrated rather than applied uniformly. Each of the framework's individual components draws on established literature; how those components are sequenced, bounded, and connected into a single architecture is developed in Section 3. For clarity, the paper uses the term high-volume transaction environment to describe institutions or platforms that process transactions on the order of hundreds of thousands to tens of millions per day, across multiple channels such as card payments, real-time and batch bank transfers, wire transfers, and, where relevant, digital-asset rails, and that require sub-second to low-single-digit-second scoring latency for at least a meaningful subset of that volume. The framework is scoped to the compliance monitoring and anomaly detection function specifically; it does not attempt to cover adjacent functions such as customer onboarding identity verification, sanctions list curation, or enterprise-wide operational risk management, except where those functions directly supply inputs to, or receive outputs from, the layers described in Section 3. Section 2 reviews the relevant literature, Section 3 presents the framework itself, Sections 4 and 5 illustrate and compare it, and Sections 6 through 8 discuss its implications, limitations, and conclusions.

2. Related Work

2.1 The Limits of Rule-Based Monitoring

Rule-based transaction monitoring remains widely used because it is transparent, auditable, and directly traceable to specific regulatory scenarios. Its structural weakness is that it depends on static rules and fixed thresholds that must be manually tuned and that struggle to keep pace with the dynamic tactics used in financial crime. Because thresholds are fixed and can often be inferred through repeated probing, actors can adjust their behavior to remain just below detection limits, while legitimate variation in customer behavior continues to trigger unnecessary alerts. The Financial Action Task Force has characterized this pattern directly, noting that traditional approaches can be both resource-intensive and limited in their ability to detect novel or evolving typologies, and identifying technology-enabled analytics as a route to improving effectiveness rather than merely reducing cost (FATF, 2021b). The resulting system is simultaneously too rigid to catch unfamiliar patterns and too coarse to avoid generating substantial volumes of alerts that do not correspond to genuine risk, a combination that motivates augmenting, rather than replacing, rule-based systems with adaptive techniques. Reviews of the banking-specific literature reach the same conclusion from the opposite direction, observing that the move from threshold-based scenarios toward model-based detection is driven less by cost than by the inability of fixed rules to represent laundering typologies that change faster than the tuning cycle (Atakpa, 2021). A further limitation is one of scope rather than of method: coordination of suspicious activity increasingly occurs through channels that sit outside the transaction record, and analytics applied to communications and other non-transactional evidence has been proposed as a complement to transaction monitoring precisely because the latter cannot observe them (Mbonu et al., 2021).

2.2 Machine Learning Approaches to Anomaly Detection

A substantial body of technical literature has moved beyond static rules toward adaptive, data-driven detection. Anomaly detection as a general problem has been surveyed extensively, with early work establishing the distinction between supervised, semi-supervised, and unsupervised formulations and cataloguing the domain-specific assumptions that determine which formulation is appropriate (Chandola et al., 2009). Because confirmed fraud and financial-crime labels are scarce, expensive to obtain, and severely imbalanced relative to legitimate activity, unsupervised and semi-supervised formulations are particularly relevant to compliance monitoring, where the analyst confirms only a small fraction of flagged activity and where confirmation may arrive months after the transaction itself (Dal Pozzolo et al., 2018).

Within this space, several model families recur. Classical shallow methods score observations by density or separability, including the local outlier factor, which compares an observation's local density to that of its neighbors (Breunig et al., 2000), and isolation forest, which scores anomalies by how few random partitions are required to isolate a point and which scales favorably to large datasets (Liu et al., 2008). Deep architectures extended this work considerably, with reconstruction-based and representation-learning approaches surveyed comprehensively by Pang et al. (2021), who note both the performance gains available from deep models and the recurring difficulty of interpreting their outputs. In the specific setting of card-transaction fraud, Dal Pozzolo et al. (2018) demonstrate that realistic evaluation requires modeling verification latency and class imbalance explicitly, since accuracy figures obtained without those constraints substantially overstate operational performance, a caution directly relevant to any claim about detection accuracy in a compliance setting.

Graph-based detection is particularly relevant to high-volume transaction environments because financial crime is frequently relational rather than purely behavioral: mule networks, layered laundering structures, and coordinated fraud rings are easier to detect through the pattern of connections between accounts than through any single account's transaction history in isolation. Pourhabibi et al. (2020) provide a systematic review of graph-based anomaly detection for fraud, distinguishing approaches that treat anomalies as unusual node attributes, unusual structural position, or a combination of the two, and noting that graph-based methods consistently surface patterns that record-level methods miss. Applied work in the financial-crime domain has demonstrated the practical viability of this approach: Weber et al. (2019) apply graph convolutional networks to a large labeled dataset of Bitcoin transactions for anti-money laundering purposes, providing one of the more widely used public benchmarks in this area and illustrating both the promise of graph learning and its sensitivity to shifts in the underlying data over time.

Review work applying these families to financial-crime settings specifically reports convergent constraints. Atakpa and Abetoh (2022) survey machine learning advances in fraud detection for banking systems and identify label scarcity, class imbalance, and performance decay as the recurring barriers to deployment rather than model capacity. Atakpa and Fobellah (2023) narrow the problem to financial time-series data, proposing a conceptual model spanning banking and healthcare payment contexts in which temporal structure rather than record-level features carries the primary detection signal. Adelanwa et al. (2023b) apply predictive analytics to financial risk detection and fraud prevention in public-sector systems, where the statistical problem is similar but the accountability requirements attaching to a false positive differ. Notably, the same detection problem and broadly the same model families recur well outside financial services: systematic reviews of machine learning for anomaly detection in industrial control systems (Adebayo et al., 2023) and of layered detection architectures combining security information and event

management, network detection and response, and anomaly detection in energy-sector networks (Adegbite et al., 2023) describe an arrangement closely analogous to the hybrid, multi-layer structure developed in Section 3. That the same layering emerges in a domain with different data, different adversaries, and different regulators suggests it is a response to a general property of high-volume anomaly detection rather than an artifact of financial compliance.

2.3 Explainability and the Regulatory Demand for Interpretability

Detection performance alone is not sufficient in a regulated environment. The Federal Reserve and the Office of the Comptroller of the Currency jointly issued Supervisory Letter SR 11-7 in 2011, establishing a three-tier framework for model governance covering development, independent validation, and ongoing monitoring, and requiring that validation evaluate a model's conceptual soundness, track its performance over time, and analyze its outcomes (Federal Reserve Board of Governors, 2011). The guidance is notable for treating ongoing monitoring as a continuous obligation rather than a one-time check: institutions are expected to maintain a systematic process for evaluating model performance over time, and material changes in that performance are expected to trigger renewed validation activity. This governance expectation has direct implications for how AI-based detection models must be built, not only how they must be documented after the fact. A substantial methodological literature now supplies post-hoc explanation techniques capable of meeting part of this requirement. Model-agnostic local explanation methods approximate a complex model's behavior in the neighborhood of a single prediction, allowing an individual decision to be explained without interpreting the entire model (Ribeiro et al., 2016). Additive attribution methods based on Shapley values provide a unified formulation with desirable consistency properties and have become widely used for feature-level attribution in financial applications (Lundberg & Lee, 2017). For graph-based models, explanation requires identifying which relationships, rather than which features, drove a prediction; GNNExplainer addresses this by identifying a compact subgraph and feature subset that account for a given prediction (Ying et al., 2019). Together these methods make it technically feasible to attach a human-readable rationale to an individual alert, though they do not by themselves guarantee that the rationale faithfully reflects the model's internal reasoning.

