GlassboxAD: An Interactive System for Dissecting Hierarchical Time-Series Anomaly Detection

Mingyi Huang
The Ohio State University
Columbus, Ohio, USA
huang.5749@osu.edu

Qinghua Liu
The Ohio State University
Columbus, Ohio, USA
liu.11085@osu.edu

Paul Boniol
Inria, ENS, CNRS, PSL
Paris, France
paul.boniol@inria.fr

John Paparrizos
The Ohio State University
and Aristotle University of
Thessaloniki
paparrizos.1@osu.edu

Abstract

Time-series anomaly detection (TSAD) is challenging in unsupervised settings because anomalies are heterogeneous and often manifest at different temporal scales. HYDRA addresses this by constructing a multi-level hierarchy of representative subsequences, computing reference-driven anomaly scores at each level, and aggregating them into a final score. In this paper, we present GlassboxAD, an interactive demo system that makes HYDRA's hierarchical behavior observable and interpretable. The system supports two entry points: instant browsing of benchmark results on TSB-AD and on-demand analysis of user-uploaded time series. Users can inspect how detections emerge across levels and assess robustness under different tolerance and threshold settings through linked views: layered visualizations of representatives, per-layer evidence for selected subsequences, score-layer switching on the timeline, and specific evaluation metrics for TSAD. We further contextualize these interactive diagnoses with benchmark-driven comparisons, showing that HYDRA attains a top overall rank on TSB-AD while enabling users to connect case-level, hierarchy-based explanations with suite-level trends across heterogeneous anomaly regimes. The webpage is available online: https://glassboxad.streamlit.app/

CCS Concepts

• **Computing methodologies** → **Anomaly detection**.

Keywords

Time-Series Analysis, Unsupervised Learning, Ensemble Learning

ACM Reference Format:
Mingyi Huang, Qinghua Liu, Paul Boniol, and John Paparrizos. 2026. GlassboxAD: An Interactive System for Dissecting Hierarchical Time-Series Anomaly Detection. In *Companion of the International Conference on Management of Data (SIGMOD Companion '26), May 31-June 05, 2026, Bengaluru, India.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3788853.3801594

1 Introduction

Time-series anomaly detection underpins a wide range of monitoring and diagnosis tasks, such as those arising in sensing, healthcare, financial analytics, and IoT systems [1, 4, 10, 11]. A key difficulty is that anomalies are rarely homogeneous: they can be short and isolated, persistent and slowly evolving, or appear as subtle deviations that are easily masked by higher-amplitude irregularities. Figure 1 illustrates how such heterogeneity undermines common unsupervised paradigms: In discord-based methods, repeated near-duplicate anomalies may become each other's nearest neighbors, suppressing their anomaly scores. In clustering-based methods, concentrated anomalies may be absorbed into the same cluster, and misspecified cluster numbers may lead to contaminated centroids that no longer represent normal subsequences well. Recent benchmark studies have shown that the best-performing models can differ substantially under different anomaly assumptions [7, 8]. In unsupervised settings, the anomaly type is typically unknown in advance, making single-resolution detectors (e.g., KMeans [9, 13], MatrixProfile [14], and NormA [3]) brittle when the anomaly scale shifts across datasets, or even within the same stream.

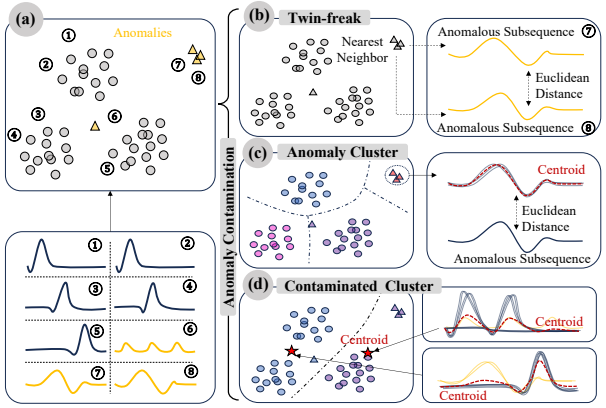

Figure 1: (a) 2D subsequence embedding contrasting normal subsequences with anomalies; (b)–(d) three anomaly-contamination patterns that violate model assumptions.

