

Anomalies in Multivariate Time Series Benchmarks Are Mostly Univariate

Marc Pinet*

marc.pinet@orange.com

Orange Research, Univ. Grenoble Alpes, CNRS, LIG
Meylan, France

Samuel Berlemont

samuel.berlemont@orange.com

Orange Research
Meylan, France

Julien Cumin

julien1.cumin@orange.com

Orange Research
Meylan, France

Dominique Vaufreydaz

dominique.vaufreydaz@univ-grenoble-alpes.fr

Univ. Grenoble Alpes, CNRS, LIG
Grenoble, France

Abstract

Many recent multivariate time series anomaly detection (MTSAD) models incorporate cross-channel modeling, under the implicit assumption that the structure of anomalies may be spread across multiple channels. We evaluate this assumption on eight widely used public benchmarks by introducing a per-segment diagnostic framework that flags, for each labeled anomaly, whether at least one channel deviates individually from its normal history, whether the cross-channel correlation structure changes, or both. The framework shows that no cross-channel rupture occurs without an accompanying univariate deviation across a range of reasonable thresholds. A complementary metric also reveals that on six of the eight benchmarks, at least half of the labeled anomaly segments deviate univariately on 89% to 100% of their timesteps, reaching 100% on three of these datasets. To verify that our framework captures cross-channel structure when present, we construct synthetic data of phase-shifted sinusoidal channels with shared noise. Each anomalous segment is altered through one of two channel-wise corruptions that preserve the per-channel marginal distribution while breaking cross-channel structure, and our framework correctly characterizes these segments as cross-channel-only. On these data, channel-dependent (CD) models successfully exploit the cross-channel signal whereas channel-independent (CI) ones fail. The CI/CD comparison of a recent SOTA detector on real benchmarks further confirms that CD modeling brings no measurable gain. We conclude that current MTSAD benchmarks are unsuitable for validating cross-channel modeling capabilities, and we call for the development of more structurally diverse evaluation sets. The code for this study is publicly available¹.

*Corresponding Author

¹<https://github.com/marcpinet/mtsad-benchmarks-are-mostly-univariate>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MILETS @ KDD'26, Jeju, Korea

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2026/08

<https://doi.org/10.48550/arXiv.2606.02670>

CCS Concepts

• **Computing methodologies** → **Anomaly detection**; • **Mathematics of computing** → *Time series analysis*; • **Information systems** → Data stream mining.

Keywords

Anomaly detection, Multivariate time series, Datasets, Evaluation methodology, Time series analysis, Benchmark reliability

ACM Reference Format:

Marc Pinet, Julien Cumin, Samuel Berlemont, and Dominique Vaufreydaz. 2026. Anomalies in Multivariate Time Series Benchmarks Are Mostly Univariate. In *Proceedings of The 12th Mining and Learning from Time Series (MILETS) Workshop at the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (MILETS @ KDD'26)*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.48550/arXiv.2606.02670>

1 Introduction

Many monitoring tasks rely on time series anomaly detection (TSAD) [18]. In the multivariate setting (MTSAD), recent architectures have increasingly turned to channel-aware mechanisms: graph-based detectors [5, 36], cross-channel modeling through learnable masked attention [32, 35], and more. These models rely on the same implicit assumption: that public benchmarks contain anomalies that are more easily detected by modeling cross-channel dependencies. The closely related field of forecasting has reached the opposite consensus, where the channel-independent (CI) strategy popularized by PatchTST [17] dominates through a capacity-robustness trade-off [9], and where the recent survey of Qiu et al. [20] formalizes the intermediate channel-partiality (CP) strategy. This shift has begun to propagate to TSAD, where several recent detectors adopt channel-independent designs [4, 12, 23, 33, 39] and report competitive or superior performance to channel-aware counterparts.

For each labeled anomalous segment in a multivariate benchmark, we ask: *is there evidence of a univariate deviation (an individual channel deviating from its recent normal history), of a change in cross-channel dependence, both, or neither?* Our per-segment diagnostic (Section 3) combines a univariate test with a cross-channel test based on the Pearson and Spearman correlations and the unbiased squared distance correlation [26], with lagged extensions. Applied to eight standard benchmarks (Section 4), the diagnostic finds no labeled segment in the strictly cross-channel category, on

any benchmark, under any of the three measures, across a sensitivity sweep over both thresholds within their reasonable range. Every cross-channel rupture co-occurs with a univariate deviation, and, on six of the eight benchmarks, at least half of the labeled anomaly segments deviate univariately between 89% and 100% of their timesteps. A synthetic protocol (Section 5) confirms that the diagnostic identifies cross-channel ruptures when they exist. Moreover, channel-dependent models trained on these data recover the anomalous segments, whereas channel-independent variants perform at random (AUC-ROC ≈ 0.5). On real benchmarks, our channel-dependent variant of the channel-independent state-of-the-art (SOTA) detector CrossAD [12] brings no measurable gain over the original (Section 6).

Our contributions are: (i) a per-segment diagnostic framework that characterizes each labeled anomaly along two axes (univariate and cross-channel), with extension to lag and a sensitivity analysis over diagnostic thresholds; (ii) an empirical evaluation of eight standard benchmarks reporting the absence of strictly cross-channel anomaly segments, complemented by a univariate-ratio metric that measures how persistent the univariate signal is within each segment; (iii) a synthetic protocol validating the proposed diagnostic framework, on which a flattened linear autoencoder outperforms two recent SOTA detectors on anomalies that require cross-channel modeling; and (iv) a comparison on real benchmarks showing that the channel-dependent variant of CrossAD does not improve over its channel-independent counterpart, and in one case collapses entirely.

2 Related Work

MTSAD models are classified by their treatment of cross-channel dependence into three strategies [20]: channel-independent (CI), where each channel is processed independently without considering interactions among channels; channel-dependent (CD), which treats all channels as a unified entity assumed to be interdependent; and channel-partiality (CP), where each channel retains some independence while interacting selectively with a subset of others. In TSAD specifically, channel-aware (CD and CP) designs have proliferated, from graph-based detectors (GDN [5], MTAD-GAT [36]) to MLP-Mixer-based designs with explicit channel mixing, where a multi-layer perceptron (MLP) operates along the channel dimension to fuse cross-channel information (PatchAD [38], CCM-TAD [16]) and cross-channel attention (XCTFormer [35], CATCH [32]). Hybrid strategies that combine a channel-independent prediction branch with a channel-mixing reconstruction branch (DBAD [25]) further illustrate the variety of channel-strategy choices currently being explored. Beyond architecture, ChInf [29] scores anomalies with per-channel influence functions, while remaining structurally channel-independent (further discussed in Section 6).

A parallel line of work has questioned the reliability of public MTSAD benchmarks. Wu and Keogh [31] identified four classes of flaws in the Yahoo, Numenta, NASA, and SMD² benchmarks (triviality, unrealistic anomaly density, mislabeled ground truth,

and run-to-failure bias) and issued five recommendations, including the abandonment of these datasets and the systematic visualization of data and algorithmic outputs. The authors target the quality of labeled segments at a per-exemplar and univariate level, whereas our evaluation is structural and explicitly multivariate: it asks, for each labeled segment, what kind of evidence (univariate or cross-channel) the data actually carries. These two works thus complement each other. Wagner et al. [28] conducted the first audit dedicated to multivariate benchmarks and concluded that SWaT, WADI (water treatment and distribution), MSL, and SMAP (spacecraft telemetry) are unsuited for MTSAD evaluation due to distribution shift, positional bias, and the presence of a single continuous sensor channel alongside binary ones in MSL and SMAP. Sarfraz et al. [21] showed that classical baselines such as Principal Component Analysis (PCA) reconstruction error and nearest-neighbor distance match or exceed SOTA deep learning models on standard MTSAD benchmarks, and further reduced trained deep TSAD models to linear distillations that retained their detection performance. Liu and Paparrizos [13] curated TSB-AD across 40 datasets and concluded that statistical methods and lightweight models often outperform SOTA deep learning detectors. Most recently, Zhou et al. [40] introduced mTSBench and observed “high-magnitude anomalies in certain channels” for PSM, SMD, and SMAP.

