Journal of Economics, Finance and Accounting Studies (JEFAS)
ISSN: 2709-0809
DOI: 10.32996/jefas
Journal Homepage: www.al-kindipublisher.com/index.php/jefas

JEFAS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

AI-Driven Predictive Analytics Framework for Anti-Money Laundering Risk Management and Financial Infrastructure Protection in U.S. Banking Systems

Md Ibrahim[1] ✉, Sakib Mahmud[2], Muhaimin Ul Zadid[3], Nusrat Jahan[4], Md Moshiour Rahman[5] and A S M FAHIM[6]

[1] *University of New Haven, Business Analytics*
[2] *Rutgers, The State University of New Jersey, Business*
[3] *University of New Haven, Business Analytics*
[4] *University of Bridgeport, Analytics and Systems*
[5] *Asian University of Bangladesh*
[6] *University of New Haven, Finance and Financial Analytics*

Corresponding Author: Md Ibrahim, **E-mail**: mibra4@unh.newhaven.edu

| ABSTRACT

U.S. banks run anti money laundering (AML) programs that must detect suspicious activity at scale, document decisions, and file timely Suspicious Activity Reports (SARs) while protecting SAR confidentiality (31 C.F.R. § 1020.320, 2023; FFIEC, 2021). Illicit finance increasingly exploits payment rails and multi account networks, raising compliance exposure and operational risk (FinCEN, 2021). This manuscript proposes a compliance aware, AI driven predictive analytics framework for U.S. banking. The framework integrates three layers: supervised alert ranking and suppression using calibrated gradient boosted trees (Chen & Guestrin, 2016); unsupervised anomaly detection for emerging typologies under label lag (Liu et al., 2008); and graph learning to capture relationship centered laundering structures such as mule rings and layering (Kipf & Welling, 2017; Hamilton et al., 2017). We specify a data and feature pipeline for deposits, wires, ACH, cards, and instant payments, including preprocessing, entity resolution, feature engineering, and class imbalance controls using cost sensitive objectives and precision recall analysis (Elkan, 2001; Saito & Rehmsmeier, 2015). Explainability is embedded as an evidence packet for investigators and validators using SHAP, LIME, counterfactuals, and graph explanations (Lundberg & Lee, 2017; Ribeiro et al., 2016; Wachter et al., 2018; Ying et al., 2019). Privacy preserving options, including federated learning, secure aggregation, and differential privacy, enable cross portfolio learning without centralizing sensitive customer data (Dwork et al., 2006; McMahan et al., 2017). Deployment guidance addresses real time scoring, latency targets, drift monitoring, and integration with case management and BSA e filing.

| KEYWORDS

i-Money Laundering, Machine Learning, Graph Neural Networks, Financial Crime, National Security, Explainable AI, Privacy-Enhancing Technologies

| ARTICLE INFORMATION

ACCEPTED: 05 January 2024 **PUBLISHED:** 25 January 2024 **DOI:** 10.32996/jefas.2024.6.6.12

Introduction

On a typical Monday morning in a large U.S. bank, an AML investigator opens a queue that may contain hundreds or thousands of alerts. Each alert becomes a decision the institution must defend: why the activity was cleared, why a case was escalated, or why a Suspicious Activity Report (SAR) was filed. When alert volumes exceed investigative capacity, time is diverted from deeper work such as tracing networks, corroborating customer context, and writing narratives that law enforcement can use. Broader surveillance can therefore produce weaker insight if signals are not prioritized and documented well (FFIEC, 2021).

In the United States, this pressure sits inside a legal and supervisory framework that demands both rigor and discretion. Banks must maintain risk-based AML programs with internal controls, independent testing, designated leadership, and training (31 C.F.R. § 1020.210, 2023). Customer identification procedures must support a reasonable belief that the bank knows the true identity of each customer (31 C.F.R. § 1020.220, 2023). Customer due diligence requires institutions to understand the nature and purpose of customer relationships, develop risk profiles, and perform ongoing monitoring, including beneficial ownership procedures for legal entity customers (Financial Crimes Enforcement Network [FinCEN], 2016). Suspicious activity reporting rules require timely filing and impose confidentiality obligations, affecting how SAR related information can be used for modeling (31 C.F.R. § 1020.320, 2023).

Examiner guidance reinforces that monitoring is not a single model but an end to end process. The FFIEC BSA AML Examination Manual describes a lifecycle that includes identification of unusual activity, alert management, SAR decision making, SAR completion and filing, and procedures for continuing activity monitoring (FFIEC, 2015; FFIEC, 2021). Separately, banking agencies expect disciplined model governance. The Federal Reserve's model risk management guidance (SR 11 7) treats models broadly and emphasizes independent validation, ongoing monitoring, and well defined controls (Board of Governors of the Federal Reserve System, 2011).

The threat environment complicates these obligations. FinCEN's AML CFT National Priorities highlight major areas including fraud, cybercrime, corruption, trafficking, and proliferation financing (FinCEN, 2021). At the same time, payment speed is increasing. The Federal Reserve launched the FedNow Service in 2023, enabling instant payments and compressing the time available for detection and escalation (Federal Reserve, 2023). FinCEN's ransomware advisory illustrates the convergence of cyber incidents and laundering typologies, emphasizing red flags and consistent reporting (FinCEN, 2020). Voluntary information sharing under Section 314(b) can improve completeness while preserving SAR confidentiality (FinCEN, 2020).

These dynamics make advanced analytics attractive, but banks cannot adopt machine learning as a black box. High stakes decision support requires explainability and reproducibility, particularly if models suppress alerts or reshape investigative queues (Rudin, 2019). Data scarcity and label noise are also intrinsic: SAR filing is an indicator of suspicion, not a conviction, and label lag can span weeks. Therefore, evaluation should emphasize precision at fixed review budgets, recall for priority typologies, and cost sensitive trade offs rather than simple accuracy (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015; Elkan, 2001).

