Distribution-Free Error Control for Fraud Alert Triage: Conformal Anomaly P-Values Under Class Imbalance, Concept Drift, and Calibration Contamination

Olamiji Onafowokan[1], Olawale Fadugba[2], Deborah Okunola[1], Alex Mendy[1]
[1] Department of Statistics, Georgia State University, USA
[2] Federal University of Agriculture, Abeokuta, Nigeria
Corresponding author: olamijionafowokan17@gmail.com
https://doi.org/10.56201/ijasmt.vol.12.no8.2026pg35.48

Abstract
Operational fraud detection almost always terminates in a fixed-quantile threshold applied to a model score, a rule that fixes alert volume but leaves the statistical properties of the alert list undefined. We study an alternative in which the alert list is produced by a multiple-testing procedure applied to split-conformal anomaly p-values, so that the expected proportion of false alarms is controlled at a level chosen in advance. Using a fully synthetic transaction generator with a hard-to-detect mimicry subpopulation, we quantify the behaviour of this design under the three conditions that characterize real fraud environments: extreme class imbalance, covariate drift, and contamination of the calibration sample by undetected fraud. Across twenty replications and four prevalence levels spanning 20 to 200 basis points, the Benjamini–Hochberg procedure applied to conformal p-values held the realized false discovery rate near its nominal 10 percent level (0.067 to 0.139) and delivered stable alert precision between 0.86 and 0.93, whereas a conventional top-one-percent score rule with the same detector produced precision ranging from 0.28 to 0.80 as prevalence varied, with realized false discovery proportions as high as 0.72. Adaptive null-proportion estimation gave no measurable power gain at these prevalences. The procedure is, however, fragile in two specific and quantifiable ways. Modest covariate drift broke exchangeability and inflated the realized false discovery rate from 0.09 to 0.63 at a drift magnitude that shifted the log-amount distribution by one fifth of a standard deviation, while recalibration on a recent window restored validity at a severe cost in power. Contamination of the calibration sample at only 50 basis points eliminated detection entirely by inflating the reference distribution. We show that trimming the upper tail of the calibration scores recovers most of the lost power, but only when the trimming fraction is strictly smaller than the true contamination rate: trimming at exactly the contamination rate inflated the false discovery rate to 0.40, and trimming at three times that rate to 0.80. We also find that a Kolmogorov–Smirnov uniformity test on the null p-values is a sensitive monitor for drift but a weak monitor for contamination, and we translate these results into a deployment protocol. All results are generated from synthetic data; no real transaction records were used.

Keywords: *conformal prediction; false discovery rate; anomaly detection; class imbalance; concept drift; calibration contamination; fraud analytics; alert triage*

1. Introduction

A fraud detection system produces a score; an operations team investigates the top of the ranking. The threshold is usually chosen so that the number of alerts matches investigative capacity, which makes the alert volume a controlled quantity and the error rate an uncontrolled one. This is an awkward inversion of priorities. Capacity is renegotiable in the short run, whereas an investigator confronted with a queue in which four alerts in five are spurious will rationally deprioritize the queue, and a supervisor asked to attest to the soundness of an automated decision process cannot do so on the basis of a quantile.

The statistical machinery for the alternative is mature. Split-conformal inference converts any score into a p-value that is marginally valid under exchangeability, without distributional assumptions on the score or the features (Vovk, Gammerman, & Shafer, 2005; Angelopoulos & Bates, 2023). Applied to the outlier-testing problem, conformal p-values computed against a calibration sample of inliers test the null hypothesis that a new observation was drawn from the inlier distribution, and Bates, Candès, Lei, Romano, and Sesia (2023) showed that the Benjamini–Hochberg procedure applied to such p-values controls the false discovery rate despite their dependence through the shared calibration sample. The resulting alert list carries a statement of the form that an investigator understands: at most a stated fraction of these cases is expected to be a false alarm. Whether this transfers to fraud is not obvious, because fraud violates both premises of the construction. Exchangeability fails because transaction distributions drift; and the calibration sample of "legitimate" cases is contaminated, because undetected fraud is unlabelled by definition. Neither failure has been quantified in a fraud-realistic setting. This paper supplies that quantification and proposes a partial remedy. Our contributions are four. First, we benchmark false discovery rate control on conformal anomaly p-values against the fixed-quantile rule that dominates practice, across a realistic prevalence range, and show that the advantage lies not in detection power, which is essentially identical, but in the stability of alert precision as prevalence varies. Second, we quantify the sensitivity of the guarantee to covariate drift and show that the failure is severe and arrives early. Third, we quantify the effect of calibration contamination, showing that it does not violate validity but destroys power at contamination rates far below those plausible in deployment, and we propose trimmed calibration as a remedy with an explicit and sharp condition for its safe use. Fourth, we evaluate a uniformity diagnostic computed on null p-values and characterize what it can and cannot detect.

