SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

A Conceptual AI/ML Framework for Fraud Detection and Anti-Money Laundering in Fintech: Architecture, Algorithms, and Regulatory Alignment

Dr. Mahesh Devidas Mahankal[1], Dr Anil Poman[2], Dr. Amol Pandurang Godge[3], Madhuri Chaudhari[4], Dr. Vaishali Dhanaji Nikam[5], Dr. Ganesh Sambhaji Lande[6], Dr. Prashant Bhavarlal Chordiya[7]

[1]Assistant Professor, Department of MBA, Yashaswi Education Society's International Institute of Management Science, Pune, India. Email: mahesh.mahankal@gmail.com, ORCID ID: 0009-0007-8821-9830
[2]Associate Professor, Department of MBA, Lotus Business School, Pune, India. E-mail: anilpoman147@gmail.com, ORCID: 0009-0009-2981-2217
[3]Assistant Professor, Department of MBA, Dr D Y Patil School Of Management Charholi(bk.), Lohegaon, Pune, India. E-mail: amol.godge@gmail.com, ORCID ID: 0009-0006-6184-3513
[4]Assistant Professor, Department of MBA, Dr. D Y Patil Center for Management and Researc , Chikhali, Pune, India E-mail: 12madhurichaudhari@gmail.com, ORCID ID - 0009-0003-0511-1842
[5]Associate Professor, Department of MBA, Shree Shivaji Maratha Society's Institute of Management and Research, Pune, India. E-mail: vaishalinikam16@gmail.com
[6]Associate Professor, Department of MBA, Dr D Y Patil School of Management, Pune, India E-mail: gslande@gmail.com, ORCID ID - 0000-0003-2036-8681
[7]Associate Professor, D.Y. Patil Institute of Master of Computer Applications and Management, Pune, India.
E-mail: prashant.chordiya@gmail.com, 0009-0005-6227-2583

Abstract
The amount of yearly transactions on financial technology platforms is counted in billions, and as much as the rapidity and accessibility make fintech attractive to legit users, it also makes it attractive to fraudsters and money launderers. The rule-based transaction monitoring tools, which still comprise the core of most companies' compliance programs, generate huge numbers of false positives and are unable to adjust to new fraud types, while black box machine learning models which detect fraud more accurately fail to provide case-to-case explanations which are necessary for Suspicious Activity Reports according to FATF and other regulations. The paper provides a conceptual 5-layer framework which comprises the layers of data ingestion and KYC processing, feature engineering, multi-family modeling layer containing supervised, unsupervised and graph-based algorithms, explainability and risk scoring layer involving SHAP and LIME and compliance/regulatory reporting layer with the concept drift feedback loop between the last two layers allowing for analyst decision routing to the retraining of models. The paper takes advantage of the knowledge about deep learning models, graph neural networks, mitigation of class imbalance problem and federated learning for anti-fraud and AML purposes in order to substantiate each of the framework layers. The taxonomies of candidate algorithms, mappings of framework layers to the corresponding regulatory requirements and a 5-phase implementation roadmap are provided as practical deliverables which an actual fintech company can utilize in its deployment. Consistent with its stated scope, this paper is conceptual: it does not report empirical detection performance on any dataset, and Section 11 explicitly identifies empirical validation as the required next step before any component of the framework could be considered production-ready.

Keywords: Fintech; Fraud Detection; Anti-Money Laundering; Machine Learning; Explainable AI; Graph Neural Networks; RegTech; Financial Crime Compliance

Introduction

Digital payments, neobanking, buy-now-pay-later credit, and cryptocurrency on-ramps have moved most retail financial activity onto platforms that authorise a transaction in milliseconds and settle it in seconds. This same infrastructure is attractive to fraud rings and money launderers precisely because it is fast, high-volume, and often spans multiple institutions and jurisdictions in a single transaction chain. Industry survey data indicates that the large majority of banks already use artificial intelligence (AI) for financial-crime and fraud detection, and that decision-makers expect both fraud and financial-crime activity to keep increasing even as AI-based defences improve, because criminal actors are adopting the same technology to design attacks.

The classical threshold-based method of transaction monitoring based on pre-defined limits regarding the size of transactions, their speed, and geography remains common owing to the simplicity of its audit and

comprehension by regulators. However, such static rules lack flexibility since they fail to account for the relations in a money laundering scheme involving several mule accounts, as well as create too many false positives which overwhelm the work of compliance officers. It has been shown in literature that machine learning techniques such as gradient-boosted trees, deep neural networks, and GNNs outperform rule-based systems when it comes to detecting fraud. Yet ML adoption in regulated financial-crime workflows is constrained by three recurring problems documented across the literature reviewed in Section 2: (i) severe class imbalance, since fraudulent or laundering transactions are typically a fraction of one percent of all activity; (ii) concept drift, since fraud tactics evolve specifically to evade whatever detection method is currently deployed; and (iii) a compliance and interpretability gap, since regulators require an auditable, case-level explanation for every reported transaction, which conflicts with the black-box nature of the most accurate model families.

