ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Cost-Aware Detector Selection for Multivariate Time-Series Anomaly Detection: The CADR Conceptual Framework

Sarthak Sunil Maral, S. Charishma, Abhinandana A. Poojary, A. N. Pawar

Department of Artificial Intelligence and Machine Learning Engineering, Navsahyadri Education Society's Group of Institutes (Savitribai Phule Pune University), Pune, Maharashtra, India

Published November 2026

Abstract

Adaptive model selection for multivariate time-series anomaly detection is addressed by existing frameworks such as RAMSES, mTSBench, and MSAD. However, many existing methods assume a stationary data distribution, immediate ground-truth feedback, and sufficient computational resources to evaluate or execute multiple detectors simultaneously. Real-world monitoring environments rarely satisfy these assumptions: operating regimes shift over time, operator confirmation arrives with delay, labeled anomalies are sparse or absent, and strict latency and compute budgets constrain inference. This paper synthesizes existing adaptive-selection literature against five practical operational criteria and identifies six interconnected deployment gaps. To address these issues jointly, we propose CADR (Cost-Aware, Drift-Adaptive, Feedback-Limited Detector Routing), a conceptual framework that dynamically selects or weights anomaly detectors over streaming multivariate time series. CADR coordinates six modules: a complementary detector library, a context- and regime-aware drift monitor, a two-level cost-aware router, a feedback-limited learning module, an online score calibration module, and a hysteresis-stabilized decision policy. We formalize the joint selection objective, detail the modular architecture, and outline a rigorous time-ordered experimental protocol to evaluate CADR against standard baselines and an oracle upper bound.

Keywords: Multivariate time-series anomaly detection; adaptive model selection; concept drift; cost-aware routing; delayed feedback; online score calibration

1 Introduction

1.1 Background

Multivariate time-series streams are continuously generated across modern operational infrastructures, including industrial plants, cyber-physical systems, and cloud computing platforms. At each time step, sensors or telemetry agents record multiple correlated measurements, and anomaly detection mechanisms are tasked with identifying patterns that deviate significantly from expected operational behavior [3]. Recent surveys of deep anomaly detection for multivariate time series highlight a broad spectrum of available detector architectures [4]. Because individual detector families exhibit distinct inductive biases and excel on different anomaly types, a monitoring platform can maintain a diverse library of detectors and dynamically select or weight them over time.

Adaptive model selection for time-series anomaly detection has gained attention through recent frameworks such as RAMSES [5], mTSBench [10], and MSAD [2]. Despite promis- ing benchmark performance, a recurring limitation in existing adaptive selection approaches is the reliance on idealized

operational assumptions: stationary data distributions, immediate label feedback after predictions, and ample compute budgets to evaluate multiple candidate detectors concurrently. Relaxing these assumptions is critical for dependable real-world deployment.

1.2 Problem statement and motivation

Operational monitoring systems contend with non-stationary environments, delayed human verification, scarce anomaly labels, and stringent latency or compute constraints. The joint problem investigated in this work is defined as follows: given an incoming stream of multivariate time-series observations and a heterogeneous library of anomaly detectors, select or weight the detectors sequentially over time such that detection accuracy, false-alarm rates, adaptation speed, and computational overhead remain acceptable, even when feedback is delayed, sparse, or entirely absent.

Four practical observations motivate this joint formulation:

- **No single detector dominates across contexts:** Individual detectors specialize in distinct anomaly structures (e.g., sudden point spikes versus subtle temporal or multivariate correlation shifts), causing detector rankings to vary substantially across domains and temporal regimes.
- **Non-stationary data behavior:** Monitoring systems must distinguish between planned normal operating-regime transitions (such as day-to-night cycles, seasonal variations, or scheduled maintenance), suspicious concept drift (unexplained distributional shifts indicating sensor degradation, faults, or cyber-physical attacks), and localized anomalies within a known regime.
- **Delayed, sparse, or absent feedback:** Assuming immediate ground-truth labels at every time step is unrealistic in production environments where operator triage is slow and costly.
- **Computational cost and switching instability:** Executing all detectors at every step violates online latency budgets, raw anomaly scores across heterogeneous models lie on incompatible scales, and erratic detector switching erodes operator trust.