2. Related work

The fraud analytics literature is dominated by comparisons of predictors under imbalance (Bhattacharyya, Jha, Tharakunnel, & Westland, 2011; Dal Pozzolo, Boracchi, Caelen, Alippi, & Bontempi, 2018; Carcillo et al., 2021), by feature construction (Bahnsen, Aouada, Stojanovic, & Ottersten, 2016; Wedge, Kanter, Veeramachaneni, Rubio, & Perez, 2019), and by surveys of algorithm families (Bolton & Hand, 2002; Abdallah, Maarof, & Zainal, 2016; Hilal, Gadsden, & Yawney, 2022). Evaluation in this literature is threshold-free or fixed-threshold; error-rate control is not addressed. Multiple testing supplies the missing element. The Benjamini–Hochberg procedure controls the expected proportion of false discoveries among rejections under independence and under positive regression dependence (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001), and adaptive procedures improve power by estimating the null proportion (Storey, 2002; Efron, 2004). Conformal inference supplies assumption-light p-values (Vovk et al., 2005), and Bates et al. (2023) establish the outlier-testing case that we use here. Robustness of conformal validity to departures from exchangeability has been studied through

reweighting and through bounds involving the total variation distance between calibration and test distributions (Barber, Candès, Ramdas, & Tibshirani, 2023); the fraud-specific magnitudes have not been characterized. Contamination of an assumed-clean reference sample is a classical robustness concern in outlier detection, but its interaction with false discovery rate procedures in the extreme-imbalance regime appears not to have been quantified.

3. Methods

3.1 Setup and notation

Let each case be a feature vector X in R^p with an unobserved label Y in {0, 1}, where Y = 1 denotes fraud. Three disjoint samples are used. A training sample fits a scoring function s: R^p → R with larger values indicating greater suspicion. A calibration sample of size n, intended to contain only legitimate cases, supplies reference scores $s_1, \ldots, s_n$. A test batch of size m consists of the cases awaiting triage in the current operating period. For each test case j we test the null hypothesis H_{0j} that case j was generated by the legitimate distribution.

3.2 Split-conformal anomaly p-values

The conformal p-value for test case j is the rank of its score within the calibration scores, with the usual finite-sample correction:

$$p_j = (1 + \#\{ i \le n : s_i \ge s(x_j) \}) / (n + 1).$$

If the calibration scores and the score of a legitimate test case are exchangeable, p_j is stochastically larger than or equal to a uniform variate on the unit interval, so the test that rejects when $p_j \le \alpha$ is valid at level α (Vovk et al., 2005; Bates et al., 2023). Two features of the construction matter operationally. The p-value is bounded below by 1/(n + 1), so the calibration sample must be large relative to m/α for a multiple-testing procedure to be able to reject at all; this is why calibration sizes of tens of thousands, which are trivially available in transaction settings, are required. And the p-value is invariant to monotone transformations of the score, so any calibration pathology of the underlying model is irrelevant: only the ranking matters.

3.3 False discovery rate control

Given p-values $p_1, \ldots, p_m$ with order statistics $p_{(1)} \le \ldots \le p_{(m)}$, the Benjamini–Hochberg procedure at level α rejects the hypotheses corresponding to the k smallest p-values, where k is the largest index with $p_{(k)} \le \alpha k/m$. Conformal p-values computed against a common calibration sample are dependent, but the dependence is positive regression dependent on the subset of nulls, so control is retained (Benjamini & Yekutieli, 2001; Bates et al., 2023). We also evaluate Storey's adaptive variant, which rescales the level by an estimate of the null proportion,

$$\hat{\pi_0} = (1 + \#\{ j : p_j > \lambda \}) / (m (1 - \lambda)), \quad \lambda = 0.5,$$

and which has no finite-sample guarantee in this dependent setting; we assess it empirically. The interpretation of the output is the operationally useful part: among the cases alerted, the expected proportion that are legitimate is at most α.

