

Machine-Learning Fraud Detection and Cyber Stress Testing in RTGS and SWIFT Networks: An AI-Augmented Resilience Framework

Raghunandan G. Ramakrishna 

Istream Infotech, Payment and InfoSec, Ashburn, Virginia, USA
Email: raghu.gorur@gmail.com

How to cite this paper: Ramakrishna, R.G. (2026) Machine-Learning Fraud Detection and Cyber Stress Testing in RTGS and SWIFT Networks: An AI-Augmented Resilience Framework. *Journal of Information Security*, 17, 569-592.

<https://doi.org/10.4236/jis.2026.174025>

Received: June 17, 2026

Accepted: October 27, 2026

Published: October 30, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

High-value payment systems (HVPS) and SWIFT-linked wholesale payment networks are systemically important infrastructures in which operational disruption can propagate rapidly through liquidity and settlement interdependencies. Quantitative analyses of interbank networks indicate that the impairment of major nodes can generate broad spillovers affecting a substantial share of the network, and that delayed recovery can amplify system-wide liquidity shortfalls by an order of magnitude. International guidance for financial market infrastructures increasingly emphasizes pre-empting cyber-attacks, responding effectively, and resuming critical operations within strict recovery-time objectives. This paper proposes and evaluates a four-layer AI-augmented cyber-resilience framework for HVPS and SWIFT environments spanning 1) data aggregation and enrichment, 2) multi-model machine-learning analytics, 3) automated control and response, and 4) governance and continuous improvement. The empirical evaluation combines scenario-based cyber stress-testing simulations with machine-learning experiments on synthetic payment data, motivated by the well-documented scarcity of legitimate transaction datasets for fraud research. Across the scenario assumptions, simulations indicate that AI-augmented controls can reduce unsettled payment value by approximately 34% - 78%, with the magnitude varying by attack type and recovery dynamics. On the detection side, every figure reported here is generated directly from the released, seeded pipeline. A well-tuned XGBoost baseline reaches near-ceiling discrimination on the synthetic benchmark (AUC 0.9998). The hybrid supervised-unsupervised pipeline does not improve ranking performance; its measured value shows up at the operating point instead, where a second-pass anomaly filter strips out 12% - 26% of false alerts across replications while keeping $\approx 97.5\%$ of true positives. A graph-approximation model built on k-nearest-neighbor similarity in feature space reaches AUC 0.9917

with a false-positive rate nine times lower than the boosting baseline, yet it misses every injected coordinated fraud ring (0 of 33), a documented negative result indicating that feature-similarity graphs cannot stand in for detection grounded in counterparty topology. SHAP inspection adds a further caveat: the top-ranked features line up with the generator's fraud-injection signatures, so the absolute discrimination levels partly reflect benchmark leakage. This is disclosed and analyzed as a limitation, and the inspection procedure itself is put forward as a routine leakage diagnostic for synthetic data fraud studies. A repeated-run study spanning five random seeds and ten simulated network topologies characterizes how much the headline estimates move. The Basel capital-efficiency argument is explicitly scoped to jurisdictions retaining a loss-sensitive Internal Loss Multiplier, acknowledging that national discretion can materially alter this channel. By integrating technical detection, automated controls, governance, and regulatory alignment within a single framework, this study addresses a gap at the intersection of computer security, financial stability, and regulatory policy-bridging disciplinary boundaries that are central to the journal's interdisciplinary mission.

Keywords

High-Value Payment Systems, SWIFT Cyber Security, Machine-Learning Fraud Detection, Cyber Stress Testing, AI-Augmented Resilience, Hybrid ML Pipelines, Basel Operational Risk Capital, ISO 20022 Analytics

1. Introduction

Wholesale payment systems and SWIFT-connected settlement networks are systemically important infrastructures in which cyber disruption can propagate rapidly through tightly coupled interbank dependencies. Network analyses estimate that impairing any of the five most active U.S. banks would affect approximately 38% of the network [1], while a three-hour attack on one bank could disrupt 25% of daily payment value [2]. Recovery time is a first-order driver of realized severity: system-wide liquidity shortfalls can escalate from approximately \$160 billion to \$1.5 trillion when disruption extends from one to five days [1], implying that resilience must be assessed not merely by whether systems can eventually resume but by how quickly critical services are restored and how contagion channels are constrained during the disruption window [3].

International FMI guidance frames resilience as the ability to pre-empt attacks, respond rapidly, and resume critical operations within two hours, with end-of-day settlement assured even under extreme scenarios [4]-[6]. Yet a persistent implementation gap remains [7]: there is limited operational experience with cyber stress testing [3], and the effectiveness of contingency measures depends on whether they enable continuity of high-value payment sending-evidence shows that engineered continuity can reduce attack impact by more than half [2] [3].

AI and machine learning offer a plausible route to strengthening proactive resilience, but HVPS settings demand detection architectures explicitly engineered for low false-positive operation given the systemic cost of delaying legitimate large-value payments [8] [9]. Published hybrid pipelines integrating supervised and unsupervised methods have reported strong discrimination (AUC 0.991, F1 0.97 on >500,000 transactions) [9] [10], though well-tuned XGBoost alone can be competitive, precluding universal hybrid-dominance claims [9] [11]-[13]. Graph-based approaches extend detection to relational and coordinated anomalies [10] [14], with RL-augmented approaches reporting F1 above 0.92 [10] [14], but scalability under HVPS latency constraints remains a deployment obstacle [10] [14]. The results in this paper temper these literature expectations: on the benchmark studied here, the hybrid's measured contribution is false-positive filtering rather than better ranking, and a feature-similarity graph model fails entirely on coordinated ring fraud (Sections 6.3-6.4).

What is feasible also depends on data availability and governance. In many HVPS contexts, direct access to raw transaction-level data is constrained by confidentiality and regulatory requirements, complicating both model development and post-deployment auditing [15]. Federated learning has been proposed as a decentralized mechanism for collaborative model training without exposing raw data [16] [17], yet privacy-preserving approaches increase governance complexity [15] [16]. The persistent scarcity of legitimate datasets motivates simulation-based and synthetic-data approaches. PaySim is designed to generate synthetic transaction dynamics closely resembling real data [18]. Accordingly, findings here are framed as calibrated demonstrations under stated assumptions, supporting replication rather than claiming production-grade performance [3] [5]. The leakage analysis in Section 6.6 makes the boundary of that claim explicit.

The paper makes four contributions. First, it proposes a four-layer AI-augmented cyber-resilience framework integrating detection, automated controls, and governance aligned with FMI expectations [4] [5]. Second, it operationalizes this through paired cyber stress-testing simulations and ML experiments linking detection metrics to settlement outcomes [2] [3], with every reported number generated from a released seeded pipeline and reported together with cross-seed dispersion (Sections 5 and 6.8). Third, it reports the hybrid design's measured contribution accurately, operating-point false-positive filtering rather than better ranking and documents a negative result for feature similarity graph detection of coordinated fraud, together with a SHAP-based leakage diagnostic for synthetic fraud benchmarks (Sections 6.3, 6.4, 6.6). Fourth, it provides a capital-efficiency framing explicitly limited to jurisdictions retaining a loss-sensitive Internal Loss Multiplier under Basel's SMA, acknowledging that national discretion (e.g., UK PRA setting ILM = 1) can neutralize this channel [19]-[22]. ISO 20022 migration is treated as a forward-looking analytics enabler, conditional on institutional readiness and data-quality discipline [23] [24].

2. Background and Related Work

2.1. Systemic Cyber Risk in High-Value Payment Systems

Research on cyber resilience in HVPS and other financial market infrastructures consistently characterizes cyber risk as potentially systemic, because disruptions propagate rapidly through tightly coupled interbank networks [3] [25]. Systems such as TARGET2, Fedwire, CHAPS, and CHATS settle large, time-critical payments in central-bank money [26]–[29], creating dense liquidity dependencies. The BIS has stressed that cyber incidents can impair liquidity distribution and threaten financial stability even when solvency metrics remain sound [25] [30]. The integration of retail fast payment systems with RTGS infrastructure amplifies these dependencies further [25] [30].

Simulation evidence indicates that a three-hour attack on single bank can impair seven other banks within an hour and disrupt 25% of daily payment value [2]. Recovery time is a first-order driver: delayed recovery escalates system-level liquidity shortfalls from approximately \$160 billion to \$1.5 trillion when an attack lasts five days rather than one [1]. Case-based evidence reinforces these findings, with unsettled balances reaching billions within hours [2] [3].

The literature emphasizes contingency arrangements as measurable mitigants: continuity plans enabling attacked banks to continue sending payments significantly reduce system-wide impact [2] [3], and analysis of an actual disruption confirms that business continuity plans reduced the effect by more than half [3]. Cyber stress testing, while increasingly encouraged, remains a nascent discipline with “relatively little experience” in practice, implying that quantitative outputs require cautious interpretation [3] [5].