1.3 Contributions and paper organization

This work presents a literature-grounded conceptual formulation rather than an empirical evaluation of a completed implementation. Its primary contributions are:

1. A structured comparative synthesis of existing adaptive model-selection paradigms against five operational deployment criteria.
2. Identification of six interconnected operational gaps that prevent standalone model-selection techniques from succeeding in constrained streaming settings.
3. Formulation of CADR (Cost-Aware, Drift-Adaptive, Feedback-Limited Detector Routing), integrating six coordinated modules under a unified mathematical objective $\pi_t = f(z_t, h_t, c_t)$.
4. Design of a leak-free, time-ordered experimental evaluation protocol covering baselines, feedback regimes, temporal drift scenarios, cross-domain generalization, and component ablations.

The remainder of this paper is organized as follows. the section 'Related work' reviews relevant literature on adaptive model selection, detector families, and concept drift. the section 'Related work'I presents a comparative analysis of existing selection strategies. the section 'Identified research gaps' identifies core research gaps. the section 'Proposed framework: CADR' details the formal problem definition and the six modules of CADR. the section 'Proposed framework: CADR'I

outlines the planned experimental design and expected outcomes, and the section 'Proposed framework: CADR'II concludes the paper.

2 Related work

2.1 Adaptive model selection frameworks

Three recent studies establish the foundation for model selection in time-series anomaly detection. Abdelmaksoud et al. [5] introduced RAMSES, focusing on robust and adaptive model selection across time-series anomaly detectors. Zhou et al. [10] developed mTSBench to benchmark multivariate time-series anomaly detection and automated model selection at scale, highlighting trade-offs between detector accuracy and computational overhead. Sylligardos et al. [2] presented MSAD, offering an in-depth empirical investigation of model selection strategies for time-series anomaly detection. While these works confirm that dynamic selection outperforms static single-model baselines, they do not jointly resolve regime-aware drift separation, delayed or absent operator feedback, online score calibration, compute-bounded routing, and switching stability.

2.2 Detector families and behavioral diversity

Prior surveys [3], [4] categorize anomaly detectors across a spectrum of modeling complexity and computational expense:

- **Statistical detectors:** Z-score and robust median/IQR detectors rapidly capture sudden univariate point deviations at negligible computational cost.
- **Subspace and projection methods:** Principal Component Analysis (PCA) detectors identify disruptions in linear multivariate correlations at low computational overhead.
- **Isolation-based methods:** Isolation Forest [6] isolates anomalous observations in high-dimensional feature space with low-to-moderate compute requirements.
- **Forecasting-based detectors:** Temporal predictors flag deviations between forecasted and observed trajectories at moderate inference cost.
- **Deep reconstruction and graph architectures:** LSTM encoder-decoder networks [7] and directed hypergraph neural networks [8] capture complex nonlinear temporal and structural dependencies at high computational cost.

Furthermore, benchmark testbeds such as the Secure Water Treatment (SWaT) dataset [9], [10] illustrate the complexity of cyber-physical multivariate streams where single detectors struggle to cover all attack vectors.

2.3 Concept drift and streaming adaptation

Handling evolving data distributions draws directly from streaming concept-drift research. Gama et al. [12] provide a comprehensive taxonomy of concept drift adaptation, while early drift detection methods [8] and adaptive windowing techniques such as ADWIN [9] enable online tracking of distributional shifts. Subsequent studies have extended drift detection to general streaming data [1] and unsupervised multi-label streams [7]. Table 1 summarizes the reviewed literature and its relationship to the proposed framework.

Table 1. Summary of reviewed literature and relevance to CADR.