The regulatory stakes of interpretability are concrete rather than abstract. Guidance from the Consumer Financial Protection Bureau states that creditors using complex algorithms, including those sometimes described as black-box models, must still provide accurate and specific reasons for adverse actions, and that the complexity of the model does not excuse noncompliance (CFPB, 2022). Separately, the Bank Secrecy Act framework requires that suspicious activity reports include a narrative describing the basis for the institution's suspicion, which means that a numerical risk score is not, on its own, a sufficient regulatory artifact; some mechanism must translate the score into a defensible account of why the activity was considered suspicious (31 U.S.C. §§ 5311-5336). In the European Union, the General Data Protection Regulation constrains decisions based solely on automated processing where those decisions produce legal or similarly significant effects, reinforcing the same requirement from a data-protection direction (European Parliament & Council, 2016). Taken together, these instruments position explainability as a compliance obligation attaching to the decision rather than an optional technical enhancement.

2.4 The Role of Human Oversight

A parallel literature treats human review not as a fallback but as a structural governance mechanism. The European Union's Artificial Intelligence Act, adopted in 2024, requires that high-

risk AI systems be designed so that they can be effectively overseen by natural persons during use, including the ability to understand the system's capacities and limitations, remain aware of automation bias, interpret system output, decide not to use the system in a given case, and intervene or halt operation (European Parliament & Council, 2024, Art. 14). This is a design obligation directed at the system rather than a procedural obligation directed at staff, which distinguishes it from earlier compliance-training approaches to human oversight.

Rather than applying uniform human review to every output, the calibration question, how much oversight to apply and to which cases, is the operationally decisive one. A monitoring system that routes every alert to full manual review recreates the capacity bottleneck that motivated automation in the first place, while one that routes nothing to human review forfeits the oversight that both the AI Act and data-protection law contemplate. The literature on verification latency in fraud detection is informative here, since it establishes that analyst confirmation is a scarce, delayed, and non-random resource, and that treating unreviewed cases as confirmed negatives introduces systematic bias into subsequent model training (Dal Pozzolo et al., 2018). This implies that the design of the human review layer is not only a workload question but a data-quality question, because the pattern of what analysts review determines the labels the detection models subsequently learn from. Work on control automation makes a structurally similar point from the audit side. Mbonu et al. (2022) develop a conceptual framework for AI-enabled information technology general controls and Sarbanes-Oxley audit automation in which automation does not remove the reviewer but changes what the reviewer is accountable for, shifting attention from inspecting individual sampled items toward evaluating whether the automated control operated as designed across the population. The compliance monitoring analogue is that once alert triage is partly automated, scarce review capacity is best directed toward the cases and the control behavior that automated scoring cannot itself attest to.

2.5 Sector Extensions: Digital Assets, Cross-Border Payments, and Insurance

The compliance monitoring problem extends beyond traditional banking rails into digital assets, where the underlying data structure differs but the governance logic is closely related. The Financial Action Task Force extended its standards to virtual assets and virtual asset service providers, requiring such providers to be licensed or registered, to conduct customer due diligence, and to obtain, hold, and transmit originator and beneficiary information for transfers, a requirement commonly referred to as the travel rule (FATF, 2021a). Implementation across jurisdictions has been uneven, producing what practitioners describe as a sunrise problem, in which a provider subject to the requirement in one jurisdiction may be unable to obtain corresponding information from a counterparty operating where the requirement has not yet taken effect (FATF, 2021a).

Because public-ledger transactions are pseudonymous rather than anonymous, and because the transaction graph is directly observable, this sector is unusually well suited to graph-based analytics, and much of the methodological work on graph learning for financial crime has been developed and benchmarked in exactly this setting (Weber et al., 2019). For a compliance monitoring framework, this extension has two direct implications: the ingestion layer must be able to consume on-chain transaction data alongside conventional banking data, and the governance layer must treat jurisdiction-specific thresholds and counterparty readiness as configurable, periodically reviewed parameters rather than fixed constants. The cross-border dimension raises a parallel problem on conventional rails. Adesuyi et al. (2024) develop an AI-driven risk scoring model for global cross-border trade payment transactions, in which a defensible score must combine counterparty, corridor, and instrument-level signals that no single institution in the

payment chain observes in full, reinforcing the case made in Section 2.6 for treating cross-institutional learning as an architectural option rather than a procedural one.

A related but distinct literature addresses fraud detection in insurance claims, which shares underlying machine learning techniques with transaction monitoring but differs in operational tempo. Ensemble tree-based methods are consistently strong performers for structured claims data, with random forests providing variance reduction through bootstrap aggregation (Breiman, 2001) and gradient-boosted tree implementations offering both accuracy and computational efficiency on tabular data of the kind typical in claims processing (Chen & Guestrin, 2016). In a peer-reviewed comparison, Debener et al. (2023) evaluate supervised and unsupervised machine learning approaches to insurance fraud detection directly against one another, providing one of the more methodologically careful benchmarks in this literature and highlighting the practical trade-offs between the two formulations when labels are incomplete. The key operational difference from real-time payments monitoring is timing: claims fraud is typically evaluated after a claim has been submitted, in a case-based review process, rather than scored in-line as a transaction is authorized. A framework built primarily around real-time transaction scoring therefore needs an explicit case-based mode to accommodate insurance and other post-event review contexts.

The same post-event review pattern recurs in adjacent domains and reinforces the need for that second mode. Atakpa et al. (2024) review artificial intelligence applied to healthcare payment fraud detection in National Health Service settings, where claims are adjudicated after submission and detection therefore operates on a review cycle rather than an authorization cycle. Eboh and Aliliele (2024) address a longer-horizon case still, proposing a data analytics framework for assessing and detecting venture capital and private equity investment fraud in United States capital markets, where the relevant evidence accumulates over months and the unit of review is an entity or a fund rather than a transaction. Across insurance, healthcare payments, and investment fraud, the detection techniques are largely shared while the latency budget and the unit of review are not, which is the distinction the case-based mode described in Section 3.7 is intended to capture.

2.6 Federated and Privacy-Preserving Learning Across Institutions

A persistent constraint on detection performance is that financial crime, particularly money laundering, is often distributed across institutions in ways that no single institution's data can fully reveal, while data protection law and competitive concerns limit direct data sharing. Federated learning offers a technical response: a shared model is trained collaboratively by exchanging model updates rather than raw data, with each participant's data remaining local (McMahan et al., 2017). Subsequent work has systematized the field, cataloguing the privacy, security, communication-efficiency, and statistical-heterogeneity challenges that arise when participants hold non-identically distributed data, as institutions with different customer bases necessarily do, and surveying the secure aggregation and differential privacy techniques used to provide formal guarantees against reconstruction of individual records from shared updates (Kairouz et al., 2021).

The relevance to compliance monitoring is direct. The Financial Action Task Force has explicitly identified privacy-preserving analytics, including techniques that permit collaborative analysis without pooling underlying customer data, as a promising avenue for improving the detection of cross-institutional laundering typologies while respecting data protection constraints (FATF, 2021b). For the framework developed in Section 3, federated learning represents an architectural option for the feature and detection layers that is specifically relevant to consortia such as correspondent banking networks or payment schemes, where an entity graph spanning multiple institutions would be analytically valuable but cannot lawfully or commercially be centralized.