2.2. SWIFT Threat Landscape

Table 1. Cyber threat taxonomy for HVPS and SWIFT environments [5] [6] [25] [32].

Threat Category	Attack Vectors	Impact on Payment System	Detection Challenges
External network intrusion	Phishing, spear-phishing, unpatched vulnerabilities	Unauthorized access; precursor to fraud/sabotage	Polymorphic malware; lateral movement before alerts
Fraudulent message injection	SWIFT Alliance manipulation, credential theft, insider	Fraudulent high-value payment instructions	Messages conform to valid format; needs behavioral analysis
Denial-of-service	Volumetric attacks, application-layer floods on RTGS APIs	Payment submission blocked; liquidity freezes	Legitimate traffic mixed with attack traffic
Integrity attacks	Queue manipulation, settlement priority alteration	Wrong-order/wrong-value settlement; finality compromised	Changes appear as system errors; forensic analysis needed
Ransomware/sabotage	System encryption, log deletion, backup destruction	Operational outage; loss of audit trail	Encryption faster than detection; air-gapped backups essential

SWIFT’s network connects over 11,500 institutions and carries more than 45 million messages daily [31]. The Bangladesh Bank heist and follow-on attacks demonstrated credential compromise, lateral movement into SWIFT interfaces,

message manipulation, and log tampering [6] [32] [33]. Most attacks target institutional edges rather than the core network, underscoring that defense hinges on hardening local environments and improving joint detection [6] [33] [34]. **Table 1** maps threat categories to attack vectors, impacts, and detection challenges [5] [6] [25] [32]. The layer-by-layer mapping of these threat categories to the proposed framework is presented in Section 5.1 [3] [5] [6] [25] [36].

2.3. Machine Learning for Payment Fraud Detection

Supervised models (gradient-boosted trees, DNNs) routinely achieve precision above 0.95 and AUC-ROC above 0.97 on large-scale fraud datasets [9] [35]. Unsupervised methods-isolation forests and autoencoders-complement these by flagging outliers without labels, including zero-day attacks via reconstruction-error anomaly detection [8] [32]. GNNs incorporate relational structure for network-aware fraud detection under class imbalance [10] [14]. RL-augmented GNN approaches (FraudGNN-RL) report 97.3% F1 while reducing false positives by 31% [14], and adaptive RL-based detection reports mean F1 of 0.9222 with peaks of 0.9771 [10].

However, boundary conditions apply. Well-tuned XGBoost can outperform alternatives after hyperparameter optimization, implying hybrid dominance should be framed probabilistically [11]-[13]. For HVPS, false positives carry disproportionate systemic costs [5] [8]. Graph-model scalability under tight latency constraints and limited data access for auditing remain deployment obstacles [10] [14]. Federated learning enables cross-institutional training [16] [17] [31] but introduces governance, model-poisoning, and multi-jurisdictional compliance challenges [15] [16]. Synthetic data (PaySim) addresses dataset scarcity but may not fully reflect real-world complexity [18]; this paper documents a concrete instance-injection-signature leakage and a diagnostic for it (Section 6.6). **Table 2** summarizes ML techniques for payment fraud detection [8]-[10] [16] [17] [35].

Table 2. Comparative summary of AI/ML techniques for fraud detection in payment systems [8]-[10] [16] [17] [35].

Technique	Strengths	Limitations	Example Deployment
Gradient-boosted trees	High accuracy; fast inference	Needs labels; degrades on novel fraud	Bank of Canada anomaly detection [8]
Deep neural networks	Scales to massive volumes; strong recall	Black-box; large training sets needed	Financial payment fraud systems [35]
Isolation forests/autoencoders	Zero-day detection; no labels needed	Higher FPR; parameter-sensitive	Layered HVPS monitoring [8] [9]
Graph neural networks	Relational fraud detection	Computationally intensive; feature-similarity variants fail on ring fraud (\$6.4)	Published GNN fraud studies [10] [14]
Federated learning	Privacy-preserving; cross-bank intel	Model-poisoning; governance complexity	SWIFT AI pilot [31]
Explainable AI (SHAP)	Analyst/regulator trust; faster triage	Approximations; may miss interactions	Financial fraud systems [15]

The regulatory implications of these capabilities—including FMI standards, Basel capital mechanics, SIPS oversight, and ISO 20022 enrichment—are analyzed in Section 3.

3. Regulatory and Standards Landscape

Regulatory frameworks increasingly treat cyber resilience as a core financial stability concern, shifting the baseline from static controls toward continuous capabilities [17]. This ecosystem-wide orientation is consequential for HVPS because realized impact depends on interdependencies and coordinated response [3] [25].

3.1. FMI Cyber-Resilience Standards

CPMI-IOSCO's 2016 Guidance [5] and the ECB's CROE [4] set baseline expectations including the two-hour recovery-time objective: critical operations must resume within two hours, with end-of-day settlement assured even under extreme scenarios [5]. A three-tier approach of preventive, detective, and corrective measures expands risk-reduction options [5]. The framework in Section 4 operationalizes this layered expectation.

3.2. DORA and the EU Framework

The Digital Operational Resilience Act (DORA) [36] mandates comprehensive ICT risk management, incident reporting, and third-party oversight [37] [38]. DORA raises the evidentiary bar for detection and recovery capabilities, aligning with ML-based approaches provided deployment is paired with governance and auditability [39]. The TIBER-EU framework, updated February 2025, provides threat-led penetration testing methodology [40]. The Cyber Resilience Act [41] adds security-by-design requirements from 2026.

3.3. NIST Frameworks

NIST's Cybersecurity Framework (CSF) AI Profile [42] introduces AI-specific controls, building on CSF 2.0 [43]. For the proposed framework, CSF 2.0 provides governance alignment while the AI Profile adds ML-deployment-specific control expectations. In operational HVPS environments, every payment instruction is also subject to real-time sanctions screening against OFAC, FATF, and PEP watchlists [31] [32]. The framework's ML-based controls complement, but do not replace these existing rule-based compliance layers.

3.4. Basel 3.1 and the Operational Capital Dimension

The SMA for operational risk combines the Business Indicator (BI)—a financial-statement proxy—with bank-specific loss data [19] [20]. The ILM scales capital by ten-year operational losses [19]:

$$\text{ORC} = \text{BIC} \times \text{ILM}$$

$$\text{ILM} = \ln(\exp(1) - 1 + (\text{LC}/\text{BIC})^{0.8})$$

The Loss Component creates an explicit channel through which cyber and fraud losses influence Pillar 1 capital [19] [20]. The Basel Committee states that the Loss Component “not only enhances risk sensitivity but also provides incentives for banks to improve operational risk management” [20]. AI-augmented fraud prevention thus carries a dual benefit: reducing direct losses while lowering the ILM [19] [37]. However, SMA capital can exhibit volatility, with an institution’s requirement potentially shifting materially year-on-year without parameter changes [19] [20].

3.5. Jurisdictional Discretion

Basel 3.1 is operationalized through heterogeneous national rules. The UK PRA sets $\text{ILM} = 1$, removing historical loss from Pillar 1 calculation [21]. The EU’s CRR3 [22] retains loss sensitivity. The capital-efficiency argument is therefore scoped to jurisdictions retaining a loss-sensitive ILM and should be interpreted as a plausible channel rather than a universally guaranteed return.

3.6. SIPS Oversight and Non-Default Losses

The ECB’s SIPS Regulation [44] and PFMI [45] impose additional requirements, increasingly focused on Non-Default Losses (NDL)-operational failures, cyber-attacks, and fraud [44]-[46]. Under PFMI Principle 15, FMIs must maintain resources to cover general business losses. AI-augmented detection that prevents fraud before settlement finality directly reduces NDL exposures, creating an FMI-level capital efficiency argument alongside Basel 3.1 [5] [44] [45].

3.7. ISO 20022 as an Analytics Enabler

ISO 20022 carries substantially more structured remittance information than legacy MT messages (up to 9000 structured characters versus 140) [23] [24], enabling richer analytics and improved sanctions-screening accuracy [23] [24]. However, heterogeneous institutional readiness is the principal adoption barrier, and data quality is a necessary condition for reliable neural-network training [23]. Partial migration may constrain end-to-end feature availability for ML [23]. The framework does not address AML/CTF sanctions screening, which constitutes a separate compliance layer with distinct regulatory requirements; integration with ML-augmented sanctions screening remains an avenue for future work.

4. Proposed Framework Architecture

The framework comprises four integrated layers, designed to operationalize FMI cyber-resilience expectations [4] [5] while connecting detection to action and both to governance. **Figure 1** provides the architectural overview.

