Retrieval Over Training
Similarity search-based Model Selection for Time Series Anomaly Detection

Christos Panourgias[1,*], Roberto Stanzione[2,*], Adrien Petralia[1], Themis Palpanas[1], Paul Boniol[2]

[1]Université Paris Cité, LIPADE [2]École Normale Supérieure, Inria

{panourgiaschris,adrien.petralia}@gmail.com {roberto.stanzione,paul.boniol}@inria.fr
themis@mi.parisdescartes.fr

[*]Equal Contribution

Abstract

Anomaly detection is a fundamental task for time-series analytics, with important implications for the downstream performance of many applications. Despite the large number of anomaly detection methods proposed in the literature, recent benchmark studies have shown that no single detector performs best across highly heterogeneous time series. Therefore, a practical and scalable solution is to develop a model-selection method that, for a given time series, selects the anomaly detector most likely to perform well. Nevertheless, the model selection approaches proposed in the literature suffer from a significant drop in accuracy when applied in Out-of-Distribution (OOD) settings. In this paper, we tackle the aforementioned limitation and propose **RAMSAD**, a retrieval-based framework for model selection in time-series anomaly detection. Given a new time series, our method queries a knowledge base of previously observed instances, retrieves the top-k most similar series, and transfers detector recommendations from their performance profiles. The framework can operate with standard similarity measures as well as embedding-based representations. Our results show that RAMSAD outperforms existing baselines in In-Distribution (ID) settings and is the most robust selector in OOD, where, when selecting a small subset of detectors, it matches or exceeds ensembles that run the entire detector pool, at a fraction of their cost. Overall, we demonstrate that similarity-based retrieval constitutes a strong and efficient foundation for model selection in time-series anomaly detection.

1 Introduction

Time-series anomaly detection is a fundamental problem in machine learning, with applications in healthcare, manufacturing, finance, energy, cybersecurity, and scientific monitoring. Its goal is to identify observations or subsequences that deviate from expected temporal behavior. Because the utility of downstream monitoring systems depends directly on the quality of anomaly detection, selecting an appropriate detector is a critical practical challenge [11, 54].

Despite substantial progress in the literature, time-series anomaly detection remains far from solved. A central difficulty is the extreme heterogeneity of time series encountered in practice. Real-world datasets differ widely in sampling rate, scale, length, noise characteristics, seasonality, dimensionality, and anomaly type. Recent benchmark studies have reinforced this observation, showing that there is no universally best anomaly detector across heterogeneous collections of datasets [46, 34]. This variability makes *model selection* a key problem in its own right.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Existing work has increasingly approached model selection with trainable strategies [58, 59]. These methods learn, from benchmark results or dataset representations, which detector to apply to a new series. While appealing, they add another layer of learning on top of an already difficult task, requiring additional supervision, tuning, and assumptions about generalization across domains. This complexity is not always justified, especially in heterogeneous settings where benchmark coverage is limited and domain shift is common. We therefore ask a simpler question: *can anomaly-detector selection be solved without training a selector at all?*

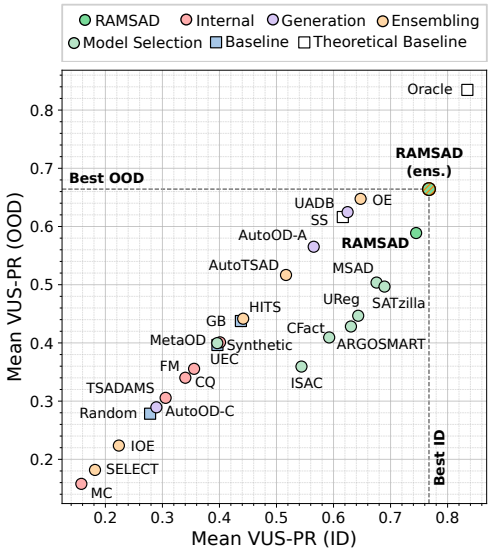


Figure 1: Comparison of RAMSAD with all automatic TSAD approaches (ID and OOD).

In this paper, we answer this question affirmatively. We propose **RAMSAD** (**R**etrieval-**A**ugmented **M**odel **S**election for **A**nomaly **D**etection), a retrieval-based framework for time-series anomaly detection based on a training-free, non-parametric selector, i.e., it does not require fitting a separate meta-learner; instead, it ranks candidate detectors directly from the performance of detectors on retrieved similar time series. The idea is simple: similar time series are likely to favor the same anomaly detector. We maintain a knowledge base of previously observed time series, together with the performance profiles of anomaly detectors when applied to these series; as for trained selectors, computing these profiles requires labels for the knowledge-base series, but never for the query. Given a new query series, we retrieve its top-k nearest neighbors and use their outcomes to rank candidate detectors. This turns model selection into a similarity-search problem, eliminating the need for a dedicated trained selector.

Our framework is flexible with respect to the representation used for retrieval. In particular, it supports multimodal embeddings obtained by contrastive alignment of time series with textual dataset descriptions. We evaluate the proposed approach in both in-distribution (ID) and out-of-distribution (OOD) settings. In addition, as illustrated in Figure 1, selecting an ensemble of $N = 10$ detectors improves robustness and achieves SOTA in ID and OOD performance while remaining substantially faster than competing train-dependent selectors. Training-free refers to the selector only: (i) like any supervised selector, RAMSAD needs a labeled knowledge base, but no labels for the query; (ii) the optional alignment module is trained offline without detector identities or performance, and queries need no description; and (iii) a frozen, off-the-shelf TSFM already outperforms all trained selectors. Our main contributions are as follows:

• We introduce a retrieval-based formulation of anomaly-detector selection for time series, casting the problem as non-parametric inference over a knowledge base of prior instances.

• We introduce a *segment-to-series* aggregation pipeline where we effectively translate segment-level similarity matches into a robust, series-level ranking.

• We propose RAMSAD[1] for both classical similarity measures and learned embedding spaces, including multimodal time-series–text representations. We experimentally show that aligning TSFM embeddings with simple dataset descriptions provides a useful semantic signal that mainly improves retrieval in OOD.

• We empirically demonstrate that RAMSAD is a strong alternative to trainable selectors, delivering higher accuracy in ID and greater robustness in OOD, while incurring substantially lower cost.

2 Foundations and Related Work

Notations. A dataset is denoted by $\mathfrak{D} = \{T_1, \ldots, T_{|\mathfrak{D}|}\}$, where each T_i is a univariate time series. Specifically, the i-th series is written as $T_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,L_i}] \in \mathbb{R}^{L_i}$, where $L_i = |T_i|$ is the length of T_i, and $x_{i,t} \in \mathbb{R}$ denotes the value of T_i at time index t. A subsequence of length ℓ

[1]Code available at: `https://github.com/Hendrix8/RAMSAD`

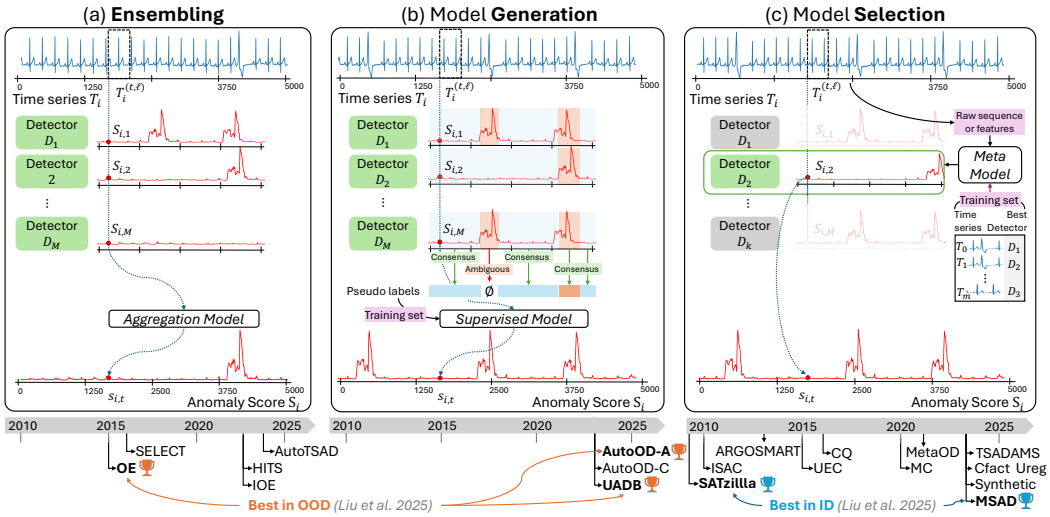


Figure 2: Automatic solutions for time series anomaly detection discussed in Section 2.

extracted from T_i and starting at position t is denoted by $T_i^{(t,\ell)} = [x_{i,t}, x_{i,t+1}, \ldots, x_{i,t+\ell-1}] \in \mathbb{R}^\ell$. The anomaly score of a detector D applied on T_i is denoted as $S_i = [s_{i,1}, s_{i,2}, \ldots, s_{i,L_i}] \in \mathbb{R}^{L_i}$.

