WHERE DOES DETECTION FAIL? A RESIDUAL-EVIDENCE DIAGNOSTIC WORKFLOW FOR INDUSTRIAL CONTROL SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Residual-based anomaly detectors for industrial control systems (ICS) convert model residuals into anomaly scores and alarms, but final detection metrics give little guidance on which component to examine when detection performance is poor. We present a retrospective diagnostic workflow built around covariance-normalized residual energy (CNRE), a measurement based on the squared Mahalanobis distance. The workflow asks whether the normal reference still describes the evaluation recording, whether attacks separate from normal operation under a stated reference, and how scoring and alarm rules use fixed residuals. Across fifteen detector–dataset pairs (five detectors on SWaT, WADI, and HAI), reference updates improve separation in some pairs but degrade it in others, and a retrospective policy that updates only where a label-free screen recommends it raises mean AUC from 0.751 (never updating) and 0.759 (always updating) to 0.809; on the twelve non-GDN pairs, which did not inform its boundaries, it gives 0.805 versus 0.759 and 0.747. Rescoring CNRE computed with a fixed reference reproduces most of the AUC gain of a reference update on WADI but not on SWaT. A detailed GDN case study connects the diagnostics: normal-only adaptation reduces normal prediction error by 27.3%, but the direction of its effect on attack separation depends on the reference protocol. With residuals fixed, stronger score ranking does not ensure that normal-calibrated alarms preserve their nominal false-positive rate. These results support choosing intervention stages and interpreting their outcomes through separate measurements of reference fit, attack evidence, and alarm calibration.

1 INTRODUCTION

Industrial control systems (ICS) rely on anomaly detectors to identify abnormal behavior in physical processes such as water treatment, manufacturing, and energy systems (Stouffer et al., 2023). Many detectors learn expected behavior, compare it with observations, and convert the resulting deviation into anomaly scores and alarms. When an attack is missed, the final detection result reveals little about which stage limited detection. The normal reference may no longer describe current operation, the deviations may separate attacks weakly, or downstream scoring and alarm decisions may use the available evidence poorly.

Component-level and benchmark studies show that model, scoring, thresholding, and protocol choices can change reported performance substantially (Garg et al., 2022; Kim et al., 2022). We ask a complementary question: which controlled comparisons distinguish a mismatched normal reference, a change in attack–normal separability, and a limitation in downstream scoring or alarm-threshold calibration. We present a retrospective diagnostic workflow built from such comparisons. The workflow identifies comparisons that can explain intervention outcomes; it does not assign a unique cause to each missed alarm.

The workflow applies to detectors that expose a *residual* before the alarm, such as prediction errors, reconstruction errors, innovations, and constraint violations (Deng & Hooi, 2021; Tuli et al., 2022; Feng & Tian, 2021; Wolsing et al., 2025). We measure residuals with *covariance-normalized residual energy* (CNRE), which accounts for the normal scale and covariance of each detector's residuals.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

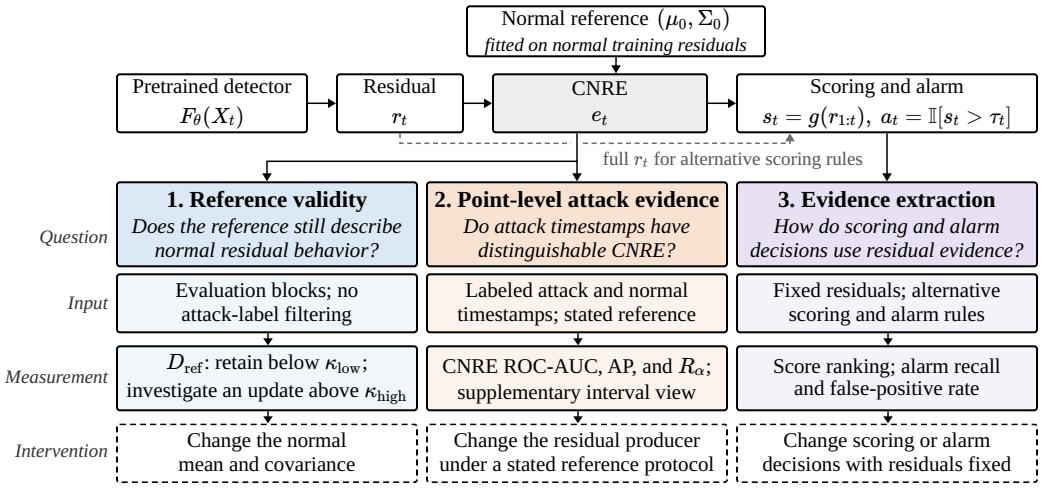

Figure 1: Residual-evidence diagnostic workflow. Diagnostics are examined in the numbered order, and the reference screen is repeated after a model or reference-protocol change. Dashed boxes are candidate interventions, each evaluated by its own controlled comparison.

Its quadratic form is the established Mahalanobis construction (Mahalanobis, 1936; Qin, 2003); we use it as a common measurement interface rather than as a new statistic. CNRE supports three diagnostic questions (Figure 1): whether the reference still describes normal residual behavior (*reference validity*), whether attack timestamps have distinguishable CNRE (*point-level attack evidence*), and how scoring and alarm decisions use residual evidence (*evidence extraction*).

We evaluate the workflow across five detectors on SWaT, WADI, and HAI (Mathur & Tippenhauer, 2016; Ahmed et al., 2017; Shin et al., 2021). Updating never helps a pair that the label-free reference screen retains, and updating only where the screen recommends it raises mean AUC from 0.751 (never updating) and 0.759 (always updating) to 0.809. Read after the reference decision, attack evidence separates pairs that the update recovers from pairs that stay weak under a suitable reference. Rescoring reproduces most of the AUC gain of a reference update on WADI but not on SWaT. A controlled case study on one detector shows that better normal prediction does not determine attack separation, that the sign of a model change depends on its reference protocol (CNRE average precision 0.573 falls to 0.452 or rises to 0.593), and that stronger ranking does not ensure calibrated alarms.

Contributions.

- **Workflow.** CNRE connects reference validity, point-level attack evidence, and evidence extraction, each with its own measurement and controlled comparison.
- **Cross-detector evidence.** On fifteen detector–dataset pairs, a label-free screen indicates where a reference update can help, a threshold-only control separates alarm failures that a threshold fixes from those that also need a reference update, and rescoring reproduces most of the update's AUC gain on WADI but not on SWaT.
- **Case study.** Reference fitting can reverse the measured effect of a model intervention, and stronger ranking does not ensure calibrated alarms.

2 Related Work

2.1 ICS Anomaly Detection

ICS anomaly detection includes process-aware monitoring and data-driven modeling. Process-aware methods examine whether observations agree with expected operating states or process behavior. State-aware detection incorporates system state into monitoring (Ghaeini et al., 2018), while GeCo automatically derives process models from benign operation (Wolsing et al., 2025).

Data-driven detectors learn temporal and cross-variable relationships: GDN predicts with a learned sensor graph (Deng & Hooi, 2021), FuSAGNet combines sparse latent graphs with reconstruction and forecasting (Han & Woo, 2022), TranAD reconstructs with Transformers (Tuli et al., 2022), and NSIBF combines neural system identification with Bayesian filtering (Feng & Tian, 2021). Together with GeCo, their different residuals test whether one measurement can organize comparisons across detector families.

2.2 Analysis and Evaluation of Anomaly Detectors

Component-level studies show that detection performance depends on more than the residual-producing model. Garg et al. (2022) separate the model, scoring function, and thresholding function, and evaluate alternative combinations. In their experiments, scoring choices strongly affect performance, and adaptive scoring can lose sensitivity when it incorporates anomalous behavior. Their anomaly-diagnosis task identifies affected channels, whereas our diagnostic target is the detector pipeline. We build on component-level evaluation by asking which controlled comparisons distinguish reference mismatch, weak attack–normal separability, and scoring or calibration limitations, with CNRE as the common measurement across these questions.

Benchmark studies further show that evaluation choices can obscure what a detector has learned. Wu & Keogh (2023) identify problems in widely used time-series anomaly-detection benchmarks. Kim et al. (2022) show that point adjustment can inflate reported performance and distort comparisons. More recently, TAB provides unified benchmarking and examines inconsistencies in evaluation pipelines (Qiu et al., 2025). These studies motivate explicit reporting of data access, temporal alignment, scoring, and threshold selection. We do not apply point adjustment, and we distinguish retrospective ROC operating points from alarms calibrated on normal data. Supplementary event-level measurements describe variation across attack intervals; the main diagnostic analysis remains point-level.

2.3 Residual Statistics and Reference Adaptation

Covariance-normalized residual statistics have established foundations in multivariate analysis and process monitoring. Mahalanobis distance measures deviation relative to a reference mean and covariance (Mahalanobis, 1936). Statistical process monitoring uses related measures to detect departures from normal operation (Qin, 2003). Shrinkage covariance estimation provides a practical way to stabilize these calculations (Ledoit & Wolf, 2004). CNRE uses this established statistical construction as the measurement core of our workflow.

The usefulness of such a measurement depends on its reference. Research on concept drift studies how changing data distributions affect learned models and motivate adaptation (Gama et al., 2014). For anomaly detection, updating the reference also creates a specific concern: abnormal behavior may enter the data used to define normality. Streaming threshold methods such as SPOT and DSPOT face the same trade-off: they fit extreme-value thresholds online and exclude flagged values from the update (Siffer et al., 2017). Our reference experiments examine this tension while keeping the residual-producing detector fixed. They connect persistent CNRE elevation to observed update outcomes, including cases where an update that pauses refitting during elevated alarm activity still loses attack separation.

3 Problem Setup

Detection pipeline. We study pretrained ICS anomaly detectors that expose a residual before producing an alarm. At time t, we write

$$\hat{y}_t = F_\theta(X_t), \qquad r_t = \psi(y_t, \hat{y}_t) \in \mathbb{R}^p. \tag{1}$$

Here, X_t is the detector input, which may include a sequence of observations. The model F_θ, with pretrained parameters θ, produces the prediction or reconstruction $\hat{y}_t$ of the corresponding observation y_t. The function ψ computes the detector's residual, and p is its dimension; for a prediction error, $\psi(y_t, \hat{y}_t) = y_t - \hat{y}_t$. We also use the term residual for innovations and constraint violations. For diagnostic analysis, we fit a separate normal reference (μ_0, Σ_0) for each detector,

the mean and covariance of its normal training residuals (Section 4.1). Downstream, an *anomaly scoring rule* and an *alarm decision rule* convert residuals into alarms:

$$s_t = g(r_{1:t}; \phi), \qquad a_t = \mathbb{I}[s_t > \tau_t]. \tag{2}$$

The scoring rule g maps the residual history $r_{1:t}$ to a scalar score s_t using fitted parameters ϕ, such as normalization statistics. The alarm decision compares that score with a threshold τ_t, which may be fixed or updated from past data; $a_t = 1$ denotes an alarm. Detection can therefore be limited at three stages: the normal reference (μ_0, Σ_0), the residual-producing model F_θ, and the scoring and alarm rules (g, τ_t).

