IDEAS: UNIVERSAL FRAMEWORK FOR MULTIVARIATE TIME SERIES ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Anomaly detection in multivariate time series is typically approached with specialized models designed to capture cross-channel dependencies, overlooking algorithms considered strictly univariate. In this work, we revisit this assumption and demonstrate that univariate detectors, when properly adapted, serve as strong and efficient alternatives to natively multivariate models. We present **IDEAS** (**I**ndependent **D**imension **E**nsemble for **A**nomaly **S**coring), a universal framework that enables the application of any anomaly detector to multivariate time series. Specifically, a detector is applied independently to each channel, and the resulting channel-wise scores are combined through a selected strategy to yield an aggregate anomaly score. To capture anomalies expressed through cross-channel relationships, the original channels are augmented with pairwise differences between highly correlated dimensions. Our experiments show that unsupervised univariate detectors adapted with IDEAS consistently outperform strong multivariate baselines spanning classical machine learning, semi-supervised deep learning, and large language models, achieving up to a 3.8× improvement in VUS-PR over the strongest baselines. Furthermore, natively multivariate detectors also improve when deployed within the IDEAS framework. Importantly, IDEAS enables detection of not only the temporal location of anomalies but also the specific channel(s) in which they occur. Finally, IDEAS can significantly accelerate computation by parallelizing a highly efficient univariate detector. These findings establish IDEAS as a highly competitive and scalable alternative to multivariate methods.

1 INTRODUCTION

Multivariate time series are widely collected from monitoring systems in domains such as industry, infrastructure, and scientific applications, where detecting abnormal behavior is important for reliability and diagnostics (Blázquez-García et al., 2021; Garg et al., 2021; Sgueglia et al., 2022; Yang et al., 2023a). Because anomalies may appear in multiple channels or relationships between channels, time series anomaly detection (TSAD) in multivariate settings is often treated as a problem that requires explicit modeling of cross-channel dependencies (Wu et al., 2025; Liu et al., 2025; Kim et al., 2025; Xie et al., 2025). This view has led to the introduction of many specialized multivariate detectors, including classical machine-learning, deep-learning, and more recent large-language-model-based approaches (Zamanzadeh Darban et al., 2024; Schmidl et al., 2022; Zhou et al., 2023).

At the same time, many widely used TSAD algorithms are fundamentally univariate and operate on single-channel inputs (Keogh et al., 2005; Senin et al., 2015; Gao et al., 2020; Braei & Wagner, 2020; Breunig et al., 2000). These methods are often overlooked in multivariate settings because they do not directly model interactions among channels. However, their simplicity, efficiency, and strong single-channel performance raise an important question: *Can univariate detectors, when properly adapted, provide competitive alternatives to natively multivariate models?* In this work, we answer this question in the affirmative and introduce **IDEAS** (Independent Dimension Ensemble for Anomaly Scoring), a universal framework for applying anomaly detectors to multivariate time series. As outlined in Figure 1, IDEAS applies a detector independently to each channel and aggregates the resulting channel-wise anomaly scores into a final multivariate anomaly score. To account for anomalies that arise from cross-channel relationships, we further augment the input space with pairwise difference channels constructed from highly correlated dimensions. This design allows both univariate and multivariate detectors to be deployed within a unified channel-wise framework.

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

Our experiments show that state-of-the-art unsupervised univariate detectors such as Ensemble Grammar Induction (EGI) (Gao et al., 2020) and Matrix Profile (MP) (Zhu et al., 2018), when adapted through IDEAS, outperform state-of-the-art multivariate anomaly detectors. We further find that natively multivariate detectors also benefit from the same framework: when applied through IDEAS, detectors such as CNN (Munir et al., 2018) and PCA (Aggarwal, 2017) improve over their standard multivariate versions. Beyond detection accuracy, IDEAS naturally provides information about which dimensions are likely to contain anomalies and supports embarrassingly parallel processing across channels, enabling efficient computation with lightweight detectors such as EGI and MP. These findings suggest that IDEAS can serve as a scalable and competitive alternative to specialized multivariate approaches. The contributions of this work are summarized as follows:

(1) We introduce IDEAS, a framework that applies arbitrary anomaly detectors, both univariate and multivariate, to multivariate time series by aggregating channel-wise anomaly scores.

(2) We show that IDEAS consistently outperforms strong multivariate baselines across benchmark settings that vary anomaly spread, dimensionality, time-series length, and cross-channel correlation structure, effectively achieving up to a $3.8\times$ increase in VUS-PR.

(3) We demonstrate that IDEAS naturally identifies which channels are likely to contain anomalies, with negligible additional computational cost. Additionally, the channel-wise design of IDEAS enables efficient parallel computation: EGI-IDEAS (the most accurate detector overall) is, on average, about an order of magnitude faster than the most efficient multivariate baseline on time series with 100 channels and 10^4 time points.

2 Related Work

Score-based detection. TSAD is commonly formulated as a score-based problem: a detector assigns an anomaly score to each time point, and binary predictions are obtained by thresholding the resulting scores (Schmidl et al., 2022; Paparrizos et al., 2022b;a; Zhou et al., 2026). This interface is shared by classical outlier detectors (Breunig et al., 2000; Liu et al., 2008), subsequence-based methods (Nakamura et al., 2020; Yeh et al., 2016; Lu et al., 2022), and deep learning methods based on reconstruction or forecasting errors (Xu et al., 2018; Munir et al., 2018). In multivariate TSAD, the input consists of multiple interacting channels, and anomalies may appear in individual channels or through violations of cross-channel dependencies. Existing methods typically produce multivariate anomaly scores by learning temporal and inter-variable structure, using recurrent or autoencoder-based models (Su et al., 2019; Audibert et al., 2020; Li et al., 2021), graph-based dependency models (Deng & Hooi, 2021), attention-based and Transformer methods (Xu et al., 2022; Tuli et al., 2022; Song et al., 2023), or recent contrastive and channel-aware frequency-domain methods (Yang et al., 2023b; Wu et al., 2025). IDEAS instead starts from the same score-based interface, but decouples the base detector from multivariate construction: a detector produces channel-wise scores, while multivariate structure is captured by channel-set construction and score aggregation.

Adapting univariate detectors to multivariate data. Recent empirical studies show that detection performance can vary substantially across datasets, anomaly patterns, thresholding choices, and evaluation metrics (Paparrizos et al., 2022a; Liu & Paparrizos, 2024b). Further analysis of common multivariate TSAD benchmarks suggests that many labeled anomalies contain strong univariate evidence, while cross-channel anomalies are underrepresented (Pinet et al., 2026). Together, these findings motivate adapting strong univariate detectors to multivariate inputs, rather than always designing a new joint architecture. Two key challenges hinder channel-wise adaptation: anomalies affecting only an unknown subset of channels can be diluted across normal dimensions, or they may manifest strictly as altered cross-channel relationships. These issues have been studied in multidimensional extensions of the MP method, which examined how to avoid diluting sparse anomalous evidence across many normal dimensions (Tafazoli & Keogh, 2023; Yeh et al., 2024); however, these approaches are designed specifically for MP-based scoring. In contrast, IDEAS is detector-agnostic and applies to any detector producing point-wise scores.

Anomalous dimension localization. From a practical perspective, predicting anomalous points/intervals on the time axis is often not enough; operators also need to know which variables or channels exhibit abnormal behavior to support diagnosis. This problem is related to anomaly localization, anomaly interpretation, and root-cause analysis, where prior work aims to identify the sensors or variables associated with an anomalous event (Ameli et al., 2022; Yang et al., 2022; Han et al., 2025). These methods often rely on a separate interpretation, attribution, or causal-analysis

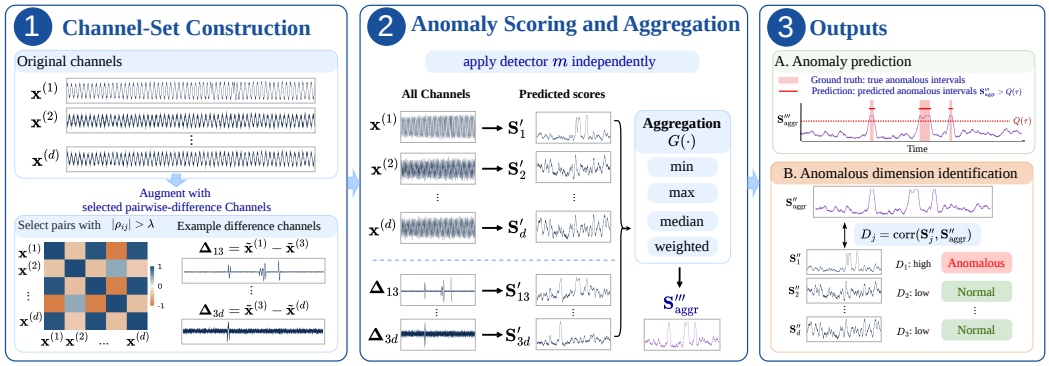


Figure 1: Overview of the IDEAS framework. IDEAS (1) constructs a set of original and pairwise-difference channels, (2) applies an anomaly detector to each channel in this set and aggregates the resulting scores, and (3) outputs both the prediction of temporal anomaly locations and the identification of dimensions that contain anomalies.

procedure after or alongside anomaly detection. In contrast, IDEAS supports temporal anomaly detection and anomalous-dimension identification within the same channel-wise scoring framework, without training an additional localization model.

3 Method

3.1 Problem Definition

A multivariate time series can be represented as a matrix

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix} = \left[\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(d)}\right] \in \mathbb{R}^{n \times d}, \tag{1}$$

where n is the time-series length (the number of time points), d is the number of dimensions (channels), $\mathbf{x}_t \in \mathbb{R}^d$ is the vector of recorded observation values at time point $t \in \{1, \ldots, n\}$, and

$$\mathbf{x}^{(j)} = \left(x_1^{(j)}, \ldots, x_n^{(j)}\right)^\top \in \mathbb{R}^n \tag{2}$$

is the univariate time series associated with dimension j. A multivariate anomaly detector applied to the time series $\mathbf{X}$ produces the vector of anomaly scores

$$\mathbf{S} = (s_1, \ldots, s_n)^\top \in \mathbb{R}^n, \tag{3}$$

where larger values indicate stronger evidence of anomalous behavior. Binary anomaly predictions are most commonly obtained from continuous anomaly scores via thresholding. For a given threshold $\tau \in (0, 1)$, we classify a time point as anomalous if its anomaly score exceeds the τ-quantile of the score vector $\mathbf{S}$. The corresponding binary prediction at time point t is then defined as $\widehat{y}_t(\tau) = \mathbb{I}[s_t > Q_{\mathbf{S}}(\tau)]$, where $Q_{\mathbf{V}}(p)$ denotes the empirical p-quantile of a given vector $\mathbf{V}$.

3.2 IDEAS: Independent Dimension Ensemble for Anomaly Scoring

Many anomaly detectors are designed for univariate inputs and cannot be directly applied to multivariate time series without an adaptation strategy. We introduce IDEAS, a detector-agnostic wrapper that applies a base detector independently to individual channels and aggregates their anomaly scores to obtain a multivariate prediction. Specifically, given a multivariate time series $\mathbf{X}$ and a base detector m, IDEAS constructs a set $\mathcal{C}$ of univariate channels from the original dimensions and pairwise differences between strongly correlated dimensions, applies m independently to each channel $c \in \mathcal{C}$ to obtain a channel-level vector of anomaly scores, $\mathbf{S}_c$, and finally aggregates the resulting channel-level scores $\{\mathbf{S}_c : c \in \mathcal{C}\}$ into the vector of anomaly scores for $\mathbf{X}$.

Channel-set construction. Let $\mathcal{C}$ denote the selected set of univariate channels. By default, this set contains the original dimensions, $\mathcal{C}_{\text{orig}} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(d)}\}$. To expose cross-channel anomalies to univariate detectors, we augment this set with pairwise-difference channels constructed from highly correlated dimensions. For each pair (i, j) satisfying $|\rho_{ij}| > \lambda$, where $\rho_{ij} = \text{corr}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ is the Pearson correlation coefficient between dimensions i and j, and $\lambda \in [0, 1]$ is a correlation threshold, we construct the difference channel $\mathbf{\Delta}_{ij} = \tilde{\mathbf{x}}^{(i)} - \text{sign}(\rho_{ij})\tilde{\mathbf{x}}^{(j)}$, where $\tilde{\mathbf{x}}^{(i)}$ and $\tilde{\mathbf{x}}^{(j)}$ are obtained by scaling: $\tilde{\mathbf{x}}^{(j)} = \left[\mathbf{x}^{(j)} - \text{mean}(\mathbf{x}^{(j)})\right] / \text{stdev}(\mathbf{x}^{(j)})$. The resulting set of channels is

$$\mathcal{C} = \mathcal{C}_{\text{orig}} \cup \{\mathbf{\Delta}_{ij} : |\rho_{ij}| > \lambda\}. \tag{4}$$

If $\lambda = 1$, then pairwise-difference channels are not used, i.e., $\mathcal{C} = \mathcal{C}_{\text{orig}}$. If $\lambda = 0$, then $\mathcal{C}$ includes all $d(d-1)/2$ pairs. Under the realistic assumption that relational anomalies are likely to occur across channels with interdependent normal signals, a high value of λ (e.g., 0.8–0.9) is appropriate.

Anomaly scoring and post-processing. For each channel $c \in \mathcal{C}$, the base detector m produces a channel-wise anomaly-score vector $\mathbf{S}_c = (s_{c,1}, \ldots, s_{c,n})^\top \in \mathbb{R}^n$. The detector is applied independently to each channel, and in each score vector larger values indicate higher anomaly likelihood. To reduce boundary effects, we trim a fixed fraction $\alpha \in [0, 0.1)$ from each end of every score vector. Let $\ell = \lfloor \alpha n \rfloor$ and $n' = n - 2\ell$. The trimmed score vector is $\mathbf{S}'_c = (s_{c,\ell+1}, \ldots, s_{c,n-\ell})^\top \in \mathbb{R}^{n'}$. We then center each trimmed vector by subtracting its mean: $\mathbf{S}''_c = \mathbf{S}'_c - \text{mean}(\mathbf{S}'_c)$. Although its importance may not be immediately apparent, centering is a critical step that amplifies contributions from channels with prominent anomaly-score peaks while suppressing those from featureless channels.

Pointwise score aggregation. The multivariate anomaly score is obtained by pointwise aggregation across channels. For $t = 1, \ldots, n'$,

$$s''_{\text{aggr},t} = G\left(\{s''_{c,t} : c \in \mathcal{C}\}\right), \tag{5}$$

where $G(\cdot)$ is one of the following aggregation operators:

$$\min: \; s''_{\text{aggr},t} = \min_{c \in \mathcal{C}} s''_{c,t}, \tag{6a}$$

$$\max: \; s''_{\text{aggr},t} = \max_{c \in \mathcal{C}} s''_{c,t}, \tag{6b}$$

$$\text{median}: \; s''_{\text{aggr},t} = \text{median}_{c \in \mathcal{C}} s''_{c,t}, \tag{6c}$$

$$\text{weighted}: \; s''_{\text{aggr},t} = \sum_{c \in \mathcal{C}} w_c \, s''_{c,t}, \;\; w_c = \frac{W_c}{\sum_{c \in \mathcal{C}} W_c}, \tag{6d}$$

where $W_c = \text{mean}\{s'_{c,t} : s'_{c,t} > Q_{\mathbf{S}'_c}(\theta)\} / \text{mean}\{s'_{c,t} : s'_{c,t} \leq Q_{\mathbf{S}'_c}(\theta)\}$ and $\theta \in (0, 1)$ is a top-score threshold. Thus, channels whose upper-tail scores are large relative to their lower-score background receive larger weights. Each aggregation operator defines one IDEAS variant for the base detector m.

Due to boundary trimming, $\mathbf{S}''_{\text{aggr}}$ has length $n' = n - 2\ell$. To restore the original time axis, we pad $\mathbf{S}''_{\text{aggr}}$ at both ends with ℓ values equal to $\min(\mathbf{S}''_{\text{aggr}})$, yielding the aggregated score vector $\mathbf{S}'''_{\text{aggr}} \in \mathbb{R}^n$. Binary anomaly predictions are obtained from $\mathbf{S}'''_{\text{aggr}}$ by thresholding.

Anomalous dimension identification. In addition to predicting temporal anomaly locations, we use the channel-wise anomaly scores produced by IDEAS to predict which dimensions contain anomalies. For dimension j, we compute the dimension anomaly score D_j as the Pearson correlation coefficient between the dimension's score vector and the aggregated score vector:

$$D_j = \text{corr}\left(\mathbf{S}''_j, \mathbf{S}''_{\text{aggr}}\right). \tag{7}$$

If a time series is known to have exactly one anomalous dimension, we predict this dimension as $\hat{j} = \arg\max_j D_j$. For a dataset that includes only such time series, we report the top-1 anomalous-dimension prediction (ADP) success rate, i.e., the fraction of time series for which $\hat{j}$ matches the ground-truth anomalous dimension. For a time series with an unknown number of anomalous dimensions, we apply Otsu thresholding (Otsu, 1979) to the set of dimension anomaly scores $\{D_j\}_{j=1}^d$. Dimensions with $D_j > \tau_{\text{Otsu}}$ are predicted as anomalous, and we evaluate dimension-level precision, recall, and F_1, averaged over all time series in a dataset.

Table 1: Summary of synthetic multivariate datasets. d is the total number of dimensions, d_{anom} is the number of dimensions containing anomalies, and n is the time-series length.

Dataset	# series	d	d_{anom}	n	Description
A	500	20	{1, 2, 5, 10, 20}	10^4	Varying the number of anomalous dimensions
B	500	{5, 10, 20, 50, 100}	1	10^4	Varying the total number of dimensions
C	500	{5, 10, 20, 50, 100}	$[1, d]$	10^4	As B but with a random number of anomalous dimensions
D	500	20	1	$\{5, 10, 20, 50, 100\} \times 10^3$	Varying the time-series length
E	500	20	$[1, d]$	$\{5, 10, 20, 50, 100\} \times 10^3$	As D but with a random number of anomalous dimensions
F	300	{5, 10, 20}	$[1, d]$	10^4	Including cross-channel anomalies

4 EXPERIMENTS

4.1 TIME-SERIES DATA

We evaluate IDEAS using six controlled synthetic datasets designed to isolate the factors that determine the difficulty of multivariate TSAD, along with 20 real-world datasets spanning diverse application domains. Synthetic construction provides exact point-level and dimension-level ground truth and allows us to vary the number of anomalous dimensions, the total dimensionality, the time-series length, and the abundance of cross-channel anomalies. These factors are difficult to disentangle in existing real-world benchmarks, where labels may be incomplete, dataset artifacts can complicate evaluation, and many reported anomalies contain strong univariate evidence (Wu & Keogh, 2023; Schmidl et al., 2022; Liu & Paparrizos, 2024b; Pinet et al., 2026).

