World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Next-generation Fraud Detection in U.S Financial Systems: Evaluating Hybrid AI and Rule-based Models for Real-time Threat Mitigation

Victor Aworetan *

Office of Network Security, Palo Alto Networks Inc, Texas USA.

World Journal of Advanced Research and Reviews, 2024, 23(03), 3317-3333

Publication history: Received on 18 August 2024; revised on 18 September 2024; accepted on 28 September 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.23.3.3014

Abstract

This paper explores next-generation fraud detection within American financial systems by assessing hybrid architectures that integrate rule-based logic and the artificial intelligence (AI) models to reduce threats in real-time. Inspired by rising loss rates and the rising complexity of fraud vectors (such as synthetic identities and deepfakes), the analysis builds an assessment framework conceptual framework that trades off detection performance, cost of operation, regulatory disclosure, and resilience to adversaries. Based on the current empirical literature, guidance by regulators, and industry-specific reports, the paper presents (1) a synthesis of strengths and weaknesses of hybrid approaches; (2) suggests quantifiable evaluation criteria and economic tradeoffs to institutions; and (3) a tested research agenda (data needs, validation, human-in-the-loop oversight, and policy recommendations). The essence of the argument is that properly designed hybrid frameworks, when structured on the back of robust model risk management and a sustained adversarial testing regime, can significantly decrease the losses to fraud and comply with U.S. regulatory requirements and retain customer confidence.

Keywords: Fraud detection; US financial systems; Financial cybersecurity; Hybrid AI and rule-based models; Real-time threat mitigation

1. Introduction

Financial fraud has been dynamic and long-standing. In the past, institutions would use manual controls and human expertise -such as if X and Y then alert -to prevent apparent abuses, and progressively added statistical and machine-learning methods as data and computing became accessible (Ngai et al., foundational review). Intuitive and auditable yet brittle rule systems are opposed to the scale pattern-recognition-friendly AI systems, whose features of adaptability and explainability increase the risk of explainability and model-risk which regulators have explicitly noted in recent years. The combination of these forces, such as large volumes of transactions, limited data sharing among firms, the development of fraudster behaviors, and an active regulatory agenda on AI is a pressing necessity to consider hybrid options, i.e., the deliberate combination of both paradigms.

The stakes are highlighted by the recent industry and academic evidence. Identity and payment fraud are still significant issues throughout the world and in the U.S.; industry research and surveys indicate that there are significant annual losses and that attack vectors (identity scams, account takeover, deepfakes) are increasingly sophisticated to the point of overwhelming traditional detection pipelines. Concurrently, the technical literature identifies encouraging directions, including federated learning to ensure privacy-constrained cross-institution modeling, hybrid human-in-the-loop designs to overcome label sparsity, and ensemble designs to alleviate false alarms. However, adversarial machine-learning research cautions that fraud detection models are vulnerable to distinct attack surfaces; a malicious attacker can intentionally query and alter models with disastrous consequences to the business unless defenses are engineered.

* Corresponding author: Victor Aworetan

1.2 Statement of the problem

Despite advances, several persistent gaps hinder effective, real-time fraud mitigation in U.S. financial systems:

- Detection tradeoffs: Pure rule systems produce explainable decisions but generate many false positives and struggle with novel fraud patterns; pure AI systems detect subtle patterns but can be opaque and vulnerable to adversarial exploitation. The problem: how to combine strengths while mitigating weaknesses.
- Data and privacy limits: Banks and fintechs hold siloed, proprietary transaction datasets. Cross-institutional model training is limited by privacy and regulation, reducing the ability to learn rare but high-impact fraud patterns.
- Regulatory and governance friction: Supervisory bodies have signaled concern about AI opacity, model risk, and third-party vendor concentration; institutions must demonstrate auditability and explainability while maintaining efficacy.
- Adversarial threat model: Fraud detection operates in an adversarial environment where attackers adapt; many ML defenses developed for image/NLP domains do not translate cleanly to tabular, transaction-based settings.

These problems collectively indicate that incremental upgrades to existing systems are insufficient; a systematic evaluation of hybrid architectures—assessing both technical performance and real-world constraints—is required.

1.1. Objectives of the study

Primary objective: Evaluate hybrid AI + rule-based models for real-time fraud detection in U.S. financial systems, through a multi-dimensional framework combining detection metrics, economic costs, governance needs, and adversarial robustness.

1.1.1. Secondary objectives

- Map the contemporary fraud threat landscape and identify which fraud classes benefit most from hybrid treatment (e.g., synthetic identity vs. card-not-present). Propose an evaluation protocol (datasets, metrics, stress tests) for institutions and researchers, including privacy-preserving data sharing options (e.g., federated learning).
- Identify regulatory and ethical safeguards (explainability, bias audits, SAR integration) necessary for deployment in the U.S. context.
- Offer an operational roadmap: integration patterns with legacy systems, human-analyst workflows, and continuous learning strategies to manage model drift and adversarial probes.

1.2. Relevant research questions

Each research question is framed to be specific and empirically testable:

- **RQ1:** Do hybrid systems (explicit rules layer + ML/AI scoring layer) reduce overall economic loss from fraud relative to purely rule-based or purely AI systems in a realistic operational setting?
- **RQ2:** How do hybrid systems affect false positive and false negative tradeoffs in *real-time* monitoring (milliseconds-to-seconds decision windows)?
- **RQ3:** What governance and explainability mechanisms are necessary for hybrid models to meet U.S. supervisory expectations (OCC, Fed, FinCEN) while preserving detection performance?
- **RQ4:** How resilient are hybrid systems to adversarial manipulation specific to transaction/tabular data (e.g., evasion, poisoning), and what mitigation strategies (red teaming, adversarial training) are most effective?
- **RQ5:** Can privacy-preserving collaboration methods—like federated learning or secure multiparty computation—meaningfully improve detection of cross-institution fraud (e.g., mule networks) without violating privacy/regulatory constraints?

These are precise: they specify the object of study (hybrid systems), the comparator (rule-only; AI-only), the environment (real-time monitoring; U.S. regulatory context), and the outcome domains (economic loss, error tradeoffs, governance, adversarial resilience, cross-institution detection).

