WHEN DOES REFERENCE PURIFICATION HELP? CONTAMINATION-AWARE REPRESENTATION LEARNING FOR UNSUPERVISED ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised anomaly detectors often learn representations and reference distributions from unlabeled data assumed to be predominantly normal. Unknown contamination can distort both the learned geometry and the reference bank used for scoring, but removing suspected anomalies can itself discard useful normal structure. We study when reference purification helps and when it does not. We introduce **CI-DCUAD**, which combines representation learning with **contamination-invariant reference selection (CI-RS)**. Preliminary unsupervised scores select observations for representation warmup and final reference construction, while main-stage training returns to the complete unlabeled dataset. We characterize *effective contamination* and show that selection reduces anomalous reference mass exactly when normal observations are retained more frequently than anomalies, $q_N > q_A$. Purification alone, however, does not guarantee better detection, which also requires preserving normal support and informative reference geometry.

A controlled separability–contamination study comprising 840 runs identifies distinct failure, success, and abstention regimes. Purification is not practically reliable at zero separation, is reliable in every valid nonempty positive-separation condition, and yields empty selections in two extreme conditions. Across five tabular benchmarks, CI-selected reference construction accounts for most of the favorable F1 change on CREDITCARD2013 and NSL-KDD, but produces unfavorable effects on the remaining datasets. Neither reference-only nor CI-Full significantly exceeds the matched baseline over the complete structural-ablation grid after Holm correction, although CI-Full significantly exceeds both incomplete CI variants. These results show that reference purification is a conditional mechanism, not a universal robustness strategy, and characterize the conditions under which it succeeds, fails, or should abstain.

1 INTRODUCTION

Unsupervised anomaly detection (UAD) is important when abnormal events matter but reliable labels are scarce, including fraud detection, cybersecurity, industrial monitoring, and Internet-of-Things systems. Many UAD methods learn a representation from unlabeled observations and subsequently detect anomalies using reconstruction, density, isolation, or distance-based scores. Their effectiveness, however, often depends on the training data being predominantly normal.

Real training data may contain an unknown and substantial fraction of anomalies, together with missingness, measurement noise, and partial observability. Contamination is then more than a threshold-calibration problem: anomalous observations can influence representation learning and the reference distribution used for scoring, thereby changing the geometry on which the detector operates. Once this influence has entered the representation or reference bank, changing the final decision threshold cannot remove it.

A natural response is to *purify* the observations that most directly influence the detector. In a fully unsupervised setting, however, neither the latent labels nor the true contamination proportion are available. Moreover, selecting a cleaner subset is not necessarily beneficial: a rule may remove anomalies while also discarding normal observations required to represent the normal distribution.

Central question. *When does reference purification help representation learning and reference-based scoring for unsupervised anomaly detection under unknown contamination?*

We study this question through **CI-DCUAD**, a contamination-aware UAD framework combining defect-aware representation learning with **contamination-invariant reference selection (CI-RS)**. CI-RS uses preliminary unsupervised scores to identify observations associated with a higher-normality score-space component. The selected observations control an early representation-learning warmup and construction of the final reference bank. Subsequent main-stage learning returns to the complete unlabeled training set, so excluded observations are not permanently removed from optimization. CI-RS requires neither ground-truth labels nor knowledge of the true contamination proportion.

Our analysis distinguishes nominal contamination from *effective contamination*, the anomalous mass remaining in the selected reference distribution. If q_N and q_A are the probabilities that normal and anomalous observations are retained, respectively, then selection reduces effective contamination exactly when $q_N > q_A$. This condition guarantees a cleaner selected distribution, not improved detection. Downstream benefit also depends on retaining useful normal support and producing informative reference geometry. Reference purification is therefore a conditional mechanism rather than a universal robustness guarantee.

We evaluate this mechanism through a controlled experiment spanning 42 separability–contamination conditions and 20 seeds per condition, together with experiments on five tabular anomaly-detection benchmarks. The evaluation covers structured defects, training contamination up to 75%, matched structural ablations, detector comparisons, and failure regimes. The controlled study distinguishes reliable purification, ineffective selection, and empty-selection abstention, while the real-data results show strongly dataset-dependent effects. These findings characterize when purification succeeds or fails rather than claiming uniform detector superiority.

Our contributions are:

- We formulate contamination-aware UAD in terms of *effective contamination* in the observations used for representation warmup and reference-based scoring, without estimating the unknown nominal contamination proportion.
- We introduce **CI-RS** and integrate it into **CI-DCUAD** to control early representation learning and final reference-bank construction using label-free score-space structure, while retaining the complete unlabeled dataset during subsequent training.
- We use an exact effective-contamination identity and a selected-gradient decomposition bound to separate preferential normal retention from selection-induced loss of normal support, and empirically characterize the corresponding success, failure, and abstention regimes.

2 Related Work

Unsupervised anomaly detection. Classical UAD methods identify unusual observations through isolation, local density, or one-class boundaries, including Isolation Forest Liu et al. (2008), Local Outlier Factor Breunig et al. (2000), and One-Class SVM Schölkopf et al. (2001). Deep methods instead learn representations through reconstruction, latent compactness, or learned anomaly scores, as in DAGMM Zong et al. (2018) and Deep SVDD Ruff et al. (2018). More recent methods develop specialized representations and scoring mechanisms, including ICL Shenkar & Wolf (2022), ECOD Li et al. (2023), DIF Xu et al. (2023a), and SLAD Xu et al. (2023b) for tabular anomaly detection. Diffusion Time Estimation (DTE) derives anomaly scores from estimated diffusion-time distributions Livernoche et al. (2024), while Deep Orthogonal Hypersphere Compression (DOHSC) learns compact orthogonally projected representations across tabular, image, and graph settings Zhang et al. (2024). Time-Conditioned Contraction Matching (TCCM) uses one-step flow-based deviation scores for scalable semi-supervised tabular detection Li et al. (2025). These advances improve representation geometry or scoring, but do not directly control how unknown anomalies in fully unlabeled training data influence both representation learning and reference construction.

Contaminated training and data refinement. Training anomalies may become reconstructable or distort the latent geometry and reference distribution used for scoring. General robust-learning strategies address corrupted observations through trimming, reweighting, robust estimation, and sample

selection Rousseeuw & Hubert (2011), while positive-unlabeled and mixture-proportion methods study mixtures with incomplete class information du Plessis et al. (2014); Jain et al. (2016). More directly, Latent Outlier Exposure (LOE) Qiu et al. (2022) jointly infers latent normal/anomalous assignments while updating a detector under contamination. Self-supervise, Refine, Repeat (SRR) Yoon et al. (2022) iteratively refines unlabeled data using an ensemble of one-class classifiers. Wu et al. Wu et al. (2022) analyze the degradation of deep detectors under contamination and mitigate it using a small auxiliary set of labeled anomalies. ADBench Han et al. (2022) further documents substantial variation across detectors, datasets, supervision regimes, anomaly types, and corruptions.

Related vision-based methods also control the observations defining normal reference structure. SoftPatch Jiang et al. (2022) identifies patch-level noise before constructing a reliability-weighted memory bank. FUN-AD Im et al. (2025) iteratively constructs confident normal sets from potentially contaminated industrial images, whereas MuSc Li et al. (2024) mutually scores neighborhoods across unlabeled test images without target-specific training. Although image-specific, these methods provide complementary evidence that the composition of the reference set materially affects anomaly scoring.

CI-DCUAD differs in both setting and intervention. It targets tabular data with unknown training contamination and structured feature defects, using preliminary score-space structure to control early representation warmup and final reference-bank construction; main-stage learning then returns to the complete unlabeled dataset. It therefore neither assigns persistent latent classes as in LOE, repeatedly refines the training set as in SRR, nor requires auxiliary labeled anomalies as in Wu et al. Unlike SoftPatch, FUN-AD, and MuSc, it does not rely on image patches or industrial-vision backbones. We characterize when this reference-purification mechanism succeeds, abstains, or fails rather than claiming general superiority over contaminated-data anomaly detection.

Data defects and post-hoc calibration. Missingness and measurement noise alter how observations are represented, whereas contamination alters which latent structures participate in learning and reference construction. CI-DCUAD combines defect-aware representation learning with contamination-aware reference selection to address both sources of distortion.

Post-hoc calibration selects an operating threshold after anomaly scores have been constructed. It cannot remove contamination already incorporated into a representation or reference bank. CI-DCUAD instead intervenes before final scoring. Its objective is therefore not universal robustness, but to determine when reference purification reduces effective contamination while retaining sufficient normal support for informative scoring.

3 Problem Formulation

Contaminated and defective observations. Let $x \in \mathbb{R}^d$ denote a sample and $y \in \{0, 1\}$ its unobserved normal ($y = 0$) or anomalous ($y = 1$) indicator. The unlabeled training distribution is

$$P_X = (1 - \alpha)P_N + \alpha P_A, \tag{1}$$

where P_N and P_A are the normal and anomalous distributions and the contamination proportion $\alpha \in [0, 1]$ is unknown. Structured defects are represented by

$$(x', m) = \mathcal{D}(x; \lambda_m, \lambda_n), \tag{2}$$

where x' is the potentially corrupted observation, $m \in \{0, 1\}^d$ is its observation mask, and λ_m and λ_n control missingness and additive-noise severity. The detector observes (x', m) but neither y nor α.

Representation and reference-based scoring. A mask-aware encoder maps each observation to

$$z = f_\theta([x', m]), \qquad f_\theta : \mathbb{R}^{2d} \to \mathbb{R}^p. \tag{3}$$

For a reference index set $I_{\text{ref}} \subseteq \{1, \ldots, n\}$, define

$$\mathcal{R}_\theta = \left\{ f_\theta([x'_j, m_j]) : j \in I_{\text{ref}} \right\}. \tag{4}$$

The detector scores a query using its mean squared distance to the k nearest reference embeddings:

$$s_\theta(x') = \frac{1}{k} \sum_{j \in \text{kNN}(z; \mathcal{R}_\theta)} \|z - z_j\|_2^2. \tag{5}$$

Consequently, anomalous observations admitted to $\mathcal{R}_\theta$ can alter the neighborhoods against which subsequent observations are evaluated.

Effective contamination and learning objective. Let $R \in \{0, 1\}$ indicate whether an observation is retained by the reference-selection rule. Its *effective contamination* is

$$\alpha_{\text{eff}} = \Pr(Y = 1 \mid R = 1). \tag{6}$$

Unlike the nominal contamination α, this quantity describes the anomalous mass that actually remains in the selected reference distribution. Although Y is unavailable during training, α_{eff} provides a mechanism-level quantity for analysis. The purification objective is

$$\alpha_{\text{eff}} < \alpha \tag{7}$$

when the observable data contain sufficient information for selection, while preserving normal support needed for informative reference-based scoring. The objective is not to recover the latent labels or estimate α.

Conditional identifiability. Let $S = s(X)$ be a preliminary anomaly score, with conditional distributions $P_{S|N}$ and $P_{S|A}$. Reference purification is possible only when their observable score-space structure permits preferential retention of normal observations. As these distributions overlap, the information available for unsupervised selection deteriorates. We therefore treat robustness as conditional on score-space identifiability and examine when reduced effective contamination improves detection, when normal-support loss offsets that benefit, and when insufficient separability prevents reliable selection.

4 METHOD: CI-DCUAD

CI-DCUAD combines defect-aware representation learning with contamination-invariant reference selection (CI-RS). An initial representation produces preliminary scores from which CI-RS selects observations with stronger normality evidence. The selected subset controls representation warmup and final reference-bank construction, while main-stage training returns to the complete unlabeled dataset. Thus, unselected observations are neither assumed anomalous nor permanently removed. The complete pipeline is illustrated in Appendix I.

Defect-aware representation learning. For each observation x_i, the defect operator $\mathcal{D}$ generates two independently corrupted observation–mask pairs $(x_i^{(v)}, m_i^{(v)})$, $v \in \{1, 2\}$. The mask-aware encoder produces

$$z_i^{(v)} = f_\theta([x_i^{(v)}, m_i^{(v)}]), \qquad v \in \{1, 2\}, \tag{8}$$

and is trained using

$$\mathcal{L}_{\text{DC}} = \text{InfoNCE}\left(z_i^{(1)}, z_i^{(2)}\right). \tag{9}$$

Views of the same observation form a positive pair and other observations provide contrasting examples. This encourages stability under the evaluated defects but does not itself prevent anomalous observations from influencing representation learning.

Contamination-invariant reference selection. An encoder initialized on the complete unlabeled training set produces

$$z_i^{(0)} = f_{\theta_0}([x_i, \mathbf{1}]), \qquad r_i = -\mathcal{S}(z_i^{(0)}), \tag{10}$$

where $\mathcal{S}$ is an unsupervised anomaly-scoring function and larger r_i indicates stronger normality evidence. CI-RS robustly normalizes these scores as

$$u_i = \frac{r_i - \text{median}(r)}{\text{MAD}(r) + \varepsilon}. \tag{11}$$

A two-component Gaussian mixture is fitted to the normalized scores:

$$p(u) = \sum_{\ell=1}^{2} \omega_\ell \mathcal{N}(u; \mu_\ell, \sigma_\ell^2), \qquad \ell^\star = \arg\max_{\ell \in \{1,2\}} \mu_\ell. \tag{12}$$

The larger-mean component is designated as normal-associated because the score orientation is fixed in advance. If $\gamma_{i\ell} = P(\ell \mid u_i)$ is its posterior responsibility, CI-RS selects

$$I_{\mathrm{CI}} = \{i : \gamma_{i\ell^\star} \geq \pi\}, \tag{13}$$

where π is a fixed posterior-confidence threshold. The mixture components describe score-space structure and are not treated as recovered ground-truth classes.

Two-stage learning and reference scoring. The selected indices define both the warmup and reference sets:

$$I_{\mathrm{warm}} = I_{\mathrm{ref}} = I_{\mathrm{CI}}, \tag{14}$$

with I_{ref} capped at N_{ref} when necessary. Let $\widehat{P}_{\mathrm{CI}}$ and $\widehat{P}_n$ denote the empirical distributions of the selected and complete training sets. The two learning stages optimize

$$\underbrace{\min_\theta \mathbb{E}_{x\sim\widehat{P}_{\mathrm{CI}}}[\mathcal{L}_{\mathrm{DC}}(x;\theta)]}_{\text{CI-controlled warmup}} \quad\longrightarrow\quad \underbrace{\min_\theta \mathbb{E}_{x\sim\widehat{P}_n}[\mathcal{L}_{\mathrm{DC}}(x;\theta)]}_{\text{main-stage training}}. \tag{15}$$

Selection therefore changes the warmup distribution but not the data used in main-stage learning.

After training, the final reference bank is

$$\mathcal{R}_\theta = \{f_\theta([x_j, \mathbf{1}]) : j \in I_{\mathrm{ref}}\}, \tag{16}$$

and anomaly scores follow the kNN rule in equation 5. Binary decisions use the fixed alert-mass protocol in Section 6. Neither selection nor scoring uses the true contamination proportion.

Component attribution. CI-RS affects both representation warmup and reference construction. We therefore compare a full-data baseline, warmup-only and reference-only variants, and complete CI-DCUAD. The variants share the encoder, contrastive objective, main-stage schedule, and kNN scorer, isolating the two uses of selection. Their formal definitions are provided in Appendix I.6; the complete training and decision pipeline is given in Appendix B.1. Ground-truth labels are used only for post-hoc evaluation and mechanism analysis.

5 Mechanism Analysis of Reference Purification

We provide mechanism-level accounting for two effects of reference selection: the anomalous mass remaining after selection and the deviation of the selected warmup gradient from its clean-normal counterpart. These results are exact identities and bounds used to organize the empirical analysis; they are not presented as convergence, generalization, or detector-optimality guarantees.

5.1 Effective Contamination After Selection

Let $Y \in \{0, 1\}$ be the latent normal/anomalous indicator and U the normalized score used by CI-RS. Its distribution is

$$P_U = (1-\alpha)P_{U|Y=0} + \alpha P_{U|Y=1}, \tag{17}$$

where $\alpha = P(Y = 1)$. For posterior threshold π, define the selection region and class-conditional retention probabilities as

$$A_\pi = \{u : \gamma_{\ell^\star}(u) \geq \pi\}, \qquad q_N = P(U \in A_\pi \mid Y = 0), \qquad q_A = P(U \in A_\pi \mid Y = 1). \tag{18}$$

The quantities q_N and q_A are used only for analysis; they are not available to CI-RS during training.

Effective-contamination identity. If $P(U \in A_\pi) > 0$, Bayes' rule gives

$$\alpha_{\mathrm{eff}} = P(Y = 1 \mid U \in A_\pi) = \frac{\alpha q_A}{(1-\alpha)q_N + \alpha q_A}. \tag{19}$$

For $0 < \alpha < 1$, direct rearrangement yields

$$\alpha_{\mathrm{eff}} < \alpha \quad\Longleftrightarrow\quad q_N > q_A. \tag{20}$$

Thus, purification occurs exactly when selection preferentially retains normal observations; it does not require estimating α. The derivation and a bounded-selection-error consequence are given in Appendix B.3.

5.2 Effect on CI-Controlled Warmup

Let $g(x;\theta) = \nabla_\theta \mathcal{L}_{\mathrm{DC}}(x;\theta)$. If $g_N^{\mathrm{sel}}(\theta)$ and $g_A^{\mathrm{sel}}(\theta)$ are the conditional mean gradients of selected normal and anomalous observations, then

$$g_{\mathrm{sel}}(\theta) = (1-\alpha_{\mathrm{eff}})g_N^{\mathrm{sel}}(\theta) + \alpha_{\mathrm{eff}} g_A^{\mathrm{sel}}(\theta). \tag{21}$$

Let $g_N(\theta) = \mathbb{E}_{X\sim P_N}[g(X;\theta)]$ be the clean normal-data gradient and define the normal-selection shift

$$\delta_N(\theta) = \left\| g_N^{\mathrm{sel}}(\theta) - g_N(\theta) \right\|_2. \tag{22}$$

Selected-gradient decomposition bound. If $\|g(x;\theta)\|_2 \le G$ for the observations under consideration, the decomposition above and the triangle inequality give

$$\|g_{\mathrm{sel}}(\theta) - g_N(\theta)\|_2 \le (1-\alpha_{\mathrm{eff}})\delta_N(\theta) + 2G\alpha_{\mathrm{eff}}. \tag{23}$$

The bound separates residual anomalous influence, represented by α_{eff}, from distortion caused by selecting a non-representative subset of the normal population, represented by $\delta_N(\theta)$. Lower effective contamination is therefore insufficient by itself: selection must also preserve useful normal support. The derivation is given in Appendix B.4.

5.3 Observable Diagnostic and Scope

For diagnostic analysis, we summarize fitted score-space separation by

$$\mathrm{Sep} = \frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}}. \tag{24}$$

This label-free statistic is not used to train the encoder or tune hyperparameters. Greater fitted separation can support more decisive posterior assignments, but does not prove that the fitted components correspond to the latent populations; the purification condition remains $q_N > q_A$.

Accordingly, the analysis does not claim that lower effective contamination must improve AUROC, PR-AUC, or F1, or that selected-data warmup converges to training on clean normal data. The experiments evaluate the connection through controlled separability conditions, structural ablations, effective- contamination measurements, representation diagnostics, and explicit failure regimes.

6 Experimental Setup

The experiments test whether CI-RS reduces effective contamination when score-space structure supports selection, whether this purification improves detection, and where the mechanism fails. Matched comparisons use identical experimental conditions, and labels are accessed only for post-hoc evaluation and mechanism analysis.

Datasets and evaluation conditions. We evaluate five tabular benchmarks spanning cybersecurity, IoT, and financial fraud: CSE-CIC-IDS2018 (hereafter CICIDS2018) Canadian Institute for Cybersecurity (CIC), University of New Brunswick (2018), CICIoT2023 Neto et al. (2023); Canadian Institute for Cybersecurity (CIC), University of New Brunswick (2023), CreditCard2013 Dal Pozzolo et al. (2018); MLG ULB (2018), IEEE-CIS Kaggle (2019), and NSL-KDD Tavallaee et al. (2009). For each of the predefined seeds $\{7, 13, 21, 37, 51\}$, we construct a fixed 70/30 train–test split and reuse it across matched comparisons.

Controlled training contamination varies over $\alpha \in \{0.01, 0.05, 0.10, 0.25, 0.50, 0.75\}$; α constructs the experimental condition but is not supplied to CI-DCUAD. Missingness, additive noise, and mixed defects are evaluated at severities $\lambda \in \{0, 0.01, 0.02, 0.05, 0.10\}$. The primary defect grid contains five datasets, five seeds, three defect types, and five severity levels, yielding 375 matched configurations. Dataset statistics, preprocessing, and perturbation details are provided in Appendices I.1 and I.2.

Configuration and comparisons. The architecture and primary training configuration are fixed across datasets. CI-RS applies Isolation Forest to the initial embeddings, robustly normalizes its scores, and fits the two-component mixture described in Section 4. Parameters are fixed without anomaly labels; posterior-threshold and reference-cap sensitivity are evaluated separately.

The matched DCAD baseline uses the same encoder, contrastive objective, optimization schedule, and kNN scorer without CI-based selection. Structural comparisons among `dcad_base`, `ci_warm_only`, `ci_ref_only`, and `ci_full` isolate the effects of selected warmup and selected reference construction. External comparisons include Isolation Forest, LOF, OC-SVM, HBOS, PCA, COPOD, ECOD, autoencoder, VAE, Deep SVDD, DIF, ICL, SLAD, LOE Qiu et al. (2022), and a paper-based raw-tabular SRR reproduction (SRR-Raw) Yoon et al. (2022). All use fixed configurations without dataset-specific search, test-label tuning, or post-hoc score-sign selection. LOE and SRR-Raw are evaluated only at the common focus condition. Exact configurations and variant definitions are reported in Appendices I.2 and I.6.

Metrics and mechanism measurements. We report AUROC and average precision (AP) as ranking metrics and anomaly-class binary F1 under a fixed alert-mass protocol with $\beta = 0.05$. In all evaluations, $Y = 1$ denotes an anomaly and larger final kNN distances denote stronger anomaly evidence; AUROC and AP are computed using this fixed orientation. The alert mass is an externally specified operating constraint, not an estimate of the contamination proportion. Mechanism measurements include effective contamination α_{eff}, purification gain $\Delta\alpha = \alpha - \alpha_{\text{eff}}$, selected fraction $\rho_{\text{sel}} = |I_{\text{ref}}|/n$, and fitted score-space separation. We also examine the post-hoc association between reference-purity gain and matched F1 change. These label-dependent measurements are diagnostic rather than deployment inputs or causal estimates. Complete definitions are given in Appendix I.3.

Controlled experiment and statistical analysis. To isolate CI-RS from dataset-specific representation effects, we construct an equal-variance Gaussian score-space experiment with standardized separation $\delta \in \{0, 0.5, 1, 1.5, 2, 3, 4\}$ and the six contamination levels above. The resulting 42 conditions are repeated over 20 seeds, producing 840 runs. Empty selections are recorded explicitly and are not treated as successful purification.

Detector comparisons are paired by dataset, defect type, severity, and seed. Statistical conclusions consider effect magnitude, direction, dataset heterogeneity, and empty-selection behavior in addition to significance. Controlled-experiment parameters, confidence intervals, multiplicity corrections, sensitivity analyses, and reproducibility controls are detailed in Appendices I.3 and I.5.

7 Results and Discussion

We examine the central question in three stages: detector-level effects under a common focus condition, controlled evidence for the purification mechanism, and structural attribution of the two CI-RS intervention points. We then summarize the observed failure regimes and statistical interpretation.

Detector-level effects are heterogeneous. Table 1 compares CI-DCUAD with the structurally matched DCAD baseline under mixed defects at $\lambda = 0.05$. Both rows of each paired comparison come from the same controlled reference-selection experiment and share the data split, encoder, training objective, optimization schedule, and kNN scorer; DCAD does not use CI-based selection. The separate four-variant structural-ablation grid is analyzed independently in Appendix E.1.

The effects differ across datasets and metrics. All three metrics improve on NSL-KDD, CreditCard2013, and IEEE-CIS, whereas all decline on CICIoT2023. On CICIDS2018, AUROC and AP increase while fixed-mass F1 decreases, showing that improved global ranking need not improve the selected operating point.

DCAD is a structurally matched attribution control rather than a competitive-performance reference. A five-seed implementation audit found no label-coding or metric-implementation error, but identified a dataset-specific reference-inversion failure on CICIoT2023. Its reconstructed training split contains 97.67% attacks; the preliminary Isolation Forest consequently assigns attacks higher normality scores, and every audited CI-RS reference set contains 100% attacks. The resulting distance score therefore reverses the intended ranking: CI-DCUAD obtains AUROC 0.112, while post-hoc score negation

Table 1: Mean change from DCAD to CI-DCUAD under mixed defects at $\lambda = 0.05$ over five seeds. Positive values favor CI-DCUAD; F1 is anomaly-class binary F1 under fixed alert-mass calibration with $\beta = 0.05$.

Dataset	ΔF1	ΔAUROC	ΔAP
CICIDS2018	-0.021	$+0.029$	$+0.004$
CICIoT2023	-0.012	-0.221	-0.024
CreditCard2013	$+0.015$	$+0.162$	$+0.006$
IEEE-CIS	$+0.015$	$+0.077$	$+0.015$
NSL-KDD	$+0.064$	$+0.108$	$+0.151$

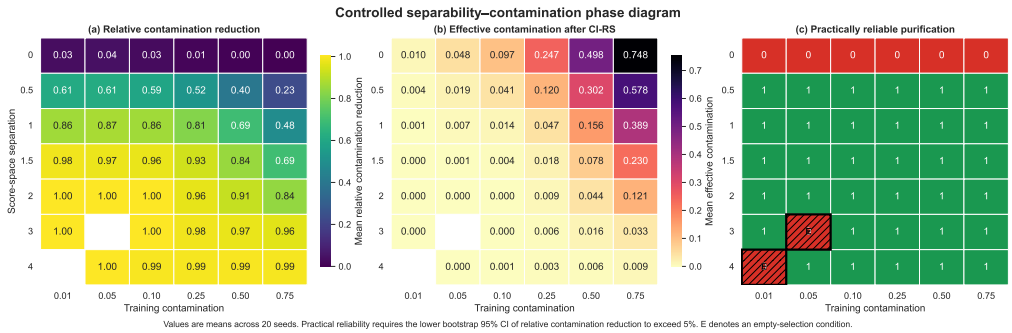


Figure 1: Controlled separability–contamination experiment over 840 runs. Panels show mean relative contamination reduction, mean effective contamination, and whether the lower bootstrap 95% confidence bound on relative reduction exceeds 5%. Hatched "E" cells are empty selections and are treated as abstention failures.

gives approximately 0.889. Activation diagnostics do not indicate severe global encoder collapse. Thus, improvements relative to DCAD isolate the effect of CI-RS within the shared pipeline but do not establish general competitiveness. Full diagnostics and external comparisons are reported in Appendix C.1 and Appendices C.4, G.1, and G.2.

For additional external context at the same focus condition, LOE Qiu et al. (2022) obtains five-dataset mean F1/AUROC/AP of $0.056/0.580/0.365$, while SRR-Raw Yoon et al. (2022) obtains $0.080/0.612/0.381$. These comparisons are descriptive rather than seed-paired and remain strongly dataset dependent. Complete per-dataset results and implementation provenance are reported in Table 21.

Purification requires informative score structure. The effective-contamination identity in Section 5.1 shows that selection reduces contamination exactly when $q_N > q_A$. We next test whether increasing score-space separation supports this preferential-retention condition.

At zero separation, mean relative contamination reduction is approximately 0–4%, and none of the six contamination levels satisfies the predefined reliability criterion. In contrast, all 34 nonempty positive-separation conditions satisfy it, with purification increasing with separation under both Spearman and Kendall tests (all $p \leq 0.031$). Two extreme conditions, $(\delta, \alpha) = (3, 0.05)$ and $(4, 0.01)$, produce empty selections across all 20 seeds, representing conservative abstention rather than perfect purification.

A matched-grid experiment with variance-standardized Student-t_5 scores tests whether this behavior survives non-Gaussian model misspecification. Among the 36 positive-separation conditions, 18 satisfy the reliability criterion; every condition with nonempty selections across all 20 seeds is reliable. The other conditions exhibit empty-selection instability or abstention. Thus, separation-dependent purification persists under heavy tails when selection remains operational, although its reliable region is narrower than in the Gaussian experiment.

Together, the experiments identify negative-control, success, instability, and abstention regimes, but do not imply that purification improves detection unless useful normal support is preserved.

Full non-Gaussian results and supporting mixture, purity, and selection-control analyses appear in Appendix I.4, Appendix C.2, Appendix C.3, and Appendix C.4.

Component attribution is dataset dependent. CI-RS controls both representation warmup and final reference construction. We compare `dcad_base`, `ci_warm_only`, `ci_ref_only`, and `ci_full`. Because warmup can alter the representation subsequently used for scoring, these variants constitute a structural ablation rather than an additive causal decomposition.

Across 75 matched conditions, warmup-only has a mean F1 difference of -0.0025 from DCAD, with 95% CI $[-0.0040, -0.0011]$. Reference-only and CI-Full have mean differences of $+0.0234$ and $+0.0262$, respectively, but neither exceeds DCAD significantly after joint Holm correction over all six pairwise structural contrasts (both adjusted $p = 0.158$); only 40% of their condition-level effects are positive.

CI-Full effects are -0.0332 on CICIDS2018, -0.0037 on CICIoT2023, $+0.0577$ on CreditCard2013, -0.0111 on IEEE-CIS, and $+0.1215$ on NSL-KDD. Reference-only follows the same qualitative pattern. Reference construction therefore accounts for most of the favorable change on CreditCard2013 and NSL-KDD, but its effect is unfavorable on the other three datasets.

CI-Full nevertheless exceeds warmup-only by $+0.0287$ (95% CI $[0.0164, 0.0416]$, Holm-adjusted $p = 0.0359$) and reference-only by $+0.0028$ (95% CI $[0.0017, 0.0039]$, adjusted $p = 1.03 \times 10^{-4}$). The complete configuration thus improves over both incomplete CI variants without establishing uniform superiority over DCAD. The complete ablation grid and threshold-specific checks are reported in Appendices E.1 and F.1.

Failure regimes and statistical interpretation. Increasing corruption can coincide with reduced score-space separation, higher effective contamination, smaller purification gain, and deteriorating AUROC, AP, and F1. This co-variation is consistent with the proposed mechanism but is not a causal decomposition: normal support, representation quality, and neighborhood geometry may deteriorate together. Complete NSL-KDD and IEEE-CIS failure panels are provided in Appendix C.5.

8 Conclusion

CI-DCUAD uses preliminary score-space structure to control representation warmup and reference construction while returning to the complete unlabeled dataset during main-stage training. Purification occurs exactly when $q_N > q_A$, but detection benefit also requires preserved normal support and informative geometry. Controlled experiments identify failure, success, and abstention regimes; across real datasets, favorable effects concentrate on CreditCard2013 and NSL-KDD, while neither CI-Full nor reference-only significantly exceeds DCAD over the complete ablation grid after Holm correction. The CICIoT2023 audit further shows that extreme anomaly-majority training can invert reference construction even without global encoder collapse. Thus, reference purification is conditional rather than universally robust. Scope, reproducibility, and extended analyses are reported in Appendix A, Appendix D.3, and Appendix G.3.

Ethics Statement

This work uses publicly available benchmark datasets and does not involve human-subject recruitment, user interaction, or the collection of new personally identifiable information. Ground-truth labels are used only for controlled data construction and post-hoc evaluation, not during unsupervised training, reference selection, scoring, or threshold calibration. Because anomaly-detection systems may affect security, fraud investigation, and other high-impact decisions, their predictions should not be treated as autonomous evidence against individuals. The observed dataset-dependent failure regimes also caution against deployment without domain-specific validation, monitoring, and appropriate human oversight.