Time-Series Anomaly Detection. The latter can be grouped according to the underlying detection principle [11]. *Distance-based* methods identify anomalous subsequences by comparing them against their nearest neighbors in the series or in a reference set [28, 35, 72, 31]. *Density-based* methods instead detect anomalous behaviors by estimating how isolated observations or subsequences are in a representation space. This family includes *distribution-based* [5], *graph-based* [9, 61], *tree-based* [32, 37], and *encoding-based* [21]. Finally, *prediction-based* methods estimate expected temporal behavior and use prediction error as the anomaly proxy. These approaches can be divided into *forecasting-based* [64] and *reconstruction-based* [57]. A detailed review is provided in Appendix A.

2.1 Automatic Solutions

Recent benchmark studies for time-series anomaly detection [54, 33] have motivated automated strategies. Existing approaches can be grouped into three main categories illustrated in Figure 2: *model selection*, *model ensembling*, and *model generation* [34].

Model Selection. Model selection methods aim to choose the best detector for a given time series. A large subset of this literature follows the *meta-learning* paradigm, where historical knowledge about detector performance is used to guide future choices. Simple meta-learners such as `ARGOSMART` [43] and `ISAC` [27] rely on similarity search over meta-features. Other approaches formulate selection through, (i) collaborative filtering [73], (ii) as a classification problem [58, 59], or (iii) as a regression problem that predicts expected detector performance scores [68, 41]. A different line of work, denoted as *Internal* in [34], estimates detector quality directly on the target series without labeled historical data, using (i) unsupervised surrogate metrics computed from anomaly scores [22], (ii) clustering-based internal validation criteria [42], (iii) synthetic anomaly injection [15], or (iv) the Model Centrality [30]. Building on multiple surrogate criteria, `TSADAMS` [24] aggregates detector rankings to obtain more robust recommendations.

Model Ensembling. Model ensembling combines the outputs of multiple detectors in order to improve robustness and detection quality. Outlier Ensemble (OE) [2] aggregates detector scores using simple operators such as the average, the maximum, or the average of maxima. More selective variants first identify a subset of reliable detectors before aggregation [50, 38]. Finally, `AutoTSAD` [55] provides a fully automated pipeline that generates synthetic data, optimizes detector configurations, and ensembles the most effective detectors into a final anomaly score.

Model Generation. Model generation typically relies on pseudo-labels that identify likely inliers and candidate outliers, and uses them to guide the combination of multiple detectors. For example, `AutoOD-A` [14] starts from a small set of pseudo-labels and iteratively expands it while pruning less effective detectors. Conversely, `AutoOD-C` [14] starts from a larger pseudo-labeled set and

progressively removes labels. Finally, UADB [71] combines information from multiple detectors into a single booster model that iteratively refines anomaly predictions.

2.1.1 Limitations of Current Strategies: *The scalability and OOD trade-off*

Existing automatic solutions for time-series anomaly detection involve a trade-off between robustness, flexibility, and computational efficiency. *Model ensembling* and *model generation* can improve robustness by combining information from multiple detectors, but they typically require running a large part—or the entirety—of the detector pool before producing a final anomaly score, **which can be prohibitively expensive for long time series or large detector collections (L1).** *Model selection* addresses this issue more directly by aiming to choose a single detector for each series, but most existing approaches rely on train-dependent classification or regression models. As a result, they are often tied to a fixed detector pool, **require retraining when the pool changes (L2)**, and may **generalize poorly under domain shift (L3)**. These limitations motivate the search for a simpler and more flexible alternative without requiring a dedicated training phase for model selection.

2.2 Problem Definition

Let $\mathcal{B} = \{D_1, \ldots, D_M\}$ be a finite pool of candidate anomaly detectors, and let $\mathfrak{D} = \{T_1, \ldots, T_{|\mathfrak{D}|}\}$ be a collection of time series. For a given series $T_i \in \mathfrak{D}$ with ground-truth anomaly labels Y_i, each detector $D_j \in \mathcal{B}$ produces an anomaly score sequence $S_{i,j} = D_j(T_i)$, whose quality is evaluated through a performance measure $\mathrm{Acc}(S_{i,j}, Y_i)$ (e.g., AUC-PR, F1, or VUS). The *model selection* problem is to construct a selector $\Phi_K : \mathfrak{D} \to 2^{\mathcal{B}}$, which maps each time series T_i to a subset of K detectors, $\Phi_K(T_i) \subseteq \mathcal{B}$. The case $K = 1$ recovers standard single-detector selection, while $K > 1$ corresponds to selecting an ensemble. Let $g(\Phi_K(T_i), T_i)$ denote the score sequence obtained by applying the selected detector set to T_i, using the ensemble aggregation rule when $K > 1$. The goal is for this selector to match, as closely as possible, the *oracle* choice $\Phi_K^\star(T_i) = \arg\max_{\mathcal{S}\subseteq\mathcal{B},\,|\mathcal{S}|=K} \mathrm{Acc}\big(g(\mathcal{S}, T_i), Y_i\big)$. Equivalently, model selection can be formulated as minimizing the expected regret: $\mathbb{E}_{(T_i,Y_i)}\big[\mathrm{Acc}\big(g(\Phi_K^\star(T_i), T_i), Y_i\big) - \mathrm{Acc}\big(g(\Phi_K(T_i), T_i), Y_i\big)\big]$. Under this formulation, the objective is to predict which detector or detector subset from $\mathcal{B}$ to select for a given series.

In this paper, we aim to address effectively model selection (addressing **L1**) using a framework based on a knowledge base and similarity search, and thus using a train-free selector (addressing **L2**). The main challenge remaining is on the definition of the similarity space and the retrieval process that minimize the effect of domain drift (i.e., OOD) on accuracy (addressing **L3**).

Scope. Throughout this paper, we consider static univariate time series, and OOD refers to *domain shift*: the query series comes from a domain that is entirely absent from the knowledge base. Temporal concept drift within a domain, streaming settings, and non-stationarity within a single long series (e.g., an anomalous pattern becoming normal over time) are outside the scope of our evaluation; we discuss them in Appendix F.1.

3 RAMSAD: Proposed Approach

To address the limitations mentioned above, we propose **RAMSAD**, a retrieval-based framework for selecting anomaly detectors for time series (cf. Figure 3). Our approach relies on the following two core principles.

Similarity Space & Alignment. From a time series database, we first partition each series into fixed-length segments and pair every segment with the textual description of its source dataset. We then encode the time-series segments with a TSFM to capture temporal structure, and encode the dataset descriptions with a sentence transformer to capture domain-level semantic information. While we demonstrate in Section 4 that our TSFM-based embedding is sufficient for in-distribution, the latter faces issues in OOD (i.e., Limitation **L3**). To cope with **L3**, we align the two modalities (i.e., text and time series) with lightweight MLP module that learns a shared vector space in which temporally and semantically related items are mapped close to one another. The aligned segment representations are then stored in a knowledge base together with the known accuracy scores across the candidate detectors. This semantically informed space is particularly useful in anomaly detection, where the interpretation of anomalous behavior depends strongly on context. For example, a sudden spike may indicate a fault in one domain but correspond to normal peak activity in another. By