Most directly relevant to our work, Garg et al. [7] claimed, based on author descriptions and visual inspection, that the standard MTSAD benchmarks “primarily contain temporal anomalies” and that “we did not find open datasets where cross-channel anomalies are known to be present”. They also observed that a univariate fully connected autoencoder beats SOTA multivariate algorithms across seven datasets. Wenig et al. [30] later reported the same pattern from a detector-comparison angle on fourteen benchmarks. Both works infer the structure of the benchmarks indirectly, through detector performance or visual inspection.

Across all these audits, the structure of the labeled anomalies themselves is inferred indirectly, through detector performance or qualitative case analysis. Our contribution is to characterize them directly through a per-segment diagnostic that operates on the data independently of any detector, and to validate it on a controlled synthetic protocol of cross-channel-only anomalies.

3 Diagnostic Framework

Our diagnostic framework operates on each anomalous segment and applies two tests: a univariate test that flags deviations from a recent normal history, and a cross-channel test that flags changes in the dependence structure between channels. *The goal is to characterize the evidence available to a detector, rather than implementing an anomaly detector.* Unlike an anomaly detector, the diagnostic uses the ground-truth labels as input: it isolates each labeled segment, builds its normal history from the preceding unlabeled timesteps, and measures the evidence carried by the segment relative to that history.

3.1 Definitions

We denote by $X \in \mathbb{R}^{T \times C}$ a multivariate time series of C channels and T timesteps indexed by $t \in \{1, \dots, T\}$, with binary labels $y \in \{0, 1\}^T$ where $y_t = 1$ marks an anomalous timestep.

²The benchmark that Wu and Keogh [31] call “OMNI” is the Server Machine Dataset introduced alongside OmniAnomaly by Su et al. [24].

An *anomalous segment* $\mathcal{A} = [t_s, t_e]$ is any maximal contiguous run of timesteps with $y_t = 1$ in $\mathbf{y}$. Its length is denoted $|\mathcal{A}| = t_e - t_s + 1$. Each segment is analyzed independently (referred to as *per-segment* throughout the paper).

For each anomalous segment $\mathcal{A}$, we define its *normal history* $\mathcal{H}_{\mathcal{A}}$ as the H closest normal timesteps preceding t_s . The history is constructed by walking backwards from t_s , skipping any anomalous timestep encountered along the way.

3.2 Univariate Test

For each anomalous segment $\mathcal{A}$, we estimate the per-channel normal mean $\mu_c^{\mathcal{H}_{\mathcal{A}}}$ and standard deviation $\sigma_c^{\mathcal{H}_{\mathcal{A}}}$ from $\mathcal{H}_{\mathcal{A}}$. We then compute, for every channel $c \in \{1, \dots, C\}$, the maximum absolute z-score reached inside the segment:

$$z_c^{\mathcal{A}} = \max_{t \in \mathcal{A}} \frac{|X_{t,c} - \mu_c^{\mathcal{H}_{\mathcal{A}}}|}{\sigma_c^{\mathcal{H}_{\mathcal{A}}}}. \quad (1)$$

The segment is classified *univariate* if there is a channel $c \in \{1, \dots, C\}$ such that $z_c^{\mathcal{A}} > \tau_z$. We use a default threshold $\tau_z = 3$. Indeed, Chebyshev's inequality guarantees that for any random variable X with finite mean μ and finite standard deviation σ :

$$P(|X - \mu| > k\sigma) \leq \frac{1}{k^2}, \quad \forall k > 0. \quad (2)$$

With $\tau_z = 3$, the probability that a single timestep deviates from its normal mean by more than $3\sigma_c^{\mathcal{H}_{\mathcal{A}}}$ under non-anomalous stationary conditions is bounded by $1/9 \approx 11\%$, regardless of the distribution (this is only an upper bound, not the actual probability). The threshold is therefore not chosen to fit any benchmark, and we report a sensitivity analysis over $\tau_z \in \{2, 3, 4, 5, 7, 10\}$ in Section 4.3.

While this condition is necessary for a segment to be classified as univariate, it is not sufficient on its own: the existence of a single timestep exceeding τ_z in one channel says nothing about how pronounced the univariate signal is across $\mathcal{A}$. We therefore complement this binary test with a per-segment intensity measure in Section 3.5.

We keep the z-score as the default test for its simplicity. When it is inapplicable, for instance on channels that are binary or nearly constant, we fall back to a Matrix Profile [34] applied on each channel independently, which captures unusual motifs rather than unusual values.

The z-score is deliberately a minimal univariate test, capturing only deviations in level relative to the recent history. Anomalies that the z-score misses but a richer univariate test would catch only strengthen our claim: they are still univariate.

3.3 Correlation Test

The cross-channel test compares the dependence structure between channels inside an anomalous segment with the one observed in its normal history. For each segment $\mathcal{A}$ and its history $\mathcal{H}_{\mathcal{A}}$, we compute two correlation matrices $\mathbf{R}^{\mathcal{A}}, \mathbf{R}^{\mathcal{H}_{\mathcal{A}}} \in \mathbb{R}^{C \times C}$ and quantify their discrepancy³ by:

$$\Delta_{\rho}^{\max} = \max_{i \neq j} \left| R_{ij}^{\mathcal{A}} - R_{ij}^{\mathcal{H}_{\mathcal{A}}} \right|. \quad (3)$$

³Channels with negligible variance ($\sigma < 10^{-4}$) in either $\mathcal{A}$ or $\mathcal{H}_{\mathcal{A}}$ are excluded from the cross-channel test, since their pairwise correlations are undefined.

The segment is classified *cross-channel* if $\Delta_{\rho}^{\max} > \tau_{\rho}$ for at least one of the three following correlation methods:

- the **Pearson** correlation and its coefficient $\rho_{ij} \in [-1, 1]$, measuring linear dependence,
- the **Spearman** rank correlation and its coefficient $\rho_{ij} \in [-1, 1]$, measuring monotone dependence,
- the **Distance correlation** [27], in its unbiased squared form [26] $\rho_{ij} = \text{dCor}_{ij}^2 \in [-1, 1]$ (negative values are bias-correction noise, clipped to zero [14]), which captures any form of statistical dependence and is harder to trigger than Pearson/Spearman at high τ_{ρ} .

The threshold $\tau_{\rho} \in [0, 2]$ controls the sensitivity of the cross-channel test. We use a permissive default threshold $\tau_{\rho} = 0.1$ and report a sensitivity analysis in Section 4.3. We use three measures rather than dCor alone because dCor is computed at lag zero only, while Pearson and Spearman allow an efficient lagged extension that captures lead-lag dependencies.

Extension to lag. Channel pairs in real-world systems often exhibit time-shifted relationships [37]. To capture such dependencies, we extend Pearson and Spearman to lagged correlations: $\hat{\rho}_{ij}(\ell)$ measures, on the training set, the correlation between channel i at time t and channel j at time $t + \ell$, for $\ell \in \{-L_{\max}, \dots, L_{\max}\}$.

