# Interpretable Multivariate Anomaly Detector Selection for Automatic Marine Data Quality Control

Ngoc-Thanh Nguyen*, Astrid Marie Skålvik†, Emmanouil Sylligardos‡, Rogardt Heldal*,
Patrizio Pelliccione§, Paul Boniol‡, Themis Palpanas¶, Sverre Jakob Alvsvåg‖
*Western Norway University of Applied Sciences, †University of Bergen, ‡DI ENS, ENS, PSL, CNRS, Inria,
§Gran Sasso Science Institute, ¶Université Paris Cité, ‖Reach Subsea AS

***Abstract*—We propose MuMSAD, an AutoML framework for automatic selection of 10 interpretable multivariate anomaly detectors. MuMSAD has two main contributions: (i) a systematic screening that results in the first curated library of 10 interpretable multivariate detectors, (ii) enabling dimension-level interpretable anomaly detection. We evaluate MuMSAD in root cause analysis of problems observed in multi-parameter smart ocean monitoring systems. The results show that the automatic model selection improves interpretability by up to 20% and accuracy by 15% compared to single anomaly detectors, while maintaining competitive runtime efficiency. Such results realize the implementation of automatic multi-parameter marine data quality control.**

***Index Terms*—interpretable multivariate anomaly detection, model selection, marine data quality control**

## I. Introduction

Multivariate time series (MTS) are a core data source for ensuring the reliable operation of cyber-physical systems [1]. Detecting anomalies in such data is crucial for preventing critical failures with potentially severe consequences [1], [2]. While existing multivariate anomaly detection (AD) methods can identify irregularities, most cannot indicate *which dimensions* are problematic, limiting their usefulness for root-cause analysis [1], [3].

Recent research has proposed interpretable multivariate AD methods that provide dimension-level anomaly scores, allowing users to trace anomalies back to their sources [1], [2]. Yet, there is no one-size-fits-all AD method for heterogeneous time series: a detector that performs well on one dataset may fail on another [4], [5]. Ensemble approaches can improve accuracy but introduce high runtime costs [4]. Automated Machine Learning (AutoML) solutions such as MSAD [4] address such problems but for univariate time series only. Consequently, practitioners lack a general solution that combines accuracy, interpretability, and efficiency for MTS anomaly detection.

To address this gap, we propose **MuMSAD** (Multivariate Model Selection for Anomaly Detectors), an AutoML framework for automatic selection of interpretable multivariate anomaly detectors. MuMSAD adapts MSAD [4] principles to MTS, and goes beyond it by supporting interpretability for ten state-of-the-art anomaly detectors reviewed in [5], which enables them to produce both overall and dimension-level anomaly scores for actionable root-cause analysis. We particularly focus on its application to automatic data quality control in marine monitoring systems, a domain where reliable anomaly detection is both practically and economically critical [6], [7].

On real-world datasets, MuMSAD improves interpretability by up to 20% and accuracy by 15% compared to the best single detector, while maintaining competitive runtime efficiency. These results demonstrate the potential of MuMSAD as a practical and general solution for interpretable anomaly detection in large-scale, real-time monitoring systems. The framework can be found at https://github.com/ntnguyen-so/MuMSAD_framework.

**Outline:** Section II details the practical need for automatic marine data quality control (DQC). Section III introduces notations and related work. Section IV formulates the problem of automatic model selection for interpretable multivariate anomaly detectors. Section V presents our proposed solution, MuMSAD. Section VI demonstrates its application to automatic multi-parameter marine DQC. Finally, Section VII concludes the paper and discusses future directions.

## II. Motivation

This section highlights the practical need for automatic data quality control within the marine domain. We then show how existing solutions fail to meet the highlighted needs.

### A. Practical Need for Automatic Marine Data Quality Control

Ocean industries are projected to contribute at least three trillion USD annually to the global economy by 2030 [8]. Achieving this growth depends on reliable marine data [7]. However, marine data is notoriously prone to errors. Factors such as biofouling, sensor drift, extreme environmental conditions, and communication failures frequently contaminate the data [6], [7], [9]. Marine DQC is therefore critical, but it remains largely manual. Expert-driven inspection can take up to six months for a single dataset [7]. The complexity arises not only from the volume of data collected by observatories, but also from the interdependence of parameters such as temperature, salinity, and conductivity.