AI USE STATEMENT

Generative AI tools were used to assist with literature discovery, manuscript organization, language editing, LaTeX formatting, code development and debugging, and feedback on experimental design and result interpretation. They were not used to generate experimental data or independently establish the paper's scientific claims. All cited works were manually verified, all AI-assisted code and analyses were executed and checked by the authors, and all reported results were obtained from the documented experimental pipeline. The authors reviewed the complete manuscript and take responsibility for its content, claims, citations, code, and experimental results.

REFERENCES

M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: Identifying density-based local outliers. In *Proc. ACM SIGMOD Int. Conf. Management of Data*, pp. 93–104, Dallas, TX, USA, 2000. doi: 10.1145/342009.335388.

Canadian Institute for Cybersecurity (CIC), University of New Brunswick. CSE-CIC-IDS2018 dataset, 2018. [Online]. Available: https://www.unb.ca/cic/datasets/ids-2018.html.

Canadian Institute for Cybersecurity (CIC), University of New Brunswick. CICIoT2023 dataset, 2023. [Online]. Available: https://www.unb.ca/cic/datasets/iotdataset-2023.html.

Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi, and Gianluca Bontempi. Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8):3784–3797, 2018. doi: 10.1109/TNNLS.2017.2736643.

M. du Plessis, G. Niu, and M. Sugiyama. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. ADBench: Anomaly detection benchmark. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Jiin Im, Yongho Son, and Je Hyeong Hong. FUN-AD: Fully unsupervised learning for anomaly detection with noisy training data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 9447–9456, 2025. doi: 10.1109/WACV61041.2025.00915.

Shantanu Jain, Martha White, and Predrag Radivojac. Estimating the class prior and posterior from noisy positives and unlabeled data. In *Advances in Neural Information Processing Systems*, volume 29, pp. 2693–2701, 2016.

Xi Jiang, Jianlin Liu, Jinbao Wang, Qiang Nie, Kai Wu, Yong Liu, Chengjie Wang, and Feng Zheng. SoftPatch: Unsupervised anomaly detection with noisy data. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15433–15445, 2022. doi: 10.52202/068431-1123.

Kaggle. IEEE-CIS fraud detection, 2019. [Online]. Available: https://www.kaggle.com/c/ieee-fraud-detection.

Xurui Li, Ziming Huang, Feng Xue, and Yu Zhou. MuSc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=AHgc5SMdtd`.

Z. Li, Y. Zhao, X. Hu, N. Botta, C. Ionescu, and G. H. Chen. ECOD: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12181–12193, 2023. doi: 10.1109/TKDE.2022.3159580.

Zhong Li, Qi Huang, Yuxuan Zhu, Lincen Yang, Mohammad Mohammadi Amiri, Niki van Stein, and Matthijs van Leeuwen. Scalable, explainable and provably robust anomaly detection with one-step flow matching. In *Advances in Neural Information Processing Systems*, volume 38, pp. 97620–97685, 2025. doi: 10.52202/085713-2936.

F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *Proc. IEEE Int. Conf. Data Mining (ICDM)*, pp. 413–422, Pisa, Italy, 2008. doi: 10.1109/ICDM.2008.17.

Victor Livernoche, Vineet Jain, Yashar Hezaveh, and Siamak Ravanbakhsh. On diffusion modeling for anomaly detection. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=lR3rk7ysXz.

MLG ULB. Credit card fraud detection dataset. Kaggle, 2018. [Online]. Available: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/data.

Euclides Carlos Pinto Neto, Sajjad Dadkhah, Raphael Ferreira, Alireza Zohourian, Rongxing Lu, and Ali A. Ghorbani. CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment. *Sensors*, 23(13):5941, 2023. doi: 10.3390/s23135941.

Karl Pearson. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895. doi: 10.1098/rspl.1895.0041.

Chen Qiu, Aodong Li, Marius Kloft, Maja Rudolph, and Stephan Mandt. Latent outlier exposure for anomaly detection with contaminated data. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18153–18167. PMLR, 2022.

P. J. Rousseeuw and M. Hubert. Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):73–79, 2011. doi: 10.1002/widm.2.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4393–4402. PMLR, 2018.

B. Schölkopf et al. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001. doi: 10.1162/089976601750264965.

T. Shenkar and L. Wolf. Anomaly detection for tabular data with internal contrastive learning. In *Proc. Int. Conf. Learning Representations (ICLR)*, 2022.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. A detailed analysis of the KDD CUP 99 data set. In *Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pp. 1–6, 2009. doi: 10.1109/CISDA.2009.5356528.

Shuang Wu, Jingyu Zhao, and Guangjian Tian. Understanding and mitigating data contamination in deep anomaly detection: A kernel-based approach. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 2319–2325, 2022. doi: 10.24963/ijcai.2022/322.

H. Xu, G. Pang, Y. Wang, and Y. Wang. Deep isolation forest for anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12591–12604, 2023a. doi: 10.1109/TKDE.2023.3270293.

H. Xu, Y. Wang, J. Wei, S. Jian, Y. Li, and N. Liu. Fascinating supervisory signals and where to find them: Deep anomaly detection with scale learning. In *Proc. Int. Conf. Machine Learning (ICML)*, volume 202 of *PMLR*, pp. 38655–38673, 2023b.

Jinsung Yoon, Kihyuk Sohn, Chun-Liang Li, Sercan O. Arik, Chen-Yu Lee, and Tomas Pfister. Self-supervise, refine, repeat: Improving unsupervised anomaly detection. *Transactions on Machine Learning Research*, 2022.

Yunhe Zhang, Yan Sun, Jinyu Cai, and Jicong Fan. Deep orthogonal hypersphere compression for anomaly detection. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=cJs4oE4m9Q.

Y. Zhao, Z. Nasrullah, and Z. Li. PyOD: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96):1–7, 2019.

B. Zong et al. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *Proc. Int. Conf. Learning Representations (ICLR)*, Vancouver, Canada, 2018.

A Detailed Limitations and Reproducibility

Limitations. The evaluated missingness, additive noise, and mixed defects are controlled stress conditions; they do not represent feature-dependent missingness, temporal drift, cross-feature shifts, or adaptive perturbations. Similarly, training contamination up to 75% tests prevalence sensitivity rather than worst-case robustness.

CI-RS may fail when score populations overlap, anomalies receive normal-like scores, or the fitted mixture is weakly identifiable. The CICIoT2023 audit demonstrates a more extreme failure mode: when the training population is overwhelmingly anomalous, the preliminary detector can treat attack structure as normal and construct an attack-dominated reference bank. CI-RS can then amplify, rather than correct, the preliminary ranking error. Moreover, lower effective contamination is insufficient if selection removes normal support needed for informative reference geometry. Empty controlled selections show that the fixed posterior threshold can also abstain too aggressively. The present study does not establish a validated label-free criterion for detecting either failure before deployment or selecting a fallback; developing such a safeguard remains future work.

The Gaussian and Student-t_5 experiments isolate separability under controlled light- and heavy-tailed score models, but do not reproduce the dimensionality, multimodality, dependence, or learning dynamics of real data. Evaluation is limited to five tabular benchmarks and one principal architecture, so transfer to other architectures, modalities, and streaming settings remains unestablished. Absolute detector performance is weak in some dataset–metric conditions, and classical or modern baselines outperform CI-DCUAD in several comparisons. The matched DCAD analysis therefore supports component attribution within the shared pipeline, not a claim of general competitiveness. Fixed alert-mass calibration standardizes the F1 operating point but cannot repair degraded representations, references, or rankings.

Reproducibility. Experiments use fixed configurations, predefined seeds $\{7, 13, 21, 37, 51\}$, shared splits for matched comparisons, and independently recomputed CI-RS selections. The primary grids contain 375 matched configurations and 840 controlled runs, including 40 explicitly recorded empty selections. Labels are used only for post-hoc evaluation and mechanism diagnostics. Processing, configurations, measurements, statistical procedures, and coverage checks are documented in Appendix I.1 and Appendix H.2; tables and figures are generated programmatically from retained run-level outputs.

B Formalization and Conditions for CI-RS

This appendix provides a formal description of the CI-DCUAD detector and characterizes the condition under which contamination-aware reference selection (CI-RS) reduces the effective contamination of the selected set. The analysis separates three questions: how the complete detector operates, how CI-RS behaves under different score-space regimes, and what condition is required for successful purification.

B.1 Formal Detector Definition and Decision Pipeline

Let

$$\mathcal{X}_{\text{train}} = \{x_i\}_{i=1}^{n} \subset \mathbb{R}^d$$

denote an unlabeled training set with unknown contamination proportion α, and let $\mathcal{X}_{\text{test}}$ denote the evaluation set. Ground-truth labels are not used during representation learning, CI-RS selection, anomaly scoring, or threshold calibration.

Initial representation. An encoder f_{θ_0} is first trained on the full unlabeled training set under the initial mild-defect stage. The resulting embeddings are evaluated under full observation:

$$z_i^{(0)} = f_{\theta_0}(x_i).$$

Preliminary normality scoring. Isolation Forest is applied to the initial embeddings to obtain a preliminary anomaly score $\mathcal{S}(z_i^{(0)})$. We convert this to a normality-oriented score

$$r_i = -\mathcal{S}(z_i^{(0)}),$$

so that larger values indicate observations that appear more normal according to the preliminary scorer.

The scores are robustly standardized using the median and median absolute deviation (MAD):

$$u_i = \frac{r_i - \mathrm{median}(r)}{\mathrm{MAD}(r) + \epsilon}.$$

CI-RS selection. A two-component Gaussian mixture model is fitted to the one-dimensional normalized score distribution:

$$p(u) = \sum_{k=1}^{2} \omega_k \mathcal{N}(u; \mu_k, \sigma_k^2),$$

where ω_k denotes the fitted mixture weight.

The component with the larger mean normality score is designated as the normal-associated component,

$$k^\star = \arg\max_{k \in \{1,2\}} \mu_k.$$

For posterior threshold π, CI-RS selects

$$\mathcal{I}_{\mathrm{CI}} = \{i : P(k^\star \mid u_i) \geq \pi\}.$$

The fitted components are statistical components of the preliminary score distribution; they are not ground-truth normal and anomalous classes.

Warmup and main training. The CI-selected observations are used during the CI-controlled warmup stage. After warmup, the main representation-learning stage returns to the full unlabeled training set. Thus, CI-RS does not restrict all training epochs to $\mathcal{I}_{\mathrm{CI}}$.

Let $f_{\theta^\star}$ denote the final encoder obtained after the main training stage.

Reference-bank construction. The selected indices are retained for construction of the final reference bank using the final encoder:

$$\mathcal{R}_{\mathrm{CI}} = \{f_{\theta^\star}(x_i) : i \in \mathcal{I}_{\mathrm{CI}}\}, \qquad |\mathcal{R}_{\mathrm{CI}}| \leq N_{\mathrm{ref}},$$

where N_{ref} is the reference-bank cap.

Anomaly scoring. For a test observation x, let

$$z = f_{\theta^\star}(x).$$

Its anomaly score is the mean distance to its k nearest neighbors in the reference bank:

$$s(x) = \frac{1}{k} \sum_{z_j \in \mathcal{N}_k(z; \mathcal{R}_{\mathrm{CI}})} \|z - z_j\|_2.$$

Table 2: Audit of the five primary CI-DCUAD runs for the CREDITCARD2013 focus condition. The alert fraction is reconstructed as $r = pR/P$.

Seed	Prevalence	Alert fraction	Precision	Recall	F1
7	0.001814	0.050011	0.016850	0.464516	0.032520
13	0.001814	0.050011	0.016382	0.451613	0.031617
21	0.001814	0.050011	0.018254	0.503226	0.035230
37	0.001814	0.050011	0.016382	0.451613	0.031617
51	0.001814	0.050011	0.018722	0.516129	0.036134

Decision rule. For a score set $\{s_i\}_{i=1}^n$ and alert-mass parameter β, the operating threshold is

$$t_\beta = \mathrm{Quantile}_{1-\beta}\left(\{s_i\}_{i=1}^n\right),$$

and the corresponding binary predictions are

$$\widehat{y}_i = \mathbb{1}[s_i > t_\beta].$$

The primary experiments use $\beta = 0.05$. Consequently, the realized predicted-positive fraction on the score set to which the quantile rule is applied is approximately 0.05, up to ties and the empirical-quantile convention.

Here, β specifies an externally chosen alert budget and is not an estimate of the unknown contamination proportion α. Thresholding determines the binary operating point only after anomaly-score construction; it does not protect the learned representation or reference bank from contamination.

CreditCard2013 operating-point audit. We additionally audited the five saved primary CI-DCUAD runs used for the CREDITCARD2013 focus condition. Let p denote anomaly prevalence, P precision, and R recall. The realized predicted-positive fraction can be reconstructed as

$$r = \frac{pR}{P},$$

because $P = \mathrm{TP}/N_{\mathrm{pred}}$ and $R = \mathrm{TP}/N_{\mathrm{pos}}$. For fixed prevalence p and alert fraction r, the maximum attainable F1 is

$$\mathrm{F1}_{\max}(p, r) = \frac{2\min(p, r)}{p + r}.$$

Across the five primary runs, $p = 0.001814$, the saved alert-mass parameter is $\beta = 0.05$, and the reconstructed alert fraction is approximately $r = 0.050011$. The mean saved F1 is 0.03342, whereas the corresponding upper bound is 0.07001. Every saved F1 value agrees with the value recomputed from its recorded precision and recall, up to floating-point precision. Thus, the reported CreditCard2013 F1 values are internally consistent with the evaluated operating point.

B.2 SCORE-SPACE SEPARABILITY REGIMES

The usefulness of CI-RS depends on whether the preliminary normality score contains sufficient information to preferentially retain normal observations. Let

$$S_\pi = \{P(k^\star \mid u) \geq \pi\}$$

denote the CI-RS selection event, and define

$$q_N = P(S_\pi \mid N), \qquad q_A = P(S_\pi \mid A),$$

where N and A denote the latent normal and anomalous populations, respectively. These quantities are used only for analysis; class labels are not available to CI-RS.

When the score distribution exhibits useful separation, posterior membership can preferentially retain normal observations, corresponding to $q_N > q_A$. In such a regime, the selected set can have lower effective contamination than the original training distribution.

As score-space overlap increases, posterior assignments become less discriminative and q_N and q_A can approach one another. In the limiting case $q_N = q_A$, selection does not change the contamination proportion. If $q_A > q_N$, the selected set can instead become more contaminated.

Consequently, fitted mixture separation should be interpreted as an empirical diagnostic rather than as a binary identifiability guarantee. Large separation between the fitted Gaussian components does not by itself establish that the components correspond to the latent normal and anomalous populations.

Moreover, contamination reduction alone is not sufficient for improved downstream detection. Let δ_N denote distortion or loss of useful normal support induced by selection. A selected set may have lower effective contamination while simultaneously removing normal regions needed for representative reference-bank coverage. CI-RS therefore depends jointly on preferential anomaly rejection and preservation of useful normal support.

B.3 Effective-Contamination Derivation

This subsection derives the effective-contamination identity in Section 5.1 and gives a bounded-selection-error consequence.

Derivation of the effective-contamination identity. By Bayes' rule,

$$\alpha_{\text{eff}} = P(Y = 1 \mid U \in A_\pi) \tag{25}$$

$$= \frac{P(U \in A_\pi \mid Y = 1)P(Y = 1)}{P(U \in A_\pi)}. \tag{26}$$

Using

$$P(U \in A_\pi \mid Y = 1) = q_A, \qquad P(Y = 1) = \alpha,$$

and the law of total probability,

$$P(U \in A_\pi) = (1 - \alpha)q_N + \alpha q_A, \tag{27}$$

we obtain

$$\alpha_{\text{eff}} = \frac{\alpha q_A}{(1 - \alpha)q_N + \alpha q_A}. \tag{28}$$

For $0 < \alpha < 1$ and $P(U \in A_\pi) > 0$,

$$\alpha_{\text{eff}} < \alpha \iff \frac{\alpha q_A}{(1 - \alpha)q_N + \alpha q_A} < \alpha \tag{29}$$

$$\iff q_A < (1 - \alpha)q_N + \alpha q_A \tag{30}$$

$$\iff (1 - \alpha)q_A < (1 - \alpha)q_N \tag{31}$$

$$\iff q_A < q_N. \tag{32}$$

The equality and reverse-inequality cases follow analogously:

$$\alpha_{\text{eff}} = \alpha \iff q_A = q_N, \qquad \alpha_{\text{eff}} > \alpha \iff q_A > q_N. \tag{33}$$

□

Corollary B.1 (Bounded selection errors)**.** *Suppose*

$$q_N \geq 1 - \eta_N, \qquad q_A \leq \eta_A. \tag{34}$$

Then

$$\alpha_{\text{eff}} \leq \frac{\alpha \eta_A}{(1 - \alpha)(1 - \eta_N) + \alpha \eta_A}. \tag{35}$$

Proof. The expression in equation 19 is nondecreasing in q_A and nonincreasing in q_N. Substituting the upper bound $q_A \leq \eta_A$ and lower bound $q_N \geq 1 - \eta_N$ therefore gives equation 35. □

The corollary describes the joint effect of anomalous observations retained by the selection rule and normal observations excluded by it. It does not require the fitted Gaussian-mixture components to recover the latent classes perfectly.

B.4 Selected-Gradient Decomposition Bound

This subsection derives the selected-gradient decomposition bound stated in Section 5.2. Define the conditional mean gradients under the normal and anomalous components of the selected distribution as

$$g_N^{\mathrm{sel}}(\theta) = \mathbb{E}\left[g(X;\theta) \mid Y = 0,\, U \in A_\pi\right] \tag{36}$$

and

$$g_A^{\mathrm{sel}}(\theta) = \mathbb{E}\left[g(X;\theta) \mid Y = 1,\, U \in A_\pi\right]. \tag{37}$$

The selected-distribution gradient is therefore

$$g_{\mathrm{sel}}(\theta) = (1-\alpha_{\mathrm{eff}})g_N^{\mathrm{sel}}(\theta) + \alpha_{\mathrm{eff}} g_A^{\mathrm{sel}}(\theta). \tag{38}$$

Derivation of the selected-gradient bound. Subtracting the clean normal-data gradient gives

$$g_{\mathrm{sel}} - g_N = (1-\alpha_{\mathrm{eff}})\left(g_N^{\mathrm{sel}} - g_N\right) \tag{39}$$

$$+\,\alpha_{\mathrm{eff}}\left(g_A^{\mathrm{sel}} - g_N\right). \tag{40}$$

Applying the triangle inequality,

$$\|g_{\mathrm{sel}} - g_N\|_2 \le (1-\alpha_{\mathrm{eff}})\|g_N^{\mathrm{sel}} - g_N\|_2 \tag{41}$$

$$+\,\alpha_{\mathrm{eff}}\|g_A^{\mathrm{sel}} - g_N\|_2. \tag{42}$$

By the definition in equation 22, the first norm is $\delta_N(\theta)$. Under the assumption $\|g(x;\theta)\|_2 \le G$, Jensen's inequality gives

$$\|g_A^{\mathrm{sel}}(\theta)\|_2 \le G, \qquad \|g_N(\theta)\|_2 \le G. \tag{43}$$

Consequently,

$$\|g_A^{\mathrm{sel}} - g_N\|_2 \le \|g_A^{\mathrm{sel}}\|_2 + \|g_N\|_2 \le 2G. \tag{44}$$

Substitution yields

$$\|g_{\mathrm{sel}}(\theta) - g_N(\theta)\|_2 \le (1-\alpha_{\mathrm{eff}})\delta_N(\theta) + 2G\alpha_{\mathrm{eff}}. \tag{45}$$

□

The bound does not establish convergence to the representation obtained from clean normal data. It isolates two contributions to the selected warmup gradient: residual anomalous influence through α_{eff} and selection-induced normal-distribution shift through $\delta_N(\theta)$.

C CI-RS Diagnostics and Failure Regimes

This section provides complementary diagnostics for understanding when contamination-aware reference selection (CI-RS) produces a useful selected set and when its benefit deteriorates. We examine four related aspects: (i) the preliminary score-space mixture structure used for selection, (ii) the empirical association between selection purity and downstream performance, (iii) a representative failure regime in which purification weakens, and (iv) a label-informed clean-reference diagnostic.

Ground-truth labels are used only in analyses explicitly identified as post-hoc oracle diagnostics. They are not used by CI-DCUAD during representation learning, CI-RS selection, reference construction, anomaly scoring, or threshold calibration.

C.1 Extended Detector-Level Baseline Results

The main paper reports compact changes relative to the structurally matched DCAD-BASE. This section provides the corresponding absolute detector results, reference-size controls, and oracle diagnostics under the same focus condition (mixed defects, $\lambda = 0.05$).

DCAD-BASE uses the complete training reference population. The random-reference control matches the size of the CI-selected reference bank without using CI-RS to determine its composition. It therefore tests whether the observed changes can be attributed only to reducing the reference-bank size. The oracle clean-reference variant uses ground-truth labels to construct a normal-only reference bank and is included solely as a post-hoc diagnostic. It is not available to the unsupervised method.

Table 3: Absolute detector performance from the paired controlled reference-selection experiment under the focus condition (mixed defects, $\lambda = 0.05$). Threshold-dependent results use fixed alert-mass calibration with $\beta = 0.05$. Reference size is reported to separate changes in reference composition from changes caused only by reference-bank size. The oracle clean-reference method is a label-informed post-hoc diagnostic. Dashes indicate that the corresponding oracle experiment was not available.

Dataset	Method	$F1_\mu$	$F1_\sigma$	$AUROC_\mu$	$AUROC_\sigma$	AP_μ	AP_σ	Seeds	Ref. size$_\mu$
CICIDS2018	DCAD-BASE	0.028	0.003	0.362	0.021	0.086	0.003	5	280000.000
CICIDS2018	Random-Ref	0.029	0.004	0.425	0.016	0.096	0.003	5	200000.000
CICIDS2018	**CI-DCUAD**	0.007	0.003	0.391	0.025	0.090	0.004	5	200000.000
CICIDS2018	Oracle Clean-Ref	0.424	0.032	0.838	0.027	0.518	0.052	5	200000.000
CICIoT2023	DCAD-BASE	0.092	0.001	0.333	0.044	0.965	0.003	5	210000.000
CICIoT2023	Random-Ref	0.092	0.001	0.326	0.108	0.965	0.007	5	131017.800
CICIoT2023	**CI-DCUAD**	0.080	0.003	0.112	0.037	0.941	0.008	5	131017.800
CICIoT2023	Oracle Clean-Ref	–	–	–	–	–	–	–	–
CreditCard2013	DCAD-BASE	0.018	0.003	0.687	0.043	0.012	0.005	5	199365.000
CreditCard2013	Random-Ref	0.021	0.004	0.710	0.054	0.017	0.008	5	174553.600
CreditCard2013	**CI-DCUAD**	0.033	0.003	0.849	0.021	0.018	0.002	5	174553.600
CreditCard2013	Oracle Clean-Ref	0.020	0.003	0.734	0.044	0.017	0.009	5	174553.600
IEEE-CIS	DCAD-BASE	0.077	0.010	0.566	0.026	0.046	0.005	5	413378.000
IEEE-CIS	Random-Ref	0.076	0.007	0.554	0.015	0.046	0.003	5	200000.000
IEEE-CIS	**CI-DCUAD**	0.092	0.004	0.643	0.020	0.061	0.005	5	200000.000
IEEE-CIS	Oracle Clean-Ref	0.086	0.003	0.574	0.016	0.050	0.002	5	200000.000
NSL-KDD	DCAD-BASE	0.074	0.007	0.471	0.035	0.448	0.018	5	103962.000
NSL-KDD	Random-Ref	0.072	0.007	0.424	0.027	0.427	0.014	5	68956.600
NSL-KDD	**CI-DCUAD**	0.138	0.014	0.579	0.024	0.599	0.028	5	68956.600
NSL-KDD	Oracle Clean-Ref	–	–	–	–	–	–	–	–

Result provenance and score-orientation audit. The absolute DCAD-BASE and CI-DCUAD results in Table 3 are paired within the same controlled reference-selection experiment family. They should not be combined with the separate four-variant structural-ablation grid, which serves a different component-attribution analysis.

We additionally audited the scoring and metric implementation. For both DCAD-BASE and CI-DCUAD, the final anomaly score is the unnegated mean distance to the k nearest reference embeddings:

$$s_i = \frac{1}{k} \sum_{z_j \in \mathcal{N}_k(z_i;\mathcal{R})} \|z_i - z_j\|_2.$$

Thus, larger scores indicate greater distance from the reference bank and are treated as stronger anomaly evidence. AUROC and AP are computed directly from these scores with $Y = 1$ denoting an anomaly, and fixed-mass predictions select the observations with the largest scores. No score negation or normality-oriented transformation is applied during final evaluation.

The low AUROC values are reproducible across all five predefined seeds and are not caused by a label-coding or AUROC-implementation error. Under the focus condition, mean AUROC is 0.362 for DCAD-BASE on CICIDS2018, 0.333 for DCAD-BASE on CICIoT2023, and 0.471 for DCAD-BASE on NSL-KDD; the corresponding CI-DCUAD values are 0.391, 0.112, and 0.579.

A post-hoc reconstruction isolates the CICIoT2023 failure. The training split contains $210{,}000$ observations with attack prevalence 0.9767. Isolation Forest therefore models the dominant attack population as normal: its normality score has AUROC 0.9872 when attack is treated as the positive class. Across all five seeds, both the warm and final CI-RS selections contain 100% attacks. The downstream kNN detector consequently treats attack embeddings as its reference distribution and assigns larger distances to the rare benign observations. This reference inversion explains why the reported anomaly score has AUROC approximately 0.112, whereas negating it post hoc yields approximately 0.889. The negated value is reported only as a diagnostic and is not substituted for the method's experimental result.

Activation-level collapse check. We additionally compared baseline and CI-full embeddings on a balanced diagnostic sample of $4{,}903$ benign and $4{,}903$ attack observations. The CI-full encoder has effective rank 2.28, compared with 1.78 for the baseline, and neither representation exhibits dead dimensions, near-constant dimensions, or duplicated embeddings under the predefined diagnostics. Although both representations are strongly variance-concentrated, CI-full does not satisfy the severe-

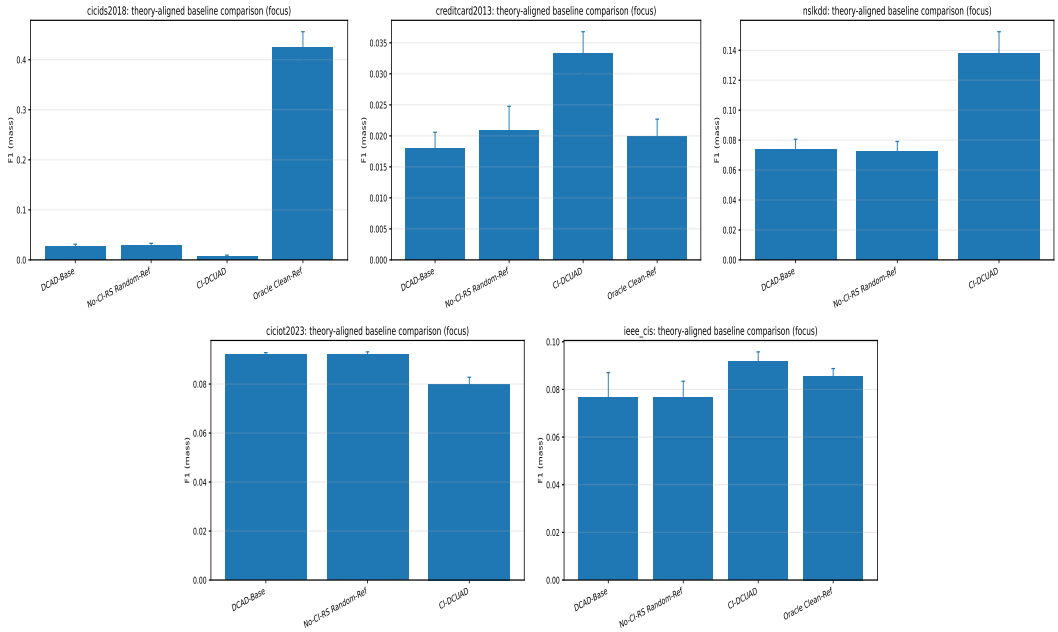


Figure 2: Per-dataset F1 comparison between CI-DCUAD and the matched DCAD-BASE under the controlled focus condition (mixed defect, $\lambda = 0.05$). Results are reported under fixed alert-mass calibration and illustrate the heterogeneous effect of CI-DCUAD across datasets.

collapse criteria. The evidence therefore does not support global encoder collapse as the primary cause of the CICIoT2023 failure; the attack-only reference construction is already sufficient to explain the observed ranking reversal.

The absolute results confirm the heterogeneous detector-level pattern summarized in Table 1. CI-DCUAD improves all three metrics on CreditCard2013, IEEE-CIS, and NSL-KDD. On CICIDS2018, AUROC and AP improve while fixed-mass F1 decreases. On CICIoT2023, all three metrics decrease.

The random-reference comparison shows that reference-bank size alone does not explain the behavior of CI-RS. Random selection can outperform CI-RS in some unfavorable settings, while failing to reproduce its gains in favorable settings such as CreditCard2013 and NSL-KDD. Reference composition, rather than only reference-set size, therefore contributes to the observed changes.

The oracle results should be interpreted cautiously. On CICIDS2018, the large oracle improvement indicates that substantially cleaner reference construction could be beneficial, but CI-RS does not recover that benefit under the evaluated score structure. On CreditCard2013 and IEEE-CIS, CI-DCUAD can exceed the reported oracle diagnostic because reference purity is not the only factor determining performance. Representation learning, retained normal support, and the resulting neighborhood geometry also affect the final detector. Accordingly, the oracle is not an upper bound on CI-DCUAD performance.

Figure 2 visualizes the corresponding per-dataset F1 results. It highlights that the aggregate improvement is concentrated in particular datasets and should not be interpreted as uniform detector dominance.

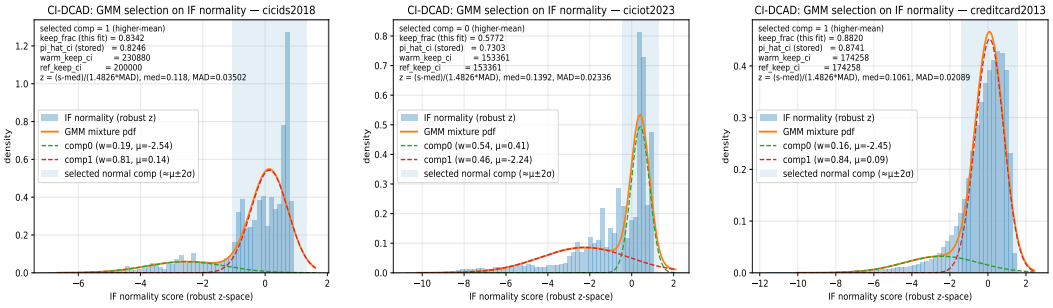
C.2 Score-Space Mixture Diagnostics

CI-RS operates on a one-dimensional preliminary normality-score distribution. For each dataset, we visualize the robustly normalized Isolation Forest (IF) normality scores together with the fitted two-component Gaussian mixture model (GMM).

The preliminary scores are obtained from embeddings produced by the initial encoder. Specifically, Isolation Forest produces a preliminary anomaly score $\mathcal{S}(\mathbf{z}_i)$, which is converted to the normality score

Table 4: Aggregate detector performance across the evaluated datasets under the focus-condition detector-comparison protocol. F1 is reported as mean ± standard deviation across datasets; AUROC and AP are dataset-level means. Results should be interpreted together with the heterogeneous per-dataset results in Table 3.