Let $\mathcal{L} = \{-L_{\max}, \dots, L_{\max}\} \setminus \{0\}$ denote the set of non-zero lags. A pair (i, j) is recorded as *lagged-dominant* if $\exists \ell_{ij}^* \in \mathcal{L}$ such that $|\hat{\rho}_{ij}(\ell_{ij}^*)| > |\hat{\rho}_{ij}(0)|$. When computing $\mathbf{R}^{\mathcal{A}}$ and $\mathbf{R}^{\mathcal{H}_{\mathcal{A}}}$ in Equation (3), the entry (i, j) for any lagged-dominant pair uses the lagged correlation at ℓ_{ij}^* rather than at lag zero.

Distance correlation is intentionally not extended to lags. Computing dCor_{ij}^2 at every lag for every channel pair was prohibitive even on the smallest benchmark in both channel count and series length.

3.4 Classification

Combining both tests yields four per-segment labels: **UNIVARIATE**, **CROSS-CHANNEL**, **BOTH**, or **UNDETECTED**. Segments shorter than $L_{\min}$ skip the cross-channel test (Section 3.3) and fall back to UNIVARIATE or UNDETECTED. A CROSS-CHANNEL segment that fails the univariate test at every timestep would thus require channel-aware modeling.

3.5 Univariate Ratio

The classification of Section 3.4 flags whether at least one timestep in a segment exhibits a univariate deviation, but it does not measure how persistent this deviation is across the segment (in the same way that a single positive prediction inflates point-adjusted metrics [11]). To quantify this persistence, we define the per-segment *univariate ratio*:

$$u^{\mathcal{A}} = \frac{1}{|\mathcal{A}|} \sum_{t \in \mathcal{A}} \mathbf{1} \left\{ \max_{c \in \{1, \dots, C\}} \frac{|X_{t,c} - \mu_c^{\mathcal{H}_{\mathcal{A}}}|}{\sigma_c^{\mathcal{H}_{\mathcal{A}}}} > \tau_z \right\}. \quad (4)$$

The ratio $u^{\mathcal{A}} \in [0, 1]$ is the fraction of timesteps within the segment where at least one channel exceeds the z-score threshold. A value of $u^{\mathcal{A}} = 1$ means that every timestep of the segment exhibits a univariate deviation on at least one channel. A value close to zero means that the segment passes the binary univariate test due to a

Table 1: Summary of the eight MTSAD benchmarks evaluated. C is the number of channels. The anomaly ratio is computed on the test split.

Dataset	C	Train size	Test size	Anomaly ratio (%)
GECCO	9	69 260	69 261	1.05
MSL	55	58 317	73 729	10.53
SMAP	25	140 825	444 035	12.87
PSM	25	132 481	87 841	27.76
SMD	38	708 405	708 420	4.16
SWAN-SF	38	60 000	60 000	32.60
SWaT	31	472 459	472 460	11.56
WADI	79	784 571	172 803	5.77

single or a few isolated timesteps. We report $u^{\mathcal{A}}$ both globally (over all long segments) and restricted to segments classified as BOTH, which directly addresses whether segments flagged by both tests retain a persistent univariate signal or a sparse one. Sensitivity to τ_z is reported in Section 4.4.

4 Empirical Evaluation

4.1 Datasets

We evaluate eight benchmarks used in the vast majority of recent MTSAD evaluations [18, 21, 22, 28]: GECCO [15], NASA (MSL and SMAP) [10], PSM [1], SMD [24], SWaT [8], SWAN-SF [3], and WADI [2]. They span water treatment and distribution (GECCO, SWaT, WADI), spacecraft telemetry (MSL, SMAP), server metrics (PSM, SMD), and solar weather (SWAN-SF). Their main characteristics are summarized in Table 1.

We adopt a uniform preprocessing pipeline: for each dataset, we fit a per-channel StandardScaler on the training set and apply the resulting transform to both train and test splits. SMD is split into 28 separate machines and each machine is processed independently, with its own scaler, as recommended by Su et al. [24]. The GECCO training set contains 1.44% of labeled anomalies, which we filter out before the preprocessing. The original train/test splits are preserved.

4.2 Setup

We apply the diagnostic of Section 3 with the following default hyperparameters: history length $H = 300$, univariate threshold $\tau_z = 3$, correlation threshold $\tau_\rho = 0.1$, maximum lag $L_{\max} = 192$, and minimum segment length $L_{\min} = 10$. We verified that the conclusions of this section are stable for $H \in \{100, 300, 1000\}$. The permissive correlation threshold τ_ρ deliberately makes the cross-channel test easy to trigger, which strengthens the central claim of this paper: the lack of segments classified as cross-channel without a univariate deviation. The maximum lag $L_{\max}$ matches the upper end of sliding window sizes used in recent MTSAD models [12, 32]. Sensitivity to τ_z and τ_ρ is reported in Section 4.3 and Section 4.4.

4.3 Classification Results

Table 2 reports the per-segment classification on each benchmark, with each of the three correlation methods. Out of 373 long segments aggregated across all datasets and methods, the CROSS-CHANNEL category is *empty*: every detected cross-channel rupture

Table 2: Per-segment classification on the eight benchmarks, for each of the three correlation methods (P = Pearson, S = Spearman, D = dCor_U²). The chosen thresholds are $\tau_z = 3$, $\tau_\rho = 0.1$. Long segments are those with $|\mathcal{A}| \geq L_{\min} = 10$. The two rightmost columns count short segments labeled UNIVARIATE or UNDETECTED. The CROSS-CHANNEL category is empty everywhere, on every method, and is therefore not shown.

Dataset	UNIVARIATE			BOTH			Short	
	P	S	D	P	S	D	UNI	UND
GECCO	0	0	0	22	22	22	0	0
MSL	4	3	31	30	31	3	0	0
SMAP	15	15	16	41	41	40	0	0
PSM	0	0	0	34	34	34	33	4
SMD	0	0	0	176	176	176	148	1
SWAN-SF	0	0	0	2	2	2	5079	1056
SWaT	0	0	0	35	35	35	0	0
WADI	0	0	0	14	14	14	0	0

co-occurs with a univariate deviation. The BOTH category dominates almost everywhere. This is consistent with our permissive correlation threshold $\tau_\rho = 0.1$.

Figure 1 reports the CROSS-CHANNEL count for each correlation method across different (τ_z, τ_ρ) , aggregated across the eight benchmarks. At $\tau_z \in \{2, 3\}$, zero CROSS-CHANNEL segments are detected at any $\tau_\rho \in [0.1, 0.8]$ on any benchmark. At $\tau_z = 4$, two CROSS-CHANNEL segments appear on a single dataset. It is only at $\tau_z \geq 7$, where Chebyshev’s bound drops below $1/49 \approx 2\%$, that CROSS-CHANNEL segments appear on a majority of benchmarks.

Even at extreme thresholds, the CROSS-CHANNEL category remains small: across the entire (τ_z, τ_ρ) sweep, the maximum number of CROSS-CHANNEL segments observed in a single configuration is 34 out of the 373 long segments aggregated across all benchmarks, that is less than $\approx 9\%$.