This paper addresses these three problems not by proposing a new algorithm, but by proposing a conceptual system architecture that positions existing, well-validated algorithmic building blocks into a coherent, regulator-aligned pipeline. The specific contributions of this paper are:

- A five-layer conceptual framework (Section 4) spanning data ingestion, feature engineering, modelling, explainability, and compliance reporting, with an explicit feedback loop for concept-drift retraining.
- A taxonomy of candidate algorithm families for the model layer (Section 5), comparing supervised, unsupervised, and graph-based approaches on interpretability, latency, and typical use case.
- A synthesis of mitigations for the three recurring technical challenges of class imbalance, concept drift, and explainability (Section 6), grounded in recent peer-reviewed and preprint literature.
- An explicit mapping between framework layers and regulatory obligations under FATF guidance, Reserve Bank of India (RBI) directions, and the EU General Data Protection Regulation / AI Act (Section 7).
- A five-phase implementation roadmap (Section 8) that a fintech institution or bank could use to plan an actual pilot deployment of the framework.

The remainder of the paper is organised as follows. Section 2 reviews related work on ML-based fraud detection, graph-based AML, explainable AI in finance, and federated learning. The Problem Statement and the justification for the Framework Approach will be covered in Section 3. Section 4 and 5 will describe the framework and the building blocks. Technical Challenges and Mitigation Measures will be addressed in Section 6. Section 7 will focus on Regulatory Compliance of the Framework. Section 9 discusses practical fintech use cases. Sections 10 and 11 discuss the framework's implications and limitations, and Section 12 concludes with directions for future empirical work.

2. Related Work

2.1 Supervised and Deep Learning Approaches to Fraud Detection

Bin Sulaiman et al. (2022) reviewed the machine learning approaches applied to credit card fraud detection and found that ensemble tree-based methods and neural architectures consistently outperform single classical classifiers, while noting that most studies still evaluate on a small number of public benchmark datasets that do not reflect the concept drift of production systems. Hilal, Gadsden, and Yawney (2022) provided a broad review of anomaly-detection techniques for financial fraud, distinguishing statistical, classical ML, and deep learning approaches and highlighting that reported performance varies substantially with how researchers handle the extreme class imbalance inherent in fraud data. More recent systematic literature reviews using the Kitchenham methodology - Hernandez Aros et al. (2024) analysing 104 articles from 2012-2023, and Chen, Zhao, Xu, and Nie (2025) analysing 57 deep-learning-specific studies from 2019-2024 - both report a marked increase in published work from 2022 onward, concentrated in the credit-card and banking sectors, and both identify model interpretability and imbalanced datasets as the most frequently cited open challenges.

2.2 Graph-Based Approaches to Anti-Money Laundering

Because money laundering is inherently relational - funds are deliberately routed through chains of accounts and intermediaries to obscure origin - a distinct line of research models transactions as graphs rather than independent feature vectors. Early work by Weber et al. applied graph convolutional networks to the Elliptic Bitcoin transaction dataset, demonstrating that relational features improve detection of illicit cryptocurrency transactions over transaction-level features alone. Other research in GNNs for financial fraud detection has mentioned applications of self-supervised graph representation learning and group-aware deep graph learning for detecting collusion laundering among multiple accounts, and gave examples of how the use of these models has varied from credit card fraud detection to crypto-currency analysis. GNNs with explainable layers have been used by other research as well; Olaniyi et al. (2026) mentioned the design of a GNN framework for financial crime networks with multiple layers which has employed SHAP, LIME, and attention explainability techniques for regulators, just like in section 4.5 of this study.

2.3 Explainable AI in Financial Risk and Compliance Systems

The regulatory requirement for explainability in credit and financial-crime decisions has produced a

substantial body of work applying post-hoc explanation methods to financial ML models. SHAP (Lundberg and Lee, 2017), based on Shapley values from cooperative game theory, and LIME (Ribeiro, Singh, and Guestrin, 2016), which fits a locally interpretable surrogate model around a single prediction, are the two most widely adopted model-agnostic explanation techniques in the finance literature. Gramegna and Giudici (2021) evaluated the discriminative power of SHAP and LIME specifically for credit-risk models, and subsequent work has extended tree-specific SHAP explanations to production-scale gradient-boosted models. Industry and professional-body guidance, including CFA Institute analysis of explainable AI in finance, identifies distinct explanation needs across consumer, analyst, and regulator stakeholder groups - a distinction this paper adopts directly in the design of the explainability layer (Section 4.5) and the stakeholder-mapped comparison in Table 3.

2.4 Privacy-Preserving and Federated Approaches

A separate constraint on fraud and AML modelling is that the most informative signal - coordinated activity across multiple institutions - is exactly the data that data-protection regulation and competitive concerns prevent institutions from sharing directly. Federated learning (FL) has been proposed by multiple recent studies as a solution: banks or fintech platforms train a shared model collaboratively while raw transaction data never leaves the originating institution, with only model parameters or gradients exchanged. Reported evaluations, for example a cross-institutional FedAvg deep-neural-network study, found that a federated global model achieved a meaningfully higher AUROC than models trained in isolation at each institution, while keeping raw data local. This line of work motivates the cross-institutional federated variant of the model layer discussed in Section 6.4.