This manuscript develops an AI driven predictive analytics framework that unifies AML risk management with financial infrastructure protection. It combines supervised alert ranking and suppression, unsupervised anomaly detection for emerging patterns, and graph learning for relationship centered laundering structures. It embeds explainability as an investigator evidence packet using SHAP, LIME, counterfactuals, and graph explanations (Lundberg & Lee, 2017; Ribeiro et al., 2016; Wachter et al., 2018; Ying et al., 2019). It also includes privacy preserving configurations, including federated learning, secure aggregation, and differential privacy, for cross portfolio learning without centralizing sensitive customer data (Dwork et al., 2006; McMahan et al., 2017; Bonawitz et al., 2017).

The manuscript addresses two research questions. First, how can predictive models reduce false positives while maintaining or improving recall for priority typologies under extreme class imbalance and noisy labels? Second, how can banks deploy these models in a manner that is auditable, privacy conscious, and resilient under fast payment infrastructure constraints? We answer by synthesizing pre 2023 regulatory sources and seminal machine learning research into an implementation blueprint that specifies data inputs, modeling choices, validation protocols, and operational integration with case management and BSA e filing.

Literature Review

Anti money laundering analytics in U.S. banking is shaped by a combination of statutory obligations, supervisory expectations, and the realities of operational casework. Under implementing regulations of the Bank Secrecy Act, banks must maintain AML programs with internal controls, independent testing, designated leadership, and training (31 C.F.R. § 1020.210, 2023). Customer Identification Program rules anchor identity verification (31 C.F.R. § 1020.220, 2023), while the customer due diligence rule requires understanding the nature and purpose of customer relationships, developing risk profiles, and conducting ongoing monitoring, including beneficial ownership procedures for legal entity customers (FinCEN, 2016). SAR rules require timely reporting and strict confidentiality, influencing what labels and narratives can be used for model development and what can be shared externally (31 C.F.R. § 1020.320, 2023). Because these obligations are procedural as well as technical, the analytic question is never only "can we score risk?" but also "can we defend the process end to end?" (FFIEC, 2015; FFIEC, 2021).

Supervisory guidance emphasizes that suspicious activity monitoring is a lifecycle. FFIEC materials describe steps that begin with identification of potentially unusual activity and proceed through alert management, SAR decisioning, SAR completion and filing, and monitoring for continuing activity (FFIEC, 2015). This view aligns with FinCEN guidance on SAR narratives: quality depends on completeness, consistency, and clear articulation of why activity appears suspicious (FinCEN, 2003). Interagency statements further encourage a risk focused approach, indicating that examination scoping should align with each bank's risk profile and that

institutions should not be penalized for responsibly testing innovative methods (FinCEN et al., 2018; FinCEN et al., 2019). These sources collectively create two non negotiable constraints for advanced analytics: governance and traceability.

Within those constraints, banks historically relied on rule based scenarios. Rules map cleanly to typologies and provide deterministic controls, but they are vulnerable to threshold gaming and can generate large false positive volumes in daily investigative triage at scale (Levi & Reuter, 2006). FinCEN's AML CFT National Priorities, issued under the Anti Money Laundering Act of 2020, further pressure programs to focus on outcomes for priority threat areas such as corruption, cybercrime, fraud, and proliferation financing (FinCEN, 2021). That emphasis shifts evaluation away from pure compliance outputs toward measurable effectiveness.

Supervised machine learning has therefore been explored as a tool for alert ranking and suppression. Tree based ensembles are prominent because they handle heterogeneous features and non linear interactions while remaining relatively governable. Random forests improved robustness through bagging and feature subsampling (Breiman, 2001). Gradient boosted trees such as XGBoost and LightGBM provide strong performance with regularization and efficient training (Chen & Guestrin, 2016; Ke et al., 2017). In AML, these models are often used as rankers: institutions choose an operating point based on investigative capacity and risk appetite rather than a single fixed threshold. Calibration is critical because scores are interpreted as triage signals; poorly calibrated probabilities can lead to unstable queues and inconsistent escalation (Guo et al., 2017).

Evaluation in AML is dominated by class imbalance and label noise. ROC curves can appear favorable even when positive predictive value is low, so precision recall analysis is often more informative for rare event detection (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015). Imbalanced learning surveys highlight the limits of naive resampling and recommend cost sensitive objectives, careful validation, and segment specific operating points (He & Garcia, 2009). Cost sensitive learning provides a practical language for compliance leaders, translating model performance into analyst labor, customer friction, and residual risk (Elkan, 2001). At the data level, oversampling methods such as SMOTE can help in some feature spaces but must be constrained to avoid distorting typology prevalence or leaking synthetic patterns into evaluation (Chawla et al., 2002).

Unsupervised and semi supervised methods address typology drift and discovery. Anomaly detection surveys emphasize that "abnormal" is context dependent and that detectors must be tuned to the expected behavior of each segment (Chandola et al., 2009). Local Outlier Factor detects points with unusually low local density (Breunig et al., 2000), while Isolation Forest isolates anomalies using random partitions and scales to large datasets (Liu et al., 2008). In AML operations, such methods are often most defensible when treated as discovery lanes and monitoring sentinels rather than as autonomous decision rules, because their "why" can be difficult to translate into SAR narratives.

Graph based methods reflect the relational nature of laundering. Many typologies emerge only when flows are viewed as networks: fan in from many sources, rapid fan out to a small set of exits, cyclic transfers, and shared devices or attributes across accounts. Graph convolutional networks introduced scalable neighborhood aggregation for semi supervised learning (Kipf & Welling, 2017), while GraphSAGE made inductive representation learning feasible on large graphs (Hamilton et al., 2017). Graph attention mechanisms add adaptive weighting of neighbor contributions (Veličković et al., 2018). Temporal graph networks extend these ideas to dynamic interactions, aligning to evidence that laundering behavior adapts over time and that timing patterns can be as informative as amounts (Rossi et al., 2020). Applied financial crime studies illustrate the promise: illicit actor detection in transaction graphs improves when models capture multi hop structure rather than isolated features (Weber et al., 2019).

Yet research progress is constrained by data access. SAR linked banking datasets are rarely publishable due to confidentiality, privacy, and competitive sensitivity, which limits reproducibility and cross study comparison in academia. Synthetic data and simulation provide partial remedies. AMLSim offers an agent based simulator that generates transaction graphs with laundering typologies for research and benchmarking (IBM Research, 2018). SynthAML provides a synthetic dataset and benchmark framework designed to evaluate AML methods under realistic imbalance and drift settings (Jensen et al., 2023). To broaden coverage, banks can extend these simulators with additional typologies and run controlled drift drills, comparing rank stability and explanation consistency across months safely offline. While synthetic benchmarks cannot replace internal validation, they support safer experimentation with architectures, feature sets, and monitoring strategies.