Data access and scope. References and fixed thresholds are fitted on normal data, and the reference diagnostic reads the evaluation recording without labels; the other two diagnostics use benchmark labels. All diagnostics are *retrospective*, computed after a recording is complete, and *point-level*, with the timestamp as the unit of evaluation.

4 Residual-Evidence Diagnostic Workflow

The workflow uses CNRE to address the three diagnostic questions of Figure 1, each with its own data, measurement, and controlled comparison. We examine the reference first, because a mismatched reference distorts both attack–normal separability and normal-calibrated alarms. We then measure attack evidence under a stated reference and finally examine scoring and alarms with the residuals fixed.

4.1 CNRE as a Measurement Interface

Residuals need not be zero during normal operation, and their scales and correlations differ across variables because of prediction error, sensor noise, and normal process variation. We therefore estimate the reference mean $\mu_0 \in \mathbb{R}^p$ and covariance $\Sigma_0 \in \mathbb{R}^{p \times p}$ from normal training residuals, using Ledoit–Wolf shrinkage for the covariance (Ledoit & Wolf, 2004), and define

$$e_t = \frac{1}{2}(r_t - \mu_0)^\top \Sigma_0^{-1}(r_t - \mu_0). \tag{3}$$

Centering removes the normal offset of r_t, and covariance normalization accounts for normal residual scales and correlations between variables. The resulting energy is half the squared Mahalanobis distance (Mahalanobis, 1936). A deviation in a normally stable residual pattern therefore contributes more than a comparable deviation in a pattern that already varies during normal operation.

An attack that disrupts the detector's learned relationships can raise CNRE above normal levels; an attack with little effect on the residuals may not. Normal operation can also produce high CNRE if the reference becomes unsuitable. When the reference equals the population normal mean and covariance, the expected energy is $p/2$. We use this value as a nominal scale for comparing residual dimensions, not as an alarm threshold. Each detector keeps its own reference, so raw energies are not compared across detectors; only the $p/2$-normalized statistic of Eq. (4) shares decision boundaries, and Section 5.2 tests that transfer.

4.2 Diagnostic Questions and Controlled Comparisons

Reference validity. Training data and later operation can differ as equipment conditions, control settings, or operating regimes change (Gama et al., 2014). A fixed reference may then assign high energy to normal behavior. To screen for persistent elevation without attack labels, we divide the evaluation recording into B consecutive blocks $\mathcal{B}_1, \ldots, \mathcal{B}_B$ and compute

$$D_{\mathrm{ref}} = \mathrm{median}_{b=1,\ldots,B} \left[\frac{\mathrm{median}_{t \in \mathcal{B}_b}\, e_t}{p/2} \right]. \tag{4}$$

The inner median limits the influence of isolated energy spikes, and the outer median summarizes the recording's typical block. Division by $p/2$ puts different residual dimensions on a common nominal scale. Because this compares medians with a nominal mean, a value of one is not a calibrated validity boundary.

D_{ref} is a screening statistic for persistent energy elevation, not a test of reference validity. The decision must account for two risks: retaining a stale reference and updating a suitable reference with contaminated data. We therefore use two empirical boundaries $\kappa_{\mathrm{low}} < \kappa_{\mathrm{high}}$: retain the reference when $D_{\mathrm{ref}} < \kappa_{\mathrm{low}}$, investigate an update when $D_{\mathrm{ref}} > \kappa_{\mathrm{high}}$, and make no recommendation between them, where the screen is inconclusive. The boundaries are set on designated development cases and held fixed for every subsequent evaluation; Section 5.2 reports their values and provenance. They define empirical regions, not significance thresholds. Because the statistic keeps attacks in the evaluation blocks, sustained attacks can also raise it. An update recommendation motivates a controlled reference comparison, not a conclusion that benign drift caused the elevation. The screen applies to one model and its reference, so it should be repeated after a model intervention or a change in the reference-fitting protocol.

Point-level attack evidence. Even under a suitable reference, attacks may not produce residuals that CNRE separates from normal operation. Let $\mathcal{A}$ and $\mathcal{N}$ contain the eligible attack and normal timestamps. We measure CNRE-based attack–normal separability by point-level ROC-AUC:

$$S_{\mathrm{atk}} = \Pr(e_i > e_j) + \tfrac{1}{2}\Pr(e_i = e_j), \qquad i \sim \mathcal{A}, \quad j \sim \mathcal{N}. \tag{5}$$

Here i and j are drawn uniformly and independently from their timestamp sets, and ties receive half credit (Fawcett, 2006). We also examine average precision (AP) and recall R_α at a specified false-positive rate (FPR) α, read retrospectively from the labeled evaluation ROC curve; the experiments use $\alpha = 1\%$ (Table 5 in Appendix A). These measurements complement overall AUC with precision–recall behavior and detection at a restricted false-positive rate.

Weak separation concerns scalar CNRE under the stated reference; it does not show that the residual vector contains no attack information, and it does not by itself localize a limitation to the residual producer. A model intervention tests whether a controlled change to the residual producer improves separation under a stated reference-fitting protocol. A supplementary interval view (Appendix C) replaces $\mathcal{A}$ with one contiguous labeled attack interval against a common normal pool to describe heterogeneity within the pooled result.

Evidence extraction. The third diagnostic holds the residuals fixed and examines the scoring and alarm rules in Eq. (2). Setting $s_t = e_t$ gives a baseline scoring rule, which we compare with other rules on the same residuals and timestamps; alternative rules may use the full residual vector or a history of CNRE values, and each comparison states its input and fitted parameters. Score comparisons assess attack–normal ranking; alarm comparisons assess recall and false positives under a stated calibration policy.

Controlled comparisons. Each intervention is evaluated by a comparison that states what is held fixed: reference interventions keep the residuals and change (μ_0, Σ_0); model interventions change the residual producer and state whether each model keeps the original reference or refits its own; scoring interventions keep the residuals and reference and change the score or alarm rule, stating any extra fitting data. No comparison assumes that a recommended intervention must succeed.

5 Experimental Evaluation

We first examine the three diagnostics across detector families and datasets, closing each with the conclusion the diagnostic supports. We then use one detector (GDN) to illustrate how the workflow interprets model and scoring interventions.

5.1 Experimental Setup and Measurements

The cross-detector study covers GDN, FuSAGNet, TranAD, NSIBF, and GeCo on SWaT, WADI, and HAI, giving fifteen detector–dataset pairs (Deng & Hooi, 2021; Han & Woo, 2022; Tuli et al., 2022; Feng & Tian, 2021; Wolsing et al., 2025). We use archived residuals with each detector's own preprocessing and temporal alignment (for TranAD, its v3 residual archive), so this is a diagnostic study of existing outputs, not a retraining comparison. Initial CNRE references use normal training residuals, and every intervention is compared with its baseline on the same eligible timestamps, those scored after the 13-hour warm-up that adaptive references need. Three evaluation supports appear:

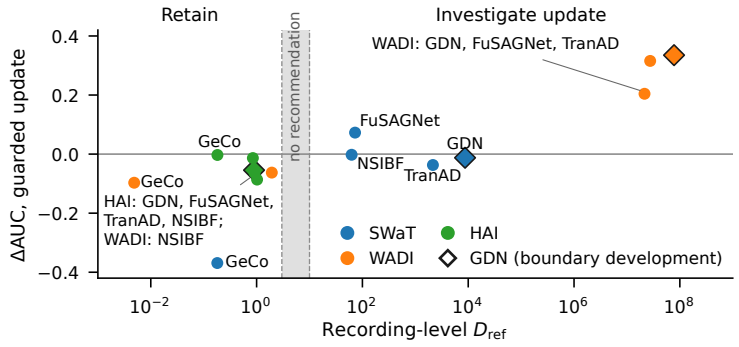

Figure 2: Reference diagnosis and guarded-update effects across the fifteen detector–dataset pairs, on all eligible timestamps after warm-up. Colors mark datasets and diamonds the GDN cases used to develop the boundaries; the other points test transfer across detector families on the same datasets. No pair lies in the shaded 3–10 no-recommendation region.

these post-warm-up timestamps with all attacks retained (reference policies, Table 1, and rescoring), the full archived test stream (Appendix C), and the case study's own timestamps (Section 5.5), whose separate normal fitting splits mean that its absolute scores need not match the cross-detector study. Appendix A defines the measurements (Table 5) and data access; $R_{1\%}$ is the retrospective recall at 1% test FPR.

5.2 Reference Validity Across Detectors and Datasets

Normal residual behavior changes across operating periods. On SWaT, the training-normal 99th-percentile CNRE threshold flags 78.3–80.3% of eligible normal test timestamps for GDN, FuSAGNet, TranAD, and NSIBF, but 1.3% for GeCo (Table 8). The size of this shift depends on the detector even on one dataset, and the measurement does not identify its physical cause.

Reference diagnosis and intervention outcomes. We compute D_{ref} using one-hour blocks. The three GDN cases informed $\kappa_{\text{low}} = 3$ and $\kappa_{\text{high}} = 10$; these boundaries remain fixed for the other twelve pairs, and no pair lies between them; any boundaries with $1.94 < \kappa_{\text{low}} \leq \kappa_{\text{high}} < 62.5$ give the same fifteen decisions. The screen varies within a dataset (Table 1): GeCo is retained on SWaT, and NSIBF and GeCo on WADI.

Figure 2 relates these measurements to guarded updates using a 12-hour trailing window, a one-hour lag, and hourly refits. Guarding pauses refits during elevated alarm activity; a rolling update refits every hour regardless, and either variant can absorb a long attack into the reference. The comparison retains all eligible attack timestamps after warm-up. All eight retain-diagnosed pairs lose AUC when updated; four of the seven update-diagnosed pairs improve, and the other three, all on SWaT, change by -0.037 to -0.002. Applying guarded updates only to the seven update-diagnosed pairs gives mean AUC 0.809, compared with 0.751 for never updating and 0.759 for always updating. Importantly, this result is not driven by the three GDN cases used to set the boundaries: Over the twelve non-GDN pairs, these means are 0.805, 0.759, and 0.747. This policy comparison is retrospective because the screen summarizes the whole evaluation recording, although it uses no attack labels.

Updating can also remove attack contrast. Guarding does not always prevent absorption: it restores FuSAGNet on SWaT but not TranAD (Table 8). Two scope checks (Appendix B): BATADAL is retained for all three GDN seeds and its updates change AUC by at most 0.055; on the NoBOOM batch-process pilot (Wagner et al., 2025) only 5 of 17 update-diagnosed recordings improve, so the rule did not transfer reliably in this short-batch pilot.

Table 1: Observed response patterns across the fifteen detector–dataset pairs, on all eligible timestamps after warm-up. AUC (fixed): CNRE ROC-AUC under the training-normal reference. ΔAUC (update): change from a guarded update, shown for every pair; the policy applies it only to update-diagnosed pairs (Table 6). Patterns are defined after the comparisons (Section 5.3) and are not failure causes identified by the screen.