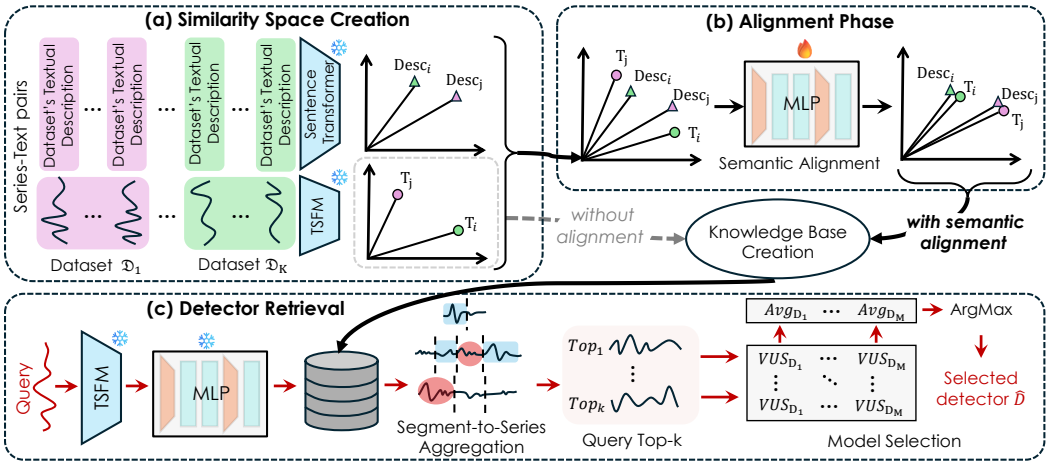


Figure 3: Overview of RAMSAD.

incorporating semantic information, the framework can better capture such differences and thus provide more reliable neighbors for model selection, especially in OOD settings. Alignment is an optional, offline refinement: the textual descriptions only shape the time-series space during training, so neither the knowledge base nor the query requires a description at inference time.

Detector Retrieval. At inference time, a query series is processed through the same segmentation and encoding pipeline and projected into the shared space using only the time-series projection head. Retrieval is first performed at the segment level, where the nearest neighbors of the query segments are identified. They are then aggregated into the top-k retrieved series. For these retrieved series, we access the precomputed VUS matrix and compute, for each candidate detector, the average VUS over the retrieved neighbors. The detector with the highest average VUS is then selected.

3.1 The Core Retrieval Pipeline

We now describe the core retrieval mechanism of RAMSAD. The same pipeline can operate on the original space, TSFM embeddings, or our aligned representations presented in Section 3.2.

Similarity Space Creation. RAMSAD is agnostic to the underlying similarity space. In the simplest setting, retrieval may be performed directly in the original time-series space, in which case segmentation is not necessary. For instance, Euclidean similarity can be computed on full series after truncating or padding them to a common length. We study such original space and distance alternatives in Appendix C.

Our main instantiation instead relies on embeddings produced by a TSFM [4]. We use TSFMs because they capture temporal structure more effectively than raw pointwise distances and standardize heterogeneous time series into fixed-dimensional representations. To build this embedding space, we partition each series into fixed-length segments of size $\min(L_i, 8192)$, where 8192 is the maximum input size supported by `Chronos2` [4] in one of our instantiations. Each segment is encoded as $z^{\mathrm{ts}}_{i,t} = f_{\mathrm{ts}}(T_i^{(t,\ell)})$, with $z^{\mathrm{ts}}_{i,t} \in \mathbb{R}^d$, where f_{ts} may be instantiated with any TSFM or other time-series embedding backbone. Segmentation is therefore required by the bounded context of the encoder: it maps series of very different lengths to fixed-dimensional vectors that can be compared directly, and the aggregation step rewards candidate series that match the query consistently and accurately. We do not resample series and therefore make no claim of invariance to the sampling rate; segmentation may also discard dependencies longer than the encoder context (Appendix F.1).

From Segment-to-Series Aggregation. Retrieval in RAMSAD is performed at the *segment* level, whereas model selection is defined at the *series* level. We therefore introduce an aggregation step that converts segment-level matches into a ranked list of candidate series. For simplicity, we use k throughout to denote the retrieval size; because different query segments may retrieve segments originating from the same series, the final number of distinct candidate series may be smaller than k.

Let T_q denote a query time series, and let $\{T_q^{(t_u,\ell)}\}_{u=1}^m$ be the m extracted segments. For each segment $T_q^{(t_u,\ell)}$, we retrieve its top-k nearest segments, $\{T_{j_{u,r}}^{(t_{u,r},\ell)}\}_{r=1}^k$, together with

their corresponding distances $d_{u,r} = d(T_q^{(t_u,\ell)}, T_{j_{u,r}}^{(t_{u,r},\ell)})$, for $r \in \{1,\dots,k\}$. Each retrieved segment $T_{j_{u,r}}^{(t_{u,r},\ell)}$ originates from the indexed time series $T_{j_{u,r}} \in \mathfrak{D}$. For each candidate series $T_j \in \mathfrak{D}$, we collect all retrieved distances associated with segments originating from T_j: $\mathcal{A}(T_j) = \{\, d_{u,r} \;:\; j_{u,r} = j \,\}$. We then define a base similarity score, where smaller values indicate stronger similarity, as $\mathrm{base}(T_j) = \frac{1}{|\mathcal{A}(T_j)|} \sum_{d \in \mathcal{A}(T_j)} d$.

A candidate series may be retrieved by many query segments or only by a few isolated ones. To distinguish consistent support from sporadic matches, we define the frequency of a candidate series T_j as the fraction of query segments for which at least one of the top-k retrieved neighbors originates from T_j: $\mathrm{freq}(T_j) = \frac{1}{m}\Big|\{\, u \in \{1,\dots,m\} : \exists r \in \{1,\dots,k\} \text{ s.t. } j_{u,r} = j \,\}\Big|$ We then combine freq with the base similarity score to obtain the final ranking score: $\mathrm{score}(T_j) = \frac{1}{\max(\epsilon, \mathrm{freq}(T_j))} \cdot \mathrm{base}(T_j)$, where $\epsilon > 0$ is a small constant. Candidate series are ranked in ascending order of $\mathrm{score}(T_j)$, and the top-k retrieved series for query T_q are given by $\mathrm{top_k}(T_q) = \arg\min_{T_j \in \mathfrak{D}}^{k} \mathrm{score}(T_j)$.

Averaging distances across retrieved segments reduces sensitivity to isolated mismatches, while the frequency-aware term favors candidate series that receive consistent support across multiple query segments. The intuition of this approach is shown in Figure 4.

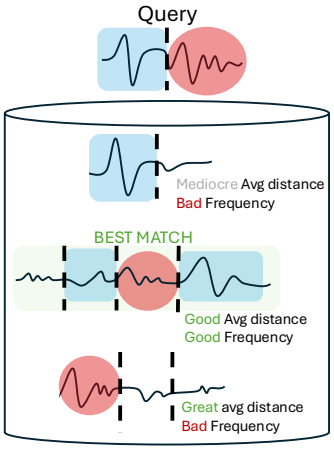


Figure 4: Segment-to-series aggregation in RAMSAD.

Model Selection. Once the top-k most similar series are retrieved, RAMSAD transfers detector recommendations from their stored performance profiles. We assume access to a precomputed performance table $\mathbf{V} \in \mathbb{R}^{|\mathfrak{D}|\times|\mathcal{B}|}$, where $\mathbf{V}(T, D)$ denotes the accuracy (e.g., VUS-PR) of detector $D \in \mathcal{B}$ on series $T \in \mathfrak{D}$. Computing $\mathbf{V}$ requires a labeled knowledge base on which every candidate detector has been evaluated; this is the same supervision consumed by trained meta-learning selectors, whereas no label is needed for the query. For each detector we aggregate its performance over the retrieved neighbors: $\bar{V}_D(T_q) = \frac{1}{|\mathrm{top}_k(T_q)|} \sum_{T \in \mathrm{top}_k(T_q)} \mathbf{V}(T, D)$. The selected detector for query T_q is then $\widehat{D}(T_q) = \arg\max_{D \in \mathcal{B}} \bar{V}_D(T_q)$. When selecting an ensemble of $N > 1$ detectors, the N detectors with the highest $\bar{V}_D(T_q)$ are selected. Because the full performance profile of all candidate detectors is averaged over the neighborhood, a detector is selected only if it is broadly strong across the retrieved series. Similar series may favor different detectors, and the retrieved neighbors therefore need not agree on their individual best detector.

3.2 Enhancing the Similarity Space: Multimodal Semantic Alignment

We now describe the semantic alignment module that incorporates dataset-level context. As illustrated in Figure 5, semantic alignment structures TSFM embeddings using dataset-level context. This motivates aligning each segment with both its paired description and related descriptions.