HYDRA [6] is motivated by the multi-scale nature of subsequence anomalies. Instead of committing to one resolution, HYDRA constructs a hierarchy of representative subsequences and integrates evidence across levels in a unified, unsupervised manner. Concretely, starting from a windowed collection, HYDRA iteratively (i) forms conservative, redundancy-aware groups of windows, (ii) selects one representative per group to create a coarser reference set, (iii) computes reference-driven (distance-based) anomaly scores at each level, and (iv) aggregates the resulting level-wise scores to capture anomalies across multiple temporal resolutions. Crucially, relying on any single level is suboptimal because the optimal resolution is unknown *a priori* and a single representative set can miss parts of the anomaly spectrum (e.g., large deviations may dominate distance-based scores and suppress subtler changes); HYDRA

SIGMOD Companion '26, Bengaluru, India

ACM ISBN 979-8-4007-2450-3/2026/05
https://doi.org/10.1145/3788853.3801594

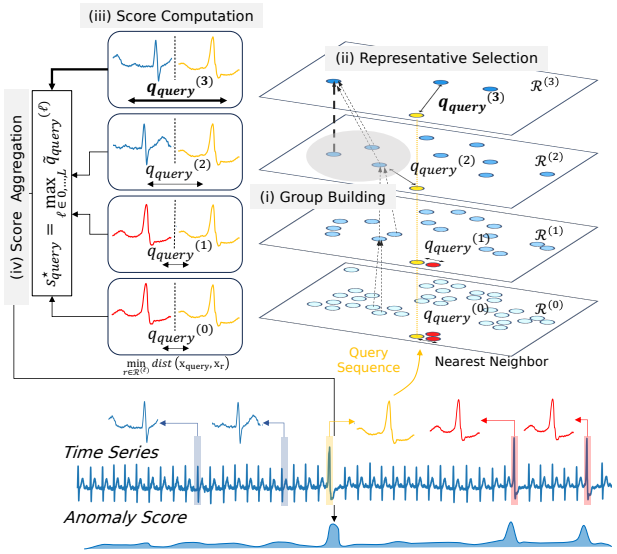

Figure 2: Overview of HYDRA. HYDRA builds a multi-level representative hierarchy, scores windows against level-wise references, and robustly aggregates scores across resolutions.

therefore leverages the hierarchy to combine fine-resolution sensitivity to localized, low-magnitude deviations with coarse-resolution emphasis on persistent irregularities via cross-level integration.

While HYDRA's hierarchical design is central to its effectiveness, it also introduces internal dynamics: grouping, representative selection, level-wise scoring, and aggregation. Those components are difficult to convey through static figures alone. A demonstration system is therefore particularly important for HYDRA: it can make the hierarchy observable and allow users to inspect how (1) the reference set shrinks across levels, (2) anomalous representatives are progressively eliminated as the hierarchy becomes "cleaner", and (3) different levels contribute complementary evidence to the final decision. This aligns with prior anomaly-detection demos that emphasize not only effectiveness, but also interpretability of intermediate steps via visualization and user-driven exploration.

In this work, we present **GlassboxAD**, an interactive demo that makes HYDRA's hierarchical behavior observable and actionable. Beyond showing a final anomaly curve, the system helps users answer *why* a detection happens and *which resolution* is responsible: users can (i) trace how representative sets contract and become progressively "cleaner" across levels, (ii) contrast localized versus persistent events to see how fine and coarse resolutions contribute complementary evidence, and (iii) diagnose failure modes that commonly confuse single-resolution detectors (e.g., clustered/duplicated anomalies and contamination-driven prototype drift). GlassboxADcouples these insights with range-aware evaluation under varying tolerance and threshold settings, and with benchmark-driven summaries (metrics and runtime) to contextualize HYDRA's behavior across diverse anomaly regimes.

We next summarize HYDRA at a level sufficient to understand the demo (Section 2), and then describe the system architecture and interaction workflow of GlassboxAD (Section 3 and 4). We conclude with a compact benchmark snapshot that contextualizes the demo with baselines and evaluation protocols.

2 HYDRA Approach

We summarize HYDRA only at the level needed to understand the demo interface; full algorithmic details and complete experiments are provided in the full HYDRA paper [6].

