OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

A Multi-Dimensional Safety Readiness Index for Machine Learning-Based Industrial Control System Anomaly Detectors

Ashwajit Warwatkar
ashvonte50@gmail.com

ABSTRACT

Purpose: Industrial control systems (ICS) require anomaly detectors that perform reliably under realistic deployment conditions, yet current evaluation practice relies almost exclusively on event-F1 computed under artificially inflated attack priors. This paper addresses this gap by introducing the Safety Readiness Index (SRI), a five-dimensional composite evaluation framework designed to assess detector quality across performance, robustness, uncertainty quantification, explainability, and fault tolerance.

Methods: The SRI framework derives dimension weights from Fussell-Vesely fault-tree importance analysis applied to IEC 62443 and NIST SP 800-82 guidelines. We evaluate five unsupervised anomaly detectors, Isolation Forest, One-Class SVM, XGBoost (lag-1 residual forecaster), LSTM Autoencoder, and USAD, on the SWaT water treatment benchmark, which comprises 51 sensors and 9 documented attack events. We compare two aggregation rules, Weighted Linear Sum and Geometric Mean, and conduct attribution localization studies against ground-truth attacked sensors.

Conclusions: The choice of aggregation rule, not model capability, determines the top-ranked model: Weighted Linear Sum ranks XGBoost first on individual dimensions yet USAD leads overall, whereas Geometric Mean collapses XGBoost to last place due to its zero Robustness score. At a realistic attack prior of 10^{-6}, XGBoost precision collapses to 7.4×10^{-5} while reconstruction-based models retain approximately 0.59 true positive rate under strict false-alarm constraints. Attribution localization reveals hit@1 equals zero for all five models, indicating that existing explainability metrics overstate interpretability by approximately 83%. We further prove via symbolic computation that any safety aggregator simultaneously satisfying strict Pareto-monotonicity and a zero-floor axiom is unsatisfiable. All findings, including negative results, are reported as observed.

Keywords: industrial control systems, anomaly detection, safety evaluation, multi-criteria decision making, fault tree analysis, conformal prediction, integrated gradients, SWaT

1. INTRODUCTION

The proliferation of machine learning-based intrusion detection systems in operational technology environments has outpaced the development of principled evaluation frameworks for their safety adequacy. Existing benchmarks including those applied to the widely-used Secure Water Treatment (SWaT) testbed (Goh et al., 2016; Mathur & Tippenhauer, 2016), report event-level F1 scores under artificially high attack priors (40.9% in the standard split), implicitly treating industrial anomaly detection as a balanced classification problem. This framing is ecologically invalid: in operational water treatment, energy distribution, or chemical processing facilities, the prior probability of a cyberattack at any given timestep is orders of magnitude lower than 0.4, and a single false alarm can trigger costly process shutdowns or unnecessary operator interventions.

Beyond the base-rate problem, F1 omits dimensions that are safety-critical in high-consequence environments. A detector that achieves high event-F1 on historical attacks but collapses under calibration drift, sensor dropout, or adversarial perturbation provides weaker operational safety guarantees than one with lower F1 but graceful degradation. Standards bodies—including IEC 62443 for industrial automation security (IEC, 2018) and NIST SP 800-82 for ICS security (NIST, 2023), articulate multi-dimensional requirements covering detection performance, resilience, uncertainty reporting, traceability, and fault tolerance, yet no published evaluation framework operationalizes these requirements into a composite score suitable for model comparison.

This paper makes four primary contributions. First, we introduce the Safety Readiness Index (SRI), a five-dimensional composite score (P, R, U, E, F) with weights derived from Fussell-Vesely fault-tree importance analysis of ICS attack scenarios rather than document page-counts or expert opinion. Second, we demonstrate that the aggregation rule is a first-order factor in model selection: the top-ranked model changes depending on whether a compensatory (WLS) or non-compensatory (GEO, MIN) aggregator is applied, and that four of five evaluated models score zero on at least one safety dimension. Third, we characterize operational deployment realism by computing precision across six decades of attack prior probability and true positive rate under strict false-alarm constraints, revealing that the event-F1 leader collapses at realistic operating points. Fourth, we conduct a ground-truth attribution localization study using Integrated Gradients (Sundararajan et al., 2017), finding that coverage-and-stability explainability metrics overstate genuine interpretability by approximately 83% when evaluated against actual sensors targeted in each documented attack.

The remainder of this paper is structured as follows. Section 2 reviews related work in ICS anomaly detection evaluation, uncertainty quantification, and explainability. Section 3 defines the SRI framework formally, including aggregators and weights. Section 4 describes the experimental setup. Section 5 presents results organized by each SRI dimension. Section 6 discusses key implications and limitations. Section 7 concludes.

2. RELATED WORK

2.1 Machine Learn ing Anomaly Detection for ICS

The SWaT dataset (Goh et al., 2016) has become the dominant benchmark for ICS anomaly detection, enabling comparison across reconstruction-based, density-based, and forecasting paradigms. LSTM Autoencoders (Malhotra et al., 2015) and the Unsupervised Anomaly Detection (USAD) architecture (Audibert et al., 2020) represent the reconstruction paradigm, learning to reconstruct normal sensor trajectories and flagging anomalies where reconstruction error exceeds a threshold. Isolation Forest (Liu et al., 2008) and One-Class SVM (Scholkopf et al., 2001) operate in feature space, estimating a density or boundary around normal operating regions. Tree-based forecasters such as XGBoost (Chen & Guestrin, 2016) predict next-step values with residual mean squared error as the anomaly score. Comparative studies on SWaT consistently find that no single model dominates across all evaluation dimensions, motivating multi-criteria assessment of the kind formalized in the SRI.

2.2 Safety and Assurance Evaluation Frameworks

The AI safety assurance literature increasingly recognizes the inadequacy of single-metric evaluation for safety-critical deployment. Goal Structuring Notation and assurance case frameworks (Kelly & Weaver, 2004) distinguish between component-level evidence and system-level safety claims; the SRI provides component-level evidence rather than system-level sufficiency proofs. Multi-Attribute Utility Theory (MAUT), from which the SRI aggregators derive their semantics, provides a principled decision-theoretic basis for multi-dimensional scoring (Keeney & Raiffa, 1993). Fault tree analysis used here to derive dimension weights is standardized in IEC 61025 and provides Fussell-Vesely importance measures that quantify each basic event's contribution to the top-event probability.