Integrating Dataset Descriptions. Each segment $T_i^{(t,\ell)}$ inherits the textual description of the dataset from which it is drawn. Let c_i denote the description associated with the dataset of series T_i, and let $z_i^{\mathrm{desc}} = f_{\mathrm{text}}(c_i)$ be its text embedding, where $z_i^{\mathrm{desc}} \in \mathbb{R}^d$. In practice, f_{text} is instantiated with the `all-mpnet-base-v2` Sentence Transformer. Since different datasets may belong to related domains, their descriptions can be similar even when the corresponding time-series segments are not identical.

Description Consistency and Multi-Positive Matching. A one-to-one alignment between time-series segments and dataset descriptions is overly restrictive, since different datasets often carry highly similar textual descriptions. In standard contrastive training, each segment has exactly one positive description, and every other description in the batch is treated as a negative and pushed away. When two datasets have near-synonymous descriptions then the description of the other dataset is wrongly declared a negative, which injects label noise in proportion to the number of near- duplicate descriptions in the benchmark. To account for this, we compute a description-similarity matrix from pairwise dot products between text embeddings. We then threshold this matrix at $\gamma = 0.95$ and treat any pair of descriptions with similarity $\geq \gamma$ as semantically compatible.

This yields a binary *multi-positive* matching matrix $P \in \{0,1\}^{B\times B}$ over a batch of B time-series-segment–description pairs. An entry $P_{ij} = 1$ indicates that the j-th text description is considered a valid positive for the i-th segment, i.e., the segment may plausibly originate from the domain or context described by that text. This matrix can therefore contain multiple positive entries per row. We additionally mask out in-batch pairs coming from the same time series, so that valid positives must originate from a different series than the anchor.

Symmetric Alignment Objective. Because one description may match multiple segments and one segment may be compatible with multiple descriptions, we optimize a symmetric multi-positive contrastive objective. The objective is *symmetric* because it is applied in both directions: from time-series segments to descriptions and from descriptions to time-series segments. The exact formulation of the loss is given in Appendix B.2.

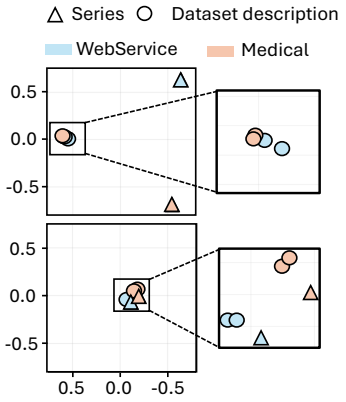


Figure 5: Before alignment (top), TSFM embeddings lack semantic structure; semantic alignment injects dataset-level description information (bottom).

Using the Aligned Space at Inference. The alignment module has two projection heads, one per modality (Appendix B.2), and the text head is only used inside the training objective. At knowledge-base construction and at query time, all series follow the same path and retrieval is entirely time-series-to-time-series: text embeddings are neither stored in the index nor computed for a query, so a query never requires a description. Moreover, the alignment objective only uses the correspondence between segments and descriptions; it never uses detector identities or detector-performance values, so it is independent of the selection task, and the selector itself remains training-free. In our OOD experiments, the alignment module is trained without the target domain and is applied to that domain without any retraining: a query from an unseen domain is mapped close to the semantically closest domains covered by the knowledge base.

3.3 System Complexity

In RAMSAD, the dominant cost arises from segment retrieval. With exact search, processing a batch of b query segments requires computing similarities against all n indexed segments, leading to $\mathcal{O}(bnd)$ query time and $\mathcal{O}(nd + |\mathfrak{D}|M)$ persistent storage for the embedding database and the performance matrix. An HNSW index avoids exhaustive comparisons at a small extra storage cost, making retrieval sublinear. A detailed analysis is deferred to Appendix B.5.

4 Experimental Evaluation

We first compare our proposed approach with SOTA automated solutions both in terms of accuracy and efficiency. We then evaluate it when using different similarity spaces. We provide additional analysis for scalability (Appendix B.5) and parameter sensitivity (Appendix C).

Experimental Setup: We evaluate on TSB-AD [33], which includes 870 univariate time series from 9 domains, using the same train/test splits as recent model selection benchmarks [34] (details in Appendix D.1). We compare against 25 automated solutions, including model selection, ensemble, model generation, and simple reference strategies such as Oracle, SS, GB, Random, and FM; all methods are described in Appendix D.3. Among these, Oracle and Supervised Selection (SS) are not deployable methods but oracle-style upper-bound references: Oracle picks the best detector of each test series, while SS selects the detector using the labels of a portion of the target series itself, which are unavailable to RAMSAD and to all other selectors. The candidate pool contains 32 univariate time-series anomaly detectors from TSB-AD [46], listed in Appendix D.2 where 19 of them are also applicable to multivariate series [34]. We evaluate accuracy using VUS-PR [45, 12]. For RAMSAD, we test $k \in [1, 10]$, and use $k = 6$ as default. In ID, the knowledge base holds the 619 training series as 3,264 segments, and the 251 test series give 1,383 query segments. In OOD, we hold out each of the 9 domains in turn, removing its series from the knowledge base and from alignment training such that no query has a same-domain neighbor, and no retraining is performed for the target domain.

In-Distribution Evaluation (ID). Figure 6 (a) shows the overall VUS-PR score on the in-distribution test split (i.e., 251 series) for all approaches, which are categorized according to the underlying

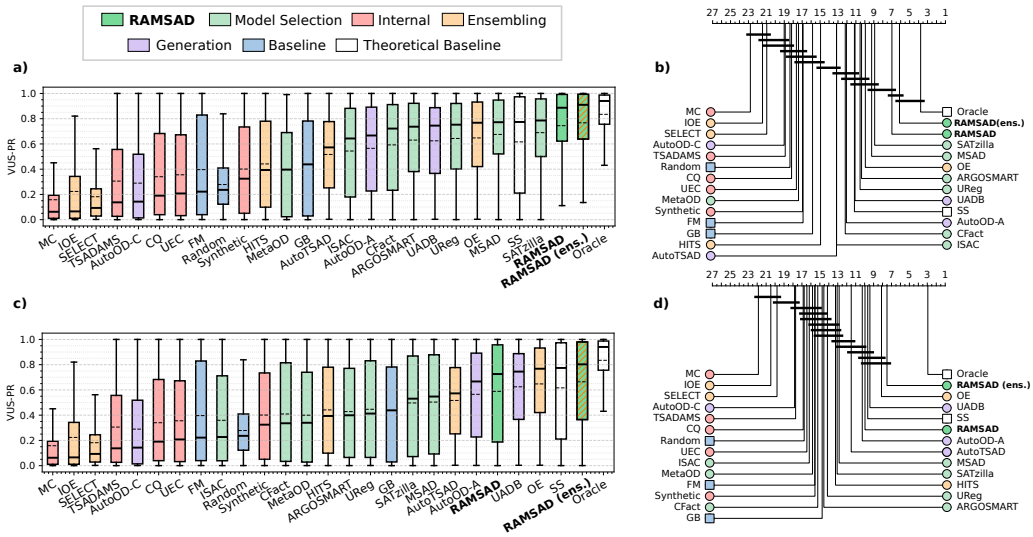


Figure 6: VUS-PR comparison for in-distribution (panel a) and out-of-distribution settings (panel c), with the associated Critical Difference Diagrams (panels b and d, *resp.*).

methodology. RAMSAD outperforms all other baselines, achieving an average VUS-PR score of 0.74. This shows that a retrieval-based selector, without any trained meta-learner, can outperform trained meta-learning selectors. Among other baselines, the best model selector is SATZilla, with an average VUS-PR of 0.69; the best ensemble approach is OE, with an average VUS-PR of 0.64; the best model generation is UADB with an average VUS-PR of 0.62. Unsupervised model selection methods lag significantly behind; the best performer, Synthetic [15], achieves only 0.40, highlighting the difficulty of unsupervised selection. In the critical difference diagram (Figure 6(b)), RAMSAD stands out as the best overall approach with an average rank of 6.97, ahead of the best baseline, SATZilla, which recorded an average rank of 8.90.