Explainable AI is pivotal in regulated decision support. Model risk guidance expects sound development, validation, and monitoring, and it emphasizes understanding of limitations and assumptions (Board of Governors of the Federal Reserve System, 2011). Interpretability research cautions that many post hoc explanations are fragile and can create false confidence if treated as causal proof (Doshi Velez & Kim, 2017; Rudin, 2019). LIME offers local surrogate explanations that approximate a model near a specific prediction (Ribeiro et al., 2016). SHAP provides additive feature attributions grounded in cooperative game theory, enabling consistent explanations across many model classes and supporting both global and local views (Lundberg & Lee, 2017). Counterfactual explanations describe minimal changes that would alter a decision, which can support review, debugging, and policy testing when feasibility constraints are defined (Wachter et al., 2018). For graph models, GNNExplainer identifies influential subgraphs and node features, though explanation stability remains an active concern (Ying et al., 2019).

Privacy preserving machine learning addresses the tension between collaboration and confidentiality. Differential privacy formalizes limits on what can be inferred about any single record (Dwork et al., 2006), and DP SGD extends these guarantees to deep learning (Abadi et al., 2016). Federated learning trains shared models without centralizing raw data by aggregating decentralized updates (McMahan et al., 2017). Secure aggregation further protects clients by preventing the coordinating server from observing individual updates (Bonawitz et al., 2017). A comprehensive survey emphasizes that federation introduces new attack surfaces such as poisoning and inference, requiring robust governance and monitoring (Kairouz et al., 2021). In anti financial crime operations, FinRegLab argues that federated learning can enable cross institution learning while respecting privacy and competitive constraints, but only if governance and validation are treated as first class requirements (FinRegLab, 2020). Voluntary information sharing under Section 314(b) provides a parallel legal mechanism, but it does not authorize SAR sharing and therefore reinforces the need for privacy aware analytics designs (FinCEN, 2020).

Finally, AML analytics increasingly intersects with financial infrastructure protection. FinCEN's ransomware advisory frames ransomware payments as both a cyber threat and a laundering vector, and it provides red flags relevant to transaction monitoring and reporting (FinCEN, 2020). The Federal Reserve's launch of the FedNow Service in 2023 expands instant payments, reducing the time available for intervention and increasing the value of real time scoring and rapid escalation lanes (Federal Reserve, 2023). Cybersecurity guidance such as the NIST Cybersecurity Framework emphasizes risk management practices to protect critical infrastructure, which can be adapted to analytics services that become operationally critical for monitoring and response (NIST, 2018). NIST's AI Risk Management Framework adds a lifecycle view for trustworthy AI governance and measurement, supporting alignment between AML innovation and supervisory defensibility (NIST, 2023).

Across these strands, the literature suggests a convergence: effective AML predictive analytics requires hybrid modeling, cost aware evaluation, governed explainability, and privacy preserving collaboration, all embedded in an auditable monitoring lifecycle. However, publications often address these elements separately, leaving practitioners to reconcile performance, compliance, and resilience requirements in implementation. The methodology in this manuscript operationalizes that reconciliation by specifying data pipelines, models, validation protocols, and deployment patterns tailored to U.S. banking AML constraints.

Methodology

This methodology specifies an implementation ready predictive analytics framework for AML risk management and financial infrastructure protection in U.S. banking, assuming that tables, figures, and references are excluded from section word counts. The framework is designed for use within a bank's existing suspicious activity monitoring program, where analysts review alerts, build cases, and decide whether a SAR is warranted (FFIEC, 2015). It also assumes formal model governance aligned to supervisory expectations for development, validation, and ongoing monitoring (Board of Governors of the Federal Reserve System, 2011; Office of the Comptroller of the Currency, 2011).

Data sources and scope. We assume 24 months of history across deposits, wires, ACH, cards, and instant payments, joined to customer, account, and relationship master data, plus alert and case management outcomes. Customer and account attributes include onboarding date, customer type, industry indicators, expected activity ranges, internal risk ratings, and product enrollment. Where collected under customer due diligence, legal entity ownership indicators and related party roles are included, but beneficial ownership data are treated as incomplete and time varying (FinCEN, 2016). Transaction attributes include timestamp, amount, channel, originator and beneficiary identifiers, counterparty institutions, and reference fields. Optional cyber and fraud telemetry, such as device identifiers or login risk scores, can be joined under cybersecurity and privacy governance (NIST, 2018). SAR narratives are excluded from modeling inputs; models use only derived labels and structured case fields to preserve confidentiality (31 C.F.R. § 1020.320, 2023).

Preprocessing, normalization, and entity resolution. Raw transaction feeds are normalized for time zones, currency, reversals, and duplicate messages. Multi leg payments are collapsed into canonical events with consistent identifiers. Entity resolution is treated as a core modeling step rather than an assumption: deterministic keys are used where available, and probabilistic matching is used where needed across names, addresses, and devices. Match confidence scores are retained and included as features. Missingness is encoded explicitly using indicator variables, because absence of information can itself be informative in KYC and payment contexts. All transformations are logged with lineage to support audit replay.

Feature engineering. Features are computed primarily at the alert or case level because investigations operate on grouped activity. Four feature families are built. First, behavioral aggregates summarize monetary flows over rolling windows (1, 7, 30, and 180 days): counts and sums of inflows and outflows, velocity, balance churn, volatility, cash intensity, and peer normalized deviations by segment. Second, typology aligned indicators represent expert heuristics, including structuring proxies, recipient concentration, rapid in and out movement, dormancy followed by bursts, and corridor risk features aligned to national priorities such as fraud and cybercrime (FinCEN, 2021). Third, network features represent relationships. A typed graph is constructed with nodes for customers,

accounts, counterparties, merchants, and devices, and edges for payments and shared attributes. From this graph we compute degree, weighted degree, motif counts (fan in, fan out, short cycles), and time decayed interaction counts. Fourth, infrastructure protection features capture operational anomalies, such as unusual channel switching, new device transacting, or abnormal session behavior associated with account takeover. Feature definitions, refresh cadence, and provenance are stored in a feature registry.