2.5 Research Gap

Taken together, the literature reviewed above establishes that (a) deep learning and graph-based models outperform classical baselines on fraud and AML detection tasks; (b) explainability methods such as SHAP and LIME are mature enough for production use in credit-risk contexts; and (c) federated learning is a viable route to cross-institutional collaboration without raw data sharing. What remains comparatively underspecified in the literature is how these three separately-validated components - a graph-capable model layer, a stakeholder-differentiated explainability layer, and a privacy-preserving training regime - should be integrated into a single pipeline that also satisfies a specific, named set of regulatory obligations end to end, from transaction ingestion through to Suspicious Activity Report (SAR) filing. This paper addresses that integration gap conceptually, rather than proposing a new point algorithm.

3. Problem Statement and Motivation

Four key aspects of the fraud-and-AML detection problem collectively justify the use of an integrated architecture as a solution rather than the top performing model alone.

- Class imbalance. The amount of transactions that are fraudulent or involve money laundering is below one percent of the total number of transactions; some benchmark datasets feature even worse class imbalance, showing that the ratio of the positive class is at most 0.17 percent. A model that seeks maximum accuracy would end up predicting that all the transactions were non-fraudulent.
- Concept drift due to adaptive adversary. This particular problem is exceptional in comparison to the majority of the other classification problems in the sense that the very idea of fraud changes as a consequence of efforts made by fraud rings to adjust themselves to the existing anti-fraud system.
- The need for explanation from the regulator. Due to the requirements for suspicious activity reports, fair lending, and artificial intelligence transparency, the bank needs to justify why the transaction is considered to be suspicious or is refused to be processed based on its probability output.
- Fragmentation of the data between institutions. The diagnostic attributes of fraud and money laundering can be extracted only in the context of relations between different institutions; however, due to data privacy and competition, they cannot share raw data.

Point solutions proposed in the existing literature focus on one of the four aspects individually – either a sampling technique to handle class imbalance, a detection algorithm to detect concept drift, an explanation mechanism for the classification output, or a federated learning framework – but none of them explain how all four aspects can be integrated together and in which order they have to be applied along with the interaction between each other. The architecture described in Section 4 attempts to answer these questions.

4. Proposed Conceptual Framework

4.1 Design Principles

- Explainability-oriented design: the downstream layers have to be designed in such a way as to support tracing of the calculated risk score to the particular combination of features and graph relations that were used for calculating the particular risk score.

- Adaptability-oriented design: there is an inherent feedback loop built into the architecture (see Figure 1 below), which captures the analyst's disposition of alerts and uses it as an input for re-training at periodic intervals, rather than assuming that the model that was trained once will always remain valid.
- Regulatory alignment-oriented design: the compliance layer is one of the primary components of the architecture, not just something added at the end; see Section 7 for more detail.
- Modular by algorithm family: the model layer is deliberately specified as a family of candidate algorithms (Section 5) rather than a single fixed model, since the appropriate algorithm depends on the specific fraud **typology, available compute, and interpretability requirement of a given deployment**

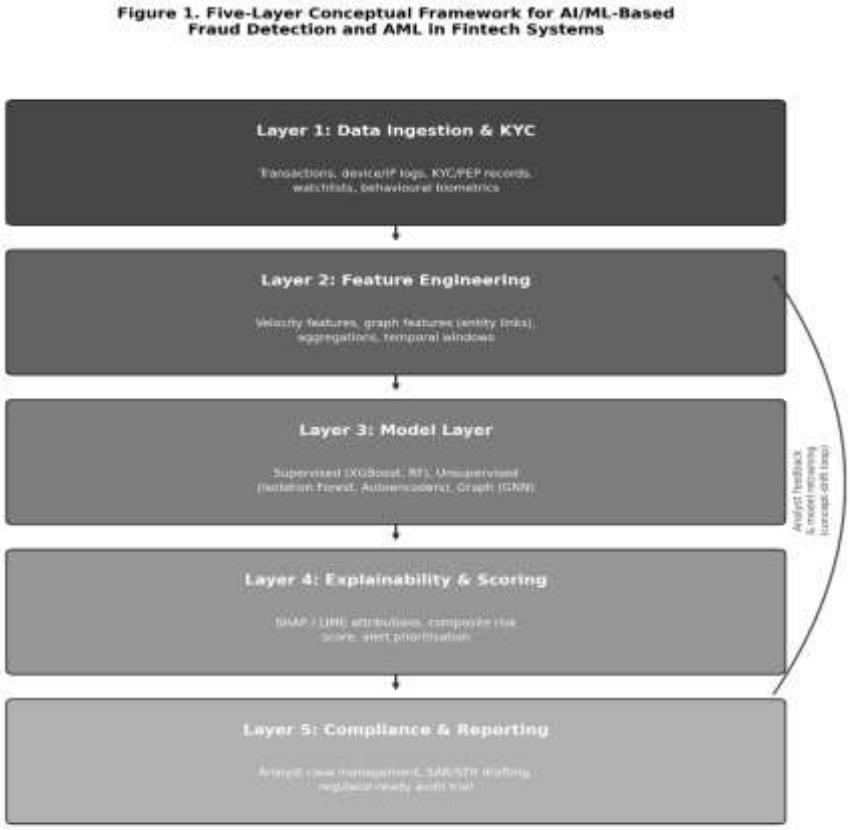

Figure 1. Five-layer conceptual framework for AI/ML-based fraud detection and AML in fintech systems, with an analyst-feedback loop supporting concept-drift retraining.