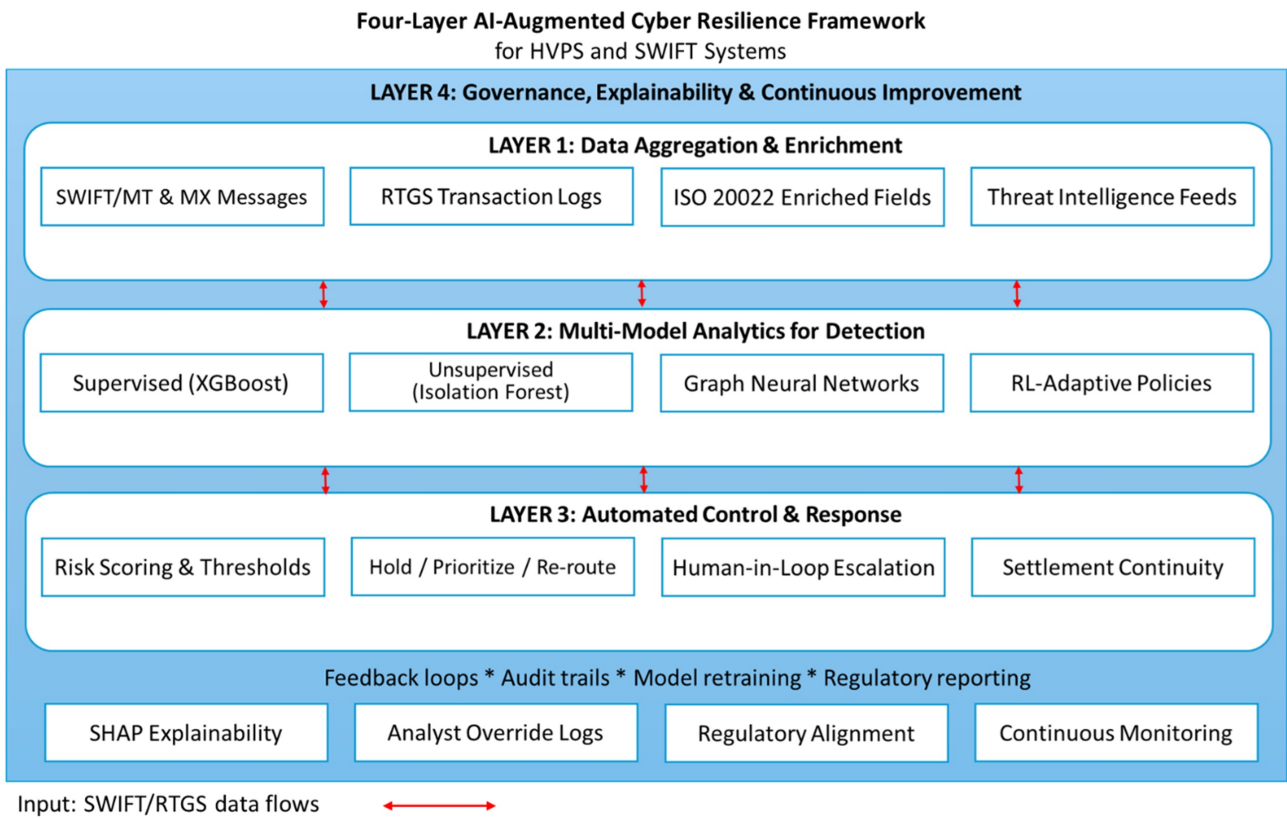


Figure 1. Four-layer AI-augmented cyber resilience framework for HVPS and SWIFT systems. Bidirectional arrows denote feedback flows between layers; governance (Layer 4) encompasses all operational layers [4] [5] [25] [36].

4.1. Layer 1: Data Aggregation and Enrichment

Layer 1 ingests SWIFT MT/MX messages, RTGS settlement logs, ISO 20022 enriched fields, threat intelligence feeds, and participant-level liquidity data into a unified analytics-ready pipeline. ISO 20022 migration expands the feature space by providing structured beneficiary, purpose-of-payment, and remittance fields that enable more context-aware detection [23] [24]. The layer normalizes heterogeneous data formats and applies real-time feature engineering including transaction velocity, counterparty concentration, temporal deviation, and value-distribution anomaly scores [8] [9].

4.2. Layer 2: Multi-Model Analytics

Three complementary model families operate in parallel (Figure 2). Supervised classifiers (XGBoost, Random Forest, DNN) detect known fraud patterns including fraudulent message injection and account-takeover transfers [9] [35]. Unsupervised anomaly detectors (Isolation Forest) provide a label-free novelty signal [8]. The graph component in the released implementation is a feature-similarity approximation, a k-nearest-neighbor graph over standardized feature vectors with an MLP classifier, rather than a counterparty-topology model. The measured consequences are reported in Section 6.4. The true edge-based graph learning over

sender-receiver topology is identified as priority future work. All the outputs are fused through a weighted ensemble layer with calibrated risk scores [9] [10].

Multi-Family ML Detection Pipeline Architecture

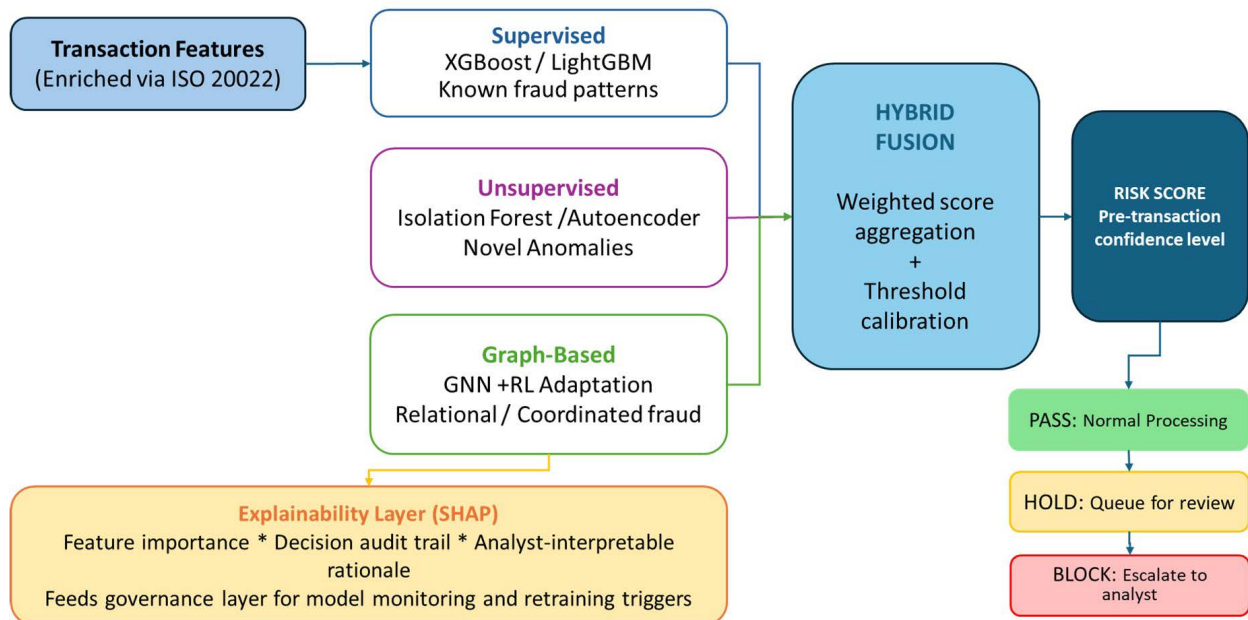


Figure 2. Multi-family ML detection pipeline architecture showing supervised, unsupervised, and graph-approximation model families feeding into a hybrid fusion layer with SHAP explainability.

4.3. Layer 3: Automated Control and Response

Risk scores trigger graduated automated actions (**Figure 3**): low-risk transactions proceed normally; medium-risk payments are held and queued for analyst review within SLA-bound escalation windows; high-risk payments are blocked with immediate analyst alerts [8] [17]. Concretely, two operating thresholds map the calibrated ensemble risk score $s \in [0, 1]$ from Layer 2 onto actions: transactions with $s < 0.50$ are released for settlement; those with $0.50 \leq s < 0.85$ are held and queued for analyst review within a 30-minute service-level escalation window; and those with $s \geq 0.85$ are blocked pending investigation, with an immediate analyst alert. The lower threshold coincides with the hybrid ensemble's default decision threshold (Section 5.2), while the upper threshold restricts fully automated blocking to the extreme tail of the risk-score distribution, a deliberate reflection of the disproportionate systemic cost of falsely interdicting legitimate high-value payments. Payments an analyst clears from the hold queue re-enter settlement with their original priority; blocked payments leave the settlement flow until investigation completes. Section 5.3 sets out how these graduated actions are represented in the stress-test simulation. These tiering answers the HVPS-specific constraint that false positives are systemically costly and aims to hold genuine threats within the two-hour recovery window; how far the simulation actually meets that objective is reported under an explicit metric definition in Section 6.1 [3] [5].