Data generation. The six synthetic datasets comprise five GutenTAG-based datasets (A–E) with a variety of features (Wenig et al., 2022) and one dataset (F) designed specifically to contain cross-channel anomalies (Pinet et al., 2026). Dataset characteristics are summarized in Table 1. Additional details are provided in Appendix A.3. In datasets A–E, normal signals are randomly sampled from 10 base oscillation families: sine, cosine, square, polynomial, CBF, ECG, random mode jump, sawtooth, Dirichlet, and MLS. Anomalies are randomly sampled from 10 compatible anomaly classes: amplitude, extremum, frequency, mean, pattern, pattern shift, platform, trend, variance, and mode-correlation. Each corrupted dimension contains 1 to 3 anomaly intervals, with their start points sampled from the central 80% of the sequence and lengths sampled from $[50, 500]$ for datasets A–C and from $[0.005n, 0.05n]$ for datasets D and E. If an anomaly occurs across multiple channels, its start point and length are the same for all these dimensions, but the anomaly class and parameters are selected independently and randomly for each dimension. Datasets A–E are designed to test detector performance as anomaly spread, dimensionality, and time-series length vary. Dataset F is generated using a method adapted from (Pinet et al., 2026): normal signals are multivariate sinusoidal with shared noise, and anomalies are injected using noise flipping or local temporal rolling to disrupt the shared cross-channel structure while largely preserving per-channel marginal behavior. The purpose of dataset F is to evaluate detection of cross-channel anomalies.

4.2 IMPLEMENTATION DETAILS

Threshold values. We set $\tau = \theta = 0.925$ and $\alpha = 0.02$ in all experiments, $\lambda = 0.9$ in experiments on datasets A–E, and $\lambda \in \{0.8, 0.9, 1.0\}$ in experiments on dataset F. Since cross-channel anomalies are rare in datasets A–E, the choice of the λ value has a small effect on detection performance for them, while it typically has a significant effect for dataset F.

Anomaly score normalization. We deploy four base detectors within IDEAS: two unsupervised univariate methods (EGI and MP), one unsupervised classical machine-learning model (PCA), and one semi-supervised deep-learning model (CNN). The raw output of EGI for channel c is the grammar-rule-density vector, $\mathbf{R}_c$, whose elements are normalized in $[0, 1]$; it is re-oriented to obtain normalized anomaly scores as $\mathbf{S}_c = \mathbf{1}_n - \mathbf{R}_c$. The raw outputs, $\mathbf{S}_{\text{raw},c}$, produced by MP, PCA, and CNN are not normalized. For MP and PCA, we normalize them as $\mathbf{S}_c = [\mathbf{S}_{\text{raw},c} - \mathbf{1}_n \min(\mathbf{S}_{\text{raw},c})]/[\max(\mathbf{S}_{\text{raw},c}) - \min(\mathbf{S}_{\text{raw},c})]$. For CNN, we do not normalize the raw output.

Aggregation strategy selection. For each base detector, we select a default aggregation strategy from the set $\{\min, \max, \text{median}, \text{weighted}\}$ using the ratio-to-winner (RTW) score described in Appendix A.6. Based on the mean RTW scores across datasets A–F, the default variant is $\max$ for EGI and PCA and weighted for MP and CNN. These results show that the aggregation strategy is a consequential design choice in IDEAS, since the same channel-wise scores can yield different

performance depending on how they are combined. We refer to detector m applied with its default aggregation variant as m-IDEAS, i.e., EGI-IDEAS, MP-IDEAS, CNN-IDEAS, and PCA-IDEAS.

Multivariate baselines. We compare IDEAS against a diverse set of state-of-the-art multivariate anomaly detectors. The classical machine learning baselines include PCA (Aggarwal, 2017), CBLOF (He et al., 2003), IForest (Liu et al., 2008), and KMeansAD (Yairi et al., 2001). The deep-learning baselines include CNN (Munir et al., 2018), LSTMAD (Malhotra et al., 2015), OmniAnomaly (Su et al., 2019), and TranAD (Tuli et al., 2022). Also, recent LLM-based methods ALLM4TS (Bian et al., 2024) and OFA (Zhou et al., 2023) are included. Lastly, we include a baseline with randomly generated predictions as a lower bound. Baseline details are provided in Appendix A.4.

Evaluation. We evaluate detectors using 12 commonly used metrics: Precision, Recall, F_1, AUC-ROC, AUC-PR, AUC-$P_T R_T$ (Tatbul et al., 2018), VUS-ROC, VUS-PR (Paparrizos et al., 2022a), Affiliation-F_1 (Huet et al., 2022), Event-based-F_1 (Garg et al., 2021), Max-F_1, and R-based-F_1 (Tatbul et al., 2018); see Appendix A.5 for details. VUS-PR is used as the main metric, while Appendix A.7 presents detailed results for all 12 metrics. All detectors are evaluated on the same train/test splits and ground-truth labels. For semi-supervised detectors requiring training, an anomaly-free training series is paired with each test series; unsupervised methods are evaluated directly on the test series.

5 Experimental Results

5.1 Performance on Synthetic Datasets

Figure 2 compares the IDEAS-deployed detectors with multivariate baselines on datasets A–E, exploring the effects of anomaly spread (dataset A), total dimensionality (datasets B and C), and time-series length (datasets D and E).

IDEAS is strong across different levels of anomaly spread. The top panel of Figure 2 shows mean VUS-PR values obtained on the groups of dataset A, where the total dimensionality is fixed at $d = 20$ and the number of dimensions containing anomalies, d_{anom}, is different for each group, increasing from 1 to 20. VUS-PR generally increases with d_{anom}, since anomalous evidence becomes less sparse across channels. EGI-IDEAS performs best in all five groups ($d_{\text{anom}} \in \{1, 2, 5, 10, 20\}$), being especially dominant relative to the baselines when anomalies are contained in a small fraction of dimensions. In particular, for $d_{\text{anom}} \in \{1, 2\}$, EGI-IDEAS achieves VUS-PR values more than three times as high as the strongest multivariate baselines. The benefit of the IDEAS formulation is also clear for detectors that already support multivariate inputs: PCA-IDEAS consistently outperforms PCA, and CNN-IDEAS consistently outperforms CNN across all five d_{anom} settings. As anomalies spread to more dimensions, several multivariate baselines become more competitive, especially CNN, PCA, OmniAnomaly, and TranAD at $d_{\text{anom}} = 20$. However, they still underperform all four IDEAS detectors. These results demonstrate that channel-wise scoring by IDEAS is uniquely effective for preserving sparse anomaly evidence, while sustaining superior performance when anomalous signals are distributed across many dimensions.

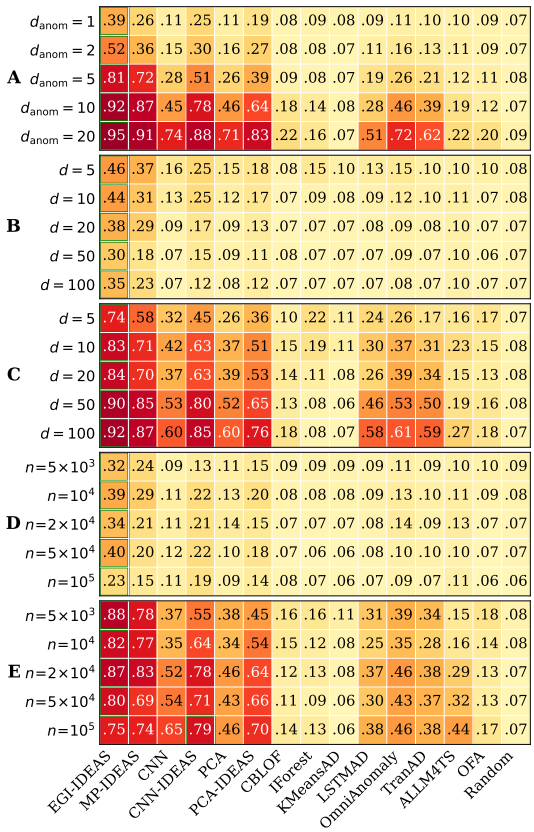


Figure 2: VUS-PR comparison across synthetic dataset groups. Dataset A varies anomalous dimensions; datasets B and C vary the total dimensionality under fixed and random anomalous-dimension settings, respectively; and datasets D and E vary the time-series length under fixed and random anomalous-dimension settings, respectively.

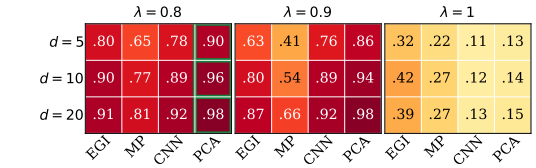
As the total dimensionality grows and one-dimensional anomalies become increasingly sparse, IDEAS dominates. The 2nd panel of Figure 2 shows mean VUS-PR values for dataset B, where the total dimensionality d is different for each group, increasing from 5 to 100. In dataset B, each time series has exactly one anomalous dimension, making the anomalous signal increasingly sparse relative to the normal background, as d grows. This is reflected in VUS-PR values that are low across all detectors and generally decrease as dimensionality increases. EGI-IDEAS and MP-IDEAS are the top and second-best performers, respectively, for all five groups. Notably, for $d \geq 10$, EGI-IDEAS achieves VUS-PR scores more than three times as high as any baseline ($3.8\times$ for $d = 20$). The other IDEAS detectors also outperform the multivariate baselines, showing that independent channel-wise scoring is especially effective when anomalies are confined to a single dimension.

Dataset C also includes five groups with different total dimensionality ($d \in \{5, 10, 20, 50, 100\}$), but the number of anomalous dimensions is a random integer sampled from $[1, d]$. Therefore, anomalous evidence is much less sparse on average, and overall performance is substantially stronger than in dataset B, as shown in the 3rd panel of Figure 2. Across all five groups, EGI-IDEAS performs best and all four IDEAS detectors outperform all multivariate baselines, showing a consistent advantage of independent channel-wise scoring in this setting. Among the baselines, CNN, PCA, OmniAnomaly, and TranAD are the strongest, with OmniAnomaly and TranAD becoming more competitive at larger d. However, across all dimensionality settings considered in our experiments, even the strongest baseline underperforms the weakest IDEAS detector.

Across time-series lengths, IDEAS is strongest, especially when anomalies are sparse. The 4th panel of Figure 2 shows mean VUS-PR values for the groups of dataset D, where the time-series length increases from $n = 5 \times 10^3$ to $n = 10^5$, while the total dimensionality is fixed at $d = 20$. Similarly to dataset B, each time series in dataset D has only one anomalous dimension, which once again results in sparsity of the anomalous signal relative to the normal background and therefore leads to low VUS-PR values across all detectors. EGI-IDEAS performs best across all five groups. The other three IDEAS detectors also outperform all multivariate baselines across all evaluated time-series lengths, although by a smaller margin than the dominant EGI-IDEAS method.

Dataset E also includes five groups with different time-series lengths that increase from $n = 5 \times 10^3$ to $n = 10^5$, but the number of anomalous dimensions is a random integer sampled from $[1, d]$. Therefore, anomalous signals span more dimensions on average, and VUS-PR values are much higher than in dataset D, as seen in the 5th panel of Figure 2. EGI-IDEAS performs best for $n \in \{5, 10, 20, 50\} \times 10^3$, while CNN-IDEAS performs best for $n = 10^5$. All four IDEAS detectors exhibit strong performance and improve upon all multivariate baselines across all groups. Among the baselines, CNN performs the best, demonstrating particularly strong results for $n \in \{50, 100\} \times 10^3$, but it still falls short of all four IDEAS detectors.

Pairwise-difference channels help detect cross-channel anomalies. Figure 3 shows mean VUS-PR values for the groups of dataset F, where three groups have different total dimensionality values, $d \in \{5, 10, 20\}$, and the number of anomalous dimensions is sampled from $[1, d]$. The key feature of dataset F is that all anomalies in its time series disrupt cross-channel dependencies while having a smaller effect on the marginal behavior of individual channels.



(a) Effect of the correlation threshold λ.



(b) VUS-PR comparison with multivariate baselines.

Figure 3: Results on dataset F. (a) VUS-PR comparison of the IDEAS detectors for different values of λ. (b) VUS-PR comparison across all detectors (with the IDEAS detectors using $\lambda = 0.8$).

Figure 3a presents the mean VUS-PR results for the four IDEAS detectors across different values of the correlation threshold, $\lambda \in \{0.8, 0.9, 1.0\}$. The smaller the λ value, the more pairwise difference channels are included in the set $\mathcal{C}$. The performance improves drastically as λ decreases from 1.0 (where only original channels are included) to 0.9. An additional improvement (although less drastic, especially for CNN-IDEAS and PCA-IDEAS) is gained as λ decreases from from 0.9 to

0.8. This result indicates that including additional difference channels helps expose the disrupted cross-channel structure, especially for EGI-IDEAS and MP-IDEAS.

Figure 3b compares the IDEAS detectors (with $\lambda = 0.8$) against multivariate baselines. PCA-IDEAS is the top performer in all groups, surpassing the other IDEAS methods and all baselines. Both PCA-IDEAS and CNN-IDEAS improve substantially over their original counterparts, PCA and CNN, respectively. Among the baselines, OFA and CBLOF are the strongest performers, but fall short of EGI-IDEAS and CNN-IDEAS for all groups, except for $d = 5$ where OFA and CNN-IDEAS are tied.

5.2 Performance on Real-World Datasets

The results on real-world datasets, shown in Figure 4, exhibit significantly greater variance than the experiments on synthetic benchmarks. The heatmap of mean VUS-PR scores across 20 datasets illustrates that most detectors yield low-to-moderate scores, punctuated by isolated performance spikes on particular datasets. Consequently, no single detector achieves consistently strong performance across the full benchmark. The final row in Figure 4 reports mean RTW scores (defined in Appendix A.8), which show that *on average* three detectors are virtually tied for the first rank: PCA, OmniAnomaly, and EGI-IDEAS, followed by PCA-IDEAS and OFA. As shown in Appendix A.8, the volatility persists within the TSB-AD-M benchmark (Liu & Paparrizos, 2024a), where detector performance varies substantially across its source datasets. This contrasts sharply with the results on the synthetic datasets, where the IDEAS detectors maintain consistent performance under controlled anomaly-generation settings. The instability observed on real-world benchmarks likely reflects inherent dataset artifacts such as noisy or incomplete labels, imprecise anomaly boundaries, and inconsistent anomaly definitions across sources, rather than true differences in detector capability. Therefore, performance on real-world benchmarks should be interpreted with caution, as isolated score fluctuations may not accurately reflect a detector's fundamental efficacy.

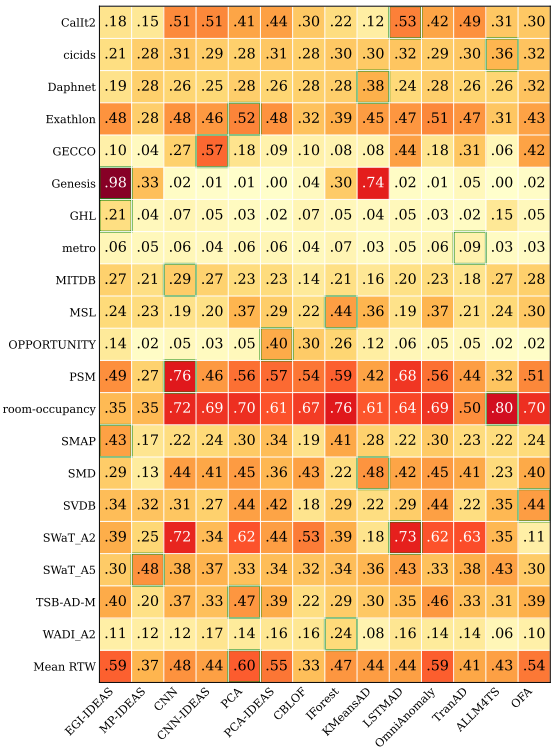


Figure 4: Mean VUS-PR on real-world datasets.

5.3 Anomalous Dimension Identification

Channel-wise scoring provides a natural way to predict which dimensions contain anomalies. Unlike detectors that only return a single anomaly score per time point, IDEAS preserves channel-wise scores before aggregation, which provides a straightforward way to predict anomalous channels.

Table 2 reports the ADP performance of the IDEAS detectors, along with a random prediction baseline, for datasets A–E (see Appendix A.9 for more details). Across all datasets, EGI-IDEAS achieves the strongest performance. CNN-IDEAS is ranked second on dataset A, MP-IDEAS on datasets B and D, and MP-IDEAS and PCA-IDEAS are vir-

Table 2: ADP performance of the IDEAS detectors, compared to a random predictor. Top-1 ADP success rate is reported for datasets B and D, which contain a single anomalous dimension, and dimension-level F_1 for datasets A, C, and E.

Dataset	Metric	EGI	MP	CNN	PCA	Random
A	Dim. F_1	**0.534**	0.324	0.429	0.382	0.352
B	Top-1 ADP	**0.376**	0.326	0.260	0.122	0.076
C	Dim. F_1	**0.659**	0.561	0.521	0.561	0.475
D	Top-1 ADP	**0.406**	0.248	0.238	0.145	0.050
E	Dim. F_1	**0.670**	0.513	0.488	0.512	0.467

tually tied for the second rank on datasets C and E. For datasets with a variable number of anomalous dimensions per time series, dimension-level F_1 scores are higher on datasets C and E than on dataset A, suggesting that channel-level evidence is easier to recover when anomalies are less sparse across dimensions. These results show that the channel-wise scoring by IDEAS supports both temporal anomaly localization and anomalous dimension identification.

5.4 Runtime Analysis

Runtime results are presented as "violin plots" in Figure 5, where the shaded shapes represent the runtime distribution density and the markers represent the median runtime value for each dataset group. The top panel shows combined results for datasets B and C to evaluate scaling with the dimensionality d, and the bottom panel shows combined results for datasets D and E to evaluate scaling with the time-series length n. The distribution for each detector is based on 1000 runs (500 runs per dataset that include 100 runs per group in a dataset). We observe that MP-IDEAS and EGI-IDEAS are consistently the fastest methods, with runtimes generally in the 1–100 second range, while most multivariate baselines are concentrated in the 100–10,000 second range. This gap is especially clear compared with CNN, LSTMAD, TranAD, ALLM4TS, and OFA, whose runtimes are one to several orders of magnitude longer. The runtime advantage of MP-IDEAS and EGI-IDEAS is expected because these methods apply highly efficient univariate detectors in parallel across independent dimensions, avoiding the cost of fitting a single joint model over the full multivariate input. The runtime advantage is particularly clear for the d-scaled setting, where MP-IDEAS and EGI-IDEAS remain much faster than most baselines as d increases. In the n-scaled setting, their runtime increases with n, but they remain competitive with or faster than most alternatives. In contrast, CNN-IDEAS and PCA-IDEAS inherit the higher cost of their base detectors, showing that the runtime of IDEAS depends on the underlying base detector. Overall, these results demonstrate that IDEAS can achieve significant computational speedups when paired with efficient base detectors. Detailed runtime results are provided in Appendix A.10 for all evaluated settings.