SWAN-SF. Out of its 6 137 labeled segments, only 2 exceed $L_{\min} = 10$. The remaining 6 135 segments are too short for the cross-channel test and thus are only classified by the univariate test (Section 3.3); their length distribution is reported in Figure 2. Among them, 5 079 (83%) are classified UNIVARIATE, 1 056 are UNDETECTED, and 2 lack sufficient context (at the beginning of the test set, where fewer than $H = 300$ normal points precede the segment). The 1 056 UNDETECTED segments do not threaten the main claim of Section 4, being too short for the cross-channel test.

MSL and SMAP. The NASA benchmarks differ from the others. All channels in MSL and SMAP are binary (with the exception of one channel each), which limits the discriminative power of the z-score test. We therefore fall back to the Matrix Profile (see Section 3.2) to confirm that these segments, especially the BOTH ones, remain detectable univariately. On MSL, it detects 27 out of 31 BOTH segments (union over the three correlation methods), while on SMAP, it identifies 37 out of 43 segments. The remaining segments are visually indistinguishable from their normal history, consistent with the labeling concerns raised by Wu and Keogh [31].

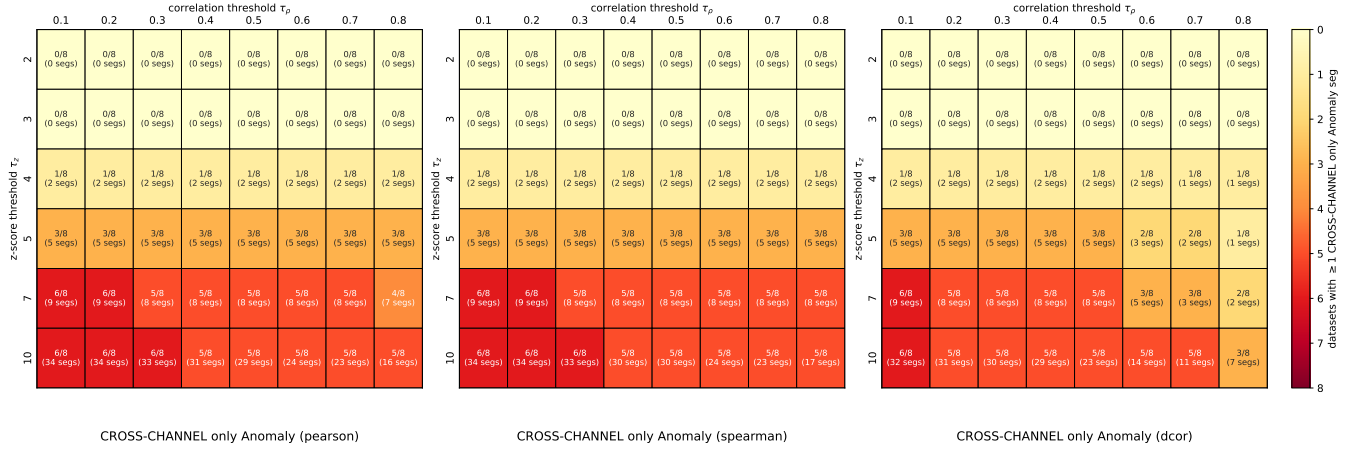


Figure 1: Sensitivity of the CROSS-CHANNEL count to both thresholds τ_z and τ_ρ , aggregated across the eight benchmarks for each of the three correlation methods. Each cell reports the number of datasets containing at least one CROSS-CHANNEL segment, followed by the total count of such segments in parentheses.

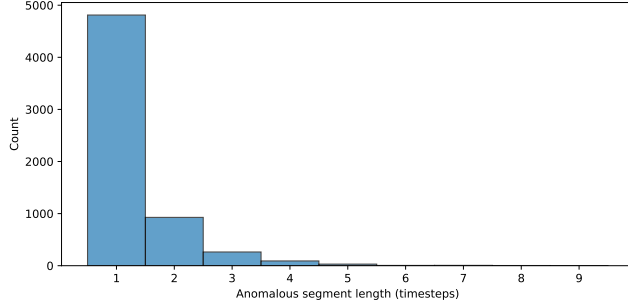


Figure 2: Distribution of segment lengths on SWAN-SF. All labeled anomalies are point-wise (length 1) or near-point-wise, with only two segments exceeding $L_{\min} = 10$.

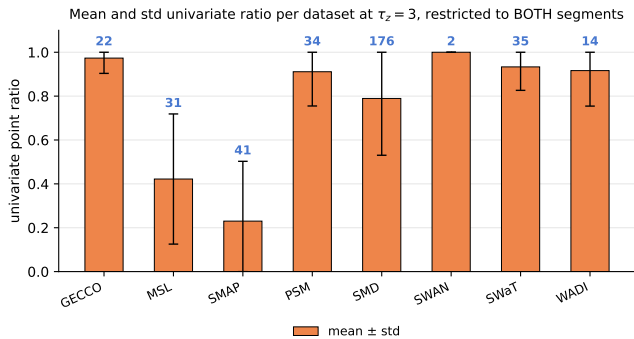


Figure 3: Mean univariate ratio (with standard deviation) per dataset at $\tau_z = 3$, restricted to BOTH segments. The number above each bar is the segment count.

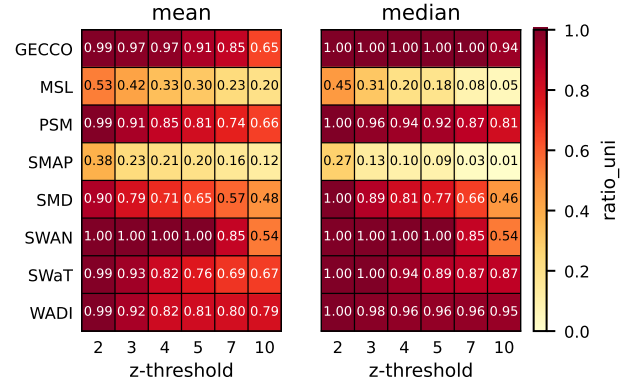


Figure 4: Sensitivity of the univariate ratio to τ_z across the eight benchmarks, restricted to BOTH segments. Mean (left) and median (right).

4.4 Univariate Ratio Results

Figure 3 reports the mean (with standard deviation) univariate ratio per dataset at $\tau_z = 3$, restricted to BOTH segments. We focus on this subset because it directly addresses our central question, i.e., whether segments flagged as both univariate and cross-channel have a dense univariate signal or not.

Six of the eight benchmarks have a mean univariate ratio between 0.79 and 1.00. Labeled anomalies on these benchmarks therefore not only deviate univariately somewhere within their anomalous segments, but deviate univariately on most of their anomalous timesteps. MSL and SMAP exhibit lower means (0.42 and 0.23 respectively) and are discussed below. Note that SWAN-SF has only two long segments (its 6 137 short segments were analyzed in Section 4.3).

Figure 4 reports the univariate ratio on BOTH segments with $\tau_z \in \{2, 3, 4, 5, 7, 10\}$. Raising τ_z from 3 to 5, which tightens Chebyshev's

bound from $1/9 \approx 11\%$ to $1/25 = 4\%$, leaves the median ratio above 0.89 on five of the eight benchmarks (GECCO, PSM, SWAN, SWaT, and WADI) and at 0.77 on SMD. The same pattern holds across the entire sweep: even at $\tau_z = 10$ where Chebyshev's bound drops to 1%, the median ratio stays at 0.94, 0.81, 0.87, and 0.95 on GECCO, PSM, SWaT, and WADI respectively.

Both NASA benchmarks exhibit low univariate ratios across all thresholds, with SMAP medians dropping to 0.13 already at $\tau_z = 3$. As explained in Section 4.3, the z-score is not suited to capture how strongly they deviate univariately. However, the complementary Matrix Profile evaluation reported in the aforementioned section already confirms that the BOTH segments remain detectable univariately.