Problem setup. Given a (uni- or multi-variate) time series $\mathbf{y}_{1:T}$, we extract an overlapping window sequence $\{\mathbf{x}_i\}_{i=1}^{N}$ using a sliding window of length m (stride $s = 1$). The task is subsequence anomaly detection: compute an anomaly score for each window and map it back to a time-indexed score curve, while accounting for the fact that anomalous behavior may manifest at different temporal scales.

A Hierarchical Time Series Anomaly Detection Approach. Our method is motivated by a reference-driven view: if we can construct a compact set of representative windows that capture *typical* behavior, then deviations from this reference can be scored as anomalies. The main challenge is that the window collection can be contaminated, and repeated/clustered anomalies may suppress one another under nearest-neighbor scoring. We address this via a conservative hierarchy that progressively filters anomalous representatives under mild rarity and majority-consistency assumptions.

Key assumption. Let each window have a latent label $a_i \in \{0, 1\}$ indicating normal (0) or anomalous (1), and define

$$\mathcal{N} = \{\, \mathbf{x}_i : \; a_i = 0 \,\}, \qquad \mathcal{A} = \{\, \mathbf{x}_i : \; a_i = 1 \,\}.$$

We assume *anomaly rarity*:

$$\frac{|\mathcal{A}|}{|\mathcal{A}| + |\mathcal{N}|} \leq \eta, \qquad \text{for a small } \eta \leq 0.5. \tag{1}$$

When partitioning the current candidate set into groups $\{C\}$, we assume a *majority-consistent* representative rule:

$$\text{if } |C \cap \mathcal{N}| > |C \cap \mathcal{A}| \;\text{ then }\; \text{Select}(C) \in C \cap \mathcal{N}. \tag{2}$$

This implies that anomalies contained in normal-dominated groups do not propagate to the next level. While anomaly-dominated groups may still pass anomalous representatives, rarity ensures that the anomalous portion of the candidate set cannot grow across levels; consequently, at some finite level the remaining candidates form an all-normal representative set that serves as an uncontaminated reference for scoring.

2.1 HYDRA Framework

Figure 2 shows how HYDRA detects anomalies by combining multi-level evidence from progressively reduced reference sets.

Group Building. At level ℓ, HYDRA forms conservative, redundancy-aware groups over the current index set $\mathcal{I}_\ell$ using a self-excluding nearest-neighbor graph and majority-favoring unions. Intuitively, highly redundant windows (often due to overlap or repeated local patterns) are merged into predominantly-normal groups under the anomaly rarity assumption, which attenuates anomalous influence before reference selection.

Representative Selection. From each group, HYDRA selects exactly one representative to form the next-level reference set $\mathcal{R}^{(\ell+1)}$ and updates $\mathcal{I}_{\ell+1} \leftarrow \mathcal{R}^{(\ell+1)}$. Repeating this procedure yields a nested hierarchy of reference sets $\mathcal{R}^{(0)} \supset \mathcal{R}^{(1)} \supset \cdots \supset \mathcal{R}^{(L)}$, whose size gradually shrinks across levels while still retaining coverage of prevalent structure; higher levels contain fewer, more stable representatives that summarize dominant patterns, while lower levels preserve finer-grained, local variations.