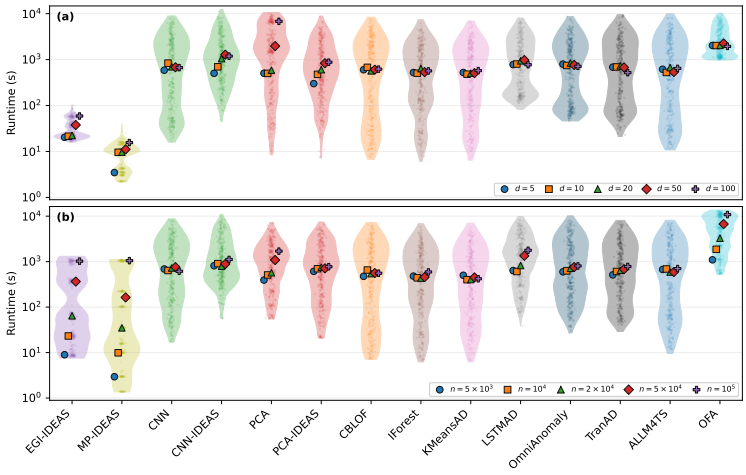


Figure 5: Runtime comparison across detectors. The shaded shapes (the "violins") represent the runtime distributions, and the markers represent the median runtime for each group. (a) Results for datasets B and C as d varies. (b) Results for datasets D and E as n varies.

6 Conclusion

This work challenges the common assumption that multivariate TSAD requires methods that explicitly model all channels jointly. We introduce IDEAS, a universal framework that enables the application of any univariate or multivariate anomaly detector to multivariate time series. Specifically, IDEAS applies a base detector independently to both the original and pairwise-difference channels, then aggregates the resulting anomaly scores. Across six controlled synthetic datasets, univariate detectors (EGI and MP) and multivariate detectors (CNN and PCA) deployed within the IDEAS framework consistently outperform strong multivariate baselines. In terms of average performance on 20 real-world datasets, EGI-IDEAS is among three detectors virtually tied for the first rank. In addition to improving temporal anomaly localization, IDEAS provides dimension-level anomaly evidence. Furthermore, it is naturally amenable to parallel computation, yielding practical speedups when efficient base detectors are used. These results suggest that simple, modular, and efficient channel-wise adaptation is a highly competitive alternative to specialized multivariate anomaly detectors.

AI USE STATEMENT

Generative AI tools were used only for minor text editing to improve grammar, clarity, and readability. We did not use generative AI tools to develop the methodology, design experiments, implement methods, analyze data, or interpret results. All AI-assisted edits were reviewed and verified by the authors. We take responsibility for the final content of this work, including any text produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects or personally identifiable information. Our experiments use publicly available and synthetically generated time-series datasets in accordance with their intended research use. The proposed methods are designed to improve anomaly detection and anomalous-dimension localization, with potential applications in monitoring and failure diagnosis. As with other anomaly detection systems, predictions may be imperfect and should not be treated as a substitute for expert judgment in safety-critical settings. We encourage users to validate model outputs and consider domain-specific risks before deployment.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce our results in the main paper and the appendix. The IDEAS methodology is described in section 3 and Appendix A.1; dataset construction, preprocessing, benchmark settings, evaluation metrics, detector configurations, and implementation details are documented in section 4 and Appendices A.3 – A.6; and analyses of obtained results are provided in section 5 and Appendices A.7 – A.10. All new synthetic datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes.

REFERENCES

Ahmed Abdulaal, Zhuanghua Liu, and Tomer Lancewicki. Practical approach to asynchronous multivariate time series anomaly detection and localization. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2485–2494, 2021. doi: 10.1145/3447548.3467174.

Charu C. Aggarwal. *Outlier Analysis*. Springer International Publishing, 2nd edition, 2017. doi: https://doi.org/10.1007/978-3-319-47578-3.

Mina Ameli, Viktor Pfanschilling, Anar Amirli, Wolfgang Maaß, and Kristian Kersting. Unsupervised multi-sensor anomaly localization with explainable AI. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 507–519. Springer, 2022.

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A Zuluaga. USAD: unsupervised anomaly detection on multivariate time series. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3395–3404, 2020. doi: 10.1145/3394486.3403392.

Marc Bachlin, Meir Plotnik, Daniel Roggen, Inbal Maidan, Jeffrey M Hausdorff, Nir Giladi, and Gerhard Troster. Wearable assistant for Parkinson's disease patients with the freezing of gait symptom. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):436–446, 2009. doi: 10.1109/TITB.2009.2036165.

Yuxuan Bian, Xuan Ju, Jiangtong Li, Zhijian Xu, Dawei Cheng, and Qiang Xu. Multi-Patch Prediction: adapting language models for time series representation learning. In *International Conference on Machine Learning*, pp. 3889–3912, 2024.

Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A Lozano. A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)*, 54(3):1–33, 2021. doi: 10.1145/3444690.

Mohammad Braei and Sebastian Wagner. Anomaly detection in univariate time-series: a survey on the state-of-the-art. *arXiv preprint arXiv:2004.00433*, 2020.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. LOF: identifying density-based local outliers. *ACM SIGMOD International Conference on Management of Data*, 29(2): 93–104, 2000. doi: 10.1145/335191.335388.

Canadian Institute for Cybersecurity. CIC-IDS-2017 Dataset. `http://205.174.165.80/CICDataset/CIC-IDS-2017/Dataset/CIC-IDS-2017/`, 2017. Accessed: 2025-05-10.

Luis M Candanedo and Véronique Feldheim. Accurate occupancy detection of an office room from light, temperature, humidity and CO_2 measurements using statistical learning models. *Energy and Buildings*, 112:28–39, 2016. doi: https://doi.org/10.1016/j.enbuild.2015.11.071.

Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *AAAI Conference on Artificial Intelligence*, volume 35, pp. 4027–4035, 2021.

Pavel Filonov, Andrey Lavrentyev, and Artem Vorontsov. Multivariate Industrial Time Series with Cyber-attack Simulation: Fault detection using an LSTM-based predictive data model, 2016.

Yifeng Gao, Jessica Lin, and Constantin Brif. Ensemble grammar induction for detecting anomalies in time series. In *International Conference on Extending Database Technology (EDBT)*, pp. 85–96. OpenProceedings.org, 2020. doi: 10.5441/002/edbt.2020.09.

Astha Garg, Wenyu Zhang, Jules Samaran, Ramasamy Savitha, and Chuan-Sheng Foo. An evaluation of anomaly detection and diagnosis in multivariate time series. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2508–2517, 2021. doi: 10.1109/TNNLS.2021.3105827.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000. doi: https://doi.org/10.1161/01.CIR.101.23.e215.

Scott David Greenwald, Ramesh S Patil, and Roger G Mark. Improved detection and classification of arrhythmias in noise-corrupted electrocardiograms using contextual information. In *Computers in Cardiology*, pp. 461–464, 1990.

GSU Data Mining Lab. SWaN dataset. `https://bitbucket.org/gsudmlab/mvtsdata_toolkit/downloads/petdataset_01.zip`, 2020. Accessed: 2025-05-10.

Xiao Han, Saima Absar, Lu Zhang, and Shuhan Yuan. Root cause analysis of anomalies in multivariate time series through granger causal discovery. In *International Conference on Learning Representations*, 2025.

Zengyou He, Xiaofei Xu, and Shengchun Deng. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9-10):1641–1650, 2003. doi: 10.1016/S0167-8655(03)00003-5.

Nikolai Helwig, Eliseo Pignanelli, and Andreas Schütze. Condition monitoring of a complex hydraulic system using multivariate statistics. In *IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pp. 210–215, 2015. doi: 10.1109/I2MTC.2015.7151267.

Alexis Huet, Jose Manuel Navarro, and Dario Rossi. Local evaluation of time series anomaly detection algorithms. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022. doi: 10.1145/3534678.3539339.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2018. doi: 10.1145/3219819.3219845.

Alexander Ihler, Jon Hutchins, and Padhraic Smyth. Adaptive event detection with time-varying poisson processes. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2006. doi: https://doi.org/10.1145/1150402.1150428.

iTrust, Singapore University of Technology and Design. Secure water treatment (SWaT) dataset characteristics. https://www.sutd.edu.sg/itrust/itrust-labs/datasets/dataset-characteristics/swat/, 2023a. Accessed 2026-08-29.

iTrust, Singapore University of Technology and Design. Water distribution (WaDi) dataset characteristics. https://www.sutd.edu.sg/itrust/itrust-labs/datasets/dataset-characteristics/wadi/, 2023b. Accessed 2026-08-29.

Vincent Jacob, Fei Song, Arnaud Stiegler, Bijan Rad, Yanlei Diao, and Nesime Tatbul. Exathlon: a benchmark for explainable anomaly detection over time series. *VLDB Endowment*, 14(11): 2613–2626, 2021. doi: 10.14778/3476249.3476307.

Eamonn Keogh, Jessica Lin, and Ada Fu. HOT SAX: Efficiently finding the most unusual time series subsequence. In *IEEE International Conference on Data Mining (ICDM)*, 2005. doi: 10.1109/ICDM.2005.79.

Hyungi Kim, Jisoo Mok, Dongjun Lee, Jaihyun Lew, Sungjae Kim, and Sungroh Yoon. Causality-aware contrastive learning for robust multivariate time-series anomaly detection. In *International Conference on Machine Learning*, pp. 30591–30608. PMLR, 2025.

Zhihan Li, Youjian Zhao, Jiaqi Han, Ya Su, Rui Jiao, Xidao Wen, and Dan Pei. Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3220–3230, 2021. doi: 10.1145/3447548.3467075.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *IEEE International Conference on Data Mining (ICDM)*, pp. 413–422, 2008. doi: 10.1109/ICDM.2008.17.

Qinghua Liu and John Paparrizos. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:108231–108261, 2024a.

Qinghua Liu and John Paparrizos. The elephant in the room: towards a reliable time-series anomaly detection benchmark. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Zehao Liu, Mengzhou Gao, and Pengfei Jiao. GCAD: Anomaly detection in multivariate time series from the perspective of granger causality. In *AAAI Conference on Artificial Intelligence*, volume 39, pp. 19041–19049, 2025. doi: 10.1609/aaai.v39i18.34096.

Yue Lu, Renjie Wu, Abdullah Mueen, Maria A Zuluaga, and Eamonn Keogh. Matrix profile XXIV: scaling time series anomaly detection to trillions of datapoints and ultra-fast arriving data streams. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1173–1182, 2022. doi: 10.1145/3534678.3539271.

Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, Puneet Agarwal, et al. Long short term memory networks for anomaly detection in time series. In *Esann*, volume 2015, pp. 89–94, 2015.

Mohsin Munir, Shoaib Ahmed Siddiqui, Andreas Dengel, and Sheraz Ahmed. DeepAnT: A deep learning approach for unsupervised anomaly detection in time series. *IEEE Access*, 7:1991–2005, 2018. doi: 10.1109/ACCESS.2018.2886457.

Takaaki Nakamura, Makoto Imamura, Ryan Mercer, and Eamonn Keogh. Merlin: Parameter-free discovery of arbitrary length anomalies in massive time series archives. In *IEEE International Conference on Data Mining (ICDM)*, pp. 1190–1195, 2020. doi: 10.1109/ICDM50108.2020.00147.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

John Paparrizos, Paul Boniol, Themis Palpanas, Ruey S Tsay, Aaron Elmore, and Michael J Franklin. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. *VLDB Endowment*, 15(11):2774–2787, 2022a. doi: 10.14778/3551793.3551830.

John Paparrizos, Yuhao Kang, Paul Boniol, Ruey S Tsay, Themis Palpanas, and Michael J Franklin. TSB-UAD: an end-to-end benchmark suite for univariate time-series anomaly detection. *VLDB Endowment*, 15(8):1697–1711, 2022b. doi: 10.14778/3529337.3529354.

Marc Pinet, Julien Cumin, Samuel Berlemont, and Dominique Vaufreydaz. Anomalies in multivariate time series benchmarks are mostly univariate. In *International Workshop on Mining and Learning from Time Series (MiLeTS)*, 2026.

Frederik Rehbach, Steffen Moritz, Sowmya Chandrasekaran, Margarita Rebolledo, Martina Friese, and Thomas Bartz-Beielstein. GECCO 2018 industrial challenge: Monitoring of drinking-water quality, 2018.

Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, et al. Collecting complex activity datasets in highly rich networked sensor environments. In *International Conference on Networked Sensing Systems (INSS)*, pp. 233–240, 2010. doi: 10.1109/INSS.2010.5573462.

Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. Anomaly detection in time series: a comprehensive evaluation. *VLDB Endowment*, 15(9):1779–1797, 2022. doi: 10.14778/3538598.3538602.

Pavel Senin, Jessica Lin, Xing Wang, Tim Oates, Sunil Gandhi, Arnold P Boedihardjo, Crystal Chen, and Susan Frankenstein. Time series anomaly discovery with grammar-based compression. In *International Conference on Extending Database Technology (EDBT)*, pp. 481–492, 2015.

Arnaldo Sgueglia, Andrea Di Sorbo, Corrado Aaron Visaggio, and Gerardo Canfora. A systematic literature review of IoT time series anomaly detection solutions. *Future Generation Computer Systems*, 134:170–186, 2022. doi: https://doi.org/10.1016/j.future.2022.04.005.

Junho Song, Keonwoo Kim, Jeonglyul Oh, and Sungzoon Cho. MEMTO: Memory-guided transformer for multivariate time series anomaly detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:57947–57963, 2023.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2019. doi: 10.1145/3292500.3330672.

Sadaf Tafazoli and Eamonn Keogh. Matrix Profile XXVIII: Discovering multi-dimensional time series anomalies with K of N anomaly detection. In *SIAM International Conference on Data Mining (SDM)*, pp. 685–693, 2023.

Nesime Tatbul, Tae Jun Lee, Stan Zdonik, Mejbah Alam, and Justin Gottschlich. Precision and recall for time series. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

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

Alexander von Birgelen and Oliver Niggemann. Anomaly detection and localization for cyber-physical production systems with self-organizing maps. *IMPROVE-Innovative Modelling Approaches for Production Systems to Raise Validatable Efficiency: Intelligent Methods for the Factory of the Future*, pp. 55–71, 2018. doi: https://doi.org/10.1007/978-3-662-57805-6_4.

Phillip Wenig, Sebastian Schmidl, and Thorsten Papenbrock. TimeEval: a benchmarking toolkit for time series anomaly detection algorithms. *VLDB Endowment*, 15(12):3678–3681, 2022. doi: 10.14778/3554821.3554873.

Renjie Wu and Eamonn Keogh. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE Transactions on Knowledge and Data Engineering*, 35 (3):2421–2429, 2023. doi: 10.1109/ICDE53745.2022.00116.

Xingjian Wu, Xiangfei Qiu, Zhengyu Li, Yihang Wang, Jilin Hu, Chenjuan Guo, Hui Xiong, and Bin Yang. CATCH: Channel-aware multivariate time series anomaly detection via frequency patching. In *International Conference on Learning Representations*, 2025.

Yongzheng Xie, Hongyu Zhang, and Muhammad Ali Babar. Multivariate time series anomaly detection by capturing coarse-grained intra-and inter-variate dependencies. In *the ACM on Web Conference*, pp. 697–705, 2025. doi: 10.1145/3696410.3714941.

Haowen Xu, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, Dan Pei, Yang Feng, Jie Chen, Zhaogang Wang, and Honglin Qiao. Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications. In *The World Wide Web Conference*, 2018. doi: 10.1145/3178876.3185996.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly Transformer: time series anomaly detection with association discrepancy. In *International Conference on Learning Representations*, 2022.

Takehisa Yairi, Yoshikiyo Kato, and Koichi Hori. Fault detection by mining association rules from house-keeping data. In *International Symposium on Artificial Intelligence, Robotics and Automation in Space (i-SAIRAS)*, volume 18, pp. 21, 2001.

Wenzhuo Yang, Kun Zhang, and Steven CH Hoi. A causal approach to detecting multivariate time-series anomalies and root causes. *arXiv preprint arXiv:2206.15033*, 2022.

Xue Yang, Xuejun Qi, and Xiaobo Zhou. Deep learning technologies for time series anomaly detection in healthcare: A review. *IEEE Access*, 2023a. doi: 10.1109/ACCESS.2023.3325896.

Yiyuan Yang, Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. DCdetector: Dual attention contrastive representation learning for time series anomaly detection. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3033–3045, 2023b. doi: 10.1145/3580305.3599295.

Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. Matrix Profile I: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. In *IEEE International Conference on Data Mining (ICDM)*, pp. 1317–1322, 2016. doi: 10.1109/ICDM.2016.0179.

Chin-Chia Michael Yeh, Audrey Der, Uday Singh Saini, Vivian Lai, Yan Zheng, Junpeng Wang, Xin Dai, Zhongfang Zhuang, Yujie Fan, Huiyuan Chen, et al. Matrix profile for anomaly detection on multidimensional time series. In *IEEE International Conference on Data Mining (ICDM)*, pp. 911–916, 2024.

Zahra Zamanzadeh Darban, Geoffrey I Webb, Shirui Pan, Charu Aggarwal, and Mahsa Salehi. Deep learning for time series anomaly detection: a survey. *ACM Computing Surveys (CSUR)*, 57(1): 1–42, 2024. doi: 10.1145/3691338.

Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. One Fits All: power general time series analysis by pretrained LM. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Xiaona Zhou, Constantin Brif, and Ismini Lourentzou. mTSBench: Benchmarking multivariate time series anomaly detection and model selection at scale. *Transactions on Machine Learning Research*, 2026.

Yan Zhu, Chin-Chia Michael Yeh, Zachary Zimmerman, Kaveh Kamgar, and Eamonn Keogh. Matrix profile xi: Scrimp++: time series motif discovery at interactive speeds. In *IEEE International Conference on Data Mining (ICDM)*, pp. 837–846, 2018. doi: 10.1109/ICDM.2018.00099.

A APPENDIX

A.1 IDEAS ALGORITHM

The IDEAS framework, described in the main text, is detailed formally in Algorithm 1.

