Strengthening U.S. Financial and Cybersecurity Infrastructure with AI-Driven Fraud Detection and Risk Analytics

1. Name Md Nazmul Hasan

Affiliation: Pompea College of Business, University of New Haven, West Haven, CT, USA

Email- mhasa9@unh.newhaven.edu

ORCID- 0009-0007-2555-6047

2. Name: Imran Hossain Rasel

Affiliation: Pompea College of Business, University of New Haven, West Haven, Connecticut, United States.

Email: irase1@unh.newhaven.edu

ORCID- 0009-0002-9225-4595

3. Name: Muhibbul Arman

Affiliation: Pompea College of Business, University of New Haven, West Haven, Connecticut, United States.

Email: iamarmanmohd@gmail.com

ORCID ID: 0009-0009-4629-3412

4. Name: Md Ibrahim

Affiliation: Pompea College of Business, University of New Haven, West Haven, Connecticut, United States.

Email: sohelibrahim200@gmail.com

ORCID ID: 0009-0008-2835-9871

5. Name: Nusrat Jahan

Affiliation: Master of Science in Analytics and Systems, University of Bridgeport, Bridgeport, CT, USA.

Email: njahan@my.bridgeport.edu

ORCID ID: https: 0009-0008-0684-1114

Abstract

The U.S. financial system faces converging threats from digital payment fraud, account takeover, and cyber intrusions that exploit highly connected infrastructure. This paper advances a practical blueprint for strengthening U.S. financial and cybersecurity infrastructure with artificial intelligence (AI) driven fraud detection and risk analytics. We synthesize pre–February 2023 research on supervised, unsupervised, temporal, and graph approaches; review evaluation for extreme class imbalance; and integrate privacy-preserving collaboration and rigorous model risk management. The proposed reference architecture combines streaming decisioning with gradient-boosted trees, sequence models, and graph risk scores; fuses cyber telemetry under zero-trust principles; and embeds governance aligned to SR 11-7/OCC 2011-12, NIST SP 800-53 and 800-207, and the NIST AI Risk Management Framework 1.0 (Federal Reserve, 2011; OCC, 2011; NIST, 2020, 2023). A cost-aware operating-point procedure uses precision–recall curves and expected-loss optimization to balance prevented fraud with analyst review costs and customer friction (Saito & Rehmsmeier, 2015; Davis & Goadrich, 2006). We also outline sector-level coordination through CISA initiatives and FFIEC and NYDFS expectations to ensure operational resilience and auditability (FFIEC, 2011, 2021; NYDFS, 2017; CISA, 2022). Illustrative figures and tables show evaluation and governance mapping using simulated data. The discussion provides implementation guidance, policy alignment, and research opportunities on drift-aware learning, adversarial robustness, and interoperable data sharing. Taken together, these measures can reduce fraud losses, compress attacker dwell time, and dampen systemic propagation, strengthening trust in digital finance without sacrificing privacy or accountability.

Keywords: fraud detection; risk analytics; zero trust; federated learning; model risk management; precision–recall.

Introduction

U.S. financial services now function like software-defined networks, connecting issuers, acquirers, wallets, merchants, core processors, cloud hosts, and analytics platforms. While this fast-paced connectivity accelerates payments and innovation, it also broadens the attack surface. Organized criminals automate card-not-present fraud, mule recruitment, and credential stuffing using everyday devices and stolen identities. Sophisticated attackers exploit software supply chains, session tokens, and misconfigurations to evade defenses and move across institutions. These attack vectors produce intertwined operational, financial, and reputational threats that no single team or firm can fully address alone. As the industry digitizes, reinforcing financial and cybersecurity infrastructure demands a design philosophy that sees fraud, cyberattacks, and systemic spread as interconnected—not just isolated compliance issues.

Machine learning provides strong tools to meet this challenge. Supervised models like gradient-boosted decision trees work well with high-cardinality, sparse behavioral features extracted from real-time payment data. Temporal models analyze short-term spending habits and session patterns. Graph analytics uncover collusive networks, synthetic identities, and compromised merchant clusters that traditional rules might miss. Crucially, these efforts operate in an environment of extreme class imbalance and constant adversary evolution, requiring institutions to prioritize precision–recall analysis and cost-sensitive thresholds over standard accuracy metrics, which may obscure risk (Saito & Rehmsmeier, 2015; Davis & Goadrich, 2006). Effective programs balance algorithmic sophistication with rigorous evaluation and clearly defined operational objectives.

However, scalable AI adoption is inseparable from governance, privacy, and security. U.S. supervisory guidance on model risk management (SR 11-7/OCC 2011-12) requires that models be built on conceptually sound methods, independently validated, monitored for drift, and governed through change control and documentation (Federal Reserve, 2011; OCC, 2011). Cross-sector security frameworks—including NIST SP 800-53 Rev. 5, SP 800-207 Zero Trust, and the NIST AI Risk Management Framework 1.0—offer control catalogs and trustworthy AI functions that map directly to data handling, identity, model inventories, and transparency (NIST, 2020, 2023). Sector rules such as 23 NYCRR 500 and the FFIEC's authentication guidance further codify layered security, incident response, and auditable records that production AI systems must satisfy (NYDFS, 2017; FFIEC, 2011, 2021).