Figure 1(a) illustrates the aforementioned challenge, with detailed explanations being included on the figure. Standard univariate detectors struggle to distinguish these cases. Figure 1(b) shows how *dimensional anomaly scores* can be helpful by highlighting not only when anomalies occur but also which parameters are responsible. They allow domain experts to perform root-cause analysis rather than relying solely on raw signal inspection.

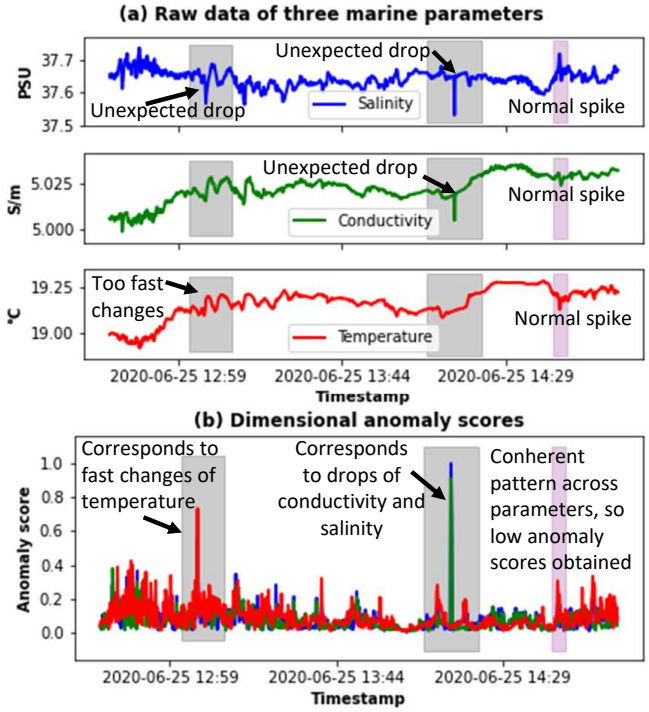


Fig. 1: (a) A three-parameter time series from marine data. (b) Dimensional anomaly scores of all three parameters. In these figures, gray subsequences are anomalous, while the purple is normal, with reasons explained.

### B. Open Challenges

Several open problems are related to the above needs:

**1) Univariate AD for MTS root cause analysis.** Applying univariate anomaly detectors independently to each dimension of an MTS is straightforward but suboptimal, as it ignores interdependencies between variables and leads to false alarms [1], [10]. Few univariate methods are tailored to marine data, and they require handcrafted thresholds and domain-specific rules [11].

**2) Limited interpretable MTS detectors.** Although interpretable AD methods have emerged [1], [3], only a handful support dimension-level anomaly scoring. Moreover, the interpretability they provide is often ad hoc and inconsistent across detectors: some return ranked features, while others produce attribution scores without clear semantics.

**3) Lack of AutoML solutions for interpretable MTS AD.** Recent AutoML frameworks such as MSAD [4] advance model selection but are limited to univariate series and ignore interpretability.

## III. Background and Related Work

### A. Notations

Let $T = \{x_1, x_2, \ldots, x_N\}$ denote a time series of length $N$, where each $x_t$ is the $t^{\text{th}}$ observation. For univariate series, $x_t \in \mathbb{R}$; for multivariate series, $x_t \in \mathbb{R}^M$ with $M \in \mathbb{N}^+$ dimensions. We refer to the latter as a MTS. In our marine context, each dimension corresponds to a parameter such as seawater temperature, salinity, or conductivity (see Figure 1).

An anomaly is defined as a subsequence $T_{i,j} = \{x_i, \ldots, x_j\}$ that deviates significantly from expected normal behavior. Most real-world datasets, including our marine data, are sampled at constant acquisition rates, which is a common assumption in existing methods [5]. For MTS $T$, an anomaly detector $D$ outputs an anomaly score sequence $S \in \mathbb{R}^N$; interpretable detectors further provide per-dimension scores $S^{(m)}$ for $m = 1, \ldots, M$.

### B. Evaluation of Anomaly Detectors

Interpretable MTS anomaly detectors are evaluated along two dimensions: **(i) Accuracy.** Given ground-truth anomaly labels $L \in \{0,1\}^N$, accuracy functions measure the alignment between $S$ and $L$. Common metrics are AUC-ROC, AUC-PR, and F-score [5]. However, in domains like ours where degenerate cases arise (e.g., entire subsequences labeled as all normal or all anomalous), these standard metrics are insufficient. In Section VI, we therefore extend them with a unified definition, $AD_{\text{acc}}$, that remains valid across all cases. **(ii) Interpretability.** Following [1], [2], we use $HitRate@K = \frac{|Hit@K|}{|GT_t|}$, where $GT_t$ is the set of true anomalous dimensions at time $t$ and $Hit@K$ is the number of those captured in the top-$K$ predicted dimensions. Ideally, $K = |GT_t|$.