Method	F1	AUROC	AP	n_{datasets}
DCAD-Base	0.0577 ± 0.0327	0.4837	0.3114	5
No-CI-RS Random-Ref	0.0581 ± 0.0314	0.4879	0.3101	5
CI-DCUAD	0.0700 ± 0.0513	0.5148	0.3416	5
Oracle Clean-Ref	0.1765 ± 0.2168	0.7152	0.1951	3



(a) CICIDS2018 (b) CICIoT2023 (c) CreditCard2013

(d) IEEE-CIS (e) NSL-KDD

Figure 3: CI-RS score-space diagnostics across the five evaluated datasets. Each panel shows the robustly normalized Isolation Forest normality-score distribution together with the fitted two-component Gaussian mixture. CI-RS designates the component with the larger mean normality score as the normal-associated component and uses posterior membership in that component for reference selection. The components are fitted without class labels and should not be interpreted as recovered normal and anomalous classes.

$$r_i = -\mathcal{S}(\mathbf{z}_i)$$

and robustly normalized using the median and MAD to obtain u_i. CI-RS then fits the two-component GMM to $\{u_i\}_{i=1}^{n}$. The component with the larger mean normality is designated as the normal-associated component, and posterior membership in this component is used for CI-RS selection.

Figure 3 shows that the observable preliminary score structure varies substantially across datasets. The fitted mixture defines a posterior-based selection rule rather than an estimate of the unknown anomaly prevalence. Observations whose posterior membership in the normal-associated component exceeds the specified threshold form the CI-selected set.

More distinguishable components can provide a stronger basis for posterior selection, whereas substantial overlap makes component membership less decisive. However, fitted mixture separation alone does not establish that the selected set is cleaner. The normal-associated component can still contain anomalous observations, and useful normal observations can be excluded.

Table 5: Association between measured reference-set purity gain Δp and fixed-mass F1 gain $\Delta\mathrm{F1}$ under the focus condition. The analysis includes five predefined seeds for each of five datasets ($n = 25$). Ground-truth labels are used only after model execution to measure reference-set purity.

Statistic	Correlation	p-value
Pearson r	0.721	4.70×10^{-5}
Spearman ρ	0.230	0.268

The relevant purification condition remains preferential normal retention:

$$q_N > q_A.$$

As established in Section B.3, this condition is equivalent to reducing effective contamination, $\alpha_{\text{eff}} < \alpha$. The GMM plots characterize the observable information available to CI-RS; they do not independently verify that this condition holds.

The CI-selected observations are used during the CI-controlled warmup stage, and the selected indices determine the final reference bank. Subsequent main representation learning returns to the full unlabeled training distribution. Thus, these diagnostics characterize the preliminary selection mechanism rather than the final anomaly-score distribution.

Ground-truth labels are not used to fit the GMM, identify the normal-associated component, or perform posterior selection. Label-based purity and effective-contamination measurements considered below are post-hoc diagnostics used only to examine the behavior of CI-RS.

C.3 Purification–Performance Association

We examine whether stronger empirical purification is associated with larger downstream improvement. This is a post-hoc mechanism analysis: ground-truth labels are used only after model execution to measure the anomaly fraction remaining in the selected reference set. Labels are not used during representation learning, CI-RS selection, reference construction, anomaly scoring, or threshold calibration.

Let the measured purity gain be

$$\Delta p = p_{\text{anom}}(\text{base}) - p_{\text{anom}}(\text{CI}), \tag{46}$$

where p_{anom} is the anomaly fraction in the corresponding reference set. Thus, $\Delta p > 0$ means that CI-RS has reduced effective reference contamination. We define the associated detector change as

$$\Delta\mathrm{F1} = \mathrm{F1}_{\text{CI}} - \mathrm{F1}_{\text{base}}. \tag{47}$$

Focus-condition association. We first analyze the exact focus condition used for the matched detector-level comparison: mixed defects, $\lambda = 0.05$, and training contamination 0.05. The analysis contains five predefined seeds for each of the five datasets, giving $n = 25$ paired observations.

Table 5 shows a strong positive pooled linear association between measured purification and F1 gain (Pearson $r = 0.721$, $p = 4.70 \times 10^{-5}$). However, the rank association is substantially weaker and is not statistically significant (Spearman $\rho = 0.230$, $p = 0.268$).

The two statistics support a deliberately limited interpretation. The Pearson result indicates that larger numerical purity changes are associated with larger F1 changes in the pooled focus-condition observations. The non-significant Spearman result does not support a stable monotonic ordering across all observations. Consequently, purification magnitude should not be treated as a dataset-independent predictor of detector improvement.

With only five seeds per dataset, dataset-specific correlations under this single condition are descriptive. Their directions are heterogeneous: positive on CICIoT2023 and CreditCard2013, near zero on IEEE-CIS, and negative on CICIDS2018 and NSL-KDD. These estimates are not treated as separate confirmatory tests.

Broader contamination-stress association. We next complement the focus-condition analysis with a broader contamination-stress analysis that pools the available contamination-stress observations

Table 6: Per-dataset associations between measured purity gain Δp and downstream F1 gain ΔF1 in the contamination-stress analysis. Parentheses report two-sided p-values. Ground-truth labels are used only for this post-hoc oracle diagnostic.

Dataset	n	Pearson r (p)	Spearman ρ (p)
CICIoT2023	30	0.898 ($< 10^{-3}$)	0.545 (0.00183)
CICIDS2018	30	-0.882 ($< 10^{-3}$)	-0.918 ($< 10^{-3}$)
NSL-KDD	29	0.393 (0.0349)	0.501 (0.00566)
IEEE-CIS	30	0.323 (0.0815)	0.387 (0.0345)
CreditCard2013	30	-0.083 (0.663)	-0.087 (0.648)

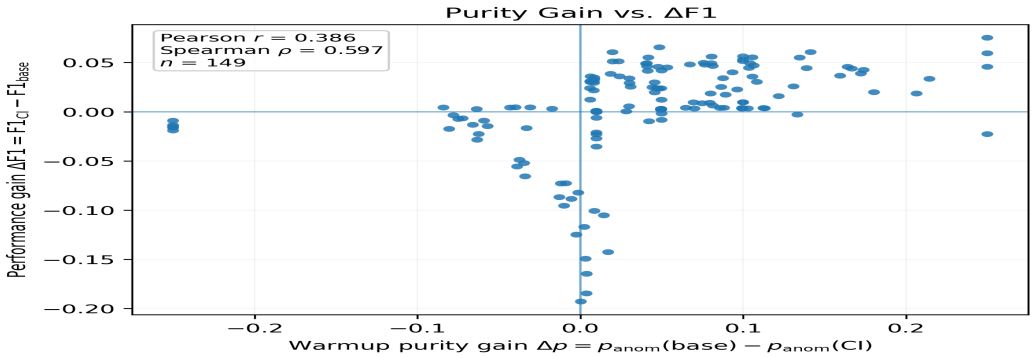


Figure 4: Association between measured warmup purity gain $\Delta p = p_{\text{anom}}(\text{base}) - p_{\text{anom}}(\text{CI})$ and downstream F1 gain $\Delta\text{F1} = \text{F1}_{\text{CI}} - \text{F1}_{\text{base}}$ over the broader contamination-stress analysis ($n = 149$ paired observations). The pooled association is positive (Pearson $r = 0.386$; Spearman $\rho = 0.597$), but the per-dataset results in Table 6 show substantial heterogeneity. Ground-truth labels are used only for this post-hoc oracle diagnostic.

within each dataset. This analysis evaluates whether the focus-condition association persists over a wider range of training contamination levels.

Pearson and Spearman correlation statistics are reported Pearson (1895); Spearman (1904). This analysis uses ground-truth labels only after model execution to measure the actual purity of the selected observations. These labels are not used by CI-RS during training, selection, reference construction, scoring, or calibration. The analysis is therefore an oracle diagnostic for mechanism interpretation rather than a component required by the deployed method.

The broader contamination-stress analysis is strongly heterogeneous across datasets. Positive purity–performance associations are observed for CICIoT2023, NSL-KDD, and, more moderately, IEEE-CIS; CreditCard2013 shows little association, whereas CICIDS2018 exhibits a strong negative association. Accordingly, the pooled positive association reported over the broader analysis should not be interpreted as a dataset-uniform relationship.

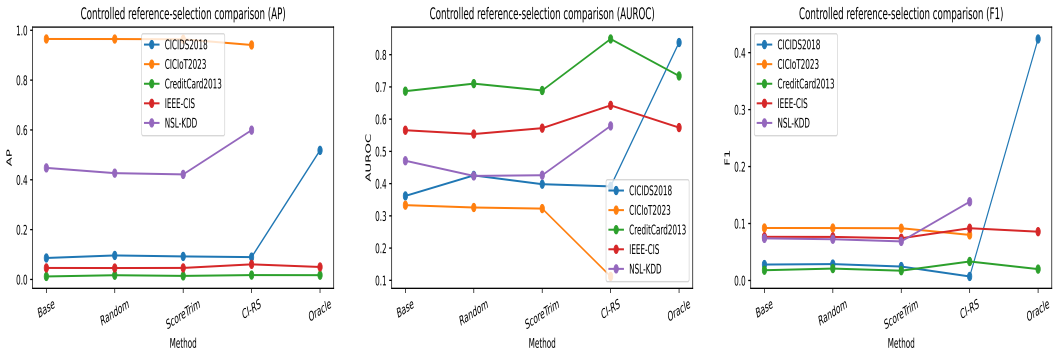
These observations reinforce the distinction between reference purification and downstream detector improvement. In the notation of the theoretical analysis, a favorable selector satisfies

$$q_N > q_A,$$

which implies

$$\alpha_{\text{eff}} < \alpha.$$

This establishes a condition for reducing effective reference contamination, not a guarantee of improved F1. Detection performance can additionally depend on preservation of useful normal support, representation quality, and the local geometry induced by the selected reference bank. We therefore use the purity–performance analysis as a post-hoc mechanism diagnostic rather than as causal evidence that purification itself produces the observed detector gains.



(a) Average Precision. (b) AUROC. (c) Fixed-mass F1.

Figure 5: Controlled reference-selection comparison under the focus condition. The matched base detector is compared with random reference selection, direct score-based trimming, CI-RS, and a label-informed oracle clean-reference diagnostic. Random-Ref controls for reference-bank size, while ScoreTrim-Ref tests whether direct preliminary-score ordering reproduces the behavior of posterior-based CI-RS.

C.4 Reference-Selection Controls

The preceding analysis examines whether measured purification is associated with downstream improvement. We next test whether the observed behavior of CI-RS can be explained by the simpler act of reducing the number of reference observations.

We compare five reference-construction strategies:

1. **DCAD-BASE**, which uses the unselected reference population;
2. **Random-Ref**, which randomly selects the same number of reference observations as CI-RS;
3. **ScoreTrim-Ref**, which directly retains observations according to their preliminary score ordering;
4. **CI-RS**, which uses posterior membership in the fitted higher-normality mixture component; and
5. **Oracle Clean-Ref**, which uses ground-truth labels to construct a normal-only reference bank and is included only as a post-hoc diagnostic.

Random-Ref controls for reference-bank size without using score information. ScoreTrim-Ref provides a stronger non-random control that uses preliminary score ordering but does not use the posterior mixture structure employed by CI-RS.

Figure 5 shows that changing reference construction can materially alter downstream detection. However, the direction and magnitude of the change remain dataset and metric dependent.

Random selection does not consistently reproduce the behavior of CI-RS. Therefore, the effect of CI-RS cannot generally be explained by reducing reference-bank size alone. Direct score trimming is also not uniformly equivalent to posterior-based CI-RS, indicating that the structure of the selection rule affects the resulting reference distribution.

In favorable settings, CI-RS recovers part of the benefit associated with a cleaner reference. Substantial gaps remain in other settings, and CI-RS can also perform worse than the simpler controls. The comparison therefore does not show that CI-RS consistently approximates the latent normal set.

The oracle clean-reference method must not be interpreted as a strict upper bound. A normal-only reference bank can still provide incomplete normal support, and CI-RS also affects the representation warmup. Final performance depends on representation quality, retained support, reference geometry, and the downstream operating point, not solely on the anomalous fraction of the reference bank.

Overall, these controls support a limited conclusion: CI-RS is not equivalent to random reference reduction or direct score trimming, but its more structured selection rule is not uniformly advantageous.

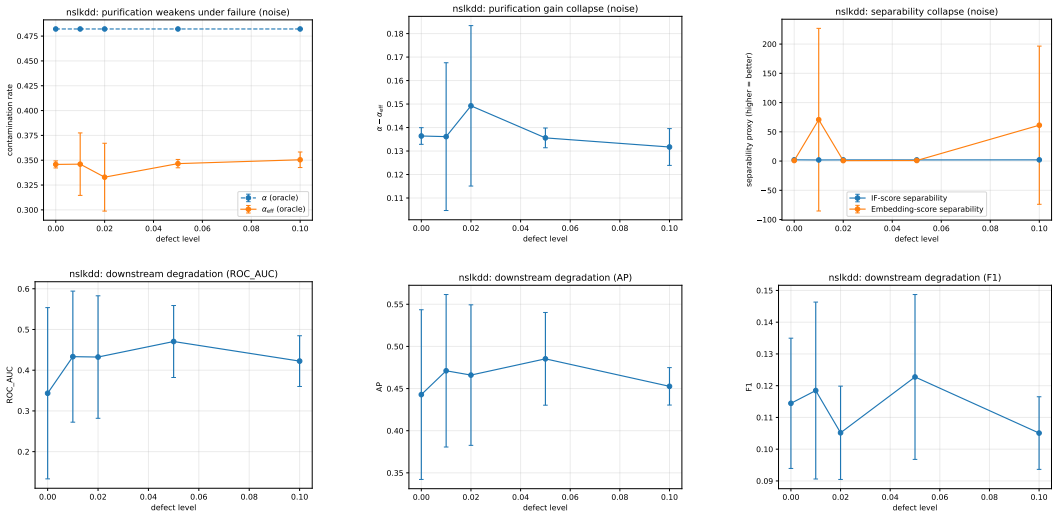


Figure 6: Representative failure-regime diagnostics on NSL-KDD. Top row: effective contamination, purification gain, and the observable score-space separability diagnostic as corruption increases. Bottom row: corresponding AUROC, AP, and fixed-mass F1. The figure shows association between deterioration of the selection mechanism and deterioration of downstream detection; it does not by itself identify a causal chain.

Its usefulness depends on whether the score distribution supports preferential normal retention without removing normal observations needed for informative reference-based scoring.

C.5 Failure-Case Diagnostics

We examine failure regimes on NSL-KDD and IEEE-CIS under increasing corruption strength. The analysis jointly tracks effective contamination, purification gain, preliminary score-space separation, and downstream AUROC, AP, and F1. Ground-truth labels are used only post hoc for the contamination and purification measurements. The trajectories describe empirical co-variation rather than establishing a unique causal pathway.

Ground-truth labels are used only post hoc to compute α_{eff} and the resulting purification diagnostic; they are not available to CI-RS during selection, representation learning, reference construction, scoring, or calibration. The separability proxy is likewise interpreted as a diagnostic of the available score structure rather than as a sufficient condition for successful purification.

The IEEE-CIS trajectories provide an additional dataset-level example of the failure behavior examined in the main text. As the preliminary score structure becomes less favorable for preferential selection, the observed purification benefit can contract while α_{eff} increases and downstream performance deteriorates. This pattern is consistent with the selection condition derived in Section B.3:

$$\alpha_{\text{eff}} = \frac{\alpha q_A}{(1-\alpha)q_N + \alpha q_A}, \qquad \alpha_{\text{eff}} < \alpha \iff q_A < q_N.$$

However, the plotted separability proxy does not directly establish $q_N > q_A$, and the joint deterioration of the diagnostics and detector metrics does not prove that reduced separability alone causes the downstream performance change. Normal-support preservation, representation geometry, and the resulting reference-based kNN structure can also affect detection.

The failure analysis therefore provides an observable diagnostic pattern associated with weakening CI-RS effectiveness under the evaluated corruption regime. It should not be interpreted as a universal predictor of failure or as evidence that every low-separability setting must produce the same downstream behavior.

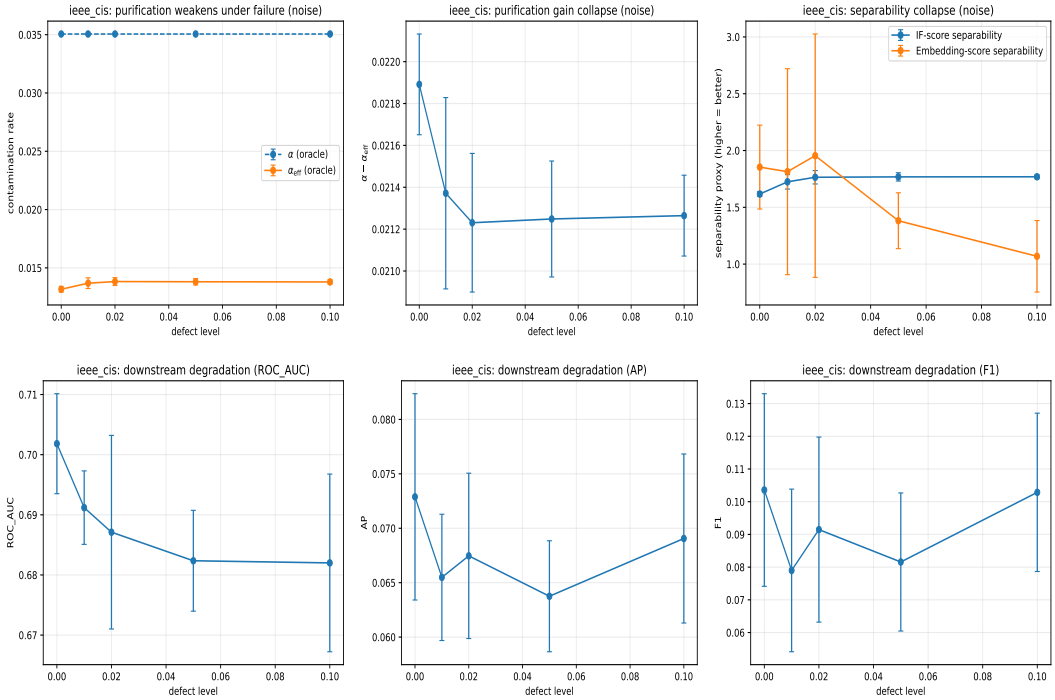


Figure 7: Failure-case diagnostics on IEEE-CIS under increasing corruption. Top row: effective contamination α_{eff} among the CI-selected observations, purification gain, and the preliminary score-space separability diagnostic. Bottom row: downstream AUROC, AP, and F1. Ground-truth labels are used only post hoc for the contamination/purification diagnostics. The panels characterize empirical co-variation between selection quality and downstream detection rather than a causal relationship.

C.6 Oracle Clean-Reference Diagnostic

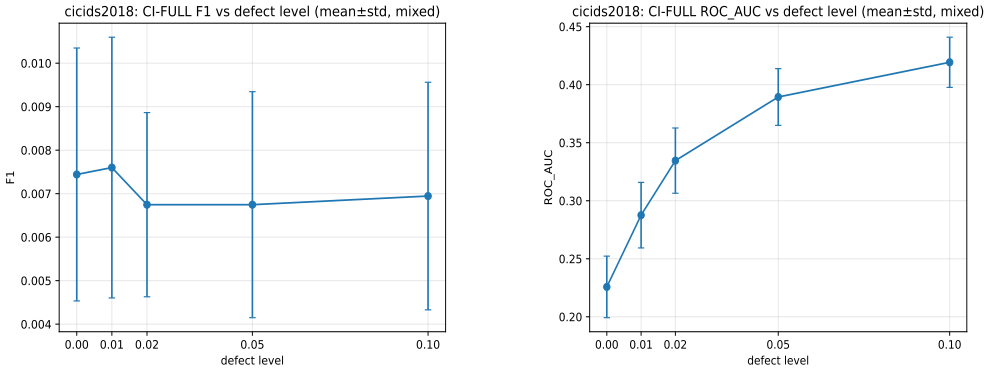
Finally, we compare CI-RS with a label-informed clean-reference construction. The oracle reference set is sampled uniformly from ground-truth normal training observations while matching the CI reference size. Because this construction uses ground-truth labels, it is unavailable to the unsupervised method and is included only as a diagnostic comparison.

Table 3 reports this oracle reference configuration together with the other theory-aligned baselines at the common focus condition.

The oracle comparison helps distinguish reference contamination from other sources of detector error. If CI-FULL approaches the oracle configuration in a particular setting, the remaining empirical gap to a label-clean reference construction is correspondingly small under that evaluated configuration. A larger gap indicates that additional improvement may be possible through reference selection or through other interacting components of the detector.

The oracle should not, however, be interpreted as a guaranteed performance upper bound. A reference bank containing only ground-truth normal observations can still provide incomplete normal-support coverage, and its downstream performance also depends on the learned representation, reference sampling, and local kNN geometry. Its role here is therefore diagnostic rather than a theoretical upper bound on achievable anomaly-detection performance.

Summary. Taken together, these diagnostics support a conditional interpretation of CI-RS. The preliminary score distribution determines the information available for posterior selection; successful preferential retention can reduce effective contamination; and greater purification can be associated with improved downstream detection. However, neither fitted-component separation nor contamination reduction alone guarantees improved performance. The failure analysis illustrates how the benefit can contract when the selection structure becomes less informative, while the oracle comparison shows the remaining gap to a label-informed clean-reference construction. These analyses therefore



(a) Fixed-mass F1. (b) AUROC.

Figure 8: Detector behavior on CICIDS2018 under increasing mixed-defect severity. F1 is evaluated under fixed alert-mass calibration, whereas AUROC measures global ranking quality. Error bars show standard deviation across the predefined seeds.

complement the formal condition $q_N > q_A$ without treating separability, purification, or oracle purity as sufficient guarantees of detector performance.

D Extended Robustness and Stress Tests

This section provides extended robustness and stress analyses across the evaluated datasets. We examine complementary sources of experimental stress: (i) increasing mixed-defect severity, (ii) increasing contamination of the unlabeled training data under a fixed nonzero evaluation defect, (iii) increasing training contamination under clean evaluation, and (iv) the range of detector performance across the evaluated contamination conditions. We conclude with the complete numerical contamination results.

These experiments characterize sensitivity under the controlled defect and contamination protocols defined in this study. They should not be interpreted as guarantees of robustness to arbitrary real-world corruption, distribution shift, or adversarial poisoning.

D.1 Robustness to Defect Severity

This section reports the complete mixed-defect severity analysis across the five evaluated datasets. The experiment varies structured-defect severity while retaining the corresponding training-contamination and evaluation protocol defined in Section 6.

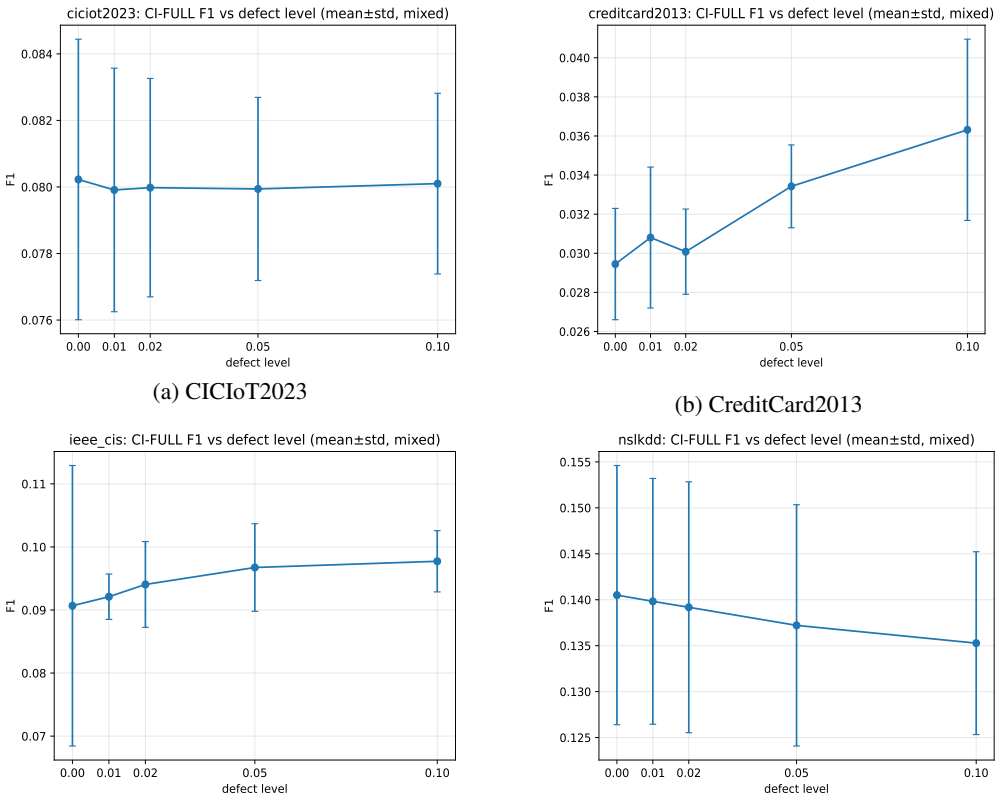
We first show CICIDS2018 as a representative paired comparison of fixed-mass F1 and AUROC. The subsequent figures report the corresponding F1 and AUROC trajectories for CICIoT2023, CreditCard2013, IEEE-CIS, and NSL-KDD.

Figure 8 shows that the detector trajectories change with increasing mixed-defect severity rather than remaining invariant. The F1 and AUROC trajectories need not move proportionally because they measure different aspects of detector behavior: AUROC evaluates global score ordering, whereas fixed-mass F1 additionally depends on performance at the selected alert budget.

The experiment should be interpreted as a controlled sensitivity analysis under the specified missingness-and-noise process. It does not establish robustness to arbitrary corruption, feature-dependent missingness, temporally correlated drift, or adversarial perturbations.

Figures 8, 9, and 10 report the mixed-defect severity analysis across all five datasets. The trajectories are dataset and metric dependent rather than uniformly monotonic. Increasing defect severity can affect the learned representation, CI-RS selection, reference composition, and final anomaly-score distribution, with the magnitude of these changes varying across datasets.

F1 and AUROC provide complementary evidence. AUROC evaluates global score ranking, whereas fixed-mass F1 additionally measures performance at the selected alert budget. The two metrics therefore need not change in parallel. Fixed alert-mass calibration standardizes operating-point



(a) CICIoT2023

(b) CreditCard2013

(c) IEEE-CIS

(d) NSL-KDD

Figure 9: Effect of mixed-defect severity on CI-DCUAD F1 for the four datasets shown. Curves report mean performance across the predefined seeds, with error bars denoting the corresponding standard deviation. F1 is evaluated at the fixed alert-mass operating point.

construction across conditions, but it does not protect the representation, CI-RS selection, reference bank, or score ranking from corruption.

These experiments characterize sensitivity to the controlled mixed-defect process and severity range evaluated in this study. They do not establish invariance to corruption, uniform superiority over the matched detector, or robustness to arbitrary missingness, noise, distribution shift, or adversarial perturbations.

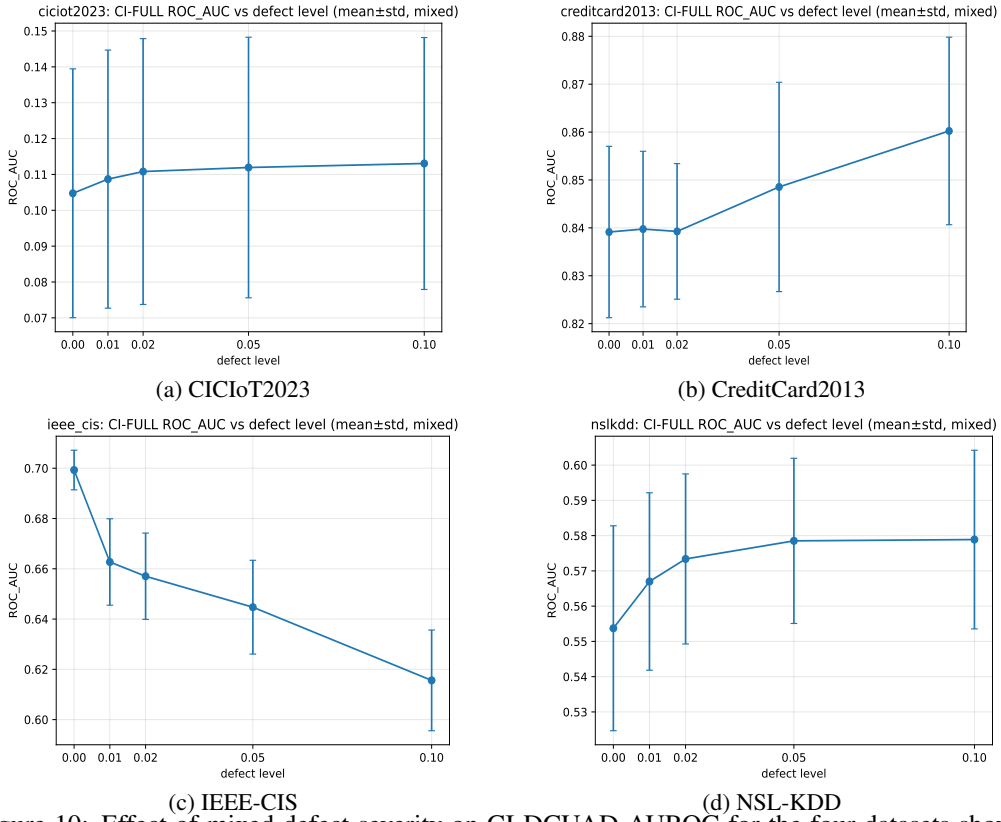
D.2 Training-Contamination Stress Under Nonzero Evaluation Defects

This section reports the complete CI-C1 training-contamination stress experiment across the five evaluated datasets. The anomaly fraction π_{train} in the unlabeled training population is varied while evaluation is held fixed under mixed defects at level 0.05.

This experiment changes the composition of the unlabeled training data while maintaining a common nonzero evaluation-defect condition. It therefore examines training-contamination sensitivity rather than isolating evaluation-time corruption. We first report the representative CICIDS2018 result, followed by the corresponding results for the remaining four datasets.

Figure 11 shows that CI-DCUAD and DCAD-BASE follow different trajectories as the anomaly fraction in the unlabeled training population increases. Any advantage observed at lower or intermediate contamination does not remain constant over the full range. At more severe contamination levels, the difference can contract as both detectors deteriorate.

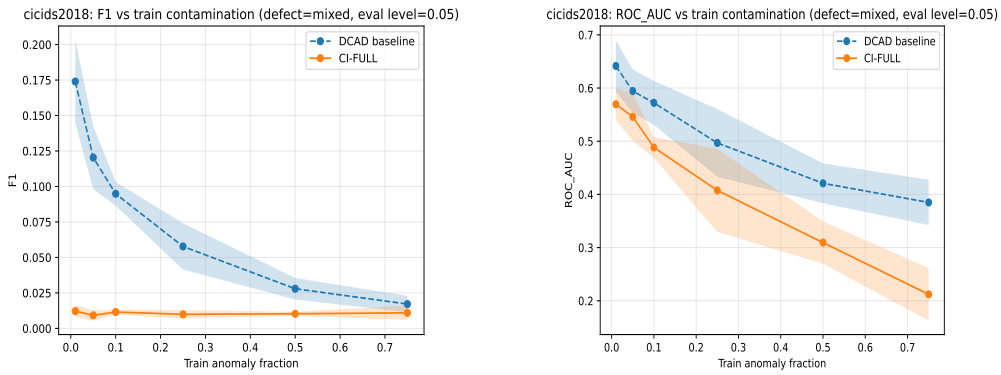
This behavior is consistent with the conditional purification mechanism. Increasing nominal contamination changes the population from which CI-RS selects candidate reference observations, but nominal contamination alone does not determine whether purification succeeds. The relevant re-



(a) CICIoT2023 (b) CreditCard2013

(c) IEEE-CIS (d) NSL-KDD

Figure 10: Effect of mixed-defect severity on CI-DCUAD AUROC for the four datasets shown. Curves report mean performance across the predefined seeds, with error bars denoting the corresponding standard deviation. AUROC evaluates global score ranking and is independent of the binary alert-mass operating threshold.