2.3 Uncertainty Quantification and Conformal Prediction

Angelopoulos and Bates (2022) survey conformal prediction as a distribution-free uncertainty quantification tool with marginal coverage guarantees. Gibbs and Candes (2021) extend conformal inference to non-exchangeable sequences via Adaptive Conformal Inference (ACI), which we apply to the SWaT attack distribution. The SRI's Uncertainty dimension draws on calibration theory (Guo et al., 2017), measuring Expected Calibration Error after post-hoc calibration via Platt scaling. Our finding that no model achieves near-target coverage under ACI on the attack out-of-distribution regime confirms that conformal guarantees do not automatically transfer to adversarial inputs.

2.4 Explainability in ICS Anomaly Detection

Integrated Gradients (Sundararajan et al., 2017) provides axiomatic attribution for differentiable models via path integration from a baseline. Prior ICS explainability work has relied primarily on SHAP (Lundberg & Lee, 2017), which encounters numerical issues for LSTM and USAD models due to singular coalition matrices on windowed inputs. We resolve this by constructing a differentiable windowed surrogate for each model and applying a uniform Integrated Gradients pipeline. Critically, prior work has evaluated attribution quality using coverage and stability metrics without validating against ground-truth attack sensor mappings the localization gap exposed in Section 5.

3. THE SAFETY READINESS INDEX FRAMEWORK

3.1 Dimensions and Measurement

The SRI scores each detector on five dimensions, each normalized to the unit interval [0, 1]. The five dimensions, with their measurement procedures, are as follows.

Performance (P) combines detection rate and response latency. Detection rate is computed as 1 minus the Wilson upper-bound false negative rate, applying a conservative confidence interval to penalize uncertain detection estimates from small attack sample sizes. The latency score is max(0, 1 − TTD/3600), where TTD is time-to-detection in seconds, capping at zero benefit for detections exceeding one hour after attack onset. The dimension P is the harmonic mean of these two sub-scores.

Robustness (R) measures event-F1 retention under five ICS-motivated failure modes: additive Gaussian noise, calibration drift (multiplicative bias), constant sensor offset, sensor signal freezing, and missing-data forward-fill imputation. Additionally, mimicry False Data Injection Attacks (FDIA) gradient-free boundary pushes toward the normal operating manifold are evaluated at three severity levels (0.05, 0.10, 0.20). Rows in which the prediction rate exceeds 0.95 (model collapse artifact) are excluded from all quantitative claims and are marked explicitly.

Uncertainty Quantification (U) is calibrated Expected Calibration Error (ECE) computed via Platt scaling and isotonic regression on a disjoint stratified test split: U = 1 − ECE_calibrated. The calibration split is held out from all threshold selection and model training procedures. All five evaluated models receive U = 0 under the corrected procedure, as no model achieves calibration within the acceptable ECE range on the attack out-of-distribution regime.

Explainability (E) is computed via Integrated Gradients through a windowed differentiable surrogate g: R^(60 x 51) to R for each model. Surrogate fidelity R^2 is reported for each model; models with $R^2 < 0.90$ are flagged and commensurability with other models is not claimed. E_uniform combines attribution coverage (top-5 sensor attribution mass share) and attribution stability (mean pairwise cosine similarity across attack events). The localization-adjusted score E_localized multiplies E_uniform by the localization factor max(2 × AUROC − 1, 0), where AUROC is computed against ground-truth attacked sensor labels.

Fault Tolerance (F) is normalized Event-F1 retention under importance-ordered iterative sensor dropout, where sensors are zeroed in order of decreasing Integrated Gradient attribution magnitude. The score is the area under the F1-vs.-k curve normalized by a reference degradation threshold. The dimension captures a model's ability to maintain detection capability under partial sensor loss, directly relevant to ICS environments where remote I/O rack failures can eliminate entire sensor groups.

3.2 Weights and Aggregators

Clause-count weights derived from the density of requirements across IEC 62443 and NIST SP 800-82 sections set P = 0.345, R = 0.230, U = 0.165, E = 0.130, F = 0.130. We additionally derive Fussell-Vesely importance weights from a fault tree grounded in the ICS attack scenario where detection failure constitutes a mitigation barrier failure. At nominal failure probabilities, FV weights are P = 0.265, R = 0.221, U = 0.155, E = 0.138, F = 0.221. The most significant difference is Fault Tolerance, which nearly doubles (0.130 → 0.221) because sensor dropout causing detection failure constitutes a single-event minimal cut set in the fault tree: one component failure alone causes the top event.

Three aggregators are defined with explicit decision-theoretic semantics from MAUT. The Weighted Linear Sum (WLS) is the additive aggregator, assuming additive independence: any dimension can compensate for any other. The Weighted Geometric Mean (GEO) assumes mutual utility independence: a zero in any dimension drives the aggregate to zero, implementing a non-compensatory zero floor. The Rawlsian Minimum (MIN) scores a model solely by its worst dimension, implementing a maximin welfare function with no inter-dimensional compensation whatsoever.

3.3 A Structural Impossibility Result

We prove via symbolic computation (SymPy) that any aggregator A simultaneously satisfying two axioms is unsatisfiable. Axiom A1 (strict Pareto-monotonicity): if dimension profile q strictly dominates profile p in every coordinate, then A(q) > A(p). Axiom A2 (zero-floor): if any component of q equals zero, then A(q) = 0. The contradiction arises as follows: consider q = (0, 1, 1, 1, 1) and p = (0, 0, 0, 0, 0). A2 forces A(q) = 0 and A(p) = 0. But q strictly dominates p in dimensions R through F, so A1 forces A(q) > A(p) = 0, yielding 0 > 0, a contradiction. This impossibility result motivates explicit aggregator selection: practitioners must choose between full compensation (WLS) and zero-penalization (GEO/MIN) as an explicit value decision, not a technical default.

This result is a direct corollary of the two stated axioms rather than a novel general impossibility theorem; it is best read as a specific instance of the broader family of impossibility results in multi-criteria decision analysis and social choice theory, most notably Arrow's impossibility theorem and related no-free-lunch results contrasting compensatory and non-compensatory aggregation rules (Keeney & Raiffa, 1993), which likewise show that no aggregation rule can simultaneously satisfy an unrestricted set of intuitively desirable axioms.