The diagnostic of Section 4 reports an empty CROSS-CHANNEL column on every benchmark, but this raises an immediate concern: would the diagnostic actually flag a CROSS-CHANNEL segment if one existed? We address this with a controlled synthetic protocol in which anomalies are strictly cross-channel, then leverage the same protocol to compare CI, CD, and CP detectors in a setting where cross-channel modeling is required.

5 Validation Protocol using Synthetic Data

The evaluation is complemented with a controlled synthetic protocol in which every anomaly is explicitly constructed to be strictly cross-channel.

5.1 Construction of Synthetic Data

We generate a synthetic dataset of $C = 9$ phase-shifted sinusoidal channels with shared additive noise. For each timestep t and channel c :

$$X_{t,c} = \sin\left(\frac{2\pi t}{P} + \frac{2\pi c}{C}\right) + \varepsilon_t, \quad \varepsilon_t \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2), \quad (5)$$

with period $P = 50$ and noise standard deviation $\sigma_\varepsilon = 0.3$. ε_t does not depend on c : the noise is shared across channels at each timestep. This induces a strong correlation between every pair of channels in the clean signal. The training set has $T_{\text{train}} = 20\,000$ anomaly-free timesteps. The test set has $T_{\text{test}} = 10\,000$ timesteps and contains four anomalous segments of lengths 400, 300, 500, and 300, distributed at fixed timesteps.

For each anomalous segment $\mathcal{A}$, a subset $\mathcal{S}_{\mathcal{A}} \subset \{1, \dots, C\}$ of channels is selected to be corrupted, uniformly at random. The size $|\mathcal{S}_{\mathcal{A}}|$ is itself drawn uniformly in $\{1, \dots, C-1\}$. Both extremes are excluded because corrupting zero channels would leave the segment unaltered, while corrupting all channels would apply the same transformation to every channel and therefore leave the cross-channel correlation structure unchanged. We consider two corruption methods:

- **NOISEFLIP.** For each $c \in \mathcal{S}_{\mathcal{A}}$ and $t \in \mathcal{A}$, set $X_{t,c} \leftarrow X_{t,c} - 2\varepsilon_t$, replacing the shared noise on channel c by its opposite $-\varepsilon_t$. By symmetry of the Gaussian, ε_t and $-\varepsilon_t$ are identically distributed, leaving the per-channel marginal unchanged while making the noise on c anti-correlated with that of unaffected channels.
- **NPROLL.** For each $c \in \mathcal{S}_{\mathcal{A}}$, the values of channel c within $\mathcal{A}$ are circularly shifted by a random offset drawn independently and uniformly in $\{1, \dots, |\mathcal{A}| - 1\}$. The segment's

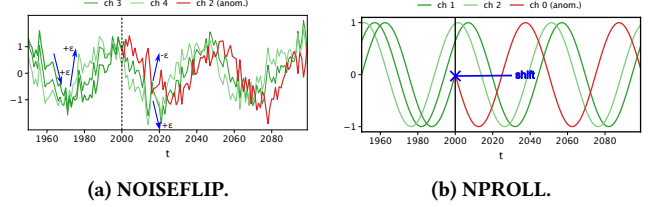


Figure 5: Synthetic protocols that generate cross-channel-only anomalies. NPROLL has its noise removed for clearer view.

Table 3: Classification counts on the synthetic protocol, aggregated over 4,000 segments per cell (1,000 seeds $\times$ 4 segments). P, S, D denote Pearson, Spearman, and dCor_U^2 . UNIVARIATE and UNDETECTED counts are zero in every cell and therefore omitted.

Category	NOISEFLIP			NPROLL		
	P	S	D	P	S	D
CROSS-CHANNEL	3 967	3 967	3 967	3 970	3 970	3 970
BOTH	33	33	33	30	30	30

per-channel marginal is preserved, but both the phase of the sinusoidal component and the shared noise are temporally misaligned with unaffected channels.

Both injection methods preserve every per-channel marginal distribution while breaking cross-channel structure. No univariate test operating on a single channel can flag the resulting segments, since each channel, whether in $\mathcal{S}_{\mathcal{A}}$ or not, exhibits the same per-segment marginal as in $\mathcal{H}_{\mathcal{A}}$. The only evidence available is the alteration of cross-channel dependence. These two protocols therefore provide a controlled cross-channel-only setting, in which our diagnostic should classify every anomalous segment as CROSS-CHANNEL.

5.2 Diagnostic Validation

For each of the two injection methods of Section 5.1, we generate 1 000 independent test sets. Each test set contains 4 anomalous segments, for a total of 4 000 segments per injection method. Every segment is then classified by the diagnostic of Section 3, under each of the three correlation methods, with default thresholds $\tau_z = 3$ and $\tau_\rho = 0.1$. We expect the classification to be CROSS-CHANNEL on every segment, since both methods break cross-channel structure while preserving every per-channel marginal.

Across the 24,000 classifications, 99.2% of segments are classified as CROSS-CHANNEL and 0.8% as BOTH, never UNIVARIATE nor UNDETECTED. NOISEFLIP produces moderate ruptures ($\Delta_\rho^{\max} \in [0.33, 0.40]$) while NPROLL produces stronger ones (≈ 0.62 for dCor , ≈ 1.5 for Pearson/Spearman), all correctly classified. The average $z_{\max} \approx 2.52$ matches the value expected on a non-corrupted segment under the sinusoidal generative process, confirming marginal preservation. The residual BOTH segments result from rare timesteps where the shared Gaussian noise ε_t happens to exceed

Table 4: Detection results on the synthetic protocol, mean $\pm$ standard deviation over 48 runs on 3 seeds.

Model	Var.	NPROLL		NOISEFLIP	
		AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
<i>Channel-dependent</i>					
LinearAE	CD	0.97 ± 0.05	0.89 ± 0.16	0.84 ± 0.17	0.64 ± 0.35
AE	CD	0.97 ± 0.05	0.88 ± 0.16	0.81 ± 0.18	0.59 ± 0.33
CrossAD	CD	0.85 ± 0.09	0.61 ± 0.19	0.55 ± 0.04	0.21 ± 0.02
<i>Channel-partiality</i>					
CATCH	CP	0.77 ± 0.11	0.37 ± 0.20	0.66 ± 0.11	0.27 ± 0.12
<i>Channel-independent</i>					
LinearAE	CI	0.57 ± 0.06	0.17 ± 0.02	0.50 ± 0.01	0.15 ± 0.01
AE	CI	0.60 ± 0.06	0.18 ± 0.03	0.52 ± 0.02	0.16 ± 0.01
CrossAD	CI	0.53 ± 0.02	0.16 ± 0.01	0.47 ± 0.01	0.15 ± 0.01

τ_z standard deviation by chance. The near-total absence of CROSS-CHANNEL segments observed in Section 4.3 on the eight real benchmarks therefore reflects the actual composition of the labeled anomalies.

5.3 Detection on Synthetic Anomalies

We use the synthetic protocol to compare channel-dependent (CD), channel-independent (CI), and channel-partiality architectures. Detectors operate on fixed-length windows: we denote a window by $\mathbf{W} \in \mathbb{R}^{L \times C}$ and its c -th channel by $\mathbf{w}_c \in \mathbb{R}^L$, where L is the window length. We evaluate four reconstruction-based detectors, two of which are recent SOTA: **LinearAE** (flattened linear autoencoder), **AE** (two-layer MLP with GELU), **CrossAD** [12] (originally CI, for which we also implemented a CD variant), and **CATCH** [32] (CP). We pick CrossAD because its native CI design allows minimal modifications to obtain a CD variant (full implementation in the released code). For LinearAE, AE, and CrossAD, the CD variant flattens the window across channels before encoding while the CI variant shares the encoder across channels.