4.2 Layer 1: Data Ingestion and KYC

This layer ingests the raw signal the rest of the framework depends on: transaction records (amount, timestamp, counterparties, channel), device and network metadata (IP address, device fingerprint, geolocation), KYC and Politically Exposed Person (PEP) records collected at onboarding, sanctions and watchlists, and, where available, behavioural biometrics such as typing or navigation patterns. Data quality and completeness at this layer bound the achievable performance of every downstream layer; in particular, KYC completeness directly determines whether the entity-resolution step required by the graph-based algorithms in Section 5 can reliably identify that two account records belong to the same beneficial owner.

4.3 Layer 2: Feature Engineering

Raw records from Layer 1 are transformed into model-ready features along three families: (i) velocity and aggregation features - transaction counts, sums, and rates over sliding time windows (1 hour, 24 hours, 7 days), which are the primary signal for classical supervised models; (ii) graph and relational features - degree centrality, shortest-path distance to a known bad actor, and community-detection membership within the transaction graph, which feed the graph-based model family; and (iii) behavioural-deviation features - distance between a transaction and the account's own historical baseline, which feed the unsupervised anomaly-detection family. Feature engineering at this layer must itself be drift-aware: a feature's normal range should be computed over a rolling, not fixed, historical window, or the feature will silently become stale as legitimate customer behaviour evolves.

4.4 Layer 3: Model Layer

The model layer hosts the three algorithm families detailed in Section 5 and Figure 2 - supervised classifiers, unsupervised anomaly detectors, and graph-based models - run in parallel rather than as a single pipeline, with their outputs combined into a composite risk score at Layer 4. Running multiple families in parallel, rather than selecting one best model, follows directly from the literature reviewed in Section 2.1: *no single family dominates across all fraud typologies*, since supervised models are strongest against known fraud patterns with sufficient labelled examples, unsupervised models are the only family capable of flagging a genuinely novel pattern with zero prior labelled examples, and graph-based models are uniquely able to surface coordinated multi-account activity that looks unremarkable at the level of any single transaction.

4.5 Layer 4: Explainability and Risk Scoring

Outputs from Layer 3 are combined into a single composite risk score, and every score above an alerting threshold is passed through a post-hoc explanation method - SHAP for the tree-based and neural components, attention-weight extraction for the graph-based component - before it ever reaches a human analyst. The order (first, explain, then alert; as opposed to first, alert, then explain) was chosen intentionally since it allows one to prepare the explanation while all the alerts are already in the analyst's queue, and therefore there is no need to undertake additional actions to prepare an explanation of the event.

4.6 Layer 5: Compliance and Regulatory Reporting

Layer 5 provides the means to manage the analyst case, creates SAR/STR stories based on information provided by Layer 4, and produces an immutable audit trail for each model run and alert and for each analyst decision. This is precisely the kind of record keeping that would be needed in a supervisory authority examination or FATF review (Section 7).

4.7 Feedback Loop and Concept-Drift Retraining

Every analyst disposition at Layer 5 (confirmed fraud, false positive, escalated for investigation) is captured and routed back to Layer 3 as a labelled training example, closing the loop shown as the red arc in Figure 1. This is the framework's explicit mechanism for addressing the concept-drift property identified in Section 3: rather than retraining on a fixed schedule alone, the model layer is retrained (or, for lighter-weight updates, recalibrated) whenever monitored performance metrics on this feedback stream - precision and recall on the analyst-confirmed subset - drift beyond a pre-agreed threshold, an approach consistent with the adaptive and continual-learning strategies discussed in the AML graph-learning literature reviewed in Section 2.2.

5. Algorithmic Building Blocks

Figure 2 organises the candidate algorithms for the Layer 3 model layer into three families. Table 1 compares them on the dimensions most relevant to a fintech deployment decision: typical use case, interpretability without **additional tooling, relative inference latency, and principal limitation.**

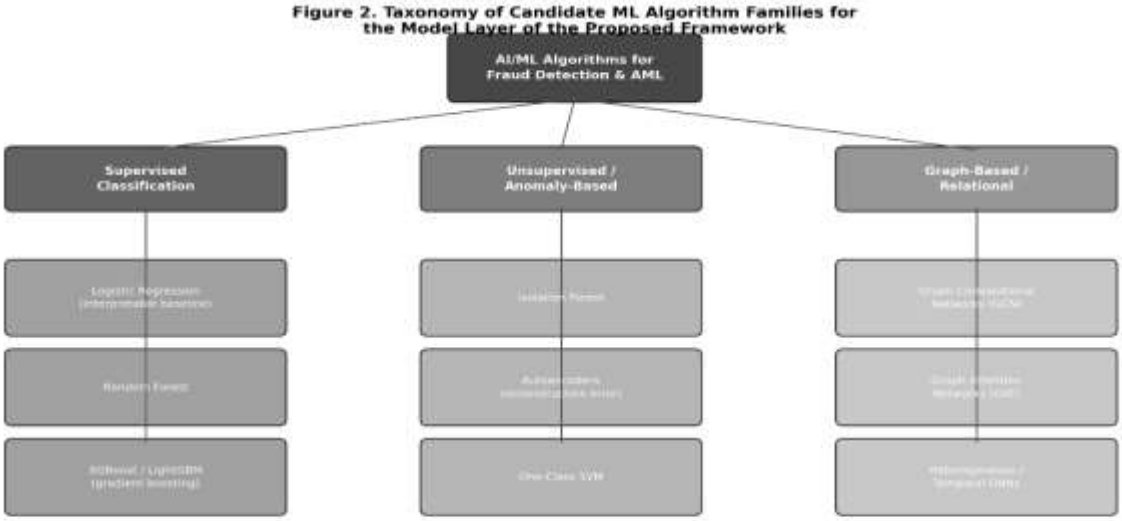

Figure 2. Taxonomy of candidate ML algorithm families for the model layer of the proposed framework.