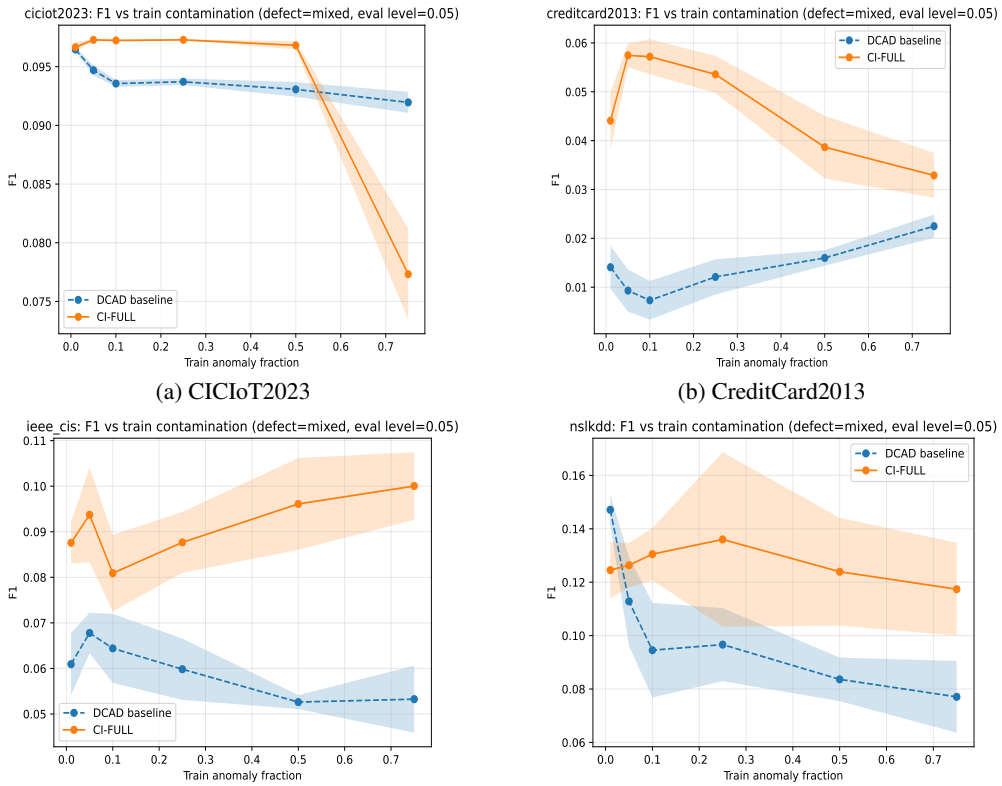


(a) Fixed-mass F1. (b) AUROC.

Figure 11: Training-contamination stress on CICIDS2018 under mixed evaluation defects at level 0.05. F1 uses fixed alert-mass calibration, whereas AUROC measures global ranking quality. Curves report performance as the training anomaly fraction π_{train} increases.

quirement is whether the available score structure continues to support preferential normal retention, $q_N > q_A$.

Figures 11, 12, and 13 report the complete CI-C1 analysis across the five datasets. The trajectories are dataset dependent: increasing π_{train} does not produce a uniform response across datasets, metrics, or contamination levels.



(a) CICIoT2023 (b) CreditCard2013

(c) IEEE-CIS (d) NSL-KDD

Figure 12: CI-C1 training-contamination stress test for the four datasets shown at evaluation defect level 0.05. F1 is reported as a function of training contamination π_{train} under the fixed alert-mass operating protocol.

F1 and AUROC provide complementary information. AUROC measures global score-ranking quality, whereas F1 evaluates performance at the fixed alert-mass operating point. Consequently, changes in the two metrics need not coincide.

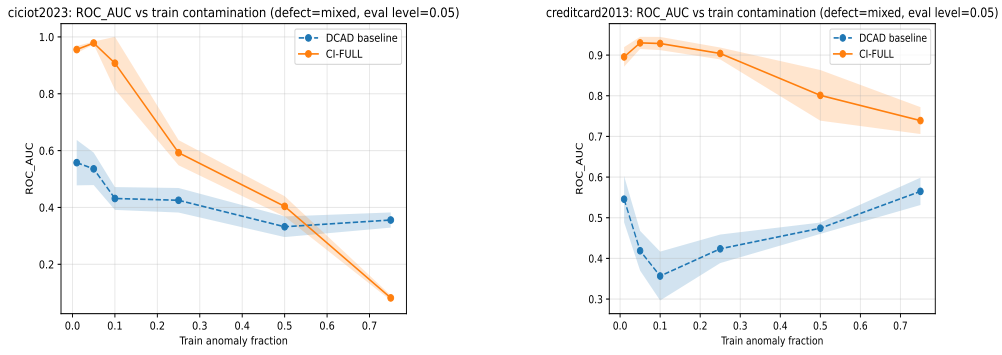
The benefit of contamination-aware reference selection is likewise condition dependent. Where CI-RS continues to construct a useful lower-contamination reference set, CI-DCUAD can retain an advantage over the matched baseline. As training contamination becomes more severe, that advantage can contract as the preliminary score structure becomes less informative for preferential reference selection.

This interpretation is consistent with, but does not by itself prove, the selection mechanism characterized in Sections B.2 and B.3. Nominal training contamination π_{train} alone does not determine whether purification succeeds; the relevant theoretical condition is whether CI-RS preferentially retains normal observations, $q_N > q_A$.

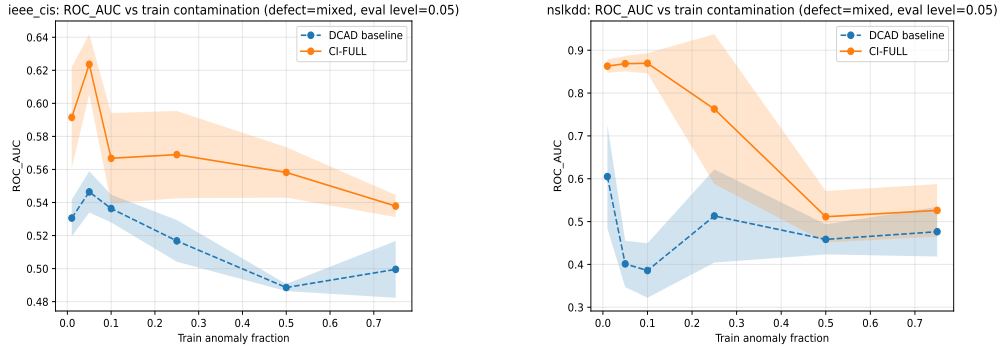
The largest values of π_{train} represent severe controlled contamination stress. They should not be interpreted as an adversarial-contamination guarantee or as evidence of robustness to arbitrary poisoning strategies.

D.3 Training-Contamination Stress Under Clean Evaluation

This analysis reports the CI-C2 training-contamination experiment across all evaluated datasets. Training contamination π_{train} is varied while evaluation is performed at defect level 0.00. In contrast to CI-C1, this setting removes evaluation-time defect corruption and therefore provides a complementary view of sensitivity to the composition of the unlabeled training data. We first report the representative CICIDS2018 result, followed by the corresponding results on the remaining datasets.

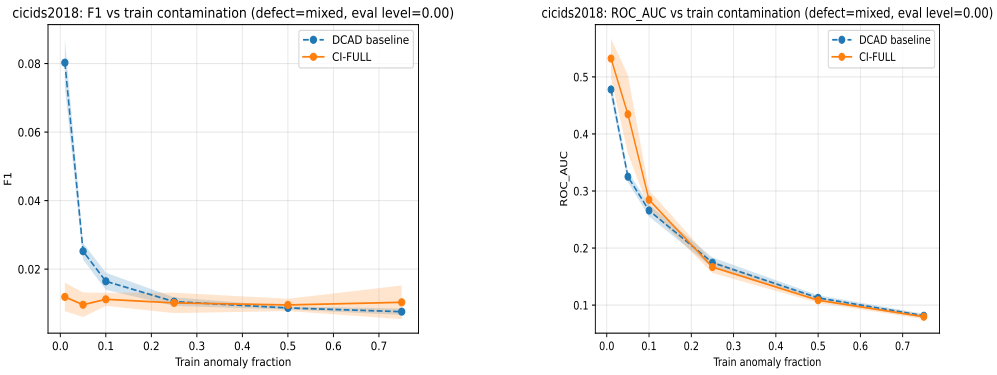


(a) CICIoT2023 (b) CreditCard2013



(c) IEEE-CIS (d) NSL-KDD

Figure 13: CI-C1 training-contamination stress test for the four datasets shown at evaluation defect level 0.05. AUROC is reported as a function of training contamination π_{train} and evaluates the global ranking induced by the anomaly scores.



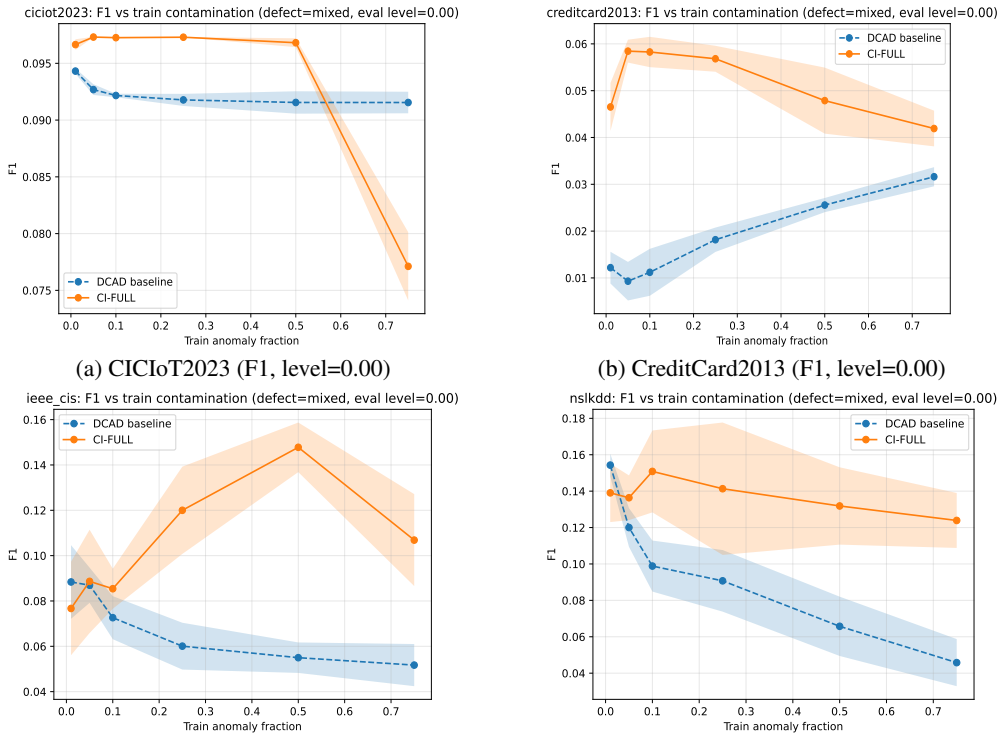
(a) F1 vs. π_{train}. (b) AUROC vs. π_{train}.

Figure 14: Clean-evaluation training-contamination stress on CICIDS2018. F1 under fixed alert-mass calibration and AUROC are reported as the training anomaly fraction π_{train} increases, with the evaluation defect level fixed at 0.00.

Figure 14 shows the representative CICIDS2018 behavior. Figures 15 and 16 extend the same clean-evaluation analysis to CICIoT2023, CreditCard2013, IEEE-CIS, and NSL-KDD.

Figures 15 and 16 show that detector behavior can change with training contamination even when evaluation-time defect corruption is absent. Because the evaluation defect level is fixed at zero throughout this experiment, changes along the horizontal axis cannot be attributed to increasing evaluation-defect severity.

Instead, this experiment characterizes sensitivity to changes in the composition of the unlabeled training data under the specified contamination protocol. Increasing π_{train} can affect the initial representation and preliminary score structure available to CI-RS, the composition of the selected set,



(a) CICIoT2023 (F1, level=0.00)

(b) CreditCard2013 (F1, level=0.00)

(c) IEEE-CIS (F1, level=0.00)

(d) NSL-KDD (F1, level=0.00)

Figure 15: CI-C2 training-contamination stress test under clean evaluation (defect level $= 0.00$) for the four datasets shown. F1 is reported as a function of training contamination π_{train} under the fixed alert-mass operating protocol. Curves show the mean across the predefined seeds, with shaded regions denoting standard deviation.

subsequent representation learning, and the final reference bank. The experiment does not isolate which of these pathways causes an individual performance change.

As in the nonzero-defect stress experiment, F1 and AUROC provide complementary views. AUROC evaluates global score ranking, whereas F1 additionally depends on performance at the fixed alert-mass operating point. Their trajectories therefore need not change in parallel.

Across the datasets shown, the response to increasing training contamination is condition dependent. In some ranges, CI-DCUAD can retain an advantage over the matched baseline, whereas at stronger contamination the advantage can contract and the methods can approach similar performance. Such convergence indicates a reduction in the observed benefit of CI-RS under those conditions; it does not directly establish a particular representation-level failure mechanism.

Together, the two training-contamination stress experiments distinguish:

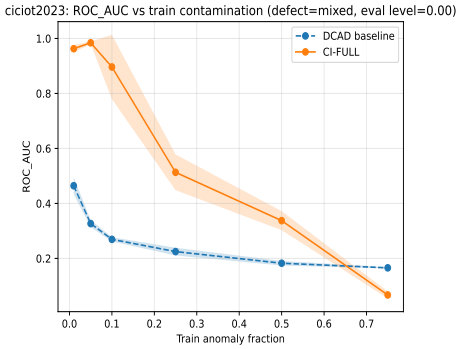
CI-C1: vary training contamination under a fixed nonzero evaluation defect,

CI-C2: vary training contamination under clean evaluation.

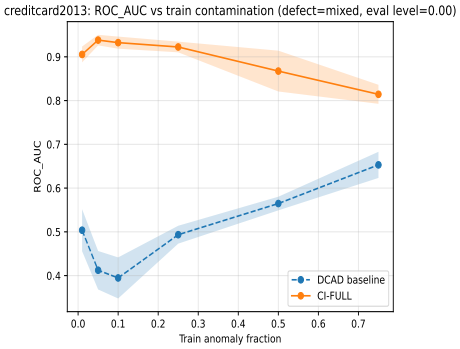
The clean-evaluation results therefore show that sensitivity to training contamination is not solely attributable to increasing evaluation-time defect severity. They do not establish robustness to arbitrary contamination or identify a single causal mechanism for the observed detector-level changes.

D.4 Detailed Contamination Results

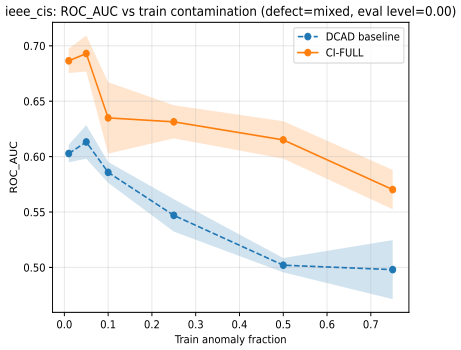
This subsection provides the detailed numerical results underlying the training-contamination analyses. The tables progress from dataset-level aggregation to per-contamination summaries and finally to the complete reported grid across training-contamination and evaluation settings.



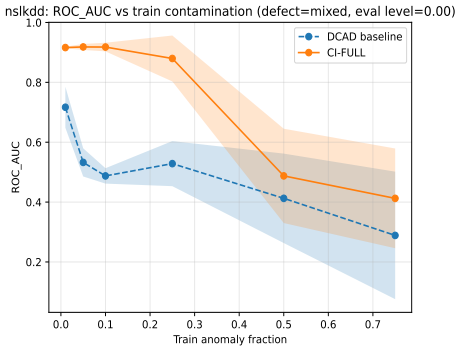
(a) CICIoT2023 (ROC-AUC, level=0.00)



(b) CreditCard2013 (ROC-AUC, level=0.00)



(c) IEEE-CIS (ROC-AUC, level=0.00)



(d) NSL-KDD (ROC-AUC, level=0.00)

Figure 16: CI-C2 training-contamination stress test under clean evaluation (defect level $= 0.00$) for the four datasets shown. AUROC is reported as a function of training contamination π_{train} and evaluates the global ranking induced by the anomaly scores. Curves show the mean across the predefined seeds, with shaded regions denoting standard deviation.

Table 7: Dataset-level performance at the fixed evaluation setting $\lambda = 0.05$, aggregated across the evaluated training-contamination levels and random seeds. Panels (a) and (b) report the F1 and AUROC results, respectively, for `dcad_base` and `ci_full`.

(a) F1 performance

Dataset	λ	K	$n_{\min}$	Base μ	CI μ	ΔF1	Rel. F1	Base σ	CI σ
cicids2018	0.0500	6	5	0.0820	0.0107	-0.0713	-0.7586	0.0149	0.0032
ciciot2023	0.0500	6	5	0.0939	0.0938	-0.0001	-0.0020	0.0005	0.0008
creditcard2013	0.0500	6	5	0.0136	0.0473	0.0338	3.2372	0.0034	0.0044
ieee_cis	0.0500	6	5	0.0598	0.0910	0.0312	0.5411	0.0057	0.0079
nslkdd	0.0500	6	4	0.1020	0.1265	0.0245	0.2933	0.0125	0.0165

(b) AUROC performance

Dataset	λ	K	$n_{\min}$	Base μ	CI μ	ΔAUROC	Rel. AUROC	Base σ	CI σ
cicids2018	0.0500	6	5	0.5184	0.4221	-0.0963	-0.2057	0.0457	0.0436
ciciot2023	0.0500	6	5	0.4397	0.6533	0.2136	0.4137	0.0472	0.0332
creditcard2013	0.0500	6	5	0.4640	0.8663	0.4023	0.9327	0.0414	0.0275
ieee_cis	0.0500	6	5	0.5197	0.5745	0.0548	0.1055	0.0107	0.0207
nslkdd	0.0500	6	4	0.4733	0.7335	0.2602	0.5920	0.0734	0.0589

Dataset-level summary. Table 7 summarizes performance at the fixed evaluation setting $\lambda = 0.05$. For each dataset, the reported values aggregate over the evaluated training-contamination levels and random seeds. Because this aggregation combines multiple contamination regimes, it provides a compact dataset-level summary rather than a characterization of behavior at any single value of π_{train}.

Per-contamination performance trends. Tables 8 and 9 disaggregate the preceding summary by training-contamination level π_{train} while retaining the fixed evaluation setting $\lambda = 0.05$. This view distinguishes aggregate behavior from condition-specific changes as the composition of the unlabeled training data varies.

Complete numerical grid. Complete seed-aggregated results across datasets, training-contamination levels, evaluation settings, and matched method variants are retained in the supplementary experimental artifacts. We omit the complete numerical grid from the manuscript because the condition-specific trends relevant to the training-contamination analysis are already reported in Tables 8 and 9.

Summary. The extended stress analyses show that CI-DCUAD's behavior is conditional on both defect severity and the composition of the unlabeled training data. Increasing mixed-defect severity can alter both ranking and fixed-operating- point performance, while increasing training contamination can reduce the advantage of CI-RS in regimes where preferential likely-normal selection becomes less effective. The corresponding clean-evaluation experiment shows that sensitivity to training contamination can persist even when evaluation-time defect corruption is removed.

These results are consistent with the formal CI-RS condition $q_N > q_A$ for contamination reduction, but they do not establish a single causal explanation for detector-level performance changes. Representation quality, preservation of normal support, selected-reference composition, and local kNN geometry remain interacting factors. The complete numerical tables above provide the underlying results across the evaluated contamination and evaluation settings.

E Ablations and Sensitivity Analysis

This section provides complementary analyses of the components, hyperparameters, operating-point choices, and stochastic stability of CI-DCUAD. We examine (i) the structural contribution of the CI-controlled warmup and CI-selected reference pathways, (ii) sensitivity to the CI-RS posterior threshold π, (iii) sensitivity to the alert-mass parameter β, (iv) sensitivity to the reference-bank cap, and (v) reproducibility of CI-selected observations across random seeds. We conclude with an explicit seed-coverage check for the reported experimental conditions.

These analyses address different sources of sensitivity and should be interpreted separately. In particular, the removal ablation changes components of the detector, the posterior-threshold sweep changes CI-RS selectivity, the alert-budget sweep changes only the post-hoc operating point, the reference-cap sweep changes the number of embeddings available to the final kNN scorer, and the Jaccard analysis measures reproducibility of selected observations.

E.1 Full Removal Ablation

This section reports the structural removal ablation used to distinguish the two pathways through which CI-RS enters the detector. We evaluate four matched configurations:

- `dcad_base`: neither CI-controlled warmup nor CI-selected reference construction;
- `ci_warm_only`: CI-based selection is used only during controlled warmup;
- `ci_ref_only`: CI-based selection is used only to construct the final reference bank; and
- `ci_full`: CI-based selection is used during both warmup and final reference construction.

All four configurations use the same encoder architecture, main-training schedule, optimization settings, defect curriculum, kNN scoring backend, reference-size cap, and fixed alert-mass calibration. Subsequent main-stage representation learning uses the complete unlabeled training set.

The variants form a structural ablation rather than an additive causal decomposition. CI-controlled warmup can change the representation subsequently used for reference scoring, so the warmup and reference-construction pathways can interact.

Table 8: F1 performance across the evaluated training-contamination levels π_{train} at the fixed evaluation setting $\lambda = 0.05$. Results include the mean (μ), standard deviation (σ), absolute difference, and relative difference between `dcad_base` and `ci_full`.

Dataset	π_{train}	λ	Base μ	Base σ	CI μ	CI σ	ΔF1	Rel. F1	n
cicids2018	0.0100	0.0500	0.1739	0.0292	0.0122	0.0042	-0.1617	-0.9298	5
cicids2018	0.0500	0.0500	0.1204	0.0223	0.0092	0.0036	-0.1112	-0.9237	5
cicids2018	0.1000	0.0500	0.0949	0.0084	0.0115	0.0019	-0.0833	-0.8785	5
cicids2018	0.2500	0.0500	0.0577	0.0163	0.0099	0.0029	-0.0478	-0.8282	5
cicids2018	0.5000	0.0500	0.0279	0.0076	0.0103	0.0014	-0.0176	-0.6314	5
cicids2018	0.7500	0.0500	0.0171	0.0058	0.0110	0.0051	-0.0062	-0.3600	5
ciciot2023	0.0100	0.0500	0.0965	0.0002	0.0967	0.0004	0.0002	0.0022	5
ciciot2023	0.0500	0.0500	0.0947	0.0004	0.0973	0.0000	0.0026	0.0274	5
ciciot2023	0.1000	0.0500	0.0936	0.0003	0.0972	0.0001	0.0037	0.0392	5
ciciot2023	0.2500	0.0500	0.0937	0.0003	0.0973	0.0000	0.0036	0.0381	5
ciciot2023	0.5000	0.0500	0.0931	0.0006	0.0968	0.0003	0.0038	0.0403	5
ciciot2023	0.7500	0.0500	0.0920	0.0009	0.0773	0.0039	-0.0146	-0.1592	5
creditcard2013	0.0100	0.0500	0.0141	0.0044	0.0441	0.0058	0.0300	2.1282	5
creditcard2013	0.0500	0.0500	0.0093	0.0043	0.0575	0.0024	0.0481	5.1748	5
creditcard2013	0.1000	0.0500	0.0073	0.0040	0.0572	0.0036	0.0499	6.8148	5
creditcard2013	0.2500	0.0500	0.0121	0.0036	0.0536	0.0038	0.0415	3.4254	5
creditcard2013	0.5000	0.0500	0.0160	0.0016	0.0387	0.0064	0.0227	1.4181	5
creditcard2013	0.7500	0.0500	0.0225	0.0024	0.0329	0.0046	0.0104	0.4618	5
ieee_cis	0.0100	0.0500	0.0609	0.0068	0.0876	0.0045	0.0266	0.4371	5
ieee_cis	0.0500	0.0500	0.0678	0.0044	0.0937	0.0104	0.0259	0.3828	5
ieee_cis	0.1000	0.0500	0.0644	0.0076	0.0809	0.0084	0.0165	0.2562	5
ieee_cis	0.2500	0.0500	0.0598	0.0067	0.0877	0.0067	0.0278	0.4653	5
ieee_cis	0.5000	0.0500	0.0526	0.0015	0.0961	0.0101	0.0435	0.8265	5
ieee_cis	0.7500	0.0500	0.0533	0.0074	0.1000	0.0074	0.0468	0.8786	5
nslkdd	0.0100	0.0500	0.1471	0.0055	0.1245	0.0105	-0.0226	-0.1535	5
nslkdd	0.0500	0.0500	0.1128	0.0167	0.1264	0.0084	0.0135	0.1199	5
nslkdd	0.1000	0.0500	0.0945	0.0177	0.1305	0.0098	0.0360	0.3805	5
nslkdd	0.2500	0.0500	0.0966	0.0137	0.1360	0.0327	0.0394	0.4076	4
nslkdd	0.5000	0.0500	0.0836	0.0082	0.1239	0.0202	0.0403	0.4821	5
nslkdd	0.7500	0.0500	0.0771	0.0134	0.1174	0.0174	0.0403	0.5231	5

Table 9: AUROC performance across the evaluated training-contamination levels π_{train} at the fixed evaluation setting $\lambda = 0.05$. Results include the mean (μ), standard deviation (σ), absolute difference, and relative difference between `dcad_base` and `ci_full`.

Dataset	π_{train}	λ	Base μ	Base σ	CI μ	CI σ	ΔAUROC	Rel. AUROC	n
cicids2018	0.0100	0.0500	0.6416	0.0487	0.5696	0.0298	-0.0720	-0.1123	5
cicids2018	0.0500	0.0500	0.5943	0.0407	0.5459	0.0449	-0.0484	-0.0815	5
cicids2018	0.1000	0.0500	0.5721	0.0413	0.4883	0.0195	-0.0839	-0.1466	5
cicids2018	0.2500	0.0500	0.4968	0.0636	0.4076	0.0781	-0.0892	-0.1796	5
cicids2018	0.5000	0.0500	0.4209	0.0374	0.3094	0.0398	-0.1115	-0.2650	5
cicids2018	0.7500	0.0500	0.3849	0.0426	0.2121	0.0492	-0.1728	-0.4490	5
ciciot2023	0.0100	0.0500	0.5576	0.0799	0.9555	0.0124	0.3979	0.7136	5
ciciot2023	0.0500	0.0500	0.5358	0.0571	0.9785	0.0052	0.4427	0.8262	5
ciciot2023	0.1000	0.0500	0.4315	0.0401	0.9076	0.0928	0.4760	1.1031	5
ciciot2023	0.2500	0.0500	0.4253	0.0432	0.5926	0.0443	0.1673	0.3935	5
ciciot2023	0.5000	0.0500	0.3321	0.0363	0.4038	0.0358	0.0718	0.2161	5
ciciot2023	0.7500	0.0500	0.3557	0.0267	0.0818	0.0091	-0.2739	-0.7701	5
creditcard2013	0.0100	0.0500	0.5457	0.0568	0.8955	0.0240	0.3499	0.6412	5
creditcard2013	0.0500	0.0500	0.4188	0.0488	0.9300	0.0149	0.5112	1.2205	5
creditcard2013	0.1000	0.0500	0.3566	0.0601	0.9285	0.0162	0.5718	1.6034	5
creditcard2013	0.2500	0.0500	0.4236	0.0351	0.9040	0.0146	0.4804	1.1340	5
creditcard2013	0.5000	0.0500	0.4742	0.0140	0.8010	0.0623	0.3268	0.6891	5
creditcard2013	0.7500	0.0500	0.5649	0.0336	0.7389	0.0331	0.1740	0.3081	5
ieee_cis	0.0100	0.0500	0.5306	0.0112	0.5915	0.0306	0.0609	0.1148	5
ieee_cis	0.0500	0.0500	0.5464	0.0125	0.6236	0.0182	0.0772	0.1413	5
ieee_cis	0.1000	0.0500	0.5363	0.0084	0.5668	0.0273	0.0304	0.0567	5
ieee_cis	0.2500	0.0500	0.5168	0.0126	0.5690	0.0264	0.0522	0.1010	5
ieee_cis	0.5000	0.0500	0.4885	0.0021	0.5582	0.0153	0.0697	0.1426	5
ieee_cis	0.7500	0.0500	0.4995	0.0172	0.5378	0.0067	0.0383	0.0767	5
nslkdd	0.0100	0.0500	0.6050	0.1208	0.8628	0.0158	0.2578	0.4262	5
nslkdd	0.0500	0.0500	0.4012	0.0542	0.8687	0.0178	0.4674	1.1650	5
nslkdd	0.1000	0.0500	0.3857	0.0637	0.8695	0.0237	0.4838	1.2544	5
nslkdd	0.2500	0.0500	0.5131	0.1088	0.7627	0.1748	0.2496	0.4865	4
nslkdd	0.5000	0.0500	0.4584	0.0351	0.5113	0.0601	0.0529	0.1154	5
nslkdd	0.7500	0.0500	0.4763	0.0579	0.5261	0.0615	0.0497	0.1044	5

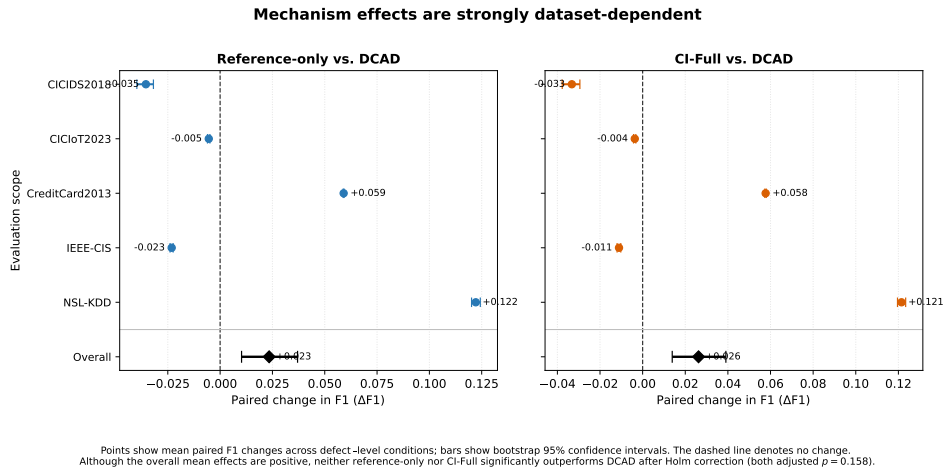


Figure 17: Dataset-specific paired F1 effects relative to DCAD. Points are means over 15 defect–severity conditions per dataset; error bars are bootstrap 95% confidence intervals. Overall estimates use all 75 condition-level units after averaging across five seeds.

Table 10: Structural ablation under mixed defects at $\lambda = 0.05$. Entries are extracted directly from the complete ablation grid and report fixed-mass F1 as mean $\pm$ standard deviation over five predefined seeds.

Dataset	DCAD-Base	CI-Warm-Only	CI-Ref-Only	CI-Full
CICIDS2018	0.058 ± 0.002	0.061 ± 0.006	0.010 ± 0.002	0.012 ± 0.001
CICIoT2023	0.086 ± 0.001	0.088 ± 0.002	0.076 ± 0.002	0.078 ± 0.005
CreditCard2013	0.002 ± 0.001	0.001 ± 0.001	0.061 ± 0.001	0.060 ± 0.002
IEEE-CIS	0.073 ± 0.010	0.056 ± 0.004	0.049 ± 0.015	0.065 ± 0.026
NSL-KDD	0.050 ± 0.002	0.052 ± 0.002	0.168 ± 0.004	0.167 ± 0.009

Focus-condition extraction. Table 10 extracts the mixed-defect, $\lambda = 0.05$ rows directly from the complete ablation grid in Table 11. Consequently, the compact table and the complete grid use the same experimental block and have identical values for the corresponding conditions.

Table 10 shows that the reference-construction pathway accounts for most of the favorable focus-condition change on CreditCard2013 and NSL-KDD. In both datasets, CI-Ref-Only substantially exceeds the matched DCAD-Base, while CI-Warm-Only remains close to the baseline.