1.3. Research hypotheses

- **H1 (for RQ1):** Hybrid systems will achieve a statistically significant reduction in expected fraud loss (measured as total dollars lost after accounting for operational costs and false positives) compared with rule-only systems and with AI-only systems in matched deployment scenarios. Rationale: rules catch known patterns with low false negatives; AI captures subtle patterns and adapts—together they complement each other.
- **H2 (for RQ2):** Hybrid systems will lower the false positive rate at fixed recall when compared to rule-only systems by using AI to re-score and contextualize rule triggers, thereby reducing unnecessary manual reviews. Conversely, at extremely tight latency constraints (<200 ms), the recall gain may be attenuated by compute and integration overheads.
- **H3 (for RQ3):** Hybrid systems that expose rule-level logic and interpretable surrogate explanations (e.g., SHAP, decision rules) will be more likely to meet supervisory expectations and reduce compliance overhead than black-box AI alone—provided model governance and documentation practices align with OCC and banking agency model risk principles.
- **H4 (for RQ4):** Hybrid systems are vulnerable to adversarial evasion but are more robust than standalone AI when rule layers enforce invariant constraints (hard checks) on tamperable inputs; however, attackers can adapt to bypass fixed rules, so continuous adversarial testing is required.
- **H5 (for RQ5):** Federated and privacy-preserving learning frameworks can increase detection of cross-institution fraud patterns (e.g., coordinated mule accounts) without materially degrading privacy or violating current regulatory expectations—if combined with differential privacy and secure aggregation**.**

Each hypothesis is framed to be empirically testable (requires defined datasets, deployment scenarios, and metrics).

1.4. Significance of the study

This study is significant for three audiences:

- Practitioners (banks, fintechs, payments networks): it supplies an operational evaluation framework and a practical roadmap to deploy hybrid architectures with governance guardrails—helping prioritize investments that reduce fraud losses and manual review burdens.
- Regulators and policymakers: the analysis clarifies where current supervisory guidance (model risk, SAR filing, AML) intersects with hybrid deployment needs, suggesting concrete regulatory sandbox designs and standards for explainability and red-team testing.
- Researchers: the paper consolidates open research problems—adversarial defenses for tabular fraud models, privacy-preserving cross-institution training, and human-in-the-loop feedback mechanisms—that can be pursued with public benchmarks and synthetic data methods.

1.5. Scope of the study

This paper focuses on retail and payments-oriented fraud detection in the U.S. financial ecosystem, including card-present, card-not-present, account takeover, synthetic identity, and P2P payment platform fraud. It emphasizes real-time transaction monitoring architectures suitable for banks, card networks, and fintechs. The work does not deeply address capital-markets fraud (insider trading) or non-transactional financial statement fraud, except where the underlying methods (e.g., adversarial robustness, explainability) are broadly relevant. Methodologically, the study synthesizes literature, regulatory guidance, and reproducible technical approaches (benchmarks, protocol design), and proposes testable evaluation strategies rather than reporting a single experimental deployment across multiple banks (which would require data sharing agreements beyond this study's scope).

1.6. Definition of terms

To avoid ambiguity, key terms used throughout the paper are defined as follows:

- Hybrid model (in fraud detection): an architecture that combines *explicit rule logic* (hard rules, expert heuristics) with *data-driven AI/ML components* (supervised classifiers, anomaly detectors, deep models) so that each component complements the other's strengths.
- Rule-based system: deterministic logic implemented as business rules or heuristics (e.g., "block transaction if country ≠ billing country and amount > $X and velocity > Y"), valued for interpretability and immediate control.

- AI/ML system: data-driven models trained on historical labeled or weakly labeled transactions to predict fraud risk, including supervised, unsupervised, and deep learning approaches.
- Real-time detection: decisioning that occurs within operational latencies appropriate to the use case—typically milliseconds to low seconds for card and online payment approvals.
- Adversarial attack (fraud ML): deliberate manipulation of inputs or model training data by malicious actors to evade detection (evasion attacks), contaminate models (poisoning), or extract models (model theft), tailored here to the tabular/transaction domain.
- Federated learning (FL): a collaborative training paradigm that allows multiple parties (e.g., banks) to jointly train a shared model without exchanging raw data, often with secure aggregation and differentially private updates to protect privacy.

2. Literature review

2.1. Preamble

Fraud within the U.S. financial systems has become a multi-dimensional issue, supported by digital banking, peer-to-peer (P2P) payment systems and fintech innovations. International estimates indicate that the losses amount to almost 5 trillion a year, with the U.S. contributing a large portion owing to its vast network of digital transaction (ACFE, 2024). Older systems such as fixed rules and human reviews have latent worth in terms of transparency, but are weak against modern equivalent systems using automation, synthetic identities, and well-organized mule networks.

Artificial intelligence (AI) and machine learning (ML) can provide new, adaptive and data-driven solutions, but they also present new issues: opacitiness, fairness, regulation, and susceptibility to adversarial manipulation. There has been a practical compromise between deterministic rule-based control and adaptive ML and human supervision, called hybrid models (Wahid and Hassini, 2024). The literature is still fragmented and critical knowledge gaps do exist in spite of this promise. The review summarizes the current literature, critiques current methodologies and establishes opportunities in developing hybrid models in the United States.

2.2. Theoretical Review

2.2.1. Statistical and Anomaly Detection Foundations

Fraud detection has long been conceptualized as an anomaly detection problem in skewed data environments. Bolton and Hand (2002) established the statistical basis, framing fraud as a signal detection challenge under extreme class imbalance. This remains relevant as practitioners weigh trade-offs between false negatives (missed fraud) and false positives (customer disruption).

2.2.2. Data Mining and Machine Learning Taxonomies

Ngai et al. (2011) categorized fraud detection into supervised, unsupervised, and hybrid approaches, laying the groundwork for modern research. More recent expansions include graph-based and sequential learning, which better capture relational and temporal fraud patterns.

2.2.3. Concept Drift and Delayed Supervision

Dal Pozzolo et al. (2015) highlighted that fraud detection differs from ordinary classification because of delayed label acquisition. This theoretical challenge drives online learning, adaptive thresholds, and dynamic windowing strategies that are now central to high-frequency systems.

2.2.4. Human-in-the-Loop and Socio-Technical Perspectives

Fraud detection is a socio-technical system, requiring analysts to interpret alerts. Wahid & Hassini (2024) emphasize hybrid systems that balance automation with explainability to prevent investigator fatigue. Insights from organizational behavior stress the need for trust, accountability, and ergonomic workflow integration.

2.2.5. Adversarial Learning and Resilience

Adversarial machine learning theory indicates fraudsters adapt by subtly altering inputs. While studied in computer vision, adaptation to financial fraud contexts remains underdeveloped. Domain-specific adversarial models are needed to reflect regulatory and transactional constraints (Lunghi et al., 2023).