Family / Algorithm	Typical Use Case	Native Interpretability	Inference Latency	Principal Limitation
Logistic Regression	Baseline credit/fraud scoring; regulator-facing benchmark	High	Very low	Limited capacity for non-linear fraud patterns
Random Forest	General-purpose transaction fraud scoring	Medium (feature importance)	Low	Can overfit on rare, highly specific fraud patterns
XGBoost / LightGBM	High-accuracy transaction and card-fraud scoring	Medium (with TreeSHAP)	Low	Requires TreeSHAP or similar for case-level explanation
Isolation Forest	Novel / zero-day fraud pattern flagging (unsupervised)	Low	Low	No use of the labelled fraud examples an institution already holds
Autoencoders	Behavioural-deviation anomaly detection	Low	Medium	Reconstruction-error threshold requires careful, ongoing calibration
One-Class SVM	Small-sample anomaly detection at account level	Low	Medium	Scales poorly to high transaction volumes
Graph Convolutional Network (GCN)	AML: detecting laundering rings and mule-account networks	Low (needs post-hoc XAI)	Medium-High	Computationally demanding on large, dense transaction graphs
Graph Attention Network (GAT)	AML with heterogeneous entity/transaction relationships	Low-Medium (attention weights)	Medium-High	Attention weights are a partial, not complete, explanation
Heterogeneous / Temporal GNN	Cross-border, multi-hop, time-evolving laundering chains	Low	High	Most data- and compute-intensive family; least mature tooling

Table 1. Comparison of candidate algorithm families for the Layer 3 model layer.

6. Core Technical Challenges and Proposed Mitigations

6.1 Class Imbalance

Because fraudulent and laundering transactions are a small minority class, models trained naively on raw transaction data tend to achieve high overall accuracy while missing most actual fraud. These are the three ways of dealing with the issue of class imbalance as described in the literature and depicted in Table 2. They may involve resampling of the data (oversampling and undersampling) or the creation of synthetic minority oversampling and the cost-sensitive learning process where there is higher cost of false negatives than the cost of false positives. The most common method of synthetic minority oversampling, as per the literature, is that of SMOTE or Synthetic Minority Oversampling Technique as proposed by Chawla et al. (2002).

Technique	Mechanism	Primary Trade-off
Random Under-sampling	Removes majority-class (legitimate) examples until classes are balanced	Discards potentially useful legitimate-transaction signal
Random Over-sampling	Duplicates minority-class (fraud) examples	Risks overfitting to duplicated examples
SMOTE	Generates synthetic minority examples by interpolating between existing fraud examples	Can generate implausible synthetic examples in sparse regions
Conservative / plausibility-	Filters synthetic examples that fall outside a	Added computational and

Technique	Mechanism	Primary Trade-off
filtered SMOTE	plausibility boundary before adding them	validation overhead
Cost-sensitive learning	Assigns a higher misclassification penalty to false negatives during training	Requires a defensible, business-agreed cost ratio
Class-weighted ensembles	Weights the minority class more heavily within tree-ensemble training (e.g., Random Forest, XGBoost)	Weight must be tuned per deployment; not a one-time setting

Table 2. Class-imbalance mitigation techniques applicable within the Layer 3 model layer.

6.2 Concept Drift and Adaptive Retraining

Concept drift in this domain is adversarial rather than incidental: fraud rings actively probe deployed detection systems and adapt their behaviour to evade them, so a model's effective accuracy can degrade faster than in typical production ML settings where the underlying data distribution shifts for unrelated reasons. The framework's feedback loop (Section 4.7) is the primary architectural response; recent continual-learning research on graph-based AML systems additionally proposes replay-based, regularisation-based, and architecture-based strategies specifically for retraining graph models without catastrophically forgetting previously learned laundering patterns, which is a relevant extension for institutions adopting the graph-based family in Table 1.

6.3 Explainability for Regulatory Compliance

Different stakeholders at Layer 4 and Layer 5 require different kinds of explanation, an observation drawn directly from CFA Institute guidance on explainable AI in finance. Table 3 maps the two dominant post-hoc explanation methods, SHAP and LIME, against the stakeholder groups the proposed framework must serve.

Stakeholder	Explanation Need	Primary Method
Front-line analyst (Layer 5)	Which features and relationships drove this specific alert, ranked by contribution	SHAP local explanation (force / waterfall plot)
Model validation / MLOps team	Global feature importance and stability across model versions	SHAP summary (beeswarm) and bar plots
Compliance officer / SAR drafter	A concise, natural-language justification suitable for a regulatory narrative	SHAP or LIME output translated into templated case-note text
External regulator / auditor	An auditable trail linking a specific decision back to specific evidence, reproducible after the fact	Stored SHAP values plus the model version and feature-snapshot at decision time
Customer (adverse-action context)	A minimal, actionable explanation of why a transaction or account action occurred	Counterfactual explanation (e.g., minimal change that would have avoided the flag)

Table 3. Mapping of explainable-AI methods to the stakeholder groups served by Layer 4 and Layer 5 of the proposed framework.