### C. Univariate and Multivariate Anomaly Detection

Univariate anomaly detectors analyze each dimension separately, which is simple but ignores cross-dimensional dependencies, leading to false alarms [1], [2]. Multivariate anomaly detectors capture correlations across variables and generally improve accuracy, but they typically lack interpretability, making root-cause analysis difficult in practice [1], [2].

### D. Interpretable Anomaly Detection

A handful of interpretable MTS anomaly detectors have been proposed [1], [2]. These methods provide dimension-level or feature-level anomaly scores, helping users identify the sources of anomalies. However, their interpretability is often inconsistent: some provide ranked dimensions, others output attribution maps without clear semantics. This heterogeneity complicates their integration into industrial workflows, where transparent explanations are essential.

### E. AutoML and Ensemble Approaches

Several AutoML frameworks have been introduced to automate anomaly detector selection and tuning. MSAD [4] learns to select suitable detectors automatically, while TimeEval [12] leverages metadata for hyperparameter tuning. Classifier-based selectors [13] improve robustness but rely on handcrafted features. Importantly, most of the existing AutoML solutions focus on univariate time series, and none addresses interpretability.

### F. Positioning of Our Work

Existing methods either (i) focus on univariate data, (ii) ignore interpretability, or (iii) trade accuracy for runtime

efficiency. Crucially, no prior AutoML framework has systematically ensured dimension-level interpretability for multivariate anomaly detection — i.e., the ability to explain which parameter caused an anomaly. MuMSAD is the first to curate a library of interpretable multivariate detectors and enforce a consistent interpretability criterion across them, enabling actionable root-cause analysis; while still ensuring competitive performance by applying the AutoML principles.

## IV. Problem Formulation

To reason about automatic model selection for interpretable multivariate anomaly detection, we first formalize the problem. The goal is to design a selector that chooses, for each given multivariate time series (MTS), the detector that yields the best balance of accuracy and interpretability.

**Input.** Let $\mathcal{S}$ be a set of MTS, with variable lengths $|T|$. Let $\mathcal{B} = \{D_1, \ldots, D_K\}$ denote a finite set of $K$ interpretable multivariate anomaly detectors.

**Automatic model selection.** We define $\mathcal{M}$ as a selector $\mathcal{M} : \mathcal{S} \to \mathcal{B}$. Given a time series $T \in \mathcal{S}$ with label $L$, $\mathcal{M}$ selects the detector maximizing a suitability function $\mathcal{F}$: $\mathcal{M}(T) = \arg\max_{D \in \mathcal{B}} \mathcal{F}(D(T), L)$. Here, $\mathcal{F}$ can capture anomaly detection accuracy, interpretability, or a combination (Section III-B). This setting can be viewed as a classification task where classes correspond to detectors in $\mathcal{B}$.

For evaluation, we compare $\mathcal{M}$ against three reference baselines: (i) an *Oracle*, which selects the detector that maximizes $\mathcal{F}$ with perfect knowledge; (ii) an *Averaging Ensemble*, which runs all detectors in $\mathcal{B}$ and averages their scores; and (iii) a *Random* selector, which chooses uniformly from $\mathcal{B}$. These baselines respectively provide upper, ensemble, and lower bounds for $\mathcal{M}$'s performance.

**Output.** Given $T$ and selected detector $D = \mathcal{M}(T)$, anomaly detection produces two outputs: (i) a sequence of point-level anomaly scores $S \in [0,1]^{|T|}$; and (ii) for interpretable detectors, per-dimension scores $S_t = \{S_t^1, \ldots, S_t^M\}$ for each observation $T_t$, with $S_t^i \in [0,1]$. These are aggregated through

$$\mathcal{H}(S_t) : [0,1]^M \to [0,1] \tag{1}$$

which combines dimension-level scores into a point-level score. For non-interpretable detectors, only scalar scores $S_t \in [0,1]$ are produced.