The same pattern is not observed on CICIDS2018, CICIoT2023, or IEEE-CIS. On these datasets, reference-only and CI-Full are below the matched base detector under the focus condition. Reference construction is therefore an important pathway in the favorable datasets, but it is not a uniformly beneficial component.

CI-Full should not be interpreted as the sum of two independently beneficial interventions. For example, CI-Full is slightly higher than CI-Ref-Only on CICIDS2018, CICIoT2023, and IEEE-CIS, but slightly lower on CreditCard2013 and NSL-KDD. This variation is consistent with interaction between representation warmup and the final reference geometry.

Complete ablation grid. Table 11 reports all evaluated datasets, defect types, and defect levels. Entries are fixed-mass F1 values reported as mean $\pm$ standard deviation over the predefined random seeds.

Table 11: Full removal ablation across datasets, defect types, and defect levels. Entries report F1 as mean ± standard deviation across the predefined random seeds under the fixed alert-mass operating protocol.

Dataset	Defect	Level	Dcad_Base	Ci_Warm_Only	Ci_Ref_Only	Ci_Full
cicids2018	missingness	0.000000	0.048 ± 0.002	0.049 ± 0.003	0.009 ± 0.002	0.012 ± 0.002
cicids2018	missingness	0.010000	0.050 ± 0.002	0.051 ± 0.003	0.010 ± 0.001	0.012 ± 0.002
cicids2018	missingness	0.020000	0.052 ± 0.003	0.054 ± 0.003	0.010 ± 0.002	0.012 ± 0.002
cicids2018	missingness	0.050000	0.054 ± 0.003	0.059 ± 0.006	0.010 ± 0.001	0.012 ± 0.001
cicids2018	missingness	0.100000	0.058 ± 0.002	0.063 ± 0.005	0.011 ± 0.002	0.012 ± 0.001
cicids2018	mixed	0.000000	0.048 ± 0.002	0.049 ± 0.003	0.009 ± 0.002	0.012 ± 0.002
cicids2018	mixed	0.010000	0.051 ± 0.002	0.052 ± 0.004	0.010 ± 0.002	0.012 ± 0.002
cicids2018	mixed	0.020000	0.052 ± 0.003	0.055 ± 0.004	0.010 ± 0.002	0.012 ± 0.002
cicids2018	mixed	0.050000	0.058 ± 0.002	0.061 ± 0.006	0.010 ± 0.002	0.012 ± 0.001
cicids2018	mixed	0.100000	0.072 ± 0.004	0.077 ± 0.013	0.011 ± 0.001	0.013 ± 0.001
cicids2018	noise	0.000000	0.048 ± 0.002	0.049 ± 0.003	0.009 ± 0.002	0.012 ± 0.002
cicids2018	noise	0.010000	0.048 ± 0.002	0.050 ± 0.003	0.009 ± 0.002	0.012 ± 0.002
cicids2018	noise	0.020000	0.048 ± 0.002	0.050 ± 0.003	0.010 ± 0.002	0.012 ± 0.002
cicids2018	noise	0.050000	0.051 ± 0.002	0.053 ± 0.003	0.010 ± 0.001	0.012 ± 0.002
cicids2018	noise	0.100000	0.056 ± 0.004	0.056 ± 0.007	0.010 ± 0.001	0.012 ± 0.001
ciciot2023	missingness	0.000000	0.087 ± 0.000	0.088 ± 0.001	0.077 ± 0.002	0.081 ± 0.004
ciciot2023	missingness	0.010000	0.087 ± 0.001	0.088 ± 0.001	0.077 ± 0.002	0.079 ± 0.004
ciciot2023	missingness	0.020000	0.086 ± 0.000	0.088 ± 0.001	0.077 ± 0.002	0.079 ± 0.004
ciciot2023	missingness	0.050000	0.086 ± 0.001	0.088 ± 0.001	0.077 ± 0.002	0.079 ± 0.004
ciciot2023	missingness	0.100000	0.085 ± 0.001	0.087 ± 0.002	0.077 ± 0.002	0.079 ± 0.004
ciciot2023	mixed	0.000000	0.087 ± 0.000	0.088 ± 0.001	0.077 ± 0.002	0.081 ± 0.004
ciciot2023	mixed	0.010000	0.087 ± 0.001	0.088 ± 0.001	0.077 ± 0.002	0.079 ± 0.004
ciciot2023	mixed	0.020000	0.087 ± 0.001	0.088 ± 0.001	0.077 ± 0.002	0.079 ± 0.004
ciciot2023	mixed	0.050000	0.086 ± 0.001	0.088 ± 0.002	0.076 ± 0.002	0.078 ± 0.005
ciciot2023	mixed	0.100000	0.087 ± 0.001	0.089 ± 0.002	0.076 ± 0.002	0.077 ± 0.004
ciciot2023	noise	0.000000	0.087 ± 0.000	0.088 ± 0.001	0.077 ± 0.002	0.081 ± 0.004
ciciot2023	noise	0.010000	0.087 ± 0.000	0.088 ± 0.001	0.077 ± 0.002	0.080 ± 0.004
ciciot2023	noise	0.020000	0.087 ± 0.000	0.088 ± 0.001	0.077 ± 0.002	0.080 ± 0.004
ciciot2023	noise	0.050000	0.087 ± 0.000	0.088 ± 0.001	0.077 ± 0.002	0.079 ± 0.004
ciciot2023	noise	0.100000	0.088 ± 0.001	0.090 ± 0.001	0.076 ± 0.003	0.078 ± 0.004
creditcard2013	missingness	0.000000	0.002 ± 0.001	0.001 ± 0.000	0.062 ± 0.001	0.061 ± 0.001
creditcard2013	missingness	0.010000	0.002 ± 0.001	0.001 ± 0.000	0.061 ± 0.001	0.061 ± 0.001
creditcard2013	missingness	0.020000	0.002 ± 0.001	0.001 ± 0.000	0.062 ± 0.001	0.060 ± 0.002
creditcard2013	missingness	0.050000	0.002 ± 0.001	0.001 ± 0.000	0.061 ± 0.001	0.060 ± 0.002
creditcard2013	missingness	0.100000	0.002 ± 0.001	0.002 ± 0.001	0.061 ± 0.001	0.059 ± 0.003
creditcard2013	mixed	0.000000	0.002 ± 0.001	0.001 ± 0.000	0.062 ± 0.001	0.061 ± 0.001
creditcard2013	mixed	0.010000	0.002 ± 0.001	0.001 ± 0.000	0.061 ± 0.001	0.060 ± 0.001
creditcard2013	mixed	0.020000	0.002 ± 0.001	0.001 ± 0.001	0.062 ± 0.001	0.060 ± 0.002
creditcard2013	mixed	0.050000	0.002 ± 0.001	0.001 ± 0.001	0.061 ± 0.001	0.060 ± 0.002
creditcard2013	mixed	0.100000	0.002 ± 0.001	0.002 ± 0.001	0.061 ± 0.001	0.059 ± 0.003
creditcard2013	noise	0.000000	0.002 ± 0.001	0.001 ± 0.000	0.062 ± 0.001	0.061 ± 0.001
creditcard2013	noise	0.010000	0.002 ± 0.001	0.001 ± 0.000	0.062 ± 0.001	0.061 ± 0.001
creditcard2013	noise	0.020000	0.002 ± 0.001	0.001 ± 0.000	0.062 ± 0.001	0.061 ± 0.001
creditcard2013	noise	0.050000	0.002 ± 0.001	0.001 ± 0.001	0.062 ± 0.001	0.061 ± 0.002
creditcard2013	noise	0.100000	0.002 ± 0.001	0.001 ± 0.000	0.062 ± 0.001	0.060 ± 0.002
ieee_cis	missingness	0.000000	0.071 ± 0.009	0.054 ± 0.004	0.049 ± 0.015	0.065 ± 0.025
ieee_cis	missingness	0.010000	0.071 ± 0.009	0.055 ± 0.004	0.049 ± 0.014	0.065 ± 0.026
ieee_cis	missingness	0.020000	0.073 ± 0.009	0.055 ± 0.006	0.049 ± 0.015	0.065 ± 0.026
ieee_cis	missingness	0.050000	0.072 ± 0.010	0.055 ± 0.006	0.048 ± 0.014	0.065 ± 0.026
ieee_cis	missingness	0.100000	0.074 ± 0.012	0.057 ± 0.006	0.048 ± 0.014	0.066 ± 0.026
ieee_cis	mixed	0.000000	0.071 ± 0.009	0.054 ± 0.004	0.049 ± 0.015	0.065 ± 0.025
ieee_cis	mixed	0.010000	0.071 ± 0.009	0.055 ± 0.004	0.049 ± 0.015	0.064 ± 0.025
ieee_cis	mixed	0.020000	0.072 ± 0.010	0.054 ± 0.004	0.049 ± 0.014	0.065 ± 0.025
ieee_cis	mixed	0.050000	0.073 ± 0.010	0.056 ± 0.004	0.049 ± 0.015	0.065 ± 0.026
ieee_cis	mixed	0.100000	0.074 ± 0.011	0.055 ± 0.003	0.049 ± 0.013	0.067 ± 0.028
ieee_cis	noise	0.000000	0.071 ± 0.009	0.054 ± 0.004	0.049 ± 0.015	0.065 ± 0.025
ieee_cis	noise	0.010000	0.071 ± 0.009	0.054 ± 0.004	0.049 ± 0.015	0.065 ± 0.025
ieee_cis	noise	0.020000	0.071 ± 0.009	0.054 ± 0.004	0.049 ± 0.014	0.065 ± 0.025
ieee_cis	noise	0.050000	0.071 ± 0.010	0.054 ± 0.004	0.049 ± 0.015	0.065 ± 0.025
ieee_cis	noise	0.100000	0.070 ± 0.009	0.054 ± 0.004	0.049 ± 0.015	0.066 ± 0.026
nslkdd	missingness	0.000000	0.044 ± 0.002	0.043 ± 0.003	0.163 ± 0.005	0.164 ± 0.009
nslkdd	missingness	0.010000	0.048 ± 0.002	0.048 ± 0.001	0.164 ± 0.005	0.165 ± 0.009
nslkdd	missingness	0.020000	0.050 ± 0.002	0.049 ± 0.002	0.164 ± 0.005	0.165 ± 0.009
nslkdd	missingness	0.050000	0.052 ± 0.003	0.053 ± 0.002	0.166 ± 0.005	0.167 ± 0.009
nslkdd	missingness	0.100000	0.056 ± 0.002	0.059 ± 0.003	0.169 ± 0.006	0.169 ± 0.008
nslkdd	mixed	0.000000	0.044 ± 0.002	0.043 ± 0.003	0.163 ± 0.005	0.164 ± 0.009
nslkdd	mixed	0.010000	0.048 ± 0.002	0.048 ± 0.003	0.164 ± 0.006	0.165 ± 0.009
nslkdd	mixed	0.020000	0.048 ± 0.001	0.049 ± 0.002	0.165 ± 0.005	0.165 ± 0.009
nslkdd	mixed	0.050000	0.050 ± 0.002	0.052 ± 0.002	0.168 ± 0.004	0.167 ± 0.009
nslkdd	mixed	0.100000	0.049 ± 0.002	0.049 ± 0.002	0.171 ± 0.004	0.168 ± 0.008
nslkdd	noise	0.000000	0.044 ± 0.002	0.043 ± 0.003	0.163 ± 0.005	0.164 ± 0.009
nslkdd	noise	0.010000	0.043 ± 0.002	0.043 ± 0.002	0.164 ± 0.005	0.164 ± 0.009
nslkdd	noise	0.020000	0.043 ± 0.001	0.043 ± 0.002	0.164 ± 0.005	0.164 ± 0.009
nslkdd	noise	0.050000	0.042 ± 0.001	0.043 ± 0.003	0.165 ± 0.004	0.164 ± 0.009
nslkdd	noise	0.100000	0.036 ± 0.003	0.038 ± 0.003	0.167 ± 0.004	0.164 ± 0.010

E.2 Posterior Threshold Sensitivity

Protocol. We evaluate sensitivity to the CI-RS posterior threshold π under the common focus condition (mixed defect, level 0.05). For each value of π, CI-RS is applied using that posterior cutoff, and the corresponding selected index set and reference bank $\mathcal{R}_{CI}(\pi)$ are reconstructed while keeping the common reference-bank cap fixed. CI-FULL is then evaluated using the same fixed alert-mass operating protocol.

Changing π modifies the selectivity of CI-RS: a larger threshold requires stronger posterior evidence for inclusion in the normal-associated component. Consequently, the sweep can affect both the number and composition of selected observations used during the CI-controlled warmup

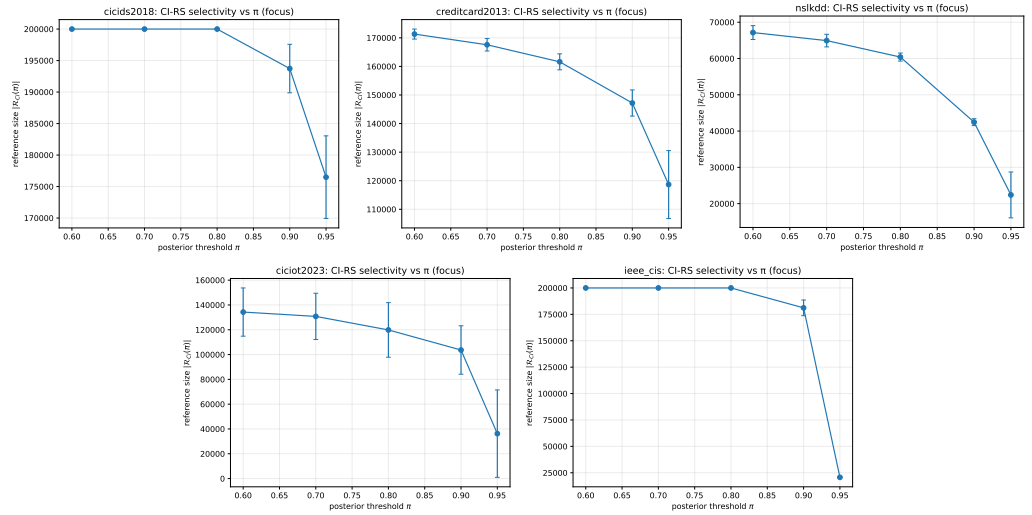


Figure 18: CI-RS selectivity under the posterior-threshold sweep at the focus condition. The plots report final reference size $|\mathcal{R}_{CI}(\pi)|$ as a function of π across the five evaluated datasets. Curves show mean $\pm$ standard deviation over the five predefined random seeds.

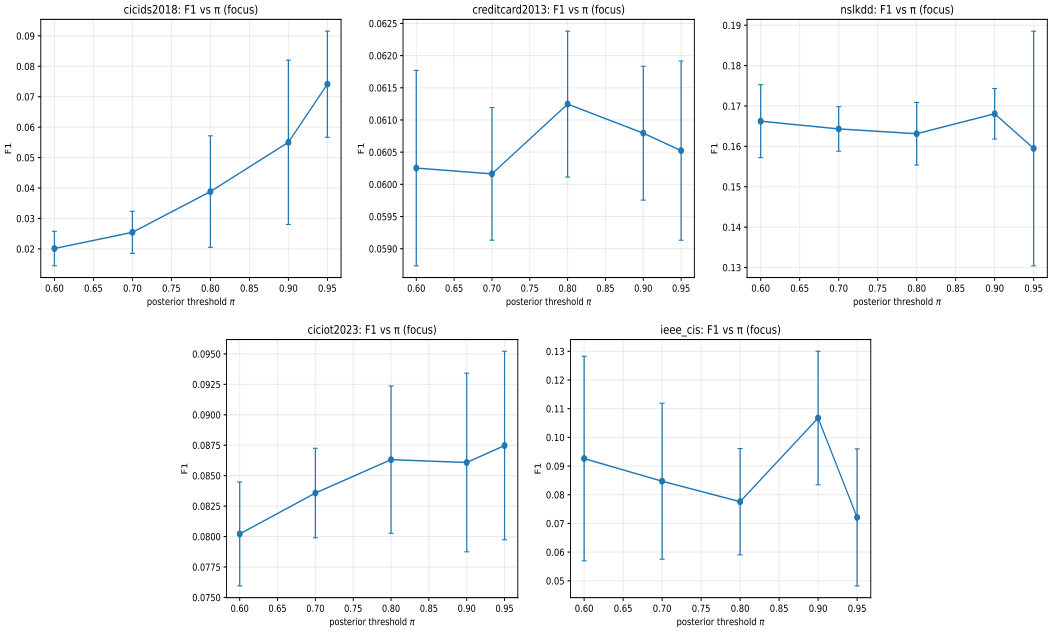


Figure 19: Sensitivity of CI-FULL F1 to the CI-RS posterior threshold π under the focus condition (mixed defect, level 0.05). F1 is evaluated under the fixed alert-mass operating protocol. Curves show mean $\pm$ standard deviation over the five predefined random seeds.

and, through the selected indices, the composition of the final reference bank. The experiment therefore measures sensitivity of the complete CI selection pathway to π, rather than reference-bank size alone.

Results. Figure 18 shows the resulting reference-bank size across the posterior-threshold sweep. As expected from the nested posterior selection rule, increasing π produces a non-increasing selected-set size before any common reference-bank cap is applied. The magnitude of this change is dataset dependent because the posterior-mass distribution differs across datasets and seeds.

Figure 19 reports the corresponding downstream F1 under the fixed alert-mass operating protocol. The results characterize how changes in CI-RS selectivity propagate to operating-point detection performance. Performance does not depend critically on a single posterior cutoff throughout the evaluated range, although the magnitude and direction of the changes are dataset dependent. Thus, the sweep supports local robustness to the posterior-threshold choice within the evaluated range rather than invariance to π.

Table 12 provides the corresponding aggregate summary across datasets and seeds. Together with the dataset-level trajectories, these results indicate that the evaluated behavior is not critically dependent on one particular value of π within the tested range. At the same time, the

Table 12: Aggregate summary across datasets and seeds for the CI-RS posterior-threshold sweep under the focus condition (mixed defects, level 0.05).

π	n	$\lvert\mathcal{R}\rvert$ Mean	$\lvert\mathcal{R}\rvert$ Std	F1 Mean	F1 Std	AUROC Mean	AUROC Std	AP Mean	AP Std
0.600	25.000	154548.440	51579.659	0.084	0.051	0.543	0.318	0.480	0.354
0.700	25.000	152659.360	52365.920	0.084	0.048	0.553	0.311	0.489	0.357
0.800	25.000	148373.680	54871.997	0.085	0.044	0.565	0.304	0.471	0.355
0.900	25.000	133644.400	57087.657	0.095	0.044	0.618	0.295	0.484	0.348
0.950	25.000	74893.080	65513.165	0.091	0.040	0.642	0.302	0.481	0.378

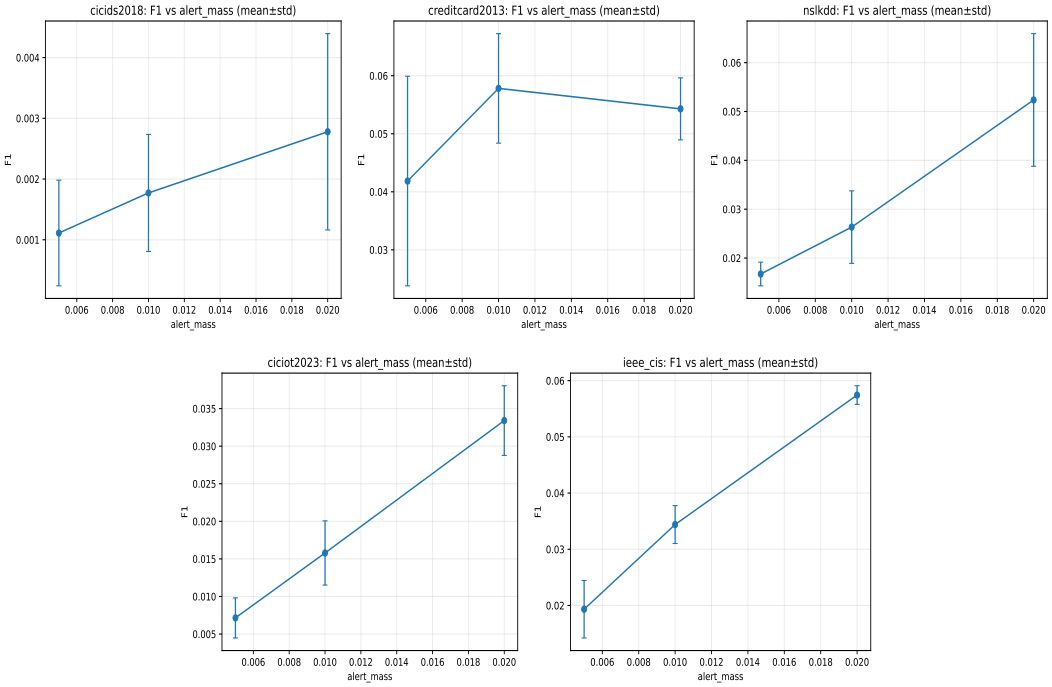


Figure 20: Sensitivity of CI-FULL F1 to the alert-mass parameter β at the focus condition (mixed defects, level 0.05). Curves show mean $\pm$ standard deviation over the five predefined random seeds.

dataset-dependent trajectories show that π controls a meaningful selection trade-off and should not be interpreted as an inconsequential hyperparameter.

The sensitivity analysis should not be interpreted as identifying an optimal posterior threshold. In particular, selecting a dataset-specific π using post-hoc test performance would introduce supervised information into an otherwise unsupervised pipeline. The purpose of this experiment is instead to determine whether the behavior observed with the predefined posterior threshold persists under reasonable perturbations of that choice.

E.3 Alert-Budget Sensitivity

We examine sensitivity of operating-point performance to the predefined alert-mass parameter β under the common focus condition (mixed defects, level 0.05). The learned representation, reference bank, and anomaly scores are unchanged across this analysis; varying β changes the post-hoc operating point at which the continuous anomaly scores are converted into binary decisions.

Figure 20 characterizes how F1 changes as the permitted alert fraction varies. Because β determines the number of observations flagged at the operating point, changing it modifies the precision–recall trade-off even though the underlying anomaly-score ranking is unchanged. Increasing β can increase recall by admitting more observations into the predicted-anomaly set, but it can also reduce precision; F1 therefore need not vary monotonically with the alert budget.

Across the evaluated range, the trajectories provide a direct sensitivity check for the operating-point choice. Gradual changes over a range of β values would indicate that the reported F1 is not critically dependent on one narrowly specified alert fraction, whereas stronger dataset-specific changes would indicate greater operating-point sensitivity. This analysis should not be interpreted as robustness of the learned representation or anomaly-score ranking to β, since those quantities are fixed before post-hoc alert-mass calibration.

E.4 Reference-Bank Size Sensitivity

Protocol. We evaluate sensitivity of CI-FULL to the reference-bank cap $\lvert\mathcal{R}\rvert$ used during final kNN scoring. For each dataset and seed, the CI-selected indices are constructed once using the maximum evaluated cap. The resulting reference pool is then deterministically subsampled without replacement to the smaller evaluated caps, using a fixed seed-dependent sampling procedure, without retraining the encoder. Detection is re-evaluated under the same experimental conditions and fixed alert-mass operating protocol.

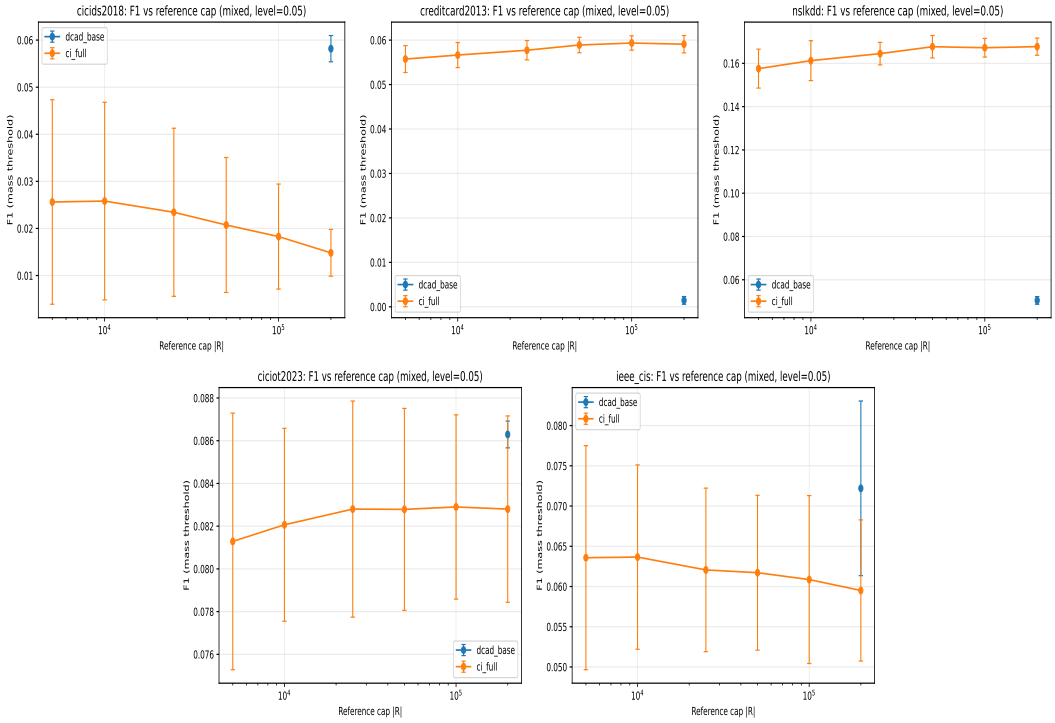


Figure 21: Sensitivity of CI-FULL F1 to the reference-bank cap $|\mathcal{R}|$ at the focus condition (mixed defects, level 0.05). The encoder and CI-selected reference pool are held fixed while the number of retained reference embeddings is varied. Curves show mean $\pm$ standard deviation over the five predefined random seeds.

Table 13: CI-FULL F1 difference relative to the matched baseline, averaged across the six evaluated reference-bank caps at the focus condition (mixed defects, level 0.05). The standard deviation summarizes variation of the mean F1 difference across caps and should not be interpreted as seed-level uncertainty.

Dataset	Mean ΔF1	SD across caps	Caps
nslkdd	$+0.1139$	0.0042	6
creditcard2013	$+0.0565$	0.0015	6
ciciot2023	-0.0039	0.0006	6
ieee_cis	-0.0103	0.0016	6
cicids2018	-0.0367	0.0044	6

This construction isolates the effect of changing the number of retained reference embeddings more directly than retraining a separate representation for every cap. The CI selection itself and the learned encoder are held fixed, while the subset of selected reference embeddings available to the final kNN scorer changes with the cap.

Findings. Figure 21 reports F1 across the evaluated reference-bank caps at the focus condition. The trajectories characterize whether detector performance depends strongly on a particular reference-bank size and whether this sensitivity differs across datasets.

Table 13 provides a complementary summary of the CI-FULL gain relative to the matched baseline across the evaluated reference caps. Averaging over caps summarizes the direction and magnitude of the detector-level difference across the tested reference-size range; it should not be interpreted as identifying an optimal reference-bank size.

Table 13 summarizes the same sensitivity analysis in terms of the CI-FULL F1 difference relative to the matched baseline. The behavior is heterogeneous across datasets. Averaged over the six evaluated caps, the difference is positive on NSL-KDD and CreditCard2013, close to zero but negative on CICIoT2023, and negative on IEEE-CIS and CICIDS2018. Together with Figure 21, these results show whether the direction and magnitude of the CI-FULL–baseline difference remain similar as the number of reference embeddings available to the kNN scorer is varied.

When a dataset-level pattern remains similar across the tested caps, the result indicates that the observed CI-FULL–baseline difference is not solely explained by one particular reference-cap choice within the evaluated range. At the same time, the heterogeneous signs across datasets reinforce the central finding that restricting the reference bank is not uniformly beneficial: the usefulness of purification depends on what reference support is retained, not only on how many reference embeddings are retained. This experiment does not establish that reference-bank size is irrelevant in general, nor does it identify dataset characteristics as the causal source of the performance difference. Reference composition, normal-support coverage, representation geometry, and local kNN structure can all contribute to the observed detector behavior.

Table 14: CI selection reproducibility across the five predefined random seeds, measured by pairwise Jaccard similarity of the selected warmup and reference index sets. Values are mean ± standard deviation across the 10 seed pairs for each dataset.

Dataset	Warmup Jaccard	Reference Jaccard
CICIDS2018	0.978 ± 0.005	0.751 ± 0.005
CICIoT2023	0.801 ± 0.116	0.801 ± 0.116
CreditCard2013	0.969 ± 0.002	0.969 ± 0.002
IEEE-CIS	0.959 ± 0.002	0.626 ± 0.003
NSL-KDD	0.956 ± 0.023	0.956 ± 0.023

Table 15: Incomplete seed coverage in the contamination-stress analysis. All conditions not listed contain the complete set of five predefined random seeds.

Dataset	π_{train}	λ	Valid seeds
NSL-KDD	0.25	0.00	4/5
NSL-KDD	0.25	0.05	4/5

E.5 CI Selection Stability Across Seeds

CI-RS contains stochastic components through representation learning and the experimental pipeline. We therefore examine whether the observations selected for the CI-controlled warmup and final reference construction are reproducible across the predefined random seeds.

Table 14 reports pairwise Jaccard similarity between the selected index sets obtained from different seeds. For two selected sets $\mathcal{I}_a$ and $\mathcal{I}_b$, the Jaccard similarity is

$$J(\mathcal{I}_a, \mathcal{I}_b) = \frac{|\mathcal{I}_a \cap \mathcal{I}_b|}{|\mathcal{I}_a \cup \mathcal{I}_b|}.$$

Values closer to one indicate greater overlap between the selected index sets, whereas values closer to zero indicate lower overlap. Reported values summarize the pairwise similarities across the evaluated seed pairs.

Across the evaluated seeds, warmup selection is highly reproducible on CICIDS2018, CreditCard2013, IEEE-CIS, and NSL-KDD, with mean pairwise Jaccard similarity above 0.95. CICIoT2023 exhibits greater seed sensitivity, with a mean warmup Jaccard of 0.801 and substantially larger variation across seed pairs.

Reference-set stability is more heterogeneous. CreditCard2013 and NSL-KDD retain high overlap across seeds, whereas CICIDS2018 and IEEE-CIS have lower mean reference-set Jaccard similarities of 0.751 and 0.626, respectively. Thus, the CI-selected sets are not uniformly seed invariant, particularly at the final reference-construction stage.

These Jaccard statistics measure reproducibility of the selected observations, not the quality of the resulting reference set. A high Jaccard similarity indicates that different seeds produce substantially overlapping selected sets, but it does not establish that those sets have low effective contamination α_{eff} or preserve the relevant normal support. Conversely, lower overlap does not by itself imply poor downstream detection. Selection stability should therefore be interpreted together with the purification and detector-level analyses.

E.6 Experimental Seed Coverage

The contamination-stress analysis uses five predefined random seeds for each dataset, training-contamination setting π_{train}, and evaluated defect level λ. Of the 60 dataset–π_{train}–λ conditions in this analysis, 58 contain all five predefined seeds. The only exceptions occur for NSL-KDD at $\pi_{\mathrm{train}} = 0.25$, where four valid seed-level results are available at both evaluated values of λ.

The two incomplete conditions should therefore be interpreted as four-seed rather than five-seed estimates. This coverage check establishes the number of valid runs contributing to each aggregate result; it does not identify the cause of the missing run, which would require the corresponding execution or failure record.

Summary. The ablation and sensitivity analyses distinguish several different sources of detector behavior. The removal ablation examines the structural roles of the CI-controlled warmup and CI-selected reference pathways. The posterior threshold π controls which observations are admitted by CI-RS and can therefore change both selection size and composition. The alert-mass parameter β acts later in the pipeline and changes only the post-hoc operating point for fixed anomaly scores. The reference-bank cap changes the number of selected embeddings available to the final kNN scorer while holding the encoder and selected reference pool fixed. Finally, pairwise Jaccard similarity measures the reproducibility of CI-selected observations across seeds, while the seed-coverage table records the number of valid runs available for the corresponding aggregate analyses.