6.4 Cross-Institutional Privacy: Federated Learning

Where the framework is extended across multiple institutions - for instance, a consortium of fintech lenders wanting to detect fraud rings that spread exposure across several platforms - Layer 3 can be instantiated as a federated learning setup, in which each participating institution trains locally on its own data and only aggregated model parameters, not raw transaction records, are exchanged with a coordinating server. This directly addresses the cross-institutional fragmentation problem identified in Section 3, and the federated-learning literature reviewed in Section 2.4 reports that a federated global model can approach or exceed the performance of any single institution's locally trained model while keeping raw data local, at the cost of additional communication overhead and sensitivity to how heterogeneous each institution's data distribution is.

6.5 Adversarial Robustness

Because the Layer 3 model outputs directly influence account-level actions (blocking a transaction, freezing an account), the model layer is itself a target for adversarial manipulation - for example, an attacker

deliberately structuring a sequence of small transactions to stay just under a learned detection threshold. The framework's mitigation is architectural rather than algorithmic: running three independent algorithm families in parallel (Section 4.4) means an adversarial pattern crafted to evade one family (say, a supervised classifier's decision boundary) still has to simultaneously evade the unsupervised anomaly detector's behavioural baseline and the graph model's relational-structure signal, which is a substantially harder joint evasion problem than evading a single model.

7. Regulatory and Compliance Alignment

An architecture capable of detecting fraud and money laundering must be implementable insofar as its architecture should be able to map the regulations that apply to financial institutions through national and international regulations. It is explicitly mentioned in FATF guidelines for 2025 on AML and financial inclusion that artificial intelligence, machine learning, and regtech for real-time transaction monitoring is one such technology that could assist in increasing compliance efficiency through reduced false positives. In India, the RBI's October 2024 guidance on internal money-laundering / terrorist-financing risk assessment requires regulated entities to maintain documented, defensible risk-assessment processes, which the audit-trail component of Layer 5 is designed to support directly. Table 4 maps each layer of the proposed framework to the regulatory obligation it is intended to satisfy.

Framework Layer	Regulatory Obligation	Governing Body / Reference
Layer 1: Data Ingestion & KYC	Customer due diligence, PEP and sanctions screening at onboarding	FATF Recommendations; RBI KYC Master Direction
Layer 2: Feature Engineering	Data minimisation and purpose limitation in derived-feature use	GDPR Articles 5-6; India DPDP Act
Layer 3: Model Layer	Risk-based, documented internal ML/TF risk assessment methodology	RBI (Oct. 2024) internal risk-assessment guidance
Layer 4: Explainability & Scoring	Right to explanation / adverse-action notice for algorithmic decisions	EU AI Act (high-risk systems); GDPR right to explanation; Fair Lending (ECOA)
Layer 5: Compliance & Reporting	Timely filing of Suspicious Activity / Transaction Reports; audit trail	FATF Recommendation 20; national SAR/STR regimes
Feedback Loop	Ongoing model validation and periodic re-assessment of AML/CFT risk	FATF risk-based approach; SR 11-7-style model-risk guidance

Table 4. Mapping of proposed framework layers to specific regulatory and compliance obligations.

8. Implementation Roadmap

Figure 3 and Table 5 propose a five-phase roadmap by which a fintech institution could move from the conceptual framework in Section 4 to a piloted, regulator-reviewed deployment. The timeline is indicative and would need to be adapted to an institution's existing data infrastructure and regulatory relationship.

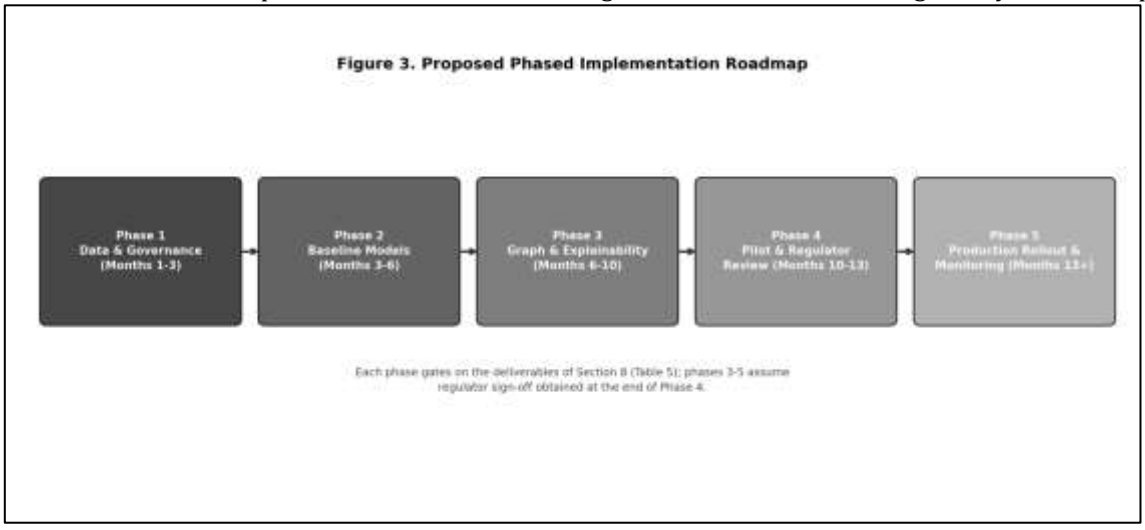

Figure 3. Proposed phased implementation roadmap.