Ecosystem coordination complements firm-level controls. CISA's Joint Cyber Defense Collaborative and Shields Up advisories provide shared situational awareness and actionable mitigations during periods of elevated risk, while Financial ISACs circulate indicators of compromise that often precede fraud surges (CISA, 2021, 2022). At the prudential level, Basel reforms and network research clarify how operational shocks can transmit across exposures and critical service dependencies, underscoring the need to embed cyber- and fraud-aware analytics into risk dashboards used for decision making (BIS, 2016, 2017; Acemoglu, Ozdaglar, & Tahbaz-Salehi, 2015).

This paper contributes a practitioner-oriented synthesis and blueprint. First, we review pre-February 2023 literature on supervised, unsupervised, temporal, and graph methods for fraud and cyber anomaly detection, emphasizing evaluation under imbalance and label latency. Second, we propose a reference architecture that unifies streaming feature stores, hybrid models, graph analytics, and cyber telemetry under zero-trust principles, with model-risk governance aligned to SR 11-7, OCC 2011-12, and the NIST AI RMF. Third, we illustrate cost-aware operating-point selection with simulated precision–recall curves and a simple expected-loss model to calibrate thresholds by channel. Fourth, we map the stack to regulatory and supervisory controls from NIST, NYDFS, FFIEC, and federal guidance. Finally, we discuss implementation, limitations, and research directions—drift-aware learning, adversarial robustness, privacy-preserving collaboration, and systemic-risk integration—that can further strengthen U.S. financial and cybersecurity infrastructure.

Literature Review

Fraud analytics has progressed from expert rules and linear scorecards to ensemble and deep architectures that learn complex interactions from high-volume event streams. Gradient-boosted decision trees became a de facto industry baseline because they handle heterogeneous features, missingness, and strong nonlinearities while delivering millisecond-level scoring suitable for real-time payments. XGBoost introduced a scalable tree boosting system with sparsity-aware split finding and regularization that improved both accuracy and efficiency (Chen & Guestrin, 2016). LightGBM contributed gradient-based one-sided sampling and exclusive feature bundling to accelerate training on large, high-cardinality datasets (Ke et al., 2017). CatBoost addressed the leakage and high variance that can arise with categorical variables by employing ordered boosting and target statistics computed in a causal, out-of-fold manner (Prokhorenkova, Gusev, Vorobev, Dorogush, & Gulin, 2018). Across card, ACH, wire, and e-commerce channels, these families consistently outperform legacy logistic-only systems when paired with domain-informed aggregates such as velocity counts, monetary dispersion, merchant entropy, and device affinity.

Temporal modeling reframes fraud as behavioral deviation over sequences of actions. Jurgovsky et al. (2018) show that recurrent neural networks leveraging transaction sequences increase discrimination and early detection relative to static features alone, especially when labels are sparse or delayed. Sequence models can also encode session patterns in online banking—logins, challenge responses, payee creation, device changes—capturing subtle escalations characteristic of account takeover. In production, hybrid stacks that combine boosted trees for robust baselines with sequence scores to represent recent behavior offer a practical compromise between accuracy and latency.

Relational structure is central to many fraud campaigns. Graph-based methods view cards, accounts, devices, IPs, merchants, addresses, and emails as nodes linked by interactions or shared attributes. Message-passing neural networks, graph attention, and inductive embedding methods propagate risk across neighborhoods to surface mule clusters,

reshipping rings, and coordinated refund abuse that may be invisible at the individual edge. While graph learning matured predominantly in recommendation and cybersecurity, financial applications increasingly report material gains when entity resolution is reliable and topologies are updated in near real time. Graph features—such as degree, motif counts, and community risk—also act as high-value inputs to non-graph classifiers.

Unsupervised anomaly detection remains vital where fraud labels are scarce or compromised by feedback loops. Isolation Forest isolates rare, high-risk points through random partitioning with strong performance and limited tuning, making it a solid baseline for event triage (Liu, Ting, & Zhou, 2008). Autoencoders flag unusual observations via reconstruction error; ensembles of shallow autoencoders have been deployed for network intrusion detection in streaming settings (Mirsky, Doitshman, Elovici, & Shabtai, 2018). DeepLog models sequences of system events to detect deviations, an approach that portably applies to authentication and authorization logs surrounding financial transactions (Du, Li, Zheng, & Srikumar, 2017). These cyber-centric techniques complement payment analytics by identifying compromised sessions and infrastructure precursors that elevate fraud risk.