For each (model, corruption) pair, we sweep latent dimensions $d_{\text{model}} \in \{16, 32, 64, 128\}$ and window sizes $L \in \{16, 32, 64, 128\}$ (16 configurations) and repeat each on three random seeds. CrossAD and CATCH use their authors' default training configurations, while LinearAE and AE train for 30 epochs. All models train on the same 20 000 anomaly-free timesteps, with stride 1 windows for training and non-overlapping windows for evaluation, so every test timestep receives exactly one reconstruction error. Experiments ran on an NVIDIA A100 (80 GB).

We report only AUC-ROC and AUC-PR. The range-based and volume-based extensions of Paparrizos et al. [19] were designed to handle range anomalies and label imprecision near anomaly boundaries; both are controlled here.

All three CI configurations and both protocols yield AUC-ROC in $[0.47, 0.60]$ and AUC-PR in $[0.15, 0.18]$ with little variance (see Table 4). As expected from their respective inductive biases, CI models cannot detect these anomalies regardless of capacity. Note that we report these results not as a contribution but as a sanity check that our protocol indeed isolates cross-channel structure.

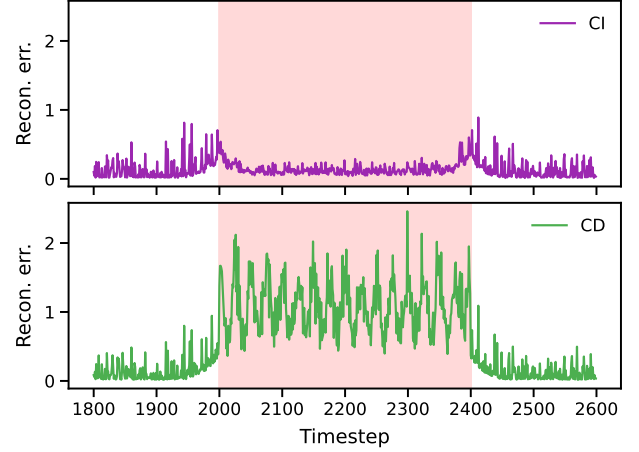


Figure 6: Reconstruction losses from LinearAE (CI and CD) on an NPROLL anomaly segment. CI captures the boundaries but not the whole segment whereas CD fully captures it. Latent dimension is 16 and window size is 64.

LinearAE-CD and AE-CD reach mean AUC-PR of 0.89 and 0.88 on NPROLL, with best configurations reaching 0.99, well above CrossAD-CD (0.61) and CATCH (0.37). The same conclusions hold for NOISEFLIP. A flattened linear autoencoder across channels is enough to reach high performance. CATCH reaches performance in between CI and CD approaches. The large standard deviation of LinearAE (CD) and AE (CD) on NOISEFLIP reflects a sharp diagonal transition in the (d, L) grid: AUC-ROC plateaus near 1.0 when $d_{\text{model}} \geq L$ and collapses below 0.7 otherwise, mixing near-perfect and near-random runs across the 16 configurations.

NPROLL is easier to detect than NOISEFLIP as it introduces sharp discontinuities at the segment boundaries, which a reconstruction-based detector easily flags. NOISEFLIP leaves both the sinusoidal signal and the segment endpoints intact, with only the small shared-noise component flipped on affected channels. We can observe this phenomenon in the reconstruction losses (see Figure 6).

A flattened linear autoencoder outperforming two SOTA models, including one specifically designed for cross-channel modeling, raises a concern: if complex channel-aware mechanisms do not help even where cross-channel modeling is necessary, what do they contribute on benchmarks where every labeled anomaly is already univariately detectable (Section 4)?

6 Detection on Real Benchmarks

From the evaluations reported in Section 4, we could hypothesize that CI and CD variants of the same model should behave similarly on real benchmarks, since no labeled anomaly requires cross-channel modeling for detection. We test this prediction directly by training CrossAD [12] in both its original CI form and our CD variant on the same datasets as in the original CrossAD evaluation (seven of the eight benchmarks of Section 4, WADI excluded). Following the CrossAD protocol, we only retain the first channel of MSL and SMAP. We train both variants with the default

Table 5: CrossAD (CI) against our counterpart CrossAD (CD). Best in bold, with ties bolded on both sides. Averaged over 3 seeds.

Dataset	VUS-ROC		VUS-PR	
	CI	CD	CI	CD
PSM	0.7401	0.7364	0.5404	0.5388
MSL	0.8091	0.8091	0.3140	0.3140
SMD	0.8586	0.8581	0.2308	0.2285
SMAP	0.5784	0.5784	0.1444	0.1444
GECCO	0.9948	0.7667	0.6215	0.0814
SWAN	0.9554	0.9554	0.9278	0.9278
SWaT	0.6006	0.5938	0.3881	0.3868

configuration of CrossAD for each dataset, using their codebase directly (and therefore do not duplicate it in our own repository). Table 5 reports the results on VUS metrics [19] as recommended by Liu and Paparrizos [13] for real-world settings where anomaly boundaries are often uncertain.

Despite its ability to recover cross-channel ruptures on the synthetic protocol of Section 5.3, the CD variant brings no gain over the CI variant on any of the seven real benchmarks. MSL and SMAP yield perfect ties, since CD is exactly identical to CI on univariate time series. SWAN also produces identical scores under both variants, without an obvious explanation. On PSM, SMD, SWaT, and GECCO, CD never improves over CI. However, GECCO is a sharper failure: CI reaches VUS-PR = 0.62 against 0.08 for CD, with a similar gap on VUS-ROC (0.99 versus 0.77). A grid sweep over $(L, d_{\text{model}}) \in \{16, 32, 64, 128\}^2$ reveals that CI’s VUS-PR rises from 0.07 to 0.62 across this range, while CD stays between 0.07 and 0.08. Added capacity benefits CI but not CD, mirroring the capacity-robustness signature reported by Han et al. [9] in forecasting, where increasing the look-back window improves CI but degrades CD. This pattern is exactly what we predicted in Section 4: if no labeled segment carries a cross-channel signature beyond what is already univariately detectable, the extra parameters CD uses to model cross-channel interaction have nothing to exploit, and can even lead to performance collapse.

Similar observations can be seen in the results reported by ChInf [29] on five of our benchmarks. Their contribution improves SOTA results, despite a structurally channel-independent design: its detection score (Eq. 4 in their paper) is the maximum, across channels, of the squared gradient norm of the per-channel reconstruction loss with respect to model parameters, and therefore never couples two channels. This reiterates our observations that CI strategies outperform CD strategies on these multivariate benchmarks.

A natural extension of this experiment would be to inspect, segment by segment, which anomalies are recovered by CD but missed by CI (and conversely) rather than comparing aggregate scores alone. We leave such explorations for future work.