3.4 Trimmed calibration under contamination

Undetected fraud in the calibration sample raises the upper tail of the reference distribution, so that genuinely suspicious test cases no longer appear extreme and their p-values are inflated. Validity is preserved — the procedure becomes conservative — but power can vanish. We therefore consider trimming: order the calibration scores and discard the largest γn of them, computing conformal p-values against the remaining n(1 − γ). Trimming removes contaminating fraud but

also removes legitimate upper-tail mass, which makes the reference distribution optimistic and the p-values anti-conservative. The trade-off is governed by the relation between γ and the true contamination rate ε, and Section 5.4 characterizes it empirically.

3.5 A uniformity diagnostic

Because validity rests on exchangeability, a natural monitor is the empirical distribution of p-values for cases known to be legitimate. Under exchangeability these are approximately uniform, so the Kolmogorov–Smirnov distance between their empirical distribution function and the uniform distribution function is a scalar drift statistic. In deployment the labels needed to isolate null cases arrive only after verification latency, so the diagnostic is retrospective; we evaluate its sensitivity to both drift and contamination.

4. Simulation design

4.1 Data-generating process

All data are synthetic. Legitimate cases are generated from a twelve-dimensional latent Gaussian vector with an autoregressive correlation structure of parameter 0.6, transformed marginally to resemble account-day transaction covariates: a log-transformed amount, a smoothed hour-of-day, a Poisson twenty-four-hour velocity count, a gamma-distributed inter-transaction interval, a merchant risk score on the unit interval, a cross-border indicator, a log device age, and a Poisson count of distinct merchant categories over seven days, together with four residual latent coordinates. Fraudulent cases are drawn from the same mechanism and then perturbed. Sixty-five percent are "overt": the amount is shifted upward by approximately two standard deviations, activity is displaced into night hours, velocity increases, merchant risk rises, and device age falls. The remaining thirty-five percent are a "mimicry" subpopulation subjected to small shifts only, and are by construction close to the legitimate manifold. This subpopulation places a ceiling on attainable power and prevents the artificially clean separation that makes many simulation studies uninformative. Covariate drift is introduced by a scalar magnitude δ that shifts the legitimate log-amount distribution upward by 0.8δ, raises merchant risk by 0.15δ, lowers device age by 0.5δ, and shifts the residual coordinates by 0.3δ. Drift is applied to legitimate cases only, so it represents a change in normal behaviour rather than a change in fraud tactics.

4.2 Detectors

Three scoring functions are compared, all trained on the same 20,000-case training sample containing one percent fraud. A histogram-based gradient boosting classifier and an L2-penalized logistic regression represent supervised scores; an isolation forest fitted to legitimate training cases only represents the unsupervised alternative (Liu, Ting, & Zhou, 2008; Breiman, 2001; Chen & Guestrin, 2016). Features are standardized using training-sample statistics. Table 1 reports discrimination on an independent 20,000-case sample with one percent prevalence, averaged over ten replications.

Table 1. Discrimination of the three scoring functions on independent test data at one percent prevalence (mean over 10 replications; standard deviations in parentheses).

Scoring function	ROC AUC	Average precision	Recall in top 1%
Gradient boosting	0.893 (0.011)	0.696 (0.007)	0.673 (0.010)
Penalized logistic	0.899 (0.010)	0.690 (0.011)	0.662 (0.009)
Isolation forest	0.797 (0.019)	0.113 (0.035)	0.187 (0.037)

The gap between the two supervised scores is negligible, as expected for a generative mechanism whose signal is largely additive; the isolation forest is markedly weaker because outlyingness and fraud are only partially aligned by construction.

4.3 Decision rules compared

- *Top one percent score rule.* The threshold is the 99th percentile of scores on an independent 10,000-case tuning sample, held fixed thereafter. This mimics the standard capacity-driven rule.
- *Benjamini–Hochberg on conformal p-values* at $\alpha = 0.10$.
- *Storey-adaptive Benjamini–Hochberg on conformal p-values* at $\alpha = 0.10$, $\lambda = 0.5$.
- *Unadjusted conformal testing*, rejecting whenever $p_j \leq 0.10$, included to show what multiplicity adjustment is doing.