2.7 Concept Drift and Model Performance Over Time

A model that performs well at deployment does not necessarily remain accurate over time. The general phenomenon, in which the statistical relationship between inputs and the target changes after training, is well characterized in the machine learning literature, along with the detection methods and adaptation strategies available to address it (Gama et al., 2014). In fraud and financial-crime detection specifically, drift is not merely incidental but adversarial: the population being modeled adapts in direct response to the detection mechanism, so that a model's very effectiveness alters the distribution it was trained on. Dal Pozzolo et al. (2018) make the related point that detection systems operate under delayed and partial feedback, which complicates both the measurement of drift and the retraining response to it, since recent performance can only be assessed on the subset of cases that have been reviewed.

This literature substantiates, and helps to operationalize, the continuous feedback loop that the framework in Section 3 assigns to its governance layer. It also aligns closely with the ongoing monitoring obligation in supervisory model risk guidance, which anticipates that model performance will be tracked over time and that material degradation will trigger revalidation rather than silent continuation (Federal Reserve Board of Governors, 2011). The methodological literature on drift and the supervisory literature on ongoing monitoring are, in effect, describing the same requirement from different directions, which is one reason the framework treats monitoring as an architectural layer rather than a periodic activity.

Taken together, the literature is strong on individual components. Adaptive detection models, explainability techniques, model risk governance, human oversight requirements, privacy-preserving collaborative learning, and drift adaptation are each reasonably well developed on their own terms, and several are supported by mature methodological work spanning two decades. What is comparatively underdeveloped, and what no single source reviewed here supplies, is an integrated architecture that connects these components end to end, from raw transaction ingestion through detection, explanation, human review, and governance, in a way that is explicitly designed for high-volume, low-latency environments and that treats regulatory defensibility as a property of the architecture rather than a compliance activity layered on top of it. The technical sources reviewed above largely stop at the detection boundary; the governance sources largely begin after the model exists. The framework developed in Section 3 is an attempt to span that boundary.

3. Framework Architecture

The framework below was developed through a conceptual synthesis of the literature reviewed in Section 2, consistent with design-science research in the information systems tradition, in which an architecture is proposed to resolve a documented tension in existing practice and justified through grounding in prior literature rather than empirical testing (Hevner et al., 2004). Recurring technical components, rule engines, unsupervised statistical models, and graph-based learning, and recurring governance components, the three-tier structure of supervisory model risk guidance, post-hoc explainability techniques, and calibrated human oversight, were mapped onto a single pipeline ordered by the sequence in which data moves through a monitoring system, then checked against the sector-specific literature in Sections 2.5 through 2.7 to confirm that digital-asset, insurance, and federated-learning use cases could be accommodated as configurations of existing layers rather than requiring additional ones. Because the framework has not been tested empirically, the design principles and layer descriptions below are traceable to cited literature describing each component's use individually, rather than to evidence that this specific combination outperforms an alternative one, a distinction returned to in Section 7.

3.1 Design Principles

Five design principles guide the framework, each grounded in the literature reviewed above.

- Detection and defensibility as joint objectives. Explainability and audit trail are treated as structural components of the architecture rather than downstream controls added after a model is built, reflecting the position in supervisory guidance that the opacity of a model does not reduce the obligation to give specific reasons for a decision (CFPB, 2022).
- Layered separation of concerns. Dividing the system into distinct layers, ingestion, features, detection, explanation, human review, and governance, allows each layer to be independently validated, consistent with the expectation that model components be assessed for conceptual soundness on their own terms (Federal Reserve Board of Governors, 2011).
- Risk-tiered human oversight. Human review effort is calibrated to case risk and model confidence rather than applied uniformly, satisfying the requirement that oversight be effective in practice rather than nominal (European Parliament & Council, 2024, Art. 14) while recognizing that analyst review is a scarce and delayed resource whose allocation shapes subsequent training data (Dal Pozzolo et al., 2018).
- Continuous feedback. Confirmed analyst dispositions and downstream outcomes are captured as structured data that feeds back into the detection engine, supporting the drift-aware adaptation that the methodological literature describes (Gama et al., 2014) and the ongoing performance monitoring that supervisory guidance requires (Federal Reserve Board of Governors, 2011).
- Cross-cutting engineering constraints. Data privacy, latency, scalability, and fairness are treated as constraints that shape every layer of the architecture, not properties added at the end, given the real-time and high-volume nature of the target environment.

3.2 Framework Overview

Figure 1 presents the resulting architecture as six layers connected by a forward data flow and a continuous feedback loop that returns confirmed outcomes to the detection engine.

Figure 1. Six-layer conceptual framework for AI-assisted compliance monitoring and anomaly detection in high-volume transaction environments.

Read from top to bottom, the framework moves from raw data acquisition through progressively more refined stages of detection, explanation, and human judgment before terminating in a governance layer that both oversees the system and feeds performance information back into it. The dashed feedback path in Figure 1 is deliberate: it represents the mechanism by which the framework is intended to adapt over time without requiring a full architectural redesign, consistent with the ongoing-monitoring obligation described in supervisory guidance (Federal Reserve Board of Governors, 2011).

3.3 Layer 1: Data Ingestion and Integration

The first layer draws together the data sources a compliance monitoring system depends on: core banking and payments data, know-your-customer and customer due-diligence records, sanctions and watchlist feeds, adverse media, and other external risk signals. Given the throughput required by high-volume transaction environments, this layer is designed around streaming ingestion rather than the batch, end-of-day processing common in legacy monitoring systems, so that anomalies can be surfaced close to the time a transaction occurs rather than the following day. This layer also supports continuous rather than periodic customer risk assessment, in which risk profiles are updated as sanctions listings, beneficial-ownership records, and adverse-media coverage change, rather than only at scheduled review intervals. Where an institution operates across the digital-asset and cross-border payments extension described in Section 2.5, this layer additionally ingests on-chain transaction data and tracks jurisdiction-specific travel rule thresholds and counterparty readiness as configurable parameters rather than fixed constants (FATF, 2021a).

3.4 Layer 2: Feature Engineering and Entity Resolution

The second layer transforms raw transaction and customer data into the behavioral, temporal, and relational features the detection layer requires. Behavioral and temporal features capture deviation from an account's own historical pattern, such as unusual velocity, amount, or timing. Relational features require entity resolution, the process of linking accounts, devices, and counterparties across channels into a coherent transaction graph in which accounts and users are represented as nodes and transactions as edges (Pourhabibi et al., 2020; Weber et al., 2019). This graph representation is what allows the detection layer to identify anomalies that are attributive, structural, or a combination of the two, patterns that would not be visible from any single account's transaction history alone (Pourhabibi et al., 2020). For institutions participating in a consortium or correspondent network, this layer can optionally be implemented in a federated configuration, in which representations are learned jointly across institutions through model-update exchange rather than raw data pooling, subject to the statistical-heterogeneity and privacy considerations catalogued in the federated learning literature (McMahan et al., 2017; Kairouz et al., 2021).