Taken together, these experiments characterize structural, hyperparameter, operating-point, reference-size, and stochastic sensitivity without treating any single analysis as evidence of universal robustness or as a causal decomposition of CI-DCUAD's performance.

F Calibration and Thresholding Analysis

This section examines the post-hoc decision and calibration stage of the evaluated anomaly-detection pipeline. We compare fixed alert-mass calibration (Mass), Peaks-Over-Threshold calibration (POT), and rolling empirical quantile calibration (RQ) from three complementary perspectives: (i) sensitivity of detector comparisons to the thresholding rule, (ii) calibration feasibility and realized operating-point behavior across the experimental grid, and (iii) alert-rate behavior as the contamination of the unlabeled training data changes.

Table 16: Thresholding comparison between CI-DCUAD and the matched baseline using fixed alert mass, POT, and Rolling Quantile (RQ). F1 is reported separately for each post-hoc thresholding rule because it depends on the selected operating point. AUROC and AP are reported once because they are computed from the same continuous anomaly scores and are unchanged by the post-hoc threshold choice.

Dataset	Method	F1 (Mass)	F1 (POT)	F1 (RQ)	AUROC	AP
cicids2018	dcad_base	0.058 ± 0.002	0.061 ± 0.002	0.060 ± 0.002	0.336 ± 0.019	0.088 ± 0.002
cicids2018	ci_full	0.015 ± 0.004	0.015 ± 0.004	0.014 ± 0.004	0.340 ± 0.035	0.085 ± 0.006
ciciot2023	dcad_base	0.086 ± 0.001	0.105 ± 0.001	0.105 ± 0.003	0.120 ± 0.009	0.946 ± 0.001
ciciot2023	ci_full	0.083 ± 0.005	0.079 ± 0.004	0.078 ± 0.006	0.074 ± 0.018	0.940 ± 0.006
creditcard2013	dcad_base	0.002 ± 0.001	0.001 ± 0.001	0.001 ± 0.001	0.323 ± 0.071	0.001 ± 0.000
creditcard2013	ci_full	0.059 ± 0.001	0.043 ± 0.002	0.044 ± 0.003	0.951 ± 0.006	0.618 ± 0.070
ieee_cis	dcad_base	0.073 ± 0.010	0.078 ± 0.011	0.080 ± 0.011	0.598 ± 0.012	0.049 ± 0.003
ieee_cis	ci_full	0.059 ± 0.010	0.060 ± 0.009	0.057 ± 0.011	0.697 ± 0.007	0.062 ± 0.002
nslkdd	dcad_base	0.050 ± 0.002	0.066 ± 0.002	0.058 ± 0.007	0.373 ± 0.022	0.393 ± 0.009
nslkdd	ci_full	0.169 ± 0.005	0.170 ± 0.005	0.165 ± 0.009	0.654 ± 0.012	0.683 ± 0.006

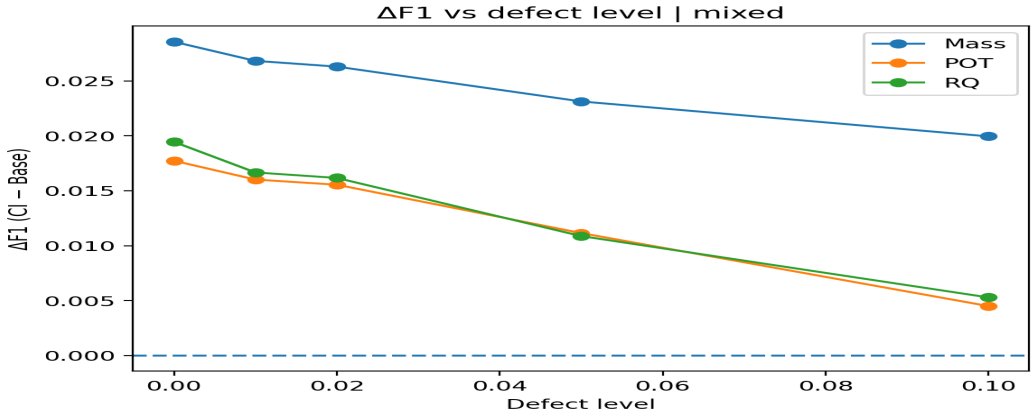


Figure 22: Aggregate ΔF1 (`ci_full` $-$ `dcad_base`) across mixed-defect levels under the evaluated thresholding rules. Positive values favor CI-DCUAD on average across the evaluated datasets; dataset-level effects remain heterogeneous.

These analyses are applied after detector score construction. Unless otherwise specified, changing the calibration procedure does not modify the learned encoder, CI-RS selection, final reference bank, or underlying continuous anomaly-score ranking. The results therefore characterize the decision layer of the pipeline rather than representation-level robustness.

F.1 Thresholding Robustness Against the Matched Baseline

We examine whether the observed detector-level differences depend strongly on the choice of decision-thresholding rule. Holding the anomaly scores fixed, we evaluate three thresholding strategies: fixed alert mass, Peaks Over Threshold (POT) with generalized Pareto tail modeling, and Rolling Quantile (RQ).

Table 16 compares CI-DCUAD (`ci_full`) with `dcad_base` under these alternative thresholding rules. The detector comparison remains heterogeneous across datasets and thresholding rules. Changing the post-hoc threshold modifies F1 in several settings, but CI-DCUAD–baseline differences remain observable under Mass, POT, and RQ. The experiment therefore tests whether the reported F1 behavior is specific to the fixed alert-mass operating rule rather than assuming threshold-rule invariance.

Figure 22 provides a complementary view by plotting the F1 difference

$$\Delta\mathrm{F1} = \mathrm{F1}_{\texttt{ci_full}} - \mathrm{F1}_{\texttt{dcad_base}}$$

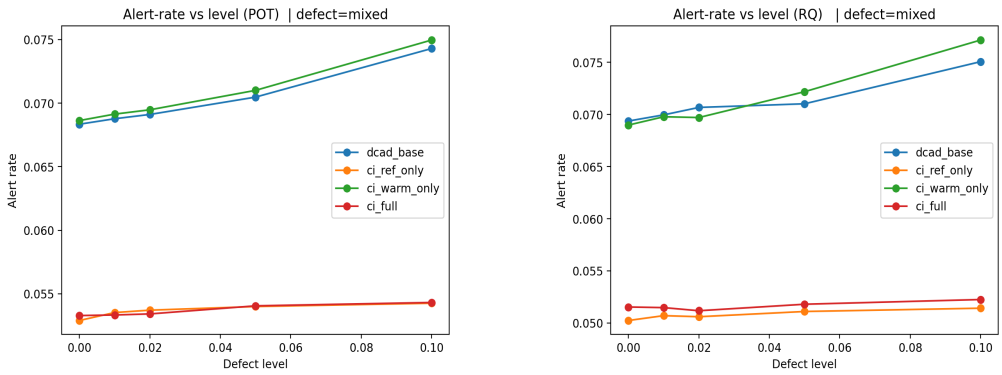
across mixed-defect levels for the evaluated thresholding rules. The aggregate F1 difference decreases as defect severity increases, particularly under POT and RQ, while the dataset-level results in Table 16 remain heterogeneous.

Interpretation. Because the underlying anomaly scores are held fixed while the thresholding rule changes, this experiment isolates sensitivity of the reported F1 comparison to the post-hoc operating rule. CI-DCUAD–baseline differences remain present under Mass, POT, and RQ, although their magnitude is dataset- and threshold-dependent. The detector-level comparison is therefore not restricted to the fixed alert-mass threshold used in the primary evaluation.

This analysis concerns only the post-hoc decision layer. It does not establish that the learned scores are invariant to corruption, that CI-RS improves score-space separability, or that one thresholding rule is uniformly preferable. Changes in the underlying anomaly-score distribution as defect severity increases can still alter the downstream performance attainable under all evaluated calibration procedures.

Table 17: Focus-condition calibration diagnostics under mixed defects at level 0.05. The table reports mean realized alert rates under Mass, POT, and RQ together with POT/RQ calibration-success rates across seeds (`pot_ok`, `rq_ok`).

Dataset	Method	N	Mass_Ar	Pot_Ok_Rate	Pot_Ar	Rq_Ok_Rate	Rq_Ar
cicids2018	ci_full	5	0.050	1.00	0.052	1.00	0.047
cicids2018	dcad_base	5	0.050	1.00	0.057	1.00	0.054
ciciot2023	ci_full	5	0.050	1.00	0.047	1.00	0.047
ciciot2023	dcad_base	5	0.050	1.00	0.062	1.00	0.062
creditcard2013	ci_full	5	0.050	1.00	0.070	1.00	0.070
creditcard2013	dcad_base	5	0.050	1.00	0.101	1.00	0.104
ieee_cis	ci_full	5	0.050	1.00	0.051	1.00	0.046
ieee_cis	dcad_base	5	0.050	1.00	0.067	1.00	0.078
nslkdd	ci_full	5	0.050	1.00	0.050	1.00	0.049
nslkdd	dcad_base	5	0.050	1.00	0.066	1.00	0.058



(a) POT (GPD tail). (b) RQ (rolling quantile).

Figure 23: Realized alert rates under mixed defects as defect severity increases for POT and RQ calibration. Unlike fixed alert-mass calibration, POT and RQ determine their operating thresholds from the corresponding tail-model and empirical-quantile procedures. The resulting alert rates therefore need not equal the predefined Mass alert fraction β.

F.2 Calibration Feasibility

We next examine calibration feasibility and realized operating-point behavior at the common focus condition. As above, Mass, POT, and RQ are applied after detector score construction. These analyses therefore characterize the decision stage and do not modify the learned encoder, CI-RS selection, final reference bank, or continuous anomaly-score ranking.

Table 17 reports realized alert rates under Mass, POT, and RQ for CI-DCUAD and the matched baseline at the focus condition (mixed defects, level 0.05), together with the corresponding POT/RQ calibration-success rates across seeds. A successful calibration indicates that the procedure produced a valid threshold for the observed score distribution; it should not be interpreted as evidence of better anomaly-score ranking, representation quality, or anomaly separation.

At the focus condition, POT and RQ calibration succeeded for all five predefined seeds for both CI-DCUAD and the matched baseline on all five datasets; consequently, no calibrated F1 values were missing in this comparison.

Full-grid results. Complete Mass/POT/RQ results across datasets, defect types, defect levels, method variants, and predefined seeds are retained in the supplementary experimental artifacts. We omit the full numerical grid here because the purpose of this analysis is to assess sensitivity to the post-hoc calibration rule rather than to duplicate the detector-level results reported elsewhere.

Alert-rate behavior under increasing defect severity. Figure 23 examines how the realized alert rate changes with mixed-defect severity under POT and RQ calibration. We omit the corresponding Mass panel because Mass fixes the alert fraction at the predefined operating budget $\beta = 0.05$ on the score set to which the quantile rule is applied; in the evaluated results, this produces the expected constant alert rate of approximately 0.05 across defect levels.

The POT and RQ results reveal method-dependent operating-point behavior as defect severity increases. In particular, the realized alert rates of `dcad_base` and `ci_warm_only` increase more noticeably with mixed-defect severity, whereas `ci_full` and `ci_ref_only` remain comparatively stable in these aggregate results. This pattern characterizes the behavior of the post-hoc calibration stage under the evaluated conditions; it should not be interpreted as evidence that CI-RS universally stabilizes anomaly-score distributions or improves continuous-score ranking.

Table 18: Alert-rate deviation under training-contamination stress. The table reports the mean absolute deviation between the realized POT/RQ alert rate and the predefined Mass operating fraction β as the training anomaly fraction α varies. Results compare `ci_full` and `dcad_base`. Because POT and RQ do not explicitly target β, these deviations characterize differences in realized operating points and should not be interpreted as measures of calibration quality.

Method	Calib	α	mean $\lvert\hat{m} - \beta\rvert$	std
ci_full	pot	0.01	0.1477	0.2799
ci_full	pot	0.05	0.1575	0.2864
ci_full	pot	0.10	0.0700	0.1123
ci_full	pot	0.30	0.0273	0.0225
ci_full	pot	0.50	0.0271	0.0155
ci_full	pot	0.75	0.0441	0.0263
dcad_base	pot	0.01	0.0541	0.0372
dcad_base	pot	0.05	0.0284	0.0079
dcad_base	pot	0.10	0.0239	0.0094
dcad_base	pot	0.30	0.0249	0.0231
dcad_base	pot	0.50	0.0386	0.0395
dcad_base	pot	0.75	0.0740	0.0793
ci_full	rq	0.01	0.1565	0.2879
ci_full	rq	0.05	0.1789	0.3254
ci_full	rq	0.10	0.1034	0.1637
ci_full	rq	0.30	0.0420	0.0500
ci_full	rq	0.50	0.0284	0.0169
ci_full	rq	0.75	0.0442	0.0279
dcad_base	rq	0.01	0.0575	0.0375
dcad_base	rq	0.05	0.0321	0.0106
dcad_base	rq	0.10	0.0237	0.0158
dcad_base	rq	0.30	0.0215	0.0216
dcad_base	rq	0.50	0.0291	0.0254
dcad_base	rq	0.75	0.0537	0.0476

F.3 Alert-Rate Behavior Under Training-Contamination Stress

Protocol. We vary the training-time anomaly fraction α and measure the realized test alert rate under the evaluated calibration procedures. The target alert fraction for fixed alert-mass calibration is denoted by β, consistent with the notation used throughout the paper. Results are reported for Mass, POT, and RQ.

This experiment varies contamination of the unlabeled training data while retaining the corresponding evaluation protocol. It therefore measures how the realized operating rate changes as training contamination varies. It is not a direct test of robustness to changes in test-time anomaly prevalence.

Mass calibration is omitted from Table 18 because its realized alert fraction equals the predefined operating budget β by construction under the evaluated protocol, yielding zero deviation throughout the sweep.

For POT and RQ, deviation from β varies with training contamination and differs between CI-Full and DCAD. The relationship is not monotonic: CI-Full exhibits larger deviations at the lower evaluated contamination levels, whereas the differences narrow or reverse in some higher-contamination settings. This indicates that the realized operating points of POT and RQ depend on the score distributions produced under the corresponding training conditions.

These deviations should not be interpreted as calibration error or anomaly-detection accuracy. POT and RQ do not explicitly target the Mass operating fraction β; consequently, a smaller deviation from β is not necessarily preferable. The experiment instead characterizes how their realized operating points vary as contamination of the unlabeled training data changes. It does not establish robustness to arbitrary distribution shift or changes in test-time anomaly prevalence.

Summary. The calibration analyses separate decision-stage behavior from the upstream representation and scoring pipeline. The matched-baseline comparison examines whether detector-level differences persist under alternative post-hoc operating rules, while the focus-condition diagnostics assess calibration feasibility and realized alert rates. The defect-severity and training-contamination analyses further characterize how realized operating points vary as the corresponding score distributions change.

Together, these experiments provide evidence about thresholding sensitivity, calibration feasibility, and operating-point behavior under the evaluated conditions. They do not establish invariance of the learned representation, reference bank, or anomaly-score ranking, nor do they imply that any calibration procedure is universally preferable.

G Extended Baseline Comparisons and Statistical Analysis

This section provides the extended comparative evidence supporting the baseline analyses in the main paper. We first report the full-grid comparison between CI-DCUAD and classical anomaly-detection baselines implemented through PyOD Zhao et al. (2019), followed by paired statistical summaries of those comparisons. We then provide additional results for the evaluated modern anomaly-detection baselines and conclude with the complete per-condition statistical comparison between `ci_full` and the matched `dcad_base`.

The analyses use different aggregation units and statistical procedures and should therefore be interpreted separately. In particular, the PyOD full-grid analysis aggregates matched experimental conditions, the modern focus-condition summary provides descriptive detector-level comparisons because the external-baseline runs use independently generated stochastic defect realizations, and the full CI-FULL–DCAD-BASE analysis performs per-condition paired tests across matched random seeds. Statistical evidence is interpreted together with the direction and magnitude of the observed differences rather than as evidence of uniform superiority.

Table 19: Aggregate comparison of CI-DCUAD and PyOD baselines across datasets under the unified evaluation protocol.

Method	F1 Mean	F1 Std	AUROC Mean	AUROC Std	AP Mean	AP Std	n_{ds}
OCSVM	0.108	0.081	0.652	0.367	0.426	0.365	5
IForest	0.106	0.093	0.636	0.377	0.426	0.392	5
PCA	0.102	0.085	0.626	0.375	0.423	0.382	5
COPOD	0.111	0.095	0.609	0.370	0.451	0.373	5
HBOS	0.111	0.091	0.594	0.366	0.413	0.386	5
ECOD	0.107	0.093	0.592	0.353	0.410	0.349	5
VAE	0.103	0.077	0.577	0.349	0.366	0.361	5
AutoEncoder	0.106	0.084	0.576	0.354	0.406	0.346	5
DeepSVDD	0.102	0.076	0.549	0.354	0.371	0.354	5
CI-DCUAD	0.077	0.057	0.543	0.340	0.478	0.388	5
LOF	0.088	0.045	0.521	0.084	0.329	0.399	5

G.1 Full-Grid Comparison with External Baselines

The full-grid analysis compares CI-DCUAD with the classical anomaly-detection baselines implemented through PyOD Zhao et al. (2019) across the evaluated datasets, defect types, defect levels, and random seeds. This analysis characterizes the heterogeneity of detector-level differences across experimental conditions.

Aggregate comparison across datasets. Before examining the condition-level paired differences, we summarize performance across the evaluated datasets under the unified evaluation protocol in Table 19. Under this aggregate averaging, CI-DCUAD has lower mean F1 and AUROC than most of the evaluated classical baselines, while attaining the highest mean AP in the table. Thus, the aggregate comparison does not establish overall detector superiority and instead shows that relative performance depends strongly on the evaluation metric. These aggregate values should be interpreted together with the condition-level paired comparisons reported below, which expose substantial dataset- and corruption-dependent heterogeneity.

For a performance metric M, the paired difference for a matched experimental condition is defined as

$$\Delta M = M_{\text{CI-DCUAD}} - M_{\text{baseline}},$$

where M denotes F1, AUROC, or AP. Positive values therefore indicate higher performance for CI-DCUAD under the corresponding metric, whereas negative values indicate higher performance for the compared baseline.

Figures 24, 25, and 26 summarize paired differences over matched (dataset, defect, level, seed) runs. Each cell reports the mean paired difference for the corresponding comparison. These figures are interpreted descriptively because observations across defect conditions and levels share datasets and random seeds and therefore should not be treated as independent experimental units.

The three metrics provide complementary views of detector performance. Figure 24 evaluates performance at the fixed alert-mass operating point, whereas Figure 25 evaluates global ranking quality and Figure 26 evaluates precision–recall ranking behavior, which is particularly informative under class imbalance.

The full-grid results are heterogeneous. CI-DCUAD exhibits positive paired differences for some baseline–dataset–metric combinations and negative differences for others. Selected F1 comparisons are favorable on CreditCard2013 Dal Pozzolo et al. (2018); MLG ULB (2018) and NSL-KDD Tavallaee et al. (2009), whereas other settings, including IEEE-CIS Kaggle (2019), contain less favorable comparisons. Similarly, the AUROC results do not indicate uniform improvement in global ranking relative to the evaluated classical detectors.

The sign reversal observed for LOF on CICIDS2018 is a post-hoc result. We currently have no label-free ex-ante diagnostic that predicts whether the associated recall gain will outweigh the corresponding precision loss; this observation is therefore not presented as a deployment rule.

AP provides an additional view of ranking behavior under class imbalance. Positive AP differences in some settings indicate higher precision–recall ranking performance for CI-DCUAD relative to the corresponding baseline, while negative differences elsewhere again demonstrate dataset- and comparator-dependent behavior. AP should not be interpreted as a direct measurement of precision within a fixed high-score tail.

Taken together, the full-grid comparison does not identify a uniformly dominant detector. Rather, the relative performance of CI-DCUAD depends on the dataset, corruption condition, metric, and comparison method. This heterogeneity is compatible with the paper's central hypothesis that the utility of contamination-aware reference selection is condition dependent, although the baseline comparison alone does not establish the mechanism responsible for any individual performance difference.

These observations are compatible with the failure-regime analysis showing that the usefulness of CI-RS can deteriorate when the preliminary score structure becomes less informative for preferential reference selection. However, the external-baseline comparisons do not directly measure q_N, q_A, normal-support distortion, or the causal contribution of representation geometry. They therefore provide comparative detector-level evidence rather than an independent test of the CI-RS mechanism.

G.2 Additional Results for Modern Baselines

The main paper reports the matched CI-DCUAD versus DCAD-BASE comparison because it provides the most direct detector-level test of CI-based reference selection. This appendix gives broader context using recent tabular anomaly-detection methods, including ECOD Li et al. (2023), DIF Xu et al. (2023a), ICL Shenkar & Wolf (2022), and SLAD Xu et al. (2023b).

The external-baseline experiments use the same nominal focus condition, train/test sizes, and evaluation metrics. However, their stochastic defect realizations were generated independently from those used for CI-DCUAD. Consequently, these results provide detector-level context rather than seed-paired estimates of method differences.

Figure 27 shows that the relative performance of CI-DCUAD depends on the dataset, metric, and comparison method. CI-DCUAD is competitive in some settings, while modern alternatives achieve stronger performance in others. The broader comparison therefore does not establish uniform detector dominance.



Figure 24: Paired full-grid F1 comparison of CI-DCUAD against PyOD baselines Zhao et al. (2019) under fixed alert-mass evaluation. Each cell shows the mean paired change $(\texttt{ci_full} - \texttt{baseline})$ over the matched experimental conditions represented by that cell. Positive values indicate higher F1 for CI-DCUAD and negative values indicate higher F1 for the corresponding baseline. The heatmap is interpreted descriptively because the full grid contains repeated observations from common datasets and random seeds across multiple corruption conditions.

Table 20: Compact descriptive comparison between CI-DCUAD and the evaluated modern baselines at the focus condition (mixed defect, $\lambda = 0.05$). Entries report differences in mean performance $(\overline{M}_{\text{CI-DCUAD}} - \overline{M}_{\text{baseline}})$ over the available runs. Positive values indicate higher mean performance for CI-DCUAD. External modern-baseline runs use independently generated stochastic defect realizations, so these differences are not interpreted as seed-paired method effects.

Compared Method	ΔF1	ΔAUROC	ΔAP	n
DCAD-Base	+0.012	+0.031	+0.030	25
Random-Ref	+0.012	+0.027	+0.031	25
ScoreTrim-Ref	+0.015	+0.033	+0.034	25
ECOD	-0.034	-0.098	-0.076	25
DIF	-0.028	-0.083	-0.054	25
ICL	-0.017	-0.108	-0.016	25
SLAD	-0.018	-0.000	-0.028	19
Oracle Clean-Ref	-0.133	-0.088	-0.139	15

The distinction between ranking and operating-point performance remains important. AUROC and AP summarize properties of the score ranking, whereas fixed-mass F1 additionally depends on the selected alert budget. A method can therefore improve a ranking metric without improving F1 at the fixed operating point.

Because the external methods were evaluated using independently generated defect realizations, differences between CI-DCUAD and these methods should not be interpreted as paired estimates of algorithmic improvement. The matched DCAD-BASE comparison in the main paper remains the primary evidence for the effect of CI-based reference selection.

Compact descriptive summary. Table 20 provides a compact descriptive comparison between CI-DCUAD and the evaluated modern anomaly-detection baselines under the common focus condition (mixed defect, $\lambda = 0.05$). For each metric, the reported difference is defined as

$$\Delta M = \overline{M}_{\text{CI-DCUAD}} - \overline{M}_{\text{baseline}},$$

where each term denotes the corresponding mean performance over the available runs. Positive values indicate higher mean performance for CI-DCUAD under the corresponding metric, whereas negative values indicate higher mean performance for the compared baseline. Because the external modern-baseline runs use independently generated stochastic defect realizations, these differences are descriptive detector-level comparisons and are not interpreted as seed-paired method effects.

The compact summary aggregates heterogeneous dataset-level comparisons. Consequently, the sign and magnitude of an aggregate gain should not be interpreted as uniform improvement over a baseline on every dataset or metric. The dataset-level results below provide the corresponding disaggregated view.

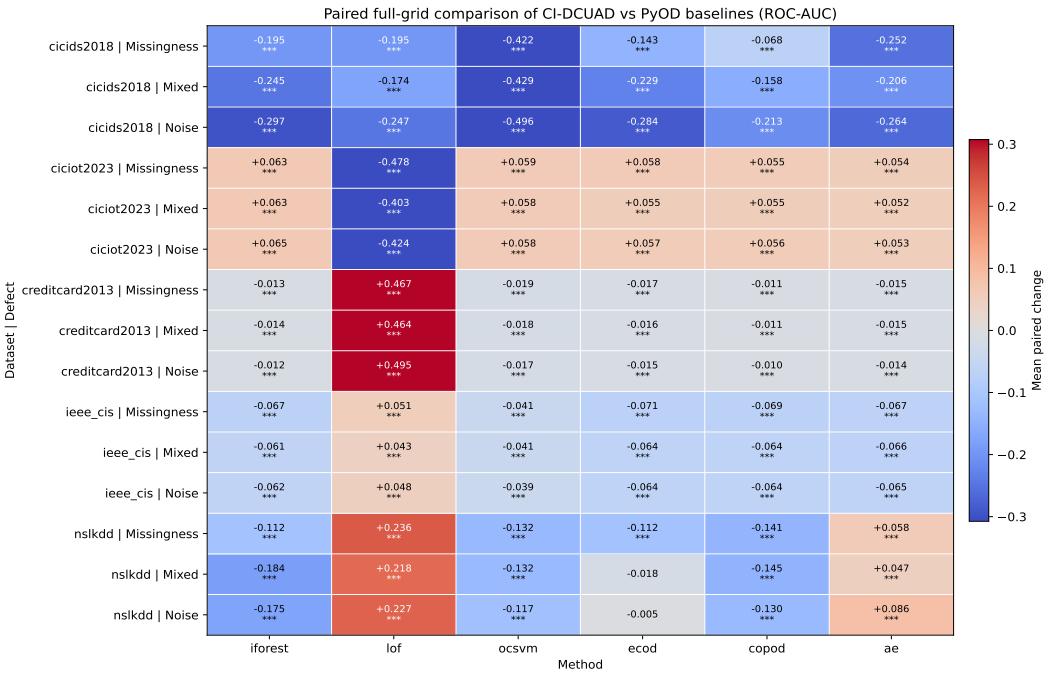


Figure 25: Paired full-grid AUROC comparison of CI-DCUAD against PyOD baselines under the evaluated experimental grid. Each cell shows the mean paired change ($\texttt{ci_full} - \texttt{baseline}$) over the matched experimental conditions represented by that cell. Positive values indicate higher AUROC for CI-DCUAD and negative values indicate higher AUROC for the corresponding baseline. The heatmap is interpreted descriptively because the full grid contains repeated observations from common datasets and random seeds across multiple corruption conditions.

Table 21: LOE and SRR-Raw results under the common focus condition. Entries are mean ± standard deviation over the five predefined seeds. LOE uses the official LatentOE-AD implementation under a fixed 20-epoch compute-matched budget; the official tabular default is 300 epochs. SRR-Raw is a paper-based raw-tabular reproduction, not an execution of official author code.

Dataset	Method	F1	AUROC	AP	Seeds
CICIDS2018	LOE	0.0280 ± 0.0494	0.4654 ± 0.0825	0.1166 ± 0.0331	5
	SRR-Raw	0.0186 ± 0.0009	0.5544 ± 0.0005	0.1213 ± 0.0004	5
CICIoT2023	LOE	0.0763 ± 0.0191	0.1213 ± 0.1354	0.9419 ± 0.0245	5
	SRR-Raw	0.0541 ± 0.0001	0.0045 ± 0.0002	0.9158 ± 0.0001	5
CreditCard2013	LOE	0.0154 ± 0.0178	0.8830 ± 0.0499	0.0311 ± 0.0444	5
	SRR-Raw	0.0616 ± 0.0004	0.9603 ± 0.0013	0.0806 ± 0.0023	5
IEEE-CIS	LOE	0.0577 ± 0.0221	0.6223 ± 0.0716	0.0559 ± 0.0123	5
	SRR-Raw	0.1562 ± 0.0010	0.7579 ± 0.0006	0.1033 ± 0.0003	5
NSL-KDD	LOE	0.1038 ± 0.0213	0.8101 ± 0.0195	0.6789 ± 0.0269	5
	SRR-Raw	0.1114 ± 0.0006	0.7838 ± 0.0006	0.6840 ± 0.0004	5

LOE and SRR-Raw focus comparison. Table 21 reports the dedicated five-seed focus-condition evaluations. All methods use the frozen train–test splits, mixed defects at $\lambda = 0.05$, and fixed alert-mass evaluation with $\beta = 0.05$.

LOE was executed through the unmodified official LatentOE-AD source. The wrapper fixed the training budget at 20 epochs for compute matching; the official tabular default is 300 epochs. This is a substantially reduced budget relative to LOE's own recommended configuration, and we did not verify that LOE has converged by epoch 20 on these datasets; LOE's reported numbers in Table 21 should therefore be read as a compute-matched comparison rather than as LOE's best achievable performance, and may understate it. SRR-Raw follows the published refine-and-repeat structure using identity representations, five disjoint diagonal Gaussian density estimators, fixed $\gamma = 0.10$, unanimous normal retention, and a final diagonal-Gaussian scorer. No official SRR repository was available from the publication page, so SRR-Raw is explicitly reported as a paper-based reproduction.

Mixed-defect stress trajectories. We next compare CI-DCUAD with the evaluated modern baselines over an extended mixed-defect stress grid,

$$\lambda \in \{0.01, 0.03, 0.05, 0.10, 0.20\}.$$

This auxiliary grid differs from the standard CI-S1 grid and is used specifically for the modern-baseline stress comparison. The experiment examines how AUROC changes as the severity of the specified structured-defect process increases.

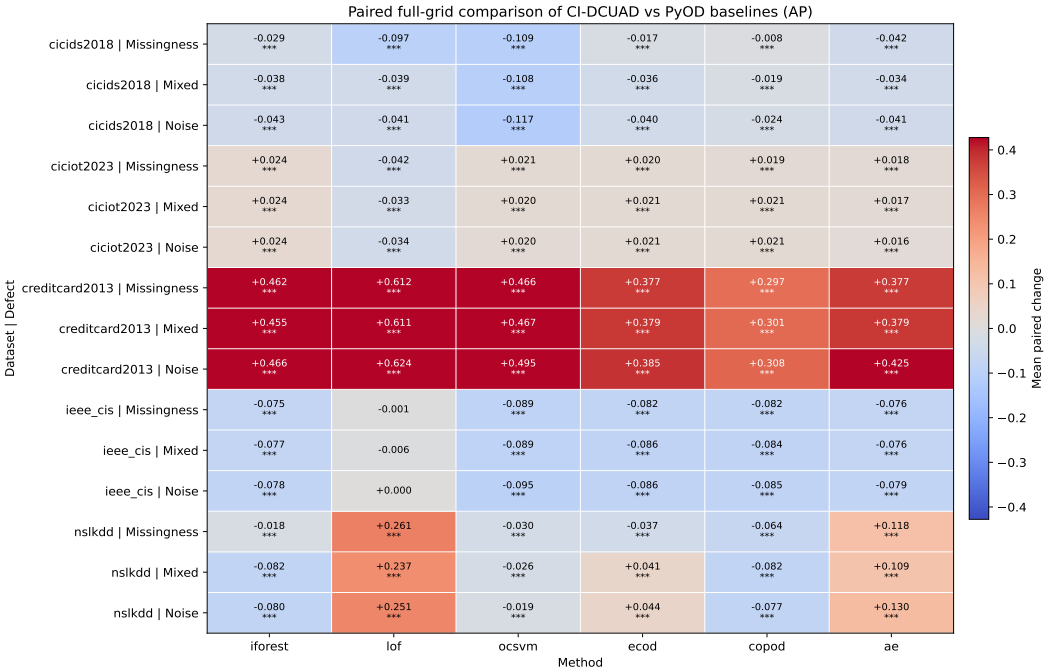
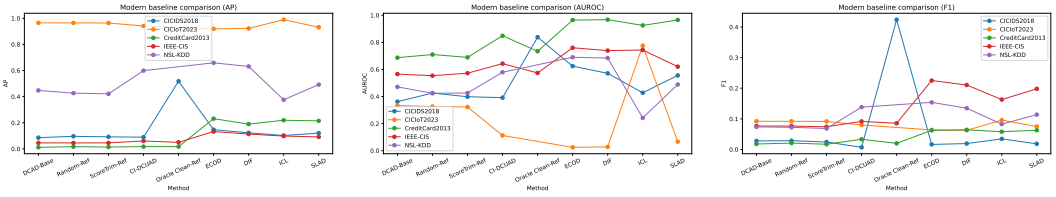


Figure 26: Paired full-grid AP comparison of CI-DCUAD against PyOD baselines under the evaluated experimental grid. Each cell shows the mean paired change (`ci_full − baseline`) over the matched experimental conditions represented by that cell. Positive values indicate higher AP for CI-DCUAD and negative values indicate higher AP for the corresponding baseline. The heatmap is interpreted descriptively because the full grid contains repeated observations from common datasets and random seeds across multiple corruption conditions.