Evaluation under extreme class imbalance requires metrics and procedures that foreground positive-class performance where it matters operationally. The precision–recall curve and its area summarize the trade-off between capturing fraud and constraining false alerts, and are more informative than ROC analysis in the low-prevalence regime typical of payments (Saito & Rehmsmeier, 2015; Davis & Goadrich, 2006). Thresholds should be selected by expected cost, accounting for average fraud loss, chargeback recovery, analyst review expense, and customer friction. Imbalance countermeasures include cost-sensitive learning, calibrated probability outputs, and resampling strategies such as SMOTE to enhance representation of minority classes during training (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). Focal loss, initially proposed for dense detection in vision, has inspired adaptations that emphasize complex examples without overwhelming training with easy negatives (Lin, Goyal, Girshick, He, & Dollár, 2017).

Label latency and concept drift complicate both training and monitoring. In payments, ground truth often arrives after chargeback windows or investigator reviews, meaning that online models operate for long stretches without fresh labels. Dal Pozzolo, Boracchi, Caelen, Alippi, and Bontempi (2018) formalize realistic evaluation protocols and propose strategies that remain stable under nonstationarity. Human-in-the-loop practices—active learning, triage queues, adjudication notes—reduce mislabeling and align learning with operational priorities (Dal Pozzolo, 2015). Streaming feature stores, challenger models, and periodic back-tests are essential to detect drift and to prevent silent performance regressions.

Trustworthy deployment depends on governance. U.S. banking guidance SR 11-7 and OCC 2011-12 describe a model life cycle with explicit roles for development, independent validation, and governance, coupled with inventories, documentation, and change control (Federal Reserve, 2011; OCC, 2011). The NIST AI Risk Management Framework 1.0 (2023) articulates functions—Map, Measure, Manage, and Govern—and characteristics (validity, reliability, safety, security, accountability, explainability, privacy, and fairness) that institutions can align to model risk processes. For infrastructure protection, NIST SP

800-53 Rev. 5 catalogs security and privacy controls across access, audit, incident response, and supply chain, while SP 800-207 codifies zero-trust architecture principles of continuous verification and least privilege (NIST, 2020, 2023).

Sector-specific expectations sharpen the control landscape. New York's 23 NYCRR 500 mandates risk assessments, multi-factor authentication, incident response, and 72-hour notifications that cover any AI-enabled fraud platform and its data flows (NYDFS, 2017). The FFIEC's 2011 Supplement and its 2021 guidance on authentication and access require layered security proportional to risk for customers, employees, and vendors, as well as governance for identification and mitigation of cyber threats (FFIEC, 2011, 2021). The BSA/AML Examination Manual provides procedures for risk-based monitoring and suspicious activity reporting; analytics that drive alerts must therefore be documented, effective, and explainable to examiners.

Privacy-preserving learning allows collaboration without wholesale data pooling. Differential privacy offers mathematically rigorous guarantees that limit the influence of any individual record on aggregate outputs (Dwork, 2006; Dwork & Roth, 2014). Federated learning enables institutions to train a shared model by exchanging model updates rather than raw data; secure aggregation prevents the server from learning any participant's update in isolation (McMahan, Moore, Ramage, Hampson, & Arcas, 2017; Bonawitz et al., 2017). Deep learning with differential privacy demonstrates that nontrivial accuracy can be achieved while bounding privacy risk (Abadi et al., 2016). These techniques are attractive for patterns—synthetic identities, merchant compromises—that span multiple firms' blind spots.

Finally, macroprudential perspectives connect micro-level detection to system resilience. Basel III capital and liquidity reforms and the Fundamental Review of the Trading Book tightened standards after the global financial crisis, while subsequent work in network science quantified conditions under which interconnectedness dampens or amplifies contagion (BIS, 2016, 2017; Acemoglu et al., 2015). DebtRank-style measures estimate the systemic importance of nodes in an exposure network, providing a lens to judge how payment disruptions or fraud spikes might cascade (Bardoscia, Battiston, Caccioli, & Caldarelli, 2015). Integrating AI-derived anomaly indicators with network metrics can make supervisory dashboards and internal stress tests more sensitive to cyber-initiated shocks.

Methodology

For regulated U.S. financial institutions, this part outlines a reference architecture and operational methodology for AI-driven fraud detection and risk analytics. The design principles are as follows: (1) risk alignment—models and thresholds are optimised for the reduction of expected losses under explicit cost parameters; (2) defence in depth—fraud analytics are fused with cyber telemetry and enforced through zero-trust controls; (3) governability—artefacts, tests, and monitoring comply with NIST AI RMF 1.0 and satisfy SR 11-7/OCC 2011-12; (4) privacy by design—federated learning and differential privacy are used for cross-institution collaboration; and (5) operational realism—streaming features, label latency, and drift are treated as first-class constraints (Federal Reserve, 2011; OCC, 2011; NIST, 2020, 2023).

Data and feature layer. Ingestion captures payment and identity events—card authorization, clearing, ACH, wires, P2P, authentication logs, device telemetry—via append-only streams, partitioned by entity keys, and time-stamped with synchronized clocks. Normalization ensures consistent units and categorical vocabularies across sources. Privacy controls tokenize sensitive identifiers, enforce role-based access, and maintain immutable lineage metadata. A feature store materializes low-latency aggregates per sliding windows (5-minute, 1-hour, 24-hour) for velocities, spending dispersion, merchant entropy, device diversity, peer risk statistics, and graph motifs. The feature code is shared between training and serving to prevent skews. Quality monitors track null ratios, distributional drift, and join coverage; alerts feed model monitors and data reliability engineering.