Dataset	Detector	D_{ref}	Screen	AUC (fixed)	ΔAUC (update)	Observed pattern
SWaT	GDN	8.71×10^3	Update	0.814	-0.012	Alarms miscalibrated
	FuSAGNet	72.6	Update	0.848	$+0.073$	Improved by update
	TranAD	2.15×10^3	Update	0.428	-0.037	Weak separation
	NSIBF	62.5	Update	0.836	-0.002	Alarms miscalibrated
	GeCo	0.181	Retain	0.887	-0.369	Strong ranking
WADI	GDN	7.77×10^7	Update	0.450	$+0.335$	Improved by update
	FuSAGNet	2.75×10^7	Update	0.458	$+0.316$	Improved by update
	TranAD	2.15×10^7	Update	0.583	$+0.204$	Improved by update
	NSIBF	1.94	Retain	0.886	-0.063	Strong ranking
	GeCo	0.00486	Retain	0.675	-0.097	Weak separation
HAI	GDN	0.895	Retain	0.889	-0.054	Strong ranking
	FuSAGNet	1.02	Retain	0.897	-0.086	Strong ranking
	TranAD	0.917	Retain	0.907	-0.063	Strong ranking
	NSIBF	0.850	Retain	0.838	-0.013	Strong ranking
	GeCo	0.182	Retain	0.865	-0.003	Strong ranking

Diagnostic implication. Elevation under the training reference marks the pairs whose reference should be examined first; updating helped only there, and the guard matters because a persistent attack can be absorbed.

5.3 Attack Evidence After the Reference Decision

Attack evidence under the fixed reference mixes how well the model exposes attacks with how stale its reference is, so we read scalar separation only after the reference comparisons. Table 1 summarizes the patterns observed after these comparisons; they describe responses to interventions, not failure causes identified by the reference screen. *Improved by update:* the three update-diagnosed WADI detectors gain 0.204–0.335, and FuSAGNet on SWaT rises from 0.848 to 0.920 (Table 8). *Alarms miscalibrated:* GDN and NSIBF on SWaT already rank attacks at 0.814 and 0.836, and the guarded update leaves ranking nearly unchanged, yet the training-normal threshold flags about 80% of normal timestamps. A trailing 99th-percentile threshold on the fixed reference separates two cases (Table 8). For NSIBF it lowers the normal FPR to 3.3% with 71.1% recall, better than the guarded update (40.2% recall at 4.7%), so its alarm failure is mainly a threshold problem. For GDN it leaves 10.4% FPR at 16.4% recall, whereas the guarded update reaches 1.3% at 35.1%, so the reference update also contributes. *Weak separation:* TranAD on SWaT stays at or below chance (0.391–0.428) under all tested references, although without the longest attack run an update raises it from 0.542 to 0.666, so part of this weakness reflects that attack; GeCo on WADI has AUC 0.675 and AP 0.099 against an attack prevalence of 0.067 with the retained reference. *Strong ranking:* the five HAI pairs, GeCo on SWaT, and NSIBF on WADI rank between 0.838 and 0.907 with a retained reference.

AP for all pairs and $R_{1\%}$ for HAI on the full archived test stream (Appendix C, Tables 10 and 11) do not always agree with AUC: on HAI, TranAD has the highest AUC but FuSAGNet the highest AP.

Diagnostic implication. These patterns prioritize further comparisons rather than locate a failure: weak separation calls for other scoring rules or model changes, strong pooled ranking for low-FPR and calibration checks, and miscalibrated alarms for separating threshold from reference effects. None of them excludes limitations at other stages.

Table 2: AUC change from trailing-percentile rescoring relative to fixed CNRE on matched timestamps, with residuals and reference unchanged. All keeps every attack; Excl. omits each recording's longest attack run.

	SWaT		WADI		HAI	
Detector	All	Excl.	All	Excl.	All	Excl.
GDN	-0.222	$+0.028$	$+0.269$	$+0.162$	$+0.004$	$+0.005$
FuSAGNet	-0.221	$+0.004$	$+0.276$	$+0.171$	-0.007	-0.007
TranAD	-0.191	$+0.004$	$+0.187$	$+0.159$	$+0.008$	$+0.009$
NSIBF	$+0.039$	$+0.028$	-0.096	-0.162	$+0.003$	-0.001
GeCo	-0.061	-0.039	-0.011	-0.017	-0.001	-0.001

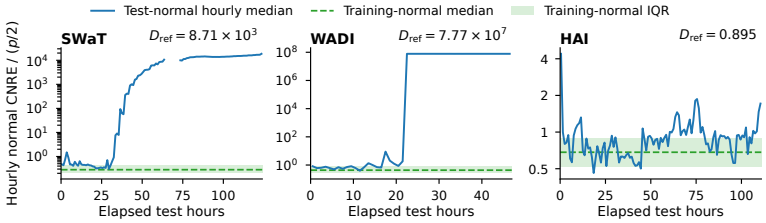

Figure 3: Hourly median CNRE of GDN's normal test timestamps (blue) under the fixed training-normal reference, against the training-normal median and interquartile range (green), all divided by $p/2$. D_{ref} uses all evaluation blocks, including attacks; gaps are hours without normal timestamps.

5.4 Evidence Extraction: Scoring and Calibration on Fixed Residuals

Cross-detector scoring intervention. To change scoring without touching the reference, trailing-percentile rescoring keeps the detector, residuals, and CNRE reference fixed and replaces each energy by its percentile among the energies of a preceding 12-hour window that ends one hour earlier and advances hourly, so it uses only past observations (Appendix D; Table 2).

On all timestamps, rescoring reproduces 80–91% of the guarded-update gain for the three update-diagnosed WADI detectors but lowers AUC by 0.191–0.222 for GDN, FuSAGNet, and TranAD on SWaT, where the guarded update changes it by -0.037 to $+0.073$. Like an adaptive reference, the trailing window absorbs the ten-hour SWaT attack, the loss of sensitivity that Garg et al. (2022) describe for adaptive scoring. Without the longest run, rescoring reproduces 57–86% of the rolling-update gain on WADI but only 2–13% on SWaT. Rescoring thus reproduces part of the point-level AUC improvement of a reference update on WADI, but not on SWaT under the evaluated protocol; because only one rescoring rule is tested, the remaining SWaT gap does not show that covariance changes are necessary. Where no update is recommended, HAI changes stay within 0.009, but the other retain-diagnosed pairs lose up to 0.096 (0.162 without the longest run).

Diagnostic implication. Rescoring after a reference flag thus shows whether the scoring stage can compensate for drift; its losses on the long SWaT attack motivate contamination-aware adaptive scorers, and whether the reference guard transfers remains to be tested. Alarm-threshold calibration, the remaining extraction question, is examined on HAI in the case study.

5.5 GDN Case Study: Interpreting Model and Scoring Interventions

The following GDN comparisons illustrate how the workflow interprets model and scoring interventions; they are separate comparisons, not a policy of successive repairs, and use the case study's HAI evaluation (Appendix A).

Reference. Figure 3 shows why the screen flags SWaT and WADI but not HAI for GDN: SWaT's normal energy rises during the recording to about 10^4 times the nominal scale, WADI's rises as a step, and HAI's stays near the training band. On SWaT, rolling updates absorb part of the ten-hour

Table 3: HAI adapter under two reference protocols. Adapter values are five-seed means on the same timestamps as the frozen GDN; Table 17 adds two test-label references.

	ROC-AUC		AP		$R_{1\%}$	
Reference	Frozen	Adapter	Frozen	Adapter	Frozen	Adapter
Each model's own refit	0.890	0.888	0.573	0.452	0.671	0.504
Frozen model's reference	0.890	0.900	0.573	0.593	0.671	0.721

attack (AUC 0.814 fixed, 0.771 rolling), and pausing refits during elevated alarm activity recovers most of that loss (0.802).

Attack evidence. On HAI, whose reference the screen does not flag, a residual adapter with a 16-unit bottleneck, inserted after GDN's node embeddings, is trained to reduce normal prediction MSE with the base checkpoint frozen, over five seeds (Appendix E). Normal MSE falls by 27.3%, but the sign of the effect on attack evidence depends on the reference protocol (Table 3): AP falls from 0.573 to 0.452 and $R_{1\%}$ from 0.671 to 0.504 when each model's reference is refitted to its own normal residuals, but they rise to 0.593 and 0.721 when both models keep the frozen model's reference. Under the refitted references, normal energy inflates sharply in two of the five HAI test recordings and only mildly in the others, whereas the frozen model's reference raises it evenly (Appendix E); this is consistent with refitting amplifying normal-to-test mismatch, but the comparison does not isolate recentering, covariance reweighting, and temporal drift. Better normal fit therefore does not determine attack evidence, a model intervention cannot be read without its reference protocol, and the reference screen must be repeated after a model change.

Extraction. On the same 40,171 HAI timestamps of the frozen GDN, CNRE raises AUC over the GDN-style scoring rule from 0.842 to 0.890, AP from 0.377 to 0.573, and $R_{1\%}$ from 0.433 to 0.671. Yet thresholds fixed at each score's 99.5th percentile on normal development data, a nominal FPR of 0.5%, give 72.1% recall at 2.19% FPR for CNRE and 38.4% recall at 0.51% FPR for the GDN-style rule: stronger ranking does not ensure that a calibrated threshold keeps its nominal FPR. Label-assisted selection on HAI recordings 1–2 chooses a robust-sum score, which re-centers and rescales each whitened coordinate before summing squares; it does not improve held-out AUC on recordings 3–5 (0.954 versus 0.957 for CNRE; Appendix D).

Outcome for GDN. On SWaT, a trailing threshold reduces but does not remove GDN's alarm miscalibration, and a reference update brings the FPR near its nominal level while partly absorbing the persistent attack; on WADI the update restores ranking; on HAI pooled ranking is strong, a model change must be read with its reference protocol, and alarm-threshold calibration remains open.

6 Discussion and Conclusion

We presented a workflow that uses CNRE to organize controlled comparisons at three stages of residual-based ICS detection. A label-free screen identified the pairs whose reference should be examined first, and a selective update policy outperformed always and never updating; a threshold-only control separated alarm failures that a threshold fixes from those that also need a reference update; and rescoring reproduced most of the update's AUC gain on WADI but not on SWaT. In the GDN case study, better normal prediction did not determine attack separation, the sign of a model change depended on its reference protocol, and stronger ranking did not ensure calibrated false-positive rates.

The reference policy is retrospective; its boundaries were developed on one detector family and tested on others over the same three datasets, and they did not transfer reliably in a short-batch pilot. The rescoring and rolling comparisons are unguarded, and the case study uses one detector with one pretrained checkpoint per dataset. Natural extensions are a prospective screen computed on a trailing window and a guard designed for adaptive scores as well as adaptive references.