3.5 Layer 3: Hybrid Detection Engine

The third layer is the analytical core of the framework and is deliberately hybrid rather than a wholesale replacement of rule-based logic. It combines three complementary components: retained rule-based scenarios for typologies that are well understood and, in some cases, regulatorily expected, such as structuring or specific sanctions-screening patterns; unsupervised and semi-supervised machine learning models, including local outlier factor (Breunig et al., 2000), isolation forest (Liu et al., 2008), and reconstruction-based deep methods (Pang et al., 2021), for detecting statistical outliers without relying on scarce labeled data; and graph-based models operating on the entity graph produced in Layer 2 for detecting relational anomalies such as mule networks or coordinated fraud rings (Weber et al., 2019; Pourhabibi et al., 2020). The outputs of these components are combined into a single ensemble risk score rather than treated as independent alerts, so that corroboration across methods is expressed in the score itself rather than left to the analyst to reconstruct. Two design cautions from the literature shape this layer. First, evaluation must account for verification latency and class imbalance, since performance estimates obtained without modeling those constraints overstate what the system will achieve operationally (Dal Pozzolo et al., 2018). Second, the layer supports two scoring modes: an in-line, real-time mode for payments and transaction monitoring, and a case-based mode for post-event review contexts such as insurance claims, where ensemble tree-based classifiers are well established on structured data (Breiman, 2001; Chen & Guestrin, 2016; Debener et al., 2023).

3.6 Layer 4: Explainability and Risk Scoring

The fourth layer attaches a post-hoc explanation to every transaction the detection engine flags, rather than producing a bare numerical score. For non-graph model components, this takes the form of attribution-based explanation identifying which input features contributed most to the score, using local model-agnostic approximation (Ribeiro et al., 2016) or additive Shapley-value attribution (Lundberg & Lee, 2017). For graph-based detections, the layer produces subgraph-level explanations identifying which relationships between accounts drove the flag, rather than attributing the decision to a single account's features in isolation (Ying et al., 2019). Explanations generated at this layer are intended for two audiences at once: a technical rationale usable by model validators conducting the conceptual-soundness review supervisory guidance requires (Federal Reserve Board of Governors, 2011), and a plain-language account usable by compliance analysts

drafting the narrative portion of a suspicious activity report (31 U.S.C. §§ 5311-5336). Positioning explanation as a layer rather than a reporting step reflects the regulatory position that model complexity does not diminish the obligation to state specific reasons for a decision (CFPB, 2022).

3.7 Layer 5: Human-in-the-Loop Review and Case Management

The fifth layer routes flagged transactions to human analysts according to a risk-tiered triage logic rather than a single undifferentiated queue. High-confidence, low-risk flags may be auto-dispositioned or routed to periodic sampled review, while high-risk or low-confidence cases are routed for full analyst review. This calibration is what makes oversight effective in practice rather than nominal, which is the standard the AI Act's human oversight provision sets for high-risk systems (European Parliament & Council, 2024, Art. 14), and it responds to the constraint that data protection law imposes on decisions taken solely by automated means (European Parliament & Council, 2016, Art. 22). Analyst decisions, confirm, dismiss, escalate, along with structured rationale codes, are captured as data rather than recorded only as free text, so that they can serve both audit purposes and as a labeled training signal. The literature on delayed and partial feedback is directly relevant to how that signal should be interpreted, since sampling only high-scoring cases produces a biased view of model performance and, if fed back uncritically, can entrench the model's existing blind spots (Dal Pozzolo et al., 2018). For post-event, case-based workloads such as insurance claims, this layer integrates with the relevant claims-adjudication workflow rather than a payment-blocking workflow, reflecting the timing distinction described earlier in this section.

3.8 Layer 6: Governance, Audit, and Model Risk Oversight

The sixth layer performs the ongoing governance function that supervisory guidance requires of any model used in a regulated financial institution: independent validation separate from the team that develops the models, continuous performance monitoring, and outcomes analysis comparing flagged cases against confirmed dispositions over time (Federal Reserve Board of Governors, 2011). This layer also monitors for concept drift and score-distribution stability, applying the detection-and-adaptation methods established in the machine learning literature at a cadence appropriate to the adversarial tempo of financial crime, where the population being modeled changes in response to the detection mechanism itself (Gama et al., 2014). It monitors for fairness and disparate outcomes as well, since a model can produce differential effects across protected groups without using protected attributes as explicit inputs, and since the obligation to give specific reasons for adverse decisions applies regardless of model complexity (CFPB, 2022). A tamper-evident audit trail records who, or which system component, took each action and when, for both automated decisions and human overrides, supporting the demonstrable oversight that the AI Act contemplates for high-risk systems (European Parliament & Council, 2024).

This layer closes the framework's feedback loop. Confirmed analyst dispositions become labeled signal that recalibrates detection thresholds and model weights on a defined, governed cadence, itself subject to the same change-management and revalidation triggers that apply to any other material change in model performance (Federal Reserve Board of Governors, 2011). Treating adaptation as a governed process rather than an ad hoc one matters because supervisory model risk guidance was written with a relatively static model lifecycle in mind, in which development, validation, and use are cleanly separated in time; a system that continues to learn from analyst feedback after deployment blurs that separation unless retraining is brought explicitly inside the governed lifecycle.

3.9 Cross-Cutting Concerns

Four concerns cut across all six layers rather than belonging to any single one. Data privacy and security must be maintained as data moves across ingestion, feature engineering, and case management, particularly where customer data crosses jurisdictional boundaries, and where federated configurations shift but do not eliminate the privacy surface (Kairouz et al., 2021). Scalability and latency must be engineered in from the ingestion layer onward, since the framework's value depends on operating at the throughput and speed of a high-volume, real-time environment. Fairness and disparate-impact controls must be monitored continuously rather than assessed only at model launch (CFPB, 2022). And regulatory alignment must be maintained across multiple, sometimes only partially overlapping, regimes at once, including Bank Secrecy Act obligations, data-protection law, AI-specific regulation, and supervisory model risk guidance.

3.10 Layer Summary and Technology Mapping

Table 1 summarizes the six layers, their core function, the representative techniques associated with each in the literature reviewed in Section 2, and the primary regulatory or governance obligation each layer is designed to satisfy.

Layer	Core Function	Representative Techniques	Primary Regulatory Hook
1. Ingestion	Acquire and stream transaction, KYC, sanctions, and risk data in real time	Streaming ingestion, continuous risk profiling, on-chain data feeds	BSA recordkeeping; FATF travel rule
2. Feature Engineering	Build behavioral, temporal, and relational features; construct entity graph	Entity resolution, transaction graphs, federated representation learning	Data protection law (cross-border data handling)
3. Detection Engine	Score transactions/cases using combined rule, statistical, and graph methods	Rules, LOF, isolation forest, deep and graph models, ensemble scoring	SR 11-7 conceptual soundness (development)
4. Explainability	Attach technical and narrative explanation to every flagged case	LIME, SHAP-style attribution, GNNExplainer subgraph explanation	SAR narrative requirement; CFPB Circular 2022-03
5. Human-in-the-Loop	Route, review, and disposition flagged cases by risk tier	Risk-tiered triage, structured disposition capture	EU AI Act Art. 14; GDPR Art. 22
6. Governance	Validate, monitor drift and fairness, audit, and govern the feedback loop	Drift detection, independent validation, audit logging	SR 11-7 ongoing monitoring; fair-lending law

Table 1. Summary of the six-layer framework, representative techniques, and primary regulatory hooks.