Author and year	Approach / focus	Key findings and limitations relative to joint deployment
Abdelmaksoud et al., 2026 [3]	RAMSES: Robust and adaptive model selection for time-series anomaly detection	Demonstrates adaptive selection gains, but typical assumptions regarding distribution stability, feedback timing, or compute budgets constrain real-world deployment.
Zhou et al., 2026 [4]	mTSBench: Benchmarking multivariate time-series anomaly detection	Benchmarks selection at scale; highlights that uncalibrated or heavy detector setups incur sharp accuracy-versus-compute trade-offs across contexts.
Sylligardos et al., 2025 [5]	MSAD: Deep dive into model selection for time-series anomaly detection	Validates model selection efficacy, yet does not jointly address drift-regime separation, delayed/missing feedback, calibration, and switch stability.
Wang et al., 2025 [2]; Chandola et al., 2009 [1]	Surveys of general and deep multivariate time-series anomaly detection	Establish taxonomy of detector families and anomaly types; focus on detection architectures rather than cost-aware, feedback-limited routing.
Liu et al., 2008 [6]; Malhotra et al., 2016 [7]; Ha, 2025 [8]	Isolation Forest, LSTM encoder-decoder, and directed hypergraph neural networks	Provide complementary low-, medium-, and high-cost detector candidates for a diverse library; represent standalone detectors rather than selection policies.
Goh et al., 2017 [9]; Inoue et al., 2017 [10]	SWaT testbed dataset and unsupervised anomaly detection on SWaT	Establish a realistic cyber-physical multivariate benchmark planned for CADR development and evaluation.
Gama et al., 2004 [12]; Gama et al., 2014 [11]; Bifet & Gavaldà, 2007 [13]; Wang & Abraham, 2015 [14]; Gulcan & Can, 2022 [15]	Concept drift detection, adaptive windowing (ADWIN), and unsupervised stream drift	Provide foundational mechanisms for sliding-window drift monitoring without immediate labels, though not tailored to anomaly detector routing.

3 Comparative analysis of existing approaches

3.1 Baseline and adaptive selection strategies

To contextualize the design requirements of CADR, seven representative detector selection paradigms are examined:

- **Best Fixed Single Detector:** Deploys a single static detector across the entire stream.
- **Random Detector Selection:** Selects detectors uniformly at random over time as an unbiased lower-bound baseline.
- **Simple Context-Based Selector:** Maps extracted window statistics directly to detector choices without drift separation or feedback adaptation.
- **Performance-Based Selector:** Updates detector weights based purely on observed historical accuracy signals.
- **Full Ensemble:** Executes all detectors in the library at every time step and aggregates their outputs.
- **Existing Adaptive Frameworks:** Approaches such as RAMSES, mTSBench, and MSAD [3], [4], [5] that dynamically select models from available context or performance signals.

- **Oracle Selector:** A non-causal theoretical upper bound with access to ground truth at every step, used strictly to measure selection regret.

3.2 Evaluation criteria and trade-offs

These strategies are evaluated along five operational criteria:

(1) Detection quality and false alarms (precision, recall, F1-score, PR-AUC, ROC-AUC, false-alarm rate, and detection delay); (2) Adaptation speed and regime awareness (distinguishing benign regime transitions from suspicious drift); (3) Feedback resilience (performance under immediate, delayed, sparse, and absent labels without future-label leakage); (4) Computational cost and latency (inference overhead and budget compliance); and (5) Score comparability and decision stability (cross-detector score calibration and low switching chatter).

Table 2 summarizes the structural advantages and limitations of each paradigm. Static and simple context-based selectors fail when anomaly types or operating regimes shift unexpectedly. Performance-based selectors degrade rapidly when operator feedback is delayed or sparse, and suffer from uncalibrated score scales. Conversely, running a full ensemble incurs prohibitive computational latency in online settings. Existing adaptive methods improve upon static baselines but generally assume stationary distributions, prompt feedback, or unconstrained compute.

Table 2. Comparative analysis of detector selection approaches.

Approach	Advantages	Limitations	Category
Best fixed single detector	Simple implementation and consistent baseline behavior	No single detector works across all regimes; fails when anomaly types or contexts shift	Static baseline
Random detector selection	Unbiased reference without model-ranking assumptions	Ignores window context, concept drift, score calibration, compute cost, and stability	Lower baseline
Simple context-based selector	Incorporates sliding-window statistical context into routing	Lacks regime-aware drift separation, feedback adaptation, score calibration, and switch stability	Heuristic baseline
Performance-based selector	Tracks empirical detector effectiveness over time	Vulnerable to delayed, sparse, or absent feedback, incomparable raw scores, and noisy switching	Adaptive baseline
Full ensemble of all detectors	Leverages complementary strengths of all available models	High computational cost and inference latency; unsuitable for strict real-time budgets	Compute-heavy baseline
Existing adaptive methods (e.g., RAMSES [3], mTSBench [4], MSAD [5])	Specifically designed for adaptive model selection across time-series tasks	Frequently assume stationary distributions, immediate feedback, and ample compute; do not optimize all operational constraints jointly	State-of-the-art literature
Oracle selector	Provides a theoretical upper bound for measuring selection regret and oracle gap	Non-causal and privileged; serves solely as an analytical benchmark rather than a deployable method	Theoretical bound

