ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

When Does Hybrid Fraud Screening Earn Its Alerts?

A Budget-Constrained Comparison of Rules, Anomaly Detection, and Supervised Learning

Theo Desai

Mays Business School, Texas A&M University, USA

Abstract

Financial fraud screening operates under fixed investigation capacity, so overall discrimination may not indicate how many fraudulent transactions reach reviewers. We evaluate a two-parameter, rule-gated ranking that combines supervised histogram gradient boosting, unsupervised Isolation Forest, and three auditable red flags. Coefficients are selected exclusively on a chronological validation split to optimize fraud recall within the highest-risk 1% of transactions. Twenty-five independently seeded, explicitly synthetic streams supply 2,250,000 transactions, each with 60/20/20 temporal train/validation/test partitions. On held-out blocks, the complete hybrid averages 14.53% recall at 1% workload versus 14.08% for boosted trees (mean paired difference +0.45 percentage points; 95% seed-level t interval −0.55 to +1.45; two-sided Monte Carlo sign-flip p≈0.376). The no-rule-bonus variant still gates anomaly evidence and reaches 14.96%; a truly ungated supervised-anomaly baseline reaches 14.05%. This separates anomaly gating from direct rule-score uplift. Budget and review-window analyses show that the marginal value of explicit rule gating depends on the allocation policy and validation-selected coefficients. The central contribution is a reproducible test of when an auditable hybrid changes a restricted review queue, including cases where added components provide no benefit. Results concern simulated, not observed financial activity.

Keywords

Financial fraud detection, anomaly detection, auditable rules, imbalanced learning, review capacity, temporal validation, synthetic data.

1. Introduction

Financial fraud screening is a ranking problem under operational constraints, not simply a binary classification exercise. A risk engine may score a large transaction stream, but a finite investigator team can examine only a small set of alerts. When genuine transactions dominate, overall accuracy may remain deceptively high even if most fraud is missed. Surveys by Ngai et al. [1], West and Bhattacharya [2], and Bolton and Hand [3] establish a broad statistical and computational foundation; operational analyses emphasize temporal change, delayed verification, and limited investigation capacity [4].

The distinction between classification quality and intervention value is consequential. A model with high receiver operating characteristic area (ROC-AUC) can allocate its highest ranks poorly, and a fixed cutoff of 0.5 ignores the daily size of the review queue. We therefore compare every candidate on the same alert budget. The primary quantity is the fraction of fraud detected among the top one percent of transactions; average precision (AP) and ROC-AUC remain secondary, threshold-independent descriptors.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

Feature aggregation and cardholder history have improved fraud detection in prior work [5], while specialized feature engineering and cost-sensitive design address the asymmetry between false alarms and missed fraud [6]. Sequence models further show that a transaction may be informative because of its neighbors rather than its individual attributes [7]. Yet a purely learned ranking can obscure the reason an individual alert is presented to an investigator. A transparent rule can provide a concise operational rationale even when its stand-alone discrimination is limited.

This work studies a budget-constrained, rule-gated ranking fusion that combines a supervised detector, a label-free anomaly detector, and an explicit rule layer. The contribution is a deliberately small and testable interface: the rule layer changes the strength of an anomaly adjustment, and only two coefficients are selected on a chronological validation block. It can be applied without training a new high-capacity end-to-end ensemble. The experiment isolates whether the additional channels materially change which cases fit into a fixed investigation queue.

The contributions are threefold. First, we formalize a review-capacity objective and an auditable, two-parameter fusion that allows an anomaly adjustment to depend on explicit rule activations. Second, we compare five stand-alone methods and five fusion and ablation configurations under chronological validation across 25 independently seeded synthetic streams, documenting all fitted coefficients and per-record predictions. Third, we quantify paired uncertainty, alert-set turnover, and sensitivity to review budgets and window sizes; the expanded replication distinguishes apparent gains in a small sample from more stable observations. No conclusion is asserted for real-world transaction populations without external testing.

2. Related Work

Fraud models differ along two axes: the source of supervision and the object scored. One class learns from verified fraudulent and legitimate transactions; a second learns typical structure and ranks deviations. Carcillo et al. [8] experimentally evaluated an integration of supervised detection with outlier representations on financial data, with particular attention to multi-granularity anomaly scores. Their method integrates derived outlier information into a trained classifier. The present study instead evaluates post-model rank fusion and makes the contribution of investigator-readable rules explicit, enabling component-level ablations.