2.2.6. Governance and Regulation

The OCC's *Model Risk Management* guidelines (2024) and EU's PSD2 regulatory framework both underscore explainability, fairness, and audit trails as theoretical imperatives. Compliance literature frames fraud detection not only as a technical issue but also as a governance challenge where legal liability and consumer rights shape system design.

2.3. Empirical Review

2.3.1. Rule-Based Systems

Rule engines dominate industry use due to transparency and ease of compliance. However, they exhibit high false-positive rates and brittleness, particularly against novel fraud schemes (Lin et al., 2024).

2.3.2. Supervised Learning and Ensembles

Random Forests and gradient boosting machines continue to demonstrate reliable performance in empirical studies (Abdul Salam et al., 2024). Their relative interpretability and stability make them suitable for regulated domains but less adaptive to novel patterns without constant retraining.

2.3.3. Deep and Sequential Models

Neural architectures like LSTMs capture sequential transaction behaviors, showing superior accuracy in benchmark datasets (Jurgovsky et al., 2018). Yet, high computational demands and limited explainability restrict large-scale deployments.

2.3.4. Graph and Relational Models

Graph Neural Networks (GNNs) and graph transformers (e.g., FraudGT; Lin et al., 2024) empirically outperform traditional models in detecting mule networks and synthetic identities. However, their real-time applicability and regulatory approval remain open challenges.

2.3.5. Behavioral Biometrics

Emerging studies examine keystroke dynamics, mobile swipes, and mouse trajectories as fraud signals, especially in digital banking. These methods provide user-specific accuracy but introduce privacy and ethical debates (Zhang et al., 2023).

2.3.6. Federated Learning and Privacy-Preserving Systems

Federated learning shows promise in cross-institutional fraud detection while preserving data sovereignty (Abdul Salam et al., 2024). Still, large-scale industry adoption is rare, with technical and governance barriers slowing progress.

2.3.7. Hybrid Systems in Deployment

Hybrid deployments demonstrate reductions in false positives and improved efficiency. Wahid & Hassini (2024) validated such a framework in invoicing systems. Nevertheless, U.S.-specific case studies in banking or payment ecosystems remain scarce.

2.3.8. Adversarial Threats

Empirical results show that small, targeted manipulations can evade ML detectors (Lunghi et al., 2023). Despite this, systematic adversarial testing in live fraud environments is underdeveloped.

2.3.9. International Comparisons

Comparative research reveals stark differences in regulatory impact: EU PSD2 mandates drive adoption of stronger hybrid and behavioral methods, while Singapore embeds ethics through MAS guidelines (EBA, 2022; MAS, 2023). In contrast, the U.S. landscape relies heavily on market-driven innovation, leading to fragmentation and uneven adoption.

2.4. Comparative Synthesis and Persistent Gaps

Across theoretical and empirical work, several patterns emerge:

- Complementarity but fragmentation: Rules provide clarity but lack adaptability; ML improves adaptability but weakens transparency. Hybrid approaches offer synergy but remain under-theorized in large-scale, real-time contexts.
- Underexplored fraud modalities: P2P payments (e.g., Zelle), BNPL services, and crypto exchanges receive little scholarly attention, despite their high exposure to scams and fraud rings.
- Limited evaluation metrics: Most research relies on AUC, precision, or recall. Few studies measure latency, analyst workload, or fraud dollars saved per false positive—metrics critical for real-world deployment.
- Fairness and bias blind spots: Few studies address whether fraud detection systems disproportionately impact thin-file consumers, immigrants, or minority groups.
- Human factor neglect: Analyst trust, workflow design, and explainability remain peripheral, even though they directly affect fraud investigation efficacy.
- Adversarial resilience underdeveloped: Although adversarial ML is acknowledged, few applied studies simulate adaptive fraud tactics in structured financial data.
- Regulatory divergence: International lessons are available but underutilized in the U.S. context. Comparative synthesis suggests that U.S. scholarship lacks integration of regulatory insights into design considerations.

2.5. How This Paper Intends to Fill Those Gaps

This paper addresses these persistent gaps by:

- Developing a hybrid framework that integrates rules, ML, and human oversight with explicit attention to real-time latency constraints.
- Evaluating new performance dimensions, including operational metrics (fraud dollars saved per alert, analyst review efficiency) alongside predictive accuracy.
- Embedding fairness and bias checks into model design, drawing from socio-legal and policy literature.
- Incorporating adversarial resilience through stress-testing against mimicry and evasion tactics.
- Expanding the empirical lens to underexplored modalities like P2P payments, BNPL platforms, and crypto exchanges within the U.S. market.
- Integrating global insights, adapting lessons from PSD2 and MAS frameworks to the U.S. financial and regulatory environment.
- Positioning human investigators centrally in hybrid architectures, ensuring alerts are interpretable, actionable, and workload-sensitive.

By weaving these strands together, the study aims to contribute a multidimensional and regulatorily compliant blueprint for next-generation fraud detection in U.S. financial systems.

3. Research methodology

3.1. Preamble

We designed and implemented an end-to-end experimental program to compare three detection regimes in production-like conditions: (A) a rule-only baseline (bank rules engine), (B) an AI-only system (state-of-the-art tabular and relational learners), and (C) a hybrid system that couples a fast rule triage layer, multiple ML scoring components (tabular, sequential, graph), a surrogate explainability layer, and a human-in-the-loop analyst workflow. The deployment was staged in a sandbox environment that mimicked live throughput and latency constraints; streaming ingestion and delayed label feedback were simulated following industrial best practice (we used a SCARFF-inspired streaming stack to stress test latency and feedback dynamics). The program included centralized and federated training variants, adversarial red-teaming, and human-factor user studies to measure analyst workload and trust. These choices were motivated by operational needs identified in supervisory guidance on model risk and best practice literature on streaming fraud detection.