7 Conclusion and Future Work

We introduced a per-segment diagnostic that classifies each labeled anomaly as univariate, cross-channel, both, or undetected, applied it to eight standard MTSAD benchmarks, and validated it

on a controlled synthetic protocol. Across the 373 long segments, none falls in the strictly cross-channel category under any of the three correlation measures, across a sensitivity sweep over both thresholds within their reasonable range. At least half of the labeled anomaly segments deviate univariately between 89% and 100% of their timesteps on six of the eight benchmarks, and the synthetic protocol confirms that the diagnostic identifies cross-channel ruptures when present, with channel-dependent models (even LinearAE) recovering them while channel-independent variants cannot. On real benchmarks, the channel-dependent variant of CrossAD brings no gain over its channel-independent counterpart. The methodological consequence is that no labeled anomaly in the MTSAD benchmarks used in this study requires cross-channel modeling for detection across the reasonable range of our sensitivity sweep. Apparent gains from channel-aware architectures on these benchmarks therefore admit a univariate explanation, and these benchmarks cannot reflect whether channel-aware modeling is actually useful or not.

Wu and Keogh [31] called for abandoning NASA and SMD in 2021, and Wagner et al. [28] independently rejected SWaT, WADI, MSL, and SMAP for MTSAD evaluation, flagging five of the eight datasets we evaluate. Yet they continue to dominate the MTSAD literature, suggesting that recommendations alone do not shift practice. Moreover, the qualitative claim of Garg et al. [7] that these benchmarks contain primarily temporal anomalies, echoed more recently by Wenig et al. [30], has remained without a quantitative analysis for four years. We reiterate their recommendations for the multivariate setting: new datasets with documented anomaly mechanisms are required if channel-aware modeling needs to be evaluated, and per-segment visualizations should accompany any reported metric. More fundamentally, aggregate metrics (F1, AUC, VUS) compress per-segment behavior into a single number and obscure whether a new method actually detects anomalies that previous methods miss. We therefore endorse previous recommendations to visualize data and algorithmic outputs, and extend them: any claim that a new architecture improves over a prior one should be supported by a per-segment analysis of *which* segments are newly recovered, not by a delta on an aggregate score alone.

The evaluation is restricted to labeled anomalies: if a benchmark contains genuine cross-channel events that were never labeled, our diagnostic cannot recover them. We do not claim that cross-channel modeling is unnecessary in real-world multivariate systems, but rather that many current benchmarks cannot be used to evaluate performance for cross-channel anomaly detection.

Finally, our diagnostic can serve as a structural validation test for any future MTSAD benchmark claiming to test channel-aware modeling, by checking whether its labeled segments include CROSS-CHANNEL anomalies or at least exhibit a low univariate ratio.

Two prerequisites would amplify this shift. First, dataset providers should release channel semantics, currently anonymized on most benchmarks, which alone prevents any interpretation of cross-channel dependence. Second, timestep-level labels $\mathbf{y} \in \{0, 1\}^T$ should be replaced by per-channel labels $\mathbf{Y} \in \{0, 1\}^{T \times C}$ documenting, at each anomalous timestep, which channels carry the anomaly. These changes could pave the way for anomaly characterization and explanation [6].