Data imbalance is also addressed through generating or resampling minority observations. Fiore et al. [9] investigated generative adversarial augmentation to increase minority examples. In a related but distinct setting, Ashtiani and Raahemi [10] systematically reviewed machine-learning approaches to fraudulent financial statements. Our analysis is restricted to transaction-level screening: corporate reporting irregularities require different sampling units, feature meanings, and decision horizons. Temporal subgraph Transformers with laundering-motif prototypes provide an effective relational perspective on coordinated transactions, complementing record-level fraud-ranking analyses [27].

A practical comparison must distinguish an interpretable decision path from a post-hoc explanation. A human-readable rule exposes precisely which conditions fired, whereas the remaining model rank may still be opaque. Accordingly, this paper does not equate an auditable gate with an explanation of every predictor. We preserve the explicit rule trace and treat more extensive model attribution as a separate requirement.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

3. Synthetic Transaction Streams

3.1. Scope and Provenance

We generate 25 independent stochastic streams (integer seeds 42–66), each containing 90,000 synthetic transactions with 13 candidate attributes; total simulated observations are 2,250,000. Fraud labels are Bernoulli draws from a latent risk function whose parameters shift across consecutive time regimes. No customer records or genuine credit-card transaction labels are involved. Generator code, frozen random seeds, feature matrices, held-out predictions, test metrics, and fitted coefficient grids are retained for audit. Streams are independent conditional on the shared generator; they do not represent 25 independent banks.

The features have concrete operational interpretations: amount, amount-to-customer-median ratio, account age, counts of transactions in one and 24 hours, merchant risk, novel-device flag, country mismatch, nighttime flag, IP activity, failed logins, online channel, and a balance ratio. Amount and ratio variables follow heavy-tailed lognormal laws; merchant risk follows a mixture of beta distributions. Discrete events combine Bernoulli, Poisson, and burst components. These are assumed statistics of a simulator, not estimates of any institution's empirical population.

3.2. Latent Fraud Mechanism And Shift

The simulator contains three overlapping attack families: (A) a novel device coupled with an unusual country and spending ratio; (B) rapid low-value transactions following failed logins; and (C) high-value transactions with a risky merchant and elevated IP activity. Continuous covariates add weak risk contributions, and Gaussian perturbation prevents the class label from being a deterministic copy of the engineered rules. If $z(x,t)$ denotes the complete latent score, each outcome obeys $y \sim \text{Bernoulli}(\text{sigmoid}(z(x,t)+b))$; b is the generator's prevalence-setting intercept, chosen before outcome sampling. These hidden coefficients are never passed to the fitted algorithms.

Four consecutive quarters define time regimes. The effects of family A diminish while the strengths of B and C increase. Marginal rates for online transactions, new devices, and geographic mismatches also increase. The deliberate shift means later data do not exactly match the earliest segment. Such imbalanced and time-varying cases are recognized as a general difficulty in fraud-learning research [11]. The experiment does not imply that the simulated change reproduces actual criminal adaptation. Temporal logical attention for distributed-system anomaly detection also shows the methodological value of analyzing changing event dependencies rather than isolated anomaly scores [28].

For each stream, the first 60% (54,000 rows) constitute training, the next 20% (18,000) validation, and the last 20% (18,000) a never-tuned test block. All partition boundaries are chronological. Across the 25 runs, overall fraud prevalence ranges from 0.610% to 0.730% (mean 0.685%), while individual test blocks contain 114–164 positive records. Each test block allocates exactly 180 alerts at 1% capacity. No oversampling is used; SMOTE [12] is therefore contextual prior work rather than an implemented control. The shared simulator provides repeated sampling evidence without removing the limitation of synthetic provenance.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

Table I. Data Generation and Temporal Evaluation

Item	Observed / specified setting
Synthetic streams	25 independent seeds (42–66)
Rows and features	90,000 per stream; 13 features
Total generated records	2,250,000 (entirely synthetic)
Train / validation / test	54,000 / 18,000 / 18,000 per seed
Test fraud records	114–164 per stream
Full-stream prevalence	0.685% mean (0.610–0.730%)
Review quota and drift	180/18,000 at 1%; four regimes

3.3. Dataset Design Choices And Observability

The generator produces observed binary labels by random sampling from risk probabilities rather than copying the visible red-flag formulas. Independent Monte Carlo seeds reveal whether detector rankings are sensitive to sample realizations and temporal changes. Nonetheless, the model family and the rules are evaluated within one designed environment; repeatability across seeds should not be confused with generalization across alternative fraud mechanisms or institutions.