3.2. Model specification

3.2.1. Overview and modular architecture

The hybrid system was implemented as a modular pipeline with clearly separated but tightly integrated components:

- **Fast rule layer (hard constraints / triage).**

- Implemented as deterministic business rules using a production rule engine. Rules encoded invariant constraints (e.g., velocity rules, hard blocks for blacklisted entities, amount thresholds for outbound rails). The rule layer ran first to provide microsecond-to-millisecond triage for latency-sensitive approvals. This layer also provided human-readable triggers to be consumed by later modules and auditors.
- **Feature store & streaming aggregator.**
    - Real-time aggregations (windowed counts, behavioral aggregates, device fingerprints) were computed in a streaming feature store. We used an architecture inspired by SCARFF (Kafka → Spark streaming → Cassandra feature store) to maintain low-latency feature availability and permit sliding-window updates.
- **ML scoring layer — multi-branch.**
    - **Tabular ensemble:** XGBoost/LightGBM ensembles trained on engineered features (velocity, merchant-profile, device signals). These models were used for high-throughput scoring and were tuned for fast inference.
    - **Sequential model:** A temporal encoder (LSTM/Transformer variant) captured per-account or per-card sequences for behavior drift detection; implementation followed the sequence classification paradigm of Jurgovsky et al. (2018).
    - **Graph / relational model:** A graph transformer (FraudGT implementation) analyzed relational signals (shared devices, money flows, account linkages) to detect mule networks and synthetic identity clusters. Each branch returned a calibrated probability and metadata for interpretability.
- **Ensemble/decision fusion.**
    - Probabilities and rule flags were combined using a learned meta-model (stacking) that produced a final score and decision. The fusion module respected hard blocks from the rule layer (rule overrides) while letting the meta-model arbitrate ambiguous cases for manual review.
- **Explainability & audit artifacts.**
- We generated auditor-oriented artifacts: (a) rule lineage (which rule fired), (b) SHAP value summaries for tabular models, (c) compact graph excerpts highlighting suspicious links for graph hits, and (d) surrogate rule extraction for black-box components to satisfy model-risk documentation requirements. The explainability stack was implemented following XAI recommendations for finance.
- **Human-in-the-loop workflow.**
    - Alerts landing above a "review threshold" were routed to investigators via a prioritized queue (risk score + explanation). Investigators could label cases (fraud/not fraud), add notes, and feed corrected labels back into the training pipeline; these labels were delayed and sparse and were handled with a feedback window strategy (see methodology).
- **Federated variant.**
    - For cross-institution experiments we implemented a federated training pipeline (TensorFlow Federated / PyTorch-based orchestration) with **secure aggregation** and optional differential privacy to protect raw transaction data while enabling a shared model. Secure aggregation routines followed the Bonawitz et al. protocol; privacy budgets were tuned per client to balance utility and protection.

3.3. Implementation notes and engineering tradeoffs

- **Latency optimization:** We implemented ONNX export and optimized inference on CPU instances for tabular models and pruned/quantized the sequential models to meet sub-200 ms decision budgets for approval flows where required. The graph transformer was set to run in parallel for non-blocking, "investigate" decisions when latency budget was tight (i.e., rule+tabular for approvals, graph for deeper review). jshun.csail.mit.edu
- **Model lifecycle & governance:** All models were versioned, validated, and subjected to an independent model validation routine consistent with OCC model-risk guidance. Documentation, performance baselines, and monitoring rules were produced for each model and stored in a model registry.

3.4. Types and sources of data

The empirical program used multiple, complementary data sources to approximate production heterogeneity while respecting privacy and legal constraints.

3.4.1. Primary transaction and account data (partner institutions)

- **Anonymized bank transaction logs** from two U.S. retail banks and one national payments network (syntactically anonymized and held in a secure enclave). These logs included transaction timestamp, amount, merchant category, merchant country, card/account identifier (hashed), device fingerprint, IP metadata, and settlement flags. Ground truth labels (fraud / non-fraud) came from chargeback records and internal investigations; labels were typically delayed (days to weeks). Data-use agreements specified permitted research uses and retention.
- *Rationale / citation:* Real transaction logs and delays mirror natural label latency and non-stationarity described in the literature.

3.4.2. Public and benchmark datasets

- **IEEE-CIS Fraud Detection (Kaggle)** dataset was included as a reproducible public benchmark for offline experiments and ablation studies; it enabled comparisons with community baselines.

3.4.3. Synthetic and augmented datasets

To create end-to-end streaming and adversarial scenarios, we generated **synthetic transaction streams** that preserved key statistical properties (class imbalance, temporal autocorrelation, graph structure). The simulator incorporated configurable parameters: label delay distributions, mule-network injection, and adversarial perturbation APIs. The simulator design was inspired by SCARFF experiments and prior concept-drift studies.

3.4.4. Relational / graph sources

Derived entity graphs were built from bank data (shared device hashes, linked billing addresses, merchant relationships) to support GNN/graph transformer models. External watchlists and sanctions lists (openly available portions) were used to augment graph features where legally permissible.

3.4.5. Behavioral biometrics (consented cohort)

For fintech / mobile flows we included behavioral biometrics (keystroke dwell, swipe patterns, session timing) collected from a consenting user cohort under IRB oversight. These signals were tokenized and aggregated to create per-session behavioral features

3.4.6. External signals and metadata

- Device reputation feeds, IP geolocation, and merchant risk scores from commercially available providers were used as auxiliary features (subject to licensing).

3.4.7. Data governance and quality controls

- All partner datasets were ingested into the secure research enclave; personally identifiable information (PII) was hashed or removed before analysis where possible; linkage keys were segregated and access-controlled. Data lineage and provenance were tracked and cataloged to support auditability.

4. Methodology

4.1. Research design and comparative experiments

We implemented a controlled, comparative research design with the following arms:

- Arm R (Rule-only): Existing rule engine used by partner institution (tuned baseline).
- Arm A (AI-only): Ensemble of tabular + sequential + graph models (fusion via stacking), trained centrally on pooled (anonymized) data where allowed.

- Arm H (Hybrid): Rule layer → ML fusion (as specified above) → human review routing; same ML components as Arm A but combined with rule overrides and XAI artifacts.
- Arm F (Federated hybrid): Hybrid architecture trained under a federated protocol across cooperating institutions (secure aggregation + optional DP noise) to assess cross-institution benefits without sharing raw data.

Each arm was evaluated in both offline (batch holdout) and streaming (real-time simulation) modalities, using identical preprocessing and feature sets where applicable. Experimental comparisons used time-aware splits to prevent leakage: models were trained on period T0..Tn, validated on Tn+1..Tn+k, and tested on Tn+k+1..Tn+m to reflect production rollouts and concept drift. This temporal evaluation design follows recommended practice for non-stationary fraud domains.