4. Illustrative Application: A High-Volume Payments Scenario

This section illustrates how the framework would operate using a hypothetical scenario. The institution, its transaction volumes, and the specific cases described are constructed for expository purposes only and are not drawn from any real institution's data; the scenario is intended to make the framework's layer interactions concrete, not to validate its performance.

4.1 Scenario Setup

Consider a hypothetical mid-size cross-border payments processor, referred to here as Northbridge Payments, that processes approximately two million transactions per day across card payments, real-time domestic transfers, and international wire and remittance rails, serving both retail customers and small and medium-sized business clients across several jurisdictions. Northbridge's compliance function currently relies primarily on rule-based scenarios and is evaluating the six-layer framework described in Section 3 as a modernization path, consistent with the incremental adoption pattern discussed in Section 6.1.

4.2 Scenario One: A Structuring-Like Pattern

A cluster of related retail accounts, linked by a shared device fingerprint and a common beneficiary account, each send transfers just below Northbridge's existing rule-based reporting threshold over a short window. Under the framework, Layer 1 ingests each transfer in near real time. Layer 2 resolves the sending accounts and the shared device and beneficiary into a single connected component of the transaction graph, a relational pattern invisible to any single account's history in isolation. Layer 3's rule component registers a near-threshold amount flag on several of the individual transfers, while its graph-based component separately flags the connected component as structurally anomalous given its unusually dense, star-shaped connectivity around a single beneficiary; the ensemble score combines both signals into a single elevated risk score for the beneficiary account. Layer 4 generates two linked explanations: an attribution-level explanation noting the near-threshold, high-frequency transfer pattern, and a subgraph-level explanation identifying the shared device and beneficiary as the specific relationship driving the graph-based flag (Lundberg & Lee, 2017; Ying et al., 2019). Layer 5 routes the case to full analyst review, because the combination of a moderate rule-based signal and a high-confidence graph-based signal places it in the framework's highest review tier under the risk-tiered triage logic described in Section 3.7. An analyst confirms the pattern as suspicious, and Layer 6 both supports the resulting suspicious activity report, using the narrative supplied by Layer 4 as a starting point, and logs the confirmed disposition as labeled feedback for the detection engine's next governed recalibration cycle.

4.3 Scenario Two: A False-Positive Suppression Case

A legitimate small e-commerce merchant client experiences a sharp, temporary increase in transaction volume and average transaction size during a seasonal sales period, a pattern that deviates markedly from its own historical baseline and would trigger a pure behavioral-deviation rule on its own. Layer 2's entity graph, however, shows no new or unusual counterparties, no shared devices with other flagged accounts, and no structural change in the merchant's surrounding network. Layer 3's rule component registers the volume deviation, but its graph-based component

does not, and the ensemble score remains below the threshold that would trigger full analyst review. Layer 4's explanation records that the flag was driven by a single behavioral rule with no corroborating relational signal, and Layer 5 routes the case to lightweight, sampled review rather than full investigation, consistent with the risk-tiered design principle in Section 3.1. This scenario illustrates the mechanism by which the framework is intended to reduce unproductive alert volume: a single deviation signal, on its own, is treated as weaker evidence than a deviation signal corroborated by a relational anomaly. Read together, the two scenarios illustrate a single underlying mechanism rather than two unrelated examples: in both cases, the decisive input to Layer 5's routing decision is not any one signal in isolation but whether a behavioral or rule-based deviation is corroborated by a relational, graph-based signal. In Scenario One, corroboration is present and the case escalates; in Scenario Two, corroboration is absent and the case is deprioritized. This is the specific architectural mechanism through which the framework pursues the joint objective set out in Section 3.1, treating detection accuracy and regulatory defensibility together, since the same corroboration signal that suppresses the unproductive alert in Scenario Two also produces the richer, subgraph-level explanation that supports the report narrative in Scenario One. It also illustrates why the sampled-review path in Scenario Two matters for reasons beyond workload: without some review of low-scoring cases, the institution would have no basis for detecting a systematic failure to escalate, a bias risk the delayed-feedback literature identifies directly (Dal Pozzolo et al., 2018).

Table 2 summarizes how each scenario moves through the six layers.

Layer	Scenario 1: Structuring-Like Pattern	Scenario 2: False-Positive Suppression
1. Ingestion	Near-real-time capture of related sub-threshold transfers	Real-time capture of merchant's seasonal volume spike
2. Feature/Graph	Shared device and beneficiary link accounts in a dense cluster	No new counterparties; network structure unchanged
3. Detection	Rule near-threshold flag and graph structural flag combine	Rule behavioral flag only; no corroborating graph signal
4. Explainability	Attribution plus subgraph explanation of shared beneficiary	Explanation notes single, uncorroborated rule flag
5. Human Review	Routed to full analyst review (high risk tier)	Routed to lightweight, sampled review (low risk tier)
6. Governance	Disposition confirmed; feeds next recalibration cycle	Dismissal logged as labeled negative feedback

Table 2. Illustrative walkthrough of two hypothetical cases through the six-layer framework.

5. Comparative Analysis with Existing Approaches

5.1 Pure Rule-Based Systems

Systems built solely on rule-based scenarios remain attractive for their transparency and direct traceability to specific regulatory typologies, and they require no model validation infrastructure beyond documenting the rules themselves. Their central limitation, discussed in Section 2.1, is

rigidity: fixed thresholds are comparatively easy for sophisticated actors to learn and evade, and legitimate variation in customer behavior continues to generate substantial volumes of unproductive alerts, a pattern supervisory bodies have identified as a barrier to effectiveness rather than merely a cost problem (FATF, 2021b).

5.2 Single-Model, Black-Box Machine Learning Systems

Systems built around a single, high-performing machine learning model, without an accompanying explainability or human-oversight architecture, can achieve strong detection performance, particularly where deep architectures are applied to rich behavioral data (Pang et al., 2021). Their limitation is not detection accuracy but defensibility. Without structural explainability, such systems struggle to satisfy the conceptual-soundness expectations of model risk guidance (Federal Reserve Board of Governors, 2011) and cannot readily produce the specific reasons that adverse-action guidance requires irrespective of model complexity (CFPB, 2022). Without demonstrable human oversight, they are poorly positioned relative to the AI Act's requirement that high-risk systems be designed for effective oversight by natural persons (European Parliament & Council, 2024, Art. 14).

5.3 Existing Point-Solution Approaches

Several strands of work reviewed in Section 2 make strong contributions on a single dimension without integrating across the full pipeline. Graph-based detection research demonstrates strong technical performance on relational fraud patterns but is typically developed and evaluated as a detection component rather than as an end-to-end compliance architecture (Weber et al., 2019; Pourhabibi et al., 2020). Explainability methods are formulated to be domain-agnostic, which is a methodological strength but leaves the mapping onto specific regulatory artifacts, such as a report narrative or an adverse-action reason, to the implementing institution (Ribeiro et al., 2016; Lundberg & Lee, 2017; Ying et al., 2019). Federated learning addresses the cross-institutional data-silo problem but is generally silent on downstream explanation and case management (McMahan et al., 2017; Kairouz et al., 2021). Supervisory guidance specifies governance obligations without prescribing an analytical architecture (Federal Reserve Board of Governors, 2011). Each is mature within its own scope; none, individually, addresses the integration gap identified in Section 2.