The simulated population has no explicit account identifier. Therefore the amount ratio and rolling-count features are generated as record attributes rather than computed from historical, identified accounts. This choice intentionally simplifies causal attribution and prevents accidental cross-entity leakage in the simulator, but it is not equivalent to a production feature store. A deployment-grade extension would calculate customer aggregates strictly from information available before each transaction and separate customer-level and transaction-level shifts. Likewise, the model cannot be considered fair, privacy-preserving, or compliant merely because the present study does not handle personal records.

4. Budget-Constrained Rule-Gated Fusion

4.1. Operational Objective And Baseline Scores

Let x_i be a transaction and y_i in {0,1} its latent fraud outcome. For n transactions available in a review batch and budget fraction q, set $k=\text{ceil}(qn)$. An allocation vector a_i in {0,1} must satisfy $\sum_i a_i=k$. The ideal, unobservable allocation maximizes $\sum_i a_i y_i$. Operationally, a scoring system orders records by estimated risk, so review is assigned to the k highest scores. The primary recall-at-budget is $\sum_{i \in \text{top-k}(s)} y_i / \sum_i y_i$; precision-at-budget divides the same numerator by k. This is a fixed-capacity assessment, not a predicted probability calibration test.

The conventional baselines are a deterministic hand-auditable rule score, class-weighted logistic regression, random forest [13], histogram-based gradient-boosted trees grounded in the boosting framework [14], and Isolation Forest [15]. All five consume exactly the same 13 features. Isolation Forest is fitted without using labels; its anomaly score is the negative of the scikit-learn decision function. The rule layer consists of three fixed cases: (A) new device plus country mismatch and amount ratio over 2; (B) at least six recent transactions, amount below 60, and at least one failed login; (C) merchant risk at least 0.70 and amount at least 250. Its severity score is $0.95A+0.80B+0.70C$ plus a small fixed amount-ratio tie-break.

4.2. Technical Integration With Forensic Workflow

Zheng et al. [16] distinguish fixed expert rules, supervised classification, unlabeled anomaly analysis, and integrated human-machine forensic workflows. We operationalize these four concrete paths: the three fixed rules stand for investigator-specified red flags; the gradient-boosted tree provides historical-label evidence; Isolation Forest contributes a label-free

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

deviation measure; and the common top-k queue constitutes the shared review interface. This mapping makes every path of the workflow an executable and separately ablatable component, coupled through the gated fusion equation introduced below.

Denote by p(x) the gradient-boosted model output and by d(x) the Isolation Forest anomaly score. Their numerical scales differ, so they are converted to validation-based empirical distribution ranks, u(x)=F_val,p(p(x)) and v(x)=F_val,d(d(x)). Each transformation is fitted only on the validation block and applied unchanged to the test block; thus the hybrid is a risk ordering score, not a calibrated posterior probability. Let h(x)=1 when the rule severity is above 0.5 and h(x)=0 otherwise. Our proposed review score is:

$$S(x) = u(x) + \alpha\,[v(x) - 0.5]\,[0.35 + 0.65\,h(x)] + \beta\,h(x) \quad (1)$$

The 0.35 floor allows anomalies to influence unruled transactions while the 0.65 increment amplifies anomaly evidence where an explicit rule also fires. The β term independently expresses a human-readable rule uplift. The restricted, low-dimensional interface is intended to make score revisions auditable: each ranked alert can show u, v, h, α, and β, with the triggered rule type recorded separately. Unlike replacing the classifier with a large stacked learner, the hybrid retains the underlying model scores.

4.3. Validation, Ablation, And Limiting Cases

The anomaly coefficient is chosen from $\alpha \in \{0, 0.05, 0.10, 0.20, 0.35, 0.50\}$ and the rule coefficient from $\beta \in \{0, 0.025, 0.05, 0.10, 0.20, 0.35\}$, using only validation labels. For each seed, the pair maximizing validation top-1% recall is selected; average precision and then the smaller combined coefficient break ties. The candidate $\alpha=\beta=0$ is always included. Supervised-only, supervised-plus-rules, and supervised-plus-anomaly controls constrain the unused coefficient to zero. All base learners retain identical fixed hyperparameters across all 25 runs. A further independently validation-tuned control removes all rule dependence, with $S_{ungated}=u+\alpha(v-0.5)$; it selects α from the same grid without using test outcomes. Setting β=0 in (1), by contrast, retains the rule-dependent factor multiplying the anomaly term.