(a) Average Precision. (b) AUROC. (c) Fixed-mass F1.

Figure 27: Comparison of CI-DCUAD with modern and classical anomaly-detection baselines under the focus condition (mixed defects, $\lambda = 0.05$). Error bars denote standard deviation across the available completed runs. External-baseline runs use the same nominal defect condition and train/test sizes but independently generated stochastic defect realizations. The comparisons therefore provide detector-level context and are not interpreted as seed-paired method effects.

Figure 28 shows substantial variation across datasets and methods. CI-DCUAD changes relatively gradually in several evaluated settings, while some external detectors show larger changes on particular datasets. The figure does not establish a single cross-dataset ordering of robustness.

The external-baseline trajectories must be interpreted descriptively. Although all methods are evaluated under the same nominal severity levels, their stochastic defect realizations were generated independently. The curves therefore do not constitute seed-paired comparisons of degradation rates.

Moreover, gradual AUROC change over this grid is not equivalent to robustness under arbitrary corruption. The analysis covers a specified structured-defect process and a finite severity range; it does not evaluate adaptive poisoning, distribution shift, or worst-case perturbations.

Stress-trajectory criterion. To summarize the trajectories in Figure 28, we apply a predefined operational criterion for large performance degradation. A trajectory is classified as a collapse when its AUROC decreases by more than 0.15 or its AP decreases by more than 0.10 between the evaluated endpoints. Table 22 reports the resulting counts over evaluable dataset–seed trajectories.

Across the 25 evaluable CI-DCUAD dataset–seed trajectories, none met the predefined collapse criterion. Among the other evaluable methods, collapses occurred for ECOD on NSL-KDD and for ICL on CICIDS2018, CICIoT2023, and NSL-KDD; DIF and SLAD had no trajectories meeting the criterion. DCAD-Base is not included in the aggregate table because no complete trajectory was available for this analysis.

This criterion is an analysis-specific summary of large metric degradation over the evaluated stress range. It is not a general definition of detector failure and does not establish robustness to arbitrary contamination, corruption, or distribution shift.

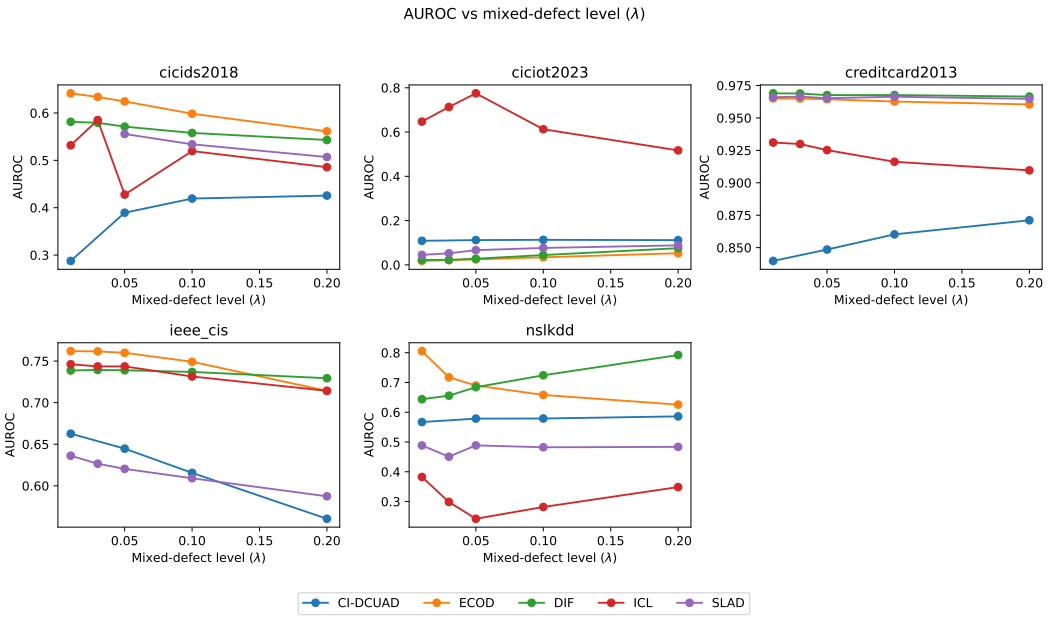


Figure 28: AUROC trajectories of CI-DCUAD and the evaluated modern anomaly-detection baselines across the extended mixed-defect stress grid $\lambda \in \{0.01, 0.03, 0.05, 0.10, 0.20\}$. Trajectories are dataset and method dependent. External-baseline runs use independently generated stochastic defect realizations and therefore provide detector-level stress context rather than seed-paired estimates of differences in degradation rate.

Table 22: Stress-trajectory summary under the predefined collapse criterion. Counts are aggregated over evaluable dataset–seed trajectories. DCAD-Base is omitted because no complete trajectory was available under this analysis.

Method	Collapses	Evaluable trajectories	Collapse rate
CI-DCUAD	0	25	0.000
ECOD	5	25	0.200
DIF	0	25	0.000
ICL	4	25	0.160
SLAD	0	19	0.000

Full numerical results. To complement the visual comparison in Figure 27, Table 23 reports the complete per-dataset results from the dedicated modern-baseline experiment at the focus condition. The table includes DCAD-Base, Random-Ref, ScoreTrim-Ref, and CI-DCUAD from the corresponding reference-selection experiment, together with ECOD, DIF, ICL, and SLAD evaluated under the same nominal mixed-defect severity and train/test sizes. Oracle Clean-Ref is included only as a label-informed diagnostic.

The external-baseline runs use independently generated stochastic defect realizations and therefore should be interpreted as broader detector-level context rather than as seed-paired comparisons with CI-DCUAD. In contrast, comparisons among CI-DCUAD and its reference-selection controls are used to study the effect of reference construction within their corresponding experimental protocol.

Taken together, the dataset-level results show substantial metric- and dataset-dependent variation. Within the reference-selection results reported in this analysis, CI-DCUAD exceeds DCAD-Base, Random-Ref, and ScoreTrim-Ref on all three reported metrics for CREDITCARD2013, IEEE-CIS, and NSL-KDD, whereas this pattern does not hold for CICIDS2018 or CICIoT2023. This heterogeneity is consistent with the central conditional question of the study: reference purification can be beneficial in favorable regimes, but its benefit is not universal.

The external-baseline results provide complementary detector-level context. Because these runs use independently generated stochastic defect realizations, we do not interpret their numerical differences from CI-DCUAD as seed-paired method effects. At the aggregate level, they nevertheless show that CI-DCUAD does not exhibit uniform metric-wise dominance over ECOD, DIF, ICL, or SLAD, reinforcing that the contribution of this work is a conditional reference-purification mechanism rather than a claim of universal detector superiority.

G.3 Full Statistical Analysis of CI-DCUAD vs. DCAD-Base

The preceding analyses compare CI-DCUAD with external detectors. We finally report the complete paired statistical analysis against `dcad_base`, the matched internal baseline used to isolate the structural changes introduced by the CI pathways.

For each dataset, defect type, defect level, and metric, runs are paired by random seed. The seed-level paired gain is

$$\Delta M_s = M_{\texttt{ci_full},s} - M_{\texttt{dcad_base},s},$$

where s indexes a matched seed. Positive gain indicates higher performance for `ci_full`, whereas negative gain indicates higher performance for `dcad_base`.

Table 23: Per-dataset results from the dedicated modern-baseline analysis at the nominal focus condition (mixed defects, $\lambda = 0.05$). Metrics are reported as mean $\pm$ standard deviation over the available predefined seeds, with n giving the number of completed runs. The external baselines use the same nominal defect severity and train/test sizes but independently generated stochastic defect realizations; their results therefore provide detector-level context rather than seed-paired comparisons with CI-DCUAD. Entries with $n = 1$ report the single available run and therefore have no across-seed standard deviation. Oracle Clean-Ref is label-informed and is included only as a diagnostic.

Dataset	Method	F1	AUROC	AP	n
CICIDS2018	DCAD-Base	0.028 ± 0.003	0.362 ± 0.021	0.086 ± 0.003	5
CICIDS2018	Random-Ref	0.029 ± 0.004	0.425 ± 0.016	0.096 ± 0.003	5
CICIDS2018	ScoreTrim-Ref	0.024 ± 0.004	0.398 ± 0.037	0.092 ± 0.006	5
CICIDS2018	CI-DCUAD	0.007 ± 0.003	0.391 ± 0.025	0.090 ± 0.004	5
CICIDS2018	Oracle Clean-Ref	0.424 ± 0.032	0.838 ± 0.027	0.518 ± 0.052	5
CICIDS2018	ECOD	0.016 ± 0.001	0.624 ± 0.002	0.147 ± 0.001	5
CICIDS2018	DIF	0.019 ± 0.001	0.571 ± 0.015	0.123 ± 0.004	5
CICIDS2018	ICL	0.035 ± 0.022	0.428 ± 0.102	0.102 ± 0.024	5
CICIDS2018	SLAD	0.019	0.556	0.121	1
CICIoT2023	DCAD-Base	0.092 ± 0.001	0.333 ± 0.044	0.965 ± 0.003	5
CICIoT2023	Random-Ref	0.092 ± 0.001	0.326 ± 0.108	0.965 ± 0.007	5
CICIoT2023	ScoreTrim-Ref	0.092 ± 0.001	0.323 ± 0.060	0.964 ± 0.005	5
CICIoT2023	CI-DCUAD	0.080 ± 0.003	0.112 ± 0.037	0.941 ± 0.008	5
CICIoT2023	ECOD	0.064 ± 0.000	0.025 ± 0.001	0.920 ± 0.000	5
CICIoT2023	DIF	0.063 ± 0.001	0.028 ± 0.002	0.922 ± 0.000	5
CICIoT2023	ICL	0.096 ± 0.001	0.775 ± 0.178	0.990 ± 0.009	5
CICIoT2023	SLAD	0.075 ± 0.002	0.066 ± 0.012	0.931 ± 0.002	5
CreditCard2013	DCAD-Base	0.018 ± 0.003	0.687 ± 0.043	0.012 ± 0.005	5
CreditCard2013	Random-Ref	0.021 ± 0.004	0.710 ± 0.054	0.017 ± 0.008	5
CreditCard2013	ScoreTrim-Ref	0.017 ± 0.003	0.689 ± 0.037	0.014 ± 0.007	5
CreditCard2013	CI-DCUAD	0.033 ± 0.003	0.849 ± 0.021	0.018 ± 0.002	5
CreditCard2013	Oracle Clean-Ref	0.020 ± 0.003	0.734 ± 0.044	0.017 ± 0.009	5
CreditCard2013	ECOD	0.063 ± 0.000	0.964 ± 0.000	0.232 ± 0.003	5
CreditCard2013	DIF	0.064 ± 0.000	0.968 ± 0.000	0.190 ± 0.012	5
CreditCard2013	ICL	0.058 ± 0.005	0.925 ± 0.020	0.219 ± 0.039	5
CreditCard2013	SLAD	0.063 ± 0.001	0.965 ± 0.004	0.215 ± 0.029	5
IEEE-CIS	DCAD-Base	0.077 ± 0.010	0.566 ± 0.026	0.046 ± 0.005	5
IEEE-CIS	Random-Ref	0.076 ± 0.007	0.554 ± 0.015	0.046 ± 0.003	5
IEEE-CIS	ScoreTrim-Ref	0.074 ± 0.013	0.572 ± 0.018	0.046 ± 0.004	5
IEEE-CIS	CI-DCUAD	0.092 ± 0.004	0.643 ± 0.020	0.061 ± 0.005	5
IEEE-CIS	Oracle Clean-Ref	0.086 ± 0.003	0.574 ± 0.016	0.050 ± 0.002	5
IEEE-CIS	ECOD	0.225 ± 0.001	0.760 ± 0.001	0.133 ± 0.001	5
IEEE-CIS	DIF	0.211 ± 0.002	0.739 ± 0.001	0.114 ± 0.001	5
IEEE-CIS	ICL	0.163 ± 0.006	0.744 ± 0.008	0.098 ± 0.004	5
IEEE-CIS	SLAD	0.198 ± 0.004	0.620 ± 0.008	0.092 ± 0.003	3
NSL-KDD	DCAD-Base	0.074 ± 0.007	0.471 ± 0.035	0.448 ± 0.018	5
NSL-KDD	Random-Ref	0.072 ± 0.007	0.424 ± 0.027	0.427 ± 0.014	5
NSL-KDD	ScoreTrim-Ref	0.069 ± 0.005	0.426 ± 0.039	0.422 ± 0.018	5
NSL-KDD	CI-DCUAD	0.138 ± 0.014	0.579 ± 0.024	0.599 ± 0.028	5
NSL-KDD	ECOD	0.154 ± 0.001	0.690 ± 0.003	0.659 ± 0.003	5
NSL-KDD	DIF	0.135 ± 0.002	0.684 ± 0.014	0.631 ± 0.010	5
NSL-KDD	ICL	0.083 ± 0.040	0.242 ± 0.130	0.376 ± 0.081	5
NSL-KDD	SLAD	0.114 ± 0.005	0.489 ± 0.086	0.492 ± 0.038	5

These quantities provide complementary information. The mean paired gain describes direction and magnitude in the original metric scale, Cohen's d_z provides a standardized paired effect size, and the inferential tests quantify evidence relative to their corresponding null hypotheses. Statistical significance should therefore be interpreted together with the direction and magnitude of the observed effect. A small p- or q-value does not by itself imply a large practical improvement or uniform superiority across conditions.

Condition-level ablation summary. We first summarize the principal F1 contrasts from the structural-ablation analysis. The five seeds are averaged within each dataset–defect–severity condition, producing 75 condition-level analysis units. This prevents repeated seeds from being treated as independent condition-level observations.

Table 24 reports the complete six-contrast family. Warmup-only does not improve over DCAD-Base ($\Delta\mathrm{F1} = -0.0025$, adjusted $p = 0.650$), and reference-only does not significantly exceed either DCAD-Base or warmup-only after correction (both adjusted $p = 0.158$). The positive mean difference of CI-Full relative to DCAD-Base also does not remain statistically significant ($\Delta\mathrm{F1} = +0.0262$, adjusted $p = 0.158$), consistent with the cross-dataset heterogeneity in Figure 17.

CI-Full nevertheless significantly exceeds both incomplete CI variants. Relative to CI-Warm-Only, its mean F1 difference is $+0.0287$, with a bootstrap 95% confidence interval of $[0.0164, 0.0416]$ and adjusted $p = 0.0359$. Relative to CI-Ref-Only, its mean difference is $+0.0028$, with a bootstrap 95% confidence interval of $[0.0017, 0.0039]$ and adjusted $p = 1.03 \times 10^{-4}$.

Table 24: Complete family of six condition-level F1 contrasts among the four structural-ablation variants. Mean differences are computed over 75 matched dataset–defect–severity conditions. Two-sided Wilcoxon p-values are adjusted jointly over all six contrasts using Holm's procedure. Positive differences favor the first configuration.

Contrast	Mean ΔF1	Holm-adjusted p
CI-Warm-Only − DCAD-Base	-0.0025	0.650
CI-Ref-Only − DCAD-Base	$+0.0234$	0.158
CI-Full − DCAD-Base	$+0.0262$	0.158
CI-Ref-Only − CI-Warm-Only	$+0.0259$	0.158
CI-Full − CI-Warm-Only	$+0.0287$	0.0359
CI-Full − CI-Ref-Only	$+0.0028$	1.03×10^{-4}

These results indicate interaction between the warmup and reference-selection pathways, but they do not establish that CI-Full uniformly improves over the matched base detector. Statistical improvement over incomplete variants is different from statistical improvement over DCAD-Base.

Full-grid statistical results. The complete seed-paired analysis spans all evaluated combinations of dataset, defect type, defect level, and metric. For each condition, we compute the mean and standard deviation of the paired gain, Cohen's d_z, a two-sided paired t-test, and a two-sided Wilcoxon signed-rank test when the paired differences permit the test. Benjamini–Hochberg false-discovery-rate correction is applied to the paired t-test p-values separately for each metric across the full grid. The complete per-condition numerical results are provided with the supplementary experimental artifacts rather than reproduced in the manuscript.

Each individual condition in this grid uses only the five predefined real-dataset seeds, so the per-condition Wilcoxon signed-rank test is subject to a discreteness floor: with $n = 5$, the minimum achievable two-sided exact p-value is $2 \times (1/2)^5 = 0.0625$, which already exceeds the conventional 0.05 threshold before any correction is applied. No single-condition Wilcoxon result in this grid can therefore reach significance at $\alpha = 0.05$ regardless of effect size. We report these per-condition Wilcoxon results for descriptive completeness and rely on the paired effect magnitude, Cohen's d_z, and the Benjamini–Hochberg corrected paired t-test (which does not share this discreteness limitation) rather than the per-condition Wilcoxon p-value for any claim of statistical significance in this appendix. The better-powered, 75-condition-level aggregation used in Table 24 does not share this floor and is the basis for the significance claims reported in the main text.

The full-grid analysis reveals substantial heterogeneity across datasets, conditions, and metrics. Favorable effects are concentrated in CREDITCARD2013 and NSL-KDD, whereas CICIDS2018, CICIoT2023, and IEEE-CIS include unfavorable or metric-dependent effects. Accordingly, aggregate positive differences should not be interpreted as uniform improvement over all evaluated conditions.

These results reinforce the conditional interpretation of the method rather than a claim of uniform superiority. In particular, improvements in ranking metrics need not translate into improvements at the fixed F1 operating point, and statistical evidence should be interpreted together with the direction and magnitude of the paired effect.

Summary. The extended baseline analyses reveal substantial heterogeneity across datasets, metrics, corruption conditions, and comparison methods. The PyOD experiments provide broad external-detector context through aggregate and condition-level descriptive comparisons, while the modern-baseline analysis provides descriptive focus-condition comparisons and an operational stress-trajectory analysis. The CI-FULL–DCAD-BASE analysis provides the corresponding seed-paired statistical evaluation within individual dataset–defect–level conditions.

Taken together, these analyses do not support a claim of uniform detector superiority. Rather, they characterize dataset-, metric-, and condition-dependent performance differences while distinguishing broad external-detector evidence from the matched internal statistical analysis. This heterogeneity is consistent with the central premise of the study that the utility of contamination-aware reference purification is condition dependent.

H Additional Representation and Stability Results

This section provides complementary analyses of representation geometry and cross-seed reproducibility. These analyses address two distinct questions. First, we examine whether detector-level behavior is accompanied by observable changes in the geometry of the learned representation. Second, we characterize the completeness of the multi-seed experimental grid and the variability of detector-level performance across the predefined random seeds.

The two analyses should not be conflated. Geometry statistics are post-hoc diagnostics of the learned embedding space, whereas cross-seed variability quantifies reproducibility of the reported detector-level metrics. Neither analysis, by itself, establishes a causal mechanism for the observed performance differences.

H.1 Additional Representation-Geometry Results

This subsection provides post-hoc embedding-geometry diagnostics that complement the mechanism and detector-level analyses reported in the main paper. Aggregate statistics are reported across the five evaluated datasets using the predefined random seeds

$$\{7, 13, 21, 37, 51\},$$

while the complete per-seed geometry measurements are retained in the supplementary experimental artifacts.

The geometry analysis is performed on clean test data using embeddings produced by the trained detectors. The reported diagnostics include within-class compactness for normal and anomalous observations, between-class separation, margin, and silhouette-based measures, together with the corresponding differences between CI-DCUAD and the matched baseline.

Ground-truth labels are required to compute these class-conditional geometry statistics. They are therefore used only for post-hoc analysis and are not available to CI-DCUAD during representation learning, CI-RS selection, reference-bank construction, anomaly scoring, or alert-mass

Table 25: Clean-test embedding-geometry comparison between `dcad_base` and CI-DCUAD (`ci_full`). Values summarize inter-class separation, margin, and silhouette statistics across the five predefined seeds. Ground-truth labels are used only for this post-hoc diagnostic analysis and are not used by the detector.

(a) Separation and margin statistics

Dataset	Base sep.	CI sep.	Base margin	CI margin	Δ margin
CICIDS2018	1.196 ± 0.020	1.149 ± 0.045	0.030 ± 0.012	0.031 ± 0.010	0.002 ± 0.013
CICIoT2023	1.275 ± 0.088	1.208 ± 0.053	0.363 ± 0.067	0.389 ± 0.037	0.026 ± 0.039
CreditCard2013	1.310 ± 0.022	1.328 ± 0.017	-0.001 ± 0.013	0.023 ± 0.033	0.024 ± 0.040
IEEE-CIS	1.265 ± 0.030	1.283 ± 0.037	-0.001 ± 0.005	0.007 ± 0.010	0.008 ± 0.008
NSL-KDD	0.587 ± 0.479	0.462 ± 0.407	0.165 ± 0.160	0.107 ± 0.102	-0.058 ± 0.059

(b) Silhouette statistics

Dataset	Base sil.	CI sil.	Δ sil.
CICIDS2018	0.036 ± 0.005	0.030 ± 0.013	-0.006 ± 0.018
CICIoT2023	0.110 ± 0.019	0.073 ± 0.022	-0.037 ± 0.028
CreditCard2013	-0.011 ± 0.027	0.018 ± 0.031	0.029 ± 0.041
IEEE-CIS	0.002 ± 0.002	0.007 ± 0.006	0.005 ± 0.007
NSL-KDD	0.239 ± 0.021	0.206 ± 0.025	-0.033 ± 0.013

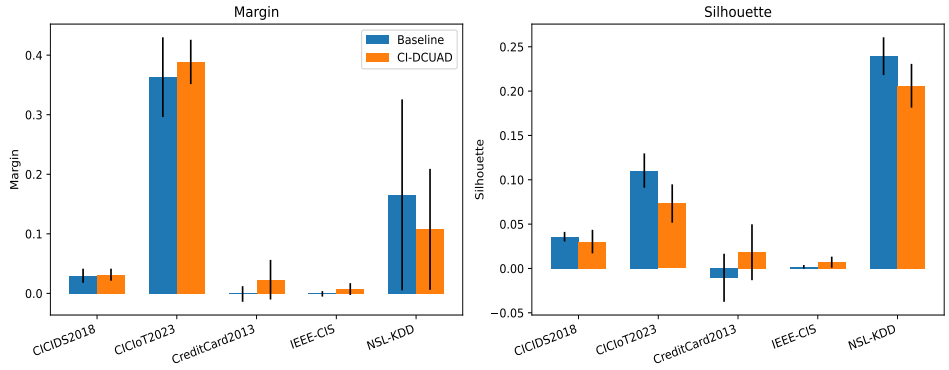


Figure 29: Embedding-geometry comparison on clean test data. CI-DCUAD and the matched baseline exhibit dataset-dependent differences in margin and silhouette, illustrating that the effect of CI-based selection on the learned representation is not uniform across datasets.

calibration. Accordingly, the geometry results should be interpreted as diagnostic measurements of the learned representations rather than quantities optimized directly by the method.

Aggregate clean-test geometry. Table 25 compares `dcad_base` and `ci_full` using intra-class compactness, inter-class separation, margin, and silhouette statistics. These quantities are computed post hoc using ground-truth labels for analysis only; labels are not used by CI-DCUAD during training, reference selection, scoring, or calibration.

Figure 29 provides a complementary visualization of margin and silhouette behavior across the evaluated datasets.

The geometry results are heterogeneous. CI-DCUAD yields positive mean margin changes on CICIDS2018 Canadian Institute for Cybersecurity (CIC), University of New Brunswick (2018), CICIoT2023 Canadian Institute for Cybersecurity (CIC), University of New Brunswick (2023), CreditCard2013 Dal Pozzolo et al. (2018); MLG ULB (2018), and IEEE-CIS Kaggle (2019), although the magnitude of these changes varies across datasets, while NSL-KDD exhibits a negative mean margin change. The silhouette statistics are similarly mixed: they increase on CreditCard2013 and IEEE-CIS but decrease on CICIDS2018, CICIoT2023, and NSL-KDD. Thus, changes in one geometry diagnostic need not be accompanied by corresponding changes in another, indicating that the CI intervention does not induce a single uniformly favorable geometric transformation.

Interpretation. These results provide representation-level evidence that CI-based selection can alter the geometry of the learned embedding, but the effect is dataset dependent. In some settings, the change is associated with increased normal–anomaly margin, whereas in others the geometric differences are small or mixed across the evaluated statistics.

The analysis should therefore be interpreted diagnostically rather than causally.

Per-seed geometry results. Complete per-seed geometry diagnostics for all five datasets, including within-class compactness, between-class separation, margin, silhouette, and their CI-DCUAD–baseline differences, are retained in the supplementary experimental artifacts. The aggregate statistics in Table 25 summarize these seed-level measurements.

The seed-level diagnostics show that changes in compactness, separation, margin, and silhouette need not have the same direction across datasets or random seeds. Improvement in an individual global geometry statistic is therefore neither necessary nor sufficient for improved anomaly detection. Because CI-DCUAD scores observations relative to the selected reference bank, downstream behavior depends jointly on the learned representation, reference-bank composition and support, and the resulting local kNN geometry.

Accordingly, these post-hoc measurements complement the reference-selection ablations and effective-contamination analysis but do not establish that CI-RS directly optimizes any particular global geometric property.

H.2 Additional Seed Coverage and Aggregation

CI-S1 completeness verification. We verified the completeness of the CI-S1 experimental grid before cross-seed aggregation. For each seed–dataset pair, the expected grid contains 15 conditions corresponding to three defect types and five defect levels. We additionally checked for duplicate run keys and missing values in the reported performance metrics.

Using the five predefined seeds

$$\{7, 13, 21, 37, 51\},$$

five datasets, three defect types, and five defect levels, the complete CI-S1 grid contains

$$5 \times 5 \times 3 \times 5 = 375$$

runs. All expected CI-S1 configurations were present in the aggregated results.

This completeness check concerns the CI-S1 grid summarized in this subsection. It is distinct from condition-specific valid-seed coverage reported elsewhere for stress experiments in which individual runs may be unavailable.

Seed protocol and aggregation. Multi-seed experiments use the same predefined train/test split for a given dataset across seeds, while stochastic components of the experimental pipeline vary with the seed. These include encoder initialization, minibatch ordering, and stochastic defect generation.

For each configuration

$$(\text{dataset, defect type, defect level}),$$

performance metrics are first computed independently for each seed and stored before cross-seed aggregation. For a metric M evaluated over $S = 5$ seeds, the reported mean is

$$\overline{M} = \frac{1}{S} \sum_{s=1}^{S} M_s,$$

and the corresponding standard deviation summarizes cross-seed variability. No seed is removed or selected according to downstream performance.

Cross-seed performance under mixed defects. Figure 30 summarizes detector-level performance across the five predefined random seeds as mixed-defect severity increases for all five datasets. Curves report the cross-seed mean, and error bars denote one standard deviation for each experimental configuration.

Seed-level variability. Cross-seed variability is metric-, dataset-, and condition-dependent. AUROC and AP characterize continuous-score ranking, whereas F1 additionally depends on the selected operating point; their seed-level variability therefore need not follow the same pattern.

The error bars should be interpreted as a detector-level reproducibility diagnostic. Small standard deviations indicate that the corresponding reported metric is relatively consistent across the five evaluated seeds, whereas larger standard deviations indicate greater sensitivity to stochastic variation under that experimental condition. Because these experiments report detector-level metrics rather than direct cross-seed similarity of learned embeddings, they do not establish invariance of the learned representation or show that CI-RS itself stabilizes representation geometry.

Summary. The representation-geometry and cross-seed analyses provide complementary diagnostics. The geometry results show that the representation changes associated with CI-DCUAD are heterogeneous across datasets and should not be interpreted as evidence of a single geometric mechanism. The seed analysis establishes complete coverage of the 375-run CI-S1 grid and characterizes the reproducibility of detector-level metrics across the five predefined random seeds.

Together, these analyses provide complementary evidence about learned representation geometry and experimental reproducibility while avoiding two stronger interpretations that the experiments do not establish: detector-level reproducibility does not imply embedding invariance, and post-hoc geometry measurements do not establish a causal mechanism for the effect of CI-RS.

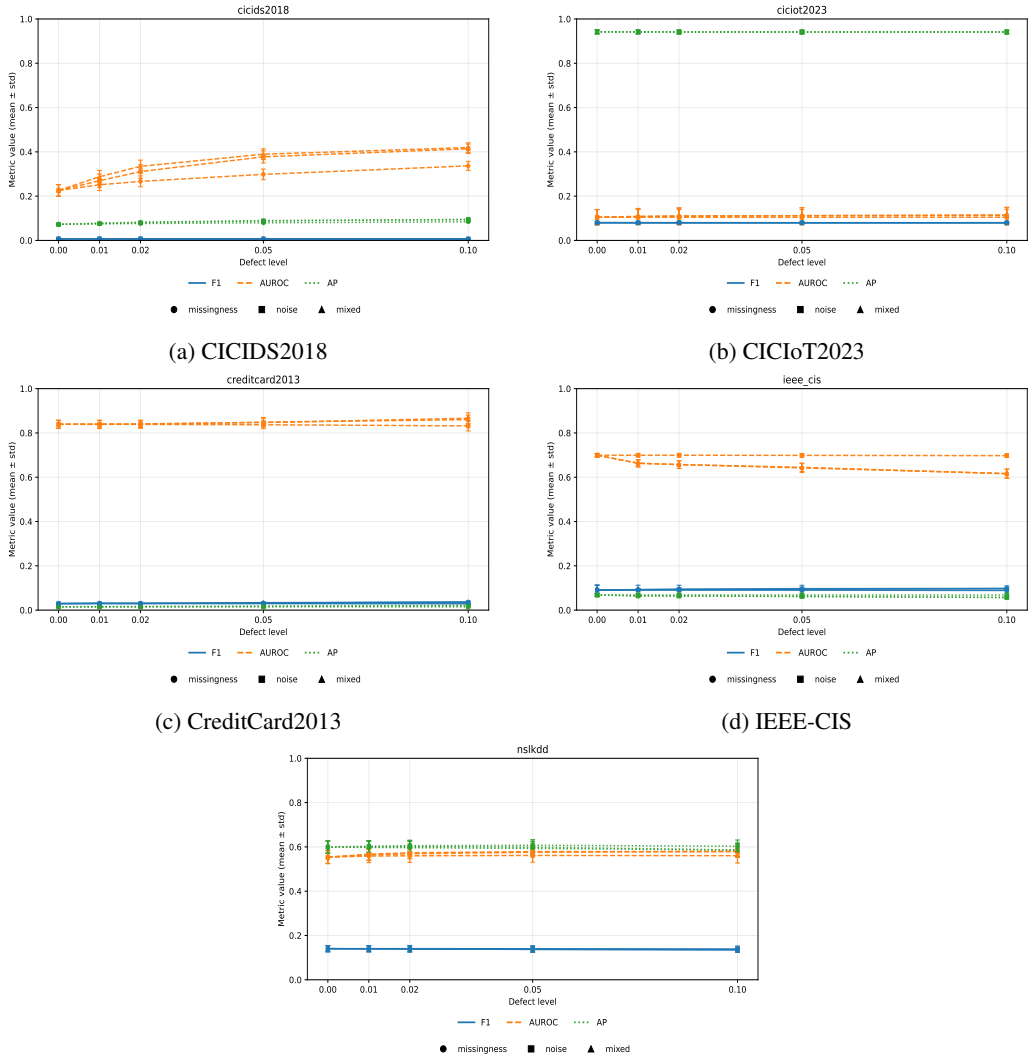
I Implementation, Variants, and Computational Details

This section provides additional implementation and computational details for the evaluated CI-DCUAD pipeline. We first define the matched baseline, CI variants, structural ablations, and post-hoc calibration variants used throughout the experiments. We then summarize empirical runtime measurements for the CI-S1 experimental grid.