Out-of-Distribution Evaluation (OOD). We now consider the out-of-distribution scenario, in which for a time series of a given domain, model selectors can only consider a training set composed of time series from other domains. This can be noticed in Figure 6 (c): all model selection approaches are affected by OOD, which is expected since every selector must generalize to a domain that is absent from its training data. For example, SATZilla has an average VUS-PR of 0.50 (-0.19 w.r.t. the ID setting), and in general all selection approaches showed VUS-PR drops of around 0.19–0.20. RAMSAD is also challenged by OOD, achieving an average VUS-PR of 0.59. However, RAMSAD is less affected than other selectors, with a drop of -0.16, showing better robustness and even managing to outperform some of ensembles and model generation approaches such as AutoTSAD and AutoOD-A. The CD plot in Figure 6 (d) summarizes the average ranking of each approach, with OE achieving the best ranking (7.5), followed by UADB (8.7). Notably, the performance of RAMSAD (avg. rank of 9.9) is not statistically different from OE and UADB, widening the gap with other model selectors that were closer in the ID scenario. RAMSAD therefore attains comparable robustness at a fraction of the cost, matching these methods while running a single detector. Despite being the best approaches in the OOD setting,

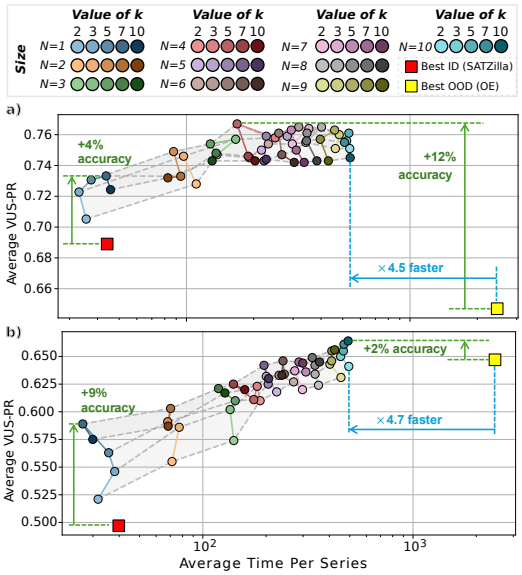


Figure 7: RAMSAD anomaly detection Runtime vs Accuracy by varying k and N.

Table 1: Latency from query series to detector recommendation, for single-detector selectors

	RAMSAD	ARGOSMART	ISAC	MSAD	SATZilla	UReg	MetaOD	CFact
Features	**19.6**				37.6 (catch22, shared)			
Inference	0.05	1.1	25.6	25.8	39.7	221.8	426.7	453.4
Median	**19.6**	38.8	63.3	63.6	77.1	258.0	463.4	489.6
Std	**0.4**	7.8	7.8	8.3	12.2	25.1	28.6	32.2

OE and UADB need to run all the 32 detectors. However, we will show in the following section that RAMSAD can outperform these approaches by selecting a much smaller subset of detectors.

Selecting Multiple Detectors. Figure 7(a) shows the average VUS-PR achieved by RAMSAD in ID scenarios when selecting ensembles of 2 to 10 detectors, illustrating the effect of different values of k and comparing against the best ID and OOD baselines, SATZilla and OE, respectively, while also reporting the execution time required to run the selected detector(s). Already with a single detector ($N = 1$), RAMSAD outperforms SATZilla by 4%. As a further benefit, selecting ensembles increases the average VUS-PR up to 0.77, widening the gap to SATZilla to 8% and achieving a 12% accuracy improvement over OE. In OOD (Figure 7(b)), RAMSAD outperforms SATZilla by 9% with $N = 1$, and accuracy keeps improving as N increases: with $N = 10$ the gap to SATZilla grows to 17%, and RAMSAD exceeds OE by 2% while being $\times 4.7$ faster. The latter is further supported by the boxplots and CD diagrams in Figure 6. The ensemble version of RAMSAD outperforms all baselines in ID and OOD, while being one order of magnitude faster than the second best in OOD. Most notably, RAMSAD ensemble has no statistically significant difference from the oracle.

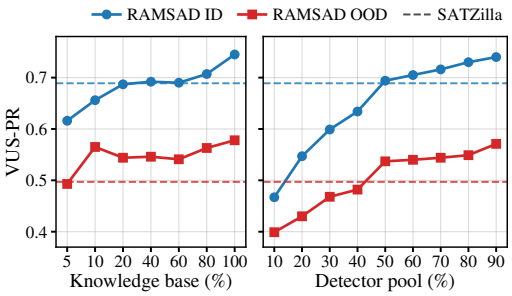


Figure 8: RAMSAD VUS-PR when randomly reducing the knowledge base (left) and the detector pool (right).

Parameter Sensitivity. Figure 7 shows the behaviour of RAMSAD for different values of k and N in both ID and OOD. Overall, RAMSAD is stable across values of k: in the ID setting, VUS-PR varies only marginally, with a maximum difference of about 0.04, while in the OOD setting the variation is slightly larger, up to 0.07 for $N = 1$, but decreases as the ensemble size increases; in all cases, RAMSAD consistently outperforms the best model selector, SATZilla.

Selection Latency. Beyond the cost of running the selected detectors, Table 1 isolates the cost of the selection itself, from a raw query series to a recommendation. RAMSAD is the fastest selector (20 ms), where its TSFM forward pass is cheaper than the catch22 extraction used by the meta-learners, which alone takes 38 ms. RAMSAD is also the most predictable as its latency varies by 0.4 ms across queries. End-to-end, single-detector selectors cost one detector run plus 39–490 ms, RAMSAD one run plus 20 ms ($N = 1$) or ten runs plus 20 ms ($N = 10$), and full-pool methods 32 runs. As OE is the cheapest of the latter, the $\times 4.7$ speed-up of RAMSAD ($N = 10$) over OE lower-bounds its speed-up over UADB, AutoTSAD, AutoOD-A, and HITS.

Knowledge-Base Coverage and Detector-Pool Size. Since RAMSAD relies on a labeled knowledge base, we study how its coverage affects selection (Figure 8; details in Appendix E). Reducing the knowledge base, RAMSAD still outperforms SATZilla in OOD with 10% of the training set, and in ID is comparable from 20% and better from 80%. Reducing the detector pool, it matches or outperforms SATZilla with half of it. With the leave-one-domain-out protocol this indicates that what matters is proximity in the representation space rather than shared domain membership.

Similarity Spaces Comparison. We compare three retrieval spaces for RAMSAD: (1) original time-series, (2) TSFM embedding, and (3) semantically aligned.

Figure 9 summarizes the main trends (full evaluation in Appendix C). First, DTW, the strongest original-space baseline, reaching VUS-PR of 0.69 in ID and 0.44 in OOD, remains below the best TSFM and aligned-TSFM spaces. By comparison, retrieval in the plain TSFM, such as `Chronos2`,

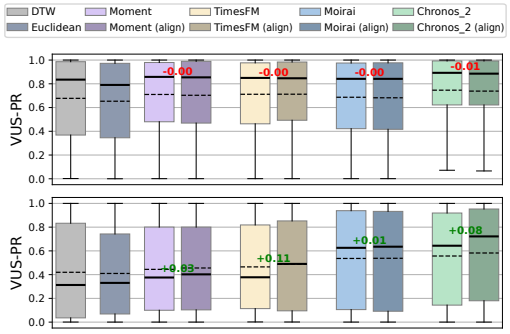


Figure 9: Comparison between original space (Euclidean and DTW), TSFM (light color), and aligned TSFM embedding spaces (darker color).

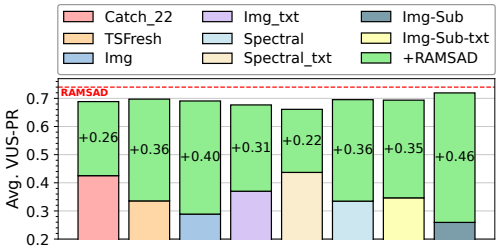


Figure 10: LLM-based model selection with and without RAG support (+RAMSAD).

reaches 0.75 in ID and 0.59 in OOD. Second, the strongest overall behavior is obtained in the aligned TSFM space. Figure 9 shows that the aligned-TSFM performances are similar to the plain TSFM space in ID, and improve on them in OOD. In OOD, the domain of the query is unknown, and temporal patterns alone may be insufficient to identify the most informative source series. In this case, incorporating semantic information becomes more important, which explains why the aligned space yields a clearer advantage in OOD. Dataset-level descriptions inject a coarse domain prior that, for some backbones, slightly lowers ID accuracy while improving cross-domain transfer (Appendix C.2). Alignment is thus an OOD-oriented refinement rather than a prerequisite.