Modeling and decisioning. The primary stack consists of a gradient-boosted tree model for robust baseline scoring, a compact sequence model that ingests most recent events, and a graph-risk component. Trees use monotone constraints and interaction controls on sensitive fields to support interpretability and stable partial effects. The sequence model consumes transaction deltas and categorical embeddings and is trained with truncated backpropagation to meet latency constraints. The graph component is computed offline per batch or via streaming updates using incremental neighborhood statistics; a lightweight graph neural network can be used where infrastructure permits. The three scores are combined by a policy engine that applies channel-specific thresholds and step-up actions—strong customer authentication, hold for review, or deny. Actions are logged with human-readable rationales to support adverse-action explanations and examiner review (Rudin, 2019).

Training protocols. Training uses time-based splits that respect causality; features are frozen relative to the decision time to avoid leakage. Class imbalance is handled with sample weighting, calibrated probabilities, and optionally SMOTE in the training fold. Hyperparameters are selected by Bayesian optimization on the area under the precision–recall curve with constraints on inference latency. Calibration is evaluated with reliability diagrams within customer segments. Where labels are delayed, semi-supervised pretraining and self-supervised sequence tasks bootstrap representation learning, while

active learning sends uncertain cases to analysts for rapid adjudication (Chawla et al., 2002; Saito & Rehmsmeier, 2015).

Evaluation and thresholding. Evaluation prioritizes precision–recall curves, recall at fixed precision targets, and cost-based selection. Let p be the fraud prevalence, L the average loss per fraud if undetected, and c the review cost per alert. For a threshold with precision π and recall r, the expected net benefit over N transactions is: Benefit = $N(r\ p\ L) - N((r\ p/\pi)\ c) - N((1 - r)\ p\ L)$. The optimal threshold maximizes this quantity, which is subject to fairness and customer-experience constraints. Figure 1 and Figure 2 illustrate precision–recall behavior and a net-benefit curve under simple assumptions, while Table 3 shows simulated performance and latency for three model families (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015).

Cyber telemetry fusion and zero trust. Fraud models consume cyber risk signals: impossible-travel flags, device health attestations, token binding failures, browser integrity, and anomalous API usage. A session risk score gates authentication and authorization decisions per FFIEC layered security expectations, while NIST SP 800-207 mandates continuous verification and segmentation to limit blast radius. Model endpoints and feature APIs are authenticated using strong service identity; policy decisions are enforced at sidecars or API gateways. Incident response runbooks integrate model drift and alert surges as triggers that page security operations and risk teams (FFIEC, 2011, 2021; NIST, 2020).

Privacy-preserving collaboration. For cross-institution patterns—synthetic identity rings, compromised merchant clusters—we employ federated learning with secure aggregation so that each bank trains locally and transmits encrypted, aggregated updates. A differential-privacy mechanism injects calibrated noise into updates or into published risk indicators, bounding the contribution of any individual. Governance records privacy budgets (ε, δ), update cadence, audit logs, and legal agreements restrict purpose and retention (Dwork, 2006; Abadi et al., 2016; McMahan et al., 2017; Bonawitz et al., 2017).

MLOps and governance. All artifacts—data schemas, features, code, models, thresholds—are versioned. Model inventory entries capture purpose, training data lineage, assumptions, and limitations. Validation tests conceptual soundness, sensitivity to feature perturbations, back-testing versus challengers, and stability under drift scenarios. Monitoring tracks AUC-PR, recall at target precision, calibration, population stability indices, and alert outcomes; guardrails halt deployment on anomalies. Documentation links model design to SR 11-7/OCC 2011-12, the NIST AI RMF 1.0 characteristics, and NYDFS/FFIEC control requirements (Federal Reserve, 2011; OCC, 2011; NIST, 2023; NYDFS, 2017; FFIEC, 2021).

Risk analytics and systemic view. Micro-level anomalies feed into portfolio loss models and network-aware stress tests. Exposure networks across correspondent banks, payment processors, and key vendors are modeled to simulate how a cyber-initiated fraud surge or processor disruption could propagate as liquidity stress. Basel and BIS frameworks inform capital and liquidity impacts, while network metrics identify critical nodes for contingency planning (BIS, 2016, 2017; Acemoglu et al., 2015; Bardoscia, Battiston, Caccioli, & Caldarelli, 2015).

Implementation roadmap. Phase 1 establishes the feature store, a baseline boosted model, and precision–recall-based thresholding with explicit cost parameters; implements model inventory and basic monitoring; and integrates with authentication for step-ups. Phase 2 adds sequence and graph components, cyber telemetry fusion, and drift controls; it begins federated pilots with privacy accounting. Phase 3 connects anomaly signals to network-risk dashboards and sector sharing playbooks (CISA/JCDC), and hardens the supply chain with software bills of materials and SP 800-53 control evidence (CISA, 2022; NIST, 2020).