This separation matters because probabilities from distinct classifiers need not share a common calibration [17]; rank transformation avoids combining their arbitrary raw scales. Class-prior correction after undersampling, as considered by Dal Pozzolo et al. [18], is unnecessary here because no class undersampling is performed. Moreover, selecting the maximum observed validation recall may itself add selection variance; independent datasets remain essential for eventual deployment.

4.4. Algorithmic Footprint And Auditability

The scoring interface requires three existing objects and a short deterministic post-processing step. Training the classifier and Isolation Forest dominates offline computation; producing their two score streams is independent and therefore parallelizable. Validation-rank transformations take $O(n_v \log n_v)$ preprocessing for n_v validation records. With their sorted reference arrays available, transforming a new observation takes $O(\log n_v)$ binary search per score, while the three binary rule checks are O(1). Sorting n scored transactions to emit a top-k queue takes $O(n \log n)$, or $O(n \log k)$ with a heap in a streaming implementation. These are asymptotic statements about the implemented steps, not measured production throughput.

A case flagged by the fixed rules can store one of three reason codes together with the model and anomaly percentile ranks. For $\alpha=\beta=0$, equation (1) reduces to the supervised rank, which is monotone in its original score. For $\beta>0$, a transaction that crosses a rule condition receives an explicit constant uplift; α controls the added contribution from deviation even when no rule fires. Since the objective concerns ordering, there is no need to interpret a fused value of 1.2 as

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

a 120% fraud probability. The implementation keeps the two channels distinct rather than performing a new training pass on test predictions.

5. Experimental Protocol

5.1. Baselines and Implementation

Logistic regression uses standardized features, class weighting, C=1, and a 400-iteration upper limit. Random forest uses 100 trees, sqrt feature sampling, minimum leaf size 12, and balanced-subset weighting. Gradient boosting uses 85 boosting iterations, learning rate 0.065, 15 maximum leaves, minimum leaf size 55, and L2 regularization 5. Isolation Forest uses standardized features, 100 trees, 512 sampled instances per tree, and an unweighted unlabeled training pool. All are implemented in scikit-learn [19]. Model fits include the random seed, and a fixed maximum of four worker threads is used for parallel ensembles.

The rule-only addition optimizes β with α fixed at zero. The β=0 ablation removes the direct rule-score uplift but deliberately retains the rule gate on the anomaly score. The separate truly ungated control optimizes α in $S_{ungated}=u+\alpha(v-0.5)$; this removes the rule indicator entirely. The full hybrid optimizes both α and β, while the supervised-only version sets both to zero. These distinctions are necessary to attribute differences correctly rather than treating the β=0 variant as independent of rules. Validation selects each configuration without examining its test labels.

5.2. Metrics and Paired Uncertainty

Average precision aggregates the precision-recall staircase over positive instances. It is more informative than accuracy at low prevalence [20, 21]. ROC-AUC is reported for continuity with prior research, while the primary constrained objective is recall among the top 1% alerts. The latter has a direct queue interpretation, whereas a cost-insensitive aggregate area may hide operational differences [22]. We additionally report precision at the same budget, and sensitivity at 0.5%, 2%, and 5% without retuning the chosen coefficients.

The 25 independent generator seeds are the paired experimental units. To describe the difference in recall-at-capacity, we calculate the mean paired difference and a two-sided 95% t interval using the 25 seed-level scores. For the same pre-specified primary comparison (full hybrid versus boosted trees at 1% review), a two-sided sign-flip test uses 200,000 seeded Monte Carlo assignments rather than treating the 450,000 held-out transactions as independent replicates. Secondary comparisons and budgets are exploratory; their unadjusted intervals do not establish familywise statistical significance.

For reproducibility we also retain 300 within-stream transaction-bootstrap resamples for the complete-hybrid versus boosted-tree contrast, using the same 1% workload each time. These describe conditional resampling variation within individual generated blocks and are reported in machine-readable results rather than in the main inferential table. The seed-level paired analysis is the relevant summary for uncertainty across the 25 generated streams.

5.3. Reproducibility and Leakage Checks

To bound information leakage, all transformations that learn statistics are fitted on training or validation portions only. StandardScaler is fitted inside each pipeline using the training block; all scikit-learn estimators are fitted before examining validation scores. Validation percentiles use only validation scores, and the two fusion parameters are chosen without reference to held-out labels. The top-k test operator uses only the number of transactions available in the batch and its scores; test labels are consulted strictly for retrospective evaluation. The generator includes chronological time and regime variables in its saved files, but neither is used as a predictive feature. This prevents the model from simply learning a regime index.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

The reproducibility archive contains generator code and 25 seeds; serialized input matrices and labels for all generated streams; individual held-out score arrays for every method; validation-selected coefficients; all metric tables, inference scripts and figures. Re-evaluation from saved predictions is separate from model fitting. A third-party auditor can reconstruct any reported mean from the 25 stored per-seed records and the stable top-k tie-breaking rule. An additional ungated model is trained and validation-selected in the same way, with each seed's test score archived separately.