Implication in Modern LLM-based Systems. RAMSAD can also serve as the retrieval component (RAG) of LLM-based anomaly-detection pipelines. To support this claim, Figure 10 reports the average VUS-PR achieved when using a high-capacity LLM (mistral-large-3, 675B) as model selector, to which we provided different types of TS representations, including features (Catch 22 [36] and TSFresh [16]), images (full, spectral, and subsequence image), and LLM-generated summaries describing the latter (denoted as *txt* in the figure). As we can see, the VUS-PR performance is rather low (between 0.28 and 0.42). Applying RAMSAD, we retrieved 5 similar training examples and provided their detectors' VUS-PR scores to the LLM. This drastically increased performances, with scores ranging from 0.66 to 0.72. While some of these results exceed the previous SOTA (0.69), they are still slightly lower than RAMSAD. RAMSAD supports post-hoc explainability, since retrieved neighbors and their detector-performance profiles can be given to ChatTS [65] to generate a rationale for the selected model.

5 Conclusion

We presented RAMSAD, a retrieval-based framework for model selection in time-series anomaly detection. By selecting detectors through similarity to previously observed series, RAMSAD avoids training a dedicated selector while remaining flexible and efficient; its optional semantic alignment is trained offline, independently of the selection task. Our key takeaways are as follows:

Takeaway 1: *RAMSAD is the strongest overall method: it achieves the best single-model performance in ID and, when used to select small ensembles, reaches SOTA results in both ID and OOD while selecting substantially fewer detectors than full-pool ensemble methods.*

Takeaway 2: *TSFM-based retrieval spaces are preferable for RAMSAD; among them, semantically aligned TSFM spaces deliver the strongest overall performance, mainly through improved OOD transfer, while unaligned TSFM spaces already outperform trained selectors.*

These findings establish retrieval as a competitive foundation for anomaly-detector selection in heterogeneous time series. We also observe (in preliminary experiments) that even when neighbours disagree, the averaged profile in RAMSAD still picks the right detector; we defer a detailed study to future work. Extending RAMSAD to multivariate series, temporal concept drift, and queries far from every covered domain are also interesting directions for future work (Appendix F.1).

Acknowledgments and Disclosure of Funding

Supported by EU Horizon project DataGEMS (101188416). This work was granted access to the HPC resources of IDRIS under the allocations 2025-A0191012641 and 2025-AD011016632 by GENCI. We thank Azedine Mani, Head of the IT team (UFR Mathématiques et Informatique, Université Paris Cité), for his pivotal role in creating and making available computational resources essential to this work. The authors declare no competing interests.

References

[1] C. C. Aggarwal. An introduction to outlier analysis. In *Outlier Analysis*, pages 1–34. Springer International Publishing, Cham, 2017. doi: 10.1007/978-3-319-47578-3_1.

[2] C. C. Aggarwal and S. Sathe. Theoretical foundations and algorithms for outlier ensembles. *Acm sigkdd explorations newsletter*, 17(1):24–47, 2015.

[3] A. F. Ansari, L. Stella, C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shchur, S. S. Rangapuram, S. P. Arango, S. Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

[4] A. F. Ansari, O. Shchur, J. Küken, A. Auer, B. Han, P. Mercado, S. S. Rangapuram, H. Shen, L. Stella, X. Zhang, M. Goswami, S. Kapoor, D. C. Maddix, P. Guerron, T. Hu, J. Yin, N. Erickson, P. M. Desai, H. Wang, H. Rangwala, G. Karypis, Y. Wang, and M. Bohlke-Schneider. Chronos-2: From univariate to universal forecasting, 2025. URL `https://arxiv.org/abs/2510.15821`.

[5] J. Antoni and P. Borghesani. A statistical methodology for the design of condition indicators. *Mechanical Systems and Signal Processing*, 114:290–327, 2019. doi: 10.1016/j.ymssp.2018.05.012.

[6] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga. Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3395–3404, 2020.

[7] A. Auer, P. Podest, D. Klotz, S. Böck, G. Klambauer, and S. Hochreiter. Tirex: Zero-shot forecasting across long and short horizons with enhanced in-context learning. *arXiv preprint arXiv:2505.23719*, 2025.

[8] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano. A review on outlier/anomaly detection in time series data. *ACM Computing Surveys*, 54(3):56:1–56:33, 2022. doi: 10.1145/3444690.

[9] P. Boniol and T. Palpanas. Series2graph: Graph-based subsequence anomaly detection for time series. *Proc. VLDB Endow.*, 13(12):1921–1934, 2020.

[10] P. Boniol, J. Paparrizos, T. Palpanas, and M. J. Franklin. Sand: Streaming subsequence anomaly detection. *Proc. VLDB Endow.*, 14(10):1717–1729, 2021.

[11] P. Boniol, Q. Liu, M. Huang, T. Palpanas, and J. Paparrizos. Dive into time-series anomaly detection: A decade review, 2024. URL `https://arxiv.org/abs/2412.20512`.

[12] P. Boniol, A. K. Krishna, M. Bruel, Q. Liu, M. Huang, T. Palpanas, R. S. Tsay, A. Elmore, M. J. Franklin, and J. Paparrizos. Vus: effective and efficient accuracy measures for time-series anomaly detection. *The VLDB Journal*, 34(3):32, 2025.

[13] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[14] L. Cao, Y. Yan, Y. Wang, S. Madden, and E. A. Rundensteiner. Autood: Automatic outlier detection. *Proceedings of the ACM on Management of Data*, 1(1):1–27, 2023.

[15] 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.

[16] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing*, 307:72–77, 2018.

[17] B. Cohen, E. Khwaja, Y. Doubli, S. Lemaachi, C. Lettieri, C. Masson, H. Miccinilli, E. Ramé, Q. Ren, A. Rostamizadeh, et al. This time is different: An observability perspective on time series foundation models. *arXiv preprint arXiv:2505.14766*, 2025.

[18] A. Das, W. Kong, R. Sen, and Y. Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

[19] F. Del Gaudio, M. Chatzakis, G. Ravendirane, B. Peng, and T. Palpanas. Daisy: A library for scalable data series similarity search. *arXiv preprint arXiv:2603.27719*, 2026.

[20] V. Feofanov, S. Wen, M. Alonso, R. Ilbert, H. Guo, M. Tiomoko, L. Pan, J. Zhang, and I. Redko. Mantis: Lightweight calibrated foundation model for user-friendly time series classification. *arXiv preprint arXiv:2502.15637*, 2025.

[21] Y. Gao, J. Lin, and C. Brif. Ensemble grammar induction for detecting anomalies in time series. In *Proceedings of the 23rd International Conference on Extending Database Technology, EDBT 2020, Copenhagen, Denmark, March 30 - April 02,2020*, pages 85–96, 2020. doi: 10.5441/002/EDBT.2020.09.

[22] N. Goix. How to evaluate the quality of unsupervised anomaly detection algorithms? *arXiv preprint arXiv:1607.01152*, 2016.

[23] M. Goldstein and A. Dengel. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: poster and demo track*, 1:59–63, 2012.

[24] M. Goswami, C. Challu, L. Callot, L. Minorics, and A. Kan. Unsupervised model selection for time-series anomaly detection. *arXiv preprint arXiv:2210.01078*, 2022.

[25] M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, and A. Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

[26] L. Graf, T. Ortner, S. Woźniak, A. Pantazi, et al. Flowstate: Sampling rate invariant time series forecasting. *arXiv preprint arXiv:2508.05287*, 2025.

[27] S. Kadioglu, Y. Malitsky, M. Sellmann, K. Tierney, et al. Isac-instance-specific algorithm configuration. In *ECAI*, volume 215, pages 751–756, 2010.

[28] E. Keogh, S. Lonardi, C. A. Ratanamahatana, L. Wei, S.-H. Lee, and J. Handley. Compression-based data mining of sequential data. *Data Mining and Knowledge Discovery*, 14(1):99–129, 2007.

[29] Z. Li, H. Ma, and Y. Mei. A unifying method for outlier and change detection from data streams based on local polynomial fitting. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 150–161. Springer, 2007.

[30] Z. Lin, K. Thekumparampil, G. Fanti, and S. Oh. Infogan-cr and modelcentrality: Self-supervised model training and selection for disentangling gans. In *international conference on machine learning*, pages 6127–6139. PMLR, 2020.