Phase	Duration	Key Deliverables	Key Risks
1. Data & Governance	Months 1-3	Data inventory; KYC data-quality audit; governance sign-off on feature use	Incomplete or siloed historical KYC data
2. Baseline Models	Months 3-6	Supervised and unsupervised baselines (Table 1) trained and back-tested	Insufficient labelled fraud examples for supervised training
3. Graph & Explainability	Months 6-10	Graph-based model integrated; SHAP/LIME explanation layer wired to alert queue	Compute cost and latency of graph inference at production volume
4. Pilot & Regulator Review	Months 10-13	Shadow-mode pilot; SAR-narrative templates reviewed with compliance and, where required, the regulator	Pilot results insufficiently documented for regulator sign-off
5. Production Rollout & Monitoring	Months 13+	Phased production rollout; feedback-loop retraining cadence established	Concept drift outpacing the agreed retraining cadence

Table 5. Deliverables and key risks for each phase of the proposed implementation roadmap.

9. Practical Applications and Use Cases

9.1 Real-Time Digital Payment Fraud (UPI / Card-Not-Present)

High-volume, low-latency payment rails such as India's Unified Payments Interface (UPI) or card-not-present e-commerce transactions are the natural fit for the supervised and unsupervised branches of Layer 3, where sub-second scoring is required and graph inference latency (Table 1) may be too high for the synchronous authorisation path; graph scoring can instead run asynchronously to flag accounts for review after authorisation.

9.2 Cross-Border Remittance and Correspondent Banking AML

Cross-border remittance chains, where laundering typically involves multiple intermediary accounts across jurisdictions, are the primary use case for the graph-based model family and the heterogeneous, temporal GNN variants in particular, since the relational structure of a layered laundering chain is precisely what transaction-level features alone cannot represent.

9.3 Neobank and Fintech-Lender Onboarding

Digital-only banks and lenders with fully remote onboarding are exposed to synthetic-identity and first-party fraud at account opening; Layer 1's KYC and device/behavioural-biometric ingestion, combined with the unsupervised anomaly family's ability to flag never-before-seen patterns, is directly applicable here.

9.4 Virtual Asset Service Provider (VASP) and Crypto On-Ramp Monitoring

FATF's Travel Rule and expanding virtual-asset oversight increasingly require VASPs to share sender/receiver information and to monitor transaction chains comparable to those a bank would monitor; the graph-based model family, already validated on blockchain transaction graphs in the literature reviewed in Section 2.2, is directly transferable to this use case.

10. Discussion

The proposed framework's central claim is architectural, not algorithmic: none of the individual components in Table 1 or Table 2 is novel, and each is independently supported by the literature reviewed in Section 2. What the framework contributes is an explicit specification of how these components should be sequenced and connected - explain-then-alert rather than alert-then-explain, three algorithm families run in parallel rather than a single selected model, and a feedback loop that treats concept drift as a first-class operational concern rather than a reason for periodic manual retraining. This sequencing has direct implications for how a fintech institution would organise its engineering and compliance teams: it implies compliance and model-risk stakeholders need to be involved from Layer 4's design (which explanation is generated for which stakeholder) rather than only at Layer 5's reporting stage, and it implies the retraining cadence in Layer 3 needs to be an agreed operational SLA between the data-science and compliance functions, not a purely technical decision.

The framework also makes an implicit prioritisation choice that deserves to be stated explicitly: it treats regulatory auditability as a hard architectural constraint that other design choices (which specific algorithm, which specific explanation method) must satisfy, rather than as a property to be retrofitted once a high-accuracy model has already been selected. This is a defensible choice given the legal consequences of an unexplainable adverse action in lending or an unsupportable SAR filing, but it is also a choice, and an

institution that weighted raw detection accuracy more heavily than explainability might reasonably favour a different point on the interpretability-accuracy trade-off documented in Table 1 - for instance, using the temporal GNN family in production despite its weaker native interpretability, and accepting a heavier post-hoc explanation burden at Layer 4 in exchange for the strongest available detection of multi-hop laundering chains.

11. Limitations

- This paper is explicitly conceptual. No component of the framework has been implemented or evaluated on any dataset, public or proprietary, and no detection-performance figures (precision, recall, F1, AUC, or false-positive rate) are claimed anywhere in this paper for the framework as a whole.
- The algorithm comparisons in Table 1 and the technique comparisons in Table 2 and Table 3 are drawn from reported findings across the separate studies cited in Section 2, each evaluated on its own dataset and protocol; they should be read as a synthesis of directional findings from the literature, not as a head-to-head benchmark conducted under a common protocol.
- The regulatory mapping in Table 4 reflects FATF guidance, RBI direction, and EU regulation as of the literature reviewed in Section 2 and Section 7; regulatory requirements in this domain change frequently, and any institution using this framework as a reference would need to verify current requirements in its own jurisdiction before relying on Table 4.
- The implementation roadmap in Section 8 is indicative rather than validated: the phase durations in Table 5 are illustrative planning estimates, not derived from any tracked real-world deployment, and would need to be adapted to a specific institution's existing infrastructure, data maturity, and regulatory relationship.
- Adversarial robustness (Section 6.5) is addressed only at the architectural level (running multiple model families in parallel); the framework does not specify or evaluate any adversarial-training or adversarial-detection technique, which remains an open technical problem in the wider ML-security literature.