Algorithm 1 IDEAS: Independent Dimension Ensemble for Anomaly Scoring

Require: Multivariate time series $\mathbf{X} = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(d)}]$, base detector m, aggregation operator $G \in \{\min, \max, \text{median}, \text{weighted}\}$, correlation threshold λ, trim ratio α, top-score threshold θ

Ensure: Length-n aggregate anomaly score vector $\mathbf{S}'''_{\text{aggr}}$, dimension anomaly scores $\{D_j\}_{j=1}^{d}$

Construct the original channel set $\mathcal{C}_{\text{orig}} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(d)}\}$ and initialize $\mathcal{C} \leftarrow \mathcal{C}_{\text{orig}}$.
for each dimension pair (i, j), $1 \le i < j \le d$ **do**
  Compute the Pearson correlation coefficient $\rho_{ij} = \text{corr}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ between dimensions i and j.
  if $|\rho_{ij}| > \lambda$ **then**
    Scale original dimensions:

$$\tilde{\mathbf{x}}^{(i)} = \frac{\mathbf{x}^{(i)} - \text{mean}(\mathbf{x}^{(i)})}{\text{stdev}(\mathbf{x}^{(i)})}, \qquad \tilde{\mathbf{x}}^{(j)} = \frac{\mathbf{x}^{(j)} - \text{mean}(\mathbf{x}^{(j)})}{\text{stdev}(\mathbf{x}^{(j)})}.$$

    Construct the pairwise-difference channel $\mathbf{\Delta}_{ij} = \tilde{\mathbf{x}}^{(i)} - \text{sign}(\rho_{ij})\tilde{\mathbf{x}}^{(j)}$ and update $\mathcal{C} \leftarrow \mathcal{C} \cup \{\mathbf{\Delta}_{ij}\}$.
  end if
end for
Set $\ell = \lfloor \alpha n \rfloor$ and $n' = n - 2\ell$.
for each channel $c \in \mathcal{C}$ **do**
  Apply m independently to channel c to obtain $\mathbf{S}_c = (s_{c,1}, \ldots, s_{c,n})$.
  Depending on m, its raw output is normalized and/or oriented to obtain $\mathbf{S}_c$, so that larger values indicate higher anomaly likelihood.
  Trim and center the score vector:

$$\mathbf{S}'_c = (s_{c,\ell+1}, \ldots, s_{c,n-\ell}), \qquad \mathbf{S}''_c = \mathbf{S}'_c - \text{mean}(\mathbf{S}'_c).$$

end for
if $G = \text{weighted}$ **then**
  for each channel $c \in \mathcal{C}$ **do**
    Compute the weight

$$W_c = \frac{\text{mean}\{s'_{c,t} : s'_{c,t} > Q_{\mathbf{S}'_c}(\theta)\}}{\text{mean}\{s'_{c,t} : s'_{c,t} \le Q_{\mathbf{S}'_c}(\theta)\}}.$$

  end for
  Normalize weights: $w_c = W_c / \sum_{c \in \mathcal{C}} W_c$.
end if
for $t = 1, \ldots, n'$ **do**

$$s''_{\text{aggr},t} = \begin{cases} \min_{c \in \mathcal{C}} s''_{c,t}, & G = \min, \\ \max_{c \in \mathcal{C}} s''_{c,t}, & G = \max, \\ \text{median}_{c \in \mathcal{C}} s''_{c,t}, & G = \text{median}, \\ \sum_{c \in \mathcal{C}} w_c s''_{c,t}, & G = \text{weighted}. \end{cases}$$

end for
Pad $\mathbf{S}''_{\text{aggr}}$ at both ends with ℓ values equal to $\min(\mathbf{S}''_{\text{aggr}})$, yielding $\mathbf{S}'''_{\text{aggr}}$.
for $j = 1, \ldots, d$ **do**
  Let $\mathbf{S}''_j$ be the trimmed and centered score vector for dimension j. Compute the dimension anomaly score $D_j = \text{corr}(\mathbf{S}''_j, \mathbf{S}''_{\text{aggr}})$.
end for
return $\mathbf{S}'''_{\text{aggr}}$ and $\{D_j\}_{j=1}^{d}$.

A.2 VISUALIZING ANOMALY SCORING AND AGGREGATION

Figures A1 and A2 visualize channel-wise anomaly scores and aggregated scores for two example time series from dataset B. Each figure contains five columns: the first column shows the channels of the multivariate time series, and columns two through five show the channel-wise anomaly scores $\mathbf{S}''_c$ produced by EGI-IDEAS, MP-IDEAS, CNN-IDEAS, and PCA-IDEAS, respectively. The top five rows correspond to individual channels, while the last row shows the ground-truth anomaly labels in

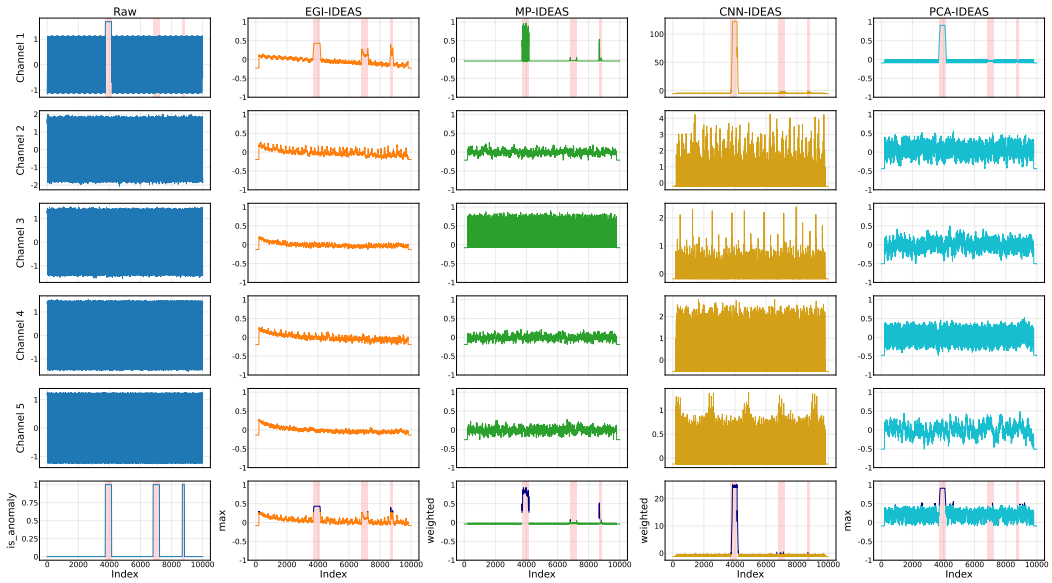


Figure A1: Channel-wise anomaly scores and the aggregated score produced by the IDEAS detectors for an example time series from dataset B. Pink shaded regions denote ground-truth anomalous intervals, and blue segments denote predicted anomalies.

the first column and the aggregated score $\mathbf{S}''_{\text{aggr}}$ for the IDEAS detectors in the remaining columns. Pink shaded regions denote ground-truth anomalous intervals, and blue segments in the last row mark the predicted anomalous regions obtained by thresholding the aggregated scores.

These examples illustrate how channel-wise scoring and aggregation can propagate anomaly signals from a single dimension into the final prediction. In Figure A1, the anomalies are located in Channel 1, and EGI-IDEAS produces clear score peaks aligned with all three ground-truth intervals. MP-IDEAS and CNN-IDEAS also produce strong responses around the most prominent interval, while PCA-IDEAS yields a sharp response for the first interval but shows no evidence for the later intervals. In Figure A2, the anomalies are in Channel 4. EGI-IDEAS produces score peaks near the anomalous regions, while MP-IDEAS and CNN-IDEAS respond most strongly to the final interval. These visualizations show that IDEAS not only produces an aggregate anomaly score, but also preserves channel-level evidence that can help predict which dimension contains the anomalies.

Figure A3 visualizes channel-wise anomaly scoring for an example time series from dataset F, where anomalies are constructed to preserve per-channel marginal distributions while disturbing cross-channel relationships. Unlike the dataset B examples, this figure includes both the original channels and the pairwise-difference channels constructed with correlation threshold $\lambda = 0.8$. The channel-wise scores show that several original dimensions already provide useful anomaly evidence, especially for EGI-IDEAS and MP-IDEAS. At the same time, multiple pairwise-difference channels produce additional localized responses near the ground-truth anomalous intervals, which further contribute to the aggregated scores. This visualization illustrates how IDEAS combines evidence from both individual channels and pairwise-difference channels when detecting cross-channel anomalies.

A.3 Data Generation

We constructed six controlled synthetic datasets, denoted A through F, whose characteristics are summarized in Table A1. Datasets A–E are based on GutenTAG (Wenig et al., 2022), while dataset F is generated using the cross-channel anomaly synthesis protocol of Pinet et al. (2026).

For datasets A–E, normal signals are generated from 10 base oscillation families: sine, cosine, square, polynomial, CBF, ECG, random mode jump, sawtooth, Dirichlet, and MLS. Anomalies are sampled from 10 compatible anomaly classes: amplitude, extremum, frequency, mean, pattern, pattern shift, platform, trend, variance, and mode-correlation. Each corrupted dimension contains 1 to 3 anomaly

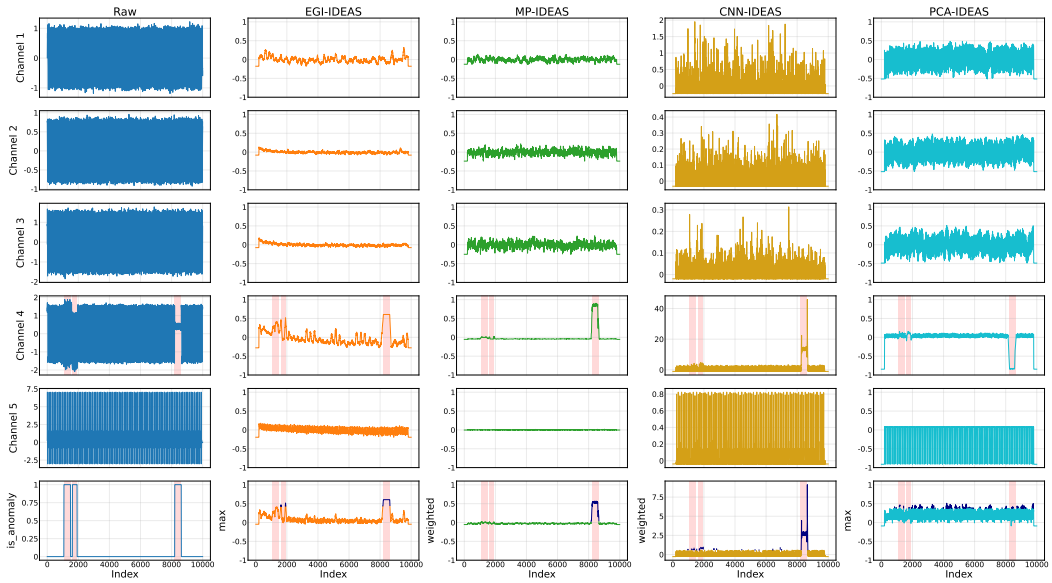


Figure A2: Channel-wise anomaly scores and the aggregated score produced by the IDEAS detectors for an example time series from dataset B. Pink shaded regions denote ground-truth anomalous intervals, and blue segments denote predicted anomalies.

intervals, with their start points sampled from the central 80% of the sequence and lengths sampled from $[50, 500]$ for datasets A–C and from $[0.005n, 0.05n]$ for datasets D and E. If an anomaly occurs across multiple channels, its start point and length are the same for all these dimensions, but the anomaly class and parameters are selected independently and randomly for each dimension.

Dataset A: varying the number of anomalous dimensions. Dataset A is used to evaluate the effect of the anomaly spread across dimensions. It contains 500 multivariate time series with fixed dimensionality of $d = 20$ and fixed time-series length of $n = 10^4$, divided into five 100-series groups with different numbers of anomalous dimensions, $d_{\text{anom}} \in \{1, 2, 5, 10, 20\}$. This setting evaluates how detection performance changes as anomaly signals become less sparse across dimensions.

Dataset B: varying the dimensionality with one anomalous dimension. Dataset B is used to evaluate the effect of the total number of dimensions, d, when anomalies are contained in a single dimension ($d_{\text{anom}} = 1$) and therefore are sparse at the channel level. It contains 500 multivariate time series with fixed time-series length, $n = 10^4$, divided into five 100-series groups with different dimensionalities, $d \in \{5, 10, 20, 50, 100\}$. This setting tests how detectors perform as the number of anomaly-free dimensions increases.

Dataset C: varying the dimensionality with a random number of anomalous dimensions. Dataset C uses the same time-series length ($n = 10^4$) and dimensionality settings as dataset B, containing 500 multivariate time series divided into five 100-series groups with $d \in \{5, 10, 20, 50, 100\}$. In contrast to dataset B, the number of anomalous dimensions, d_{anom}, is a random integer sampled from $[1, d]$ for each time series. Comparing results on datasets B and C isolates the effect of channel-level anomaly sparsity under increasing dimensionality.

Dataset D: varying the time-series length with one anomalous dimension. Dataset D is used to evaluate the effect of the time-series length on detection performance when anomalies are contained in a single dimension ($d_{\text{anom}} = 1$). It contains 500 multivariate time series with fixed dimensionality, $d = 20$, divided into five 100-series groups with different time-series lengths, $n \in \{5, 10, 20, 50, 100\} \times 10^3$. To maintain a comparable temporal anomaly ratio across all groups, anomalous interval lengths are sampled from $[0.005n, 0.05n]$.

Dataset E: varying the time-series length with a random number of anomalous dimensions. Dataset E uses the same fixed dimensionality ($d = 20$) and time-series-length settings as dataset D, with 500 multivariate time series divided into five 100-series groups with $n \in \{5, 10, 20, 50, 100\} \times 10^3$. The difference is that in dataset E, the number of anomalous dimensions, d_{anom}, is a random

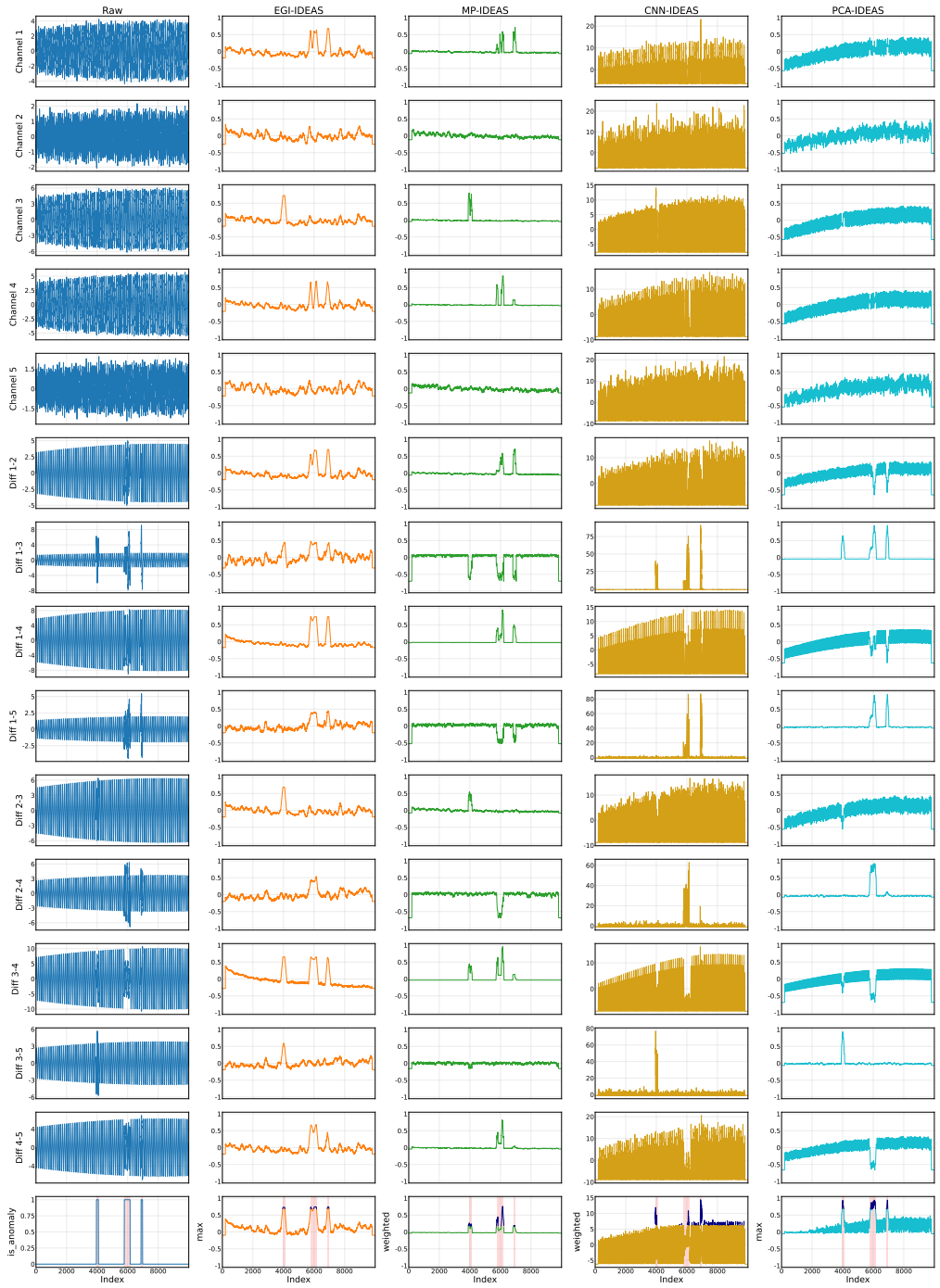


Figure A3: Channel-wise anomaly scores and the aggregated score produced by the IDEAS detectors for an example time series from dataset F. Pairwise-difference channels are constructed with correlation threshold $\lambda = 0.8$. Pink shaded regions denote ground-truth anomalous intervals, and blue segments denote predicted anomalies.

integer sampled from $[1, d]$ for each time series. Comparing detection results on datasets D and E isolates the effect of channel-level anomaly sparsity under increasing time-series length.

Dataset F: including cross-channel anomalies. Dataset F evaluates the effect of increasing dimensionality in a setting where anomalies preserve per-channel marginal distributions but disrupt cross-channel dependencies. Following the synthesis protocol of Pinet et al. (2026), normal signals

Table A1: Statistics of datasets A–F, for the test split. Each of datasets A–E contains five groups and dataset F contains three groups. Each row summarizes one group, which consists of 100 time series. $\bar{K}_{\text{anom}}$ is the mean number of anomalous intervals per time series, $\bar{\ell}_{\text{anom}}$ is the mean anomalous-interval length, and r_{anom} is the temporal anomaly ratio (the number of anomalous time points divided by the total number of time points). Each test set has a matched anomaly-free training set with the same number of time series, dimensionality d, and time-series length n.