6. Results

6.1. Stand-Alone Models Versus The Hybrid

Table II presents held-out performance over 25 separately trained synthetic streams. Mean top-1% fraud recall is 14.53% for the full hybrid and 14.08% for histogram GBDT, a difference of only +0.45 percentage points, with 180 reviewed items in every test block. Their mean average precisions are 0.0522 and 0.0499, respectively. Random forest reaches 13.27% recall, while the transparent rule-only system reaches 12.79%. The small hybrid gain relative to supervised ranking illustrates why multiple independent runs are important before inferring practical advantages.

Table II. Held-Out Performance Across 25 Synthetic Streams (Mean ± Sd)

Method	AP (mean±SD)	R@1% (%)	P@1% (%)
Rules	0.038±0.010	12.79±2.60	9.78
Logistic	0.034±0.010	9.61±2.77	7.33
Isolation forest	0.033±0.008	9.17±2.23	7.04
Random forest	0.047±0.013	13.27±3.97	10.16
Hist. GBDT	0.050±0.013	14.08±3.28	10.78
Full hybrid	0.052±0.012	14.53±2.74	11.11

Figure 1 displays all 25 paired recall differences rather than a single favorable aggregate. The full hybrid is better than GBDT on nine seeds, worse on seven, and tied on nine at the fixed top-1% queue. Several negative realizations counterbalance the largest positive results. The paired 95% t interval for the primary difference is [−0.55, +1.45] percentage points, with a two-sided Monte Carlo sign-flip p≈0.376; therefore, these tests do not establish superiority over the supervised baseline. The validation grid sometimes selects zero fusion coefficients, which limits unproductive additional model complexity.

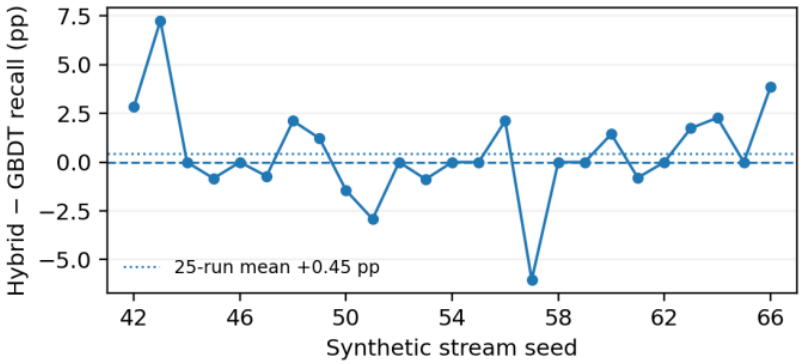


Figure 1. Per-stream paired difference in top-1% fraud recall, full rule-gated hybrid minus histogram GBDT (25 independent synthetic streams).

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

6.2. Ablation and Alert-Capacity Sensitivity

Table III distinguishes rule gating from a direct rule bonus. When β=0, anomaly evidence remains gated by h(x), giving 14.96% mean recall. A validation-tuned, truly ungated ranking S=u+α(v−0.5) reaches 14.05%, close to GBDT (14.08%). The full hybrid achieves 14.53%, and supervised-plus-rules achieves 14.56%. Thus gating an anomaly adjustment and adding a fixed rule uplift have different effects; full fusion does not establish superiority over supervised GBDT.

Table III. Component Ablation On Held-Out Streams

Ranking combination	Mean AP	Recall (%)	SD (pp)
Supervised rank	0.0490	14.04	3.31
Sup. + rules	0.0508	14.56	3.06
Gated anomaly, β=0	0.0519	14.96	2.75
Ungated anomaly	0.0504	14.05	3.28
Full hybrid	0.0522	14.53	2.74

Figure 2 evaluates unchanged rankings at four review capacities. At 0.5%, 1%, 2% and 5%, full-hybrid recall is 8.61%, 14.53%, 21.84% and 29.99%, compared with 8.34%, 14.08%, 20.26% and 29.04% for GBDT. The exploratory 2% paired advantage is +1.58 points (95% t interval +0.60 to +2.56). These post-hoc budget contrasts were not separately optimized. Each capacity corresponds to 90, 180, 360 or 900 alerts per block, respectively.