4 Identified research gaps

Synthesizing the literature and comparative analysis reveals six core operational issues that remain unresolved when considered jointly:

1. **Unrealistic operational assumptions:** Existing adaptive selectors often presuppose stationary distributions, immediate ground-truth labels, and sufficient compute to evaluate multiple models.
2. **Context-dependent detector rankings:** Relying on a single detector leads to missed anomalies when fault modes or sensor dynamics change.
3. **Conflation of normal regime shifts and drift:** Without explicit regime tracking, planned operational shifts (e.g., load changes or maintenance modes) trigger false alarms or corrupt model adaptation.
4. **Delayed, sparse, or absent labels:** Production monitoring pipelines rarely provide immediate labels; selectors must learn from delayed confirmations and unsupervised proxy signals.
5. **Prohibitive ensemble overhead:** Executing high-cost deep detectors continuously violates latency constraints on edge or high-throughput gateways.
6. **Score incomparability and switching instability:** Heterogeneous detectors produce uncalibrated raw scores, and unregularized routing policies oscillate noisily between de- tectors.

The central research gap is not merely designing any single component in isolation, but jointly optimizing all six mechanisms under realistic deployment constraints. Table 3 maps each operational issue to the corresponding architectural module in the proposed CADR framework.

Table 3. Mapping of six core operational issues to proposed CADR modules.

Core issue	CADR module	Underlying mechanism	Expected improvement
No single detector works everywhere	Detector library with complementary profiles	Maintains a compact set of diverse detector families paired with adaptive weighting	Matches detector inductive bias to current stream context
Data behavior changes over time	Context- and regime-aware drift monitor	Separates planned normal operating regimes from suspicious drift and localized anomalies	Adapts smoothly without mislabeling normal regime shifts
Labels are delayed, sparse, or absent	Feedback-limited learning module	Updates trust from delayed labels when available, falling back on internal consistency signals	Enables reliable online adaptation under realistic label scarcity
Running all detectors is costly	Two-level cost-aware router / cascade	Runs inexpensive detectors continuously and escalates to costly models only when triggered	Substantially lowers average compute cost and inference latency
Detector scores are incomparable	Online score calibration module	Transforms raw scores onto a common scale via rolling median/ IQR normalization	Ensures consistent score fusion and threshold application
Selector switches too often or noisily	Stable decision policy	Enforces multi-window hysteresis margins alongside an explicit switch-cost penalty	Produces stable, interpretable decisions that preserve operator trust

5 Proposed framework: CADR

5.1 Formal problem definition

Consider a discrete-time stream of multivariate observations $x_t \in \mathbb{R}^d$ at time t, and let $W_t = \{x_{t-L+1}, \ldots, x_t\}$ denote a sliding window of length L. A feature extractor $\phi(\cdot)$ maps the window into a context representation $z_t = \phi(W_t)$. Let h_t denote accumulated historical performance and regime evidence, c_t denote current computational and latency constraints, and $\mathcal{D} = \{D_1, \ldots, D_K\}$ denote a library of K heterogeneous anomaly detectors.

The objective of CADR is to learn an adaptive selection policy

$$\pi_t = f(z_t, h_t, c_t) \tag{1}$$

that assigns non-negative routing weights $w_{k,t}$ to detectors over time. Each executed detector D_k produces a raw anomaly score $s^{(k)}_t = D_k(W_t)$, which is mapped to a calibrated score $\tilde{s}^{(k)}_t$. The combined ensemble anomaly score is computed as

$$S_t = \sum_{k=1}^{K} w_{k,t}\,\tilde{s}^{(k)}_t, \; w_{k,t} \geq 0, \; \sum_{k=1}^{K} w_{k,t} = 1 \tag{2}$$