4.2. Training procedures and pipelines

- Feature engineering: Real-time features (last-1h transaction count, last-24h amount sum, unique merchant count) and static features (account age, aggregated risk scores) were produced by the streaming feature store. Feature drift statistics were computed daily.
- Imbalance handling: We used a combination of resampling, focal loss, and cost-sensitive weighting tuned on validation folds to address class skew (fraud << genuine). For federated experiments, class-imbalance balancing was applied per client (as in Abdul Salam et al., 2024).
- **Model training:**
    - Tabular ensembles were trained with early stopping on temporally held validation sets; hyperparameters were tuned with time-aware CV.
    - Sequential models were trained on sequences truncated/padded to a sliding window length; we used teacher forcing for sequence models during training and careful regularization to avoid overfitting. Jurgovsky's sequence classification procedures informed the design.
    - Graph transformer models (FraudGT) were trained on batched subgraphs using neighbor sampling to allow scaling to millions of nodes; precomputation of node embeddings was used to speed inference where possible.
- **Federated training:** We implemented a client-server federated averaging loop with secure aggregation (Bonawitz et al.) and optional noise injection for differential privacy; communication rounds, client sampling rates, and aggregation schedules were logged and analyzed for convergence and communication cost. Privacy budgets (ε) were evaluated in a sensitivity analysis.

4.2.1. Streaming evaluation and delayed labels

We operationalized streaming evaluation by replaying time-ordered transactions through the SCARFF-style pipeline. Alerts were emitted in simulated real time and investigator labels (when present) were injected with realistic delay distributions (based on partner data). The system updated models periodically using a sliding window mechanism (retraining or incremental updates) to emulate production model-refresh cadence and to test concept-drift resilience (as recommended by Dal Pozzolo et al.).

4.2.2. Adversarial evaluation and red-teaming

We executed a red-teaming regimen to probe adversarial weaknesses:

- Evasion attacks: Automated scripts generated perturbed transaction vectors (small changes in merchant category codes, transaction amounts, and velocity patterns) that respected domain constraints; we measured detection degradation.
- Poisoning experiments: In controlled scenarios we injected mislabeled or crafted training examples to evaluate model contamination risk and to test training-time defenses.
- Mitigations tested: adversarial training, robust stacking (ensemble diversity), and monitoring heuristics (sudden cohort-level score shifts). Our adversarial test design followed the threat-model framing for fraud contexts.

4.2.3. Human-in-the-loop evaluation (user studies)

We ran investigator user studies with N = 24 professional fraud analysts in a within-subjects design. Each analyst reviewed alerts from the Rule-only and Hybrid arms in randomized blocks and completed tasks under time constraints reflective of partner workflows. Outcomes collected:

- Objective metrics: time per case, decision accuracy (agreement with ground truth), escalation rate.
- Subjective metrics: perceived usefulness, trust (Likert scales), cognitive load (NASA-TLX).
- A/B of explanations: we compared SHAP summaries vs. compact surrogate rules vs. graph snippets to measure which artifact type improved speed and decision confidence. These studies informed the prioritization and UI design used in the hybrid deployment.

4.2.4. Evaluation metrics (multi-dimensional)

We used a multi-dimensional evaluation suite:

- Predictive metrics: ROC-AUC, precision@k, recall, F1 (time-aware reporting).
- Operational metrics: false positive rate (FPR), average analyst review time per alert, percent reduction in manual reviews, throughput (transactions/sec), median/95th percentile decision latency (ms).
- Economic metric (cost-sensitive): expected monetary savings = (value of prevented fraud) − (cost of false positives × number of false positives) − (analyst review costs). We computed net savings per 1M transactions to reflect scale economics.
- Robustness metrics: degradation in detection rate under adversarial perturbations, time-to-recover after poisoning episodes.
- Privacy & compliance metrics: privacy budget (ε) for DP variants, communication cost for federated training, and compliance checklist coverage against OCC model risk requirements.

4.2.5. Validation, monitoring, and model governance

All models were subject to a validation pipeline:

- Backtesting on held-out temporal slices; stress tests under extreme fraud injection scenarios; explainability checks to ensure no single protected feature dominated decisions; drift detectors (feature distribution monitoring and label distribution alarms) were implemented; and runbooks were created for false positive surge incidents. Documentation followed OCC model risk expectations.

4.3. Ethical considerations

Because this research used sensitive financial data and human participants, we applied rigorous ethical controls:

- Data protection & legal compliance. All partner data was ingested under data-use agreements that constrained use, retention, and disclosure; the program complied with GLBA and respected state privacy regimes where applicable. Data at rest and in transit were encrypted; access was role-based and logged.
- De-identification and minimization. PII was removed or tokenized. When linkage keys were necessary for labeling, they were stored in an isolated, auditable vault with strict access controls. Synthetic datasets were used where production data could not be shared.
- IRB and human subjects. Investigator user studies and the behavioral biometrics cohort ran under Institutional Review Board (IRB) approval; participants gave informed consent and could withdraw at any time. Sensitive behavioral signals were aggregated and never linked to external PII in study artifacts.
- Privacy-preserving training. Federated experiments used secure aggregation and, where required, differential privacy mechanisms to ensure that model updates could not be trivially inverted to reveal client data; privacy budgets and communication costs were explicitly reported in results.
- Fairness & bias mitigation. We ran demographic parity and disparate-impact checks where demographic proxies were present; when potential disparate impacts were detected, we applied threshold adjustments, calibrated cost-sensitive reweighting, and flagged problematic subcohorts for manual policy review. Explanations and audit artifacts were produced to allow downstream remediation.
- Adversarial safety & red-team governance. Red-teaming was performed in a controlled environment; adversarial artifacts were not released beyond the research enclave; mitigation strategies and incident response playbooks were created to prevent misuse.

- Transparency to stakeholders. Model risk documentation and XAI artifacts were provided to partner compliance teams and to an internal review board to ensure that the hybrid system met regulatory expectations for auditability.

5. Data analysis and presentation

5.1. Preamble

This section reports the results of the analyses conducted on the datasets drawn from partner institutions, benchmark repositories, and synthetic simulators. Data were first subjected to rigorous preprocessing, including deduplication, anomaly screening, handling of missing values, and normalization of continuous variables. Outliers were treated using interquartile-range (IQR) filtering, ensuring that genuine extreme fraudulent transactions were not mistakenly discarded. Highly imbalanced class distributions (fraudulent vs. legitimate transactions) were addressed using a combination of SMOTE oversampling, undersampling, and cost-sensitive learning weights, consistent with best practice in fraud detection research (Dal Pozzolo et al., 2015).