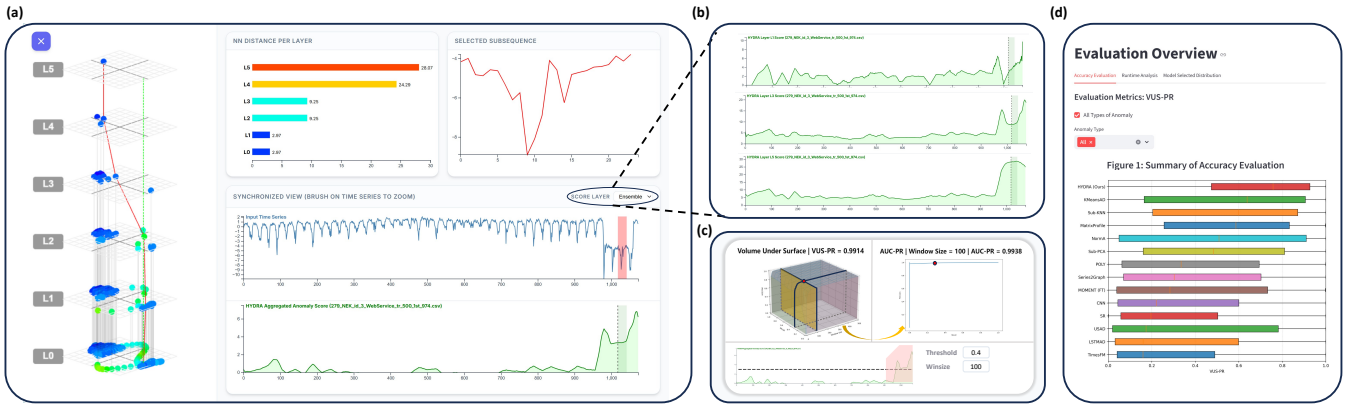

Figure 3: System overview of GlassboxAD: (a) the main interactive visualization that reveals hierarchical representative selection and window reduction; (b) layer-wise views that allow users to switch across levels and inspect per-layer anomaly scores; (c) evaluation with VUS-PR, visualizing the *volume under surface* across decision thresholds; and (d) an overview comparison against top-performing baselines on the benchmark.

Per-level Score Computation. Given the level-ℓ representatives $\mathcal{R}^{(\ell)}$, each window receives a distance-to-reference score:

$$q_i^{(\ell)} = \min_{r \in \mathcal{R}^{(\ell)}} d(\mathbf{x}_i, \mathbf{x}_r), \qquad i = 1, \ldots, N, \tag{3}$$

where $d(\cdot, \cdot)$ is a window distance (e.g., Euclidean on raw-scale windows). This reference-driven design avoids direct comparison among near-duplicate anomalous windows and reduces sensitivity to global prototype choices.

Score Aggregation. To balance score-scale disparities across levels, HYDRA standardizes scores within each level and aggregates across levels via a max rule:

$$s_i^\star = \max_{\ell=0,\ldots,L} \frac{q_i^{(\ell)} - \mu_\ell}{\sigma_\ell}, \mu_\ell = \frac{1}{N}\sum_{i=1}^{N} q_i^{(\ell)}, \ \sigma_\ell^2 = \frac{1}{N}\sum_{i=1}^{N}\left(q_i^{(\ell)} - \mu_\ell\right)^2, \tag{4}$$

where $\delta > 0$ is a small constant for numerical stability. The fused window scores $\{s_i^\star\}$ can then be mapped back to a time-indexed anomaly curve by aggregating the scores of windows that cover each timestamp (e.g., via overlap averaging).

These steps produce three key artifacts that our demo exposes end-to-end: (i) level-wise representative sets, (ii) per-level window/series scores, and (iii) an aggregated score curve. We next describe how GlassboxAD links these artifacts into coordinated views and an interactive workflow.

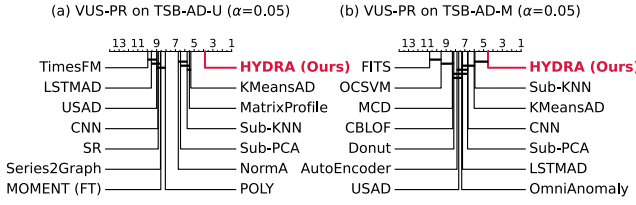

Figure 4: Critical Difference Diagrams under (a) univariate, (b) multivariate settings.

Experimental Results. Figure 4 depicts the critical difference diagrams computed using a Wilcoxon paired signed-rank test over the TSB-AD benchmark [8]. Our baseline selection is guided by TSB-AD: we prioritize top-tier methods on the benchmark and include representative detectors spanning different anomaly assumptions and paradigms. The results show that HYDRA attains the top average rank and outperforms these strong baselines.

3 GlassboxAD: System Overview

Building on the hierarchy and scores produced by HYDRA, GlassboxAD is a stand-alone web application that exposes HYDRA as an end-to-end pipeline with two usage modes: **(i) benchmark browsing mode** for TSB-AD and **(ii) on-demand mode** for user-uploaded series. In both modes, the system produces the same set of artifacts—a hierarchy of representatives, per-level anomaly scores, an aggregated score, and evaluation summaries—and routes them to a set of linked visual views.