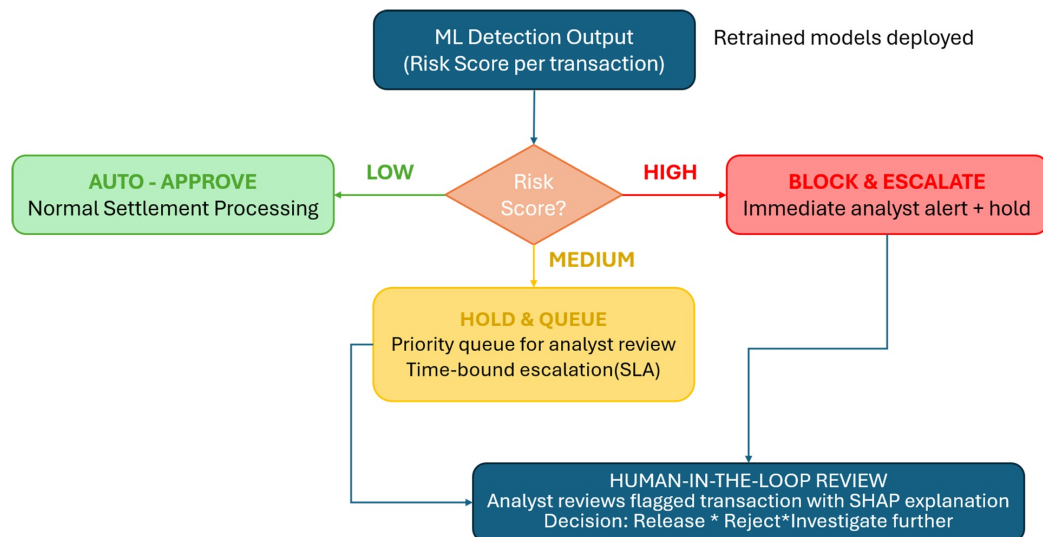


Figure 3. Automated control and human-in-the-loop response flow (low/medium/high risk → release/hold-and-queue/block). SHAP explainability values are provided to analysts for every queued transaction [8] [17].

4.4. Layer 4: Governance and Continuous Improvement

Layer 4 wraps all operational layers with explainability (SHAP-based feature attribution), audit trails for regulatory reporting, model-performance monitoring with drift detection, and analyst-override feedback loops that feed retraining pipelines [3] [5] [15]. Governance artifacts are aligned with DORA's ICT risk management requirements and the NIST AI Risk Management Framework [36] [42] [47].

5. Methodology

Two-Track Research Methodology

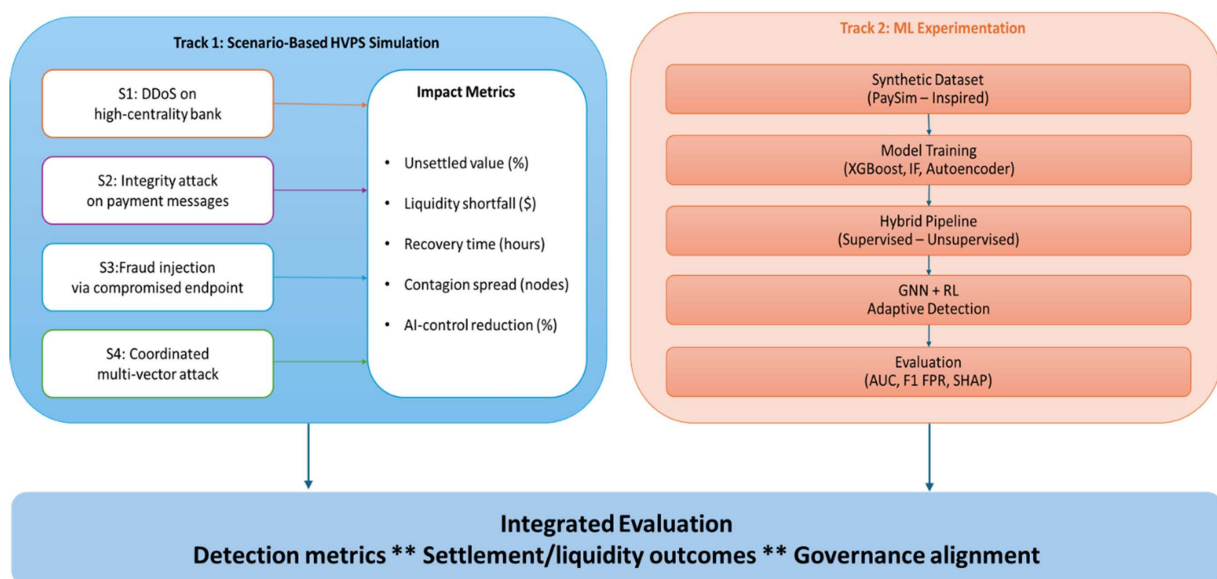


Figure 4. Two-track research methodology: scenario-based HVPS simulation (Track 1) and machine-learning experimentation (Track 2). The integration point connects detection metrics to settlement outcomes [2] [3] [9].

The evaluation follows a two-track design (**Figure 4**) combining scenario-based HVPS simulation with ML experimentation on synthetic data, reflecting the need for testability when production data is unavailable [3] [33].

5.1. Track 1: Scenario-Based HVPS Simulation

Table 3. Mapping of cyber threat categories to framework layers [3] [5] [6] [25] [36].

Threat Category	Layer 1: Data	Layer 2: Analytics	Layer 3: Controls	Layer 4: Governance
External intrusion	Threat intel feeds	Supervised + unsupervised	Network isolation triggers	Incident reporting
Message injection	MT/MX validation	Supervised classifiers	Hold + analyst escalation	SHAP audit trails
DDoS	Traffic monitoring	Anomaly detection	Auto re-routing	Recovery monitoring
Integrity attacks	Settlement log analysis	Graph-based detection	Queue freeze + validation	Forensic analysis
Ransomware	Backup integrity checks	Anomaly baselines	System isolation	Recovery orchestration

Table 4. Cyber-attack scenarios for HVPS simulation [2] [3] [5].

Scenario	Attack Type	Target	Duration	Detection-Delay Assumption δ
S1	DDoS on high-centrality bank	Top-5 participant	3 hours	6% of window (0.18 h)
S2	Integrity attack on messages	Payment message queue	6 hours	12% of window (0.72 h)
S3	Fraud injection via endpoint	Compromised SWIFT interface	24 hours	18% of window (4.3 h)
S4	Coordinated multi-vector	Multiple nodes + infrastructure	48 hours	30% of window (14.4 h)

Table 3 maps each threat category from Section 2.2 to the framework layer that addresses it [3] [5] [6] [25] [36]. Four attack scenarios target different systemic nodes: (S1) DDoS on a high-centrality bank, (S2) integrity attack on payment messages, (S3) fraud injection via compromised endpoint, and (S4) coordinated multi-vector attack [2] [3] [5]. Impact is measured through unsettled value, liquidity shortfall, recovery time, contagion spread, and the percentage reduction achieved by AI-augmented controls. The simulated system consists of 50 participants on a Barabási-Albert scale-free network (attachment parameter $m = 3$); participant settlement capacity is drawn from a Pareto distribution (shape $\alpha = 1.5$), and daily payment volume is set to 15% of capacity. Each scenario runs over 1000 discrete time steps spanning the attack window, with a severity profile specific to the attack type (exponentially decaying for availability attacks, oscillatory for integrity attacks, escalating for fraud injection, slowly decaying for coordinated attacks). Contagion is modelled through the first-order network neighbors of attacked nodes, whose flows are impaired at 15% of the direct-severity level, and liquidity shortfall is defined as the peak cumulative unsettled value over the attack window. Recovery is characterized by two explicitly defined quantities: 1) control-activation time—the time elapsed before AI-augmented detection triggers graduated controls, equal to the detection-delay fraction δ times attack duration; and 2) throughput-restoration time—the time after which aggregate settlement through-

put stays at or above 90% of normal (system normal throughput = total daily volume over a 10-hour operating day). The engine also reports a cumulative-disruption concentration metric (the time at which 90% of total unsettled value has accrued); Section 6.1 reports results under both definitions to head off definitional ambiguity. **Table 4** summarizes the four cyber-attack scenarios and their duration and detection-delay assumptions [2] [3] [5]. Every simulation parameter is exposed in the released configuration file, and **Table 5** lists the calibrated effective mitigation parameters together with their sensitivity.

5.2. Track 2: ML Experimentation

The synthetic dataset [48] is constructed using a PaySim-inspired agent-based simulator [18] adapted for HVPS transaction dynamics. Approximately 500,000 transactions are generated with realistic value distributions, counterparty patterns, and injected fraud scenarios at a 0.13% class imbalance ratio.