6.3. Review-window Stability And Seed-Level Inference

To evaluate allocation timing, we apply independent 1% quotas to consecutive windows of 500, 1,000, 3,000, 6,000 or 18,000 transactions. Every policy still selects exactly 180 records per test block. Models, features and fusion settings are frozen; only the distribution of the review budget over time changes (Table IV).

Table IV. Fixed 1% Quota By Review Window; 25-Run Mean Recall (%)

Window	GBDT	Hybrid	Δ (pp)	95% CI (pp)
500	13.57	14.21	+0.65	[-0.55, +1.84]
1,000	13.77	14.56	+0.79	[-0.23, +1.81]
3,000	13.51	14.31	+0.80	[-0.22, +1.82]
6,000	13.95	14.61	+0.66	[-0.40, +1.72]
18,000	14.08	14.53	+0.45	[-0.55, +1.45]

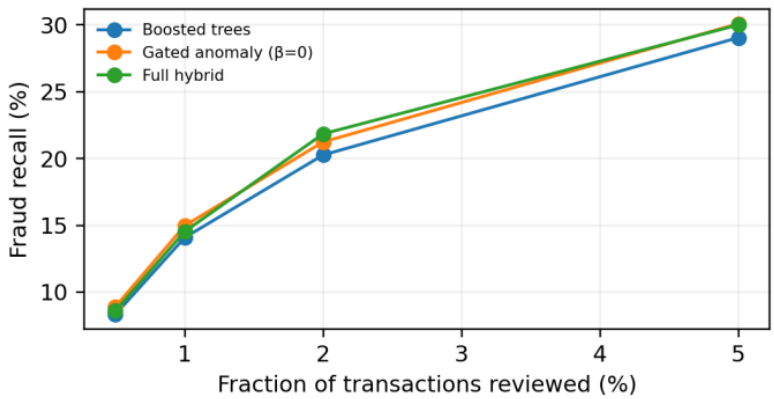


Figure 2. Mean held-out fraud recall at four review fractions with model scores and validation-selected parameters fixed across operating points.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

For 500-transaction review windows, mean recall is 14.21% for the hybrid versus 13.57% for GBDT; at 3,000 rows these means are 14.31% versus 13.51%. Across all five window sizes, the observed hybrid mean is higher, but the paired t intervals on the five window-specific differences all contain zero. Smaller windows reduce the risk of retrospectively reallocating review slots from early to late events, while introducing local quota competition. This sensitivity is distinct from any ability of a detector to retrain under concept drift.

6.4. Precision, Compute Budget, And Uncertainty

At 1% review workload, hybrid precision averages 11.11%, versus 10.78% for GBDT, 10.16% for random forest, and 9.78% for the fixed rule score. The full hybrid retains an AP advantage over GBDT in the 25-run mean (0.0522 versus 0.0499), but its operational recall improvement is uncertain. Precision, recall and alert turnover must be interpreted at the same fixed 180-alert quota; changing a score threshold without controlling queue size would not provide an equivalent resource comparison.

Observed training-time means across the 25 fits are 0.036 s for logistic regression, 0.232 s for histogram GBDT, 2.499 s for random forest, and 0.383 s for Isolation Forest on the measurement CPU. These figures measure local model-fitting time only, not production feature collection or scoring latency. All ranking and audit analyses operate on retained test predictions, allowing the operational objective to be varied without refitting the models.

Table V reports seed-level paired effects. The full hybrid exceeds the rule-only system by +1.73 points on average but differs from GBDT by only +0.45 points, with a 95% interval including zero. Relative to the β=0 rule-gated anomaly configuration the full method is −0.44 points; relative to the truly ungated supervised-anomaly model it is +0.48 points (95% t interval −0.68 to +1.63). All alternatives use independent validation selection, but these multiple secondary contrasts are exploratory. In this simulator, modifying anomaly weight in rule-activated regions can help, whereas an additional fixed rule bonus does not reliably improve the review queue.