12. Conclusion and Future Work

This paper has proposed a five-layer conceptual framework integrating data ingestion, feature engineering, a multi-family model layer, an explainability and risk-scoring layer, and a compliance and regulatory-reporting layer, connected by a concept-drift feedback loop, for AI/ML-based fraud detection and anti-money laundering in fintech systems. The framework's contribution is the explicit, regulator-aligned integration of components - supervised, unsupervised, and graph-based modelling; SHAP/LIME explainability mapped to specific stakeholder needs; class-imbalance mitigation; and optional federated learning for cross-institutional collaboration - each of which is independently supported by the literature reviewed in Section 2, but which have not previously been specified together as a single deployable architecture mapped explicitly to FATF, RBI, and GDPR/EU AI Act obligations, as done in Table 4 of this paper.

The most direct next step, and the limitation identified first in Section 11, is empirical validation. Future work should instantiate the Layer 3 model layer described in Section 5 against at least one public benchmark dataset (for example, a card-fraud or Elliptic-style AML benchmark) and, subject to data-sharing agreements, at least one proprietary institutional dataset, following the protocol implied by Sections 4-6 of this paper, and report the resulting precision, recall, and false-positive-rate figures under the class-imbalance mitigations discussed in Section 6.1. A second direction for future work is a formal cost-benefit study of the parallel three-family model layer proposed in Section 4.4 against a single best-performing model, to quantify whether the adversarial-robustness benefit argued qualitatively in Section 6.5 is large enough in practice to justify the added inference cost of running three algorithm families rather than one. A third direction is a live regulator-facing pilot of the Layer 4/5 explanation-to-SAR-narrative pipeline, to establish empirically whether SHAP- and LIME-derived case notes reduce analyst investigation time relative to current manual case-note drafting.

References

1. Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. Journal of Network and Computer Applications, 68, 90-113.
2. Alhasawi, Y., et al. (2025). A federated approach to scalable and trustworthy financial fraud detection. Security and Privacy. Wiley Online Library.
3. Andrade-Arenas, L., & Yactayo-Arias, C. (2025). Comparative analysis of machine learning models for credit card fraud detection using SMOTE for class imbalance. International Journal of Safety and Security Engineering, 15(5), 893-901.

4. Becerra-Suarez, F. L., Arones-Perez, P. P., Zamora-Del-Aguila, F. F., & Forero, M. G. (2026). Conservative plausibility-filtered SMOTE for credit card fraud detection under extreme class imbalance. Frontiers in Artificial Intelligence, 9, 1871972.
5. Bin Sulaiman, R., Schetinin, V., & Sant, P. (2022). Review of machine learning approach on credit card fraud detection. Human-Centric Intelligent Systems, 2(1), 55-68.
6. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
7. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.
8. Chen, J., Chen, Q., Jiang, F., Guo, X., Sha, K., & Wang, Y. (2024). SCN_GNN: A GNN-based fraud detection algorithm combining strong node and graph topology information. Expert Systems with Applications, 237, 121643.
9. 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.
10. Chen, Y., Zhao, C., Xu, Y., & Nie, C. (2025). Year-over-year developments in financial fraud detection via deep learning: A systematic literature review. arXiv:2502.00201.
11. Financial Action Task Force. (2021). Opportunities and challenges of new technologies for AML/CFT. FATF, Paris.
12. Financial Action Task Force. (2025). Guidance on AML and financial inclusion. FATF, Paris.
13. Gramegna, A., & Giudici, P. (2021). SHAP and LIME: An evaluation of discriminative power in credit risk. Frontiers in Artificial Intelligence, 4, 752558.
14. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. arXiv:1802.01933.
15. Hernandez Aros, L., Bustamante Molano, L. X., Gutierrez-Portela, F., Moreno Hernandez, J. J., & Rodriguez Barrero, M. S. (2024). Financial fraud detection through the application of machine learning techniques: A literature review. Humanities and Social Sciences Communications, 11, 1130.
16. Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: A review of anomaly detection techniques and recent advances. Expert Systems with Applications, 193, 116429.
17. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations (ICLR).
18. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 30.
19. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56-67.
20. Moody's Analytics. (2025). AML in 2025: How AI, real-time monitoring, and global regulation are transforming compliance. Moody's Insights.
21. Olaniyi, O. M., Aroh, I. S., Metibemu, O. C., & Akinola, O. I. (2026). Graph neural networks for multi-layered financial crime network detection: An explainable AI framework for anti-money laundering. Journal of Engineering Research and Reports, 28(2), 18-36.
22. Reserve Bank of India. (2024). Guidance note on internal money laundering / terrorist financing risk assessment for regulated entities. RBI, Mumbai.
23. 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.
24. 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.
25. West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. Computers & Security, 57, 47-66.
26. Yang, W., & Wu, D. (2024). AI-driven fraud detection models in financial networks: A comprehensive systematic review. IEEE Access, 12, 98765-98790.