AI USE STATEMENT

We used generative AI tools to assist the experiments, to support literature discovery, and to aid writing. For the experiments, they helped design control analyses, wrote and debugged the analysis and plotting code, ran it on the saved detector residuals, and summarized and helped interpret the outputs. They also suggested related work, helped edit the manuscript and review drafts for clarity and consistency, and drafted the workflow diagram in Figure 1. We have not used generative AI tools to generate synthetic datasets or to prove mathematical claims. The authors checked the analysis code, verified every reported number against the saved outputs, and take responsibility for the full content of this paper, including text, claims, and artifacts produced with AI assistance.

ETHICS STATEMENT

This work analyzes public ICS security benchmarks (SWaT, WADI, HAI, BATADAL, and NoBOOM) under their data-access terms. It studies the detection of attacks and does not introduce new attack techniques. A diagnostic recommendation can be wrong in either direction, delaying a needed reference update or triggering an unnecessary one, so recommendations should be checked with controlled comparisons before they inform an operational change.

REPRODUCIBILITY STATEMENT

Appendix A specifies data access, normal development splits, eligible targets, label aggregation, and metric definitions. Appendix B gives the reference-update and guard settings, Appendix D the scoring rules and calibration policies, and Appendix E the adapter architecture, training, and reference protocols. The code at `https://anonymous.4open.science/r/iclr2027-residual-evidence-diagnostics-E5D2` regenerates every table and figure from the saved residuals and checks each reported value against the regenerated results. The residual archives include per-timestamp data derived from SWaT and WADI, whose licenses do not allow redistribution; they are available from the authors to license holders.

REFERENCES

Chuadhry Mujeeb Ahmed, Venkata Reddy Palleti, and Aditya P. Mathur. WADI: A water distribution testbed for research in the design of secure cyber physical systems. In *Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks*, pp. 25–28. ACM, 2017. doi: 10.1145/3055366.3055375.

Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4027–4035, 2021. doi: 10.1609/aaai.v35i5.16523.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010.

Cheng Feng and Pengwei Tian. Time series anomaly detection for cyber-physical systems via neural system identification and Bayesian filtering. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2858–2867. ACM, 2021. doi: 10.1145/3447548.3467137.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44:1–44:37, 2014. doi: 10.1145/2523813.

Astha Garg, Wenyu Zhang, Jules Samaran, Ramasamy Savitha, and Chuan-Sheng Foo. An evaluation of anomaly detection and diagnosis in multivariate time series. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2508–2517, 2022. doi: 10.1109/TNNLS.2021.3105827.

Hamid Reza Ghaeini, Daniele Antonioli, Ferdinand Brasser, Ahmad-Reza Sadeghi, and Nils Ole Tippenhauer. State-aware anomaly detection for industrial control systems. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pp. 1620–1628. ACM, 2018. doi: 10.1145/3167132.3167305.

Siho Han and Simon S. Woo. Learning sparse latent graph representations for anomaly detection in multivariate time series. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2977–2986. ACM, 2022. doi: 10.1145/3534678.3539117.

Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. Towards a rigorous evaluation of time-series anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7194–7201, 2022. doi: 10.1609/aaai.v36i7.20680.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

Prasanta Chandra Mahalanobis. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55, 1936.

Aditya P. Mathur and Nils Ole Tippenhauer. SWaT: A water treatment testbed for research and training on ICS security. In *2016 International Workshop on Cyber-Physical Systems for Smart Water Networks (CySWater)*, pp. 31–36. IEEE, 2016. doi: 10.1109/CySWater.2016.7469060.

S. Joe Qin. Statistical process monitoring: Basics and beyond. *Journal of Chemometrics*, 17(8–9):480–502, 2003. doi: 10.1002/cem.800.

Xiangfei Qiu, Zhe Li, Wanghui Qiu, Shiyan Hu, Lekui Zhou, Xingjian Wu, Zhengyu Li, Chenjuan Guo, Aoying Zhou, Zhenli Sheng, Jilin Hu, Christian S. Jensen, and Bin Yang. TAB: Unified benchmarking of time series anomaly detection methods. *Proceedings of the VLDB Endowment*, 18(9):2775–2789, 2025. doi: 10.14778/3746405.3746407.

Hyeok-Ki Shin, Woomyo Lee, Jeong-Han Yun, and Byung-Gil Min. Two ICS security datasets and anomaly detection contest on the HIL-based augmented ICS testbed. In *Proceedings of the 14th Cyber Security Experimentation and Test Workshop*, pp. 36–40. ACM, 2021. doi: 10.1145/3474718.3474719.

Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouët. Anomaly detection in streams with extreme value theory. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1067–1075. ACM, 2017. doi: 10.1145/3097983.3098144.

Keith Stouffer, Michael Pease, CheeYee Tang, Timothy Zimmerman, Victoria Pillitteri, Suzanne Lightman, Adam Hahn, Stephanie Saravia, Aslam Sherule, and Michael Thompson. Guide to operational technology (OT) security. Technical Report NIST Special Publication 800-82 Rev. 3, National Institute of Standards and Technology, 2023.

Riccardo Taormina, Stefano Galelli, Nils Ole Tippenhauer, Elad Salomons, Avi Ostfeld, Demetrios G. Eliades, et al. Battle of the attack detection algorithms: Disclosing cyber attacks on water distribution networks. *Journal of Water Resources Planning and Management*, 144(8):04018048, 2018. doi: 10.1061/(ASCE)WR.1943-5452.0000969.

Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. TranAD: Deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment*, 15(6):1201–1214, 2022. doi: 10.14778/3514061.3514067.

Dennis Wagner, Fabian Hartung, Justus Arweiler, Aparna Muraleedharan, Indra Jungjohann, Arjun Nair, Steffen Reithermann, Ralf Schulz, Michael Bortz, Daniel Neider, Heike Leitte, Joachim Pfeffinger, Stephan Mandt, Sophie Fellenz, Torsten Katz, Fabian Jirasek, Jakob Burger, Hans Hasse, and Marius Kloft. NoBOOM: Chemical process datasets for industrial anomaly detection. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3499. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/96f8c5e879c339dae55e6c2188b02a33-Abstract-Datasets_and_Benchmarks_Track.html.

Konrad Wolsing, Eric Wagner, Luisa Lux, Klaus Wehrle, and Martin Henze. GeCos replacing experts: Generalizable and comprehensible industrial intrusion detection. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 7153–7172. USENIX Association, 2025.

Renjie Wu and Eamonn J. Keogh. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE Transactions on Knowledge and Data Engineering*, 35 (3):2421–2429, 2023. doi: 10.1109/TKDE.2021.3112126.

A Data Access and Evaluation Protocol

The experiments follow the three diagnostic questions in the main text. Reference interventions change the normal residual reference while keeping the detector fixed. The adapter case study changes a restricted part of the residual-producing model. Scoring comparisons keep the residuals fixed and change their conversion into scores or alarms. This separation makes the location of each intervention explicit.

Residual sources and benchmark support. The cross-detector study uses archived residuals from GDN, FuSAGNet, TranAD, NSIBF, and GeCo on SWaT, WADI, and HAI (Deng & Hooi, 2021; Han & Woo, 2022; Tuli et al., 2022; Feng & Tian, 2021; Wolsing et al., 2025). The datasets are described by Mathur & Tippenhauer (2016), Ahmed et al. (2017), and Shin et al. (2021), respectively. We retain each detector's archived preprocessing, residual dimensions, sampling interval, and timestamp alignment. TranAD results consistently use the v3 residual archive. These experiments compare reference choices within each detector; they do not retrain every detector under a common architecture or preprocessing pipeline.

The adapter and detailed extraction analyses use the GDN case study, which fixes one pretrained checkpoint and its evaluation timestamps. Its normal fitting split differs from the full training-normal reference used in the archived reference study. Consequently, the two GDN evaluations need not give identical scores or AUCs. Every intervention comparison uses the same eligible timestamps for its paired models or reference choices. We do not interpret differences between unmatched evaluation supports as intervention effects.

GDN preprocessing and eligible targets. In the GDN case study, inputs aggregate ten consecutive raw samples. Each processed target is labeled as an attack if any constituent raw sample is labeled as an attack. The model uses five processed observations to predict the next observation. HAI has 79 variables and 40,171 eligible test targets, including 935 attack targets and 39,236 normal targets. SWaT has 51 variables and 44,986 test targets, including 5,493 attack targets and 39,493 normal targets.

Windows crossing gaps within recordings or boundaries between recordings are removed from the HAI evaluation. A normal target whose input history contains an attack remains a normal target in the primary point-level metrics. The clean-normal pool used to fit the held-out scoring rules additionally excludes such histories. We distinguish these denominators when reporting false-positive rates in Appendix D.

Normal development splits. The normal recordings are divided chronologically into fitting, early-stopping, configuration-selection, and retrospective-audit regions. Five candidate window starts are purged between adjacent regions. Table 4 gives the retained counts. The fitting region supplies gradient updates and the normal residual reference. The early-stopping region selects the epoch and calibrates the fixed alarms described in Appendix D. Normal prediction MSE in the main adapter comparison is measured on the configuration-selection split.

Table 4: Retained normal development windows for the GDN case study. The HAI fitting count excludes 12 windows that cross recording gaps.

Dataset	Fitting	Early stopping	Config. selection	Audit
HAI	55,281	9,211	9,205	18,431
SWaT	28,509	4,746	4,747	9,498

The inherited input scaler was fitted on the full raw normal recording, and these normal data had been inspected during earlier development. The precise original checkpoint validation split is not fully resolved by the archived metadata. These regions therefore support a normal-only development comparison, but they are not pristine prospective holdouts. The adapter objective does not use attack labels.

Where labels enter the analysis. The initial CNRE reference uses normal training residuals. The recording-level reference statistic uses evaluation blocks without removing attacks. Labels are used

retrospectively to isolate normal targets in the drift visualization and to measure attack–normal separation. The adapter training uses normal data only. In contrast, the held-out scoring experiment uses labels from HAI recordings 1–2, the scorer-selection recordings, to select a scoring rule and identify its clean fitting samples. Recordings 3–5 evaluate the frozen choice. The test-regime and oracle references in Appendix E also use test labels and are retrospective diagnostic comparisons.

The trailing-percentile control in Section 5.4 uses past evaluation energies, including attacks, without label filtering. Labels are used only to evaluate the resulting scores and, for the Excl. columns of Table 2, to exclude the longest attack run from metric calculation. This exclusion does not remove observations from the scoring history.

Measurements. Table 5 summarizes how each measurement connects the experiments to the diagnostics; the definitions below make them precise.

Table 5: Measurements used to connect the experiments to the diagnostics. Higher values indicate stronger ranking or detection for AUC, AP, and recall; neither CNRE magnitude nor D_{ref} alone establishes an attack.