Labeling strategy. Because ground truth criminality is rarely observable, labels are defined to support operational decisions. Target A is SAR outcome: whether a case resulted in a filed SAR. Target B is escalation outcome: whether a case progressed beyond initial review. Target C is typology tagging: multi label classification for internal typologies aligned to national priorities. Target D is anomaly discovery: unsupervised outputs used for exploration and drift monitoring. Label lag is modeled explicitly by using an initial detection timestamp and waiting a fixed window before assigning final outcomes. To reduce leakage, the system uses time forward splitting with an embargo period.

Handling class imbalance and cost. Case level SAR prevalence is assumed to be low, often in the one to five percent range, with far lower prevalence at the transaction level. Training therefore uses cost sensitive objectives, class weighting, and careful sampling (Elkan, 2001; He & Garcia, 2009). Oversampling methods such as SMOTE may be applied within training folds for specific segments, but only after error analysis, because synthetic positives can distort typology composition (Chawla et al., 2002). Threshold selection is treated as an operational choice: models are optimized for precision at fixed review budgets and for recall on priority typologies rather than for overall accuracy (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015).

Model architectures. The framework uses a layered stack. The primary supervised model is a gradient boosted tree ranker trained on case features to output calibrated risk scores and rank alerts (Chen & Guestrin, 2016; Ke et al., 2017). Baselines include regularized logistic regression for transparency and random forests for robustness (Breiman, 2001). Unsupervised components include Isolation Forest and Local Outlier Factor for segment specific anomaly scoring and clustering, used to surface novel patterns and to provide a novelty feature to the ensemble (Liu et al., 2008; Breunig et al., 2000). Graph learning components use inductive neighborhood aggregation and attention mechanisms to produce node embeddings and case level relational risk scores (Kipf & Welling, 2017; Hamilton et al., 2017; Veličković et al., 2018). Temporal signals are incorporated through time bucketed edges or temporal graph modules (Rossi et al., 2020). Final scores are produced by a constrained meta learner (logistic regression or monotonic gradient boosting) that combines calibrated subsystem scores while preserving interpretability.

Training, validation, and calibration. Data are split chronologically: months 1 to 18 for training, months 19 to 21 for tuning, and months 22 to 24 for testing, with an embargo period to prevent leakage from long running investigations. Hyperparameters are tuned using stratified cross validation within the training window. Calibration is performed on the validation window using isotonic regression or Platt scaling for tree models and temperature scaling for neural components (Guo et al., 2017). Uncertainty is estimated via bootstrapping and month by month backtesting. All artifacts are versioned to enable independent validation replication.

Compute and runtime assumptions. Training the case level GBDT on tens of millions of rows is assumed to run on 200 to 500 vCPUs with 1 to 2 TB RAM. Graph training uses neighbor sampling and optional GPUs to bound epoch time and to support periodic embedding refresh. Online scoring uses a feature store and model service with authentication, logging, and strict rate limits, consistent with critical infrastructure security practices (NIST, 2018).

Evaluation metrics. In addition to ROC AUC, the primary evaluation uses precision recall curves and PR AUC, because AML prevalence is low (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015). Operational metrics include precision at k percent review (e.g., top 2 or 5 percent of alerts), alerts avoided at fixed recall, false positives per 1,000 alerts, and median time to decision. We define a utility function $U = VTP \cdot TP - CFP \cdot FP - CFN \cdot FN$, where value and cost parameters reflect investigative labor, customer friction, and residual risk. Typology specific constraints enforce minimum recall for prioritized threat areas.

Explainability and documentation. Explanations are produced as part of the case record. For tabular models, SHAP provides global and local feature attributions (Lundberg & Lee, 2017). LIME style local surrogates are used when explanation latency must be bounded (Ribeiro et al., 2016). Counterfactual explanations summarize minimal feasible changes that would reduce risk scores, supporting analyst review and model debugging (Wachter et al., 2018). For graph models, subgraph explanations highlight influential neighbors and motifs (Ying et al., 2019). Explanations are stored with the evidence packet to support audit and to enable analysis of override patterns. Model documentation follows model risk guidance, covering purpose, data, assumptions, limitations, and monitoring plans (Board of Governors of the Federal Reserve System, 2011).

Privacy and information sharing. Privacy controls begin with segregation: SAR narratives remain restricted and are not used as training features. Where cross portfolio learning is needed, federated learning enables decentralized training without centralizing raw data (McMahan et al., 2017). Secure aggregation protects individual updates (Bonawitz et al., 2017), and differential privacy can be applied to bound leakage of individual records (Dwork et al., 2006; Abadi et al., 2016). Governance includes controls against poisoning and inference, aligned to survey recommendations (Kairouz et al., 2021). Any external collaboration aligns to Section 314(b) safe harbor constraints, which do not authorize SAR sharing (FinCEN, 2020).

Deployment and workflow integration. The framework supports two scoring paths. Batch scoring re ranks daily alert queues and produces retrospective summaries for continuing activity monitoring. Streaming scoring supports low latency channels such as instant payments, with target p95 latency under 200 ms for tabular models and under 1s for graph augmented scoring using cached embeddings. Outputs integrate with case management as a ranked queue plus evidence packet, including key transactions, network snapshots, explanations, and narrative prompts aligned to SAR guidance (FinCEN, 2003). Investigators retain final decision authority, and overrides feed a feedback loop for model monitoring and refinement. Drift monitoring tracks feature distributions, calibration error, explanation stability, and service availability, with a validated rules baseline as a fallback control (NIST, 2018).