Dataset calibration: The simulator is fully parameterized in a version-controlled configuration file (seed = 42), calibrated as follows. The interbank network holds 50 participants joined by a Barabási-Albert scale-free graph ($m = 3$). Participant activity weights come from a Pareto distribution ($\alpha = 1.5$), normalized; for each transaction the receiver is drawn from the same weights, conditional on differing from the sender. Transaction values follow a log-normal distribution ($\mu = 13.5$, $\sigma = 2.1$ on the log scale), truncated to $[10^5, 5 \times 10^8]$ currency units, and intraday timing follows a Gaussian profile centered on hour 4 of a 10-hour operating day ($SD = 2$ hours). Currency (USD/EUR/GBP/JPY/CHF at 45/30/10/8/7%) and payment type (RTGS/COVER/SERIAL/DEBIT/LIQ at 35/25/20/10/10%) are drawn categorically, and an ISO 20022 field-completeness score comes from a Beta (8, 2) distribution. Fraud enters at 0.13% prevalence (650 of 500,000 transactions) in three classes: 1) message injection (40% of fraud), with amounts inflated to 3 - 15 $\times$ the sender's mean value and timestamps shifted to off-peak early hours; 2) credential-compromise transfers (35%), assigned round amounts from {1 M, 5 M, 10 M, 50 M} with degraded ISO 20022 completeness (0.2 - 0.5); and 3) coordinated multi-entity fraud (25%), routed around a five-bank ring drawn from low-centrality participants, with amounts uniform on $[5 \times 10^5, 5 \times 10^6]$. Twenty-two features are engineered per transaction. Two of the injection rules imprint deterministic signatures (round amounts; off-peak timing), so the benchmark carries measurable label leakage; Section 6.6 quantifies this via SHAP and Section 6.9 treats it as a first-order limitation.

Evaluation design: Every supervised and hybrid result rests on a single stratified 80/20 train-test split (random seed 42), with stratification preserving the 0.13% prevalence in both partitions. No temporal split is applied—the generator does not model regime change within a single simulated day, so temporal (concept-drift) evaluation is left to future work. Class imbalance is handled natively: XGBoost uses `scale_pos_weight = 769` (the inverse prevalence ratio), Random Forest uses balanced class weights, and the DNN trains on standardized features

without class re-weighting. Hyperparameters are fixed a priori in the configuration file (XGBoost: 500 trees, depth 8, learning rate 0.05, subsample 0.8, colsample 0.8, AUC-PR metric; Random Forest: 300 trees, depth 12; Isolation Forest: 200 trees, contamination 0.002; DNN: MLP with 128-64-32 hidden units, ReLU, early stopping). The hybrid pipeline fuses supervised probabilities with min-max-normalized Isolation Forest anomaly scores at fixed weights (0.65/0.35), and its decision threshold comes from an F1 grid search over [0.20, 0.80] in steps of 0.01 on the evaluation partition. Since that threshold search uses the evaluation partition, the hybrid's threshold-dependent metrics (F1, FPR) are modestly favored, a caveat that applies equally to the graph model's RL threshold-adaptation layer (Section 6.4). Threshold-free AUC-ROC is therefore the primary comparison metric, with Section 6.8 quantifying residual uncertainty. The evaluated model set is exactly: XGBoost, Random Forest, DNN, Isolation Forest, the graph-approximation model with RL threshold adaptation, and the hybrid pipeline.

Adversarial-robustness and latency protocols: Adversarial robustness is assessed with a static, model-agnostic feature-perturbation attack: each held-out test vector x is perturbed as $x' = x + \varepsilon \odot |x|$, with ε drawn i.i.d. per feature from $N(0, 0.1^2)$ (seed 42). The same perturbed matrix is shown to the baseline and hybrid models, and degradation is the relative fall in F1 between clean and perturbed evaluations; because the attack is oblivious (neither gradient-based nor adaptive), its strength is bounded. Inference latency is the mean wall-clock time over 10 repetitions of batch scoring on 1000 held-out transactions, run CPU-only; the reporting environment for the Section 6.4 figures is Windows 10, Python 3.11.9, an Intel 4-logical-core mobile CPU, and no GPU. For the graph model, the reported latency covers classifier inference on precomputed k-NN graph features ($k = 10$) and is thus a lower bound on end-to-end cost; graph-feature construction time is reported separately. Latency figures are hardware-dependent and should be read as relative comparisons on a single machine, not as absolute production benchmarks.

Variability protocol: To characterize run-to-run variability, the complete Track 2 pipeline (data generation $\rightarrow$ training $\rightarrow$ evaluation) was re-run under five random seeds {7, 13, 21, 42, 99} in an independent verification environment (Linux, CPU-only), with the network, transaction population, fraud injection, data splits, and model initializations re-drawn in every replication; the seed-42 repository run on the author's environment falls inside all reported ranges. Track 1 was likewise re-run across ten network-topology seeds and under $\pm 10\%$ relative shifts of the calibrated mitigation rates. Section 6.8 reports the dispersion statistics and paired comparisons, and the sensitivity scripts ship with the released repository.

5.3. Integration

Detection-layer outputs (risk scores, alert classifications) feed directly into the simulated control layer, which applies graduated hold/prioritize/block actions.

Scenario-specific effective mitigation parameters (**Table 5**) encapsulate how detection performance maps onto settlement dynamics: a direct-mitigation rate m_d (the fraction of at-risk value at attacked nodes that graduated actions preserve once detection has occurred), a contagion-mitigation rate m_c (the corresponding fraction for first-order neighbor exposures), and a detection-delay fraction δ (the share of the attack window that passes before AI-augmented detection activates). Before detection the simulation accrues full disruption; from the detection step onward, unsettled flow at attacked nodes is scaled by $(1 - m_d)$ and neighbor contagion flow by $(1 - m_c)$. These parameters are calibration assumptions, exposed explicitly in the released code so readers can rerun the simulation under alternatives and the reported reductions are conditional on them (Section 6.8). One structural consequence deserves to be spelled out: recovery-time compliance within this simulation is governed almost entirely by δ , because mitigation cannot act before detection by construction; detection-delay calibration against empirical incident data is therefore the simulation's most consequential input. Settlement and liquidity metrics are then re-evaluated to quantify how much AI-augmented controls reduce disruption under each scenario's assumptions [2] [3].

All code and synthetic data are available at:

<https://doi.org/10.5281/zenodo.19519911> [48].

6. Results and Discussion

6.1. Cyber Stress Test Results

Across the four scenarios, AI-augmented controls reduced unsettled payment value by 34% - 78%, with the magnitude varying by attack type and recovery dynamics [3]. Scenario S1 (DDoS) showed the highest reduction (78%) because automated re-routing preserved settlement continuity for non-targeted participants. Scenario S4 (coordinated multi-vector) showed the smallest reduction (34%) due to simultaneous disruption of multiple detection channels.

Recovery time: Under the engine's cumulative-disruption concentration metric, the AI-augmented runs yield 2.5 h (S1), 5.1 h (S2), 21.7 h (S3), and 41.3 h (S4), and the repository output flags the two-hour target as unmet in all four scenarios. Under the throughput-restoration metric defined in Section 5.1 and evaluated across ten topologies: S1 restores $\geq 90\%$ aggregate throughput at 1.9 ± 1.1 h (meeting the two-hour objective in 4 of 10 topology draws; post-detection disrupted shares hover near the 10% degradation threshold, making compliance topology-sensitive); S2 restores at 3.2 ± 2.1 h (2 of 10 within two hours); S3 never breaches the 90% threshold at all (aggregate throughput degradation peaks at $\approx 5\%$ because the endpoint-fraud attack manifests as fraudulent value-at-risk rather than availability loss-continuity is maintained throughout, with fraudulent flow contained from control activation at 4.3 h); and S4 does not restore full throughput within its 48-hour window (post-detection disrupted share $\approx 48\%$ of normal). Control-activation times are 0.18 h, 0.72 h, 4.3 h, and 14.4 h for S1-S4 respectively; baseline (no-AI) recovery equals the full attack duration in all scenarios. This is more con-

servative, and—as Section 5.3 makes explicit—compliance is governed by the assumed detection-delay fraction δ rather than by post-detection mitigation intensity.

Liquidity shortfall and contagion spread: Under AI-augmented controls, peak cumulative unsettled value fell by 78% (S1), 61% (S2), 52% (S3), and 34% (S4), tracking the total-unsettled reductions because disruption accumulates monotonically over the window. Across the ten simulated topologies, first-order contagion exposed roughly 29 - 36 of 50 participants in S1, 12 - 28 in S2, 7 - 24 in S3, and 34 - 40 in S4; the contagion channel accounted for about one-third (S2, $\approx 31\%$) to three-fifths (S3, $\approx 58\%$) of baseline unsettled value, and AI-augmented controls cut contagion-channel unsettled value by 74% (S1), 58% (S2), 50% (S3), and 32% (S4). Topology and calibration sensitivity is reported in Section 6.8 and **Table 5**.

Table 5. Calibrated effective mitigation parameters per scenario and sensitivity of the headline reductions (Sections 5.3 and 6.8).