Measurement	Meaning and role
CNRE, e_t	Residual deviation relative to normal scale and covariance.
D_{ref}	Persistent CNRE elevation across evaluation blocks; motivates checking the reference.
CNRE ROC-AUC	How often attack CNRE ranks above normal CNRE.
Average precision (AP)	Precision summarized over increasing recall; complements ROC-AUC.
$R_{1\%}$	Retrospective recall at 1% test FPR, interpolated from the test ROC curve.
Fixed-threshold recall / FPR	Attack fraction detected / normal fraction flagged after normal-data calibration.
Normal MSE	Prediction fit on the normal selection split; not an attack-detection metric.

Point-level metrics. Let $\mathcal{T}$ be the eligible evaluation timestamps, $\ell_t \in \{0, 1\}$ the label at timestamp t, and s_t a score whose larger values indicate stronger anomaly evidence. An alarm at threshold τ is $a_t(\tau) = \mathbb{I}[s_t > \tau]$. Its recall and false-positive rate are

$$R(\tau) = \frac{\sum_{t\in\mathcal{T}} \ell_t a_t(\tau)}{\sum_{t\in\mathcal{T}} \ell_t}, \qquad F(\tau) = \frac{\sum_{t\in\mathcal{T}} (1-\ell_t) a_t(\tau)}{\sum_{t\in\mathcal{T}} (1-\ell_t)}. \tag{6}$$

Thus, R measures the fraction of attack targets detected, while F measures the fraction of normal targets flagged. We do not apply point adjustment or expand an alarm to cover its entire attack interval.

ROC-AUC summarizes the ranking obtained as τ varies (Fawcett, 2006). Average precision (AP) sums precision at each distinct score threshold, weighted by the corresponding increase in recall. It therefore summarizes the precision–recall curve rather than the ROC curve. Tied scores are handled together. The reported $R_{1\%}$ is recall at $F = 0.01$, obtained by linear interpolation of the empirical test ROC curve. It uses test labels retrospectively and is not a threshold selected from normal development data.

For the held-out scoring comparison, we also use

$$\mathrm{nPAUC}_\alpha = \frac{1}{\alpha} \int_0^\alpha \mathrm{TPR}(u)\, du, \qquad \alpha = 0.01. \tag{7}$$

Here u is the false-positive rate, $\mathrm{TPR}(u)$ is recall on the interpolated ROC curve, and α is the upper false-positive rate of interest. This quantity is the average recall over the first 1% of the ROC curve. It is area divided by α, not a standardized partial AUC that assigns random ranking a value of 0.5.

Aggregation and variability. Adapter summaries use optimization seeds 5, 1014, 2023, 3032, and 4041. All five runs share the same pretrained base checkpoint. Where shown, $\pm$ denotes the sample standard deviation across these five runs; it does not measure uncertainty over independently pretrained detectors. Reference-policy means give equal weight to each detector–dataset pair. Numerical outcome bands used for reference comparisons are descriptive categories, not statistical significance tests. Intervals for reference effects come from a circular moving-block bootstrap of each

Table 6: Reference diagnosis and guarded-update effects on all eligible timestamps. ΔAUC is guarded minus fixed AUC on the same eligible timestamps. The GDN rows informed the decision boundaries. None of the fifteen pairs lies in the 3–10 no-recommendation region. Values are rounded. The 95% interval comes from a circular moving-block bootstrap of the scored timestamps (one-hour blocks, 200 resamples) and reflects within-recording sampling variability only.

Dataset	Detector	D_{ref}	Recommendation	ΔAUC	95% interval
SWaT	GDN	8.71×10^3	Investigate update	-0.012	$[-0.134, 0.153]$
	FuSAGNet	72.6	Investigate update	$+0.073$	$[0.027, 0.146]$
	TranAD	2.15×10^3	Investigate update	-0.037	$[-0.128, 0.059]$
	NSIBF	62.5	Investigate update	-0.002	$[-0.100, 0.106]$
	GeCo	0.181	Retain	-0.369	$[-0.524, -0.152]$
WADI	GDN	7.77×10^7	Investigate update	$+0.335$	$[0.073, 0.558]$
	FuSAGNet	2.75×10^7	Investigate update	$+0.316$	$[0.054, 0.534]$
	TranAD	2.15×10^7	Investigate update	$+0.204$	$[-0.023, 0.386]$
	NSIBF	1.94	Retain	-0.063	$[-0.306, 0.065]$
	GeCo	0.00486	Retain	-0.097	$[-0.230, 0.002]$
HAI	GDN	0.895	Retain	-0.054	$[-0.115, -0.020]$
	FuSAGNet	1.02	Retain	-0.086	$[-0.166, -0.036]$
	TranAD	0.917	Retain	-0.063	$[-0.124, -0.027]$
	NSIBF	0.850	Retain	-0.013	$[-0.045, 0.009]$
	GeCo	0.182	Retain	-0.003	$[-0.010, 0.002]$

recording's scored timestamps with one-hour blocks and 200 resamples; recordings are resampled independently when effects are combined across pairs. These intervals describe within-recording sampling variability, not variation across detectors, datasets, or retraining.

B Reference Interventions and Their Limits

A reference intervention asks whether the same residuals become more informative when normal behavior is represented differently. Distribution change is a common reason to reconsider a learned reference (Gama et al., 2014). The intervention must account for two competing effects: an old reference can overstate deviation during normal operation, while an adaptive reference can incorporate an ongoing attack. The following results examine both effects.

Reference statistic and decision regions. The fixed reference consists of the training-normal residual mean and a Ledoit–Wolf covariance estimate (Ledoit & Wolf, 2004). For residual dimension p, the nominal energy scale is $p/2$. This is the expected CNRE when the reference equals the population normal mean and covariance; an estimated shrinkage reference need not attain it exactly. We compute D_{ref} as the median of the within-block median energies, divided by $p/2$. Blocks span one hour in the primary SWaT, WADI, and HAI study. Attack labels do not enter this calculation. The normal-only curves in Figure 3 (Section 5.5) are a separate, label-assisted visualization.

The empirical rule recommends retaining the reference when $D_{\text{ref}} < 3$, investigating an update when $D_{\text{ref}} > 10$, and makes no recommendation in the intervening region. The three GDN cases informed these boundaries. The other twelve pairs test transfer across detector families on the same three datasets. No pair lies in the no-recommendation region, so the experiments do not establish optimal boundaries or evaluate behavior within that region. The diagnostic policy below retains the fixed reference when no update is recommended. Table 6 lists the diagnostic value, recommendation, and guarded-update effect for each pair, and Figure 2 in Section 5.2 plots them.

Causal rolling and guarded references. The rolling reference is refitted hourly using the preceding 12-hour residual window, ending one hour before the current scoring block. Within that fitting window, each residual coordinate is clipped at its empirical 1st and 99th percentiles. The clipped residuals determine the mean and Ledoit–Wolf covariance. This winsorization limits extreme samples but does not use attack labels to identify contamination. Fixed, rolling, and guarded scores are compared after the same initial 13-hour warm-up and on their common eligible timestamps.

Table 7: Reference policies on all eligible timestamps. Means are unweighted across detector–dataset pairs. Improvement and degradation count changes of at least $+0.01$ and at most -0.01, respectively, relative to never updating. The twelve non-GDN pairs share the same three datasets as the boundary-development cases.

Policy	AUC: all 15	AUC: non-GDN 12	Improved / degraded
Never update	0.751	0.759	0 / 0
Always guarded update	0.759	0.747	4 / 9
Diagnostic guarded update	0.809	0.805	4 / 2

The guarded version checks the preceding hour's alarm fraction. If more than 20% of that hour exceeds its trailing 99.5th-percentile threshold, the next reference refit is skipped. The mean and covariance can remain frozen for at most 24 consecutive hourly updates, after which refitting is allowed. The operational alarm threshold is the training-normal 99th percentile for the fixed reference and the trailing-window 99th percentile for an adaptive reference. The latter is recalculated under the current mean and covariance even while those parameters are frozen. Thus, this guard freezes the reference, not the entire alarm pipeline.

Policy comparison on all eligible timestamps. Table 7 summarizes the consequences of applying the diagnostic recommendation. The diagnostic policy uses the guarded update only for update-diagnosed pairs. Its mean AUC is 0.809, compared with 0.751 for a fixed reference and 0.759 for always applying the guarded update. Resampling each recording in one-hour blocks (Appendix A), the policy exceeds never updating by 0.058 (95% interval 0.016–0.109) and always updating by 0.050 (0.030–0.076); on the twelve non-GDN pairs the margins are 0.046 (0.014–0.086) and 0.058 (0.033–0.090). However, it still degrades two pairs by at least 0.01. The result supports selective intervention, not a guarantee that every recommended update helps.

The policy uses a diagnostic calculated over the evaluation recording. Although this calculation does not use attack labels, the policy comparison is retrospective. It is not an online decision made before the recording begins. Furthermore, the three GDN boundary-development cases remain part of the fifteen-pair mean.

Why the guard is only a partial remedy. Table 8 retains all eligible SWaT timestamps. FuSAGNet benefits from guarding: AUC rises from 0.709 with ordinary rolling to 0.920 with guarding, above the fixed-reference value of 0.848. GDN recovers part of its rolling-reference loss but remains below its fixed-reference AUC. TranAD does not improve with the guard. GeCo, for which the diagnostic recommends retaining the reference, loses substantial separation when its reference is updated.

The alarm columns of Table 8 separate threshold adaptation from reference updating. For the four update-diagnosed detectors, a fixed training-normal threshold flags about 80% of normal timestamps. A trailing 99th-percentile threshold on the fixed reference lowers this to 2.0–10.4%; for NSIBF it keeps 71.1% recall, more than the guarded reference (40.2%), whereas for GDN and FuSAGNet the guarded reference reaches a lower FPR with higher recall. A low adaptive FPR does not ensure high recall: the trailing variants detect 1.0–71.2% of attack timestamps.

Sensitivity to long attacks and normalization. A supplementary evaluation removes the longest contiguous attack run from each stream when calculating metrics. The references are still constructed from the original stream: the excluded run is not removed from rolling-reference fitting. Under this evaluation, mean AUC is 0.700 for the fixed reference, 0.779 for always rolling, and 0.808 for diagnostic rolling. These numbers describe a different evaluation support and a different update variant from Table 7.

For the coarse outcome analysis, an AUC change above $+0.05$ is classified as helpful, below -0.05 as harmful, and otherwise as neutral. An update recommendation agrees with a helpful outcome; a retain recommendation agrees with a neutral or harmful outcome. Agreement is 15/15 after longest-run exclusion with ordinary rolling, but 12/15 when long attacks are retained and the guard is used. The three latter disagreements are GDN, TranAD, and NSIBF on SWaT: their guarded effects are neutral rather than helpful. This outcome definition is distinct from the 0.01 counting rule in Table 7.

Table 8: SWaT reference and threshold interventions on matched, all-eligible support. ROC-AUC uses the fixed, rolling, or guarded reference. Alarm columns give recall / normal FPR in percent for the fixed reference with the training-normal 99th percentile, the fixed reference with a trailing 12-hour 99th percentile, and the guarded reference with its trailing 99th percentile.