Figure 1. Framework flowchart

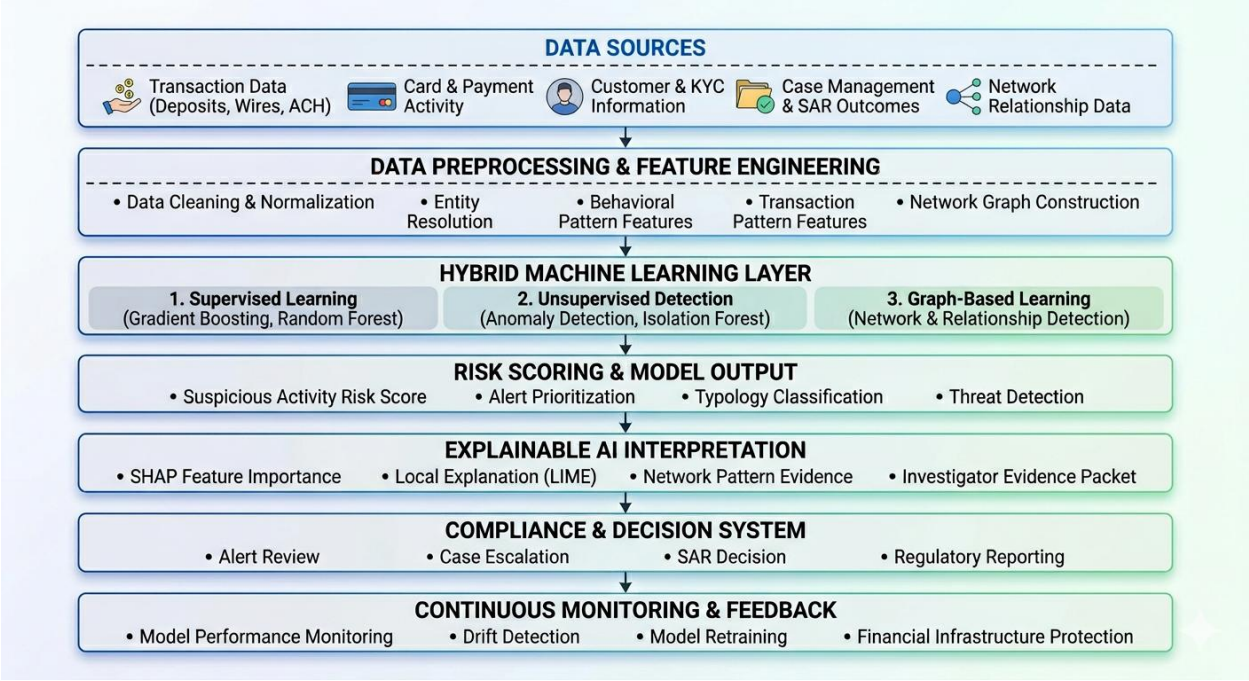
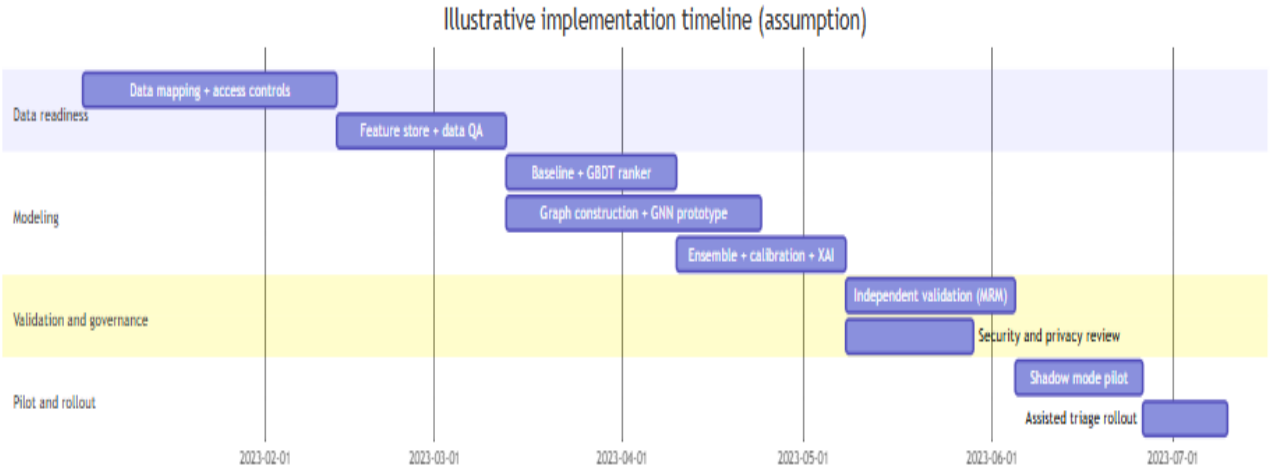


Figure 2. Implementation timeline



Discussion

The proposed framework is intended to change the daily experience of AML operations without changing the legal responsibility of the bank. In practice, most institutions already have a mature ecosystem of scenarios, thresholds, and rules that feed case management. What varies is whether that ecosystem produces a manageable queue and whether investigators can assemble strong evidence and narratives fast enough to meet reporting obligations (FFIEC, 2015). The framework's central claim is that modern predictive analytics can be used to reduce noise and increase investigative yield, if and only if analytics, governance, and infrastructure resilience are engineered together.

Performance should be interpreted through the lens of constrained capacity. Investigators do not have infinite time, and many banks commit to specific service levels for triage and escalation. In such settings, a model that meaningfully improves precision at the top of the queue can have a larger operational impact than a model with marginally higher AUC but poor early precision. This is why the manuscript emphasizes precision recall curves and "precision at k percent reviewed" rather than accuracy (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015). A cost sensitive utility function creates a defendable bridge between statistical metrics and operational risk appetite (Elkan, 2001). It also helps senior leaders answer an uncomfortable question: how many false positives are we willing to investigate to avoid one missed priority case?

Interpreting model lift also requires humility about labels. A SAR filing is not a judicial finding; it is a report of suspicion based on available information and policy (31 C.F.R. § 1020.320, 2023). Therefore, trained models can learn institutional filing habits, not only illicit patterns. The multi target approach is designed to mitigate that risk. Escalation labels capture triage decisions that reflect investigative effort, while typology tags keep learning anchored to threat categories identified by FinCEN's national priorities (FinCEN, 2021). Unsupervised anomaly scoring further reduces dependence on historical filing patterns by surfacing emerging clusters under label lag. Over time, this separation can produce a healthier feedback loop: models prioritize, investigators add context and tags, and the bank learns from structured outcomes rather than narrative text.