4.4 Experiments

Each replication draws a fresh training sample, fits the three scorers, and then evaluates on independent samples. Calibration samples contain 20,000 cases and test batches 5,000 cases. Experiment 1 varies test prevalence over 20, 50, 100, and 200 basis points with a clean calibration sample and no drift. Experiment 2 fixes prevalence at 100 basis points and contaminates the calibration sample at rates of 0, 50, 100, 200, and 500 basis points. Experiment 3 fixes prevalence at 100 basis points and varies drift over $\delta \in \{0, 0.25, 0.5, 1.0\}$, comparing a static calibration sample drawn before drift with rolling recalibration on a 20,000-case post-drift window contaminated at 20 basis points. Experiment 4 examines trimming, first at a contamination rate of 100 basis points without drift, then in the rolling-recalibration configuration of Experiment 3 at $\delta = 0.5$. Twenty independent replications are run for each configuration.

4.5 Metrics

For each replication we record the number of alerts, the realized false discovery proportion (false alerts divided by alerts, defined as zero when no alerts are issued), power (detected fraud divided by fraud present), and precision. The false discovery rate is the mean of the false discovery proportion across replications, and we report its Monte Carlo standard error. We also record the Kolmogorov–Smirnov statistic of the p-values of the legitimate test cases against the uniform distribution.

4.6 Implementation

All experiments were implemented in Python using NumPy, SciPy, and scikit-learn, with a fixed master seed and replication-indexed streams. The generating code and the replication-level output are provided as supplementary material.

5. Results

5.1 Error control and power across prevalence

Table 2 reports Experiment 1. For both supervised scores, the Benjamini–Hochberg procedure on conformal p-values held the realized false discovery rate close to the nominal level of 0.10 at every prevalence, with all deviations within roughly one Monte Carlo standard error of the target. The Storey-adaptive variant was numerically indistinguishable from the unadaptive procedure, differing by at most one alert on average; at these prevalences the estimated null proportion is essentially one, so adaptivity has nothing to recover. Unadjusted testing at the nominal level produced roughly five hundred alerts per five thousand cases with false discovery proportions between 0.87 and 0.99, which is precisely the regime that multiplicity adjustment exists to prevent, and which corresponds closely to what an uncalibrated score threshold delivers in practice.

Table 2. Experiment 1. Realized false discovery rate, power, and precision at $\alpha = 0.10$ across test prevalence, with a clean calibration sample and no drift (20 replications; Monte Carlo standard errors in parentheses).

Score / rule	Prevalence	Alerts	FDR	Power	Precision
Gradient boosting, BH	0.002	7.2	0.092 (0.026)	0.623 (0.011)	0.908
	0.005	16.7	0.067 (0.015)	0.652 (0.003)	0.933
	0.010	35.9	0.110 (0.013)	0.660 (0.004)	0.890
	0.020	74.4	0.108 (0.010)	0.659 (0.004)	0.892
Logistic, BH	0.002	6.8	0.139 (0.033)	0.629 (0.018)	0.861
	0.005	17.0	0.088 (0.019)	0.643 (0.005)	0.912
	0.010	35.7	0.093 (0.012)	0.644 (0.004)	0.907
	0.020	72.3	0.089 (0.008)	0.655 (0.003)	0.911
Gradient boosting, top 1%	0.002	24.4	0.712 (0.020)	0.667 (0.012)	0.288
	0.005	33.0	0.509 (0.020)	0.671 (0.006)	0.491
	0.010	50.7	0.351 (0.024)	0.673 (0.005)	0.649
	0.020	84.3	0.202 (0.012)	0.670 (0.005)	0.798
Logistic, unadjusted $p \le 0.10$	0.002	501.3	0.986 (0.001)	0.775 (0.026)	0.014
	0.010	529.4	0.927 (0.002)	0.777 (0.010)	0.073
	0.020	573.8	0.867 (0.003)	0.760 (0.007)	0.133
Isolation forest, BH	0.002–0.020	0.0–0.6	0.000–0.053	0.000–0.004	—

Score / rule	Prevalence	Alerts	FDR	Power	Precision
Isolation forest, top 1%	0.002	42.4	0.951 (0.010)	0.175 (0.031)	0.049
	0.020	55.9	0.675 (0.016)	0.180 (0.008)	0.325

Storey-adaptive results are omitted because they coincided with the unadaptive procedure to within one alert at every configuration.