Fairness and consumer impact. Although fraud models aim to protect customers, they can unintentionally shift friction unequally across populations. We mitigate this by segment-specific calibration, monotonicity constraints on sensitive proxies, periodic fairness reviews, and human escalation paths. Adverse action communications explain proximate reasons when transactions are declined, and appeals processes collect additional context for rapid remediation. Documentation aligns with the fairness and transparency characteristics in the NIST AI RMF (NIST, 2023).

Security of the model pipeline. Model pipelines face security threats: data poisoning, evasion, and model extraction. Defensive measures include strict data provenance, canary features to detect poisoning, adversarial training for common evasion tactics, rate-limiting and watermarking for inference APIs, and red-team exercises scoped to models and features. Controls map to SP 800-53 families for configuration management, secure development, and incident response (NIST, 2020).

Table 1. Representative AI methods and properties.

Class	Representative methods	Strengths	Gaps / cautions	Key refs
Supervised	Gradient boosting (XGBoost/LightGBM/CatBoost); logistic	High accuracy with engineered features; fast scoring	Needs labels; sensitive to drift	Chen & Guestrin (2016); Ke et al. (2017); Prokhorenkova et al. (2018)
Temporal	LSTM/GRU sequence models	Captures behavior; complements trees	Statefulness; latency constraints	Jurgovsky et al. (2018)
Graph	Message-passing/attention on entity graphs	Detects rings and mules; inductive generalization	Entity resolution; evolving topology	—
Unsupervised	Isolation Forest; autoencoders; log-sequence models (DeepLog)	Works with sparse labels; online ready	Higher false positives; explainability	Liu et al. (2008); Du et al. (2017); Mirsky et al. (2018)
Training aids	SMOTE; focal loss; cost-sensitive learning	Handles imbalance; cost-aware training	Calibration; metric selection	Chawla et al. (2002); Lin et al. (2017)

Table 2. Governance and compliance alignment for the proposed stack.

Layer	Control objective	Example controls & artifacts	Key refs
Data & features	Lawful, least-privilege processing	Data classification; tokenization; access reviews; lineage	NIST SP 800-53; NYDFS 500
Modeling	Validated, explainable decisions	Model inventory; validation reports; challenger logs; drift dashboards	SR 11-7; OCC 2011-12; NIST AI RMF 1.0
Decisioning & authn	Layered controls & MFA	Risk-based step-ups; session risk; strong authentication	FFIEC (2021); NYDFS 500
Infra security	Zero-trust segmentation	Strong identity, micro-segmentation, continuous verification	NIST SP 800-207
Incident response	Timely detection & reporting	SOAR playbooks; 72-hour notifications; tabletop exercises	NYDFS 500; EO 14028
Sector coordination	Shared situational awareness	JCDC collaboration; Shields Up mitigations	CISA

Table 3. Simulated fraud model performance metrics.

Model	AUCPR (simulated)	Recall @ 10% Precision	FPR @ 95% Recall	P99 Latency (ms)
Logistic Regression	0.28	0.81	0.020	3.8
Gradient Boosting	0.46	0.92	0.012	12.5
Autoencoder	0.18	0.68	0.035	5.6

Discussion

Economic impact and operating posture. The simulated analysis in Figure 2 demonstrates how cost-aware thresholding maximizes net benefit under realistic review expenses and average fraud values. In practice, institutions should stratify thresholds by corridor and segment. Card-not-present flows, cross-border transfers, and new-to-bank segments support higher recall even with modest precision; low-risk domestic corridors may emphasize precision to reduce customer friction and false-positive fatigue. Precision–recall targets tied to explicit business constraints align model performance with operational service levels and consumer protection goals.

Customer experience and friction. Protective controls must minimize disruption to legitimate users. Blending probabilistic scores with policy rules enables stepped-up challenges only when the overall risk is elevated. Dynamic messaging and alternative verification paths reduce abandonment. Calibration must be segment-aware so that similarly risky users experience similar friction; where disparities appear, governance should require root-cause analysis and remediation plans.

Fusion with cyber telemetry. Fraud and cyber incidents share precursors such as compromised credentials, anomalous device posture, and automation patterns. Integrating session risk, token binding, and device health with transaction analytics supports adaptive authentication consistent with FFIEC layered security and NIST SP 800-207 zero-trust tenets. Conversely, sudden drifts in fraud score distributions or graph-risk topologies should trigger cyber incident response playbooks and sector notifications, shrinking attacker dwell time.

Collaboration and privacy. Cross-institution partnership is essential for detecting synthetic identities and merchant compromises that span multiple banks. Federated learning with secure aggregation and differential privacy enables applicable global models and risk indicators without exchanging raw personal data. Legal agreements should pin down permitted purposes, retention, auditability, and redress. Effective consortia also standardize schemas, entity resolvers, and evaluation protocols so that shared models are reproducible and interpretable across members.