	ROC-AUC			Alarm recall / normal FPR (%)		
Detector	Fixed	Rolling	Guarded	Fixed, training	Fixed, trailing	Guarded
GDN	0.814	0.771	0.802	93.3 / 79.6	16.4 / 10.4	35.1 / 1.3
FuSAGNet	0.848	0.709	0.920	94.6 / 79.2	19.5 / 2.0	71.2 / 1.4
TranAD	0.428	0.391	0.391	93.4 / 78.3	1.0 / 2.2	7.1 / 2.1
NSIBF	0.836	0.823	0.834	94.1 / 80.3	71.1 / 3.3	40.2 / 4.7
GeCo	0.887	0.518	0.518	68.8 / 1.3	16.2 / 0.7	5.2 / 1.2

Table 9: BATADAL reference check on the organizers' second training dataset. Each update effect is paired with the fixed reference on the same timestamps. The two window lengths can have different eligible supports because of their different warm-up periods.

Seed	D_{ref}	Recommendation	ΔAUC: 168 h	ΔAUC: 336 h
5	2.45	Retain	$+0.037$	$+0.040$
1014	1.06	Retain	-0.035	-0.027
2023	0.89	Retain	-0.055	-0.047

A diagonal-normalization control retains each residual mean and variance but removes cross-variable covariance. In the longest-run- excluded analysis, it gives agreeing outcomes for all fourteen decided pairs. NSIBF–SWaT changes from $D_{\mathrm{ref}} = 62.5$ under full covariance to 8.71 under diagonal normalization and enters the no-recommendation region. Thus, these reference results do not establish that full covariance is essential for every diagnostic decision.

B.1 Additional Dataset Checks

BATADAL. This check applies the existing decision boundaries to a separately trained GDN on BATADAL (Taormina et al., 2018). The model is trained on the organizers' first training dataset and evaluated on their second training dataset, which contains seven attacks. This is not the withheld official test set. The chronological normal split allocates 80% to fitting and 20% to validation. GDN uses a history of five observations, hidden width 64, five graph neighbors, and dropout 0.2. Training uses Adam with learning rate 10^{-3}, batch size 128, at most 600 epochs, and early-stopping patience 20.

Because BATADAL observations are hourly, the diagnostic uses 24-hour blocks. References are updated daily using a 168- or 336-hour trailing window with a 24-hour lag. The decision boundaries remain 3 and 10. All three training seeds receive a retain recommendation (Table 9). For the 168-hour comparison, the matched-support mean AUC is 0.798 for the fixed reference and 0.780 for rolling. Seed 5 nevertheless improves by 0.037. Agreement with retain therefore refers to the coarse 0.05 outcome band, not to the claim that every update reduces AUC.

NoBOOM. We also examine the ternary batch-distillation plant from NoBOOM (Wagner et al., 2025). These recordings contain process faults and should not be treated as interchangeable with the cyberattack datasets. The pilot uses five-minute diagnostic blocks and a ten-minute rolling window with a one-minute lag and one-minute refits. It clips fitting residuals at the 1st and 99th percentiles.

For each operating point, GDN is fitted to its normal recordings using one-second observations and a history of ten samples. Channels with training standard deviation below 10^{-9} are removed; the remaining channels use training-normal min–max scaling. The chronological split uses 80% for fitting and 20% for validation. The model has hidden width 64, up to five graph neighbors, and dropout 0.2. Training uses seed 5, Adam at 10^{-3}, batches of 128, at most 199 epochs, and early-stopping patience 10. The fixed reference uses fitting-split residuals. Fault targets are identified by a positive value of the dataset's common fault label.

Table 10: Point-level CNRE evidence under fixed training-normal references for all fifteen pairs. Each pair uses its full archived test support, retaining all attacks. Preprocessing, sampling intervals, and attack prevalence can differ between detector outputs; this is a diagnostic summary of the artifacts. The AP of a random score equals the attack prevalence: 0.121–0.124 on SWaT, 0.058–0.074 on WADI, and 0.022–0.036 on HAI. Unlike Table 1, this table includes timestamps before the 13-hour warm-up of the adaptive references.

	SWaT		WADI		HAI	
Detector	AUC	AP	AUC	AP	AUC	AP
GDN	0.7877	0.6320	0.5302	0.1998	0.8897	0.5685
FuSAGNet	0.8310	0.7192	0.5295	0.2056	0.8985	0.6512
TranAD	0.4754	0.1071	0.6117	0.3427	0.9095	0.5850
NSIBF	0.8083	0.7000	0.7958	0.4419	0.8441	0.5595
GeCo	0.8770	0.7540	0.6236	0.0779	0.8679	0.5535

Table 11: Point-level CNRE evidence for the archived HAI detector outputs, each with its training-normal reference.

Detector	ROC-AUC	AP	$R_{1\%}$
GDN	0.8897	0.5685	0.6578
FuSAGNet	0.8985	0.6512	0.7134
TranAD	0.9095	0.5850	0.7227
NSIBF	0.8441	0.5595	0.5246
GeCo	0.8679	0.5535	0.5521

After excluding each recording's longest fault run from evaluation, 24 recordings remain evaluable. Of the 17 that receive an update recommendation, five improve, eight worsen, and four are neutral under the 0.05 outcome band; six fall in the no-recommendation region and one is retained. This is a scope limitation, including under the excluded-run protocol. The pilot changes both the recording structure and the adaptation time scale. It therefore does not isolate one cause of failure, and it does not justify transferring the continuous-recording rule to short batches without further validation.

C Supplementary Attack-Evidence Results

This appendix reports point-level CNRE evidence for the archived detector outputs of the cross-detector study and a supplementary interval-level view for the GDN case study.

Archived detector outputs. Table 10 reports point-level ROC-AUC and AP for all fifteen archived detector–dataset pairs under each detector's training-normal reference, on the full archived test support with all attacks retained. For most SWaT and WADI pairs these values are dominated by the reference mismatch of Section 5.2; the grouping in Table 1 reads them together with the diagnosed reference. Table 11 adds $R_{1\%}$ for HAI, where the reference is retained: TranAD has the largest ROC-AUC and $R_{1\%}$, while FuSAGNet has the largest AP. These values describe archived residual artifacts and references, not a retraining comparison under identical preprocessing. They need not match the GDN case study, whose fitting split and eligible targets differ.

Interval-level view. Point-level metrics are the primary evidence measurements in this paper. An interval-level view supplements them by showing whether pooled separation is shared across attacks or concentrated in a subset. It uses the same scalar CNRE and fixed normal reference as the point-level analysis.

Definition and interval construction. Let A_k be the eligible timestamps in the kth contiguous labeled attack interval, and let N contain all eligible normal targets in the same GDN case-study evaluation. We define

$$S(A_k) = \frac{1}{|A_k||N|} \sum_{t \in A_k} \sum_{u \in N} \left(\mathbb{I}[e_t > e_u] + \frac{1}{2} \mathbb{I}[e_t = e_u] \right). \tag{8}$$

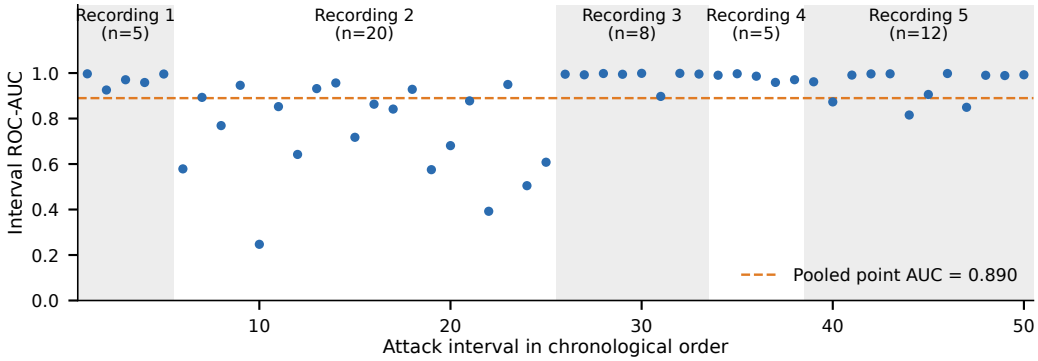

Figure 4: Descriptive interval AUCs for frozen GDN on HAI. Every interval is compared with the same full pool of eligible normal targets. All intervals, including short intervals, are retained and split at sample gaps and recording boundaries. The dashed line shows pooled point-level AUC, which is the duration-weighted average of these interval AUCs. No duration-matched null test or interval-level significance label is used.

Table 12: Descriptive interval AUC for frozen GDN on HAI. Minima, medians, and maxima summarize intervals without duration weighting. All rows use the same full normal comparison pool.

Recording	Intervals	Attack targets	Minimum	Median	Maximum
1	5	66	0.9255	0.9706	0.9966
2	20	362	0.2471	0.8053	0.9564
3	8	159	0.8973	0.9952	0.9987
4	5	120	0.9589	0.9861	0.9969
5	12	228	0.8155	0.9895	0.9980
All	50	935	0.2471	0.9531	0.9987

Here e_t and e_u are attack and normal CNRE values under the same reference. The statistic is the fraction of attack–normal pairs in which attack energy is larger, with half credit for ties. It is the ROC-AUC obtained by comparing one attack interval against the common normal pool. A large value indicates that this interval usually has higher energy than normal targets in that pool.

For frozen GDN on HAI, we split labeled attack runs at recording boundaries and gaps in eligible sample identifiers. All intervals are retained, including short intervals. This gives 50 intervals containing 935 attack targets. Every interval uses the same 39,236 normal targets as the primary point-level evaluation of the GDN case study. No interval-specific reference or threshold is fitted.

Heterogeneity within the point-level result. Table 12 and Figure 4 show substantial variation. The interval AUCs range from 0.2471 to 0.9987, while the pooled point-level AUC is 0.8898. Recording 2 contains the lowest-separation interval and has a median interval AUC of 0.8053. Most intervals in the other recordings rank well against the common normal pool. Thus, strong pooled evidence does not imply equally strong evidence for every labeled attack.

The relation to the point-level result is exact when the intervals partition the attack targets and share the same normal pool:

$$S_{\text{atk}} = \sum_{k=1}^{K} \frac{|A_k|}{\sum_{j=1}^{K} |A_j|} S(A_k). \tag{9}$$

Here K is the number of intervals and S_{atk} is the pooled point-level ROC-AUC of Eq. (5). Each interval contributes in proportion to its number of eligible attack targets. At a fixed sampling interval this is duration weighting. It explains why pooled AUC differs from the unweighted median interval AUC, which is 0.9531 in this evaluation.

Interpretation and limits. These interval scores are descriptive. They do not compare an attack against a duration-matched distribution of normal windows, and we do not use them to assign statistical evidence-present or evidence-absent labels. In particular, a short interval can have an extreme AUC without providing a stable estimate of attack behavior. An AUC below 0.5 means that CNRE ranks this interval below the normal pool; it does not establish that the residual vector contains no information.