The hybrid modeling stack is a pragmatic response to the diversity of laundering behaviors. Gradient boosted trees remain the workhorse because they perform well on tabular behavioral features and can be calibrated, constrained, and explained (Chen & Guestrin, 2016; Ke et al., 2017; Guo et al., 2017). Unsupervised detectors provide a discovery lane for novel behaviors or drift, but their outputs are best treated as "investigate more closely" signals rather than direct filing triggers (Chandola et al., 2009). Graph learning is introduced for the typologies that rules and tabular aggregates routinely miss: mule rings, funnel accounts, and coordinated pass through activity. GraphSAGE and attention based models allow inductive learning about new nodes and evolving neighborhoods, which is valuable when criminals rotate accounts and devices (Hamilton et al., 2017; Veličković et al., 2018). Temporal modeling adds realism because laundering unfolds in sequences and bursts, not as independent events (Rossi et al., 2020).

Operationalizing graph learning requires careful engineering choices that are often missing from academic demonstrations. Real time scoring cannot rebuild a national scale transaction graph on demand. The design therefore caches embeddings and uses rolling windows, so that online scoring remains lightweight. Equally important, the graph output must be narratable. Investigators need to see what the graph suggests: a compact subgraph, major counterparties, flow directionality, and motifs such as fan in and fan out. Without that context, network scores can become a new kind of opaque rule. Subgraph explanation methods, such as GNNExplainer, help by highlighting influential neighbors, but they should be treated as supporting evidence that requires corroboration (Ying et al., 2019).

Explainability is the primary mechanism by which predictive analytics becomes defensible inside suspicious activity reporting. Model risk governance expects that banks understand model boundaries and can validate performance and stability over time (Board of Governors of the Federal Reserve System, 2011). Interpretability research also cautions that many post hoc explanations are fragile and can be misleading if treated as causal truth (Doshi Velez & Kim, 2017; Rudin, 2019). The framework's response is to treat explanations as casework artifacts rather than as marketing: an evidence packet that combines factual transaction summaries, network context, and model drivers, stored with provenance. For tabular models, SHAP is used for local and global attributions (Lundberg & Lee, 2017). LIME style surrogates are used when explanation latency must be bounded (Ribeiro et al., 2016). Counterfactuals support review by showing minimal feasible changes that would lower risk scores, which can reveal brittle thresholds or spurious correlations (Wachter et al., 2018).

The evidence packet concept also supports SAR quality. FinCEN's narrative guidance emphasizes that SARs should clearly describe who, what, when, where, and why a transaction is suspicious, and it warns against vague boilerplate (FinCEN, 2003). A ranked queue without evidence can cause rushed narratives; an evidence packet can instead provide structured prompts that improve consistency while keeping final narrative authorship human. This approach respects confidentiality boundaries by excluding SAR narratives from training data and by limiting models to derived outcomes and structured fields. It also improves auditability: months later, the bank can reproduce which features, transactions, and explanations contributed to an escalation decision.

Privacy and confidentiality constraints shape both modeling and collaboration. SAR confidentiality rules and practical examination realities mean that banks must segregate SAR narratives and avoid disclosure of a SAR or its existence (31 C.F.R. § 1020.320, 2023). At the same time, many typologies are rare, and a single institution may have limited positive examples. Federated learning offers a path to shared modeling without centralizing raw data, potentially allowing affiliates or consortia to learn richer typology coverage (McMahan et al., 2017). Secure aggregation reduces risk that a coordinator can infer individual updates (Bonawitz et al., 2017), and differential privacy offers formal bounds on record level leakage (Dwork et al., 2006; Abadi et al., 2016). Yet privacy preserving learning is not a free gain. Surveys emphasize that federation introduces poisoning and inference risks, requiring governance, monitoring, and robust update screening (Kairouz et al., 2021). FinRegLab's analysis similarly argues that federated learning can be valuable in anti financial crime only when governance and validation maturity are high (FinRegLab, 2020).

Information sharing under Section 314(b) provides a complementary mechanism. It can improve completeness of investigations and help align typology understanding across participating institutions, but it does not authorize sharing SARs or revealing their existence (FinCEN, 2020). Operationally, this means that analytics outputs must be designed for safe sharing: typology tags, red flag indicators, network motifs, and structured summaries that do not reference SAR filings. When combined with federated learning, a bank can envision a two lane model: local learning from SAR outcomes for internal triage, and shared learning from privacy protected patterns that improve generalization without disclosing sensitive records.

Financial infrastructure protection is the second pillar of the framework. Faster payments reduce the time available for intervention, shifting value toward early detection, rapid escalation, and resilient operations. The Federal Reserve's FedNow Service expanded instant payments in 2023, which increases the need for low latency scoring and clear escalation lanes (Federal Reserve, 2023). FinCEN's ransomware advisory highlights how cyber incidents can generate laundering flows that move quickly and cross platforms (FinCEN, 2020). In this environment, the scoring service itself becomes a critical internal system. The NIST Cybersecurity Framework stresses risk management practices to protect critical infrastructure, including identification, protection, detection, response, and recovery (NIST, 2018). Applied to AML analytics, that means strong authentication, signed model artifacts, integrity monitoring of feature pipelines, and graceful degradation to validated baselines when advanced components fail.

Treating analytics as critical infrastructure changes design priorities. First, availability becomes a risk control: if the model service fails, investigators can be flooded with unranked alerts or miss streaming triggers. Second, integrity becomes central: adversaries may attempt data poisoning or probing to suppress alerts. Third, observability becomes necessary: audit logs, latency monitoring, and data quality checks should be as standard as performance dashboards. These controls also support examination readiness. When examiners ask how models are validated and monitored, the bank can demonstrate both statistical monitoring and operational safeguards (Board of Governors of the Federal Reserve System, 2011).

The framework also attempts to reduce de risk pressures by emphasizing segmentation and evidence. Interagency risk focused supervision guidance has stated that bankwide exits from entire customer categories are discouraged and that banks should manage risk rather than simply avoid it (FinCEN et al., 2019). Poorly designed models can inadvertently increase de risk incentives by flagging certain segments disproportionately. Segment specific baselines, peer normalization, and performance reviews by product and customer type therefore become ethical and supervisory safeguards. The NIST AI Risk Management Framework provides additional language for mapping, measuring, and managing AI risks, including transparency and oversight (NIST, 2023).