Dataset	d	d_{anom}	n	$\bar{K}_{\text{anom}}$	$\bar{\ell}_{\text{anom}}$	r_{anom}
A	20	1	10^4	2.0	284	5.7%
	20	2	10^4	2.0	272	5.2%
	20	5	10^4	2.0	257	5.0%
	20	10	10^4	2.0	268	5.3%
	20	20	10^4	2.0	275	5.5%
B	5	1	10^4	2.0	276	5.2%
	10	1	10^4	2.0	267	5.2%
	20	1	10^4	2.0	271	5.2%
	50	1	10^4	2.0	283	5.1%
	100	1	10^4	2.0	280	5.5%
C	5	$[1, d]$	10^4	1.9	272	5.3%
	10	$[1, d]$	10^4	2.1	285	5.9%
	20	$[1, d]$	10^4	1.9	286	5.4%
	50	$[1, d]$	10^4	1.9	288	5.3%
	100	$[1, d]$	10^4	2.0	296	6.0%
D	20	1	5×10^3	2.0	140	5.4%
	20	1	10^4	2.0	287	5.6%
	20	1	2×10^4	2.0	560	5.5%
	20	1	5×10^4	2.0	1,347	5.4%
	20	1	10^5	2.0	2,803	5.0%
E	20	$[1, d]$	5×10^3	2.0	134	5.4%
	20	$[1, d]$	10^4	1.9	279	5.3%
	20	$[1, d]$	2×10^4	2.1	538	5.8%
	20	$[1, d]$	5×10^4	1.9	1,368	5.3%
	20	$[1, d]$	10^5	2.0	2,739	5.6%
F	5	$[1, d]$	10^4	2.0	280	5.5%
	10	$[1, d]$	10^4	2.1	272	5.7%
	20	$[1, d]$	10^4	2.0	287	5.6%

are generated as phase-shifted sinusoids with shared additive noise, inducing cross-channel dependence. This dataset contains 300 time series with fixed time-series length, $n = 10^4$, divided into three 100-series groups with different dimensionalities, $d \in \{5, 10, 20\}$. Each time-series contains 1–3 anomalous intervals with lengths sampled from $[50, 500]$ and start points sampled from the central 80% of the sequence. For each anomalous interval, a random subset of dimensions is modified using either NOISEFLIP or NPROLL procedure. In NOISEFLIP, the shared noise in selected dimensions is negated within the anomaly interval. In NPROLL, the values within the anomalous interval are circularly shifted by a random offset. These transformations preserve the per-channel marginal distribution while breaking the cross-channel structure. To verify that dataset F contains strong cross-channel dependence, Table A2 summarizes the number of dimension pairs whose Pearson correlations exceed $\lambda = 0.9$ or $\lambda = 0.8$. Counts are computed for each time series and their statistics are summarized by train/test split and groups with different dimensionalities. The number of strongly correlated pairs increases with d, and the $\lambda = 0.8$ threshold consistently selects more pairs than $\lambda = 0.9$, matching the threshold settings used in the main part of the paper.

Training sets. For each dataset group, we also generate a corresponding anomaly-free training set with the same dimensionality, time-series length, and the number of series as in the test set. Thus, datasets A–F all provide matched train/test pairs, where the training sequences contain only normal behavior and are used to train semi-supervised detectors, which include CNN (Munir et al., 2018),

Table A2: Summary of pairwise dimension correlations for dataset F. For each time series, we count the number of dimension pairs with $|\rho_{ij}| > 0.9$ and $|\rho_{ij}| > 0.8$, then summarize their statistics by train/test split and dimensionality.

Split	d	Mean		Min		Max		Stdev	
		0.9	**0.8**	**0.9**	**0.8**	**0.9**	**0.8**	**0.9**	**0.8**
train	5	3.64	3.95	2	2	4	4	0.69	0.26
train	10	18.53	19.82	12	16	20	20	2.05	0.61
train	20	83.32	89.65	65	85	90	90	6.55	0.99
test	5	3.52	3.95	1	2	4	4	0.80	0.26
test	10	17.08	19.61	7	16	20	20	3.23	0.87
test	20	76.37	88.31	30	76	90	90	13.05	3.02

LSTMAD (Malhotra et al., 2015), OmniAnomaly (Su et al., 2019), TranAD (Tuli et al., 2022), ALLM4TS (Bian et al., 2024), and OFA (Zhou et al., 2023).

Real-world datasets. We evaluate IDEAS and baselines on 20 real-world multivariate time-series datasets spanning diverse domains and anomaly characteristics. The datasets are described below.

CALIT2 (Ihler et al., 2006) records the number of people entering and exiting the main door of the CalIt2 building at UCI over 15 weeks, with 48 time slices per day. The goal is to detect events, such as conferences, indicated by unusually high people counts on specific days or times.

CIC-IDS-2017 (Canadian Institute for Cybersecurity, 2017) contains realistic network traffic data, including both benign behavior and a wide range of attack scenarios such as DDoS, brute-force, and infiltration. It provides labeled flow-based features extracted from raw PCAP files, enabling evaluation of intrusion detection systems.

DAPHNET (Bachlin et al., 2009) consists of annotated readings from three acceleration sensors placed on the hip and leg of Parkinson's disease patients who experience freezing of gait (FoG) during walking tasks, such as straight line walking and walking with numerous turns.

EXATHLON (Jacob et al., 2021) is a benchmark for explainable anomaly detection in high-dimensional time series, based on real Apache Spark stream processing traces. Executions were intentionally disturbed by introducing six types of anomalous events (bursty input, bursty input until crash, stalled input, CPU contention, driver failure, and executor failure).

GECCO (Rehbach et al., 2018) was originally introduced in the GECCO Challenge 2017. This dataset contains sensor readings from water treatment plants and distribution systems, capturing parameters such as water quality, flow rates, and pressure.

GENESIS (von Birgelen & Niggemann, 2018) contains sensor readings collected from a portable pick-and-place demonstrator, a cyber-physical system that sorts two different materials (conductive and non-conductive) from a magazine into their corresponding target locations. The dataset contains five primary continuous signals, thirteen discrete signals, and a single Unix timestamp. Two out of 42 production cycles contain anomalous behavior.

GHL (Filonov et al., 2016) records the operational status of three reservoirs. Anomalies correspond to shifts in maximum temperature or pump frequency.

METRO (Helwig et al., 2015) contains hourly traffic volume data from a highway in Minneapolis–St Paul, collected by a single loop detector over several years. It includes features such as weather conditions, time-based attributes, and traffic volume counts.

MITDB (Goldberger et al., 2000) consists of 48 half-hour of two-channel ambulatory ECG recordings collected from 47 individuals by the BIH Arrhythmia Laboratory during the years 1975 to 1979.

MSL (Hundman et al., 2018) originates from NASA's Mars Science Laboratory mission and contains spacecraft telemetry data collected from the Curiosity rover.

OPPORTUNITY (Roggen et al., 2010) is a benchmark dataset for human activity recognition, encompassing tasks such as classification, segmentation, sensor fusion, and feature extraction. It consists of motion sensor recordings collected as users performed routine daily activities.

Table A3: Overview of multivariate anomaly detector baselines, encompassing a diverse range of techniques, including foundation models, distance-based, forecasting-based, and reconstruction-based approaches, ensuring a comprehensive representation of different methodological paradigms.

Learning	Anomaly Detection Method	Area	Method Family
Unsupervised	CBLOF (He et al., 2003)	Outlier Detection	**Distance**
	IForest (Liu et al., 2008)	Outlier Detection	**Tree**
	KMeansAD (Yairi et al., 2001)	Classic ML	**Distance**
	PCA (Aggarwal, 2017)	Classic ML	**Distance**
Semi-supervised	ALLM4TS (Bian et al., 2024)	LLM	**Foundation Model**
	CNN (Munir et al., 2018)	Deep Learning	**Forecasting**
	LSTMAD (Malhotra et al., 2015)	Deep Learning	**Forecasting**
	OFA (Zhou et al., 2023)	LLM	**Foundation Model**
	OmniAnomaly (Su et al., 2019)	Deep Learning	**Reconstruction**
	TranAD (Tuli et al., 2022)	Deep Learning	**Reconstruction**

PSM (Abdulaal et al., 2021) contains data collected over 21 weeks from multiple application server nodes at eBay. It includes 26 variables that describe server machine metrics.

OCCUPANCY (Candanedo & Feldheim, 2016) consists of indoor environmental sensor readings, including temperature, humidity, light, and CO_2 levels, collected from a monitored room to predict occupancy status. The labeled, time-stamped data supports tasks such as binary classification, energy modeling, and real-time occupancy prediction.

SMAP (Hundman et al., 2018) originates from from NASA's Soil Moisture Active Passive satellite mission. It contains telemetry data collected from the satellite's sensors.

SMD (Su et al., 2019) is composed of time series data collected over a five-week period from server machines in a data center. It captures metrics such as CPU usage, memory consumption, disk I/O, and network traffic, making it a representative dataset for monitoring IT infrastructure systems.

SVDB (Greenwald et al., 1990) contains 78 half-hour ECG recordings featuring a combination of supraventricular and ventricular ectopic beats within a normal sinus rhythm.

SWAN-SF (GSU Data Mining Lab, 2020) comprises data derived from solar photospheric vector magnetograms in the SHARP data series, designed for space weather analytics.

SWAT_A2 (iTrust, Singapore University of Technology and Design, 2023a) derived from the Secure Water Treatment (SWaT) testbed, a cyber-physical water treatment system with multivariate sensors.

SWAT_A5 (iTrust, Singapore University of Technology and Design, 2023a) derived from the SWaT testbed, providing another real-world multivariate water-treatment telemetry dataset.

WADI_A2 (iTrust, Singapore University of Technology and Design, 2023b) derived from the Water Distribution (WaDi) testbed, a cyber-physical water distribution system with multivariate sensors.

TSB-AD-M (Liu & Paparrizos, 2024a) is the multivariate part of the TSB-AD benchmark, which contains curated time-series anomaly detection datasets from diverse domains.

A.4 BASELINE MULTIVARIATE DETECTORS

We compare the IDEAS detectors against 10 multivariate anomaly detector baselines, summarized in Table A3. These baselines cover four broad areas: outlier detection, classical machine learning, deep learning, and foundation-model-based time-series methods. We further group them by their primary detection mechanism. **Distance-based methods** identify anomalies using distances to clusters, centroids, or neighboring samples. **Tree-based methods** isolate anomalies through recursive partitioning of the feature space, based on the intuition that rare observations are easier to separate. **Reconstruction-based methods** model normal patterns and assign higher anomaly scores to observations with larger reconstruction errors. **Forecasting-based methods** predict future values from historical context and detect anomalies as deviations from the forecast. **Foundation model-based methods** adapt pretrained models or LLM-based time-series frameworks for anomaly

detection. The table also distinguishes between unsupervised methods, which do not require any training, and semi-supervised methods, which are trained using anomaly-free time-series data.

CBLOF (He et al., 2003) is a clustering-based outlier detector that assigns larger anomaly scores to observations that belong to small clusters or lie far from large, dense clusters.

IForest (Liu et al., 2008) detects anomalies by recursively partitioning the feature space with random splits, where observations isolated with shorter path lengths receive higher anomaly scores.

KMeansAD (Yairi et al., 2001) applies k-means clustering to time-series windows and scores each window by its distance to the assigned cluster centroid.

PCA (Aggarwal, 2017) projects data onto a low-dimensional principal subspace and uses large deviations from this subspace to flag potential anomalies.

ALLM4TS (Bian et al., 2024) adapts a pretrained LLM to time-series representation learning through self-supervised patch prediction and uses the learned representations for anomaly detection.

CNN (Munir et al., 2018) follows the DeepAnT framework, which uses a convolutional neural network to forecast time-series values and treats large forecasting errors as anomalies.

LSTMAD (Malhotra et al., 2015) trains stacked Long Short-Term Memory (LSTM) networks on normal sequences and detects anomalies from deviations between predicted and observed values.

OFA (Zhou et al., 2023) adapts a pretrained Transformer-based LLM to time-series analysis by fine-tuning the pretrained backbone for downstream tasks including anomaly detection.

OmniAnomaly (Su et al., 2019) combines recurrent modeling with a variational autoencoder to model normal multivariate dynamics and detect anomalies using reconstruction probability.

TranAD (Tuli et al., 2022) employs a Transformer-based encoder with self-conditioning and adversarial training to model multivariate temporal patterns and detect anomalous observations.

A.5 Evaluation Metrics

All evaluation metrics are computed from the anomaly score vector, $\mathbf{S} = (s_1, \ldots, s_n)^\top$ for baseline methods or $\mathbf{S}'''_{\text{aggr}} = (s'''_{\text{aggr},1}, \ldots, s'''_{\text{aggr},n})^\top$ for IDEAS detectors, where larger scores indicate stronger evidence of anomalous behavior. For threshold-based metrics, binary predictions are

$$\widehat{y}_t(\tau) = \mathbb{I}[s_t > Q_{\mathbf{S}}(\tau)] \quad \text{or} \quad \widehat{y}_t(\tau) = \mathbb{I}[s'''_{\text{aggr},t} > Q_{\mathbf{S}'''_{\text{aggr}}}(\tau)], \quad t = 1, \ldots, n, \tag{A1}$$

where $Q_{\mathbf{V}}(p)$ denotes the empirical p-quantile of a given vector $\mathbf{V}$. We denote the vector of prediction labels for threshold τ as $\widehat{\mathbf{y}}(\tau) = \big(\widehat{y}_1(\tau), \ldots, \widehat{y}_n(\tau)\big)^\top$, and the vector of ground-truth labels as $\mathbf{y} = \big(y_1, \ldots, y_n\big)^\top$, where $y_t \in \{0, 1\}$. Consecutive predicted anomalous points are grouped into predicted anomaly intervals, which form the set $\mathcal{P}(\tau) = \{P_1, \ldots, P_{M(\tau)}\}$, and consecutive ground-truth anomalous points are grouped into ground-truth anomaly intervals, which form the set $\mathcal{G} = \{G_1, \ldots, G_K\}$. Point-wise metrics evaluate predictions at individual time points, while range-based and event-based metrics evaluate how well predicted intervals align with ground-truth.

Precision, Recall, and $\mathbf{F_1}$. For a fixed threshold τ, let $\mathrm{TP}(\tau)$, $\mathrm{TN}(\tau)$, $\mathrm{FP}(\tau)$, and $\mathrm{FN}(\tau)$, denote the counts of point-wise true positive, true negative, false positive, and false negative detections, respectively:

$$\mathrm{TP}(\tau) = \widehat{\mathbf{y}}(\tau) \cdot \mathbf{y}, \tag{A2}$$
$$\mathrm{TN}(\tau) = (\mathbf{1}_n - \widehat{\mathbf{y}}(\tau)) \cdot (\mathbf{1}_n - \mathbf{y}), \tag{A3}$$
$$\mathrm{FP}(\tau) = \widehat{\mathbf{y}}(\tau) \cdot (\mathbf{1}_n - \mathbf{y}), \tag{A4}$$
$$\mathrm{FN}(\tau) = (\mathbf{1}_n - \widehat{\mathbf{y}}(\tau)) \cdot \mathbf{y}. \tag{A5}$$

Point-wise precision, recall, and F_1 score are defined as

$$\text{Precision}(\tau) = \frac{\text{TP}(\tau)}{\text{TP}(\tau) + \text{FP}(\tau) + \epsilon}, \tag{A6}$$

$$\text{Recall}(\tau) = \frac{\text{TP}(\tau)}{\text{TP}(\tau) + \text{FN}(\tau) + \epsilon}, \tag{A7}$$

$$\text{F}_1(\tau) = \frac{2 \cdot \text{Precision}(\tau) \cdot \text{Recall}(\tau)}{\text{Precision}(\tau) + \text{Recall}(\tau) + \epsilon}, \tag{A8}$$

where ϵ is a small constant used to avoid division by zero.

AUC-ROC measures the area under the receiver operating characteristic (ROC) curve, which plots true positive rate against false positive rate as the threshold varies:

$$\text{TPR}(\tau) = \frac{\text{TP}(\tau)}{\text{TP}(\tau) + \text{FN}(\tau) + \epsilon}, \tag{A9}$$

$$\text{FPR}(\tau) = \frac{\text{FP}(\tau)}{\text{FP}(\tau) + \text{TN}(\tau) + \epsilon}. \tag{A10}$$

The area under this curve is

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}). \tag{A11}$$

AUC-PR measures the area under the precision-recall curve:

$$\text{AUC-PR} = \int_0^1 \text{Precision}(\text{Recall})\, d(\text{Recall}). \tag{A12}$$

It is useful for anomaly detection because anomalies are rare and true negatives dominate the data.

AUC-$\boldsymbol{P_T R_T}$ is a range-based version of AUC-PR, which evaluates predictions as contiguous anomaly intervals Tatbul et al. (2018). For each threshold τ, range-based precision and recall are computed using the set of predicted intervals, $\mathcal{P}(\tau)$, and the set of ground-truth intervals, $\mathcal{G}$. The interval-level overlap and existence terms are defined as

$$p_i = \max_{1 \le j \le K} \frac{|P_i \cap G_j|}{|P_i|}, \tag{A13}$$

$$e_i^P = \mathbb{I}(\exists j : P_i \cap G_j \neq \emptyset), \tag{A14}$$

$$r_j = \max_{1 \le i \le M(\tau)} \frac{|P_i \cap G_j|}{|G_j|}, \tag{A15}$$

$$e_j^R = \mathbb{I}(\exists i : P_i \cap G_j \neq \emptyset). \tag{A16}$$

Here, p_i measures the best overlap quality of predicted interval P_i, e_i^P indicates whether P_i overlaps any ground-truth interval, r_j measures the best coverage of ground-truth interval G_j, and e_j^R indicates whether G_j is detected by any predicted interval. Range-based precision is

$$P_T(\tau) = \frac{1}{M(\tau)} \sum_{i=1}^{M(\tau)} \left[(1 - \alpha_p) p_i + \alpha_p e_i^P \right], \tag{A17}$$

where $\alpha_p \in [0, 1]$ controls the trade-off between overlap quality and existence in range-based precision. Range-based recall is

$$R_T(\tau) = \frac{1}{K} \sum_{j=1}^{K} \left[(1 - \alpha_r) r_j + \alpha_r e_j^R \right], \tag{A18}$$

where $\alpha_r \in [0, 1]$ controls the trade-off between overlap quality and existence in range-based recall. The range-based precision-recall curve is obtained by varying τ, and the area is AUC-$P_T R_T$:

$$\text{AUC-}P_T R_T = \int_0^1 P_T(R_T)\, dR_T. \tag{A19}$$

VUS-ROC and VUS-PR extend AUC-based evaluation by averaging performance over temporal buffer lengths, making the evaluation more robust to small temporal misalignments Paparrizos et al. (2022a). For each buffer length $b \in \mathcal{B}$, ground-truth anomaly intervals are expanded before computing the corresponding AUC value. In a discrete implementation, we compute

$$\text{VUS-ROC} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \text{AUC-ROC}(b), \tag{A20}$$

$$\text{VUS-PR} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \text{AUC-PR}(b). \tag{A21}$$

Event-based $\mathbf{F_1}$ combines point-wise precision with event-level recall Garg et al. (2021). A ground-truth interval G_j is considered detected if at least one predicted anomalous point overlaps it:

$$\text{Recall}_{\text{event}}(\tau) = \frac{1}{K} \sum_{j=1}^{K} \mathbb{I}(\exists t \in G_j : \widehat{y}_t(\tau) = 1)\,. \tag{A22}$$