Table 3 compares the six-layer framework against these alternative approaches across five dimensions relevant to a high-volume transaction environment: adaptivity of detection to novel patterns, structural explainability, design of human oversight, alignment with existing regulatory expectations, and suitability for high-volume, low-latency operation.

Approach	Detection Adaptivity	Explainability	Human Oversight Design	Regulatory Defensibility	High-Volume Suitability
Pure rule-based	Low	High (by construction)	Ad hoc	High, but static	High
Black-box single ML model	High	Low	Ad hoc	Low	High
Graph-only detection methods	High (relational)	Moderate	Not addressed	Moderate	Moderate to high

Approach	Detection Adaptivity	Explainability	Human Oversight Design	Regulatory Defensibility	High-Volume Suitability
Governance guidance alone	Not addressed	Not addressed	High	Moderate	Not addressed
Proposed six-layer framework	High	High (structural)	High (risk-tiered)	High (by design)	High (by design)

Table 3. Comparison of the framework against alternative approaches across five design dimensions.

The comparison in Table 3 is qualitative and derived from the characteristics reported in the literature reviewed in Section 2, rather than from a controlled empirical evaluation; it is offered to position the framework relative to alternative designs, not to demonstrate its superiority in measured performance terms, a distinction elaborated in Section 7.

Reading across the rows of Table 3, the pattern that emerges is not that any single alternative performs poorly in an absolute sense, but that each excels along a narrower set of dimensions than the framework attempts to cover simultaneously. Pure rule-based systems remain a reasonable choice where transparency and predictable regulatory treatment matter more than adaptivity, for instance in a narrowly scoped sanctions-screening function. Black-box machine learning systems remain attractive where detection performance dominates and the model's outputs do not directly drive an adverse action against a customer, reducing the practical urgency of explanation. The framework is best understood as a design intended for the specific, and increasingly common, situation described in Section 1: high transaction volume, low tolerance for latency, and outputs that can plausibly lead to an adverse action such as a blocked payment or a filed report, a combination under which adaptivity, explainability, and human oversight are simultaneously necessary rather than substitutable for one another.

6. Discussion

The framework's principal contribution is integrative rather than novel at the level of any individual component. Detection models, explainability techniques, model risk governance, and human oversight design are each established separately in the literature; what is offered here is their organization into a single architecture in which each layer's requirements are shaped by the layers around it. The fragmentation this responds to is visible in the structure of the literature itself, reviewed in Section 2: the technical sources largely conclude at the detection boundary, reporting performance metrics without addressing what an institution must be able to say about a flagged case, while the governance sources largely begin after a model exists, specifying validation and documentation obligations without prescribing how detection should be organized. By treating explainability and human oversight as architectural layers rather than compliance activities appended after model development, the framework offers one way of connecting those two literatures operationally.

6.1 Practical Implications

For compliance functions facing sustained alert volumes and constrained review capacity, the layered structure is intended to support incremental rather than wholesale modernization. Institutions can introduce machine learning or graph-based detection components in Layer 3

alongside existing rule engines without discarding rules that remain regulatorily useful; they can improve explainability and case-management tooling in Layers 4 and 5 somewhat independently of changes to the underlying detection models; and they can mature their human-in-the-loop calibration over time as accumulated feedback data in Layer 6 becomes available. This incrementalism matters practically because wholesale replacement of monitoring infrastructure in a regulated environment carries validation and transition risk that few institutions can absorb quickly. It is also consistent with the supervisory posture expressed in the 2018 joint statement on innovation, which encouraged institutions to pilot new approaches alongside existing systems rather than replacing them outright, and indicated that unsuccessful pilots would not in themselves attract supervisory criticism (FinCEN et al., 2018).

6.2 Regulatory Alignment

The framework's layers map reasonably closely onto the three-tier structure of development, independent validation, and ongoing monitoring set out in supervisory model risk guidance (Federal Reserve Board of Governors, 2011): Layers 1 through 3 correspond broadly to development, Layer 6 to independent validation and ongoing monitoring, and Layers 4 and 5 provide the explanation and human-judgment inputs on which both validation and monitoring depend. At the same time, that guidance was drafted with a relatively static model lifecycle in mind, in which development, validation, and use are separable in time and organizational responsibility. A system that continues to learn from analyst dispositions after deployment complicates that separation. The continuous feedback loop in Layer 6 is one attempt to reconcile the two: rather than treating adaptation as something occurring outside the governed lifecycle, the framework brings retraining inside it, subject to the same change-management and revalidation triggers as any other material model change. The European Union's Artificial Intelligence Act adds a further layer of alignment to consider. The Act designates certain AI systems used in financial services as high-risk, including those used to evaluate the creditworthiness of natural persons, while providing that systems used for the purpose of detecting financial fraud are not classified as high-risk on that basis alone (European Parliament & Council, 2024, Annex III). For the framework proposed here, this means that Layer 3's detection engine, used purely for financial-crime signal generation, may sit outside the high-risk classification that attaches to credit scoring, while related uses, such as risk profiling that materially influences a customer's access to financial services, may still trigger high-risk obligations depending on how the output is used downstream. Given this ambiguity at the margin, building Layer 5's human oversight to a consistently high standard, rather than one that varies with formal classification, is a prudent default: the actions that can follow from a Layer 3 flag, including account restriction or a filed report, can carry significant consequences for a customer regardless of how the underlying model is classified.

Table 4 collects the key regulatory or legal reference most directly associated with each layer, alongside a brief compliance note. It is intended as a starting point for scoping an implementation, not as a substitute for jurisdiction-specific legal advice; several of these instruments, particularly the travel rule and the AI Act's classification boundaries, are subject to staged implementation and ongoing interpretation, and should be reverified against current guidance before being relied on operationally.

Layer	Key Regulatory or Legal Reference	Compliance Note
1. Ingestion	BSA recordkeeping (31 U.S.C. §§ 5311-5336); FATF travel rule (FATF, 2021a)	Retain source data and, for digital-asset rails, originator and beneficiary information
2. Feature Engineering	GDPR principles governing cross-border data linkage (European Parliament & Council, 2016)	Entity resolution must respect data-residency and minimization constraints
3. Detection Engine	SR 11-7 model development standards (Federal Reserve Board of Governors, 2011)	Document conceptual soundness of each component, including the ensemble logic
4. Explainability	SAR narrative requirement; CFPB Circular 2022-03 (CFPB, 2022)	Explanations must serve both a technical validator and an analyst-facing narrative
5. Human-in-the-Loop	EU AI Act Art. 14; GDPR Art. 22 (European Parliament & Council, 2024, 2016)	Maintain meaningful human review capable of overriding or halting the system
6. Governance	SR 11-7 ongoing monitoring; fair-lending obligations	Track drift, fairness, and classification status as regulation evolves

Table 4. Regulatory mapping reference by layer.