Governance, explainability, and supervision. Examiners expect clear documentation of model purpose, conceptual soundness, data lineage, performance, limitations, and change history. For boosted trees, reason codes derived from constrained features can be mapped to actionable controls, while challenger models and back-testing guard against silent regressions. For deep and graph models, documentation should include architecture, training objectives, and stability analyses. The NIST AI RMF offers a helpful checklist for trustworthy AI characteristics, and SR 11-7/OCC 2011-12 provides the organizational scaffolding for independent validation and ongoing monitoring.

Resilience and systemic risk. AI-detected anomalies are not only operational signals; they are also early indicators of systemic stress. Embedding alerts into network-aware dashboards that visualize dependencies among banks, processors, and shared service providers can highlight critical nodes where mitigation or traffic shaping will have the most

significant macro effect. Scenario design should include blended events—payment-processor outages coinciding with peak shopping, plus targeted phishing—so that executive teams can exercise decision rights and contingencies before a crisis. BIS standards and network research provide calibration points for translating operational disruption into liquidity and capital impacts.

Human-in-the-loop operations. Skilled investigators remain central to sustainable performance. Well-designed queues prioritize cases by the marginal value of information, and active learning routes ambiguous alerts for rapid labeling. Reviewer guidance, rationale capture, and appeal workflows improve decision consistency and create training data with lower noise. Operational analytics should track capacity, backlog aging, duplicate alerts, and the stability of key features to prevent label drift and to focus engineering on real bottlenecks.

Security of models and data. Attackers increasingly target data pipelines and model endpoints. Defenses include strong service-to-service authentication, encryption in transit and at rest, supply-chain hardening with software bills of materials, dependency pinning, and automated patching. Adversarial testing should probe evasion (e.g., small feature changes), poisoning (e.g., tainted feedback loops), and extraction (e.g., query-based stealing). Findings flow into mitigation backlogs and validation reports under SR 11-7 and SP 800-53 control evidence.

Pitfalls and trade-offs. Standard failure modes include training–serving skew from divergent feature code, label leakage from post-decision fields, and miscalibrated scores due to nonstationary sampling. Overly aggressive thresholds can inflate costs and customer attrition even when headline recall improves. Conversely, under-triggered models can allow fraud spikes that exceed operational response capacity. Institutions should adopt decision reviews that consider not only AUC-PR but also time-to-detect, alert growth rates, analyst throughput, and downstream recovery. Business cases must include avoided losses, reduced manual reviews, and improved customer retention.

Policy implications. Regulators and standard setters can accelerate safe adoption by publishing reference taxonomies for fraud events and evaluation metrics; by encouraging privacy-preserving collaboration through clear safe harbors; and by aligning incident reporting so that cyber and fraud signals flow rapidly to sector partners. Shared exercises through CISA's JCDC and Financial ISACs can help prioritize defenses around critical nodes and validate that cross-firm mitigations work under pressure. Together, these steps translate technical progress in AI into durable public benefits.

Conclusion

AI, when deployed with rigorous governance and security controls, can materially strengthen U.S. financial infrastructure. Ensemble trees, sequence learning, and graph analytics jointly surface individual and networked fraud patterns; precision–recall

optimization aligns model thresholds with economic value and customer experience; and privacy-preserving collaboration extends coverage beyond a single institution's field of view. Yet model risk management, explainability, and secure engineering are non-negotiable: SR 11-7/OCC 2011-12, NIST SP 800-53 and SP 800-207, and the NIST AI Risk Management Framework provide practical scaffolding for trustworthy, hardened systems. At the ecosystem level, integrating AI alerts into network-based risk analytics and coordinating through CISA programs elevates sector resilience against multi-vector campaigns. The blueprint presented here—architecture, cost-aware evaluation, governance mapping, and sector coordination—offers a pragmatic path to reduce fraud losses, contain intrusions, and curb systemic propagation. Continued investment in open standards, privacy-preserving methods, adversarial robustness, and workforce training will help ensure that AI enhances security and public trust at the pace of innovation.

Limitations and Future Directions

This paper synthesizes literature and proposes an architecture with simulated evaluation figures; it does not publish proprietary transaction data or head-to-head benchmarks. While the methods and governance align with pre-February 2023 guidance, rapidly evolving threats and tools require ongoing validation against current attacker tradecraft and regulatory updates. Performance depends on data quality, entity resolution, and operational maturity; many institutions face constraints in feature engineering across siloed systems and third parties. We emphasize precision–recall-based evaluation, but organizations should perform complete cost modeling, including customer friction, chargeback cycles, and downstream recovery operations—parameters that vary materially by product and channel. Privacy-preserving collaboration introduces communication and accuracy trade-offs; its effectiveness hinges on harmonized schemas, legal agreements, and secure aggregation protocols. From a security perspective, we note—without formal experiment—susceptibility to data and model attacks (poisoning, evasion, extraction) and the need for robust MLOps security controls under Zero Trust and SP 800-53 families. Future work should validate the reference stack on public or consortial datasets with reproducible pipelines; evaluate adversarial robustness tailored to fraud and identity signals; advance drift-aware, self-supervised pretraining that reduces label latency dependence; refine interoperable data sharing (governed identifiers, privacy budgets, auditable APIs) across institutions; integrate AI alerts with network systemic-risk dashboards for joint exercises with sector partners; and assess human-factors design for analysts and customers to minimize false-positive fatigue and friction.