Table V. Paired Hybrid Recall Differences At 1% Budget (25 Seeds)

Versus	Δ R (pp)	95% CI (pp)	p (MC)
GBDT	+0.45	[-0.55, +1.45]	0.376
RF	+1.26	[-0.12, +2.63]	0.071
Rules	+1.73	[+0.58, +2.89]	0.003
Gated β=0	-0.44	[-0.84, -0.03]	0.036
Ungated anom.	+0.48	[-0.68, +1.63]	0.414
Sup.+rules	-0.04	[-0.44, +0.36]	0.847

Taken together, these controls evaluate each path of the forensic workflow of Zheng et al. [16] at the same review budget: fixed red-flag rules, supervised and unsupervised scores, the fully ungated two-channel fusion, rule-gated anomaly modulation with β=0, and joint fusion. The rule-only baseline recalls 12.79% of fraud at the fixed budget, compared with 14.53% for joint fusion. The β=0 gated-anomaly case recalls 14.96%, versus 14.05% for the true no-rule fusion, showing why separate gates and rule bonuses must be examined individually. Within this workflow, the rule path is most useful when it modulates anomaly evidence rather than when it adds a fixed score bonus.

6.5. Alert Turnover And Rule Enrichment

We compare the identity of each seed's 180 selected transactions under boosted trees and the full hybrid. Averaged across 25 streams, the fusion replaces 51.2 alerts per queue and exchanges a net 0.6 fraudulent alerts for a mean +0.45 percentage-point recall change. Which coefficients are chosen matters: Table VI partitions the 25 fitted systems by whether validation

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

chose only the rule adjustment, only the anomaly adjustment, both, or neither. These groups are post-selection descriptions, not randomized assignments or evidence of causal improvements by a component.

Only 0.91% of held-out synthetic transactions activate at least one fixed rule, yet average fraud prevalence is 10.60% among rule-activated events compared with 0.68% otherwise. The rule-active fraction of the hybrid alert queues averages 72.9%, up from 44.9% for boosted trees. This enrichment is deliberately associated with the generator's attack patterns, so it is an internal explanation of the alert swaps and not an externally measured industry signal. High rule enrichment alone does not guarantee that its extra selected alerts are more useful than the supervised top ranks.

Table VI. Fusion Coefficients And Alert-Queue Turnover (25 Seeds)

Activated controls	Runs	Avg. replaced	Avg. net TP
None	6	0.5	-0.17
Anomaly-only	5	53.4	+0.60
Rule-only	3	109.7	+1.00
Both	11	61.9	+0.91

Validation chooses the exact supervised fallback $\alpha=\beta=0$ in six streams, anomaly-only influence in five, rule-only influence in three, and both adjustments in eleven. Average changed-alert counts range from almost zero for the fallback to about 110 for rule-only augmentation. The largest rank changes need not generate correspondingly large recall gains because all variants compete for the same 180 slots. Thus, coefficient fallback and queue movement are useful audit signals, but selection stability across independent temporal environments must be tested rather than assumed.

7. Discussion and Limitations

Three design lessons emerge. First, review capacity should be enforced identically when comparing rankings. Second, an auditable rule can act either as a modifier of anomaly importance or as a direct score bonus; these mechanisms are not equivalent. Third, removing the bonus while retaining the gate delivers the highest mean top-1% recall in the expanded simulation, whereas removing the gate as well drops mean recall close to supervised GBDT. These are controlled synthetic observations, not validated real-world recommendations; larger cross-distribution and operational tests remain necessary.

Rule traceability should not be confused with complete model transparency. The gate permits an auditor to inspect activated thresholds, but it does not identify which predictor drove an arbitrary gradient-boosted score. Model-agnostic explanations such as LIME [23] and additive attributions such as SHAP [24] are relevant options for an independently designed explanation layer; neither is evaluated in this experiment. A production evaluation would also need explanations that remain stable under correlated features and data drift.

The principal limitation remains data provenance. All 2.25 million transactions are generated from one synthetic family with visible, intentionally rule-related fraud mechanisms; repeated seeds sample uncertainty within that family, not distinct financial markets. Features such as prior spending ratios and short-term activity counts are generated attributes, not temporally computed customer aggregates. There are no account identities, delayed adjudication, reviewer behavior, intervention costs or genuine regulatory constraints. Seed-level confidence intervals quantify simulation variability but do not replace a real external benchmark.

The ULB/Worldline credit-card benchmark, which provides anonymized principal components, Amount, Time and fraud labels, is the natural external test; because its PCA features cannot

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

express the simulator's device and location rules, it requires an independent train-period statistical-rule control. Evaluating this frozen protocol on the real, time-ordered records, and then testing label delays and institution-specific intervention costs, is the next step. Paired ROC procedures [25] and paired classification-proportion tests [26] assess different objectives and should not replace uncertainty analysis for the fixed-capacity policy.

The batch-window check allocates the same total number of alerts among windows drawn from the unchanged final simulation regime. It examines review scheduling, not adaptive retraining. The selected fusion policy is fixed using validation labels available before the test interval; any prospective implementation would also have to document feedback delays and feature-availability timestamps.