Likewise, a high interval AUC does not ensure an alarm at a threshold calibrated on normal development data. The primary extraction analysis therefore remains the point-level scoring and calibration comparison in Appendix D. The interval view identifies heterogeneous cases for inspection without changing that diagnostic criterion or making a claim about temporal-accumulation benefits.

D Scoring Rules and Alarm Calibration

Evidence extraction concerns what happens after the residuals have been produced. CNRE provides a direct scalar score, $s_t = e_t$, against which we compare other scoring rules applied to the same residuals. A scoring rule produces an anomaly score; the alarm decision rule compares that score with a threshold. The comparisons below separate score ranking from the behavior of a threshold fixed on normal data. A score can rank attacks well while a transferred threshold produces an unsuitable false-positive rate.

Cross-detector trailing-percentile control. This control retains the pretrained detector, residuals, and training-normal CNRE mean and covariance. For each scoring block b, let $\mathcal{H}_b$ contain the timestamps in the preceding 12-hour window, ending one hour before the block begins. For each timestamp t in that block, the rescored value is

$$s_t^{\mathrm{local}} = \frac{1}{|\mathcal{H}_b|} \sum_{u \in \mathcal{H}_b} \mathbb{I}[e_u \le e_t]. \tag{10}$$

Here e_t and e_u are energies computed under the same fixed CNRE reference, and $|\mathcal{H}_b|$ is the number of historical samples. The score is the fraction of historical energies no larger than the current energy. Including ties matches the saved implementation's empirical cumulative distribution. The history window advances hourly and uses all past residuals, without filtering attack timestamps.

Fixed CNRE and the rescored values are evaluated on the same timestamps after the initial 13-hour warm-up. Table 2 reports two evaluation supports: all scored timestamps, and the same timestamps without each recording's longest contiguous attack run. In the second, the excluded run remains in the original stream and can enter later history windows; labels determine this evaluation mask but do not construct the scores. The longest SWaT run lasts about ten hours and covers about 70% of the scored attack timestamps, so a 12-hour history window absorbs much of it, which is consistent with the losses in the All columns. Without the longest run, unguarded rolling updates change AUC by $+0.328$, $+0.209$, $+0.125$, and $+0.220$ for GDN, FuSAGNet, TranAD, and NSIBF on SWaT, and by $+0.283$, $+0.274$, and $+0.184$ for GDN, FuSAGNet, and TranAD on WADI. These are the denominators of the recovered shares in Section 5.4.

Because the empirical distribution changes between blocks, this transformation can change ranking across time. It should not be described as selecting a different constant threshold for an otherwise unchanged score. The comparison evaluates scalar rescoring; it does not compare all detectors' native scoring pipelines.

GDN-style scoring rule on fixed residuals. The GDN-style comparator follows the channel-normalization and aggregation structure of GDN (Deng & Hooi, 2021), with its normalization statistics fitted on the normal fitting split. For residual coordinate $r_{t,j}$, define

$$u_{t,j} = \frac{|r_{t,j}| - m_j}{q_{0.75,j} - q_{0.25,j} + 0.01}, \qquad s_t^{\mathrm{GDN}} = \max_{j=1,\dots,p} \left(\frac{1}{4} \sum_{k=0}^{3} u_{t-k,j} \right). \tag{11}$$

Here j indexes the p residual coordinates. The quantities m_j, $q_{0.25,j}$, and $q_{0.75,j}$ are the median, first quartile, and third quartile of $|r_{t,j}|$ on normal fitting data. The constant 0.01 stabilizes normalization

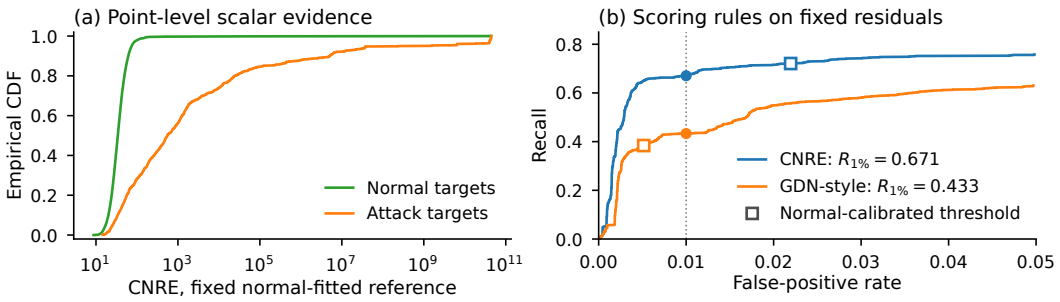

Figure 5: Point-level evidence and extraction for the frozen GDN on HAI in the GDN case study. (a) Empirical CNRE distributions for eligible attack and normal targets under the same normal-fitting reference. (b) Low-FPR ROC curves for CNRE and GDN-style scores on identical residuals and timestamps. Filled circles show retrospective recall at 1% test FPR, read from the test ROC curve; open squares show where each score's threshold calibrated on normal development data lands (Table 14).

Table 13: Score ranking on the same frozen GDN HAI residuals. $R_{1\%}$ uses test ROC interpolation. Table 14 gives the alarms with thresholds fixed on normal development data.

Scoring rule	ROC-AUC	AP	$R_{1\%}$
GDN-style	0.8425	0.3768	0.4332
CNRE	0.8898	0.5734	0.6706

when a channel has a small interquartile range. The four-sample trailing average smooths each normalized channel, and the maximum selects the largest smoothed channel response.

This equation describes the comparator used here rather than every native GDN evaluation procedure. In that implementation, the first three scores are zero. Smoothing follows the concatenated eligible sample sequence and does not reset at recording boundaries or removed-window gaps. We preserve this behavior when reproducing the archived comparison. CNRE and the GDN-style scoring rule are evaluated on identical saved residuals and target labels (Table 13). Their difference concerns the scoring rule as a whole; it does not isolate covariance normalization from smoothing or channel aggregation.

Thresholds fixed on normal development data. For each model and score, the fixed threshold is the 99.5th percentile on the early-stopping normal split. The quantile uses the upper empirical order statistic (the `higher` convention), and an alarm requires a score strictly above that threshold. No test labels select this threshold. Table 14 reports the realized recall and false-positive rate on all HAI test timestamps of the GDN case study.

For frozen GDN, CNRE detects more attack targets but also flags more normal targets at these separately calibrated thresholds. Section 5.5 and Figure 5 therefore also compare their retrospective ROC curves and $R_{1\%}$ on the same residuals. Those matched-FPR comparisons address ranking quality; Table 14 addresses normal-to-test threshold transfer. Neither measurement replaces the other.

Held-out scoring-rule selection. A separate experiment asks whether a more elaborate scoring rule improves on CNRE when selected on earlier recordings. The scorer-selection recordings, HAI 1–2, provide 16,189 eligible targets, and recordings 3–5 provide 23,982 held-out targets. The residual-producing GDN and its normal-fitting CNRE reference stay fixed. Scorer statistics are estimated only from clean normal samples of the scorer-selection recordings. Here, clean excludes normal targets whose model-input histories contain an attack.

The candidate with the largest point-level selection AUC is frozen before held-out evaluation. Since selection AUC uses attack labels, this is a label-assisted benchmark intervention. It is not a normal-

Table 14: HAI alarms with thresholds fixed on normal development data. FPR counts all normal-labeled eligible test targets. Adapter rows average five optimization seeds. These are different operating points, not comparisons at a matched test false-positive rate.

Model	Score	Recall	Test FPR
Frozen GDN	CNRE	0.721	0.0219
Primary adapter	CNRE	0.716	0.0375
Frozen GDN	GDN-style	0.384	0.0051
Primary adapter	GDN-style	0.422	0.0399

only deployment procedure. All target labels, including normal targets with attack-containing histories, remain in the point-level AUC calculations.

Candidate definitions. Let $\Sigma_0 = V\Lambda V^\top$ be the eigendecomposition of the fixed normal covariance, with eigenvalues in descending order. We compute

$$z_t = \Lambda^{-1/2} V^\top (r_t - \mu_0), \qquad E_{t,j} = z_{t,j}^2, \qquad \sum_{j=1}^{p} E_{t,j} = 2e_t. \tag{12}$$

The columns of V are covariance eigenvectors, Λ contains their eigenvalues, and μ_0 is the normal residual mean. Each $z_{t,j}$ is a covariance-normalized residual coordinate, while $E_{t,j}$ is its squared contribution to total energy. The sum score has exactly the same ranking as CNRE. Coordinates are used here to specify alternative downstream scoring rules, not to introduce a separate directional attack-evidence diagnosis.

The ten candidates are the sum, maximum, top-three sum, and top-five sum of $E_{t,j}$; robust sum and robust maximum; two weighted sums; and sum and maximum over the lower-eigenvalue half of the coordinates. The top-k rules sum the k largest values at each timestamp. The lower-eigenvalue rules use $j = \lfloor p/2 \rfloor + 1, \ldots, p$; for HAI this retains 40 of 79 coordinates.

For the robust rules, let b_j be the median of $z_{t,j}$ on the scorer-selection clean-normal pool and d_j its median absolute deviation about b_j. Then

$$\widetilde{z}_{t,j} = \frac{z_{t,j} - b_j}{1.4826 d_j + 10^{-9}}, \qquad s_t^{\mathrm{robust}} = \sum_{j=1}^{p} \widetilde{z}_{t,j}^2. \tag{13}$$

The normal-consistency factor 1.4826 rescales the median absolute deviation, and 10^{-9} prevents division by zero. The robust-maximum candidate uses $\max_j |\widetilde{z}_{t,j}|$ instead of the sum of squares. These rules fit additional centering and scaling parameters on the scorer-selection recordings while leaving the original residual producer and the CNRE whitening transform unchanged.

For the instability-weighted sum, the clean selection indices are split chronologically into twenty approximately equal groups. Let $m_{g,j}$ and $v_{g,j}$ denote the mean and population variance of $z_{t,j}$ in group g. The unnormalized weight is

$$\widetilde{w}_j = \left[\mathrm{Var}_g(m_{g,j}) + \frac{1}{20} \sum_{g=1}^{20} |v_{g,j} - 1| + 10^{-3} \right]^{-1}. \tag{14}$$

The two terms penalize changes in the normal coordinate mean and departures from unit variance. The constant 10^{-3} stabilizes the inverse. We divide these weights by their mean across coordinates and use $\sum_j w_j E_{t,j}$ as the score. The kurtosis-weighted candidate instead starts from $\widetilde{w}_j = 1/\max(\gamma_j, 0.5)$, where γ_j is the selection-normal excess kurtosis of coordinate j. It is computed from the fourth central moment divided by the squared population variance plus 10^{-12}, minus three. Its weights are also normalized to have mean one.

Held-out ranking and calibration. The selected robust sum improves selection AUC from 0.8430 to 0.8456, but its held-out AUC is 0.9539, below CNRE's 0.9572 (Table 15). Held-out $\mathrm{nPAUC}_{0.01}$ also falls, from 0.6410 to 0.6292. The lower-eigenvalue-half sum has a higher held-out AUC, but choosing it after viewing that result would violate the selection protocol.