4. EXPERIMENTAL SETUP

4.1 Dataset

The Secure Water Treatment (SWaT) dataset (Mathur & Tippenhauer, 2016) records 51 sensors and actuators across six process stages (P1–P6) of a water treatment plant at one-second resolution. Normal partition: 496,800 timesteps (approximately 7 days). Attack partition: 449,919 timesteps containing 9 documented attack events across 11 attack scenarios targeting sensors and actuators at various stages. The

standard evaluation split uses 60% of normal data for training, 20% for validation, and combines the remaining 20% of normal data with the full attack partition for testing, yielding a test attack prior of 40.9%, noted as ecologically invalid and addressed in Section 5.3.

4.2 Models

All five models are unsupervised, trained on normal-condition data only, with no access to attack labels during training. Isolation Forest uses 100 trees with subsampling fraction 0.1. One-Class SVM uses a radial basis function kernel with $\nu = 0.05$. XGBoost is a lag-1 next-step predictor with 200 boosting rounds; the anomaly score is the per-timestep prediction residual mean squared error across all 51 sensors. All three classical-model hyperparameters were fixed at these values rather than tuned via the validation split; no validation-based hyperparameter search was performed for the reported results. We note this as a factor that could affect the fairness of the robustness comparison, particularly for XGBoost, whose lag-1 forecasting formulation is the most sensitive of the five models to boosting-round count and could plausibly exhibit different robustness characteristics under a tuned configuration. LSTM Autoencoder uses a 60-timestep sliding window with two LSTM encoder layers (64, 32 hidden units) and symmetric decoder. USAD uses the same 60-timestep window with two adversarially trained autoencoder heads. All neural models were trained with Adam optimizer on a single GPU and frozen at checkpoint after validation loss convergence.

4.3 Evaluation Protocol

All evaluation is performed on the frozen test set without any form of look-ahead or threshold re-optimization on test data. Anomaly thresholds are set as the 99th percentile of the validation-set anomaly score distribution. Conformal prediction quantiles are fit on a disjoint held-out split. Calibration ECE is computed on a disjoint stratified 50/50 partition of the test set. Attribution localization is evaluated against the 9 documented attack-to-sensor mappings extracted from the SWaT attack documentation, with one domain-inferred mapping (Attack 7, P3 stage) recorded with explicit provenance. All model collapse cells (prediction rate > 0.95 artifact) are explicitly identified and excluded from quantitative claims. The SRI is computed under both clause-count and Fussell-Vesely weights, and rank stability is assessed via 5,000 Monte Carlo draws from the FV weight uncertainty intervals.

5. RESULTS

5.1 Aggregation Rule Determines the Winner

Figure 1 presents SRI scores under WLS, GEO, and MIN aggregators for all five models. The headline finding is that the top-ranked model changes depending on the aggregation rule. Under WLS, USAD leads (0.718) followed by LSTM-AE (0.642), OneClassSVM (0.614), IsolationForest (0.601), and XGBoost (0.512). Under GEO, the ranking changes markedly: USAD (0.660) and LSTM-AE (0.539) retain their positions, but OneClassSVM collapses to 0.037 because SRI-F = 0 drives the geometric mean to near-zero, and XGBoost collapses to 0.003 because SRI-R = 0. IsolationForest (0.504) rises to third under GEO as the only model with no zero dimension. The MIN aggregator reveals the most severe

finding: three of five models score exactly zero, meaning each has at least one safety dimension where performance is effectively absent. Only USAD (0.267), LSTM-AE (0.259), and IsolationForest (0.248) have nonzero MIN scores. XGBoost's SRI-R = 0 and OC-SVM's SRI-F = 0 produce zero MIN regardless of other dimension values. The Spearman rank correlation between WLS and GEO rankings is +0.90 on the corrected canonical data—high overall but masking the catastrophic individual collapses.

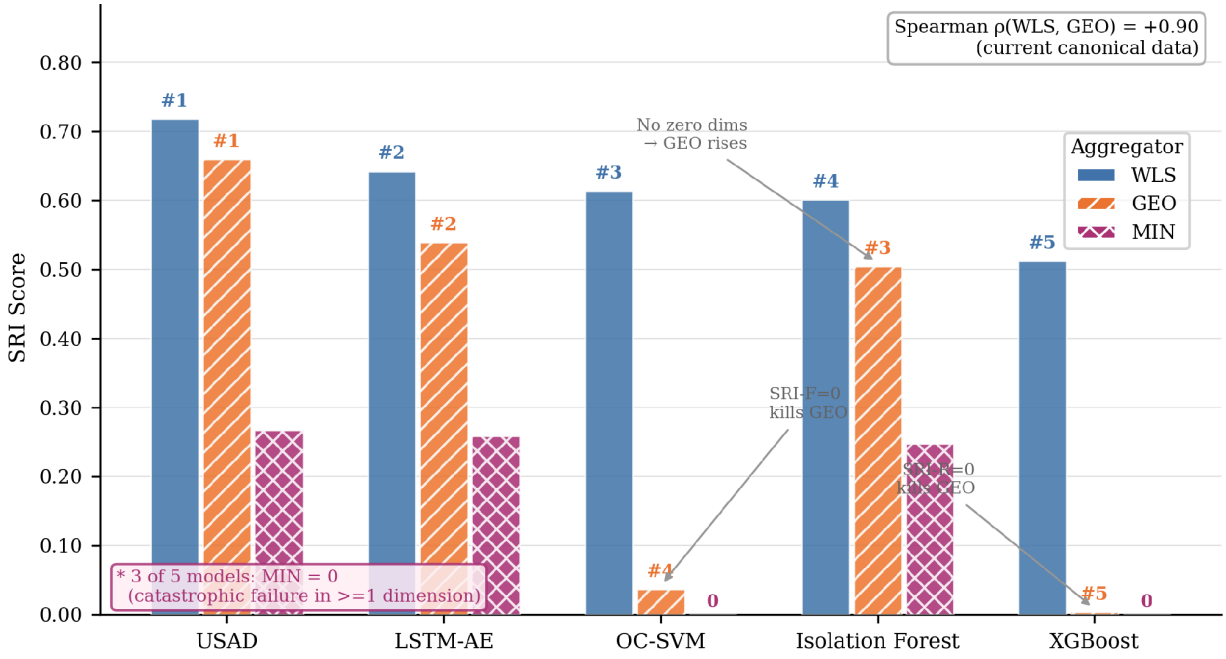

Figure 1. *SRI scores under WLS (additive), GEO (geometric, non-compensatory), and MIN (Rawlsian worst-case) aggregators. Rank numbers are shown above each bar. OneClassSVM falls from WLS rank 3 to GEO rank 4 because SRI-F = 0 collapses the geometric mean. XGBoost GEO score collapses to 0.003 because SRI-R = 0. Three of five models have MIN = 0, indicating catastrophic failure in at least one dimension. Spearman ρ(WLS, GEO) = +0.90 on corrected canonical data.*