Figures & Tables

Figure 1. Precision–Recall curves (simulated) comparing logistic regression, gradient boosting, and autoencoder models on an imbalanced transactions setting. See Saito & Rehmsmeier (2015) and Davis & Goadrich (2006) for metric guidance.

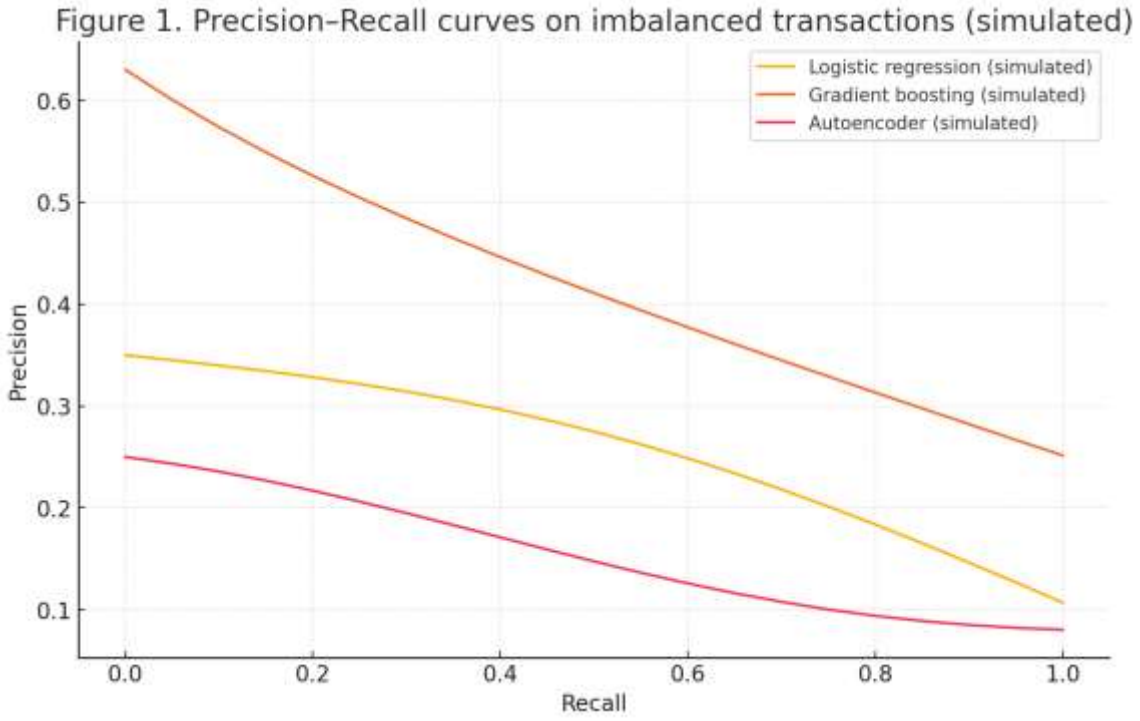

Figure 2. Net economic benefit versus recall under simple assumptions (10M transactions, 0.2% base fraud rate, $250 average loss, $5 review). The "knee" indicates a favorable operating point that balances preventing loss and review cost.