Data and execution modes. In the *benchmark mode*, the system offers a library of TSB-AD case studies with corresponding HYDRA outputs for interactive exploration: (1) the full hierarchy (representative sets and metadata per level), (2) per-level and aggregated anomaly score curves, (3) PCA projections for each level, and (4) evaluation outputs (VUS-PR surfaces/slices and benchmark-level summaries). This enables low-latency browsing across a large number of datasets. In the *on-demand mode*, users can upload their own time series and the system will compute the same artifacts.

Core computation and artifacts. Given a time series, the backend (1) constructs hierarchical groups and selects one representative per group to form level-wise reference sets; (2) computes reference-driven distance scores for all windows at each level; and (3) normalizes and aggregates level-wise scores into a final anomaly score curve. These artifacts are stored with consistent identifiers so that a selected window/subsequence can be traced across levels and linked to its scores. Each window is assigned a persistent identifier that is shared across views, so that a selection in the PCA panel immediately retrieves (i) the raw subsequence, (ii) its per-layer nearest-representative distances, and (iii) the corresponding timestamps on the score curve.

Visualization and evaluation modules. The front-end consumes these artifacts through four modules: a *hierarchy module* that renders per-level PCA embeddings with representative highlighting; a

scoring module that exposes level-wise and aggregated scores and supports querying the per-level evidence of a selected subsequence; an *metrics module* that visualizes VUS-PR and its slices under user-chosen tolerance/window settings; and a *benchmark module* that summarizes HYDRA's comparative performance against top-tier baselines on univariate and multivariate suites. Figure 3 provides a schematic view of how these modules are organized in the UI.

In the next section, we describe how users navigate these modules in a small set of task-oriented scenarios.

4 Demonstration Scenarios

The demo supports two entry points: users can either select a time series from TSB-AD in *benchmark mode* or upload one or more series in *on-demand mode*. After loading a series, users follow a guided workflow through the hierarchy, scoring, and evaluation views, organized into four scenarios in Figure 3.

Scenario 1: Overview. Users begin by either selecting a time series from a benchmark suite for immediate exploration or uploading their own time series for on-demand analysis. They then enter the hierarchical view, where a stacked PCA embedding is rendered for each level with representative subsequences explicitly exposed (Fig. 3a). Clicking any point selects a concrete subsequence; the selection is linked across levels so users can trace how representative structure evolves and how the candidate/reference set contracts as the hierarchy coarsens. For the selected subsequence, the system displays its nearest-representative distance per layer as a compact bar chart alongside the raw subsequence shape. Finally, users switch the displayed anomaly curve between individual layers (Layer_0–Layer_l) and the ensemble score to see which resolutions contribute the strongest evidence for the highlighted regions (Fig. 3b). For interactive use, the system adopts HYDRA_approx [6], which is substantially faster on long series without sacrificing accuracy.

Scenario 2: Resolution Contrast (event-level diagnosis). This scenario emphasizes that different anomaly regimes are captured at different levels. Users select two regions on the timeline (e.g., a localized deviation versus a persistent irregularity) and, for each event, repeat the same linked inspection: locate representative candidates in the layered PCA views, select a subsequence, and compare its per-layer distance profile (Fig. 3a). By toggling the displayed score layer on the timeline, users can directly observe how fine levels and coarse levels emphasize different regimes and how the ensemble score preserves evidence visible at any resolution (Fig. 3b).

Scenario 3: VUS-PR Evaluation. Users switch to the evaluation panel to assess range-aware detection quality. Following the recommendation in TSB-AD [8], we adopt VUS-PR [2] as a robust and practically meaningful metric for anomaly detection. VUS-PR extends AUC-PR by introducing a buffer (tolerance) region around each anomaly interval to credit near-miss detections, and adds this buffer size as a third axis to yield a volume-under-surface evaluation across tolerance levels. The system visualizes this volume and allows users to interactively vary key settings such as the decision threshold and buffer size, then inspect the resulting VUS slices (e.g., the AUC at a fixed window size), making it clear how evaluation changes across tolerance/threshold configurations rather than relying on a single operating point (Fig. 3c).

Scenario 4: Benchmark Summary. Finally, users open the benchmark tab to contextualize the above interactions at the dataset-suite level. The panel summarizes HYDRA's performance against strong baselines on a large-scale benchmark suite in both univariate and multivariate settings (Fig. 3d). Beyond simply showing numbers, it supports (i) runtime analysis across the benchmark; (ii) multi-metric comparisons across range-aware, point-wise, and event-based protocols [5, 8, 12]; and (iii) stratification by anomaly types, enabling users to examine trade-offs across heterogeneous regimes and connect case-level explanations to suite-level trends.