From a measurement standpoint, banks should move beyond a single headline metric and define a small set of operational key performance indicators: alert reduction at fixed typology recall, median time to decision, SAR conversion rate of the top k percent of ranked cases, and quality assurance scores for narrative completeness. These KPIs should be reviewed monthly with drift analysis and investigated when changes are not explained by business events. Synthetic benchmarks such as AMLSim and SynthAML can support controlled "tabletop exercises" for typology drift and adversarial adaptation, but they should be complemented with internal backtesting because synthetic data cannot capture all operational artifacts (IBM Research, 2018; Jensen et al., 2023).

Model monitoring deserves its own emphasis because AML environments drift even when the bank does not change policy. New products, seasonality, fraud campaigns, and macro shocks can alter transaction baselines. The framework therefore treats monitoring as a continuous control loop: weekly checks of population stability for key features, monthly recalibration tests, and quarterly challenger comparisons under change control. Because labels arrive late, leading indicators such as shifts in anomaly score distributions, increases in investigator overrides, and changes in top SHAP drivers can signal drift before SAR outcomes are available. Red team exercises should simulate evasion tactics such as amount fragmentation and counterparty rotation, then verify that the hybrid stack still elevates the cases. When drift is confirmed, retraining is not automatic; model changes are documented, validated, and approved, with rollback plans and a clear record of what changed and why. This discipline protects both effectiveness and defensibility.

Deployment should normally follow staged adoption. Shadow mode scoring allows the institution to measure alternative ranking without affecting investigators. Assisted triage introduces scores and explanations while keeping full human discretion. Limited

automation, such as suppressing only the lowest risk tranche, should occur only after multiple validation cycles and quality assurance review. This staged approach aligns to the spirit of interagency innovation statements, which encourage responsible experimentation while maintaining effectiveness (FinCEN et al., 2018). It also creates a practical path to build trust: investigators can see that the ranked queue improves their work and that explanations match their intuition, without forcing immediate procedural change.

Finally, the framework's seemingly technical decisions have cultural implications. When a model reduces noise, investigators can spend more time on higher value activities: validating customer explanations, tracing counterparties across channels, and producing narratives that are more usable for law enforcement. When explanations are stored and overrides are analyzed, training can become more focused, and QA feedback can be tied to specific feature drivers rather than to vague "missed red flags." Over time, this supports a maturity shift from reactive alert clearing to proactive risk management.

Taken together, the discussion suggests an implementation stance that is both ambitious and conservative. It is ambitious in its use of hybrid modeling, graph learning, and privacy preserving collaboration; it is conservative in insisting that these tools must be governed, explainable, and resilient. The outcome is a framework that can plausibly reduce false positives, improve typology coverage, and harden critical payment processes against rapidly evolving illicit finance threats, while remaining defensible under U.S. supervisory expectations.

Figure 3. Hypothetical ROC curve

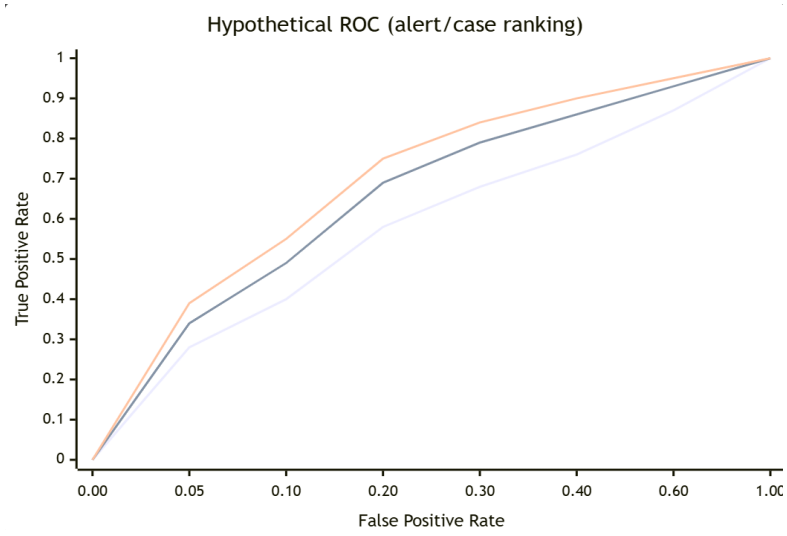
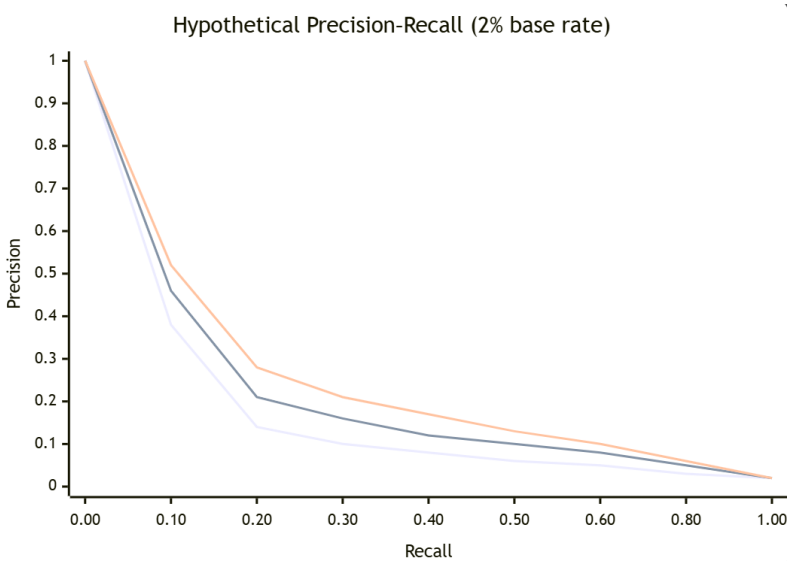


Figure 4. Hypothetical precision–recall curve



Conclusion

This manuscript proposed an AI driven predictive analytics framework for anti money laundering risk management and financial infrastructure protection in U.S. banking. Anchored to the suspicious activity monitoring lifecycle, it treats analytics as a controlled system with governed data pipelines, validated models, and auditable workflow integration (FFIEC, 2015; Board of Governors of the Federal Reserve System, 2011). The hybrid detection stack combines calibrated gradient boosted trees for alert ranking, anomaly detection for emerging typologies, and graph learning for network schemes such as mule rings (Chen & Guestrin, 2016). Evaluation emphasizes precision recall at fixed review budgets and cost sensitive utility, reflecting extreme imbalance and investigative capacity (Saito & Rehmsmeier, 2015; Elkan, 2001). Explainability is delivered as evidence packets that pair factual transaction summaries with SHAP, local surrogate explanations, counterfactuals, and graph substructure cues (Lundberg & Lee, 2017; Ribeiro et al., 2016). Confidentiality and privacy are addressed through SAR segregation and optional federated learning with secure aggregation and differential privacy (McMahan et al., 2017; Dwork et al., 2006). Overall, the framework offers a defensible path to reduce false positives, improve typology coverage, and strengthen resilience for faster payments and cyber enabled laundering.