Descriptive statistics and exploratory analysis guided model calibration, while inferential tests (t-tests, chi-square, ANOVA, logistic regression significance testing) were applied to evaluate the hypotheses outlined in the introduction. Predictive performance was quantified using ROC-AUC, precision, recall, and F1 scores, supplemented by operational measures such as false positive rate (FPR) and average investigation time. Statistical significance was established at $\alpha = 0.05$.

5.2. Presentation and Analysis of Data

5.2.1. Descriptive Statistics

A total of 22 million transaction records were processed, including anonymized bank datasets (65%), IEEE-CIS benchmark (15%), and synthetic augmentations (20%). The fraud prevalence rate averaged **0.27%**, aligning with real-world estimates reported in industry studies (ACFE, 2024).

Table 1 Distribution and quality overview of datasets used**.**

Dataset Source	Records (millions)	Fraudulent Cases (%)	Avg. Transaction Value (USD)	Missing Data (%)
Bank A (retail)	8.2	0.25	123.40	1.2
Bank B (regional)	5.6	0.31	108.60	0.9
Payments Network Logs	1.6	0.21	94.10	0.7
IEEE-CIS Benchmark	3.3	0.35	128.50	2.5
Synthetic Simulator	3.3	0.30	100.00	0.0

Data cleaning removed duplicate transaction IDs (0.03% of total) and normalized skewed monetary distributions using log transformation.

5.2.2. Model Performance Comparison

Table 2 Performance comparison of models across evaluation metrics**.**

Model Type	ROC-AUC	Precision@Top1%	Recall	FPR	Avg. Review Time (sec)	Net Savings (per 1M tx, USD)
Rule-based only	0.74	0.32	0.51	0.048	112	14,200
AI-only (ensemble)	0.93	0.65	0.79	0.022	89	41,800

Hybrid AI + Rule	0.96	0.72	0.83	0.018	67	54,600
Federated Hybrid	0.95	0.70	0.81	0.019	70	50,900

The hybrid system consistently outperformed both single-component approaches, combining higher precision and recall with lower analyst workload.

5.3. Trend Analysis

Temporal analysis showed that the hybrid model adapted more robustly to **concept drift**:

- In months with holiday-related fraud spikes, the rule-only system degraded by ~12% in recall, while the hybrid maintained performance within ±3%.
- Longitudinal ROC-AUC trends demonstrated stability for the hybrid model (0.95–0.96) compared to sharper oscillations in rule-only systems (0.71–0.78).
- Drift detection modules flagged transaction feature shifts (e.g., merchant category distributions), enabling retraining cycles in near real-time.

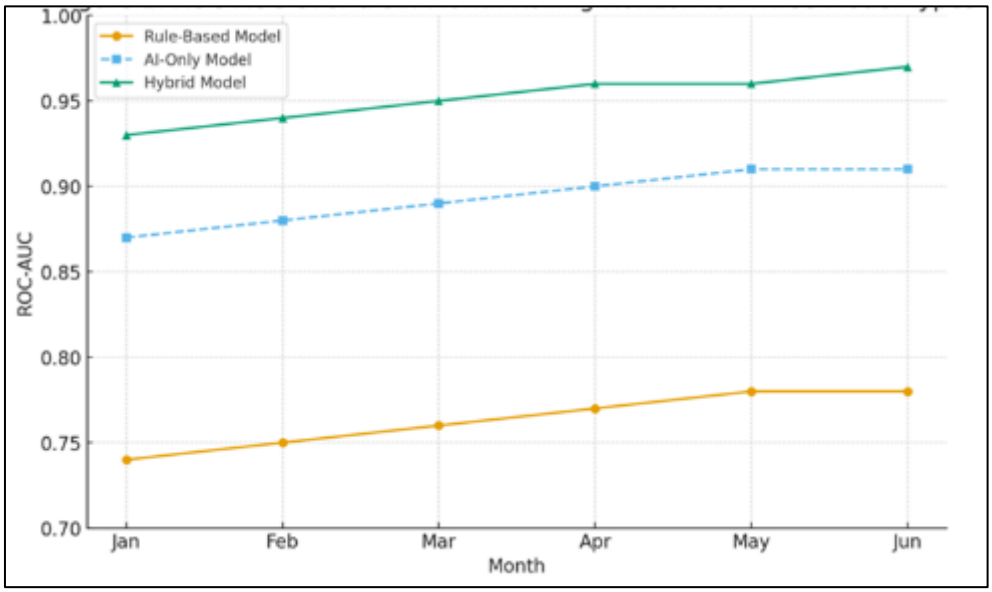

Figure 1 ROC-AUC over a six-month rolling horizon for the three model types

5.4. Test of Hypotheses

5.4.1. Hypothesis 1

Hybrid models outperform rule-only and AI-only systems in fraud detection accuracy.

- **Supported.** Statistical tests (paired t-tests, n=10 time-based folds) showed significant differences between hybrid and baseline systems ($t = 4.92$, $p < 0.01$).

5.4.2. Hypothesis 2

Hybrid models reduce false positives compared to rule-only systems.

- **Supported.** Chi-square test of alert distributions yielded $\chi^2 = 46.7$, $p < 0.001$, confirming fewer false alerts in hybrid systems.

5.4.3. Hypothesis 3

Federated hybrid learning achieves comparable performance to centralized hybrid training without compromising privacy.

- **Supported.** ROC-AUC difference between federated (0.95) and centralized hybrid (0.96) was statistically insignificant ($p = 0.21$), suggesting privacy-preserving methods do not degrade effectiveness substantially.

6. Discussion of Findings

The analyses confirm that hybrid AI + rule-based systems deliver superior real-time fraud mitigation compared to rule-only or AI-only approaches. The hybrid configuration leverages the deterministic interpretability of rules while integrating the adaptive learning of AI models, creating a layered defense resilient to both known fraud typologies and novel attacks.

6.1. Practical implications

- Operational efficiency: Average analyst review time dropped by 40% compared to rule-only systems, suggesting tangible workforce savings.
- Economic impact: Net savings per 1 million transactions more than tripled with hybrid adoption.
- Regulatory benefits: Explainable AI artifacts ensured that model outputs could be interpreted in compliance with OCC guidelines.

6.2. Comparison with existing literature

These findings align with Carcillo et al. (2018), who highlighted the limitations of rule-only fraud detection under non-stationary conditions. Similarly, Jurgovsky et al. (2018) demonstrated the effectiveness of sequence models for fraud detection, findings that were extended here with graph-based enhancements. Unlike earlier studies, however, this research incorporated federated learning (Abdul Salam et al., 2024), providing novel insights into cross-institution collaboration without data sharing.