## V. Proposed Solution: Interpretable Multivariate Anomaly Detector Selection

We introduce **MuMSAD**, a framework for automatic selection of interpretable multivariate anomaly detectors. MuMSAD extends MSAD [4] by (i) supporting heterogeneous multivariate time series, (ii) incorporating extended feature extraction and flexible labeling, (iii) enforcing dimension-level interpretability, and (iv) reducing runtime costs compared to ensembles. Figure 2 shows the architecture.

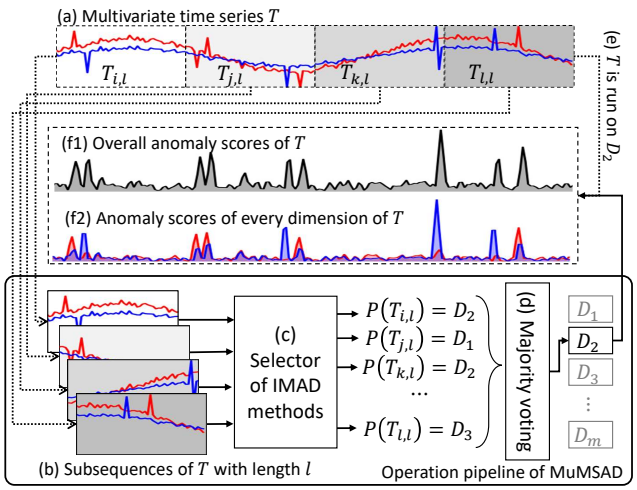


Fig. 2: Pipeline of MuMSAD.

### A. Pipeline Overview

MuMSAD consists of four stages:

1) **Preprocessing:** Segments input MTS into fixed-length subsequences, enabling robust feature extraction and handling variable-length inputs.
2) **Prediction:** Each subsequence is encoded into features and passed to a selector (feature-based classifier or deep model) that predicts the most suitable detector.
3) **Selection:** Subsequence-level predictions are aggregated via majority voting into a single detector. We choose majority voting for robustness and runtime simplicity compared to weighted voting.
4) **Anomaly Detection:** The selected detector is applied to the full MTS to generate (i) point-level scores $S$ and (ii) dimension-level scores $S_t^i$, enabling root-cause analysis.

We direct readers to the detailed workflow of MSAD at [4]. Below we outline the novelty of MuMSAD.

### B. Extended Feature Extraction and Labeling

**Feature extraction** is a key bottleneck for model selectors. MSAD [4] relied on nine simple TSFresh features, mostly basic statistics (mean, median, variance, min, max), which are often insufficient for heterogeneous MTS. In MuMSAD, we expand the feature space along three axes, while also reconfiguring the preprocessing steps so that classifiers originally designed for univariate inputs can operate effectively on MTS, maintaining backward compatibility with univariate series.

*(i) Statistical features.* We retain the original TSFresh statistics and add two measures of local variation: the mean of absolute differences between consecutive values, capturing sudden irregular changes, and the mean of signed differences, highlighting persistent drifts in stationary parameters such as temperature. These are relevant in marine data streams, where short-term turbulence and long-term drift can coexist.

*(ii) Shape-based features.* We incorporate the number of peaks in a subsequence, which identifies unusual oscillations. Such peaks often indicate sensor malfunctions or disturbances

caused by environmental factors such as biofouling or mechanical noise during maintenance [6].

*(iii) Distributional features.* We adopt Benford's correlation coefficient, widely used in anomaly detection [14], to capture deviations in the expected distribution of measurement values. This measure has shown strong performance in domains such as fraud detection, and here it helps identify cases where sensor outputs follow unnatural digit distributions.

Beyond TSFresh, we also evaluate catch22 [15], which computes 22 handcrafted features designed for broad applicability across time series domains. Catch22 is considerably more efficient than TSFresh [16], enabling MuMSAD to reduce feature extraction overhead when runtime is critical.

**Normalized vs. raw data.** While MSAD operated solely on normalized data, we observed that normalization can sometimes suppress patterns that are important for root-cause analysis. For example, absolute shifts in salinity or conductivity are meaningful in marine data, but these can be attenuated by normalization. Therefore, MuMSAD extracts features from both raw and normalized inputs.