A binary alert decision $y_t \in \{0, 1\}$ is generated against a dynamically calibrated decision threshold τ_t:

$$y_t = \begin{cases} 1, & S_t > \tau_t \\ 0, & S_t \leq \tau_t \end{cases} \tag{3}$$

The global selection policy is governed by a multi-objective optimization criterion balancing detection quality against operational penalties:

$$J = \text{Detection Quality} - \lambda_1 \text{False Alarms} - \lambda_2 \text{Detection Delay} - \lambda_3 \text{Computation} - \lambda_4 \text{Switching} \tag{4}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \geq 0$ are application-specific trade-off weights.

5.2 End-to-end architecture flow

As the multivariate stream x_t arrives, sliding-window context features $z_t = \phi(W_t)$ are extracted and passed to the Context and Drift Monitor. This monitor distinguishes normal operating-regime updates from suspicious drift and produces an uncertainty or disagreement estimate. Both signals feed the Cost-Aware Router, which dynamically allocates compute across three execution tiers: a single inexpensive detector, an expanded subset of detectors, or a selective ensemble. Raw outputs from active detectors pass through the Online Score Calibration module before being combined into S_t and evaluated against τ_t to yield alert y_t. Whenever delayed or sparse operator feedback arrives, the Feedback-Limited Learning module updates historical detector trust h_t without leaking future labels.

5.3 Module 1: Complementary detector library

CADR maintains a compact library of four to five detector families offering complementary detection capabilities across varying computational budgets, summarized in Table 4. Combining fast statistical and subspace detectors with medium-cost isolation/forecasting models and high-cost deep architectures allows the router to trade off sensitivity against latency.

Table 4. Proposed detector library profiles in CADR.

Detector family	Primary strength	Relative cost
Z-score / robust statistical	Sudden univariate point deviations	Very low
PCA-based detector	Shifts in linear multivariate correlation structure	Low
Isolation Forest [6]	Isolated anomalies in high-dimensional feature space	Low to medium
Forecasting-based detector	Deviations from expected short-term temporal dynamics	Medium
Autoencoder / LSTM detector [7]	Complex nonlinear temporal and reconstruction anomalies	High

5.4 Module 2: Context and drift monitor

The Context and Drift Monitor extracts summary statistics from W_t, including mean, variance, skewness, kurtosis, trend strength, autocorrelation, seasonality indicators, signal entropy, cross-variable correlation matrices, and streaming drift statistics [12], [9], [1], [7]. Using these indicators, the module classifies the current window into one of three states:

- **Normal regime change:** A recurring or planned operational transition (e.g., diurnal load cycles or maintenance states). The monitor maintains a library of known normal-regime profiles and updates baseline expectations prior to scoring anomalies.
- **Suspicious drift:** An unexplained distributional or structural shift indicative of sensor degradation, system fault, or emerging attack.
- **Point or collective anomaly:** Localized abnormal behavior occurring within an established normal regime.

5.5 Module 3: Two-level cost-aware router

To avoid executing all K detectors at every step, the router implements a two-level cascade:

- **Level 1 (Always-on baseline):** One or two low-cost detectors (e.g., robust statistical and PCA detectors) run continuously to provide baseline anomaly scores and rapid point-deviation screening.
- **Level 2 (On-demand escalation):** Higher-cost detectors are triggered only when an escalation condition is met: low confidence from Level 1 detectors, disagreement among active cheap detectors, elevated drift indicators, entry into a historically difficult operating regime, high domain criticality, or expiration of a maximum interval since the last deep evaluation.

Based on uncertainty and budget constraints c_t, Level 2 dispatches either a single specialized detector (low uncertainty, clear regime match), a small ensemble of two to three detectors (moderate disagreement or transitional context), or a full ensemble (high-risk or high-uncertainty episodes).