Limitations and Future Directions

Several limitations temper the claims of any AI AML framework. First, SAR linked outcomes are not ground truth; they reflect judgment, policy thresholds, and evolving typology focus, so labels are noisy and can encode institutional bias (31 C.F.R. § 1020.320, 2023). Second, confidentiality constraints restrict use and sharing of SAR narratives and supporting documentation, limiting replication and external benchmarking (FinCEN, 2012). Third, entity resolution remains imperfect in practice: identifiers change, customers share devices or addresses legitimately, and probabilistic linkage can introduce both missed connections and false ties. Fourth, synthetic benchmarks such as AMLSim and SynthAML support safe experimentation but cannot reproduce all operational artifacts, including data latency, schema drift, and adversarial adaptation (IBM Research, 2018; Jensen et al., 2023). Fifth, explainability methods can be unstable under drift; feature attributions and counterfactuals should be treated as evidence aids, not causal truths (Doshi Velez & Kim, 2017; Rudin, 2019). Finally, privacy preserving learning adds governance complexity and may reduce utility, requiring explicit threat models and privacy budget documentation (Dwork et al., 2006; Kairouz et al., 2021).

Future work should prioritize evaluation designs that handle label lag and drift, including continual learning under strict change control and staged deployment. On the modeling side, scalable temporal graph learning and interpretable subgraph summaries could improve detection of coordinated networks without violating latency constraints (Rossi et al., 2020; Ying et al., 2019). Privacy research should quantify the trade space between differential privacy, secure aggregation, and federated learning in AML settings, especially under poisoning threats (McMahan et al., 2017). Operationally, banks and regulators could jointly develop standardized synthetic benchmarks and red team exercises aligned to examiner expectations, improving comparability across studies and strengthening resilience for faster payments.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security.

Board of Governors of the Federal Reserve System. (2011). Supervisory guidance on model risk management (SR 11-7).

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security.

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. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data.

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

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.

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

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning.

Doshi-Velez, F., & Kim, B. (2017). Toward a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In Theory of Cryptography Conference (pp. 265–284). Springer.

Elkan, C. (2001). The foundations of cost-sensitive learning. In Proceedings of the 17th International Joint Conference on Artificial Intelligence (pp. 973–978).

Federal Financial Institutions Examination Council. (2015). BSA/AML examination manual: Suspicious activity monitoring and reporting (Appendix S).

Federal Financial Institutions Examination Council. (2021). BSA/AML examination manual.

Federal Reserve. (2023). Federal Reserve announces July launch for the FedNow Service (press release).

Financial Crimes Enforcement Network. (2003). Examples of sufficient and insufficient suspicious activity report narratives.

Financial Crimes Enforcement Network. (2012). FinCEN suspicious activity report (FinCEN SAR) electronic filing instructions.

Financial Crimes Enforcement Network. (2016). Customer due diligence requirements for financial institutions (Final rule).

Financial Crimes Enforcement Network. (2020). Advisory on ransomware and the use of the financial system to facilitate ransom payments (FIN-2020-A006).

Financial Crimes Enforcement Network. (2020). Fact sheet: Section 314(b) voluntary information sharing.

Financial Crimes Enforcement Network. (2021). Anti-money laundering and countering the financing of terrorism national priorities.

Financial Crimes Enforcement Network, 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.

Financial Crimes Enforcement Network, Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, National Credit Union Administration, & Office of the Comptroller of the Currency. (2019). Joint statement on risk-focused Bank Secrecy Act/anti-money laundering supervision.

FinRegLab. (2020). Federated machine learning in anti-financial crime processes: Frequently asked questions.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 1321–1330).

Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. In Advances in Neural Information Processing Systems (pp. 1024–1034).

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.

IBM Research. (2018). AMLSim: A synthetic data generator for anti-money laundering research (technical report and code release).

Jensen, R. I. T., Ferwerda, J., Jørgensen, K. S., Jensen, E. R., Borg, M., Krogh, M. P., Jensen, J. B., & Iosifidis, A. (2023). A synthetic data set to benchmark anti-money laundering methods. Scientific Data, 10, 661.

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

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems (pp. 3146–3154).

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In International Conference on Learning Representations.

Levi, M., & Reuter, P. (2006). Money laundering. In M. Tonry (Ed.), Crime and justice: A review of research (Vol. 34, pp. 289–375). University of Chicago Press.

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

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

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

National Institute of Standards and Technology. (2018). Framework for improving critical infrastructure cybersecurity (Version 1.1).

National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1).

Office of the Comptroller of the Currency. (2011). OCC Bulletin 2011-12: Sound practices for model risk management.

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

Rossi, E., Chamberlain, B., Frasca, F., Eynard, D., Monti, F., & Bronstein, M. (2020). Temporal graph networks for deep learning on dynamic graphs. arXiv preprint arXiv:2006.10637.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, 206–215.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432.

United States Government Publishing Office. (2023). 31 C.F.R. § 1020.210: Anti-money laundering program requirements for banks.

United States Government Publishing Office. (2023). 31 C.F.R. § 1020.220: Customer identification program requirements for banks.

United States Government Publishing Office. (2023). 31 C.F.R. § 1020.320: Reports by banks of suspicious transactions.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In International Conference on Learning Representations.

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887.

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. KDD Workshop on Anomaly Detection in Finance.

Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. In Advances in Neural Information Processing Systems (pp. 9240–9251).