### C. Integrating Interpretability

Interpretability is central to MuMSAD. We require each interpretable detector $D$ to satisfy the condition of Eq. 1. In practice, we adopt $\mathcal{H}(S_t) = \sum_{i=1}^{M} S_t^i$, where $S_t^i \in [0, 1]$ is the anomaly score of dimension $i$ at time $t$. This simple aggregation ensures that every point-level anomaly score $S_t$ is supported by a decomposition across dimensions, enabling root-cause analysis. While other aggregation strategies are possible (e.g., max, weighted sum), we adopt the summation because it is widely supported by existing detectors and yields consistent semantics across different methods.

From the 71 anomaly detectors reviewed in [5], we systematically identified the subset of 10 multivariate methods that meet this interpretability criterion. They can be grouped by family according to the taxonomy of [17]: *Prediction-based* (Torsk, AutoEncoder, DenoisingAutoEncoder, PCC), *Distance-based* (CBLOF, COF, LOF), and *Density-based* (RandomBlackForest, HBOS, COPOD).

We emphasize that this screening is itself a contribution. Although prior surveys [5], [17] categorize detectors by algorithmic family, none distinguish which methods produce dimension-level scores. By applying a consistent interpretability criterion, MuMSAD identifies the detectors that can support root cause analysis.

The functionality of interpretability is critical in practice. In industrial contexts such as subsea monitoring, operators not only need to know *what* an anomaly occurred but also *which parameter* (e.g., salinity, pressure) caused it. Dimension-level scores are therefore essential for isolating faulty sensors and reducing downtime.

While the replication package of MuMSAD consists of 10 interpretable detectors, the framework remains general and can still support non-interpretable methods. Specifically, when a non-interpretable method is selected, only point-level scores $S_t$ are produced, without dimension-level explanations.

## VI. Experimental Evaluation

We evaluate **MuMSAD** as an automatic multi-parameter marine DQC module for smart ocean observatories.

*1) Data Preparation:* We use data from the OBSEA cabled observatory (https://obsea.es), located 4 km off the Barcelona coast. Two SeaBird CTD nodes, deployed at 20 meters depth, continuously recorded temperature, conductivity, and salinity at 10-second intervals between 2020–2021. These parameters are strongly interdependent (e.g., temperature and conductivity jointly determine salinity), making the dataset well-suited for evaluating multivariate anomaly detection.

For near-real-time DQC operations, we segmented the stream into one-day multivariate time series (MTS), resulting in 554 labeled series, each of length $8640$ and with three dimensions. Labels of bad data were created using the physics-informed diagnostic rules of [6].

*2) Technical Setup:*

*a) Detector execution:* We ran all ten interpretable multivariate anomaly detectors identified in Section V-C on every series in the OBSEA dataset. Hyperparameters were tuned automatically using the TimeEval framework [12], which optimizes settings based on meta-information of each MTS. The computations were carried out on two Linux servers, each equipped with 48 cores of Intel(R) Xeon(R) Gold 6136 CPUs at 3.00GHz and 64GB RAM.

*b) Model selector training:* To build MuMSAD selectors, we split the OBSEA dataset into 85% training and 15% testing sets, stratified to maintain similar class distributions of anomalies across both splits. All supported selector architectures were evaluated, including both feature-based (e.g., Decision Tree, RF, SVM, AdaBoost, MLP, kNN, QDA, Naïve Bayes) and deep learning–based selectors (ConvNet, ResNet, InceptionTime, SiT variants).

We tested input window sizes from $\{4, 8, 16, 32, 64, 128, 256, 512, 768, 1024\}$, covering both short local subsequences and long-range temporal contexts. For feature-based selectors, we evaluated both TSFresh and catch22 feature extractors. To understand the effect of preprocessing, we extracted features from both raw and normalized data. In total, 390 distinct selector configurations were trained, giving us a broad empirical basis for comparing design choices.

*c) Evaluation metrics:* We evaluated MuMSAD in two aspects: anomaly detection accuracy and interpretability. Accuracy was assessed with the domain-specific function $AD_{acc}$:

$$AD_{acc}(T, L) = \begin{cases} \textit{AUC-PR}, & \text{if } \{0, 1\} \in L \\ 1 - \textit{FPR}, & \text{if } \{0\} \in L \\ \textit{TPR}, & \text{if } \{1\} \in L, \end{cases} \tag{2}$$

where $T$ is a multivariate time series and $L$ its label sequence. This formulation is necessary because many OBSEA series are degenerate cases: some days contain only normal data, while others contain only anomalies (e.g., due to sustained communication failure). In such cases, conventional precision/recall-based metrics become undefined. Equation 2 ensures comparability across all three scenarios, with outputs always in $[0, 1]$.