5.6 Module 4: Feedback-limited learning

When delayed ground-truth confirmation or incident feedback arrives for an earlier window, CADR updates the estimated trust $\hat{J}_{k,t}$ solely for the detectors that were active during that historical window, strictly preventing future-label leakage:

$$\hat{J}_{k,t} \leftarrow \left(1 - \alpha_t\right)\hat{J}_{k,t-1} + \alpha_t r_{k,t} \tag{5}$$

where $r_{k,t}$ is the feedback-derived reward and α_t is an adaptive learning rate that increases during detected drift and decays during stationary regimes. When external labels are unavailable, the module relies on internal unsupervised consistency signals, including inter-detector agreement, temporal score stability, reconstruction or forecasting error residuals, change-point evidence, and context similarity to previously validated regimes. Table 5 outlines this feedback hierarchy.

Table 5. Hierarchical feedback handling in CADR.

Feedback type	Availability	CADR update mechanism
Immediate ground-truth labels	Rare in practice	Direct high-confidence trust update via Eq. (5)
Delayed operator confirmation	Common	Retroactive trust update restricted to historically active detectors
Sparse confirmed incidents	Common	Localized trust refinement within matching context regimes
Downstream alarms / maintenance logs	Intermittent	Weak contextual supervision and regime-profile verification
No operational labels (proxy only)	Common	Unsupervised detector agreement, score stability, and conservative adaptation

5.7 Module 5: Online score calibration

Because raw scores from statistical, tree-based, and neural detectors span incompatible numerical ranges, CADR maintains a rolling window of recent score distributions per detector and applies robust median and interquartile range (IQR) normalization:

$$\tilde{s}_t^{(k)} = \frac{s_t^{(k)} - \mathrm{median}\left(s^{(k)}\right)}{\mathrm{IQR}\left(s^{(k)}\right)} \tag{6}$$

Compared to mean and standard deviation scaling, Eq. (6) resists distortion from extreme anomaly spikes. Calibration statistics are updated periodically and reset immediately upon confirmed regime transitions, keeping score alignment decoupled from threshold selection τ_t.

5.8 Module 6: Stable decision policy

To prevent noisy oscillations between candidate detectors, CADR enforces a hysteresis switching rule: a transition from active detector A to candidate detector B occurs only if

$$\hat{J}_{B,t} > \hat{J}_{A,t} + \delta \tag{7}$$

holds consistently across m consecutive sliding windows, where $\delta > 0$ is a minimum performance margin. Combined with the explicit switch-cost penalty λ_4 in Eq. (4), this mechanism suppresses transient switching chatter and preserves operational interpretability.

6 Planned experimental protocol and expected outcomes

6.1 Experimental design

Because CADR is currently a conceptual architecture, no empirical benchmark numbers are claimed in this paper. Instead, we define a strict, leak-free experimental protocol to test the central hypothesis

that CADR can approach the detection quality of an oracle selector while consuming substantially less computation than a full ensemble under delayed, sparse, or absent feedback.

CADR will be benchmarked against the seven selection strategies detailed in the section 'Baseline and adaptive selection strategies' using two complementary metric suites:

- **Detection metrics:** Precision, recall, F1-score, PR-AUC, ROC-AUC, false alarms per unit time, detection delay, and event-level F1.
- **Selector and efficiency metrics:** Selection regret relative to the oracle, oracle gap, selector accuracy, switching frequency (stability), post-drift adaptation time, average inference latency, FLOPs/compute overhead, and budget constraint violations.

All experiments will enforce strict temporal ordering so that future labels are never exposed to preprocessing, calibration, or routing modules. Table 6 summarizes the planned evaluation axes.

Table 6. Planned evaluation scenarios and ablation studies.

Experiment axis	Planned variants and configurations
Feedback regimes (time-ordered)	Idealized (immediate labels), Restricted (delayed labels), Sparse (limited confirmed incidents), Unsupervised (zero operational labels)
Temporal adaptation	Stationary context, gradual drift, abrupt regime change, evolving attack/anomaly types, shifting normal behavior
Cross-domain generalization	Development and hyperparameter tuning on SWaT [9], [10], followed by zero-shot or transfer evaluation on unseen multivariate datasets
Ablation studies	CADR without context features; without drift/regime monitor; without feedback updates; without cost-aware routing; without online score calibration; without hysteresis; and full-ensemble always-on baseline

6.2 Expected outcomes and future scope

Empirical validation of CADR is expected to demonstrate three primary outcomes: (1) competitive F1 and PR-AUC scores approaching the oracle bound at a fraction of full-ensemble compute cost; (2) substantial reductions in false alarms during planned operational regime shifts due to explicit regime profiling; and (3) stable detector trajectories without erratic switching under sparse or delayed feedback. Future research directions include multi-domain empirical benchmark- ing beyond SWaT, automated Pareto tuning of the objective weights (λ_1, ..., λ_4) for safety-critical versus costsensitive deployments, and stress testing under adversarial drift condi- tions.