5.2 Robustness Under Adversarial Perturbation

Figure 2 presents Event-F1 degradation under mimicry FDIA across tested severity levels. XGBoost, the model with the highest event-F1 under standard evaluation (baseline F1 = 0.798), collapses to F1 = 0.040 at severity 0.05, a 95.0% drop and does not recover at higher severity levels. This collapse occurs because the lag-1 forecasting approach is uniquely vulnerable to smooth, bounded perturbations: a small but targeted push toward the normal operating manifold renders the attack indistinguishable from normal-operating-condition residuals.

Deep reconstruction models demonstrate substantially greater mimicry resistance. USAD retains 93.4% of baseline F1 at severity 0.05 (F1 = 0.458), and LSTM-AE retains 90.4% (F1 = 0.385). IsolationForest shows intermediate degradation (24.0% retention, F1 = 0.080) and OneClassSVM retains 40.8% (F1 =

0.206). Of 90 total robustness evaluation rows across all failure modes, 46 exhibited model collapse (prediction rate > 0.95 artifact) and were excluded from quantitative claims; all mimicry-FDIA rows shown in Figure 2 are valid.

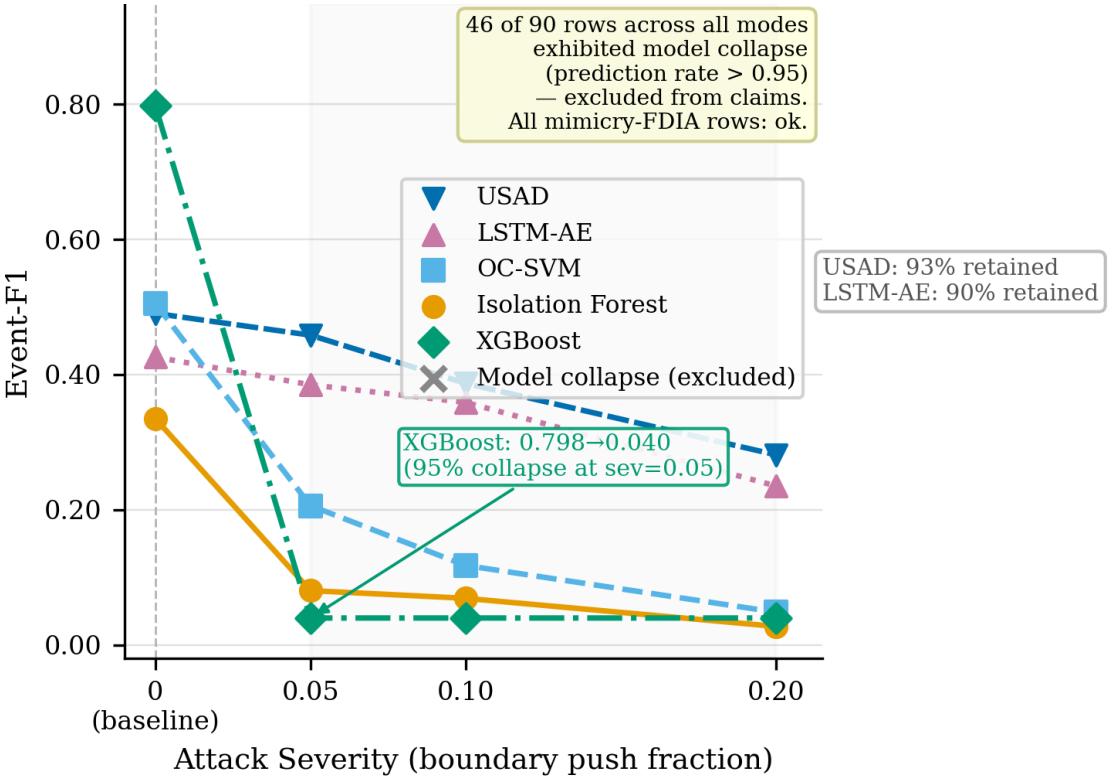

Figure 2. *Event-F1 under mimicry FDIA (gradient-free boundary push toward the normal operating manifold) at tested severity levels 0.05, 0.10, and 0.20. XGBoost baseline F1 = 0.798 collapses to 0.040 at severity 0.05 (95.0% drop) and does not recover. USAD and LSTM-AE retain approximately 90% of baseline F1 at severity 0.05. All 15 mimicry-FDIA rows have artifact_flag = ok; 46 of 90 rows across all modes exhibited model collapse and were excluded.*

5.3 Operational Deployment Realism

Figure 3 quantifies the base-rate fallacy inherent in the standard SWaT evaluation. The left panel (Figure 3a) plots Precision (PPV) against the log-scale attack prior at an operating point constrained to one false alarm per day. At the test prior of $\pi = 0.409$, all models appear near-perfect because the Bayesian denominator is dominated by the artificially high prior. At a realistic operational prior of $\pi = 10^{-6}$, XGBoost precision collapses to 7.4×10^{-5} fewer than one true alarm per ten thousand alerts—while the best-performing models (USAD, OneClassSVM, LSTM-AE) reach a maximum PPV of approximately 0.045.

The right panel (Figure 3b) constrains the operating point to at most one false alarm per day and plots True Positive Rate (TPR) for three false-alarm budgets (1/hour, 1/day, 1/week). XGBoost TPR under this constraint is 0.0009, virtually zero, while USAD, LSTM-AE, and OneClassSVM maintain TPR ≈ 0.59 across all three budget levels. IsolationForest occupies an intermediate position (TPR = 0.038). This reversal from F1 rank 2 to TPR rank 5 under a realistic false-alarm constraint—is an operationally significant finding invisible under standard F1 reporting.