The comparison with the top one percent rule is the substantive finding, and it is not about power. At one hundred basis points the two rules detect almost the same fraud: 0.660 against 0.673 for gradient boosting. What differs is the composition of the queue. As prevalence falls from 200 to 20 basis points, the precision of the fixed-quantile rule degrades from 0.798 to 0.288, so that the same operating rule delivers a queue that is four fifths true positives in one period and less than one third true positives in another, with no signal to the operations team that anything has changed. Under the Benjamini–Hochberg rule precision remained between 0.861 and 0.933 across the same tenfold change in prevalence, and the adaptation occurred through alert volume, which fell from 74 to 7. In other words, the fixed-quantile rule holds volume constant and lets quality float; the testing rule holds quality constant and lets volume float. Only the second exposes the change in the environment to the people responsible for it.

The isolation forest results are a negative finding worth stating plainly. Its ranking is far too coarse for the procedure to reject anything: because the smallest attainable p-value is $1/(n + 1)$ and the Benjamini–Hochberg threshold at the first rejection is α/m, a detector must place a sufficient number of true positives strictly above the entire calibration sample before any discovery is possible. The unsupervised score never did so. It nonetheless produced alerts under the fixed-quantile rule — at a realized false discovery proportion of 0.95 at low prevalence. The testing framework converts an unusable detector into an empty alert list rather than a misleading one, which we regard as the correct behaviour.

5.2 Alert-volume stability

Table 3 reports the replication-level dispersion of alert counts under the two rules, at the same nominal settings. The fixed-quantile rule is not in fact volume-stable at low prevalence, because a fixed score cut-off applied to a finite batch produces a random number of alerts whose variability is dominated by the legitimate tail; its standard deviation was 6.5 alerts at 20 basis points against a mean of 24.1. The testing rule produced fewer alerts with lower dispersion (mean 7.3, standard deviation 2.1) because it responded to the evidence actually present in the batch.

Table 3. Experiment 4. Alert volume, precision, and realized error rates for the gradient boosting score under the two decision rules (20 replications).

Rule	Prevalence	Mean alerts	SD of alerts	Precision	FDR
BH on conformal p	0.002	7.3	2.1	0.911	0.089
	0.005	18.4	3.0	0.934	0.066
	0.010	38.5	5.5	0.907	0.093
	0.020	75.9	7.7	0.897	0.103

Rule	Prevalence	Mean alerts	SD of alerts	Precision	FDR
Top 1% score rule	0.002	24.1	6.5	0.296	0.704
	0.005	34.7	7.1	0.531	0.469
	0.010	52.0	4.9	0.688	0.312
	0.020	86.2	11.4	0.801	0.199

5.3 Sensitivity to covariate drift

Table 4 reports Experiment 3. Drift breaks exchangeability between calibration and test samples, and the resulting failure of error control is both severe and early. At $\delta = 0.25$, a shift of one fifth of a standard deviation in the legitimate log-amount distribution together with proportionally smaller shifts in four other coordinates, the realized false discovery rate for the gradient boosting score rose from 0.094 to 0.628 and alert volume rose from 34 to 148. At $\delta = 1.0$ the procedure alerted on nearly a fifth of the entire batch at a false discovery rate of 0.954. The logistic score degraded more slowly — 0.204 at $\delta = 0.25$ — but failed in the same manner. Power was essentially unaffected throughout, which is the diagnostic signature of the failure: the additional alerts are drift-induced false positives, not newly detected fraud.

The uniformity diagnostic tracked this closely. The Kolmogorov–Smirnov statistic on null p-values rose from about 0.015 with no drift to 0.092 at $\delta = 0.25$ and 0.330 at $\delta = 1.0$ for gradient boosting, against an approximate five percent critical value of 0.019 at the relevant sample size. Drift of a magnitude sufficient to destroy error control is therefore comfortably detectable from the p-value distribution alone, once labels are available.