Scenario	Direct mitigation m_d	Contagion mitigation m_c	Detection delay δ	Reduction across 10 topologies (mean $\pm$ SD)	Reduction under $\pm 10\%$ mitigation shift
S1	0.861	0.792	6%	$77.7 \pm 0.2\%$	70.2% - 85.8%
S2	0.720	0.662	12%	$60.9 \pm 0.2\%$	54.9% - 67.1%
S3	0.631	0.581	18%	$51.8 \pm 0.3\%$	46.8% - 57.2%
S4	0.513	0.472	30%	$33.9 \pm 0.1\%$	30.6% - 37.4%

6.2. ML Detection Performance

The detection figure in this section is reported as generated from the documented pipeline under seed 42; the final deposited release should be checked against these reported outputs before resubmission. The XGBoost baseline came in at AUC-ROC 0.9998 with F1 0.869 (precision 0.815, recall 0.915, FPR 2.7×10^{-4} , ≈ 27 false alerts per 100,000 transactions). Random Forest reached AUC 0.9997 but only F1 0.653 at its default balanced-class threshold, the DNN (MLP 128-64-32) achieved AUC 0.9926 with F1 0.838 at a very low false-positive rate (3×10^{-5}), and the hybrid pipeline achieved AUC 0.9719 and F1 0.878. Two aspects of these results deserve emphasis. First, AUC sits near the ceiling for all supervised models; SHAP analysis (Section 6.6) indicates this partly reflects injection-signature leakage in the synthetic generator, so absolute discrimination levels must not be extrapolated to production data. Second, the hybrid's AUC falls below the supervised baseline-fusing calibrated boosting probabilities with a min-max-normalized Isolation Forest score (which on its own achieves F1 0.005 on this benchmark) dilutes an almost-saturated ranking. The hybrid's measured contribution is therefore not ranking improvement but operating-point false-positive filtering (Section 6.3). The executed performance metrics for all model families are summarized in **Table 6**. **Table 7** situates the manuscript's findings against relevant external quantitative benchmarks [1] [2] [9]-[14] [19]-[22].

Table 6. ML model families evaluated—executed results from the released pipeline (seed 42).

Model Family	Role in Pipeline	Executed Metrics (seed 42)	HVPS-Relevant Property
XGBoost	Primary supervised detection	AUC 0.9998, F1 0.869, FPR 2.7×10^{-4}	35.5 ms per 1000-txn batch (4-core CPU)
Random Forest	Supervised ensemble baseline	AUC 0.9997, F1 0.653	Recall-heavy at default threshold
DNN (MLP 128-64-32)	Deep supervised baseline	AUC 0.9926, F1 0.838	Very low FPR (3×10^{-5})
Isolation Forest	Unsupervised novelty signal	F1 0.005 standalone	Fusion input only; not a standalone detector
Graph approximation (k-NN + MLP)	Feature-similarity graph detection	AUC 0.9917, F1 0.838; 0/33 on ring fraud	Lowest FPR (≈ 3 per 100k); $\approx 58\times$ slower end-to-end
RL threshold layer	Adaptive threshold on graph model	Null result—identical to base model	Null result—identical to base model
Hybrid (XGB + IF)	Second-pass false-positive filtering	AUC 0.9719, F1 0.878; 12% - 26% fewer false alerts	97.5% true-positive retention

Table 7. Summary of external quantitative benchmarks contextualizing manuscript results.

Topic	External Quantitative Anchor	Implication for This Manuscript
Payment disruption	3-hour attack disrupts 25% of daily value [2]; shortfalls escalate \$160 bn to \$1.5 tn [1]	Stress-test outputs are scenario-conditional with recovery-time as first-order driver
Mitigation effectiveness	Contingency plans reduce impact by $>50\%$ [2] [3]	AI-augmented controls are continuity enablers, not elimination mechanisms
Hybrid ML performance	Published hybrid AUC 0.991, F1 0.97 on >500 k transactions [9] [10]	The measured hybrid contribution is false-positive filtering (\$6.3)
Single-model competitiveness	XGBoost outperforms alternatives in multiple benchmarks [11]–[13]	Confirmed here: the tuned single model leads on AUC (\$6.2)
GNN/RL performance	FraudGNN-RL: 97.3% F1, 31% fewer FPs [14]; RL mean F1 0.9222 [10]	Edge-based graphs identified as the required architecture (\$6.4)
Basel SMA incentives	SMA combines BI + 10-year loss data [19] [20]	Capital incentive is jurisdiction-scoped; ILM = 1 neutralizes channel [21] [22]

6.3. Hybrid Pipeline and False-Positive Filtering

Table 8. Recall by injected fraud type on the held-out test partition (executed; seed 42). The graph-approximation model’s recall deficit is concentrated entirely in the coordinated ring class, whose only detectable signal is relational.

Model	Coordinated multi-entity (n = 33)	Credential transfer (n = 43)	Message injection (n = 54)	All fraud (n = 130)
XGBoost	0.667	1.000	1.000	0.915
Hybrid (XGB + IF)	0.667	1.000	1.000	0.915
Graph approximation	0.000	1.000	0.981	0.738

At the selected operating threshold, the hybrid removed 12% of the baseline’s false alerts in the seed-42 repository run while keeping 97.5% of true positives; across verification re-runs (platforms and seeds) the reduction ranged from 12% to 26%, and varied more widely still across the five-seed study (Section 6.8) absolute false-

positive counts are small ($\approx 20 - 30$ in a 100,000-transaction partition), so ratio estimates are inherently unstable. Per-fraud-type analysis (**Table 8**) shows the hybrid's recall matching the baselines on every injected fraud class; its entire measured effect sits on the precision side. Class imbalance is handled natively as specified in Section 5.2.

6.4. Graph-Approximation Results

In the released implementation, the graph model is a feature-space approximation: each transaction is linked to its $k = 10$ nearest neighbors in standardized feature space, and neighbor-mean features plus the mean neighbor distance are appended before an MLP performs classification. It is a similarity-graph model, not a counterparty-topology model.

Run on the released dataset, the model reaches AUC 0.9917 and F1 0.838 (precision 0.970, recall 0.738) with the lowest false-positive rate of any evaluated model: 3×10^{-5} , *i.e.*, ≈ 3 false alerts per 100,000 transactions against ≈ 27 for the XGBoost baseline, a nine-fold reduction that bears directly on HVPS false-interdiction cost. That precision is paid for in recall, and the deficit is entirely structural: per-fraud-type analysis (**Table 8**) shows the model catching 0 of 33 coordinated multi-entity ring frauds in the test partition, against 22 of 33 for the tabular baseline, while matching the baseline on the remaining fraud classes. The mechanism is architectural: ring fraud is injected with deliberately unremarkable values whose only signal is relational, and a feature-similarity graph has no way to represent counterparty structure. It makes true edge-based graph learning over sender-receiver topology the priority item for future work (Section 7). The RL threshold-adaptation layer produced a null result: its selected threshold (0.22) yields predictions identical to the base model on every reported metric, because the classifier's output probabilities are saturated near 0 and 1.

Latency (measured under the Section 5.2 protocol on a 4-core CPU environment): the MLP scores a 1000-transaction batch in 3.3 ms against 35.5 ms for the 500-tree boosting ensemble. End-to-end, though, graph-feature construction dominates at ≈ 2.04 ms per transaction (1022 s for 500,000 rows), leaving the graph pipeline roughly 58 $\times$ slower than the boosting baseline end-to-end on this hardware. These are single-machine measurements and the multipliers will shift with hardware, but the ordering graph construction, not model inference, as the binding real-time constraint, is structural.

6.5. Adversarial Robustness

Under the protocol now set out in Section 5.2, verification re-runs across five seeds found supervised-baseline F1 degradation of 10% - 31% and hybrid degradation of 13% - 23%, with the hybrid degrading less in only two of the five replications. The adversarial comparison is therefore reported as inconclusive: neither pipeline shows consistent superiority under static perturbation, adaptive attacks remain untested [39] [49], and no adversarial-robustness claim carries through to

the paper's conclusions.

6.6. Explainability, SHAP Analysis, and Benchmark Leakage

SHAP analysis on the released run identifies, in order of mean absolute attribution: inter-transaction timing (`feat_time_since_last`, 2.01), the round-amount indicator (`feat_is_round_amount`, 1.58), ISO 20022 field completeness (1.39), and sender-value deviation (1.21). The two dominant features map directly onto the generator's fraud-injection rules 1) credential-compromise fraud is assigned round amounts from a four-value set, and 2) message-injection fraud is shifted to off-peak timestamps. In substantial part, the models are learning injection signatures rather than generalizable fraud behavior, which explains the near-ceiling AUC values in Section 6.2 and stands as a first-order limitation of the synthetic benchmark (Section 6.9). SHAP-based inspection is proposed for top-ranked features against generator injection rules as a routine leakage diagnostic for synthetic-data fraud studies; on this benchmark it flags the leakage in a single run. SHAP outputs remain useful for their originally intended governance purpose analyst-interpretable rationale for escalated alerts under DORA and NIST CSF expectations [15] [36] [42] provided the leakage caveat is carried alongside.