6.3 Sector-Specific Implications

The framework's core design, described in Section 3, targets conventional banking and payments monitoring most directly. The sector-specific extensions discussed in Sections 2.5 through 2.7 imply adjustments in practice rather than a different architecture. For digital-asset and cross-border payment operations, Layer 1 must be extended to ingest on-chain transaction data, and Layer 6 must maintain jurisdiction-specific travel rule thresholds and counterparty readiness as governed, updatable parameters, given the staged and uneven implementation of those requirements across jurisdictions (FATF, 2021a). For insurance and other post-event review contexts, Layer 3 requires its case-based scoring mode rather than its real-time mode, and Layer 5's case management must integrate with claims-adjudication workflows rather than payment-blocking workflows (Debener et al., 2023). For institutions operating within a consortium or correspondent banking network, Layers 2 and 3 can be implemented in a federated configuration, allowing shared detection capability without raw data pooling, subject to the heterogeneity and privacy considerations that literature identifies (Kairouz et al., 2021).

6.4 Organizational and Change-Management Considerations

Adopting the framework raises organizational considerations beyond the technical architecture itself. Compliance analysts working within Layer 5 need to be equipped, and in many cases retrained, to interpret model-generated explanations rather than working from raw transaction data alone, a shift in skill set from case investigation toward structured judgment and feedback provision. That shift carries a specific risk the AI Act names directly, automation bias, in which a reviewer defers to a system's output rather than independently evaluating it; the Act's oversight

provision requires that operators remain aware of this tendency, which implies training and interface design choices rather than policy statements alone (European Parliament & Council, 2024, Art. 14). Layer 6's governance function requires a clearly defined committee or oversight structure with authority to approve model changes, consistent with the independent-validation principle in supervisory guidance (Federal Reserve Board of Governors, 2011); institutions without an existing model risk management function serving other lines of business may need to build this capability from a low base.

Table 5 summarizes, for each layer, the function or team typically responsible for it and the primary risk that would materialize if the layer were absent or poorly implemented, offered as a planning reference for institutions considering implementation.

Layer	Primary Owner/Stakeholder	Key Risk if Layer Fails
1. Ingestion	Data engineering / IT operations	Missed or delayed transactions; incomplete regulatory data trail
2. Feature Engineering	Data science / analytics engineering	Weak or missing relational signal; degraded downstream detection
3. Detection Engine	Model development team	Missed typologies; excessive unproductive alerts
4. Explainability	Model validation / compliance analytics	Undocumented decisions; report narratives lack basis; validation gaps
5. Human-in-the-Loop	Compliance operations / case management	Analyst overload or under-review of high-risk cases
6. Governance	Model risk management / internal audit	Undetected drift or disparate impact; loss of audit trail

Table 5. Layer ownership and failure-risk summary.

6.5 Cost and Resourcing Considerations

Although this paper does not attempt a quantified cost-benefit analysis, the structure of the framework has implications for how such an analysis should be organized. The layers are not equally expensive to build or to operate, and the ordering of investment affects when returns materialize. Layers 1 through 3, ingestion, feature engineering, and detection, typically represent the largest upfront capital investment, since they require streaming infrastructure and model development effort. Layers 4 through 6, explainability, human review, and governance, are comparatively less capital-intensive to build but more labor-intensive to operate on a continuing basis, since they depend on sustained analyst and validation staffing rather than one-time engineering work. A practical implication is that the return profile shifts over time. Early phases dominated by Layer 1 through 3 investment are likely to show returns primarily through improved detection of patterns that record-level rules miss. Later phases, once Layers 4 through 6 are mature, are more likely to show returns through reduced unproductive alert volume and improved analyst throughput, because those gains depend on explanation quality and feedback-loop maturity that only develop once the governance layer has been operating long enough to accumulate reliable disposition data. Institutions evaluating the framework on a short horizon should therefore expect

its benefits to be back-loaded relative to its costs, which is a material consideration for how a pilot is scoped and how long it must run before its results are informative.

6.6 Vendor and Build-versus-Buy Considerations

Institutions adopting the framework face a further practical choice cutting across all six layers: whether to build each component internally, license it from a specialist vendor, or combine the two. Ingestion, feature engineering, and detection, Layers 1 through 3, are the layers most commonly available as mature third-party platforms, since the underlying techniques, streaming pipelines, entity resolution, and ensemble anomaly scoring, are broadly similar across institutions. Explainability and governance, Layers 4 and 6, are harder to purchase off the shelf in a form that satisfies a specific institution's validation practice, because conceptual-soundness review is inherently tied to how that institution's model risk function is organized (Federal Reserve Board of Governors, 2011). Layer 5 sits in between: workflow and routing logic is often vendor-supplied, but the risk-tier thresholds embedded in that workflow, and the analyst training needed to use it well, are institution-specific. A pragmatic implication is that institutions should expect to build or closely customize Layers 4 through 6 even where Layers 1 through 3 are substantially purchased, since those governance-facing layers are precisely where the integration gap identified in Section 2 leaves generic tooling least adequate.

7. Limitations and Future Work

As a conceptual contribution, this framework has not been empirically validated against live transaction data. Its components are each grounded in prior literature describing their use individually, but the specific combination and sequencing proposed here has not been tested, and no performance claim should be attributed to the framework as a whole on the basis of results reported for its parts. The illustrative scenarios presented in Section 4 are hypothetical and constructed for expository purposes; they demonstrate how the layers are intended to interact, not that the framework performs as described when implemented.

A related caution concerns how any future evaluation should be conducted. Performance estimates in fraud and financial-crime detection are highly sensitive to how verification latency and class imbalance are modeled, and figures obtained without accounting for those constraints systematically overstate operational performance (Dal Pozzolo et al., 2018). Any empirical test of this framework would need to adopt an evaluation protocol that reflects the delayed, partial, and non-random nature of analyst confirmation, rather than reporting accuracy against a fully labeled retrospective dataset. The framework also does not resolve, and does not claim to resolve, the underlying tension between model complexity and interpretability. Deep and graph-based components generally offer stronger detection performance than simpler models (Pang et al., 2021), but post-hoc explanation methods for these architectures approximate rather than reproduce the model's internal reasoning, and the question of how faithfully an attribution or subgraph explanation reflects the actual computation remains methodologically open (Ribeiro et al., 2016; Ying et al., 2019). An institution relying on such explanations to satisfy a regulatory obligation is relying on an approximation whose fidelity it should independently assess.

Calibrating the risk tiers used for human-in-the-loop routing in Layer 5 requires institution-specific data on the relative cost of different error types and on available reviewer capacity, information a conceptual framework cannot supply; each implementing institution would need to determine its own thresholds. Cross-jurisdictional regulatory divergence is a further limitation: obligations under the Bank Secrecy Act framework, data-protection regimes, AI-specific regulation, and

supervisory model risk guidance are not fully harmonized, and institutions operating across borders may require jurisdiction-specific variants of the governance layer, a challenge made more concrete by the staged implementation of the travel rule across jurisdictions (FATF, 2021a) and by the classification questions discussed in Section 6.2. The comparative analysis in Section 5 is qualitative rather than quantitative, drawn from characteristics reported across separate bodies of literature rather than from a controlled evaluation of the approaches on a common dataset; Table 3 should be read as a positioning device rather than as evidence of measured superiority. Similarly, the federated configuration described for Layers 2 and 3 has been validated in the cited literature primarily on benchmark rather than production data, and its behavior under a live, regulated, multi-institution deployment, including its communication overhead and its sensitivity to heterogeneous data distributions across participants, remains untested in this setting (Kairouz et al., 2021). Finally, the framework's design assumptions suit institutions large enough to justify the specialized development, validation, and case management functions described across the six layers. The paper does not address how a smaller institution operating at the lower end of the volume range defined in Section 1 should scale the framework down. Such institutions might implement the same six layers with lighter-weight components, or combine some layers organizationally while keeping them conceptually distinct, but neither possibility has been examined here.