Interpretability was measured using the $HitRate@K$ metric (see Section III-B). We set $K = 1$, corresponding to the most operationally relevant use case: whether the method correctly identifies the single most problematic parameter (e.g., conductivity vs. temperature).

*3) Results:*

*a) Interpretability:* Figure 3 reports interpretability across all tested detectors and selectors. The variation among individual detectors is striking: the AutoEncoder (AE) achieves a score around 0.5, while others such as CBLOF or Torsk fall much lower. This shows that simply choosing a single method yields inconsistent interpretability. The *Oracle* establishes the upper bound by always picking the optimal detector. There remains a substantial gap between Oracle and AE. Conversely, *Random* collapses performance, confirming that uninformed choice is close to useless. Interestingly, the *Averaging Ensemble* (Avg Ens), which aggregates anomaly scores across detectors, performs worse than every individual method. This result contrasts with [4], where score averaging improved accuracy in the univariate setting.

MuMSAD selectors close much of the gap to the Oracle. The kNN selector trained on TSFresh features extracted from raw data, with window size 256, yields the highest gains, improving interpretability by 10–20% over AE depending on whether mean or median scores are considered. Among deep learning selectors, four out of seven are competitive, though in aggregate the feature-based family remains stronger.

Two trends emerge. First, *raw features outperform normalized*. Normalization erases correlations and scale differences between parameters, which are often diagnostic in marine sensor data (e.g., conductivity drifting while temperature remains stable). Retaining raw features preserves these cues and improves interpretability. Second, *smaller windows outperform larger windows*. Shorter segments provide more subsequences for training, which increases the selector's ability to generalize across heterogeneous conditions. This is particularly important in marine data, where anomalies may be short-lived but repeated across time. These results confirm MuMSAD's value: careful selection, rather than naive aggregation or reliance on a single detector, substantially enhances interpretability.

*b) Anomaly detection accuracy:* Figure 4 summarizes anomaly detection accuracy using $AD_{acc}$ (Equation 2). The trade-off with interpretability becomes clear: AE ranks best for interpretability but only mid-pack for accuracy (fourth worst overall). In contrast, detectors like Torsk, CBLOF, and PCC achieve high accuracy but are among the least interpretable. This illustrates the central challenge: no single detector provides strong performance on both dimensions simultaneously.

MuMSAD selectors mitigate this trade-off effectively. Feature-based selectors such as DT, MLP, RF, kNN, and AdaBoost consistently achieve both higher accuracy and higher interpretability than any individual detector. This result is crucial for operational settings, where reliability and explainability must co-exist.

Several patterns are robust across the experiments. First, *catch22 outperforms TSFresh*: although TSFresh with our extensions is strong, the compact and efficient catch22 feature set appears more suited to small datasets and real-time constraints. Second, *raw data outperforms normalized data*, mirroring the interpretability results. Third, *smaller windows outperform larger ones*, again reflecting that more subsequences yield more training examples for selectors. Finally, the *Avg Ens* baseline again underperforms, confirming that aggregation is not a substitute for selection in multivariate anomaly detection.

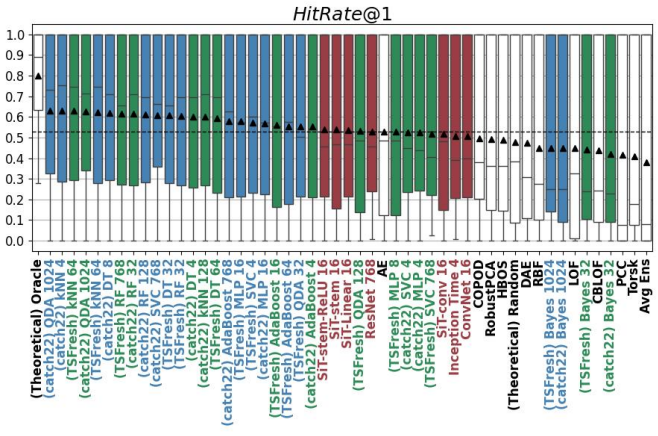


Fig. 3: Interpretability ($HitRate@1$). White bars: individual detectors/baselines; blue/green bars: feature-based selectors (raw/normalized); red bars: deep learning. Black triangles: mean interpretability.