The reported Event-based F_1 is the maximum value over the evaluated thresholds:

$$\text{Event-F}_1 = \max_{\tau} \frac{2 \cdot \text{Precision}(\tau) \cdot \text{Recall}_{\text{event}}(\tau)}{\text{Precision}(\tau) + \text{Recall}_{\text{event}}(\tau) + \epsilon}. \tag{A23}$$

Max-$\mathbf{F_1}$ is the best point-wise F_1 score over a sweep of decision thresholds:

$$\text{Max-F}_1 = \max_{\tau} \text{F}_1(\tau). \tag{A24}$$

R-based $\mathbf{F_1}$ is a range-level F_1 score based on range precision and range recall Tatbul et al. (2018). For each ground-truth interval G_j, let

$$k_j = \sum_{i=1}^{M(\tau)} \mathbb{I}(P_i \cap G_j \neq \emptyset) \tag{A25}$$

denote the number of predicted intervals overlapping G_j. The existence reward is

$$\text{ER}_j = \mathbb{I}(k_j > 0)\,, \tag{A26}$$

and the overlap reward is

$$\text{OR}_j = \begin{cases} \dfrac{1}{k_j} \displaystyle\sum_{i:P_i \cap G_j \neq \emptyset} \frac{|P_i \cap G_j|}{|G_j|}, & k_j > 0, \\ 0, & k_j = 0. \end{cases} \tag{A27}$$

The range-based recall is

$$\text{Recall}_R(\tau) = \frac{1}{K} \sum_{j=1}^{K} \left[\alpha\, \text{ER}_j + (1-\alpha)\, \text{OR}_j\right], \tag{A28}$$

where $\alpha \in [0,1]$ controls the trade-off between existence reward and overlap reward. For each predicted interval P_i, let

$$\ell_i = \sum_{j=1}^{K} \mathbb{I}(G_j \cap P_i \neq \emptyset)\,. \tag{A29}$$

The range-based precision is

$$\text{Precision}_R(\tau) = \frac{1}{M(\tau)} \sum_{i=1}^{M(\tau)} q_i, \tag{A30}$$

where

$$q_i = \begin{cases} \dfrac{1}{\ell_i} \displaystyle\sum_{j:G_j \cap P_i \neq \emptyset} \frac{|P_i \cap G_j|}{|P_i|}, & \ell_i > 0, \\ 0, & \ell_i = 0. \end{cases} \tag{A31}$$

The threshold-specific R-based F_1 score is

$$\text{R-F}_1(\tau) = \frac{2 \cdot \text{Precision}_R(\tau) \cdot \text{Recall}_R(\tau)}{\text{Precision}_R(\tau) + \text{Recall}_R(\tau) + \epsilon}. \tag{A32}$$

The reported score is

$$\text{R-F}_1 = \max_{\tau} \text{R-F}_1(\tau). \tag{A33}$$

Affiliation-F_1 evaluates temporal alignment between predicted and ground-truth anomaly intervals using affiliation zones around true events Huet et al. (2022). The time axis is partitioned into zones $\{I_1, \ldots, I_K\}$, where each zone I_j is associated with the closest ground-truth interval G_j. For threshold τ, let

$$\text{pred}_\tau = \bigcup_{i=1}^{M(\tau)} P_i \tag{A34}$$

denote the union of predicted anomaly intervals, and let $\text{pred}_{\tau,j} = \text{pred}_\tau \cap I_j$ denote the predictions inside zone I_j. Affiliation precision and recall are computed locally within each zone by measuring directed temporal distances between $\text{pred}_{\tau,j}$ and G_j, then normalizing these distances against a random prediction baseline within the same zone. Let $A_j^{\text{prec}}(\tau)$ and $A_j^{\text{rec}}(\tau)$ denote the resulting zone-level affiliation precision and recall scores. The global affiliation precision and recall are

$$\text{Precision}_{\text{aff}}(\tau) = \frac{1}{|\mathcal{J}_\tau|} \sum_{j \in \mathcal{J}_\tau} A_j^{\text{prec}}(\tau), \tag{A35}$$

$$\text{Recall}_{\text{aff}}(\tau) = \frac{1}{K} \sum_{j=1}^{K} A_j^{\text{rec}}(\tau), \tag{A36}$$

where $\mathcal{J}_\tau = \{j : \text{pred}_{\tau,j} \neq \emptyset\}$. The threshold-specific Affiliation-F_1 score is

$$\text{Affiliation-F}_1(\tau) = \frac{2 \cdot \text{Precision}_{\text{aff}}(\tau) \cdot \text{Recall}_{\text{aff}}(\tau)}{\text{Precision}_{\text{aff}}(\tau) + \text{Recall}_{\text{aff}}(\tau) + \epsilon}. \tag{A37}$$

A.6 Aggregation Strategy Analysis

Aggregation strategy selection. For a base detector deployed within IDEAS, we select a default aggregation strategy from the set of four variants, $\mathcal{V} = \{\min, \max, \text{median}, \text{weighted}\}$, which is akin to choosing a default hyperparameter value for a model. The selection procedure is described below. Let $M_{z,r,m,v,q}$ be the value of metric q obtained by detector m with aggregation variant v on time series r in dataset z. We define the ratio-to-winner (RTW) score as $M_{z,r,m,v,q}$ divided by the highest value in the set $\{M_{z,r,m,v,q} : v \in \mathcal{V}\}$:

$$\text{RTW}_{z,r,m,v,q} = \frac{M_{z,r,m,v,q}}{\max_{v \in \mathcal{V}} M_{z,r,m,v,q}}. \tag{A38}$$

Thus, the best variant for a given time series and metric receives an RTW score of 1, and all other variants are scored relative to that winner. For each variant v, we average RTW over the set $\mathcal{T}_z$ of time series in dataset z and the set $\mathcal{Q}$ of evaluation metrics:

$$\overline{\text{RTW}}_{z,m,v} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{|\mathcal{T}_z|} \sum_{r \in \mathcal{T}_z} \text{RTW}_{z,r,m,v,q}. \tag{A39}$$

Finally, we average over the set $\mathcal{T} = \cup_{z=\text{A}}^{\text{F}} \mathcal{T}_z$ of time series in datasets A–F:

$$\overline{\text{RTW}}_{m,v} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{|\mathcal{T}|} \sum_{r \in \mathcal{T}} \text{RTW}_{z,r,m,v,q}. \tag{A40}$$

The default aggregation variant selected for detector m is

$$v_m^* = \arg\max_{v \in \mathcal{V}} \overline{\text{RTW}}_{m,v}. \tag{A41}$$

Table A4: Effect of aggregation strategy on IDEAS detectors across all datasets. Values are $\overline{\mathrm{RTW}}_{z,m,v}$ scores of equation A39 for each of the datasets A–F and $\overline{\mathrm{RTW}}_{m,v}$ scores of equation A40 averaged over all six datasets. The best aggregation variant for each detector is highlighted in bold.

Dataset	Operator	EGI	MP	CNN	PCA
A	min	0.325	0.453	0.680	0.385
	max	**0.956**	0.783	0.885	**0.869**
	median	0.468	0.678	0.707	0.635
	weighted	0.640	**0.896**	**0.912**	0.785
B	min	0.506	0.552	0.706	0.643
	max	**0.882**	0.722	**0.829**	**0.801**
	median	0.469	0.574	0.675	0.706
	weighted	0.558	**0.782**	0.828	0.737
C	min	0.282	0.434	0.704	0.338
	max	**0.972**	0.812	0.909	**0.874**
	median	0.498	0.730	0.751	0.648
	weighted	0.690	**0.927**	**0.935**	0.840
D	min	0.493	0.551	0.724	0.632
	max	**0.895**	0.724	0.829	**0.813**
	median	0.459	0.582	0.698	0.707
	weighted	0.515	**0.748**	**0.835**	0.742
E	min	0.272	0.413	0.710	0.297
	max	**0.971**	0.804	0.885	**0.902**
	median	0.476	0.734	0.766	0.625
	weighted	0.705	**0.933**	**0.926**	0.820
F	min	0.324	0.283	0.571	0.355
	max	**0.987**	**0.958**	**0.980**	**0.972**
	median	0.539	0.377	0.564	0.439
	weighted	0.723	0.900	0.867	0.959
$\overline{\mathrm{RTW}}_{m,v}$	min	0.370	0.460	0.691	0.448
	max	**0.941**	0.789	0.879	**0.865**
	median	0.481	0.629	0.703	0.640
	weighted	0.633	**0.862**	**0.885**	0.804

The best aggregation strategy is detector-dependent. Table A4 reports, for each of the IDEAS detectors, $\overline{\mathrm{RTW}}_{z,m,v}$ scores of equation A39 for each of the datasets A–F (i.e., $z \in \{\mathrm{A}, \ldots, \mathrm{F}\}$) and $\overline{\mathrm{RTW}}_{m,v}$ scores of equation A40 averaged over all six datasets. It shows that no single aggregation operator is best for all detectors. The $\max$ operator is consistently the strongest choice for EGI and PCA across all six datasets, suggesting that these detectors benefit from preserving the largest channel-wise anomaly response rather than averaging it with weaker channels. In contrast, the $\mathrm{weighted}$ operator is best for CNN on four of the six datasets, and is best for MP on five of the six datasets, indicating that these detectors benefit from emphasizing channels with stronger anomaly evidence while reducing the influence of less informative channels. The $\min$ and median operators are never the best variant, suggesting that overly conservative aggregation can dilute sparse anomaly signals. Overall, these results motivate selecting the default aggregation variant separately for each detector rather than using a fixed operator across all of them. If a practitioner considers deploying a new base detector within IDEAS and prefers skipping the selection of the best aggregation strategy, we recommend using the $\max$ variant as the default, since it always produces strong results.

A.7 Performance Comparison with Baselines for Synthetic Datasets

To provide a comprehensive comparison of the IDEAS detectors with multivariate baselines, we report results for 12 metrics described above, across datasets A–F. Figures A4–A9 illustrate the mean performance over 100 time series in each group of each dataset, across 12 evaluation metrics

and 15 detectors (four IDEAS detectors, 10 multivariate baselines and the random baseline). These figures complement the VUS-PR analysis in the main paper by showing trends across threshold-based, threshold-independent, and range-based metrics. The final row in each block in Figures A4–A9 reports the mean RTW scores $\overline{\mathrm{RTW}}_{z,g,m}$, calculated as follows: Let $M_{z,g,r,m,q}$ denote the value of metric q obtained by detector m (with its default aggregation variant v_m^* if it is an IDEAS detector) on time series r in group g of dataset z. We define the ratio-to-winner (RTW) score as the ratio of $M_{z,g,r,m,q}$ to the highest value in the set $\{M_{z,g,r,m,q} : m \in \mathcal{M}\}$:

$$\mathrm{RTW}_{z,g,r,m,q} = \frac{M_{z,g,r,m,q}}{\max_{m \in \mathcal{M}} M_{z,g,r,m,q}}, \tag{A42}$$

where $\mathcal{M}$ is the set of all 15 detectors. Thus, the best detector for given time series and metric receives an RTW score of 1, and all other detectors are scored relative to that winner. For each detector m, we average RTW scores over all time series in the group and all evaluation metrics:

$$\overline{\mathrm{RTW}}_{z,g,m} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{|\mathcal{T}_g|} \sum_{r \in \mathcal{T}_g} \mathrm{RTW}_{z,g,r,m,q}, \tag{A43}$$

where $\mathcal{T}_g$ is the set of time series in group g and $\mathcal{Q}$ is the set of 12 evaluation metrics.

Overall, IDEAS variants remain strong across different metrics, and the best IDEAS detector often outperforms all baselines within a group. In most cases, the IDEAS adaptation improves over the corresponding base detector, while the best method among the IDEAS detectors varies across groups and metrics. Across datasets A–E, the strongest IDEAS detector depends on how anomaly signals are distributed across dimensions. In dataset A, where d_{anom} increases while d is fixed, EGI-IDEAS is consistently strong across groups, while CNN-IDEAS and PCA-IDEAS become more competitive as anomalies appear in more dimensions. Dataset B shows a similar advantage for EGI-IDEAS across different dimensionalities when anomalies remain confined to one dimension. In dataset C, where d_{anom} is sampled from $[1, d]$, EGI-IDEAS remains the strongest overall, while MP-IDEAS, CNN-IDEAS, and PCA-IDEAS also show strong performance, especially as d increases. For the datasets with varying time-series length, EGI-IDEAS is strongest in dataset D for most groups and metrics, while dataset E shows a transition toward better performances of PCA-IDEAS and CNN-IDEAS as n increases. Dataset F differs from the other datasets because anomalies in its time series are designed to preserve per-channel marginal distributions while disrupting cross-channel dependencies. In this setting, PCA-IDEAS is consistently the strongest across different dimensionalities, while CBLOF and OFA remain competitive on some of the metrics, in particular AUC-ROC and VUS-ROC.

A.8 Performance Comparison with Baselines for Real-World Datasets

For comparison on real-world data, we use benchmarks that include multiple datasets and focus on results for a particular metric (specifically, VUS-PR). Accordingly, for this comparison, we use a different version of the RTW score: Let $M_{b,r,m,q}$ denote the value of metric q obtained by detector m (with its default aggregation variant v_m^* if it is an IDEAS detector) on time series r in benchmark b. We define the RTW score as the ratio of $M_{b,r,m,q}$ to the highest value in the set $\{M_{b,r,m,q} : m \in \mathcal{M}\}$:

$$\mathrm{RTW}_{b,r,m,q} = \frac{M_{b,r,m,q}}{\max_{m \in \mathcal{M}} M_{b,r,m,q}}, \tag{A44}$$

where $\mathcal{M}$ is the set of the 14 detectors (not including the random baseline). Thus, the best detector for given time series and metric receives an RTW score of 1, and all other detectors are scored relative to that winner. For each detector m, we average RTW scores over all time series in the benchmark:

$$\overline{\mathrm{RTW}}_{b,m,q} = \frac{1}{|\mathcal{T}_b|} \sum_{r \in \mathcal{T}_b} \mathrm{RTW}_{b,r,m,q}, \tag{A45}$$

where $\mathcal{T}_b$ is the set of all time series in all datasets of benchmark b. Specifically, the final row in Figure 4 reports the mean RTW scores $\overline{\mathrm{RTW}}_{b,m,q}$ of equation A45 for $q =$ VUS-PR and b being the benchmark consisting of 20 real-world datasets. These results demonstrate that in terms of average performance on this benchmark, PCA, OmniAnomaly, and EGI-IDEAS are virtually tied for the top rank, with PCA-IDEAS and OFA following closely behind.

Results on TSB-AD-M. As shown in Figure A10, detector performance varies substantially across the source datasets within TSB-AD-M, and the best-performing detector changes from one source

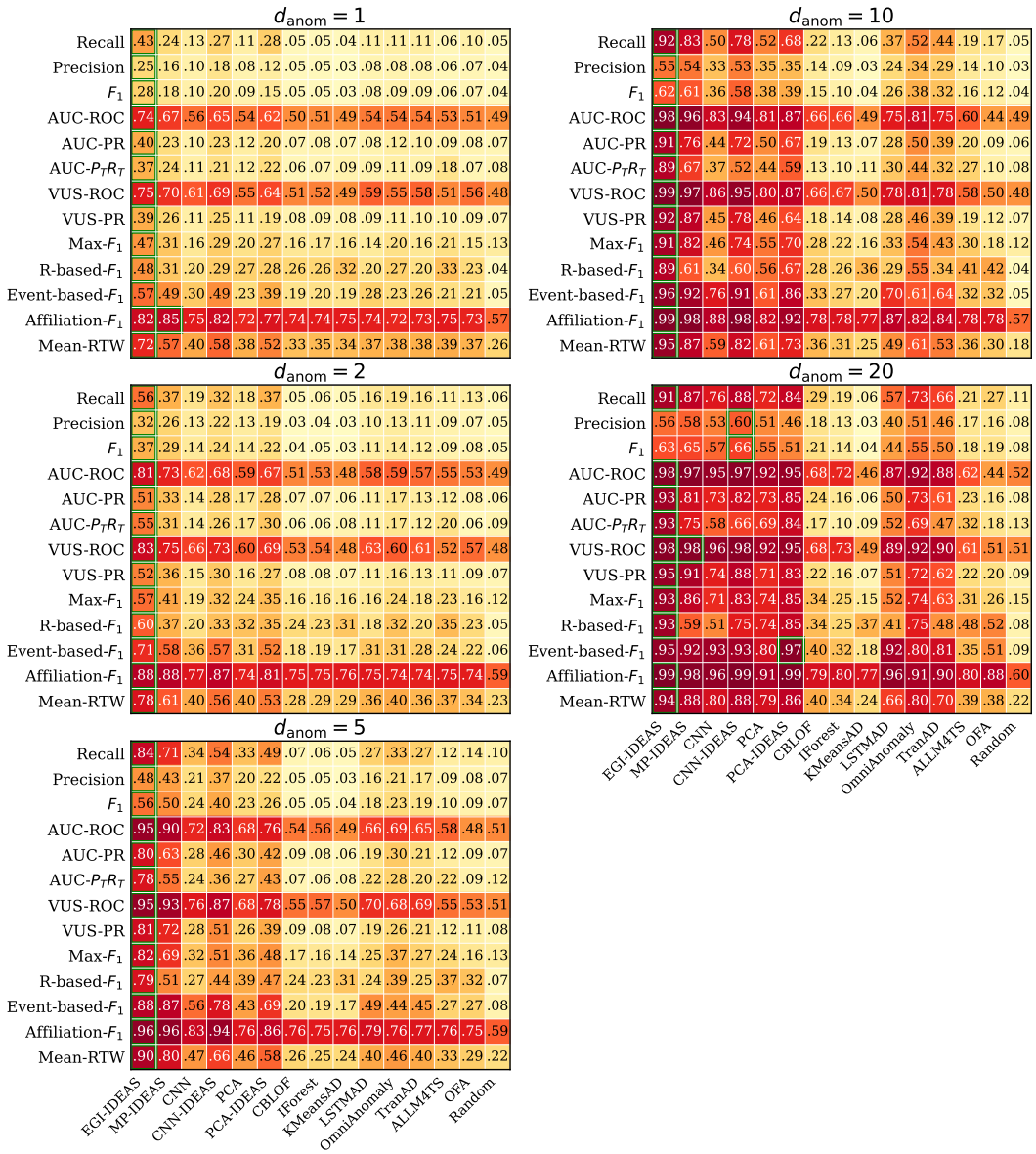


Figure A4: Detailed performance comparison on dataset A. Each block represents a group in the dataset, corresponding to one of the values $d_{\text{anom}} \in \{1, 2, 5, 10, 20\}$ of the number of anomalous dimensions. The total number of dimensions is fixed at $d = 20$ and the time-series length is fixed at $n = 10^4$. For the IDEAS detectors, the correlation threshold is set to $\lambda = 0.9$. Rows correspond to evaluation metrics, columns to detectors, and cell values report metric values averaged over all time series in the group. The final row in each block reports $\overline{\text{RTW}}_{z,g,m}$ scores of equation A43 for $z = \text{A}$.

dataset to another. This dataset-level variability is consistent with the broader results on real-world datasets discussed in the main paper. The final row in Figure A10 reports the mean RTW scores $\overline{\text{RTW}}_{b,m,q}$ of equation A45 for $q = $ VUS-PR and b being the TSB-AD-M benchmark. These results are qualitatively similar to those shown in Figure 4 for the benchmark of 20 real-world datasets. In terms of average performance on the TSB-AD-M benchmark, PCA performs the best, closely followed by OmniAnomaly and EGI-IDEAS.