7 Conclusion

Existing adaptive model-selection frameworks for time-series anomaly detection, including RAMSES, mTSBench, and MSAD, demonstrate the value of dynamic detector selection but typically rely on idealized assumptions of stationary distributions, immediate feedback, and unconstrained compute. To bridge the gap toward practical deployment, this paper formulated CADR, a conceptual framework that jointly coordinates a complementary detector library, a context- and regime-aware drift monitor, a two-level cost-aware router, feedback-limited trust learning, rolling robust score calibration, and a hysteresis-stabilized decision policy. By formalizing the multi-objective trade-off across detection accuracy, false alarms, latency, compute cost, and switching stability, CADR provides a structured foundation for dependable streaming anomaly detection. Implementing and validating the frame-

work across the proposed time-ordered drift, feedback, and ablation benchmarks remains the immediate focus of future work.

References

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput*. Surv., vol. 41, no. 3, Art. no. 15, Jul. 2009.

[2] F. Wang, Y. Jiang, R. Zhang, A. Wei, J. Xie, and X. Pang, "A survey of deep anomaly detection in multivariate time series: Taxonomy, applications, and directions," *Sensors*, vol. 25, Art. no. 190, 2025.

[3] M. Abdelmaksoud, S. Ding, A. Morozov, and Z. Abedjan, "RAMSES: Robust and adaptive model selection for time-series anomaly detection algorithms," in *Proc*. 2026 IEEE 42nd Int. Conf. Data Engineering (ICDE), 2026, pp. 2135–2149.

[4] X. Zhou, C. Brif, and I. Lourentzou, "mTSBench: Benchmarking multivariate time series anomaly detection and model selection at scale," *Trans*. Mach. Learn. Res., 2026, arXiv:2506.21550.

[5] E. Sylligardos, J. Paparrizos, T. Palpanas, P. Senellart, and P. Boniol, "MSAD: A deep dive into model selection for time series anomaly detection," *The VLDB Journal*, vol. 34, Art. no. 72, 2025.

[6] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc*. 2008 Eighth IEEE Int. Conf. Data Mining (ICDM), 2008, pp. 413–422.

[7] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "LSTMbased encoder-decoder for multisensor anomaly detection," *arXiv preprint arXiv:1607.00148*, 2016.

[8] T. W. Ha, "Multivariate time series anomaly detection using directed hypergraph neural networks," *Appl*. Artif. Intell., 2025.

[9] J. Goh, S. Adepu, K. N. Junejo, and A. Mathur, "A dataset to support research in the design of secure water treatment systems," in *Critical Information Infrastructures Security (CRITIS* 2016), LNCS vol. 10242, 2017, pp. 88–99.

[10] J. Inoue, Y. Yamagata, Y. Chen, C. M. Poskitt, and J. Sun, "Anomaly detection for a water treatment system using unsupervised machine learning," in *Proc*. 2017 IEEE Int. Conf. Data Mining Workshops (ICDMW), 2017, pp. 1058–1065.

[11] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Comput*. Surv., vol. 46, no. 4, Art. no. 44, 2014.

[12] J. Gama, P. Medas, G. Castillo, and P. Rodrigues, "Learning with drift detection," in *Advances in Artificial Intelligence – SBIA* 2004, 2004, pp. 286–295.

[13] A. Bifet and R. Gavaldà, "Learning from timechanging data with adaptive windowing," in *Proc*. 2007 SIAM Int. Conf. Data Mining (SDM), 2007, pp. 443–448.

[14] H. Wang and Z. Abraham, "Concept drift detection for streaming data," *arXiv preprint arXiv:1504.01044*, 2015.

[15] E. B. Gulcan and F. Can, "Unsupervised concept drift detection for multi-label data streams," *Artif*. Intell. Rev., vol. 56, pp. 2401–2434, 2022.