5 Conclusion

We presented GlassboxAD, a stand-alone interactive demo system for exploring HYDRA's multi-level anomaly detection. By coupling layered representative visualizations, per-layer scoring evidence, and range-aware VUS-PR evaluation, the demo helps users inspect how different resolutions contribute to detections and how performance changes across tolerance and threshold settings. The system supports both benchmark browsing and on-demand user uploads, while also providing benchmark summaries with metric- and runtime-based comparisons across diverse anomaly types.

References

[1] Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A Lozano. 2021. A Review on outlier/Anomaly Detection in Time Series Data. *ACM Computing Surveys (CSUR)* 54, 3 (2021), 1–33.

[2] Paul Boniol, Ashwin K Krishna, Marine Bruel, Qinghua Liu, Mingyi Huang, Themis Palpanas, Ruey S Tsay, Aaron Elmore, Michael J Franklin, and John Paparrizos. 2025. VUS: effective and efficient accuracy measures for time-series anomaly detection. *The VLDB Journal* 34, 3 (2025), 32.

[3] Paul Boniol, Michele Linardi, Federico Roncallo, Themis Palpanas, Mohammed Meftah, and Emmanuel Remy. 2021. Unsupervised and scalable subsequence anomaly detection in large data series. *The VLDB Journal* (2021), 1–23.

[4] S. Chauhan and L. Vig. 2015. Anomaly Detection in ECG Time Signals via Deep Long Short-Term Memory Networks. In *Proceedings of the International Conference on Data Science and Advanced Analytics (DSAA)*. 1–7. doi:10.1109/DSAA.2015.7344872

[5] Astha Garg, Wenyu Zhang, Jules Samaran, Ramasamy Savitha, and Chuan-Sheng Foo. 2022. An Evaluation of Anomaly Detection and Diagnosis in Multivariate Time Series. *IEEE Transactions on Neural Networks and Learning Systems* 33, 6 (June 2022), 2508–2517. doi:10.1109/tnnls.2021.3105827

[6] Mingyi Huang, Qinghua Liu, Paul Boniol, and John Paparrizos. 2026. HYDRA: A Multi-Level Hierarchy-Driven Approach for Robust Anomaly Detection in Time Series. *Proceedings of the ACM on Management of Data* 4, 3, Article 197 (June 2026), 30 pages. doi:10.1145/3802074

[7] Qinghua Liu, Seunghak Lee, and John Paparrizos. 2025. Tsb-autoad: Towards automated solutions for time-series anomaly detection. *Proceedings of the VLDB Endowment* 18, 11 (2025), 4364–4379.

[8] Qinghua Liu and John Paparrizos. 2024. The Elephant in the Room: Towards A Reliable Time-Series Anomaly Detection Benchmark. *Advances in Neural Information Processing Systems* 37 (2024), 108231–108261.

[9] Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE transactions on information theory* 28, 2 (1982), 129–137.

[10] John Paparrizos, Paul Boniol, Qinghua Liu, and Themis Palpanas. 2025. 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*. 6151–6161.

[11] Peter Rousseeuw, Domenico Perrotta, Marco Riani, and Mia Hubert. 2019. Robust monitoring of time series with application to fraud detection. *Econometrics and statistics* 9 (2019), 108–121.

[12] Nesime Tatbul, Tae Jun Lee, Stan Zdonik, Mejbah Alam, and Justin Gottschlich. 2018. Precision and Recall for Time Series. In *Advances in Neural Information Processing Systems*, Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/hash/8f468c873a32bb0619eaeb2050ba45d1-Abstract.html

[13] Takehisa Yairi, Yoshikiyo Kato, and Koichi Hori. 2001. Fault detection by mining association rules from house-keeping data. In *proceedings of the 6th International Symposium on Artificial Intelligence, Robotics and Automation in Space*, Vol. 18. Citeseer, 21.

[14] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. 2016. Matrix profile I: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. In *2016 IEEE 16th international conference on data mining (ICDM)*. Ieee, 1317–1322.