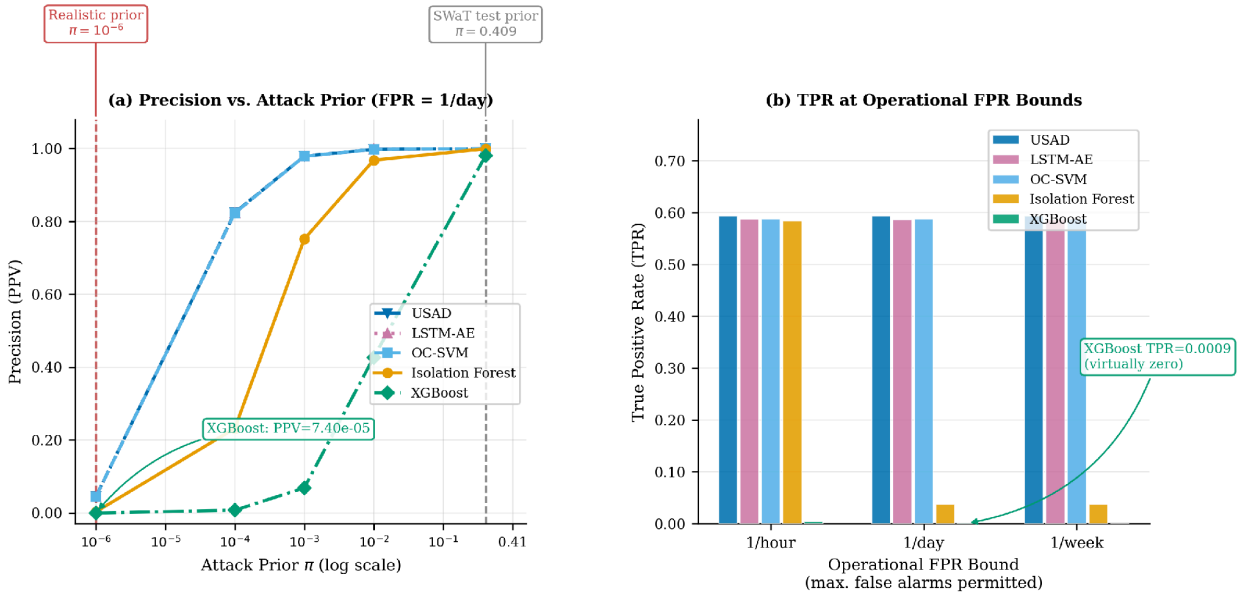

Figure 3. *(a) Precision (PPV) vs. attack prior $\log_{10}(\pi)$ at one false alarm per day operating point. At the SWaT test prior (π = 0.409) all models appear near-perfect. At realistic prior $\pi = 10^{-6}$, XGBoost PPV collapses to 7.4×10^{-5} while best models reach 0.045. (b) TPR at three operational FPR bounds. XGBoost TPR = 0.0009 at 1/day (virtually zero); USAD, LSTM-AE, and OneClassSVM maintain TPR ≈ 0.59 across all three bounds.*

5.4 Fault Tolerance Under Sensor Dropout

Figure 4 presents fault tolerance results under two dropout protocols. The left panel (Figure 4a) shows a heatmap of F1 retention under stage-group dropout, where all sensors in each of SWaT's six process stages (P1–P6) are simultaneously zeroed to simulate a remote I/O rack loss. Of 30 total cells (5 models × 6 stages), 14 exhibit model collapse (prediction rate > 0.95 artifact), shown as X-marked cells and excluded from all quantitative claims.

IsolationForest is the most fault-tolerant, surviving all six stage-group dropouts without collapse and maintaining F1 between 0.186 and 0.401. Deep models (LSTM-AE, USAD) collapse when sensor-dense P4 (9 sensors) or P5 (13 sensors) stages are lost, likely because the windowed reconstruction objective requires coherent multi-sensor trajectories. OneClassSVM collapses under five of six stage groups, surviving only P6 (3 sensors, 5.9% of the feature space). The right panel (Figure 4b) shows iterative

dropout over a Fibonacci-spaced grid of nine k values (k = 1, 2, 3, 5, 8, 13, 21, 34, 51) applied to each of the five models, yielding 45 (model, k) cells in total. OneClassSVM collapses at k = 3 sensors (5.9% of 51), LSTM-AE and USAD at k = 5 (9.8%), XGBoost at k = 13 (25.5%), and IsolationForest at k = 21 (41.2%)—providing operational thresholds for minimum sensor-set requirements.