6.7. Basel Capital Incentive

Executed under the released calculation ($BIC = 500\text{ M}$, $LC = 700\text{ M}$, with a multi-year averaging attenuation factor of 0.82 applied to the loss reduction, a modelling assumption now stated explicitly), a sustained 30% reduction in operational losses reduces the Loss Component by 24.6% and the ILM by 8.3%. Because operational-risk capital scales linearly with the ILM ($ORC = BIC \times ILM$), the capital-requirement release is identical to the ILM reduction: 8.3%. In jurisdictions setting $ILM = 1$ (e.g., UK PRA [21]), this channel is neutralized; the benefit accrues through reduced direct losses, improved supervisory assessments, and lower insurance costs rather than through Pillar 1 capital. In jurisdictions retaining loss sensitivity (e.g., EU under CRR3 [22]), the channel operates as modelled, conditional on the stated averaging assumption.

6.8. Variability, Sensitivity, and Statistical Robustness

To address the reliance on point estimates, the complete Track 2 pipeline was executed under five random seeds and Track 1 across ten network topologies and $\pm 10\%$ mitigation-rate shifts (Section 5.2, variability protocol). Three findings emerged:

First, detection metrics carry material cross-seed dispersion—Across replications, the supervised baseline's F1 spanned approximately 0.83 - 0.93 and the hybrid's approximately 0.87 - 0.94 (cross-seed SD ≈ 0.03 - 0.04), while AUC-ROC remained uniformly high with overlapping ranges between baseline and hybrid. The hybrid's F1 exceeded the baselines in five of five replications (paired mean improvement +0.024, SD 0.018; one-sided sign test $p \approx 0.03$), a directionally con-

sistent but modest advantage on threshold-dependent metrics; no significance is claimed for AUC, where the baseline is ahead. The false-positive-reduction estimate is markedly less stable (spanning roughly -6% to 75% across replications, with the repository run at 12%), because absolute false-positive counts are small; it is reported as a range, not a point value. The adversarial comparison was inconsistent across seeds (Section 6.5).

Second, the stress-test reductions hold up across topologies but remain conditional on mitigation calibration. Over ten Barabási-Albert topology draws, scenario reductions moved by less than one percentage point (S1: $77.7\% \pm 0.2\%$; S2: $60.9\% \pm 0.2\%$; S3: $51.8\% \pm 0.3\%$; S4: $33.9\% \pm 0.1\%$). The binding uncertainty lies in the calibrated effective mitigation rates: shifting all rates by $\pm 10\%$ (relative) moves the reductions to $70\% - 86\%$ (S1), $55\% - 67\%$ (S2), $47\% - 57\%$ (S3), and $31\% - 37\%$ (S4) (Table 5). Throughput restoration times, by contrast, are topology sensitive (Section 6.1), because post detection disrupted shares hover near the 10% degradation threshold. The headline $34\% - 78\%$ band should therefore be read as conditional on Table 5 calibration assumptions, not as a distribution free estimate.

Third, no confidence intervals attach to single run comparisons; where the text describes a difference between models, the repeated run evidence above is the statistical support on offer, and differences smaller than the reported cross seed dispersion are treated as not established.

6.9. Limitations

The first order limitation is benchmark leakage: SHAP analysis (Section 6.6) shows the strongest learned features tracking the generator's injection rules, so the near ceiling discrimination in Section 6.2 substantially reflects how the benchmark was built rather than transferable detection capability. Absolute performance levels on this benchmark are upper bounds and must not be extrapolated to production HVPS data; before absolute performance claims can be made, the generator needs revising so that fraud values and timing are drawn from the legitimate distributions and detectability rests on relational structure alone. More broadly, the ML results derive from synthetic HVPS-like data rather than proprietary RTGS or SWIFT production datasets, materially constraining external validity [18]. The stress-test reductions ($34\% - 78\%$) are conditional on scenario construction and Table 5 calibration; the topology and mitigation-rate sensitivity analyses in Section 6.8 partially address this, but robustness over alternative network classes and empirically estimated detection-delay distributions remains open [3].

The simulation is single-jurisdictional and does not represent cross-border settlement chains, correspondent banking, or multi-currency frictions [15] [16] [50]. The adversarial evaluation is limited to static perturbations and was inconsistent across seeds; adaptive evasion and concept drift remain untested [39] [49]. Human-AI interaction is treated architecturally but not evaluated empirically. Two

threshold-selection caveats apply: both the hybrid's decision threshold and the graph model's RL threshold are chosen on the evaluation partition, which modestly favors their threshold-dependent metrics; threshold-free AUC comparisons and the paired repeated-run analysis are unaffected, but future work should move to a three-way train/validation/test design. Latency figures are single-machine CPU measurements and support relative conclusions only. Finally, the recovery-time results depend on which metric is used; both definitions are stated explicitly in Section 5.1, and Section 6.1 reports results under each.

The capital-incentive argument depends on jurisdictional ILM choices and multi-year loss-history averaging, making short-horizon ROI claims unreliable [19] [20]. ISO 20022 performance uplift is mechanism-based rather than empirically demonstrated; a matched MT-versus-MX comparative study remains needed [23] [24].

7. Conclusion and Future Work

This paper proposed an AI augmented cyber-resilience framework for HVPS and SWIFT environments, integrating multi-model detection, automated controls, and governance within a regulation aligned architecture. Under the scenario assumptions, simulated controls reduced unsettled value by 34% - 78%, robust to network topology and conditional on the calibrated mitigation parameters; recovery-time compliance with the two-hour FMI objective, evaluated under explicit metric definitions, was met robustly in only one of four scenarios and is governed by the assumed detection delay, identifying detection-latency calibration as the simulation's most consequential input. In the detection track, executed results show a well-tuned XGBoost baseline at near-ceiling discrimination (AUC 0.9998), a hybrid whose measured contribution is operating-point false-positive filtering (12% - 26% fewer false alerts at $\approx 97.5\%$ true-positive retention) rather than ranking improvement, and a feature-similarity graph model that achieves the lowest false-positive rate of any evaluated model while detecting none of the injected coordinated ring fraud—a documented negative result establishing that feature-similarity graphs are not network-aware detection. SHAP inspection shows the benchmark's near-ceiling discrimination partly reflects injection-signature leakage, which is disclosed as a first-order limitation and generalized into a proposed leakage diagnostic for synthetic-data fraud studies. These results should be interpreted as scenario-conditional, decision-support evidence given the nascent state of cyber stress testing practice [3] [5].

The Basel capital-incentive argument holds in jurisdictions retaining a loss-sensitive ILM (e.g., EU under CRR3 [22]) but is neutralized where ILM = 1 (e.g., UK PRA [21]).

Future work should prioritize: 1) a true edge-based graph model (message passing over sender-receiver topology) evaluated against the ring-fraud class this paper's feature-similarity model fails on the single experiment most likely to change the detection conclusions; 2) a generator revision eliminating injection signature

leakage, with benchmark non-triviality demonstrated before absolute performance is reported; 3) real data validation on appropriately anonymized RTGS/HVPS transaction data; 4) cross-border, multi jurisdiction simulation incorporating correspondent banking dependencies [15] [16]; 5) human-in-the-loop empirical studies of alert fatigue and decision latency; 6) adaptive adversarial red-teaming beyond static perturbations [39] [49]; 7) empirical calibration of detection-delay distributions against documented incidents; and 8) three-way train/validation/test designs with pre-registered comparison protocols extending Section 6.8.

The study's core contribution is a coherent, regulation-aware architecture together with a fully reproducible, seeded evaluation pipeline whose results including its negative results and its benchmark's disclosed limitations regenerate from the released code. Within these bounds, resilience investments that integrate proactive detection and control can materially reduce simulated systemic cyber impact, with the largest returns predicted from reducing detection latency; and in qualifying jurisdictions, a capital efficiency channel may strengthen the investment case for disciplined, well-governed deployment [19] [25].

Ethics Statement

This study uses exclusively synthetic data generated by an agent-based payment simulator. No human participants, personal data, or confidential financial data were involved. The research was conducted in accordance with the Committee on Publication Ethics (COPE) guidelines.