Rolling recalibration on a recent window restored validity: the realized false discovery rate returned to zero at every drift level. It did so by destroying power, which fell to between 0.004 and 0.087. The reason is that the rolling window is drawn from live traffic and therefore contains undetected fraud at 20 basis points, and the next section shows that contamination at that level is by itself sufficient to suppress detection. The naive remedy for drift thus fails for the reason examined next, and the two problems must be solved together.

Table 4. Experiment 3. Effect of covariate drift on error control at $\alpha = 0.10$ and one percent prevalence, with static and rolling calibration (20 replications).

Score	Calibration	Drift δ	Alerts	FDR	Power	KS statistic
Gradient boosting	Static	0.00	33.9	0.094 (0.010)	0.658	0.015
		0.25	148.0	0.628 (0.066)	0.661	0.092
		0.50	350.0	0.800 (0.054)	0.677	0.175
		1.00	980.4	0.954 (0.007)	0.699	0.330
	Rolling	0.25	1.2	0.000	0.023	0.014
		1.00	0.3	0.000	0.005	0.016
Logistic	Static	0.00	33.3	0.093 (0.010)	0.649	0.014

Score	Calibration	Drift δ	Alerts	FDR	Power	KS statistic
		0.25	42.6	0.204 (0.017)	0.651	0.111
		0.50	60.0	0.443 (0.024)	0.656	0.218
		1.00	628.2	0.941 (0.004)	0.687	0.414
	Rolling	0.25	4.6	0.000	0.087	0.013
		1.00	2.9	0.000	0.060	0.014

The rolling window comprises 20,000 recent cases containing fraud at 20 basis points. Approximate 5% critical value for the Kolmogorov–Smirnov statistic at this sample size is 0.019.

5.4 Calibration contamination and trimmed calibration

Table 5 reports Experiment 2. Contamination does not violate the guarantee; it makes the procedure conservative to the point of uselessness. At a contamination rate of only 50 basis points — that is, one hundred undetected fraud cases in a calibration sample of twenty thousand — the gradient boosting detector issued no alerts at all in any replication, against 37.6 alerts and power of 0.657 with a clean sample. The mechanism is arithmetic. Contaminating cases occupy the extreme upper tail of the calibration score distribution, so the smallest attainable conformal p-value rises to approximately ε, and the Benjamini–Hochberg threshold for the first rejection is α/m; detection becomes impossible unless $\varepsilon < \alpha/m$, a condition violated by four orders of magnitude at realistic contamination rates. The uniformity diagnostic barely registered the problem: the Kolmogorov–Smirnov statistic rose only from 0.015 to 0.039 even at 500 basis points of contamination, so contamination is not detectable by the monitor that detects drift.

Table 5. Experiment 2. Effect of calibration contamination on the Benjamini–Hochberg procedure at $\alpha = 0.10$ and one percent test prevalence (20 replications).

Score	Contamination ε	Alerts	FDR	Power	KS statistic
Gradient boosting	0.000	37.6	0.093 (0.014)	0.657 (0.004)	0.015
	0.005	0.0	0.000	0.000	0.015
	0.010	0.3	0.000	0.006	0.016
	0.020	0.0	0.000	0.000	0.020
	0.050	0.0	0.000	0.000	0.039
Logistic	0.000	35.0	0.094 (0.015)	0.648 (0.003)	0.014
	0.005	0.4	0.000	0.007	0.013
	0.010	0.0	0.000	0.000	0.015
	0.050	0.0	0.000	0.000	0.039

Table 6 reports Experiment 4, in which the upper γ fraction of calibration scores is discarded before computing p-values. The pattern is sharp and identical across both supervised scores and both

configurations. With true contamination at 100 basis points, trimming at 50 basis points — strictly less than ε — restored 0.548 power for gradient boosting with a realized false discovery rate of exactly zero and precision of 1.000, recovering more than four fifths of the power available under clean calibration while remaining conservative. Trimming at exactly ε inflated the false discovery rate to 0.395, and trimming at 2ε and 3ε pushed it to 0.692 and 0.796. The same threshold behaviour appeared in the drifted rolling-recalibration configuration, where the true contamination rate was 20 basis points: trimming at zero was valid but nearly powerless, and trimming at 50 basis points — now above ε — produced a false discovery rate of 0.415.