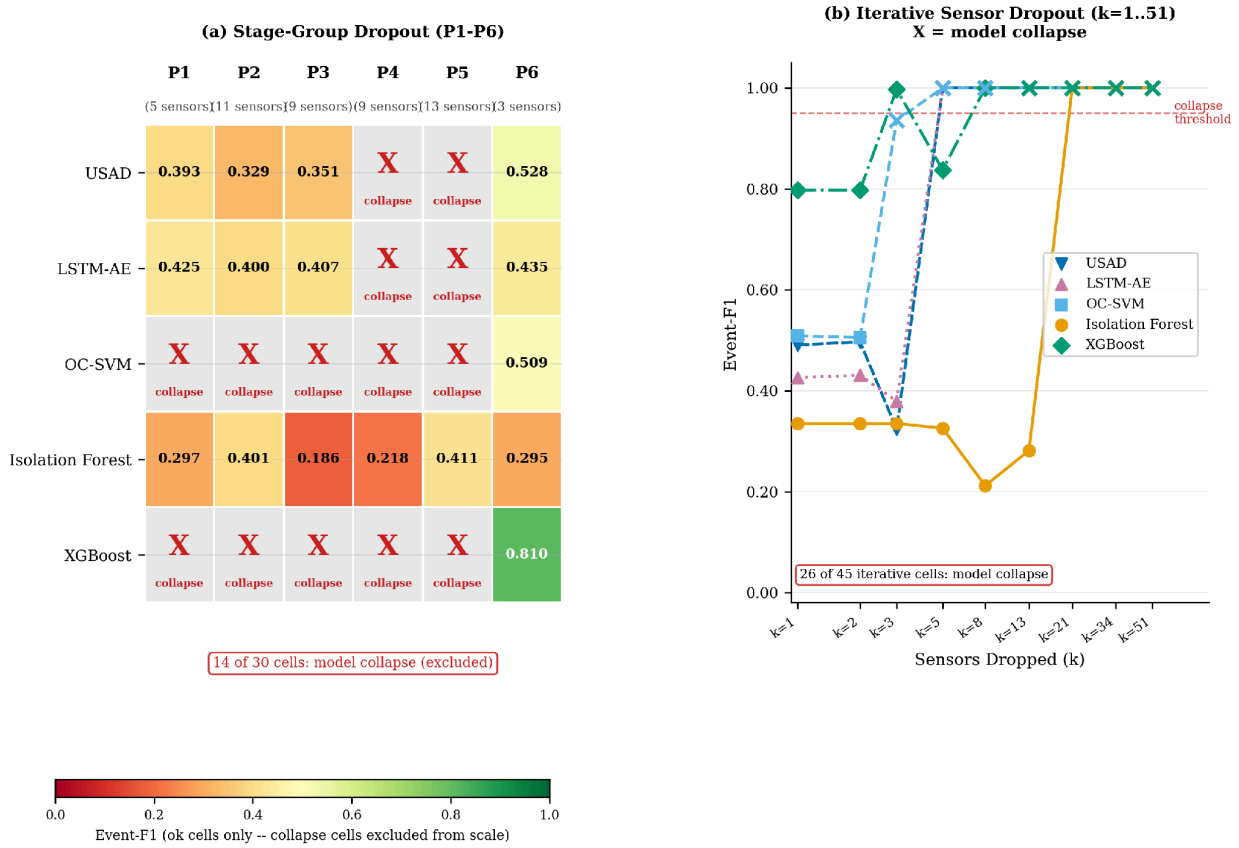

Figure 4. *(a) Stage-group dropout: F1 when all sensors in each SWaT stage (P1–P6) are zeroed. X = model collapse (prediction rate > 0.95 artifact; 14 of 30 cells). IsolationForest survives all stages; deep models collapse when sensor-dense stages are lost. (b) Iterative sensor dropout (log₂ scale). OneClassSVM collapses at k = 3 (5.9% of 51 sensors); LSTM-AE and USAD at k = 5 (9.8%); IsolationForest survives to k = 21 (41.2%). 24 of 45 iterative cells exhibit model collapse.*

5.5 Attribution Localization Failure

Figure 5 presents the primary explainability finding. We applied a single Integrated Gradients pipeline through a windowed differentiable surrogate (window W = 60 timesteps, 51 sensors) to all five models, producing 51-sensor attribution vectors for each of 9 attack events. Surrogate fidelity exceeded $R^2 = 0.988$ for four of five models; XGBoost surrogate fidelity was $R^2 = 0.794$, flagged as below the 0.90 threshold, and commensurability with other models is not claimed for that row.

The localization results are uniformly negative. For all five models, hit@1 = 0, no model correctly identifies the primary attacked sensor as its top attribution for any of the 9 attack events. AUROC values range from 0.548 (USAD) to 0.633 (XGBoost), marginally above the 0.5 chance baseline but without statistical significance given the 9-event sample. The bottom panel (Figure 5b) shows the consequence for SRI-E: E_uniform scores (0.548–0.606) collapse to E_localized values of 0.057–0.161 after multiplying by the localization factor max(2 × AUROC − 1, 0), representing an approximately 83% reduction in claimed explainability.

This finding validates a fundamental criticism of attribution-based explainability in anomaly detection: stability across attack events and top-k mass concentration do not guarantee that attributed sensors correspond to ground-truth manipulated sensors.

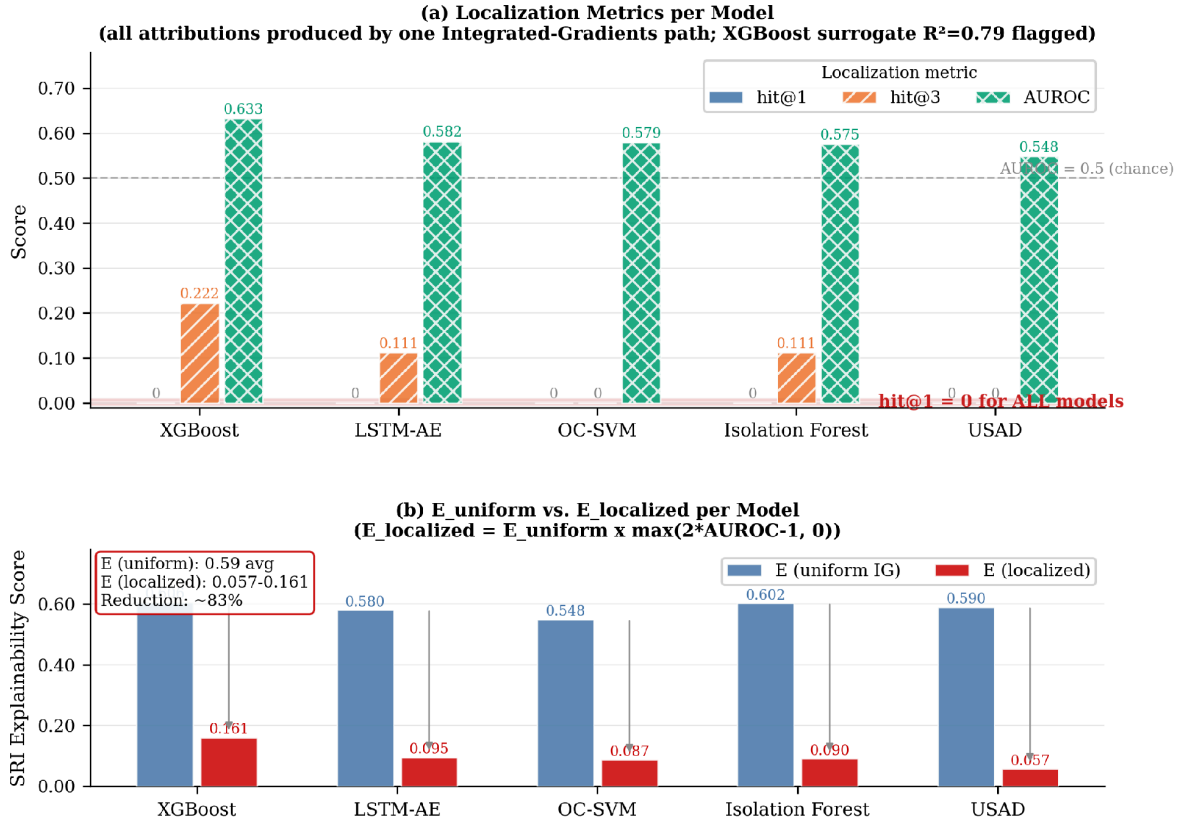

Figure 5. *(a) Attribution localization metrics per model, scored against ground-truth attacked sensors (8 explicit + 1 domain-inferred; all 9 events mapped). hit@1 = 0 for all five models; AUROC ranges 0.548–0.633 (marginally above 0.5 chance baseline). (b) SRI Explainability score before and after localization adjustment. E_uniform (0.55–0.61) collapses to E_localized (0.057–0.161) after incorporating localization factor max(2×AUROC − 1, 0). Mean reduction: approximately 83%. XGBoost surrogate R² = 0.79 is flagged; result is consistent with high-fidelity models.*