Table 15: HAI scoring candidates. Robust sum was selected by AUC on recordings 1–2. Other held-out results were examined only after this choice was frozen and do not define a new selected rule.

Rule	Selection AUC	Held-out AUC
Robust sum (selected)	0.8456	0.9539
Sum (CNRE ranking)	0.8430	0.9572
Top-three sum	0.8424	0.9494
Maximum	0.8423	0.9419
Top-five sum	0.8423	0.9522
Kurtosis-weighted sum	0.8376	0.9393
Robust maximum	0.8362	0.9396
Lower-eigenvalue half: maximum	0.8350	0.9504
Lower-eigenvalue half: sum	0.8335	0.9599
Instability-weighted sum	0.7994	0.8751

For alarms in this experiment, each score uses its empirical 99th percentile on scorer-selection clean normals, with NumPy's linear quantile interpolation and strict threshold exceedance. The held-out clean-normal FPR is 0.0027 for CNRE and 0.0068 for robust sum. These rates exclude attack-containing input histories and therefore use a narrower normal denominator than Table 14. They should not be silently substituted for the all-normal-target FPR in the primary metrics.

The positive fixed-residual comparison and the unsuccessful held-out selection answer different questions. The first shows that the scoring rule can limit extraction of available residual evidence. The second shows that changing the scoring rule does not guarantee improvement on later recordings. These results motivate evaluating scoring interventions on held-out timestamps and reporting their realized false-positive rates.

E Adapter Case Study and Reference Sensitivity

The adapter is a controlled model intervention. Its purpose is to test whether improving normal prediction also improves attack evidence. It is not trained to optimize attack detection, and its results do not establish a limit on all possible representation changes.

Architecture and insertion site. The primary adapter is inserted after GDN's node-embedding multiplication, at a late representation with 64 features per node. It applies the same residual transformation to every node:

$$h' = h + W_{\mathrm{up}}\,\mathrm{GELU}(W_{\mathrm{down}}\,\mathrm{LN}(h) + b_{\mathrm{down}}) + b_{\mathrm{up}}. \tag{15}$$

Here $h \in \mathbb{R}^{64}$ is a node's original hidden representation and h' is its adapted representation. Layer normalization LN uses no trainable affine parameters and numerical stabilizer 10^{-5}. The matrix $W_{\mathrm{down}} \in \mathbb{R}^{w\times 64}$ compresses the normalized features to bottleneck width w, and $W_{\mathrm{up}} \in \mathbb{R}^{64\times w}$ maps the nonlinear correction back to the original dimension. The vectors b_{down} and b_{up} are the corresponding biases. GELU supplies the nonlinearity. Adding the correction to h lets the model begin from the original detector and learn a restricted change to its prediction.

The primary width is $w = 16$, giving 2,128 trainable parameters. The upward projection and its bias are initialized to zero, so the initial adapter is the identity. The base detector's parameters and buffers remain frozen. Its modules stay in evaluation mode, so base dropout is disabled and normalization buffers do not change. The adapter itself has no dropout.

Training and reference fitting. Only normal fitting windows provide gradient updates. The objective is prediction MSE; there is no attack supervision or residual-energy regularizer. Adam (Kingma & Ba, 2015) uses learning rate 10^{-3}, batch size 256, coefficients $(0.9, 0.999)$, numerical stabilizer 10^{-8}, and zero weight decay. Runs last at most 120 epochs, with early-stopping patience 15. The identity model at epoch zero remains eligible for selection. The five optimization seeds are listed in Appendix A.

Table 16: HAI point-level CNRE results on identical test targets of the GDN case study. Each model uses its own normal-fitting reference. Adapted rows report mean ± sample standard deviation over five optimization seeds. The frozen row is one shared checkpoint.

Intervention	ROC-AUC	AP	$R_{1\%}$
Frozen GDN	0.8898	0.5734	0.6706
Output head only	0.8851 ± 0.0003	0.5635 ± 0.0004	0.6539 ± 0.0006
Late adapter, $w = 4$	0.8847 ± 0.0046	0.4848 ± 0.0140	0.5621 ± 0.0192
Late adapter, $w = 8$	0.8856 ± 0.0035	0.4712 ± 0.0155	0.5330 ± 0.0257
Late adapter, $w = 16$	0.8877 ± 0.0028	0.4523 ± 0.0120	0.5037 ± 0.0284
Post-activation, $w = 16$	0.8846 ± 0.0031	0.4722 ± 0.0139	0.5521 ± 0.0232

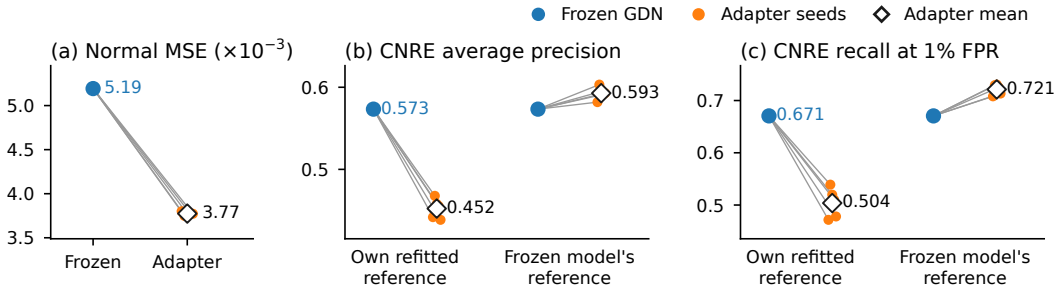

Figure 6: Normal fit and point-level attack evidence for the primary HAI adapter under two reference protocols. Blue markers denote the frozen GDN checkpoint, orange markers the five adapter optimization seeds sharing that checkpoint, and white diamonds their means. (a) Normal MSE on the configuration-selection split. (b, c) CNRE AP and recall at 1% FPR on the test targets of the GDN case study, when each model uses a reference refitted to its own normal residuals and when both use the frozen model's reference. Vertical axes do not start at zero.

After checkpoint selection, each model receives its own CNRE mean and covariance fitted to that model's residuals on the normal fitting split. This keeps the data used for reference fitting consistent while allowing the residual distribution to change after adaptation. The experiment therefore evaluates the adapted residuals together with their own normal reference; it does not hold the numerical covariance matrix fixed across different residual-producing models. Table 17 reports that alternative for HAI.

HAI controls. For the primary adapter, mean selection-normal MSE falls from 5.19×10^{-3} to 3.77×10^{-3}, a reduction of approximately 27.3%. However, CNRE AP falls from 0.5734 to 0.4523, and retrospective $R_{1\%}$ falls from 0.6706 to 0.5037. Figure 6 shows the five individual runs under both reference protocols. Table 17 contrasts this protocol with retaining the frozen model's reference, as discussed in Section 5.5. The small change in overall AUC does not capture the larger loss at low false-positive rates.

Table 16 provides nearby controls: smaller bottlenecks at the same insertion site, a width-16 adapter after the output normalization and activation, and training the existing output head alone. These alternatives do not reverse the primary conclusion for AP or $R_{1\%}$. Their role is to check the scope of the case study, not to support a claim that all adapter architectures fail.

HAI reference protocols. The primary comparison refits each model's CNRE mean and covariance to that model's residuals on the normal fitting split. Table 17 scores the same models under three other references. The frozen model's reference applies the mean and covariance fitted to the frozen GDN residuals to both models, so it holds the numerical reference fixed and uses only normal training data. The test-regime reference is fitted to clean test-normal residuals before the median clean-normal index and scores only the later half; the oracle reference uses all clean test-normal residuals. Both use test labels retrospectively. Under the frozen model's reference, the adapter improves all three metrics, and every seed has a higher AP than the frozen model. Under both test-normal references, its AP and $R_{1\%}$ losses are much smaller than under the primary protocol.

Table 17: HAI adapter results under four reference protocols. Adapter entries are five-seed means. Frozen and adapted models share evaluation timestamps within each row; the test-regime row scores only the later half of the test stream.

	ROC-AUC		AP		$R_{1\%}$	
Reference	Frozen	Adapter	Frozen	Adapter	Frozen	Adapter
Own refit (primary)	0.8898	0.8877	0.5734	0.4523	0.6706	0.5037
Frozen model's	0.8898	0.9002	0.5734	0.5928	0.6706	0.7213
Test-regime, later half	0.9159	0.9153	0.6092	0.5820	0.6529	0.6059
Oracle test-normal	0.8547	0.8624	0.5079	0.4963	0.5797	0.5549

Table 18: 99th percentile of CNRE over the normal timestamps of each HAI test recording. Adapter columns use each adapter's own refitted reference or the frozen model's reference and give the range over the five seeds.

Recording	Normal timestamps	Frozen GDN	Adapter, own refit	Adapter, frozen reference
1	4,249	171	1511–2017	310–418
2	11,512	81	99–108	274–398
3	10,635	110	135–170	302–404
4	3,834	107	456–660	315–430
5	9,006	131	169–215	329–423

The sign and size of the adapter effect therefore depend on the reference-fitting protocol; normal-fit improvement alone does not determine them.

Per-recording normal energy. The HAI test stream consists of five recordings. Table 18 gives the 99th percentile of CNRE over each recording's normal timestamps. Under each adapter's own refitted reference it rises about ninefold to twelvefold in recording 1 and fourfold to sixfold in recording 4 relative to the frozen GDN, but by less than 1.7 times in the other three recordings. Under the frozen model's reference, all five recordings stay within 274–430. This uneven inflation is consistent with the lower precision under the primary protocol; it does not separate recentering from covariance reweighting.

SWaT: a reference-sensitive companion. The same primary adapter reduces mean normal MSE from 1.71×10^{-3} to 1.45×10^{-3}, approximately 15.6%. With each model's own normal-fit reference, whole-stream AUC falls from approximately 0.764 to 0.664. We then change the reference protocol to examine whether this large gap persists when the reference represents test-normal behavior.

The test-regime and oracle references are defined as for HAI. Both require test labels and are retrospective diagnostic comparisons, not deployable normal-training-only interventions.

The much smaller within-row gaps under test-normal references show that the large own-reference gap is sensitive to reference choice. They do not identify a universal cause of adapter degradation. In particular, subtracting values across rows would mix different evaluation supports and reference-fitting data. Together with HAI, this case study suggests checking reference validity before interpreting a model intervention, and evaluating attack separation directly rather than using improved normal MSE as its proxy.

Table 19: Reference sensitivity of the SWaT adapter case study. Adapted results are five-seed mean AUCs. Frozen and adapted models share evaluation timestamps within each row. Different rows need not share the same support or data access.

Reference and evaluation support	Frozen	Adapter
Own normal-fit reference; full test	0.764	0.664
Test-regime reference; later-half test	0.912	0.904
Oracle test-normal reference; full test	0.907	0.904