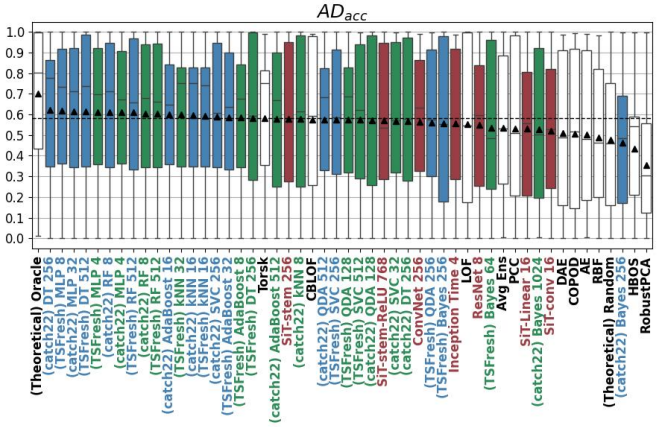


Fig. 4: Anomaly detection accuracy ($AD_{acc}$). Same color coding as Figure 3.

*c) Runtime efficiency.:* Figure 5 summarizes runtime costs. A key motivation for MuMSAD is that brute-force strategies such as the *Averaging Ensemble* require executing all $K$ detectors on each time series, which quickly becomes infeasible. The complexity grows linearly with the number of detectors, $O(KT)$, where $T$ is the length of the time series. In contrast, MuMSAD adds only a lightweight selection step and then runs a single detector, reducing the cost to $O(T)$.

To provide a concrete example, running all $K = 10$ detectors on a single one-day series ($|T| = 8640$) requires on average 3165 seconds CPU-time on our evaluation hardware. Across the 554 days of the OBSEA dataset, this corresponds

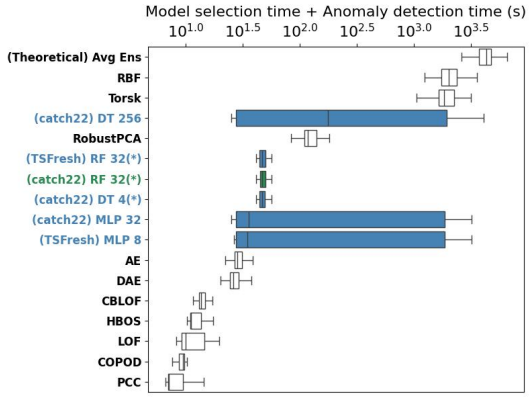


Fig. 5: Runtime efficiency: total time = selection + detection. Asterisks = top-3 selectors for accuracy or interpretability.

to roughly 20 days of computation. By comparison, MuMSAD selectors add only 35–100 seconds for the selection step, plus the runtime of a single chosen detector, bringing the per-series cost down by nearly an order of magnitude. Over the full dataset, this translates into hours rather than days of computation.

## VII. Conclusions and Future Work

We introduced **MuMSAD**, an AutoML framework for automatic selection of interpretable multivariate anomaly detectors. By extending prior AutoML solutions such as MSAD, MuMSAD integrates two advances: (i) support for heterogeneous multivariate series rather than only univariate, and (ii) dimension-level interpretability for root cause analysis.

Our experimental study demonstrated MuMSAD's value in a real-world setting where we built an automatic data quality control module for smart ocean observatories. Here, MuMSAD improved interpretability by up to 20% while simultaneously increasing anomaly detection accuracy by 15% compared to the best single detector, all while maintaining runtime comparable to the fastest interpretable baseline.

**Future work.** Several directions follow from this work. (i) We will improve robustness to out-of-distribution settings by exploring domain-invariant features, meta-learning strategies, and lightweight calibration to adapt selectors to new environments. (ii) We plan to expand the library of interpretable multivariate detectors, in particular incorporating emerging neural methods that support attention mechanisms. (iii) We aim to extend MuMSAD to streaming operation, with incremental feature extraction and rolling model selection suitable for online data pipelines. (iv) Finally, we are preparing deployments in offshore infrastructure integrity monitoring, as requested by our industrial partners, to further validate scalability and operational impact.

**Acknowledgements:** This paper is in memory of Sverre Jakob Alvsvåg, who passed away during the final stage of the work. We acknowledge SFI SmartOcean NFR Project 309612/F40. Work supported by EU Horizon projects AI4Europe (101070000), TwinODIS (101160009), DataGEMS (101188416) and RECITALS (101168490), and by $Y\Pi AI\Theta A$ & NextGenerationEU project HARSH ($Y\Pi 3TA-0560901$) that is carried out within the framework of the National Recovery and Resilience Plan "Greece 2.0" with funding from the European Union – NextGenerationEU. This work was granted access to the HPC resources of IDRIS under the allocation 2025-A0191012641 made by GENCI.