The explanation is that the upper tail of a contaminated calibration sample is a mixture. Because thirty-five percent of fraud in our generator is mimicry, contaminating cases do not fully occupy the top ε fraction of the score distribution; trimming at $\gamma = \varepsilon$ therefore discards a substantial mass of genuine legitimate extremes along with the fraud, and the resulting reference distribution understates how extreme legitimate behaviour can be. The practical rule that follows is unambiguous and asymmetric in its consequences: trim conservatively, at a fraction materially below any plausible lower bound on the contamination rate. Under-trimming costs power in a controlled and observable way; over-trimming silently voids the guarantee that motivated the entire construction.

Table 6. Experiment 4. Trimmed calibration. Panel A: contamination $\varepsilon = 0.010$, no drift. Panel B: drift $\delta = 0.5$ with rolling recalibration, $\varepsilon = 0.002$. Nominal $\alpha = 0.10$, 20 replications.

Panel / score	Trim γ	Alerts	FDR	Power	Precision
A. Gradient boosting (ε = 0.010)	0.000	0.0	0.000	0.000	—
	0.005	26.9	0.000	0.548 (0.012)	1.000
	0.010	54.2	0.395 (0.016)	0.669 (0.005)	0.605
	0.020	107.9	0.692 (0.010)	0.680 (0.006)	0.308
	0.030	164.1	0.796 (0.007)	0.687 (0.006)	0.204
A. Logistic (ε = 0.010)	0.005	24.8	0.000	0.537 (0.018)	1.000
	0.010	53.9	0.436 (0.018)	0.656 (0.004)	0.564
	0.030	162.6	0.803 (0.007)	0.697 (0.007)	0.197
B. Gradient boosting (ε = 0.002)	0.000	3.9	0.000	0.073 (0.034)	1.000
	0.005	56.5	0.415 (0.014)	0.654 (0.003)	0.585
	0.010	84.0	0.605 (0.011)	0.656 (0.003)	0.395
	0.030	195.7	0.826 (0.006)	0.674 (0.005)	0.174
B. Logistic (ε = 0.002)	0.000	2.5	0.000	0.051 (0.015)	1.000
	0.005	55.6	0.414 (0.020)	0.651 (0.003)	0.586
	0.030	197.4	0.829 (0.006)	0.678 (0.006)	0.171

Configurations in which $\gamma < \varepsilon$ are conservative; configurations in which $\gamma \geq \varepsilon$ are anti-conservative, with the violation growing rapidly in γ.

5.5 Summary of the diagnostic

The Kolmogorov–Smirnov statistic on null p-values was well below its critical value in every valid configuration (0.013 to 0.016), rose by a factor of six or more under the drift levels that broke error control, and was almost unresponsive to contamination. It is therefore a good drift alarm and a poor contamination alarm. This asymmetry is intrinsic: drift moves the whole null distribution, which a distributional test detects, whereas contamination perturbs only the extreme upper tail, which contributes negligibly to a supremum-norm statistic computed over the full unit interval. A contamination monitor must therefore look at the tail directly — for example, by tracking the eventual confirmed-fraud rate among historical calibration windows once verification completes.

6. Discussion

6.1 What the framework buys, and what it does not

It does not buy detection power. At every prevalence tested, the testing rule and the fixed-quantile rule detected nearly the same proportion of fraud, and neither exceeded roughly two thirds, the ceiling imposed by the mimicry subpopulation. What it buys is a queue whose composition is known and stable, an operating point that adapts automatically to changes in prevalence, and an auditable statement about expected false alarms. In a supervisory environment that increasingly demands documented error characteristics for automated decisions, that statement has value independent of any improvement in accuracy.

It also buys a form of safety. The isolation forest results illustrate the point: a detector whose ranking cannot support discoveries produces no discoveries, rather than an alert list that looks the same as any other while being ninety-five percent noise.