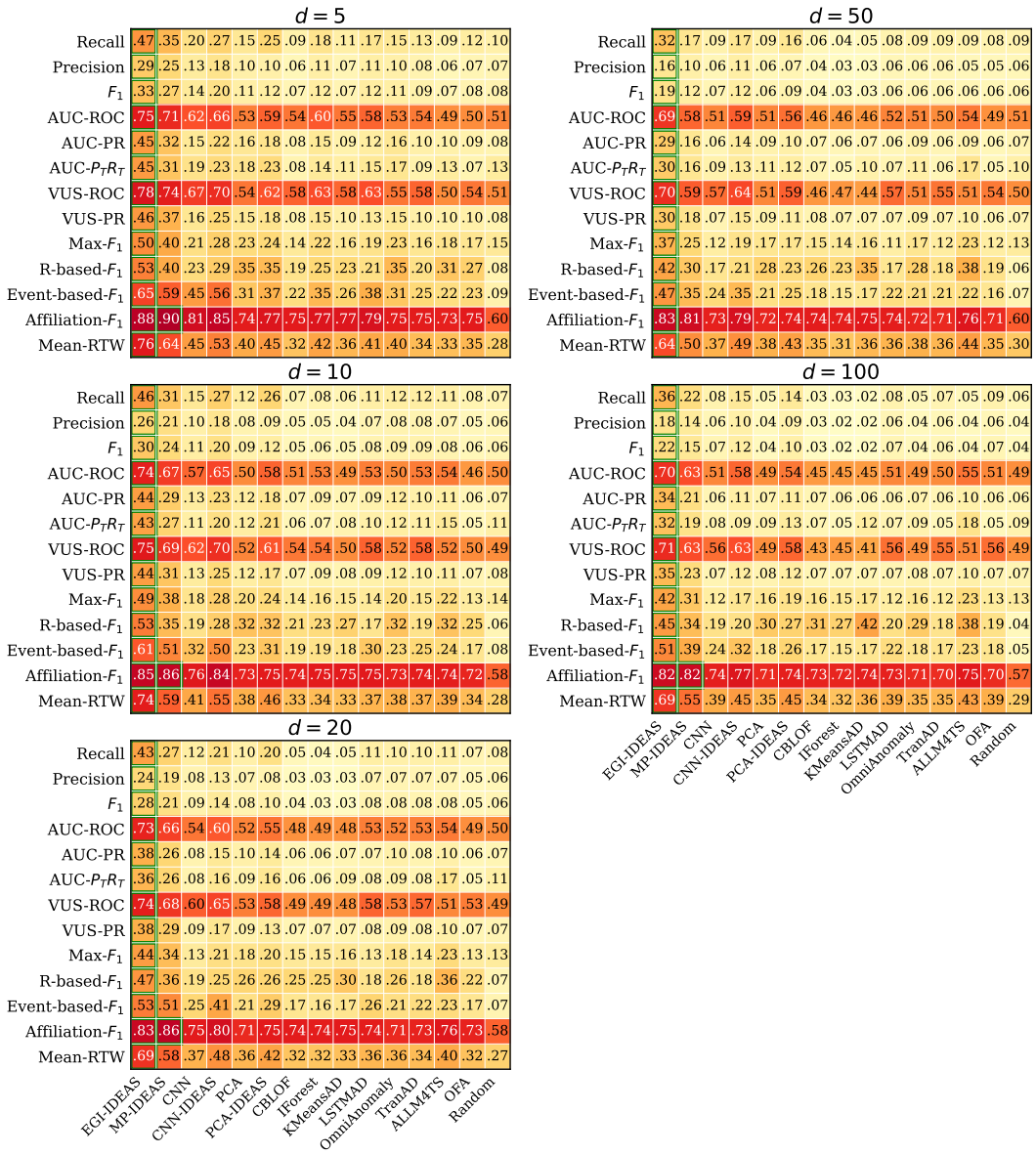


Figure A5: Detailed performance comparison on dataset B. Each block represents a group in the dataset, corresponding to one of the values $d \in \{5, 10, 20, 50, 100\}$ of the total number of dimensions. The number of anomalous dimensions is fixed at $d_{\text{anom}} = 1$ and the time-series length is fixed at $n = 10^4$. For the IDEAS detectors, the correlation threshold is set to $\lambda = 0.9$. Rows correspond to evaluation metrics, columns to detectors, and cell values report metric values averaged over all time series in the group. The final row in each block reports $\overline{\text{RTW}}_{z,g,m}$ scores of equation A43 for $z = \text{B}$.

A.9 Identification of Anomalous Dimensions

Dimension anomaly scores provide meaningful channel-level evidence. Table A5 reports the results of anomalous dimension identification by the four IDEAS detectors, for each group in datasets A–E. The results are represented by the top-1 ADP success rate for datasets B and D, and by the dimension-level F_1 score for datasets A, C, and E. Across the datasets, EGI-IDEAS gives the strongest performance in 22 out of 25 groups. On dataset B, where each time series contains anomalies in only one dimension, EGI-IDEAS achieves the highest top-1 ADP success rate for four of the five dimensionalities, with MP-IDEAS slightly better at $d = 50$. On dataset D, EGI-IDEAS is best across all time-series lengths, indicating that dimension anomaly score remains effective as

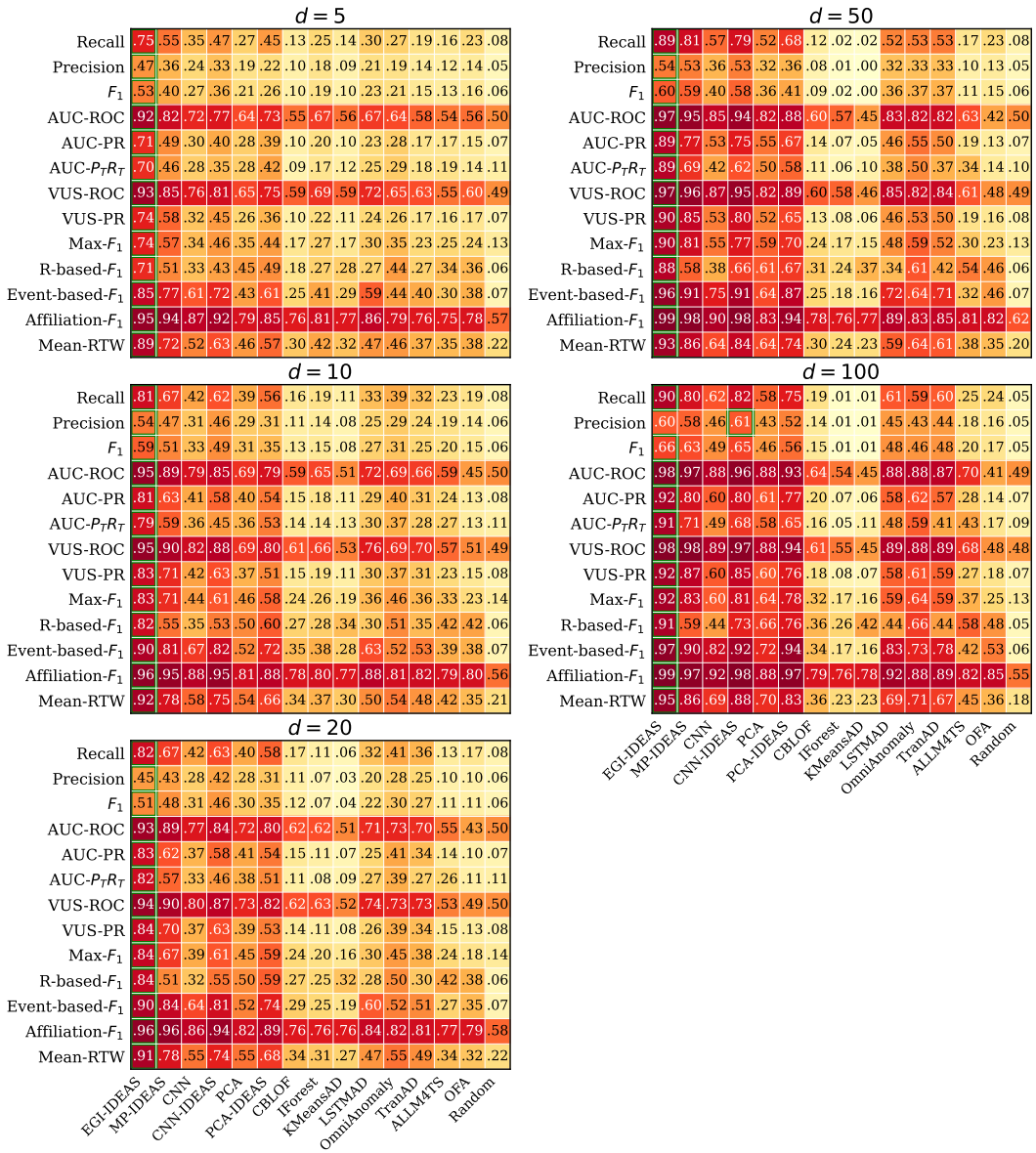


Figure A6: Detailed performance comparison on dataset C. Each block represents a group in the dataset, corresponding to one of the values $d \in \{5, 10, 20, 50, 100\}$ of the total number of dimensions. The number of anomalous dimensions is sampled from $[1, d]$ and the time-series length is fixed at $n = 10^4$. For the IDEAS detectors, the correlation threshold is set to $\lambda = 0.9$. Rows correspond to evaluation metrics, columns to detectors, and cell values report metric values averaged over all time series in the group. The final row in each block reports $\overline{\mathrm{RTW}}_{z,g,m}$ scores of equation A43 for $z = \mathrm{C}$.

n increases. For dimension-level F_1, EGI-IDEAS is also strongest in most groups: it performs best for $d_{\mathrm{anom}} \geq 5$ in dataset A and for all groups of datasets C and E. The random baseline increases in tandem with the fraction of anomalous dimensions because random selections are statistically more likely to overlap with the anomalous set. Nevertheless, EGI-IDEAS substantially outperforms the random baseline in every group, suggesting that the channel-wise scores used by IDEAS provide useful evidence for identifying which dimensions contain anomalous behavior.

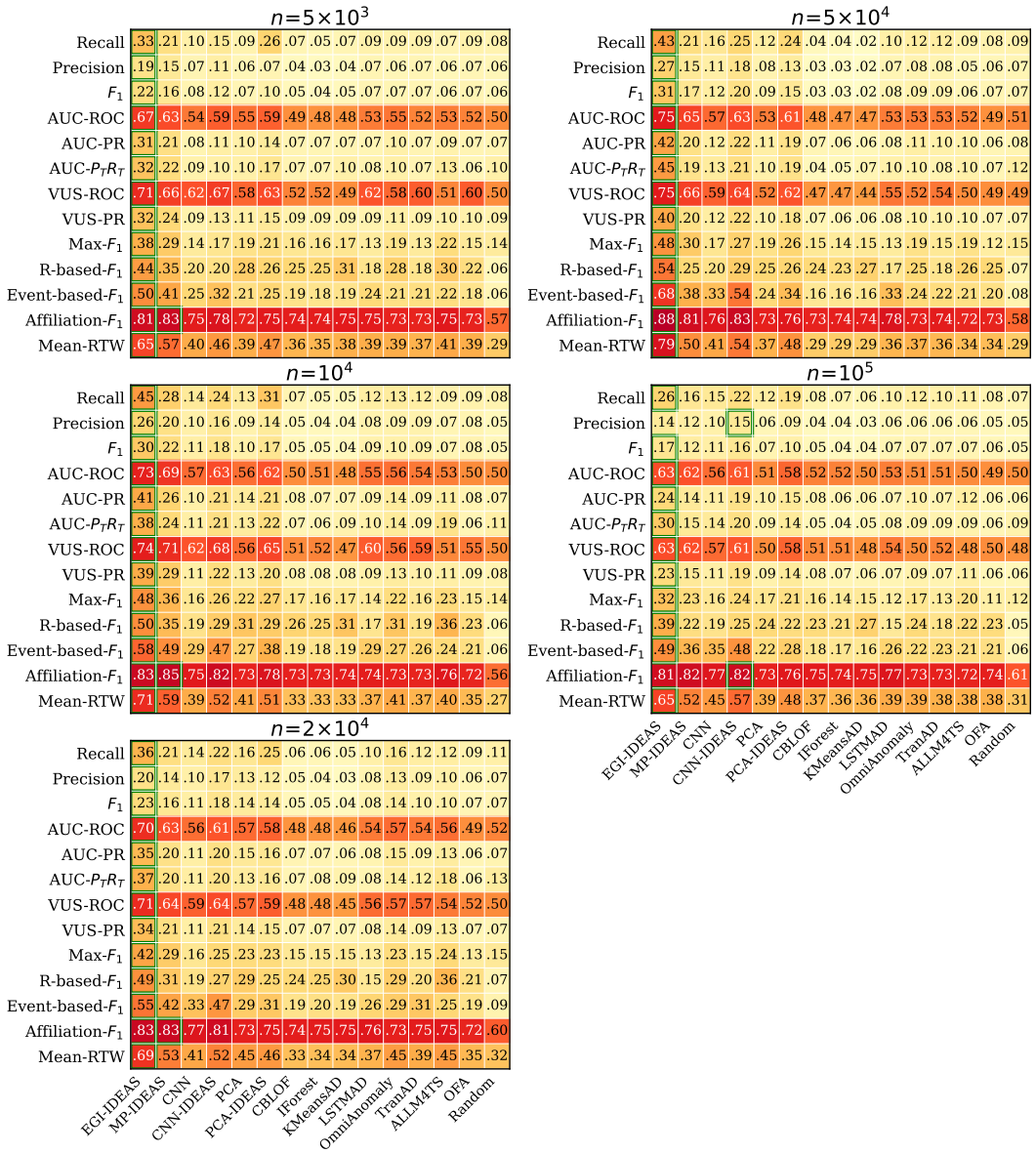


Figure A7: Detailed performance comparison on dataset D. Each block represents a group in the dataset, corresponding to one of the values $n \in \{5, 10, 20, 50, 100\} \times 10^3$ of the time-series length. The total number of dimensions is fixed at $d = 20$ and the number of anomalous dimensions is fixed at $d_{\text{anom}} = 1$. For the IDEAS detectors, the correlation threshold is set to $\lambda = 0.9$. Rows correspond to evaluation metrics, columns to detectors, and cell values report averages over all time series in the group. The final row in each block reports $\overline{\text{RTW}}_{z,g,m}$ scores of equation A43 for $z = \text{D}$.

Random baseline. For datasets B and D, where each time series contains a single anomalous dimension, a random prediction of this dimension has the success rate

$$\mathcal{R}^{(\text{rand})}_{\text{Top-1 ADP}} = \frac{1}{d}, \tag{A46}$$

where d is the total number of dimensions. For datasets A, C, and E, we assume that the predictor does not know *a priori* the number of anomalous dimensions,d_{anom}, and therefore selects $d/2$ dimensions as anomalous uniformly at random. Therefore, the expected number of correctly predicted anomalous dimensions is $d_{\text{anom}}/2$. For given d and d_{anom}, this yields the following result for the dimension-level F_1 score:

$$\text{F}^{(\text{rand})}_{1,\text{dim}}(d, d_{\text{anom}}) = \frac{d_{\text{anom}}}{d_{\text{anom}} + d/2}. \tag{A47}$$

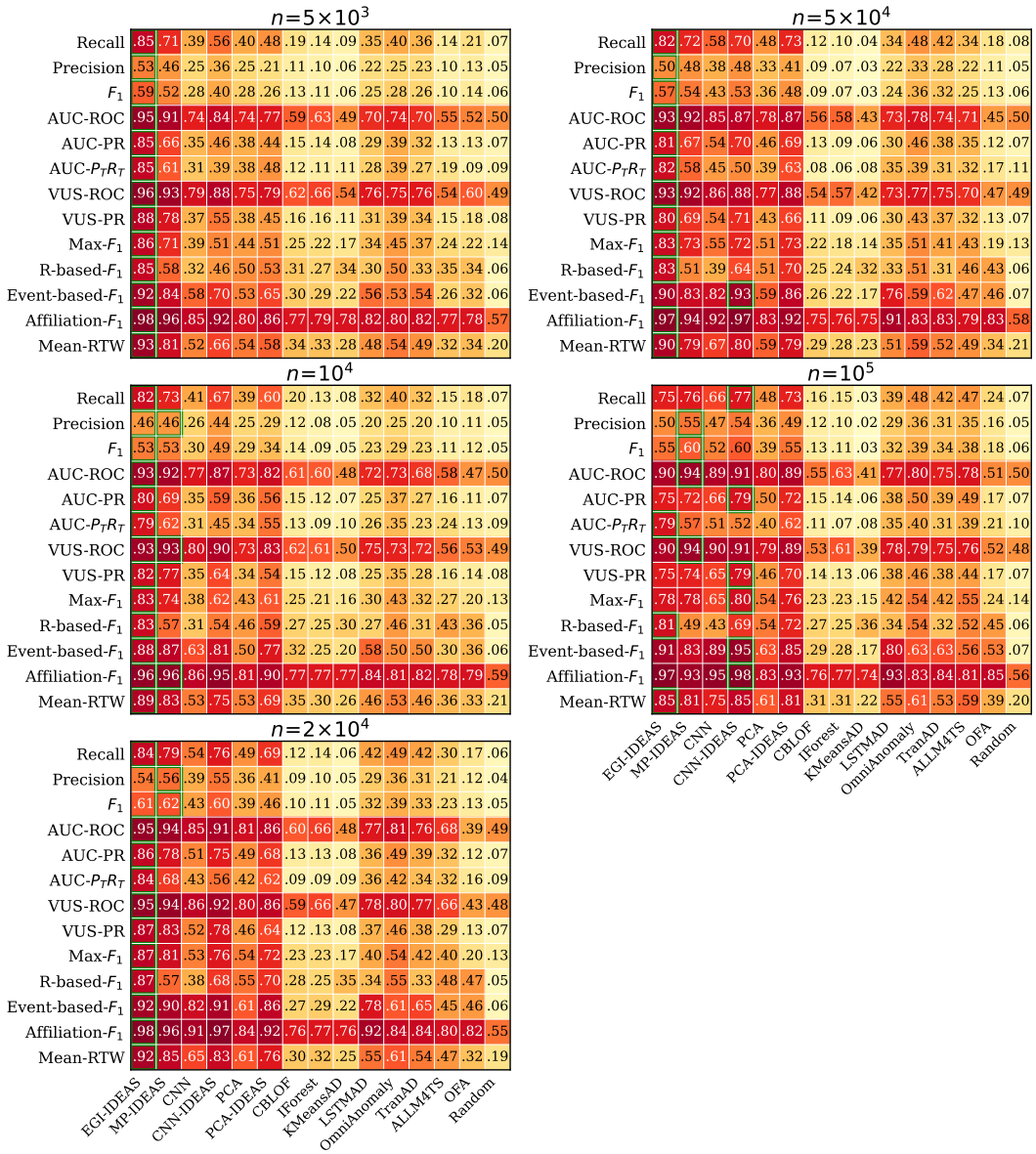


Figure A8: Detailed performance comparison on dataset E. Each block represents a group in the dataset, corresponding to one of the values $n \in \{5, 10, 20, 50, 100\} \times 10^3$ of the time-series length. The total number of dimensions is fixed at $d = 20$ and the number of anomalous dimensions is sampled from $[1, d]$. For the IDEAS detectors, the correlation threshold is set to $\lambda = 0.9$. Rows correspond to evaluation metrics, columns to detectors, and cell values report averages over all time series in the group. The final row in each block reports $\overline{\mathrm{RTW}}_{z,g,m}$ scores of equation A43 for $z = \mathrm{E}$.