References

- [1] Ahmed Abdulaal, Zhuanghua Liu, and Tomer Lancewicki. 2021. Practical Approach to Asynchronous Multivariate Time Series Anomaly Detection and Localization. In *KDD (Virtual Event Singapore, 2021-08-14)*. ACM, 2485–2494. doi:10.1145/3447548.3467174
- [2] Chuadhyr Mujeeb Ahmed, Venkata Reddy Palleti, and Aditya P. Mathur. 2017. WADI: A Water Distribution Testbed for Research in the Design of Secure Cyber Physical Systems. In *CySWater (Pittsburgh Pennsylvania, 2017-04-21)*. ACM, 25–28. doi:10.1145/3055366.3055375
- [3] Rafal A. Angryk, Petrus C. Martens, Berkay Aydin, Dustin Kempton, Sushant S. Mahajan, Sunitha Basodi, Azim Ahmadzadeh, Xumin Cai, Soukaina Filali Boubrahimi, Shah Muhammad Hamdi, Michael A. Schuh, and Manolis K. Georgioulis. 2020. Multivariate Time Series Dataset for Space Weather Data Analytics. 7, 1 (2020), 227. doi:10.1038/s41597-020-0548-x
- [4] Xiao Chen, Tongxiang Li, Zuoqiao Ma, Jing Chen, Jingfeng Guo, and Zhiliang Liu. 2025. CiTranGAN: Channel-Independent Based-Anomaly Detection for Multivariate Time Series Data. 14, 9 (2025), 1857. doi:10.3390/electronics14091857
- [5] Ailin Deng and Bryan Hooi. 2021. Graph Neural Network-Based Anomaly Detection in Multivariate Time Series. In *AAAI (2021)*, Vol. 35. 4027–4035. doi:10.1609/aaai.v35i5.16523
- [6] Audrey Der, Chin-Chia Michael Yeh, Yan Zheng, Junpeng Wang, Zhongfang Zhuang, Liang Wang, Wei Zhang, and Eamonn Keogh. 2024. *PUPAE: Intuitive and Actionable Explanations for Time Series Anomalies*. 37–45. doi:10.1137/1.9781611978032.5
- [7] Astha Garg, Wenyu Zhang, Jules Samaran, Ramasamy Savitha, and Chuan-Sheng Foo. 2022. An Evaluation of Anomaly Detection and Diagnosis in Multivariate Time Series. 33, 6 (2022), 2508–2517. doi:10.1109/TNNLS.2021.3105827
- [8] Jonathan Goh, Sridhar Adepu, Khurum Nazir Junejo, and Aditya Mathur. 2017. A Dataset to Support Research in the Design of Secure Water Treatment Systems. In *CRITIS (Cham, 2017) (Lecture Notes in Computer Science, Vol. 10242)*. Springer International Publishing, 88–99. doi:10.1007/978-3-319-71368-7_8
- [9] Lu Han, Han-Jia Ye, and De-Chuan Zhan. 2024. The Capacity and Robustness Trade-Off: Revisiting the Channel Independent Strategy for Multivariate Time Series Forecasting. 36, 11 (2024), 7129–7142. doi:10.1109/TKDE.2024.3400008
- [10] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. 2018. Detecting Spacecraft Anomalies Using LSTMs and Non-parametric Dynamic Thresholding. In *KDD (London United Kingdom, 2018-07-19)*. ACM, 387–395. doi:10.1145/3219819.3219845
- [11] Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. 2022. Towards a Rigorous Evaluation of Time-Series Anomaly Detection. In *AAAI (2022-06-30)*, Vol. 36. 7194–7201. doi:10.1609/aaai.v36i7.20680
- [12] Beibu Li, Qichao Shentu, Yang Shu, Hui Zhang, Ming Li, Ning Jin, Bin Yang, and Chenjuan Guo. 2025. CrossAD: Time Series Anomaly Detection with Cross-Scale Associations and Cross-Window Modeling. In *NeurIPS (2025)*. <https://openreview.net/forum?id=fkYFDxlpI>
- [13] Qinghua Liu and John Paparrizos. 2024. The Elephant in the Room: Towards a Reliable Time-Series Anomaly Detection Benchmark. In *NeurIPS (2024)*, Vol. 37. Curran Associates, Inc., 108231–108261. doi:10.52202/079017-3437
- [14] Blanca E. Monroy-Castillo, M. Amalia Jácome, and Ricardo Cao. 2025. Improved distance correlation estimation. *Applied Intelligence* 55, 4 (2025), 263. doi:10.1007/s10489-024-05940-x
- [15] Steffen Moritz, Frederik Rehbach, Sowmya Chandrasekaran, Margarita Rebollo, and Thomas Bartz-Beielstein. 2018. *GECCO Industrial Challenge 2018 Dataset: A Water Quality Dataset for the 'Internet of Things: Online Anomaly Detection for Drinking Water Quality' Competition at the Genetic and Evolutionary Computation Conference 2018, Kyoto, Japan*. doi:10.5281/ZENODO.3884398
- [16] Md Mahmuddun Nabi Murad and Yasin Yilmaz. 2026. Cluster-Aware Causal Mixer for Online Anomaly Detection in Multivariate Time Series. In *ICML (2026)*. arXiv:2506.00188 [cs] doi:10.48550/arXiv.2506.00188
- [17] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A Time Series Is Worth 64 Words: Long-Term Forecasting with Transformers. In *ICLR (2023)*. <https://openreview.net/forum?id=Jbdc0vTOcol>
- [18] John Paparrizos, Paul Boniol, Qinghua Liu, and Themis Palpanas. 2025. Advances in Time-Series Anomaly Detection: Algorithms, Benchmarks, and Evaluation Measures. In *KDD (Toronto, ON, Canada, 2025-08-03)*. ACM, 6151–6161. doi:10.1145/3711896.3736565
- [19] John Paparrizos, Paul Boniol, Themis Palpanas, Ruey S. Tsay, Aaron Elmore, and Michael J. Franklin. 2022. Volume under the Surface: A New Accuracy Evaluation Measure for Time-Series Anomaly Detection. 15, 11 (2022), 2774–2787. doi:10.14778/3551793.3551830
- [20] Xiangfei Qiu, Hanyin Cheng, Xingjian Wu, Junkai Lu, Jilin Hu, Chenjuan Guo, Christian S. Jensen, and Bin Yang. 2026. *A Comprehensive Survey of Deep Learning for Multivariate Time Series Forecasting: A Channel Strategy Perspective*. arXiv:2502.10721 [cs] doi:10.48550/arXiv.2502.10721
- [21] M. Saquib Sarfraz, Mei-Yen Chen, Lukas Layer, Kunyu Peng, and Marios Koulikakis. 2024. Position: Quo Vadis, Unsupervised Time Series Anomaly Detection?. In *ICML (2024) (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 43461–43476. <https://proceedings.mlr.press/v235/sarfraz24a.html>
- [22] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. 2022. Anomaly Detection in Time Series: A Comprehensive Evaluation. 15, 9 (2022), 1779–1797. doi:10.14778/3538598.3538602
- [23] Qichao Shentu, Beibu Li, Kai Zhao, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. 2025. Towards a General Time Series Anomaly Detector with Adaptive Bottlenecks and Dual Adversarial Decoders. In *ICLR (2025)*. <https://openreview.net/forum?id=aKcd7ImG5e>
- [24] Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. 2019. Robust Anomaly Detection for Multivariate Time Series through Stochastic Recurrent Neural Network. In *KDD (Anchorage AK USA, 2019-07-25)*. ACM, 2828–2837. doi:10.1145/3292500.3330672
- [25] Yang Sun, Ning Zhang, Chun Zhang, Minghui Wang, and Chenhao Shi. 2025. DBAD: A Dual-Branch Time Series Anomaly Detection Method Based on Transformer and MLP. In *AICI (Kuala Lumpur Malaysia, 2025-02-14)*. ACM, 440–446. doi:10.1145/3730436.3730510
- [26] Gábor J. Székely and Maria L. Rizzo. 2013. The Distance Correlation T-test of Independence in High Dimension. 117 (2013), 193–213. doi:10.1016/j.jmva.2013.02.012
- [27] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. 2007. Measuring and Testing Dependence by Correlation of Distances. 35, 6 (2007), 2769–2794. doi:10.1214/009053607000000505
- [28] Dennis Wagner, Tobias Michels, Florian C. F. Schulz, Arjun Nair, Maja Rudolph, and Marius Kloft. 2023. TimeSeAD: Benchmarking Deep Multivariate Time-Series Anomaly Detection. (2023). <https://openreview.net/forum?id=iMmsCl0JsS>
- [29] Muyao Wang, Zeke Xie, Bo Chen, Hongwei Liu, and James Kwok. 2026. Channel Matters: Estimating Channel Influence for Multivariate Time Series. In *NeurIPS*. arXiv:2408.14763 [cs] doi:10.48550/arXiv.2408.14763
- [30] Phillip Wenig, Sebastian Schmidl, and Thorsten Papenbrock. 2024. Anomaly Detectors for Multivariate Time Series: The Proof of the Pudding is in the Eating. In *ICDEW*. 96–101. doi:10.1109/ICDEW61823.2024.00018
- [31] Renjie Wu and Eamonn Keogh. 2021. Current Time Series Anomaly Detection Benchmarks Are Flawed and Are Creating the Illusion of Progress. (2021), 1–1. doi:10.1109/TKDE.2021.3112126
- [32] Xingjian Wu, Xiangfei Qiu, Zhengyu Li, Yihang Wang, Jilin Hu, Chenjuan Guo, Hui Xiong, and Bin Yang. 2025. CATCH: Channel-Aware Multivariate Time Series Anomaly Detection via Frequency Patching. In *ICLR (2025)*. <https://openreview.net/forum?id=m08aK3xxdJ>
- [33] Yiyuan Yang, Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. 2023. DCDetector: Dual Attention Contrastive Representation Learning for Time Series Anomaly Detection. In *KDD (2023-08-06)*. 3033–3045. arXiv:2306.10347 [cs] doi:10.1145/3580305.3599295
- [34] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. 2016. Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets. In *ICDM (Barcelona, Spain, 2016-12)*. IEEE, 1317–1322. doi:10.1109/ICDM.2016.0179
- [35] Israel Zexer and Omri Azencot. 2026. XCTFormer: Leveraging Cross-Channel and Cross-Time Dependencies for Enhanced Time-Series Analysis. (2026). <https://openreview.net/forum?id=TEfyR4t0Tw>
- [36] Hang Zhao, Yujing Wang, Juanyong Duan, Congrui Huang, Defu Cao, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. 2020. Multivariate Time-Series Anomaly Detection via Graph Attention Network. In *ICDM (2020-11)*. IEEE, 841–850. doi:10.1109/ICDM50108.2020.00093
- [37] Lifan Zhao and Yanyan Shen. 2024. Rethinking Channel Dependence for Multivariate Time Series Forecasting: Learning from Leading Indicators. In *ICLR (2024)*. <https://openreview.net/forum?id=JiTVtCUOPs>
- [38] Zhijie Zhong, Zhiwen Yu, Yiyuan Yang, Weizheng Wang, Kaixiang Yang, and C. L. Philip Chen. 2025. PatchAD: A Lightweight Patch-Based MLP-Mixer for Time Series Anomaly Detection. 11 (2025), 3460–3473. doi:10.1109/TBDATA.2025.3596745
- [39] Quan Zhou, Changhua Pei, Fei Sun, Han Jing, Zhengwei Gao, Haiming Zhang, Gaogang Xie, Dan Pei, and Jianhui Li. 2025. KAN-AD: Time Series Anomaly Detection with Kolmogorov–Arnold Networks. In *ICML (2025) (Proceedings of Machine Learning Research, Vol. 267)*. PMLR, 79136–79149. <https://proceedings.mlr.press/v267/zhou25u.html>
- [40] Xiaona Zhou, Constantin Brif, and Ismini Lourentzou. 2026. mTSBench: Benchmarking Multivariate Time Series Anomaly Detection and Model Selection at Scale. *TMLR* (2026). <https://openreview.net/forum?id=8Lfb8HD1WU>