5.6 Weight Sensitivity and Rank Stability

Figure 6 presents the weight sensitivity analysis. The left panel (Figure 6a) compares clause-count and Fussell-Vesely weights across all five dimensions. The most significant shift is FaultTolerance, which increases from 0.130 (clause-count) to 0.221 (FV) because the sensor-dropout basic event constitutes a single-event minimal cut set in the fault tree. Performance decreases from 0.345 to 0.265 because its fault-tree basic event shares a gate with other barriers.

The right panel (Figure 6b) shows P(rank 1) across 5,000 Monte Carlo weight draws from the FTA uncertainty intervals (seed 42). Under WLS with clause-count weights and nominal FV weights, USAD leads in the nominal evaluation (WLS = 0.718). However, under weight uncertainty, LSTM-AE has higher P(rank 1) at 0.615 compared to USAD's 0.384, reflecting that the two models are statistically tied: the 95% confidence interval for the WLS score difference (USAD − LSTM-AE) is [−0.117, +0.080], which includes zero. Under GEO, USAD leads with P = 0.634; under MIN, USAD always leads (P = 1.000) because its minimum dimension score is consistently the highest. Results are reproducible across seeds 42 and 7, with a maximum P(rank 1) discrepancy of 0.016.

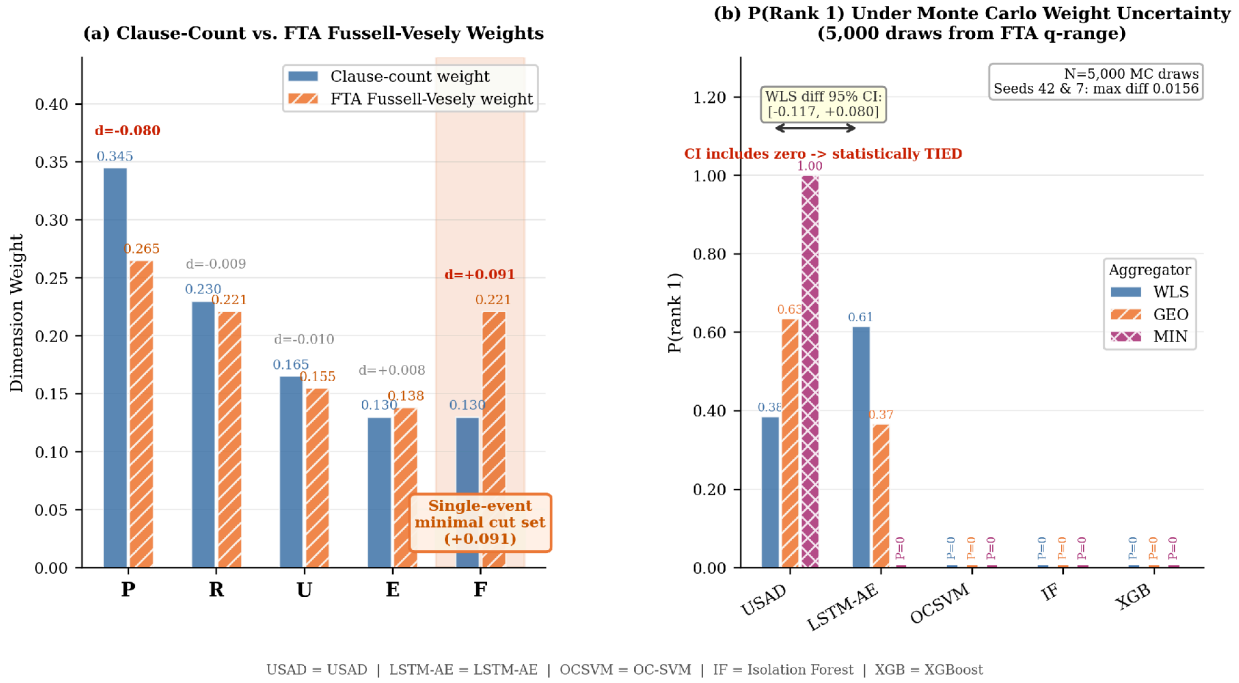

Figure 6. *(a) Clause-count vs. FTA Fussell-Vesely dimension weights. FaultTolerance nearly doubles (0.130 → 0.221) because the sensor-dropout basic event is a single-event minimal cut set. (b) P(rank 1) under 5,000 Monte Carlo weight draws. Under WLS, LSTM-AE leads (P = 0.615) but is statistically tied with USAD (95% CI [−0.117, +0.080] includes zero). Under GEO, USAD leads (P = 0.634); under MIN, USAD always ranks first (P = 1.000). Results reproducible across seeds 42 and 7 (max discrepancy 0.016).*

6. DISCUSSION

6.1 The Aggregation Problem is Primary

The most consequential finding of this study is that the choice of aggregation rule has a first-order effect on model selection, comparable to or exceeding the effect of model architecture differences. This is not merely a sensitivity concern: four of five evaluated models score exactly zero on at least one safety dimension, producing stark GEO and MIN penalties that WLS masks through cross-dimensional compensation. Practitioners who default to WLS because it is computationally convenient are implicitly assuming that any safety deficit in one dimension can be fully compensated by excellence in another, an assumption that is philosophically questionable in safety-critical deployment and formally undefended.

The structural impossibility result (Section 3.3) makes explicit what the aggregation sensitivity analysis implies: no aggregator can simultaneously satisfy strict Pareto-monotonicity and a zero-floor penalty. Practitioners must choose. We recommend: (1) GEO as the primary aggregator when zero-tolerance for complete failure in any safety dimension is required; (2) WLS as a secondary view for understanding relative strengths and weaknesses; (3) MIN as an alert mechanism for detecting catastrophic single-dimension failure; and (4) the Pareto frontier as the assumption-free view when no aggregator can be defended on principled grounds for the deployment context.