In particular, for dataset A, where $d = 20$ in all groups, equation A47 yields $\mathrm{F}^{(\mathrm{rand})}_{1,\mathrm{dim}}(20, d_{\mathrm{anom}}) = d_{\mathrm{anom}}/(d_{\mathrm{anom}} + 10)$. For datasets C and E, where d_{anom} is sampled from $[1, d]$, we average the result of equation A47 over all possible values of d_{anom}, which yields

$$\mathrm{F}^{(\mathrm{rand})}_{1,\mathrm{dim}}(d) = \frac{1}{d} \sum_{d_{\mathrm{anom}}=1}^{d} \frac{d_{\mathrm{anom}}}{d_{\mathrm{anom}} + d/2}. \tag{A48}$$

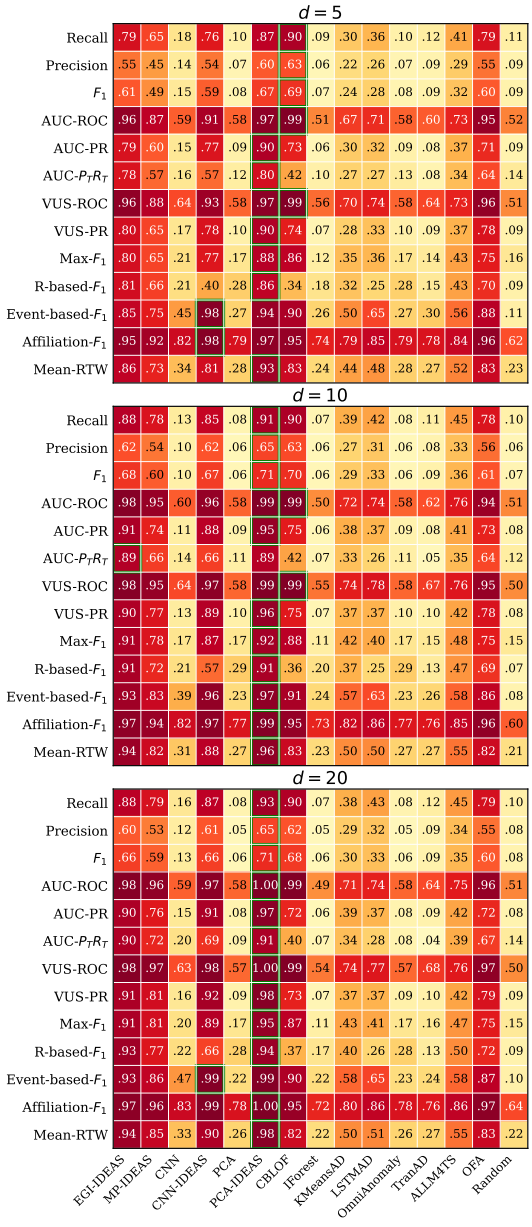


Figure A9: Detailed performance comparison on dataset F. Each block represents a group in the dataset, corresponding to one of the values $d \in \{5, 10, 20\}$ of the total number of dimensions. The number of anomalous dimensions is sampled from $[1, d]$ and the time-series length is fixed at $n = 10^4$. For the IDEAS detectors, the correlation threshold is set to $\lambda = 0.8$. Rows correspond to evaluation metrics, columns to detectors, and cell values report metric values averaged over all time series in the group. The final row in each block reports $\overline{\mathrm{RTW}}_{z,g,m}$ scores of equation A43 for $z = \mathrm{F}$.

A.10 Runtime Experiments

Computational setup. All experiments were conducted on a high-performance computing cluster using a single Intel CPU node with 48 cores at 2.9 GHz. Each method was evaluated under the same node allocation. For the IDEAS detectors, we parallelized channel-wise detector calls across up to 48 CPU cores, since each selected channel can be scored independently. Baseline detectors were run using their default settings and allocated the same amount of CPU cores.

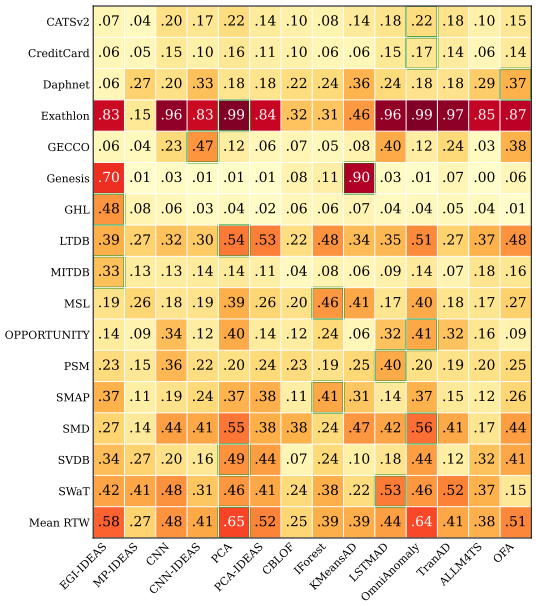


Figure A10: Mean VUS-PR performance on TSB-AD-M, separated by source dataset.

Table A5: Results of anomalous dimension identification by the four IDEAS detectors using the dimension anomaly score of equation 7, across all groups of datasets A–E. We report the top-1 anomalous-dimension prediction (ADP) success rate for datasets B and D, and the dimension-level F_1 score for datasets A, C, and E. The random baseline is included for comparison. The best result in each row is highlighted in bold.

Dataset	Metric	Group	EGI	MP	CNN	PCA	Random
A	Dim. F_1	$d_{\text{anom}} = 1$	0.239	0.082	**0.275**	0.160	0.091
		$d_{\text{anom}} = 2$	0.328	0.157	**0.329**	0.258	0.167
		$d_{\text{anom}} = 5$	**0.586**	0.350	0.535	0.392	0.333
		$d_{\text{anom}} = 10$	**0.710**	0.497	0.545	0.495	0.500
		$d_{\text{anom}} = 20$	**0.810**	0.534	0.461	0.605	0.667
B	Top-1 ADP	$d = 5$	**0.590**	0.510	0.480	0.210	0.200
		$d = 10$	**0.500**	0.440	0.390	0.080	0.100
		$d = 20$	**0.390**	0.350	0.250	0.090	0.050
		$d = 50$	0.180	**0.190**	0.110	0.100	0.020
		$d = 100$	**0.220**	0.140	0.070	0.133	0.010
C	Dim. F_1	$d = 5$	**0.706**	0.635	0.627	0.475	0.512
		$d = 10$	**0.720**	0.558	0.579	0.453	0.483
		$d = 20$	**0.620**	0.524	0.483	0.461	0.467
		$d = 50$	**0.630**	0.545	0.458	0.489	0.457
		$d = 100$	**0.621**	0.541	0.455	0.488	0.454
D	Top-1 ADP	$n = 5 \times 10^3$	**0.350**	0.260	0.170	0.082	0.050
		$n = 10^4$	**0.370**	0.310	0.290	0.081	0.050
		$n = 2 \times 10^4$	**0.380**	0.300	0.290	0.180	0.050
		$n = 5 \times 10^4$	**0.540**	0.200	0.260	0.230	0.050
		$n = 10^5$	**0.390**	0.170	0.180	0.150	0.050
E	Dim. F_1	$n = 5 \times 10^3$	**0.633**	0.606	0.477	0.474	0.467
		$n = 10^4$	**0.668**	0.553	0.454	0.473	0.467
		$n = 2 \times 10^4$	**0.734**	0.553	0.498	0.548	0.467
		$n = 5 \times 10^4$	**0.683**	0.446	0.487	0.518	0.467
		$n = 10^5$	**0.634**	0.409	0.522	0.545	0.467

EGI-IDEAS and MP-IDEAS are generally much faster than multivariate baselines. Figure A11 reports runtimes on datasets B and C as the dimensionality d varies, while Figure A12 reports runtimes on datasets D and E as the time-series length n varies. To ensure a fair comparison across detectors,

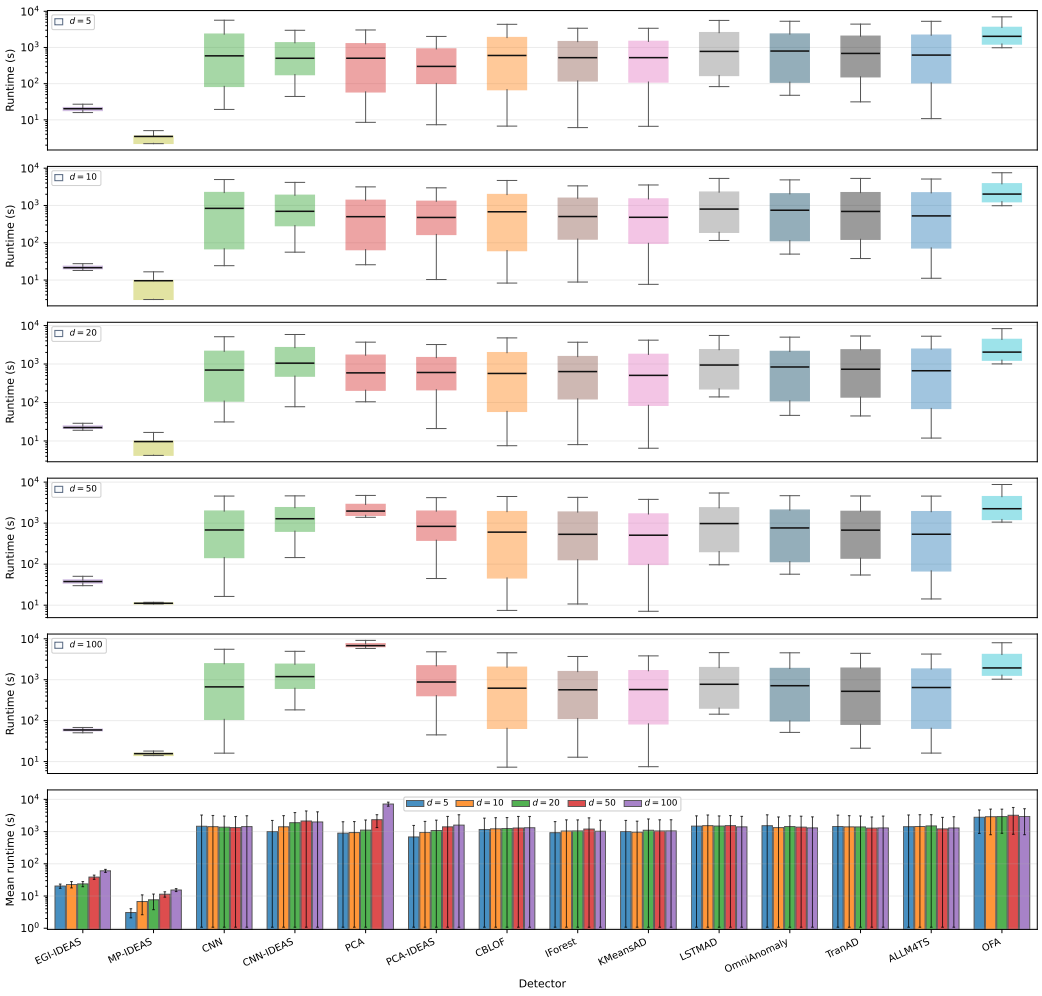


Figure A11: Runtime comparison of all detectors on datasets B and C across five groups with different dimensionalities ($d \in \{5, 10, 20, 50, 100\}$). For each detector and each value of d, runtimes are summarized over 200 runs, consisting of 100 runs from each dataset. In the box plots, the center line denotes the median runtime and each box spans the interquartile range (25th–75th percentiles). The bottom panel shows mean runtimes with standard-deviation error bars.

we selected one representative time series from each of the 20 groups in datasets B–E: five groups with varying dimensionalities from each of datasets B and C, and five groups with varying time-series lengths from each of datasets D and E. Each detector was run 100 times on each selected time series, resulting in 2,000 runtime measurements per detector. Within each batch, jobs for all detectors were submitted simultaneously so they experienced similar computing conditions. Thus, the observed runtime spreads capture the combined variability of time-varying system loads while preserving comparability across detectors within each batch.

Across all experiments with varying dimensionality, MP-IDEAS and EGI-IDEAS achieved the lowest and second-lowest median runtimes, respectively, and were, in general, significantly faster than the multivariate detectors. The runtime gap between these two IDEAS detectors and the multivariate baselines remains substantial as d increases from 5 to 100. This reflects the fact that IDEAS applies the base detector independently to each channel, making the computation highly parallelizable across dimensions. Therefore, if the base detector is efficient, its IDEAS implementation will be correspondingly fast. Among the evaluated methods, PCA shows the steepest increase in runtime with dimensionality, while most other detectors are less sensitive to d variations over this range.

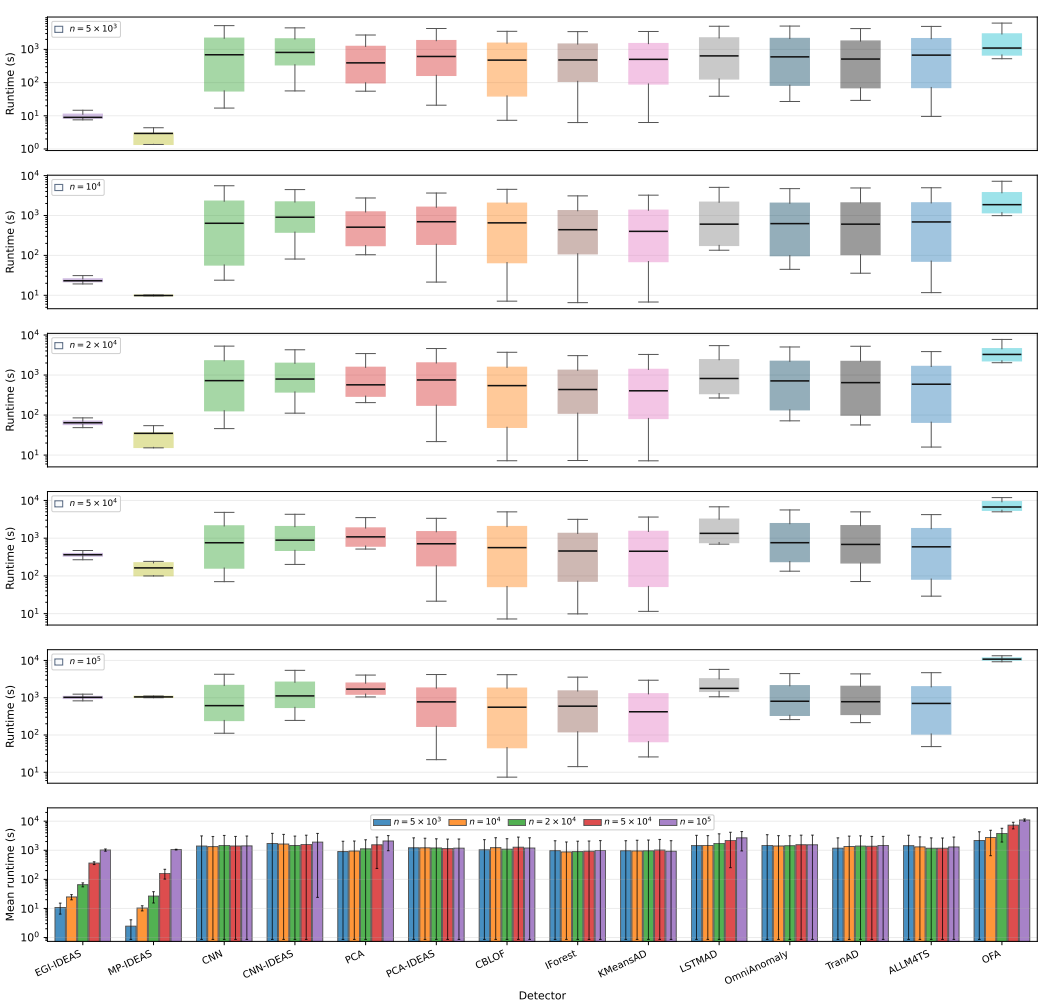


Figure A12: Runtime comparison of all detectors on datasets D and E across five groups with different time-series lengths ($n \in \{5, 10, 20, 50, 100\} \times 10^3$). For each detector and each value of n, runtimes are summarized over 200 runs, consisting of 100 runs from each dataset. In the box plots, the center line denotes the median runtime and each box spans the interquartile range (25th–75th percentiles). The bottom panel shows mean runtimes with standard-deviation error bars.

In experiments with varying time-series length, runtimes increase for several methods as n grows, most notably for MP-IDEAS and EGI-IDEAS, but also for PCA, LSTMAD, and OFA. While MP-IDEAS and EGI-IDEAS remain the fastest and second-fastest detectors, respectively, for short and medium-length time series, the runtime gap between them and the multivariate baselines narrows as n increases, practically disappearing at $n = 10^5$.

It is worth noting that MP-IDEAS and EGI-IDEAS exhibit very little run-to-run variability, whereas the opposite is true for the multivariate detectors (including their IDEAS implementations), as is evident from interquartile ranges and standard-deviation error bars in Figures A11 and A12. Based on the experimental design described above, we conclude that this wide runtime spread reflects variability within the shared computing environment rather than differences in the experimental configuration. Specifically, we observed that runtimes can vary substantially depending on when jobs are submitted due to fluctuating cluster loads and resource availability.

To summarize, these runtime experiments demonstrate that combining the independent channel-wise IDEAS framework with a highly efficient univariate baseline detector provides a significant practical speedup over natively multivariate methods, alongside superior anomaly detection performance.