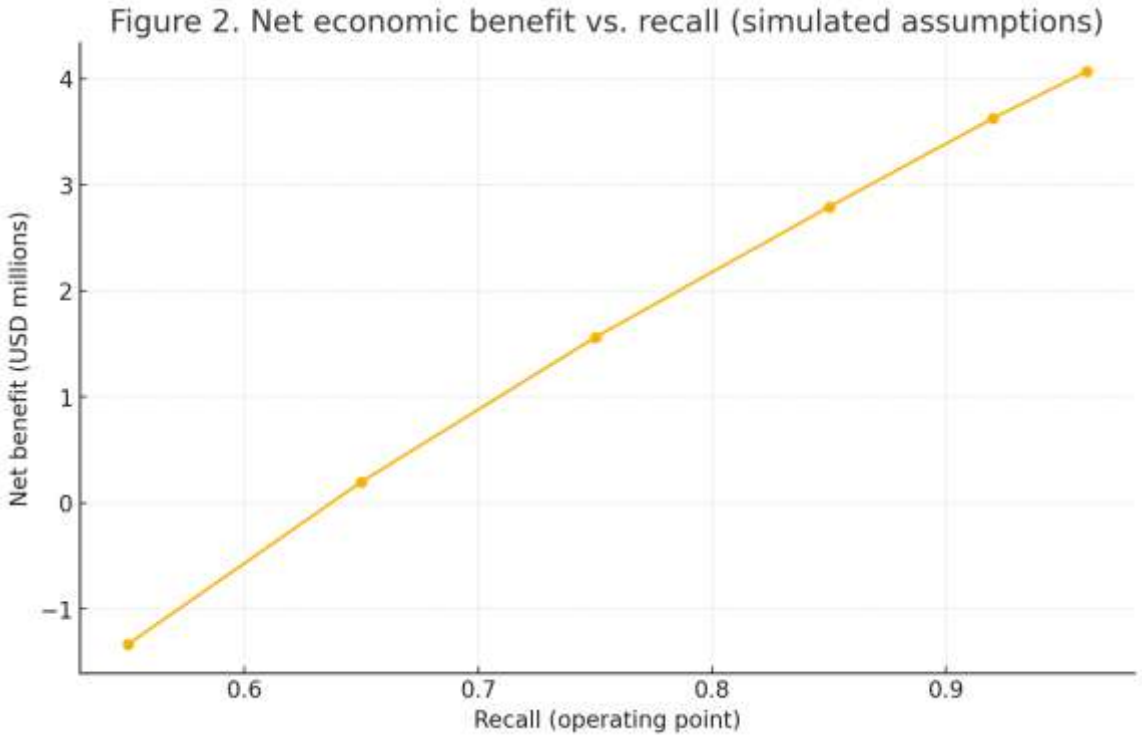

Table 3. Simulated fraud model performance metrics.

Model	AUCPR (simulated)	Recall @ 10% Precision	FPR @ 95% Recall	P99 Latency (ms)
Logistic Regression	0.28	0.81	0.02	3.8
Gradient Boosting	0.46	0.92	0.012	12.5
Autoencoder	0.18	0.68	0.035	5.6

References

1. 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.
2. Acemoglu, D., Ozdaglar, A., & Tahbaz-Salehi, A. (2015). Systemic risk and stability in financial networks. American Economic Review, 105(2), 564–608.
3. Bardoscia, M., Battiston, S., Caccioli, F., & Caldarelli, G. (2015). DebtRank: A microscopic foundation for shock propagation. PLOS ONE, 10(6), e0130406.
4. Bank for International Settlements (BIS). (2016). Minimum capital requirements for market risk. Basel Committee on Banking Supervision.
5. Bank for International Settlements (BIS). (2017). Basel III: Finalising post-crisis reforms. Basel Committee on Banking Supervision.
6. Bonawitz, K., et al. (2017). Practical secure aggregation for privacy-preserving machine learning. In Proceedings of CCS 2017.
7. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of KDD 2016.
8. 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.
9. CISA. (2022). Shields Up. Cybersecurity and Infrastructure Security Agency.
10. CISA. (2021). Joint Cyber Defense Collaborative (JCDC). Cybersecurity and Infrastructure Security Agency.
11. Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In Proceedings of ICML 2006.
12. Du, M., Li, F., Zheng, G., & Srikumar, V. (2017). DeepLog: Anomaly detection from system logs via deep learning. In Proceedings of CCS 2017.
13. Dwork, C. (2006). Differential privacy. In Proceedings of ICALP 2006.
14. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407.
15. Federal Financial Institutions Examination Council (FFIEC). (2011). Supplement to Authentication in an Internet Banking Environment.
16. FFIEC. (2021). Authentication and access to financial institution services and systems.
17. FFIEC. (2020–2021). BSA/AML Examination Manual.
18. Federal Reserve. (2011). SR 11-7: Guidance on model risk management.
19. Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P.-E., He-Guelton, L., & Caelen, O. (2018). Sequence classification for credit-card fraud detection. Expert Systems with Applications, 100, 234–245.
20. Ke, G., Meng, Q., Finley, T., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Proceedings of NeurIPS 2017.
21. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In Proceedings of ICDM 2008.
22. Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of ICCV 2017.

23. Mirsky, Y., Doitshman, T., Elovici, Y., & Shabtai, A. (2018). Kitsune: An ensemble of autoencoders for online network intrusion detection. In Proceedings of NDSS 2018.
24. NIST. (2020). SP 800-207: Zero Trust Architecture. National Institute of Standards and Technology.
25. NIST. (2020). SP 800-53 Rev. 5: Security and privacy controls for information systems and organizations. National Institute of Standards and Technology.
26. NIST. (2023). AI Risk Management Framework 1.0. National Institute of Standards and Technology.
27. New York State Department of Financial Services (NYDFS). (2017). 23 NYCRR 500: Cybersecurity requirements for financial services companies.
28. Office of the Comptroller of the Currency (OCC). (2011). Bulletin 2011-12: Sound practices for model risk management.
29. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In Proceedings of NeurIPS 2018.
30. Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.
31. Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems, 29(8), 3784–3797.
32. Dal Pozzolo, A. (2015). Adaptive machine learning for credit card fraud detection (Doctoral dissertation). Université Libre de Bruxelles.
33. Bardoscia, M., Battiston, S., Caccioli, F., & Caldarelli, G. (2015). DebtRank: A microscopic foundation for shock propagation. PLOS ONE, 10(6), e0130406.