Future research could pursue at least five directions: institution-level piloting of the framework against live or historical transaction data, using an evaluation protocol that models verification latency explicitly; comparative evaluation of the hybrid rule, statistical, and graph ensemble described in Layer 3 against single-method baselines under matched conditions; study of the feedback loop itself, including how frequently recalibration should occur, how that cadence affects drift and stability, and how it can be governed so as to remain auditable; empirical testing of the federated variant of Layers 2 and 3 in a genuine multi-institution consortium setting; and development of drift-monitoring practice calibrated specifically to adversarial financial-crime settings, where the general methodological literature on concept drift (Gama et al., 2014) has not been tailored to the particular tempo and regulatory obligations of compliance monitoring.

8. Conclusion

High-volume transaction environments have outgrown the practical capacity of static, rule-based compliance monitoring, and the combination of sustained alert volumes, constrained review capacity, and persistently low interception rates has pushed institutions toward artificial intelligence as a complement to existing controls. This paper has developed a conceptual framework organizing that response into six interdependent layers, spanning data ingestion, feature and entity-graph construction, hybrid detection, explainability, calibrated human review, and governance with continuous feedback, built around the principle that detection performance and regulatory defensibility should be treated as a single design problem rather than as separate technical and compliance concerns. By mapping these layers onto existing model risk guidance and human-oversight requirements, extending them to digital-asset, insurance, and cross-institutional contexts, illustrating their operation through a hypothetical payments scenario, and positioning them against alternative approaches, the framework offers compliance functions a way to think about AI adoption architecturally rather than as a series of point solutions. Its central limitation, that it has not been tested empirically, is also its central invitation: the framework is offered as a structure for institutional piloting and academic evaluation, not as a finished or validated system.

References

Adebayo, A., Adegbite, M. P., & Ahmed, M. O. (2023). AI augmented threat detection in industrial control systems: A systematic review of machine learning approaches for ICS anomaly detection. *International Journal of Engineering and Modern Technology, 9*(3), 287-340. https://doi.org/10.56201/ijcsmt.v9.no3.2023.pg287.340

Adegbite, M. P., Adebayo, A., & Ahmed, M. O. (2023). ICS and SCADA threat detection architectures in energy sector networks: A systematic review of SIEM, NDR, and anomaly detection approaches. *World Journal of Innovation and Modern Technology, 7*(2), 121-182. https://doi.org/10.56201/wjimt.v7.no2.2023.pg121.182

Adelanwa, A., Basnet, A., & Anene, U. N. (2023b). Predictive analytics models for financial risk detection and fraud prevention in public systems. *International Journal of Advanced Multidisciplinary Research and Studies, 3*(6). https://doi.org/10.62225/2583049X.2023.3.6.5969

Adesuyi, M. O., Akomolafe, O., Olaogun, B. O., Ndukwe, V. U., & Sakyi, J. K. (2024). AI-driven risk scoring model for global cross-border trade payment transactions. *International Journal of Advanced Multidisciplinary Research and Studies, 4*(1), 1569-1581. https://doi.org/10.62225/2583049X.2024.4.1.5281

Atakpa, M. I. (2021). A review of predictive analytics models for anti-money laundering detection in commercial banking systems. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 7*(2), 778-804. https://doi.org/10.32628/CSEIT2064813

Atakpa, M. I., & Abetoh, N. F. (2022). A systematic review of machine learning advances in financial fraud detection for banking systems. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8*(2), 771-801. https://doi.org/10.32628/CSEIT23906220

Atakpa, M. I., Abetoh, N. F., & Akeju, B. (2024). Advances in artificial intelligence for healthcare payment fraud detection: A review of NHS applications. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10*(4), 1161-1194. https://doi.org/10.32628/CSEIT26123240

Atakpa, M. I., & Fobellah, A. N. (2023). Anomaly detection in financial time-series data: A conceptual model for healthcare and banking applications. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 9*(2), 967-997. https://doi.org/10.32628/CSEIT2342441

Bank Secrecy Act, 31 U.S.C. §§ 5311-5336 (and implementing regulations at 31 C.F.R. Chapter X).

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 93-104.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), Article 15.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Consumer Financial Protection Bureau [CFPB]. (2022). *Consumer Financial Protection Circular 2022-03: Adverse action notification requirements in connection with credit decisions based on complex algorithms*.

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.

Debener, J., Heinke, V., & Kriebel, J. (2023). Detecting insurance fraud using supervised and unsupervised machine learning. *Journal of Risk and Insurance, 90*(3), 743-768.

Eboh, E. E., & Aliliele, C. (2024). AI-driven data analytics framework for risk assessment and detection of venture capital and private equity investment fraud in U.S. capital markets. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10*(6), 2624-2665. https://doi.org/10.32628/CSEIT2410787

European Parliament & Council of the European Union. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation), Article 22. *Official Journal of the European Union*, L 119.

European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Article 14 and Annex III. *Official Journal of the European Union*.

Federal Reserve Board of Governors & Office of the Comptroller of the Currency. (2011). *SR 11-7: Guidance on model risk management*.

Financial Action Task Force [FATF]. (2021a). *Updated guidance for a risk-based approach to virtual assets and virtual asset service providers*. FATF.

Financial Action Task Force [FATF]. (2021b). *Opportunities and challenges of new technologies for AML/CFT*. FATF.

Financial Crimes Enforcement Network [FinCEN], Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, National Credit Union Administration, & Office of the Comptroller of the Currency. (2018). *Joint statement on innovative efforts to combat money laundering and terrorist financing*.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys, 46*(4), Article 44.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75-105.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning, 14*(1-2), 1-210.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. *Proceedings of the 8th IEEE International Conference on Data Mining*, 413-422.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765-4774.

Mbonu, I. S., Aliliele, C., Iwuanyanwu, U., & Uzoka, E. (2022). A conceptual framework for AI enabled IT general controls and SOX audit automation processes. *Gyanshauryam, International Scientific Refereed Research Journal, 5*(5), 384-414. https://doi.org/10.32628/GISRRJ2256239

Mbonu, I. S., Iwuanyanwu, U., Aliliele, C., & Uzoka, E. (2021). A review of VoIP forensic analytics models for financial fraud detection and regulatory compliance monitoring.

International Journal of Multidisciplinary Research and Growth Evaluation, 2(6), 711-730. https://doi.org/10.54660/.IJMRGE.2021.2.6.711-730

McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 1273-1282.

Pang, G., Shen, C., Cao, L., & van den Hengel, A. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys, 54*(2), Article 38.

Pourhabibi, T., Ong, K.-L., Kam, B. H., & Boo, Y. L. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems, 133*, 113303.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.

United Nations Office on Drugs and Crime [UNODC]. (2011). *Estimating illicit financial flows resulting from drug trafficking and other transnational organized crimes*. UNODC.

Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., & Leiserson, C. E. (2019). *Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics*. arXiv:1908.02591.

Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. *Advances in Neural Information Processing Systems, 32*, 9244-9255.