8. Conclusion

This study tests conventional rules, supervised models, Isolation Forest and a compact two-parameter rule-gated ranking under a fixed investigation budget. Across 25 independently seeded, 90,000-event synthetic streams with chronological holdouts, the full hybrid captures 14.53% of fraud within the top 1% of alerts, versus 14.08% for histogram GBDT (95% seed-level interval on the difference [−0.55, +1.45] percentage points). A $\beta=0$ model that retains rule-conditioned anomaly weighting reaches 14.96%, while an independently validation-tuned ungated supervised-anomaly model reaches 14.05%. The main reproducible finding is therefore the distinction between rule-conditioned anomaly modulation and a direct rule-score bonus, not categorical superiority of the most complex detector. Secondary budget and review-window tests quantify how 180 review slots are allocated. Real transaction data, delayed labels and investigator costs are required for external validation.

References

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

[2] West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. Computers & Security, 57, 47–66. https://doi.org/10.1016/j.cose.2015.09.005

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

[4] 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. https://doi.org/10.1109/TNNLS.2017.2736643

[5] Whitrow, C., Hand, D. J., Juszczak, P., Weston, D., & Adams, N. M. (2009). Transaction aggregation as a strategy for credit card fraud detection. Data Mining and Knowledge Discovery, 18, 30–55. https://doi.org/10.10618-008-0116-z

[6] Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. Expert Systems with Applications, 51, 134–142. https://doi.org/10.1016/j.eswa.2015.12.030

[7] Chen, J., Han, X., Zhang, X., & Yang, Y. (2023). Exact concept-level explanations for early detection of financial misstatements. Eurasia Journal of Science and Technology, 5(2), 34–48.

[8] Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oble, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. Information Sciences, 557, 317–331. https://doi.org/10.1016/j.ins.2019.05.042

[9] Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. Information Sciences, 479, 448–455. https://doi.org/10.1016/j.ins.2017.12.030

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

[10] Ashtiani, M. N., & Raahemi, B. (2022). Intelligent fraud detection in financial statements using machine learning and data mining: A systematic literature review. IEEE Access, 10, 72504–72525. https://doi.org/10.1109/ACCESS.2021.3096799

[11] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

[12] 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. https://doi.org/10.1613/jair.953

[13] Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324

[14] Zhou, F. (2026). Moment-constrained side-force allocation for wind-assisted merchant ships: Hydrodynamic performance of retractable anti-leeway fins and flap rudders. Innovation and Technology Studies, 3(1), 27–41.

[15] Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. Proceedings of the 8th IEEE International Conference on Data Mining, 413–422. https://doi.org/10.1109/ICDM.2008.17

[16] Chen, J., Zhang, X., & Han, X. (2023). Explainable identification of tax-related financial reporting anomalies in U.S. public companies: Cross-fitted tax-consistency residuals and Owen-value Shapley propagation. World Journal of Economics and Business Research, 1(1), 33–45.

[17] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. Proceedings of the 22nd International Conference on Machine Learning, 625–632. https://doi.org/10.1145/1102351.1102430

[18] Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. Proceedings of the IEEE Symposium Series on Computational Intelligence, 159–166. https://doi.org/10.1109/SSCI.2015.33

[19] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[20] Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning, 233–240. https://doi.org/10.1145/1143844.1143874

[21] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), Article e0118432. https://doi.org/10.1371/journal.pone.0118432

[22] Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. Machine Learning, 77, 103–123. https://doi.org/10.1007/s10994-009-5119-5

[23] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

[24] Wang, J., Ge, Y., Liu, J., & Jiang, B. (2024). Early warning of liquidity stress in U.S. banks with seasonally aligned frailty hazards and delayed-feedback conformal alarms. Journal of Trends in Finance and Economics, 1(1), 85–100.

[25] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics, 44(3), 837–845. https://doi.org/10.2307/2531595

[26] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2), 153–157. https://doi.org/10.1007/BF02295996

[27] Lin, Y., Zhang, H., & Li, M. (2026). Subgraph-Level Temporal Graph Transformer With Laundering-Motif Prototypes for Anti-Money Laundering on Blockchains. IEEE Access. https://doi.org/10.1109/ACCESS.2026.3731160

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

[28] Liu, Y., Xu, K., & Chen, L. (2024). Temporal Logical Attention Network for Log-Based Anomaly Detection in Distributed Systems. Sensors, 24, Article 7949. https://doi.org/10.3390/s24247949