These two analyses serve complementary reproducibility purposes. The method map specifies which components change across experimental variants and which components remain matched, while the runtime analysis documents the empirical computational cost observed under the evaluated implementation. Runtime differences should therefore be interpreted in the context of the exact variant definitions rather than as evidence of algorithmic acceleration.

I.1 Datasets and Preprocessing

We evaluate CI-DCUAD on five real-world tabular anomaly-detection benchmarks spanning cybersecurity, IoT monitoring, and financial fraud.



(a) CICIDS2018

(b) CICIoT2023

(c) CreditCard2013

(d) IEEE-CIS

(e) NSL-KDD

Figure 30: Cross-seed detector performance under increasing mixed-defect severity across the five evaluated datasets. Each panel reports mean F1, AUROC, and AP across the five predefined random seeds, with error bars denoting one standard deviation. These plots characterize detector-level reproducibility across seeds rather than invariance of the learned representations.

Table 26: Summary of the evaluated datasets. Test prevalence is reported descriptively from ground-truth labels and is not supplied to CI-DCUAD during training or selection.

Dataset	n_{train}	n_{test}	d	Test prevalence
CICIDS2018	280000	120000	80	0.117983
CICIoT2023	210000	90000	40	0.977656
CreditCard2013	199365	85442	30	0.001810
IEEE-CIS	413378	177162	433	0.034833
NSL-KDD	103962	44555	43	0.478869

CICIDS2018. CSE-CIC-IDS2018 (hereafter CICIDS2018) Canadian Institute for Cybersecurity (CIC), University of New Brunswick (2018) is a network-intrusion benchmark containing benign traffic and multiple attack categories. Its heterogeneous attack structure provides a setting in which anomalous observations need not form a single homogeneous group.

CICIoT2023. CICIoT2023 Neto et al. (2023); Canadian Institute for Cybersecurity (CIC), University of New Brunswick (2023) contains benign and malicious traffic generated in an IoT environment and includes multiple attack classes. It provides a large-scale IoT benchmark with heterogeneous anomalous behavior.

CreditCard2013. CREDITCARD2013 Dal Pozzolo et al. (2018); MLG ULB (2018) is a financial transaction benchmark in which fraudulent transactions constitute a small fraction of the observations. It represents the rare-anomaly setting commonly considered in anomaly detection.

NSL-KDD. NSL-KDD Tavallaee et al. (2009) is a network-intrusion benchmark containing normal connections and multiple attack categories. Its distributional structure differs from the newer network and IoT datasets and provides an informative stress case for the reference-purification mechanism.

IEEE-CIS. IEEE-CIS Kaggle (2019) is a large transaction-fraud benchmark containing high-dimensional transaction and identity attributes with substantial missingness. It provides a financial setting with heterogeneous feature availability.

Preprocessing. All datasets are processed using a fixed preprocessing pipeline. Numerical transformations are fitted using the training partition and then applied to the corresponding evaluation partition. Categorical variables are encoded deterministically according to the dataset-specific preprocessing configuration. The preprocessing artifacts and train–test indices are stored and reused across matched comparisons.

Ground-truth anomaly labels are excluded from preprocessing decisions, representation learning, CI-RS selection, reference-bank construction, anomaly scoring, and threshold calibration. They are used only after model construction for performance evaluation and mechanism-level measurements.

The naturally observed test prevalence reported in Table 26 is distinct from the experimentally controlled training contamination α. Each benchmark is therefore evaluated over the prescribed contamination conditions rather than being assigned a single fixed contamination regime.

I.2 EXPERIMENTAL PROTOCOL AND CONFIGURATION

Data splitting and seeds. For each dataset and random seed, we construct a fixed 70/30 train–test split. The same split is reused across matched method comparisons. The predefined primary seeds are

$$\{7, 13, 21, 37, 51\}.$$

The unlabeled training partition is used for representation learning and reference selection. Labels are withheld from all training, selection, scoring, and threshold-calibration procedures.

Training contamination. The anomalous fraction of the unlabeled training distribution is varied over

$$\alpha \in \{0.01, 0.05, 0.10, 0.25, 0.50, 0.75\}.$$

The true value of α is used only to construct controlled experimental conditions and is not supplied to CI-DCUAD. Conditions with large α are interpreted as stress tests rather than assumptions about typical deployment prevalence.

Structured defects. We evaluate three controlled feature-corruption processes:

1. **Missingness**, in which a controlled fraction of feature observations is removed;
2. **Additive noise**, in which observed feature values are perturbed; and
3. **Mixed defects**, in which missingness and additive noise are applied within the same condition.

Defect severity is varied over

$$\lambda \in \{0, 0.01, 0.02, 0.05, 0.10\}.$$

Training follows the curriculum described in Section 4: mild defect augmentation is used during initialization and CI-controlled warmup, followed by stronger and mixed defects during main-stage learning. Evaluation corruption is generated for the controlled defect condition being tested.

The primary CI-S1 robustness grid contains

$$5 \text{ datasets} \times 5 \text{ seeds} \times 3 \text{ defect types} \times 5 \text{ severity levels} = 375$$

matched configurations.

CI-DCUAD configuration. The encoder produces a 128-dimensional embedding. Unless explicitly varied, the batch size is 256, CI-controlled warmup uses 10 epochs, main-stage training uses 20 epochs, the contrastive temperature is 0.1, and final scoring uses $k = 20$ neighbors. The reference bank is capped at $N_{\text{ref}} = 200{,}000$.

CI-RS uses Isolation Forest as the preliminary unsupervised scoring function on the initial embeddings. The scores are oriented as normality scores, normalized using the median and median absolute deviation, and modeled using a two-component Gaussian mixture. The component with the larger mean normality score is designated as the normal-associated component. Selection uses posterior membership in this component.

All CI-RS parameters are fixed without anomaly labels. The posterior threshold and reference-bank cap are varied only in the corresponding sensitivity experiments rather than tuned independently using test performance.

Matched baseline and structural variants. The primary controlled baseline, DCAD, uses the same mask-aware encoder, defect-contrastive objective, optimization schedule, and kNN scoring mechanism as CI-DCUAD but does not apply CI-based subset selection. The structural variants are:

- `dcad_base`: full-data warmup and full-data reference bank;
- `ci_ref_only`: full-data warmup and CI-selected reference bank;
- `ci_warm_only`: CI-selected warmup and full-data reference bank;
- `ci_full`: CI-selected warmup and CI-selected reference bank.

These variants use the same encoder architecture, contrastive objective, main-stage training schedule, and scoring rule. Their comparison isolates the two pathways through which CI-RS enters the detector.

Table 27: Fixed configurations of the 15 external detectors. Parameters not listed were inherited from the corresponding package implementation. Device selection was automatic for the DeepOD methods.

Method	Implementation	Fixed configuration
IForest	PyOD	$n_{\text{estimators}} = 200$, contamination $= 0.1$, random state $= 7$
LOF	PyOD	$n_{\text{neighbors}} = 35$, contamination $= 0.1$
OC-SVM	PyOD	RBF kernel, $\nu = 0.1$, $\gamma = \texttt{scale}$
HBOS	PyOD	contamination $= 0.1$
PCA	PyOD	contamination $= 0.1$
COPOD	PyOD	package defaults
ECOD	PyOD	contamination $= 0.1$
Autoencoder	PyOD	20 epochs, batch size 256, learning rate 10^{-3}, contamination $= 0.1$
VAE	PyOD	20 epochs, batch size 256, learning rate 10^{-3}, contamination $= 0.1$
Deep SVDD	PyOD	20 epochs, batch size 256, learning rate 10^{-3}, contamination $= 0.1$
DIF	DeepOD	package defaults with seeded initialization
ICL	DeepOD	30 epochs, batch size 128, seeded initialization
SLAD	DeepOD	30 epochs, batch size 128, seeded initialization
LOE	Official LatentOE-AD	Official tabular configuration; unmodified source; 20 epochs (official default: 300); batch size 128; learning rate 10^{-3}; fixed contamination parameter 0.10
SRR-Raw	Paper-based reproduction	Identity representation; $K = 5$ disjoint diagonal-Gaussian models; fixed $\gamma = 0.10$; unanimous normal retention; final diagonal-Gaussian scorer

External detector configurations. External comparisons include Isolation Forest, LOF, OC-SVM, HBOS, PCA, COPOD, ECOD, autoencoder, VAE, Deep SVDD, DIF, ICL, SLAD, LOE, and SRR-Raw. The configurations in Table 27 were fixed before evaluation. We did not perform dataset-specific hyperparameter search, model selection using test labels, post-hoc score-sign selection, or retrospective selection using AUROC, AP, or F1. LOE and SRR-Raw are evaluated only at the common focus condition. Consequently, these comparisons are reproducible fixed-configuration baselines rather than exhaustively tuned estimates of each method's best attainable performance.

Matched comparisons reuse the same data splits and evaluation conditions. The modern external-baseline runs use the same nominal defect conditions and train/test sizes but independently generated stochastic defect realizations; they are therefore interpreted as detector-level context rather than seed-paired component comparisons.

I.3 Mechanism Measurements and Controlled Experiment

Effective contamination and purification gain. Using labels only after training, we measure

$$\alpha_{\text{eff}} = P(Y = 1 \mid C(X) = 1) \tag{48}$$

and define

$$\Delta\alpha = \alpha - \alpha_{\text{eff}}. \tag{49}$$

Positive $\Delta\alpha$ indicates that the selected reference distribution contains less anomalous mass than the original training distribution.

Selection retention. We report the selected fraction

$$\rho_{\text{sel}} = \frac{|I_{\text{ref}}|}{n}. \tag{50}$$

This measurement distinguishes changes in reference composition from changes that may be attributable only to reducing reference-bank size.

Score-space separation. The fitted-component separation is measured using

$$\text{Sep} = \frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}}, \tag{51}$$

where the two components are fitted by CI-RS. This quantity uses no labels and is treated as a diagnostic of the fitted score-space structure rather than a guarantee of latent class separation.

Purification–performance association. For matched CI-DCUAD and DCAD-BASE configurations, define

$$\Delta p = p_{\text{CI}} - p_{\text{DCAD}}, \tag{52}$$

where p_{CI} and p_{DCAD} are the post-hoc measured normal-sample fractions in the corresponding reference sets. The matched fixed-alert F1 difference is

$$\Delta\text{F1} = \text{F1}_{\text{CI-DCUAD}} - \text{F1}_{\text{DCAD-BASE}}. \tag{53}$$

Their association is evaluated using Pearson and Spearman correlation. Because reference purity requires labels, this is a post-hoc mechanism analysis and is not an unlabeled deployment rule. The associations are interpreted diagnostically rather than causally.

Controlled separability–contamination experiment. To isolate reference selection from dataset-specific representation effects, we construct equal-variance Gaussian score distributions. Standardized separation is varied over

$$\delta \in \{0, 0.5, 1, 1.5, 2, 3, 4\},$$

and contamination is varied over

$$\alpha \in \{0.01, 0.05, 0.10, 0.25, 0.50, 0.75\}.$$

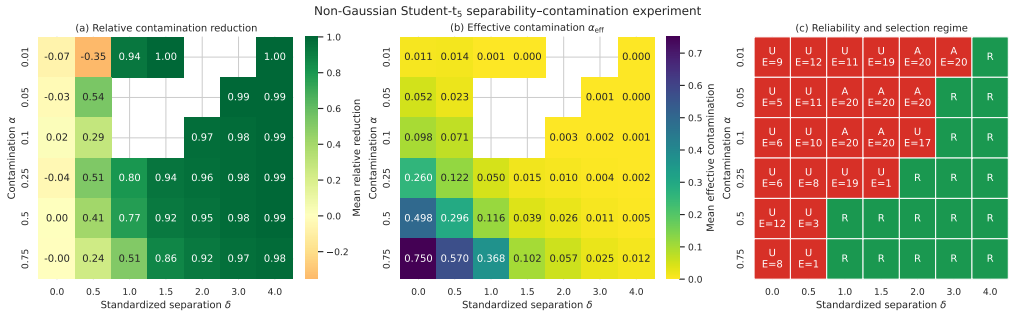


Figure 31: Controlled non-Gaussian stress test using variance-standardized Student-t_5 scores. Panels report mean relative contamination reduction, mean effective contamination, and the reliability/selection regime. "R" denotes reliable conditions, "U" denotes conditions with unstable empty selections, and "A" denotes abstention.

Each of the 42 conditions contains 20,000 observations and is repeated over 20 random seeds, producing 840 runs. CI-RS uses a two-component Gaussian mixture with 10 initializations, at most 500 iterations, and posterior selection threshold $\pi = 0.9$.

For nonempty selections, relative contamination reduction is

$$R_{\text{rel}} = \frac{\alpha - \alpha_{\text{eff}}}{\alpha}. \tag{54}$$

A condition is classified as practically reliable when the lower endpoint of the bootstrap 95% confidence interval for mean R_{rel} exceeds 0.05. Conditions producing no selected observations are recorded as empty-selection failures and are not assigned perfect purification. Confidence intervals use 5,000 bootstrap resamples. Spearman and Kendall tests assess the monotonic association between separation and purification independently at each contamination level.

I.4 Non-Gaussian Synthetic Stress Test

To test whether the controlled mechanism depends on model-matched Gaussian scores, we repeat the full separability–contamination grid using variance-standardized Student-t_5 score distributions. The experiment retains the seven separation values, six contamination levels, 20 seeds, 20,000 observations per condition, two-component Gaussian-mixture selector, and posterior threshold $\pi = 0.9$, producing 840 additional runs. A condition is classified as reliable when the lower endpoint of a 10,000-resample bootstrap 95% confidence interval for mean relative contamination reduction exceeds 0.05. Empty selections are treated as failures rather than perfect purification.

No zero-separation condition is reliable. Among the 36 positive-separation conditions, 18 are reliable, while 17 conditions across the full grid exhibit unstable empty selections and seven produce abstention. Every positive-separation condition that remains nonempty across all 20 seeds satisfies the reliability criterion. Thus, preferential purification survives heavy-tailed misspecification when selection remains operational, but heavy tails enlarge the instability and abstention regions.

Sensitivity and failure analysis. Posterior-threshold and reference-cap sweeps evaluate the trade-off between reference coverage and effective contamination without selecting hyperparameters using test labels. Low-separability and unfavorable configurations are retained rather than excluded so that the conditional robustness claim is tested directly.

I.5 Statistical Procedures and Reproducibility

Evaluation orientation. Throughout evaluation, label $Y = 1$ denotes the anomalous class. Detector outputs are oriented once according to their documented scoring conventions so that larger values represent stronger anomaly evidence; score direction is not selected post hoc using test AUROC. For CI-DCUAD and DCAD, the final score is the kNN distance in equation 5, for which larger distance indicates stronger anomaly evidence. LOE uses the official anomaly-score output without score negation. For SRR-Raw, larger diagonal-Gaussian energy represents stronger anomaly evidence. No score direction for either method was selected using test labels or post-hoc AUROC.

Matched detector comparisons. Comparisons between CI-DCUAD and matched alternatives use observations from the same dataset, defect type, severity, and seed. Principal full-grid comparisons report paired mean differences, standard deviations, Cohen's d_z, two-sided paired t-tests, and Benjamini–Hochberg correction within each predefined metric family.

Structural mechanism ablation. For structural ablations, the five matched seeds are averaged within each dataset–defect–severity condition, yielding 75 condition-level analysis units. Mean paired F1 effects are accompanied by 10,000-resample bootstrap 95% confidence intervals. The four structural variants induce six possible pairwise contrasts. All six contrasts are prespecified, tested using two-sided Wilcoxon signed-rank tests, and corrected jointly using Holm's procedure within each overall or dataset-specific analysis scope.

Controlled phase diagram. The controlled experiment uses bootstrap confidence intervals across its 20 seeds per condition and rank-based separation-trend tests. Statistical significance is interpreted together with effect magnitude, direction, dataset heterogeneity, and empty-selection behavior rather than as evidence of universal superiority.

Reproducibility controls. For each run, we retain the dataset split, contamination configuration, defect type and severity, random seed, CI-RS selection indices, embeddings, and evaluation metrics. CI-RS selection is recomputed independently for each applicable dataset–seed

Table 28: Detailed configuration of the evaluated method variants and thresholding mechanisms. Mass denotes fixed alert-mass calibration using alert fraction β; POT denotes Peaks-Over-Threshold calibration using a Generalized Pareto tail model; and RQ denotes rolling empirical quantile calibration. CI denotes CI-RS-based contamination-aware reference selection.

Method	Warm set	Ref. set	Training	Scoring	Thresholding
`dcad_base`	`warm_all` (all train)	`ref_full` (all train)	Baseline-warm encoder	kNN scores vs. full reference bank	Mass (top-β), POT (GPD tail), RQ (rolling quantile)
`ci_ref_only`	`warm_all` (all train)	`ref_idx_ci` (CI reference)	Baseline-warm encoder	kNN scores vs. CI reference bank	Mass (top-β), POT (GPD tail), RQ (rolling quantile)
`ci_warm_only`	`warm_idx_ci` (CI warm)	`ref_full` (all train)	CI-warm encoder	kNN scores vs. full reference bank	Mass (top-β), POT (GPD tail), RQ (rolling quantile)
`ci_full`	`warm_idx_ci` (CI warm)	`ref_idx_ci` (CI reference)	CI-warm encoder	kNN scores vs. CI reference bank	Mass (top-β), POT (GPD tail), RQ (rolling quantile)

configuration rather than reused across seeds. The LOE and SRR-Raw focus-condition comparisons each contain exactly 25 completed runs: five datasets multiplied by five predefined seeds. For every run, the saved score vector was checked for finite values and exact agreement with the corresponding test-set size.

Per-seed outputs and aggregate tables are generated programmatically from the stored experimental results. Ground-truth labels are excluded from representation learning, CI-RS selection, reference-bank construction, anomaly scoring, and threshold calibration. They are accessed only for post-hoc performance evaluation, effective-contamination measurement, and mechanism diagnostics.

I.6 Detailed Method Variant Definitions

The evaluated method variants are designed to separate the roles of CI-controlled warmup, CI-selected reference construction, and post-hoc threshold calibration while otherwise preserving the matched experimental pipeline.

Across the representation-learning variants, the encoder architecture, optimization settings, main-training schedule, final embedding dimension, and reference-based anomaly-scoring rule are held fixed unless explicitly stated otherwise. Final anomaly scores are computed using kNN distance to the corresponding reference bank.

Matched baseline and full CI-DCUAD. The matched baseline and CI variants differ in how observations are used during warmup and how the final reference bank is constructed. In `dcad_base`, warmup uses Isolation Forest ranking on the raw training features, while the final reference bank is constructed from the full training set subject to the common reference-size cap.

In `ci_full`, CI-RS-selected observations are used during the CI-controlled warmup, and the CI-selected indices determine which training observations contribute embeddings to the final reference bank. Importantly, subsequent main representation learning returns to the full unlabeled training set. CI-RS therefore does not restrict all representation-learning epochs to the selected subset.

Structural ablations. The warmup-only and reference-only variants modify these two CI pathways separately. They are used to examine whether detector-level changes are associated with CI-controlled warmup, CI-selected reference construction, or their combination.

These variants should be interpreted as structural ablations rather than as a causal decomposition of representation and reference effects. The two pathways occur within the same learning and scoring pipeline and can interact through the final representation and reference-based anomaly scores.

Post-hoc calibration variants. Thresholding is applied after anomaly-score construction. The evaluated calibration strategies are:

- **Mass**: fixed alert-mass calibration using the predefined alert fraction β;
- **POT**: Peaks-Over-Threshold calibration using a Generalized Pareto tail model; and
- **RQ**: rolling empirical quantile calibration.

For the POT support audit, fitting is attempted only when the calibration/reference score count satisfies $N_{\mathrm{ref}} \geq 100$ and at least 10 valid threshold exceedances are available. Otherwise, the POT result is recorded as invalid.

These calibration variants operate on the resulting anomaly scores and do not alter the learned encoder, CI-RS selection, or final reference bank. They are included to evaluate whether operating-point results are specific to the fixed alert-mass decision rule.

Table 28 provides the corresponding configuration map. The structural ablations should be read relative to the full CI-DCUAD pipeline. CI-controlled warmup modifies the early representation-learning stage, whereas CI-selected reference construction determines which selected

Table 29: Empirical CI-S1 training runtime by dataset. Runtime values are reported as mean ± standard deviation across the five predefined random seeds. Within each dataset, the recorded runtime summary is identical across the evaluated defect types and defect levels; therefore, the repeated condition-level rows are collapsed into a single dataset-level entry.

Dataset	Runtime (s)	n_{seeds}
CICIDS2018	363.877 ± 104.602	5
CICIoT2023	500.509 ± 102.197	5
CreditCard2013	502.588 ± 114.461	5
IEEE-CIS	732.669 ± 109.101	5
NSL-KDD	244.257 ± 55.225	5

training indices contribute embeddings to the final reference bank. Because main training subsequently uses the full unlabeled training set, CI-based selection should not be interpreted as purification of the complete representation-learning distribution.

Similarly, the Mass, POT, and RQ comparisons isolate the post-hoc decision-calibration stage. If a relative detector difference has a similar direction across several thresholding rules, this provides evidence that the difference is not solely an artifact of the fixed alert-mass operating rule. Such agreement does not establish general threshold invariance and does not imply that calibration protects or improves the learned representation, reference bank, or continuous anomaly-score ranking.

I.7 Additional Runtime Summaries

This subsection provides additional empirical runtime measurements that complement the computational-cost analysis in Appendix I.8. The measurements are wall-clock observations under the evaluated implementation and computational environment and complement, rather than replace, the asymptotic complexity analysis reported there.

The recorded CI-S1 runtime summaries are identical across defect types and defect levels within each dataset. We therefore report one dataset-level summary rather than repeating the same values for all 15 corruption conditions.

Interpretation. Table 29 reports empirical wall-clock runtime rather than a theoretical complexity guarantee. Runtime differs substantially across datasets, reflecting differences in dataset scale together with implementation- and system-level characteristics.

CI-DCUAD introduces additional CI-RS operations relative to the matched learning pipeline, including preliminary scoring, robust score normalization, two-component GMM fitting, posterior-based selection, and construction of the selected reference bank. These operations constitute additional computational work associated with contamination-aware reference selection.

At the same time, CI-controlled warmup can operate on a selected subset, whereas subsequent main representation learning returns to the full unlabeled training set. Consequently, measured wall-clock differences between CI-DCUAD and another variant should not be interpreted as direct evidence of lower asymptotic complexity or intrinsic algorithmic acceleration.

The runtime measurements should therefore be interpreted together with Appendix I.8, which describes how computational cost depends on the relevant pipeline components, while Table 29 documents the empirical runtime observed for the evaluated CI-S1 implementation.

I.8 Computational Cost and Practicality

CI-DCUAD uses the same encoder architecture and nominal training schedule as the matched DCAD baseline. Its additional computation arises primarily from CI-RS: preliminary Isolation Forest scoring, robust normalization of the one-dimensional score distribution, two-component Gaussian mixture modeling, and posterior-based reference selection.

Training and selection complexity. Let n denote the number of training observations, C_{enc} the per-sample encoder cost, and E_w and E_m the warmup and main epoch counts. The encoder-training cost is approximately

$$\mathcal{O}(E_w n_w C_{\text{enc}} + E_m n C_{\text{enc}}) ,$$

where n_w is the number of observations used during warmup. In our experiments, $E_w = 10$ and $E_m = 20$. The main representation-learning stage uses the full unlabeled training set; CI-selected observations affect the CI-controlled warmup and the construction of the final reference bank. Thus, CI-DCUAD neither expands the encoder architecture nor increases the nominal number of training epochs relative to the matched baseline.

For CI-RS, Isolation Forest fitting and scoring can be characterized approximately as

$$\mathcal{O}(Tn \log \psi) ,$$

for T trees and subsample size ψ. Robust score normalization, two-component one-dimensional GMM fitting, posterior computation, and threshold-based selection are linear in n for fixed mixture size, apart from the number of EM iterations. CI-RS therefore adds a preliminary scoring stage and lightweight one-dimensional score-space processing without changing the asymptotic multi-epoch encoder-training structure.

Reference-bank memory and scoring. The final reference bank is capped at

$$N_{\text{ref}} \leq 200{,}000.$$

With embedding dimension $p = 128$ and 32-bit floating-point storage, the maximum reference-embedding array requires approximately

$$200{,}000 \times 128 \times 4 = 102.4\,\text{MB}$$

(≈ 97.7 MiB), excluding nearest-neighbor and other implementation overhead.

Final anomaly scoring uses exact brute-force kNN with $k = 20$. For m evaluation observations, reference size N_{ref}, and embedding dimension p, direct scoring has complexity

$$\mathcal{O}(mN_{\mathrm{ref}}p)\,.$$

Hence, although the reference cap bounds memory and reference-set size, exact nearest-neighbor scoring can remain computationally important when the bank approaches its maximum size. Approximate nearest-neighbor methods could reduce this cost at larger scale, but they are not evaluated here.

Observed runtime. Table 29 summarizes empirical wall-clock runtime for the evaluated CI-S1 implementation across the five datasets. These measurements document the computational scale of the experiments but are not used to claim a matched runtime advantage over the baseline.

Runtime varies substantially across datasets because of differences in dataset size, selected warmup size, reference-bank construction, and implementation behavior. On the largest evaluated dataset, IEEE-CIS, the complete CI-S1 implementation requires approximately 733 seconds on average under the evaluated hardware and implementation. This value includes training, CI-RS, embedding extraction, and evaluation and should be interpreted as an implementation-specific empirical measurement rather than a hardware-independent complexity estimate.

Overall, CI-RS introduces contamination-aware selection without increasing network depth, width, or the nominal number of training epochs. The empirical runtime measurements in Table 29 document the computational scale of the evaluated implementation, while the reference cap bounds embedding storage. The principal scaling concern is exact kNN scoring as the reference bank and evaluation set grow.

Summary. The method-variant definitions establish the structural differences required to interpret the ablation and calibration experiments. In particular, CI-selected observations affect the CI-controlled warmup and final reference construction, while subsequent main training uses the full unlabeled training set. Mass, POT, and RQ act only at the post-hoc decision stage.

The runtime analysis complements these definitions by documenting empirical computational cost for the evaluated CI-S1 implementation. These measurements support reproducibility of the computational analysis but should not be interpreted as evidence of asymptotic acceleration or as a general runtime guarantee.

J Deployment Considerations

CI-DCUAD is motivated by unsupervised settings in which anomaly prevalence is unknown and data quality may change over time. We therefore distinguish the evaluated deployment-facing components of the pipeline from possible monitoring and updating procedures that remain outside the experimental scope.

Alert-rate–controlled thresholding. Threshold-dependent anomaly detection requires an operating point for converting continuous anomaly scores into alerts. For a user-specified alert mass $\beta \in (0, 1)$, CI-DCUAD applies the empirical quantile rule

$$t_\beta = \mathrm{Quantile}_{1-\beta}\left(\{s_i\}_{i=1}^{n}\right), \qquad \widehat{y}_i = \mathbb{1}[s_i > t_\beta]. \tag{55}$$

The primary experiments use $\beta = 0.05$. The realized alert fraction on the score set to which this rule is applied is therefore approximately 0.05, up to score ties and the empirical-quantile convention. Operationally, β represents an externally specified alert budget determined by available review or downstream-processing capacity; it is not an estimate of the anomaly prevalence.

This construction does not require anomaly labels or an estimate of the contamination proportion. For reproducibility, the alert mass, quantile convention, score-set size, score orientation, and resulting threshold should be recorded together with the model and reference-bank versions.

Alert-rate control should not be confused with detection quality. It standardizes the binary operating point but does not guarantee precision, recall, F1, AP, or AUROC. Because thresholding occurs after representation learning and reference-bank construction, it cannot repair contamination-induced degradation in either component.

Operating under unknown contamination. CI-DCUAD combines contamination-aware reference selection, a capped reference bank ($N_{\mathrm{ref}} \leq 200{,}000$), and label-free alert-rate calibration. These components address different operational concerns: CI-RS attempts to construct a more useful reference set, the cap bounds reference storage and exact-scoring workload, and alert-rate calibration controls the operational alert allocation.

The stress experiments in Appendix D show that this behavior remains condition dependent. When the preliminary score structure supports preferential selection, CI-RS can retain a useful lower-contamination reference set; when that structure deteriorates, effective contamination can rise, purification gain can contract, and downstream performance can degrade (Section 7). CI-RS should therefore not be interpreted as providing immunity to unknown or shifting contamination.

Monitoring and updating. A deployed system could monitor unlabeled quantities already produced by the pipeline, including score-distribution summaries, selected-set size and posterior behavior, reference-set characteristics, embedding summaries, and realized alert volume. Distributional changes between historical and current windows could additionally be summarized using statistics such as the Kolmogorov–Smirnov statistic or Wasserstein distance.

Such diagnostics indicate distributional change, not anomaly-detection quality. A score shift can result from changing normal behavior, anomaly prevalence, data defects, or other distribution shifts, while harmful changes need not appear strongly in every monitored statistic. Trigger thresholds and response policies are therefore application dependent.

Because CI-DCUAD does not require anomaly labels for representation learning, CI-RS selection, reference construction, or alert-rate calibration, a deployment could periodically rerun selection, reconstruct the reference bank, and recalibrate the operating threshold using newly collected unlabeled data. Retraining frequency, historical-window design, drift triggers, and update policies are not optimized in this study.

Importantly, the present experiments evaluate controlled contamination and defect shifts; they do *not* evaluate an online drift-detection, automatic-retraining, or continual-learning system. The monitoring and updating procedures above are therefore deployment considerations rather than additional empirically validated mechanisms.

K SCOPE BEYOND TABULAR DATA

CI-DCUAD is developed and empirically evaluated only on tabular datasets in this work. The reported results therefore support conclusions only within the evaluated tabular setting; they do not establish generalization to images, sequences, graphs, or other data modalities.

Modality-independent pipeline components. At an algorithmic level, several components of the pipeline are not inherently tied to raw tabular inputs:

- contamination-aware reference selection from a preliminary normality-score distribution;
- reference-based anomaly scoring in a learned representation space;
- post-hoc alert-mass calibration of the resulting anomaly scores.

These components are defined at the level of representations or anomaly scores rather than raw tabular features. Applying them to another modality would nevertheless require modality-appropriate representations and preliminary scores, and their effectiveness outside the evaluated tabular setting remains untested.

Modality-specific components. Transfer to another modality would require redesign and validation of components that depend directly on the input representation, including:

- the encoder architecture and representation-learning procedure;
- corruption or augmentation mechanisms used to construct training views;
- representation-space distance or similarity functions.

Changing these components would also change the representation and may alter the preliminary score structure on which CI-RS depends, as well as the geometry of the resulting reference-based scoring problem.

Implication. The present experiments provide no evidence that CI-DCUAD preserves its behavior outside tabular data. Applying the framework to another modality would require a modality-appropriate representation-learning pipeline and would still depend on the conditions underlying CI-RS: informative preliminary score structure, preferential retention of normal observations, and adequate preservation of normal support.

Evaluation on images, sequences, graphs, and other modalities is therefore a direction for future work rather than a generalization claim supported by the current experiments.