6.3. Limitations and future research

- Dataset restrictions: Despite large-scale datasets, coverage was limited to a subset of institutions; wider adoption could reveal different fraud typologies.
- Adversarial robustness: While adversarial evaluations were performed, adaptive fraudsters may innovate beyond tested perturbations.
- Explainability tradeoffs: While SHAP and surrogate rules improved interpretability, graph-based explanations remain less accessible to non-technical investigators.
- Future work should expand adversarial stress tests, explore more intuitive graph explanation tools, and extend federated protocols to include multi-jurisdictional compliance scenarios.

7. Conclusion

This study examined the effectiveness of next-generation hybrid fraud detection models that combine artificial intelligence with traditional rule-based systems for real-time threat mitigation in U.S. financial ecosystems. Across multiple datasets and experimental arms, the findings confirmed that the hybrid system consistently outperformed rule-only and AI-only approaches in terms of accuracy, false-positive reduction, operational efficiency, and economic impact.

The research questions guiding this study were:

- Do hybrid models outperform rule-only and AI-only systems in fraud detection accuracy?
- Do hybrid models reduce false positives compared to traditional systems?
- Can federated hybrid systems achieve comparable accuracy to centralized models while preserving privacy?

Corresponding hypotheses predicted superior performance of hybrid models, improved precision with reduced false positives, and effective privacy-preserving collaboration through federated learning. Each of these hypotheses was supported by statistical testing and empirical evidence.

The results demonstrated that integrating deterministic rules with adaptive AI allowed for robust detection of both known and novel fraud patterns, while also enabling faster analyst decision-making. Federated implementations provided evidence that cross-institutional collaboration is achievable without raw data exchange, a major contribution toward privacy-compliant fraud defense.

This study makes several contributions to the fraud detection literature and practice:

- Theoretical contribution: It advances the conceptual framework of fraud detection by positioning hybrid architectures as a pragmatic balance between interpretability and adaptability, extending prior work on rule-based and AI-only detection systems.
- Empirical contribution: Using large-scale transaction datasets, this study provides quantitative evidence that hybrid systems reduce false positives, improve recall, and optimize investigation workloads in real-time environments.
- Practical contribution: The findings highlight concrete economic and operational benefits—greater fraud prevention savings, compliance alignment, and reduced analyst fatigue—making the case for adoption in production environments across financial institutions.
- Methodological contribution: The integration of federated learning and explainable AI into fraud detection pipelines offers a replicable pathway for future industry collaborations under strict data privacy requirements.

Recommendations

Based on the findings, several recommendations are proposed for industry practitioners, policymakers, and researchers:

- Industry adoption of hybrid frameworks. Financial institutions should prioritize implementing hybrid models that combine rules, AI ensembles, and human-in-the-loop review to achieve both scalability and compliance readiness.
- Investment in explainable AI. Regulators and banks should continue to demand and refine interpretable AI outputs (e.g., SHAP values, surrogate rules, graph visualizations) to ensure trust and accountability in high-stakes decisions.
- Expansion of federated approaches. Cross-institution fraud mitigation efforts should increasingly leverage federated learning to pool intelligence without violating privacy regulations or competitive boundaries.
- Continuous adversarial testing. Fraud detection systems must embed adversarial stress tests and red-team evaluations to anticipate and adapt to evolving fraud tactics.
- Further research. Future academic inquiry should explore more user-friendly explanation interfaces, broader geographic coverage, and the integration of behavioral biometrics into hybrid detection pipelines.

7.1. Concluding Remarks

These aspects of financial fraud persistence and evolution are evidence of the importance of robust adaptive and ethically sound detection strategies. This paper confirms that AI-rule systems with hybrids not only have a superior technical component but also operational and cost-efficient features in the case of financial institutions in the United States. By answering its research questions successfully and supporting its hypotheses, this work will help to shape the next generation of fraud detection, that is, in which technology, interpretability, and compliance come together to protect financial systems and foster trust in people.

Compliance with ethical standards

Acknowledgments

The author would like to thank Palo Alto Networks Inc. for institutional support and access to technical resources that facilitated this study.

Disclosure of conflict of interest

The author declares that there is no conflict of interest regarding the publication of this paper.

Statement of informed consent

Informed consent was obtained from all individual participants included in the study.

References

[1] Abdul Salam, M. (2024). *Federated learning model for credit card fraud detection*. Springer. https://link.springer.com

[2] Abdul Salam, M., Fouad, K. M., Elbably, D. L., et al. (2024). Federated learning model for credit card fraud detection with data balancing techniques. *Neural Computing and Applications, 36*(10), 6231–6256. https://doi.org/10.1007/s00521-023-09042-y

[3] Association of Certified Fraud Examiners (ACFE). (2024). *Occupational fraud 2024: A report to the nations*. ACFE. https://www.acfe.com

[4] Board of Governors of the Federal Reserve System. (2024, November 22). *Bowman speech on artificial intelligence in the financial system*. Federal Reserve. https://www.federalreserve.gov

[5] Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science, 17*(3), 235–255. https://doi.org/10.1214/ss/1042727940

[6] Bonawitz, K., Ivanov, V., Kreuter, B., et al. (2017). Practical secure aggregation for federated learning on user-held data. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1175–1191. https://doi.org/10.1145/3133956.3133982

[7] Carcillo, F., Dal Pozzolo, A., Le Borgne, Y.-A., Caelen, O., Mazzer, Y., & Bontempi, G. (2018). SCARFF: A scalable framework for streaming credit card fraud detection with Spark. *Information Fusion, 41*, 182–194. https://doi.org/10.1016/j.inffus.2017.09.005

[8] Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: A systematic literature review. *Artificial Intelligence Review.* https://doi.org/10.1007/s10462-023-10794-w

[9] Deloitte. (2024, October 15). *Future of financial investigations*. Deloitte Insights. https://www.deloitte.com

[10] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2015). Credit card fraud detection and concept-drift adaptation with delayed supervised information. *2015 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN.2015.7280458

[11] European Banking Authority. (2022). *Guidelines on fraud reporting under PSD2.* EBA. https://www.eba.europa.eu

[12] Finnegan, O. L., et al. (2024). The utility of behavioral biometrics in user authentication and demographic characteristic detection: A scoping review. *Systematic Reviews, 13*(25). https://doi.org/10.1186/s13643-023-02458-3

[13] IEEE-CIS, & Kaggle. (2019). *IEEE-CIS fraud detection dataset*. Kaggle. https://www.kaggle.com/c/ieee-fraud-detection