6.2 A deployment protocol

- Maintain a calibration sample of at least 20,000 cases, and in any event large relative to m/α, drawn from the most mature verification vintage available so that residual contamination is as low as possible.
- Estimate a lower bound on the residual contamination rate ε from the eventual confirmed-fraud rate in fully verified historical windows, and trim at γ no greater than half that bound. Never trim at a fraction that could equal or exceed ε.
- Recalibrate on a rolling window at a frequency governed by monitored drift rather than by the calendar, recognizing that recalibration imports contamination and must be paired with trimming.
- Monitor the Kolmogorov–Smirnov statistic of null p-values retrospectively as each verification vintage matures; treat any sustained excursion above the critical value as invalidating the error guarantee until recalibration.
- Report realized precision on completed vintages against the nominal level. A persistent gap in either direction indicates contamination (conservative) or drift (anti-conservative), and the two are distinguishable by the diagnostic.
- Retain a capacity-driven fallback. When the procedure returns fewer alerts than capacity, the residual capacity is best spent on randomized audit sampling of low-score cases, which is the only mechanism that supplies unbiased labels for the selectively unlabelled region.

6.3 Relation to the wider literature

Our findings are consistent with the theoretical results of Bates et al. (2023) in the exchangeable regime and extend them by quantifying the two departures that matter in fraud. The drift results are the empirical counterpart of the coverage-gap bounds of Barber et al. (2023): validity degrades continuously in the distance between calibration and test distributions, but our results show that in the extreme-imbalance regime the practically relevant degradation occurs at distances that would be considered small in most applications. The contamination results connect the conformal literature to the classical robustness concern that a reference sample assumed clean rarely is, and identify a regime — small α, large m, small ε — in which the loss is total rather than gradual.

7. Limitations

The evidence here is simulation evidence. The generator captures imbalance, a hard mimicry subpopulation, correlated mixed-type covariates, and controlled drift, but it does not reproduce the temporal dependence of real transaction streams, the network structure of organized fraud, verification latency as a dynamic process, or the feedback loop by which the incumbent detector determines which labels are ever observed. Drift is modelled as covariate shift in legitimate behaviour only; adversarial drift, in which fraud tactics move in response to detection, is likely to be more damaging and is not addressed. Detection thresholds and trimming rules were evaluated at a single nominal level, and the sharp $\gamma < \varepsilon$ boundary we report is likely to depend on the mimicry fraction, which governs how completely contaminating cases occupy the calibration upper tail. Finally, we treat the false discovery rate as the operational target; where losses are strongly instance-dependent, a value-weighted error criterion would be more appropriate, and extending the procedure to weighted discovery rates is a natural next step.

8. Conclusion

Replacing a capacity threshold with a multiple-testing procedure on conformal anomaly p-values converts a fraud alert queue from an object with unknown statistical properties into one carrying an explicit, interpretable guarantee, at no cost in detection power. The guarantee is real but conditional, and this study locates the conditions precisely. Covariate drift of a magnitude that would pass unnoticed in most monitoring regimes inflated the realized false discovery rate from 0.09 to 0.63; contamination of the calibration sample at 50 basis points suppressed detection entirely. Trimming the calibration tail recovers most of the lost power provided the trimming fraction stays strictly below the true contamination rate, and voids the guarantee otherwise. For practitioners, the message is that distribution-free error control is achievable in fraud triage but requires active management of calibration hygiene and drift, and that a uniformity diagnostic on null p-values monitors one of these two failure modes well and the other not at all.

References

Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications, 68*, 90–113.

Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning, 16*(4), 494–591.

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.

Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. *Annals of Statistics, 51*(2), 816–845.

Bates, S., Candès, E., Lei, L., Romano, Y., & Sesia, M. (2023). Testing for outliers with conformal p-values. *Annals of Statistics, 51*(1), 149–178.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*(1), 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics, 29*(4), 1165–1188.

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems, 50*(3), 602–613.

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science, 17*(3), 235–255.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences, 557*, 317–331.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), 1–58.

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

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.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233–240).

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association, 99*(465), 96–104.

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

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys, 46*(4), 1–37.

Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning, 77*(1), 103–123.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature, 585*, 357–362.

Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: A review of anomaly detection techniques and recent advances. *Expert Systems with Applications, 193*, 116429.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *Proceedings of the 8th IEEE International Conference on Data Mining* (pp. 413–422).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

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

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B, 64*(3), 479–498.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods, 17*, 261–272.

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world.* New York: Springer.

Wedge, R., Kanter, J. M., Veeramachaneni, K., Rubio, S. M., & Perez, S. I. (2019). Solving the false positives problem in fraud prediction using automated feature engineering. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2018)* (pp. 372–388).