[31] M. Linardi, Y. Zhu, T. Palpanas, and E. Keogh. Matrix profile goes mad: variable-length motif and discord discovery in data series. *arXiv preprint arXiv:2008.13447*, 2020.

[32] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[33] Q. Liu and J. Paparrizos. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. *Advances in Neural Information Processing Systems*, 37:108231–108261, 2024.

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

[35] Y. Liu, X. Chen, F. Wang, and J. Yin. Efficient detection of discords for time series stream. In *Asia-Pacific Web Conference*, pages 629–634. Springer, 2009.

[36] C. H. Lubba, S. S. Sethi, P. Knaute, S. R. Schultz, B. D. Fulcher, and N. S. Jones. catch22: Canonical time-series characteristics. *arXiv preprint arXiv:1901.10200*, 2019.

[37] H. Ma, B. Ghojogh, M. N. Samad, D. Zheng, and M. Crowley. Isolation mondrian forest for batch and online anomaly detection. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3051–3058. IEEE, 2020.

[38] M. Q. Ma, Y. Zhao, X. Zhang, and L. Akoglu. The need for unsupervised outlier model selection: A review and evaluation of internal evaluation strategies. *ACM SIGKDD Explorations Newsletter*, 25(1):19–35, 2023.

[39] P. Malhotra, L. Vig, G. M. Shroff, and P. Agarwal. Long short term memory networks for anomaly detection in time series. In *The European Symposium on Artificial Neural Networks*, 2015. URL `https://api.semanticscholar.org/CorpusID:43680425`.

[40] M. Munir, S. A. Siddiqui, A. Dengel, and S. Ahmed. Deepant: A deep learning approach for unsupervised anomaly detection in time series. *IEEE access*, 7:1991–2005, 2018.

[41] J. M. Navarro, A. Huet, and D. Rossi. Meta-learning for fast model recommendation in unsupervised multivariate time series anomaly detection. In *International Conference on Automated Machine Learning*, pages 24–1. PMLR, 2023.

[42] T. T. Nguyen, U. Q. Nguyen, et al. An evaluation method for unsupervised anomaly detection algorithms. *Journal of Computer Science and Cybernetics*, 32(3):259–272, 2016.

[43] M. Nikolić, F. Marić, and P. Janičić. Simple algorithm portfolio for sat. *Artificial Intelligence Review*, 40(4):457–465, 2013.

[44] J. Paparrizos and L. Gravano. k-shape: Efficient and accurate clustering of time series. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, pages 1855–1870, 2015.

[45] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. J. Elmore, and M. J. Franklin. Volume under the surface: A new accuracy evaluation measure for time-series anomaly detection. *Proc. VLDB Endow.*, 15(11):2774–2787, 2022.

[46] J. Paparrizos, Y. Kang, P. Boniol, R. S. Tsay, T. Palpanas, and M. J. Franklin. Tsb-uad: An end-to-end benchmark suite for univariate time-series anomaly detection. *Proc. VLDB Endow.*, 15(8):1697–1711, 2022.

[47] B. Peng, P. Fatourou, and T. Palpanas. Fast data series indexing for in-memory data. *The VLDB Journal*, 30(6):1041–1067, 2021.

[48] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 427–438, 2000.

[49] K. Rasul, A. Ashok, A. R. Williams, H. Ghonia, R. Bhagwatkar, A. Khorasani, M. J. D. Bayazi, G. Adamopoulos, R. Riachi, N. Hassen, et al. Lag-llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*, 2023.

[50] S. Rayana and L. Akoglu. Less is more: Building selective anomaly ensembles. *Acm transactions on knowledge discovery from data (tkdd)*, 10(4):1–33, 2016.

[51] H. Ren, B. Xu, Y. Wang, C. Yi, C. Huang, X. Kou, T. Xing, M. Yang, J. Tong, and Q. Zhang. Time-series anomaly detection service at microsoft. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3009–3017, 2019.

[52] P. J. Rousseeuw and K. V. Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

[53] M. Sakurada and T. Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis*, pages 4–11, 2014.

[54] S. Schmidl, P. Wenig, and T. Papenbrock. Anomaly detection in time series: A comprehensive evaluation. *Proceedings of the VLDB Endowment*, 15(9):1779–1797, 2022. doi: 10.14778/3538598.3538602.

[55] S. Schmidl, F. Naumann, and T. Papenbrock. Autotsad: unsupervised holistic anomaly detection for time series data. *Proceedings of the VLDB Endowment*, 17(11):2987–3002, 2024.

[56] B. Schölkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt. Support vector method for novelty detection. *Advances in neural information processing systems*, 12, 1999.

[57] 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*, pages 2828–2837, 2019.

[58] 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*, 16(11):3418–3432, 2023.

[59] E. Sylligardos, J. Paparrizos, T. Palpanas, P. Senellart, and P. Boniol. MSAD: A deep dive into model selection for time series anomaly detection. *The VLDB Journal*, 34(6):72, 2025. ISSN 0949-877X. doi: 10.1007/s00778-025-00949-1. URL `https://doi.org/10.1007/s00778-025-00949-1`.

[60] S. Tuli, G. Casale, and N. R. Jennings. Tranad: Deep transformer networks for anomaly detection in multivariate time series data. *arXiv preprint arXiv:2201.07284*, 2022.

[61] P. Wenig and T. Papenbrock. Series2graph++: Distributed detection of correlation anomalies in multivariate time series. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 225–230. Springer, 2024.

[62] G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, and D. Sahoo. Unified training of universal time series forecasting transformers. In *Forty-first International Conference on Machine Learning*, 2024.

[63] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

[64] W. Wu, L. He, W. Lin, Y. Su, Y. Cui, C. Maple, and S. Jarvis. Developing an unsupervised real-time anomaly detection scheme for time series with multi-seasonality. *IEEE Transactions on Knowledge and Data Engineering*, 34(9):4147–4160, Sept. 2022. ISSN 2326-3865. doi: 10.1109/tkde.2020.3035685. URL `http://dx.doi.org/10.1109/TKDE.2020.3035685`.

[65] Z. Xie, Z. Li, X. He, L. Xu, X. Wen, T. Zhang, J. Chen, R. Shi, and D. Pei. Chatts: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. *arXiv preprint arXiv:2412.03104*, 2024.

[66] H. Xu, W. Chen, N. Zhao, Z. Li, J. Bu, Z. Li, Y. Liu, Y. Zhao, D. Pei, Y. Feng, et al. Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. In *Proceedings of the 2018 world wide web conference*, pages 187–196, 2018.

[67] J. Xu, H. Wu, J. Wang, and M. Long. Anomaly transformer: Time series anomaly detection with association discrepancy. *arXiv preprint arXiv:2110.02642*, 2021.

[68] L. Xu, F. Hutter, H. H. Hoos, and K. Leyton-Brown. Satzilla: portfolio-based algorithm selection for sat. *Journal of artificial intelligence research*, 32:565–606, 2008.

[69] Z. Xu, A. Zeng, and Q. Xu. Fits: Modeling time series with $10k$ parameters. *arXiv preprint arXiv:2307.03756*, 2023.

[70] T. Yairi, Y. Kato, and K. Hori. 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*, volume 18, page 21. Citeseer, 2001.

[71] H. Ye, Z. Liu, X. Shen, W. Cao, S. Zheng, X. Gui, H. Zhang, Y. Chang, and J. Bian. Uadb: Unsupervised anomaly detection booster. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 2593–2606. IEEE, 2023.

[72] C.-C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, D. F. Silva, A. Mueen, and E. Keogh. 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)*, pages 1317–1322. Ieee, 2016.

[73] Y. Zhao, R. A. Rossi, and L. Akoglu. Automatic unsupervised outlier model selection. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pages 4489–4502, 2021.

[74] T. Zhou, P. Niu, L. Sun, R. Jin, et al. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36:43322–43355, 2023.