6.2 The Operational Realism Gap

XGBoost's collapse under realistic deployment conditions illustrates a recurring pattern in machine learning safety benchmarking: a model carefully optimized for the benchmark distribution, here, a 40.9% attack prior with no evasion mechanism can fail catastrophically when benchmark assumptions are relaxed. The 95.0% F1 collapse under mimicry FDIA (Section 5.2) and the effective TPR = 0 under one-false-alarm-per-day constraints (Section 5.3) are not edge cases; they characterize the realistic deployment regime for continuous monitoring systems. The SRI's Robustness dimension partially captures this through structured adversarial evaluation, but the base-rate realism finding argues for making the operational attack prior explicit in all ICS anomaly detection benchmark protocols.

6.3 The Explainability Illusion

The attribution localization failure is perhaps the most practically alarming result. Current explainability evaluation in anomaly detection typically measures how concentrated and how stable attribution vectors are, properties that are necessary but not sufficient for genuine interpretability. The ground-truth localization study demonstrates that attribution mass can be concentrated and stable ($E_{uniform} \approx 0.59$) while pointing to entirely wrong sensors. An operator responding to an explainability output that confidently highlights sensor X when sensor Y is actually being manipulated is worse positioned than an operator who knows the system cannot explain its detections. The $E_{localized}$ metric proposed here provides a principled correction, but the 83% reduction in claimed explainability that it produces means that no current unsupervised model meets a reasonable explainability threshold for autonomous ICS monitoring.

6.4 Limitations

Several limitations constrain the generalizability of these findings. First, all results derive from a single dataset (SWaT) with 9 documented attack events. The statistical power for ranking claims is correspondingly limited: bootstrap confidence intervals on pairwise event-F1 differences can span up to 0.67. Second, deep model results depend on frozen checkpoints that cannot be exactly reproduced from training seed alone due to GPU nondeterminism; mean WLS variation across re-training runs is 0.069 (LSTM-AE) and 0.137 (USAD). Third, the attribution localization study relies on a domain-inferred sensor mapping for one of 9 attack events, recorded with explicit provenance but constituting an assumption. Fourth, the expert elicitation component envisioned for SRI threshold calibration (deploy/no-deploy thresholds) remains unimplemented, as a formal expert panel study was outside the scope of this work.

7. CONCLUSION

We introduced the Safety Readiness Index, a five-dimensional multi-criteria framework for evaluating machine learning-based ICS anomaly detectors against safety-relevant requirements from IEC 62443 and NIST SP 800-82. Our evaluation of five unsupervised detectors on the SWaT benchmark yields four findings that contradict the prevailing benchmark narrative. First, the aggregation rule not model capability, determines the top-ranked model: the F1 leader collapses to last place under a non-compensatory aggregator because its Robustness score is zero. Second, the event-F1 leader collapses under both adversarial perturbation (95.0% drop) and realistic false-alarm constraints (TPR ≈ 0.001 at one alarm per day). Third, attribution-based explainability metrics overstate interpretability by approximately 83% when evaluated against ground-truth attacked sensors, and no model achieves hit@1 > 0 across any of the 9 documented attack events. Fourth, four of five evaluated models catastrophically fail at least one safety dimension, a fact invisible under the compensatory WLS aggregator but revealed immediately by GEO or MIN.

These findings collectively argue for a fundamental reorientation of ICS anomaly detection evaluation: from single-metric accuracy benchmarking toward multi-criteria safety assurance assessment that explicitly addresses robustness, calibration, localization-validated explainability, and fault tolerance. The SRI framework provides an operational structure for this reorientation, with the critical acknowledgment that the framework's own current findings carry important caveats, particularly the single-dataset limitation, that future work must address through multi-facility evaluation, expert elicitation for threshold calibration, adversarial evaluation with physically grounded attack models, and extension to supervised and semi-supervised detection paradigms.

DATA AND CODE AVAILABILITY STATEMENT

The SWaT dataset used in this study is not freely downloadable; it is available upon request from iTrust, Centre for Research in Cyber Security, Singapore University of Technology and Design (SUTD), subject to completion of a data-sharing agreement (https://itrust.sutd.edu.sg/itrust-labs_datasets/dataset_info/). The analysis code used to produce the results reported in this paper is available from the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS

A. Warwatkar conceived the study, implemented the experimental pipeline, conducted the analysis, and wrote the manuscript. S. P. C. Balla supervised the project as mentor, providing guidance on the experimental design, methodology, and manuscript revisions.

FUNDING

This research received no external funding.

REFERENCES

1. Angelopoulos, A. N., & Bates, S. (2022). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv:2107.07511.
2. Audibert, J., Michiardi, P., Guyard, F., Marti, S., & Zuluaga, M. A. (2020). USAD: UnSupervised Anomaly Detection on multivariate time series. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 3395–3404).
3. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).
4. Gibbs, I., & Candes, E. (2021). Adaptive conformal inference under distribution shift. Advances in Neural Information Processing Systems, 34, 1660–1672.
5. Goh, J., Adepu, S., Junejo, K. N., & Mathur, A. (2016). A dataset to support research in the design of secure water treatment systems. In Proceedings of the International Conference on Critical Information Infrastructures Security (LNCS 10242, pp. 88–099). Springer.
6. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning (pp. 1321–1330).
7. IEC 62443. (2018). Industrial automation and control systems security, Parts 3-3, 4-1. International Electrotechnical Commission.
8. Keeney, R. L., & Raiffa, H. (1993). Decisions with Multiple Objectives: Preferences and Value Trade-offs. Cambridge University Press.
9. Kelly, T. P., & Weaver, R. (2004). The goal structuring notation: A safety argument notation. Workshop on Dependable and Reusable Safety Cases, DSN 2004.

10. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (pp. 413–422).
11. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.
12. Malhotra, P., Vig, L., Shroff, G., & Agarwal, P. (2015). Long short term memory networks for anomaly detection in time series. Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2015).
13. Mathur, A. P., & Tippenhauer, N. O. (2016). SWaT: A water treatment testbed for research and training on ICS security. Proceedings of the 2016 International Workshop on Cyber-physical Systems for Smart Water Networks (CySWater), IEEE.
14. NIST SP 800-82r3. (2023). Guide to operational technology (OT) security. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-82r3
15. Scholkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. Neural Computation, 13(7), 1443–1471.
16. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. Proceedings of the 34th International Conference on Machine Learning (ICML 2017, pp. 3319–3328).