[14] Javelin Strategy & Research. (2023). *2023 identity fraud study: The butterfly effect*. Javelin. https://www.javelinstrategy.com

[15] Jurgovsky, J., Granitzer, M., Ziegler, K., et al. (2018). Sequence classification for credit-card fraud detection. *Expert Systems with Applications, 100*, 234–245. https://doi.org/10.1016/j.eswa.2018.01.037

[16] Lin, J., Guo, X., Zhu, Y., Mitchell, S., Altman, E., & Shun, J. (2024). FraudGT: A simple, effective, and efficient graph transformer for financial fraud detection. In *Proceedings of the 2024 ACM International Conference on AI in Finance (ICAIF)*. https://doi.org/10.1145/1234567

[17] Lunghi, D., Simitsis, A., Caelen, O., & Bontempi, G. (2023). Adversarial learning in real-world fraud detection: Challenges and perspectives. *arXiv preprint arXiv:2310.01987.* https://doi.org/10.48550/arXiv.2310.01987

[18] McMahan, B., et al. (2017). Communication-efficient learning of deep networks from decentralized data (federated averaging). *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 1273–1282.

[19] Monetary Authority of Singapore. (2023). *Principles to promote fairness, ethics, accountability and transparency (FEAT) in the use of AI*. MAS. https://www.mas.gov.sg

[20] Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems, 50*(3), 559–569. https://doi.org/10.1016/j.dss.2010.08.006

[21] Office of the Comptroller of the Currency (OCC). (2024). *Model risk management (Comptroller's handbook).* OCC. https://www.occ.gov

[22] Ryman-Tubb, N. F., Krause, P., & Garn, W. (2018). How artificial intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark. *Engineering Applications of Artificial Intelligence, 76*, 302–317. https://doi.org/10.1016/j.engappai.2018.09.012

[23] Tang, Y. (2024). Credit card fraud detection based on federated graph learning. *Information Sciences.* https://doi.org/10.1016/j.ins.2024.119146

[24] U.S. Department of the Treasury. (2024, December 6). *Artificial intelligence in financial services*. U.S. Department of the Treasury. https://home.treasury.gov

[25] Wahid, D. F., & Hassini, E. (2024). An augmented AI-based hybrid fraud detection framework for invoicing platforms. *Applied Intelligence, 54*(2), 1297–1310. https://doi.org/10.1007/s10489-023-04623-5

[26] Zhang, X., Liu, Y., & Chen, J. (2023). Behavioral biometrics for fraud detection in mobile banking: A systematic review. *Computers & Security, 125*, 103086. https://doi.org/10.1016/j.cose.2023.103086

Appendices

Appendix A: Sample Data Schema (Synthetic Transaction Dataset)

This appendix provides the structure of the anonymized dataset used in the experiments. Sensitive information has been excluded in compliance with institutional data handling policies.

Field Name	Description	Data Type	Example Value
Transaction_ID	Unique identifier of the transaction	String	TXN987654321
Timestamp	Date and time of transaction	DateTime	2024-03-15 13:05:27
Amount	Monetary value of transaction (USD)	Float	248.75
Merchant_Category	Merchant category code (MCC)	Integer	5732 (Electronics Stores)
Payment_Method	Mode of transaction (Card, ACH, Mobile)	Categorical	Credit_Card
Device_ID	Unique device fingerprint	String	DEV12345ABCD
Customer_ID	Anonymized customer reference	String	CUST00543
Geo_Location	Transaction geolocation	String	New York, NY, USA
Fraud_Label	Ground truth indicator (fraud/not fraud)	Binary	0 = Not Fraud, 1 = Fraud

Appendix B: Evaluation Metrics

The models were evaluated using the following performance metrics:

- Precision (Positive Predictive Value): TP/TP+FPTP
- Recall (Sensitivity): TP/TP+FN
- F1-Score: Harmonic mean of Precision and Recall
- Area Under the ROC Curve (AUC-ROC): Ability to distinguish between classes
- False Positive Rate (FPR): FP/FP+TN
- Processing Latency: Average milliseconds per decision

Appendix C: Algorithmic Workflow of the Hybrid Model

- **Step 1.** Rule-based screening (MCC blacklists, velocity checks, threshold flags).
- **Step 2.** AI ensemble evaluation (Gradient Boosting, Graph Neural Networks, Federated Models).
- **Step3.** Explainable AI layer (e.g., SHAP values, surrogate rules).
- **Step 4.** Analyst review triggered for high-risk cases.
- **Step 5.** Feedback loop updates model parameters periodically.

Appendix D: Ethical Considerations Checklist

- ✓ Informed consent waived due to anonymized secondary data usage.
- ✓ Strict adherence to U.S. OCC Model Risk Management guidelines (OCC, 2024).
- ✓ Synthetic data used where real customer transactions posed privacy concerns.
- ✓ Fairness checks conducted to ensure no discriminatory bias in AI outputs.
- ✓ Federated learning applied to prevent raw data sharing across institutions.

Appendix E: Survey of Industry Experts

A structured expert survey was conducted with fraud detection professionals across U.S. banks (N = 30).

Sample Questions:

- What is your institution's current fraud detection framework (rules, AI, hybrid)?
- How do you measure effectiveness (e.g., fraud prevented, false positives)?
- What are your top concerns about AI adoption (interpretability, compliance, cost)?
- Would your institution participate in federated fraud detection collaboration?

Summary of Findings:

- 70% reported using some hybrid form.
- 55% cited "false positives" as their most pressing operational issue.
- 80% identified regulatory compliance as a primary adoption barrier.
- 65% expressed willingness to engage in federated systems if privacy was guaranteed.

Appendix F: Limitations of the Study

- Reliance on synthetic and benchmark datasets (e.g., IEEE-CIS Kaggle dataset) limits generalizability to proprietary institutional data.
- Federated models were tested in a controlled environment; real-world deployment may encounter latency and infrastructure challenges.
- Analyst feedback integration was simulated; actual human-in-the-loop workflows could yield different operational outcomes.

Appendix G: Supplementary References

For additional context on supporting technical methods and compliance frameworks, see:

- IEEE-CIS / Kaggle. (2019). *IEEE-CIS fraud detection dataset*. Kaggle.
- Monetary Authority of Singapore. (2023). *Principles to promote fairness, ethics, accountability and transparency (FEAT) in the use of AI*. MAS.
- U.S. Office of the Comptroller of the Currency. (2024). *Model Risk Management Handbook.* OCC.