## References

[1] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei, "Robust anomaly detection for multivariate time series through stochastic recurrent neural network," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2828–2837.

[2] Z. Li, Y. Zhao, J. Han, Y. Su, R. Jiao, X. Wen, and D. Pei, "Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding," in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2021, pp. 3220–3230.

[3] X. Chen, L. Deng, F. Huang, C. Zhang, Z. Zhang, Y. Zhao, and K. Zheng, "Daemon: Unsupervised anomaly detection and interpretation for multivariate time series," in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 2021, pp. 2225–2230.

[4] E. Sylligardos, P. Boniol, J. Paparrizos, P. Trahanias, and T. Palpanas, "Choose wisely: An extensive evaluation of model selection for anomaly detection in time series," *Proceedings of the VLDB Endowment*, vol. 16, no. 11, pp. 3418–3432, 2023.

[5] S. Schmidl, P. Wenig, and T. Papenbrock, "Anomaly detection in time series: a comprehensive evaluation," *Proceedings of the VLDB Endowment*, vol. 15, no. 9, pp. 1779–1797, 2022.

[6] A. M. Skålvik, R. N. Bjørk, E. Martinez, K.-E. Frøysa, and C. Saetre, "Multivariate, automatic diagnostics based on insights into sensor technology," *Journal of Marine Science and Engineering*, vol. 12, no. 12, 2024.

[7] N.-T. Nguyen, K. Lima, A. M. Skålvik, R. Heldal, E. Knauss, T. D. Oyetoyan, P. Pelliccione, and C. Sætre, "Synthesized data quality requirements and roadmap for improving reusability of in-situ marine data," in *2023 IEEE 31st International Requirements Engineering Conference (RE)*. IEEE, 2023, pp. 65–76.

[8] OECD, *The ocean economy in 2030*. OECD, 2016.

[9] N.-T. Nguyen, R. Heldal, K. Lima, T. D. Oyetoyan, P. Pelliccione, L. M. Kristensen, K. W. Hoydal, P. A. Reiersgaard, and Y. Kvinnsland, "Engineering Challenges of Stationary Wireless Smart Ocean Observation Systems," *IEEE Internet of Things Journal*, 2023.

[10] X. Wang, J. Lin, N. Patel, and M. Braun, "Exact variable-length anomaly detection algorithm for univariate and multivariate time series," *Data Mining and Knowledge Discovery*, vol. 32, pp. 1806–1844, 2018.

[11] N.-T. Nguyen, R. Heldal, and P. Pelliccione, "Concept-drift-adaptive anomaly detector for marine sensor data streams," *Internet of Things*, p. 101414, 2024.

[12] P. Wenig, S. Schmidl, and T. Papenbrock, "Timeeval: A benchmarking toolkit for time series anomaly detection algorithms," *Proceedings of the VLDB Endowment*, vol. 15, no. 12, pp. 3678–3681, 2022.

[13] S. Chatterjee, R. Bopardikar, M. Guerard, U. Thakore, and X. Jiang, "Mospat: Automl based model selection and parameter tuning for time series anomaly detection," *arXiv preprint arXiv:2205.11755*, 2022.

[14] T. development team, "Tsfresh: Feature extraction api," https://tsfresh.readthedocs.io/en/latest/api/tsfresh.feature_extraction.html, 2021, accessed: 2024-11-14.

[15] C. H. Lubba, S. S. Sethi, P. Knaute, S. R. Schultz, B. D. Fulcher, and N. S. Jones, "catch22: Canonical time-series characteristics: Selected through highly comparative time-series analysis," *Data Mining and Knowledge Discovery*, vol. 33, no. 6, pp. 1821–1852, 2019.

[16] A. Dempster, D. F. Schmidt, and G. I. Webb, "Minirocket: A very fast (almost) deterministic transform for time series classification," in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2021, pp. 248–257.

[17] J. Paparrizos, P. Boniol, Q. Liu, and T. Palpanas, "Advances in time-series anomaly detection: Algorithms, benchmarks, and evaluation measures," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 2025, pp. 6151–6161.