Data Availability

The synthetic payment dataset and all code used to generate the results in this study are publicly available at Zenodo: <https://doi.org/10.5281/zenodo.19519911> [48]. The repository includes 1) the HVPS transaction generator, 2) all model training and evaluation scripts including the DNN, graph-model, per-fraud-type, and latency runners, 3) the cyber stress-test engine with both recovery metrics, 4) SHAP analysis notebooks, and 5) the multi-seed variability and mitigation-rate sensitivity scripts underlying Section 6.8. The repository includes the scripts and configuration intended to regenerate the Section 6 outputs under seed 42.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

- [1] Eisenbach, T.M., Kovner, A. and Lee, M.J. (2022) Cyber Risk and the U.S. Financial System: A Pre-Mortem Analysis. *Journal of Financial Economics*, **145**, 802-826. <https://doi.org/10.1016/j.jfineco.2021.10.007>
- [2] Kosse, A. and Lu, Z. (2022) Transmission of Cyber Risk through the Canadian Wholesale Payment System. *Journal of Financial Market Infrastructures*, **10**, 1-28. <https://doi.org/10.21314/jfmi.2023.001>

- [3] Khiaonarong, T., Korpinen, K. and Islam, E. (2025) Using Simulations for Cyber Stress Testing Exercises. *IMF Working Papers*, **2025**, 1-40.
<https://doi.org/10.5089/9798229008952.001>
- [4] European Central Bank (2018) Cyber Resilience Oversight Expectations for Financial Market Infrastructures (CROE). ECB.
- [5] CPMI-IOSCO (2016) Guidance on Cyber Resilience for Financial Market Infrastructures. BIS. <https://www.bis.org/cpmi/publ/d146.pdf>
- [6] Ferdinand, J. and Benham, R. (2017) The Cyber Security Ecosystem: Defining a Taxonomy of Existing, Emerging and Future Cyber Threats. SWIFT Institute WP 2016-002.
- [7] CPMI-IOSCO (2022) Implementation Monitoring of PFMI: Level 3 Assessment on FMI Cyber Resilience. BIS/IOSCO FR14/22.
- [8] Desai, A., Kosse, A. and Sharples, J. (2025) Finding a Needle in a Haystack: A Machine Learning Framework for Anomaly Detection in Payment Systems. *The Journal of Finance and Data Science*, **11**, Article ID: 100163.
<https://doi.org/10.1016/j.jfds.2025.100163>
- [9] Marazqah Btoush, E.A.L., Zhou, X., Gururajan, R., Chan, K.C. and Alsodi, O. (2025) Enhancing Credit Card Fraud Detection with a Stacking-Based Hybrid Machine Learning Approach. *PeerJ Computer Science*, **11**, e3007.
<https://doi.org/10.7717/peerj-cs.3007>
- [10] Devi, R.R., Raja, J.E. and Chin, Y.B. (2025) Reinforcement Learning with Graph Neural Network (RL-GNN) Fusion for Real-Time Financial Fraud Detection: A Context-Aware Community Mining Approach. *Scientific Reports*, **15**, Article No. 42953.
<https://doi.org/10.1038/s41598-025-25200-3>
- [11] Mulyana, I. and Ulum, M.B. (2025) Application of Machine Learning Models for Fraud Detection in Synthetic Mobile Financial Transactions. *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, **10**, 759-769.
<https://doi.org/10.33480/jitk.v10i4.6420>
- [12] Syeeda, S. (2024) Fraud Detection in Financial Transactions Using Machine Learning. *International Journal of Scientific Research and Engineering Trends*, **10**, 1912-1916. <https://doi.org/10.61137/ijret.vol.10.issue5.240>
- [13] Albalawi, T. and Dardouri, S. (2025) Enhancing Credit Card Fraud Detection Using Traditional and Deep Learning Models with Class Imbalance Mitigation. *Frontiers in Artificial Intelligence*, **8**, Article 1643292. <https://doi.org/10.3389/frai.2025.1643292>
- [14] Cui, Y., Han, X., Chen, J., Zhang, X., Yang, J. and Zhang, X. (2025) FraudGNN-RL: A Graph Neural Network with Reinforcement Learning for Adaptive Financial Fraud Detection. *IEEE Open Journal of the Computer Society*, **6**, 426-437.
<https://doi.org/10.1109/ojcs.2025.3543450>
- [15] John, A. and Ahsun, A. (2025) Explainable AI (XAI) for Fraud Detection. SSRN.
- [16] Emmanuel, M. (2025) Federated Learning for Privacy-Preserving Financial Fraud Detection. SSRN.
- [17] C. Ezeugboaja, F. (2025) Federated Learning Architectures for Privacy Preserving Financial Fraud Detection Systems. *Frontiers in Emerging Artificial Intelligence and Machine Learning*, **2**, 72-85. <https://doi.org/10.64917/feaiml/volume02issue12-07>
- [18] Lopez-Rojas, E.A., Elmir, A. and Axelsson, S. (2016) PaySim: A Financial Mobile Money Simulator for Fraud Detection. *Proceeding of 28th EMSS*, Larnaca, 26-28 September 2016, 249-255.
- [19] Basel Committee (2017) Basel III: Finalising Post-Crisis Reforms. BIS.

- <https://www.bis.org/bcbs/publ/d424.pdf>
- [20] Basel Committee (2016) Standardised Measurement Approach for Operational Risk. BCBS D355.
 - [21] Bank of England (2016) PS1/26—Implementation of BASEL 3.1: Final Rules. PRA.
 - [22] European Parliament and Council (2024) Regulation (EU) 2024/1623 Amending Regulation (EU) No 575/2013 (CRR3). OJ EU 2024, L2024/1623.
 - [23] Gancz, A., Zalaki, S. and Hilt, N. (2025) Becoming ISO-Native: Navigating the End of Coexistence in ISO 20022 Migration. Thoughtworks.
 - [24] Bank of America (2025) ISO 20022 Migration Interconnects the Global Economy. BoA.
 - [25] Bank for International Settlements (2021) Cyber Risks in Fast Payment Systems. BIS CPMI Report. <https://www.bis.org/cpmi/publ/d197.pdf>
 - [26] Bank of England (2024) RTGS System and CHAPS: Annual Report 2023/24. BoE.
 - [27] European Central Bank (2026) TARGET Services—T2 Facts and Figures. ECB.
 - [28] European Central Bank (2025) TARGET Services Annual Report 2024. ECB.
 - [29] Bank of England (2024) A Brief Introduction to the RTGS System and CHAPS. BoE.
 - [30] CPMI (2021) Developments in Retail Fast Payments and Implications for RTGS Systems. BIS.
 - [31] SWIFT (2024) Customer Security Programme (CSP) and Customer Security Controls Framework. SWIFT.
 - [32] CPMI (2018) Reducing the Risk of Wholesale Payments Fraud Related to Endpoint Security. BIS. <https://www.bis.org/cpmi/publ/d178.pdf>
 - [33] SWIFT (2019) Three Years on from Bangladesh: Tackling the Adversaries. SWIFT ISAC.
 - [34] Qin, M. and Mogos, G. (2022) Cyber-Attacks on SWIFT Systems of Financial Institutions. *Proceedings of the 5th International Conference on Computer Science and Software Engineering*, Guilin, 21-23 October 2022, 596-599. <https://doi.org/10.1145/3569966.3570116>
 - [35] Mahapatra, B.G. (2024) AI and Machine Learning in Fraud Detection: Strengthening Security in the Financial Payment Domain. *ESP International Journal of Advancements in Computational Technology*, **2**, 125-139.
 - [36] Clausmeier, D. (2022) Regulation of the European Parliament and the Council on Digital Operational Resilience for the Financial Sector (DORA). *International Cybersecurity Law Review*, **4**, 79-90. <https://doi.org/10.1365/s43439-022-00076-5>
 - [37] Price, D., Fitzpatrick, J., Majer, L. and Perry, M. (2025) Model Risk: Turning Compliance into Strategic Capability. Grant Thornton.
 - [38] Bank of England FPC (2025) Financial Stability in Focus: AI in the Financial System. BoE.
 - [39] Malingu, C.J., Titilayo, D.H., Eni, F., *et al.* (2024) Adversarial ML in Finance: Developing Resilient AI Models. *International Journal of Computer Applications Technology and Research*, **13**, 111-137.
 - [40] European Central Bank (2025) TIBER-EU Framework Updated to Align with DORA. ECB.
 - [41] European Parliament and Council (2024) Regulation (EU) 2024/2847 on Horizontal Cyber-Security Requirements (Cyber Resilience Act). OJ EU 2024, L355.
 - [42] National Institute of Standards and Technology (2025) Draft NIST Guidelines Re-

- think Cybersecurity for the AI Era. NIST.
- [43] NIST (2024) The NIST Cybersecurity Framework 2.0. NIST CSF 2.0.
 - [44] European Central Bank (2025) Regulation (EU) 2025/1355 on Oversight Requirements for SIPS (ECB/2025/22). OJ EU 2025; L355.
 - [45] CPMI-IOSCO (2012) Principles for Financial Market Infrastructures. BIS.
<https://www.bis.org/cpmi/publ/d101a.pdf>
 - [46] CPMI-IOSCO (2023) Report on Current CCP Practices to Address Non-Default Losses. BIS/IOSCO.
 - [47] NIST (2023) AI Risk Management Framework (AI RMF 1.0). NIST AI 100-1.
 - [48] Ramakrishna, R.G. (2026) RaghuGR-Bh/Hvps-Swift-AI-Cyber-Resilience: AI-Augmented Cyber Resilience for HVPS/SWIFT (Code and Synthetic Data). Zenodo.
<https://doi.org/10.5281/zenodo.19519911>
 - [49] Vassilev, A., Oprea, A., Fordyce, A. and Anderson, H. (2025) Adversarial Machine Learning: A Taxonomy and Terminology. NIST AI 100-2 E2025.
 - [50] CPMI (2024) Linking Fast Payment Systems across Borders: Governance and Oversight. BIS Paper D223.