Appendices

A Literature Review **17**

B RAMSAD Method Details **18**

B.1 Notation 18

B.2 Semantic Alignment Training 18

B.3 Segment-to-Series Aggregation 21

B.4 RAMSAD Model Selection 22

B.5 System Complexity and Scalability 23

C Similarity Spaces: Performance and Runtime Analysis **24**

C.1 Parameter Sensitivity with Respect to k 24

C.2 ID and OOD Performance Tables 24

C.3 Retrieval Runtime Across TSFM Similarity Spaces 27

C.4 Retrieval Runtime of Original Similarity Space 28

D Experimental Setup and Benchmark Details **28**

D.1 Dataset Details 28

D.2 List of Candidate Detectors 28

D.3 Automated TSAD Baselines 29

D.4 Hardware Details 29

E Additional Ablation Studies **30**

E.1 Effect of Reducing the Training Size 30

E.2 Effect of Reducing the Pool Size of Detectors 30

F Additional Discussions **30**

F.1 Limitations 30

F.2 Broader Impacts 33

A Literature Review

Time-series anomaly detection (TSAD) has been studied extensively in recent years because of its importance in monitoring, diagnosis, and decision support across domains such as healthcare, manufacturing, finance, energy, and cybersecurity [8, 54]. A central challenge in TSAD is that anomalies may take very different forms, and the statistical properties of time series vary widely across datasets. As a result, the literature has produced a large variety of detectors, each relying on different assumptions about temporal structure, recurrence, noise, and anomaly manifestation. Recent surveys and benchmark studies consistently show that this diversity prevents the emergence of a single universally best detector across heterogeneous benchmark collections [46, 54, 34, 59].

Taxonomy of TSAD methods. A common organization of the TSAD literature distinguishes three broad families: *distance-based*, *density-based*, and *prediction-based* methods [8, 54, 59]. Distance-based approaches identify anomalous observations or subsequences by measuring dissimilarity with respect to previously observed patterns. This family includes nearest-neighbor and discord-style methods, as well as matrix-profile-based approaches that search for subsequences with unusually large distances to their closest matches [28, 35, 72, 31]. These methods are often effective when anomalies correspond to rare subsequences that clearly depart from the dominant motifs of the series.

Density-based methods instead assess how isolated an observation or subsequence is in an induced representation space. Rather than relying only on raw temporal distances, they estimate local or global concentration patterns and flag sparse or weakly supported regions as anomalous. This category spans several subfamilies, including distribution-based approaches, graph-based methods, tree-based methods, and encoding-based methods [5, 9, 61, 32, 37, 21, 59]. Such methods are attractive when the notion of anomaly is better captured through structural relations or learned representations than through direct pointwise comparison.

Prediction-based methods model expected temporal behavior and use prediction error as the anomaly signal. They are commonly divided into *forecasting-based* and *reconstruction-based* approaches [64, 57, 59]. Forecasting-based detectors predict future values from the past and treat large residuals as indicators of anomalous behavior, whereas reconstruction-based methods learn a compact representation of normal temporal patterns and detect anomalies via reconstruction error. These approaches are particularly appealing when normal dynamics are regular enough to be predicted or reconstructed reliably.

Why detector performance varies across datasets. The lack of a universally dominant detector is now a well-established empirical finding in TSAD benchmarks [46, 54, 34]. A key reason is the heterogeneity of anomaly types. Following prior work, anomalies are often categorized as *point*, *contextual*, or *collective* anomalies [59]. Point anomalies correspond to isolated values that strongly deviate from the bulk of the series, contextual anomalies are only abnormal relative to a local temporal context, and collective anomalies arise when a subsequence is abnormal as a whole even if individual points are not. Different detector families are naturally better suited to different anomaly types, which already makes detector performance highly dataset dependent.

A second source of variability comes from the structure of the time series itself. Even within the same anomaly category, datasets may differ in length, periodicity, stationarity, noise level, sampling characteristics, and the number and diversity of normal regimes [54, 59]. For example, a detector that works well on time series with a single stable normal pattern may degrade substantially when applied to non-stationary data or to series containing multiple normalities. Similarly, methods that are effective for isolated rare events may behave differently when anomalies are frequent, repeated, or highly heterogeneous. These observations explain why benchmark studies repeatedly conclude that detector rankings are unstable across domains and that model choice should depend on the target series rather than on a single globally best detector [46, 34].

Automatic detector selection. These benchmark findings have motivated a growing literature on *automatic* approaches for choosing or combining detectors [34], which is summarized in Figure 2. The first line of work is *model selection*, where the objective is to choose the detector expected to perform best on a given time series. Many of these methods follow a meta-learning perspective: they use benchmark results together with representations of datasets or time series to learn how detector performance varies across problems. Classical algorithm-selection approaches such as `ARGOSMART` and `ISAC` rely on similarity in a meta-feature space, selecting either the best detector on the nearest training instance or the best detector inside a nearest cluster [43, 27]. Other approaches formulate

selection through collaborative filtering, regression, or classification, predicting detector identity or detector performance from dataset characteristics [73, 68, 41, 58, 59].

A complementary group of methods avoids supervised meta-learning and instead evaluates candidate detectors directly on the target time series using internal criteria. These approaches include unsupervised score-evaluation techniques, clustering-based validation criteria, and detector-assessment schemes based on synthetic perturbations or rank aggregation [22, 42, 15, 24]. Their main appeal is that they do not require labeled historical performance data. However, they generally still require executing multiple detectors on the target series, which can substantially reduce their practical advantage when the candidate pool is large.

Model ensembling and model generation. A second major line of work is *model ensembling*, where multiple detectors are executed and their scores are aggregated to improve robustness [2]. Simple averaging or max-based ensembles often provide a strong baseline because they hedge against detector-specific failures. More selective ensembling methods attempt to identify a subset of reliable detectors before aggregation [50, 38]. At a larger systems level, `AutoTSAD` automates detector configuration, synthetic data generation, and final ensembling within a unified pipeline [55]. Although ensembles can be effective, their main limitation is computational: they typically require running many or all detectors before producing a final anomaly score.

A third direction is *model generation*, where multiple detector outputs are used to build a new detector, usually through pseudo-labeling or boosting. Examples include `AutoOD-A` and `AutoOD-C`, which iteratively refine pseudo-labels and detector subsets, and `UADB`, which learns a booster from the outputs of several detectors [14, 71]. These approaches can improve robustness and adaptivity, but they are also more complex, and like ensembling methods they often require evaluating a substantial fraction of the detector pool.

Position of our work. Our work is most closely related to the *model selection* line, but it departs from most recent TSAD selectors in an important way. Whereas much of the recent literature formulates detector selection as a trainable classification or regression problem [58, 59], we revisit a simpler non-parametric principle: similar time series should tend to prefer similar detectors. This intuition is related to earlier similarity-based algorithm selection methods such as `ARGOSMART` and `ISAC` [43, 27], but our framework is tailored to TSAD and is explicitly designed to operate directly on time-series similarity spaces, including raw-distance retrieval and embedding-based retrieval. In contrast to trainable selectors, retrieval-based model selection does not require fitting a separate predictive selector, can adapt naturally when the detector pool changes, and offers a simple way to leverage rich pretrained time-series representations.

Overall, the literature suggests two conclusions that motivate our approach. First, detector performance in TSAD is fundamentally heterogeneous, so automatic adaptation is necessary [46, 34]. Second, while many existing automatic solutions are effective, they often rely either on heavy training pipelines or on the execution of many detectors at inference time [59, 55]. This leaves room for lightweight alternatives that exploit similarity structure in previously observed benchmark instances. Our proposed framework addresses this gap by casting detector selection as retrieval over a knowledge base of time series and detector-performance profiles.

B RAMSAD Method Details

B.1 Notation

Table 2 summarizes the main notation used throughout the paper and appendix. We group the symbols into two parts. The first part introduces the general time-series anomaly-detection setting, including the collection of time series, candidate detectors, detector scores, labels, and performance measures. The second part defines the notation specific to RAMSAD, including the time-series and text encoders, segment and description embeddings, retrieval scores, the detector-performance matrix, and the semantic-alignment objective.

B.2 Semantic Alignment Training

Trainable parameters. The time-series and text embeddings are precomputed using `Chronos2` and `all-mpnet-base-v2`, respectively. During alignment training, these backbone embeddings

Table 2: Table of main symbols used in RAMSAD.

Symbol	Description
	General Notation
$\mathfrak{D}$	Collection of time series
T_i, L_i	The i-th univariate time series and its length
$T_i^{(t,\ell)}$	A subsequence of length ℓ extracted from T_i starting at index t
$\mathcal{B}$	Finite pool of candidate anomaly detectors
D_j	The j-th candidate anomaly detector
$S_{i,j}, Y_i$	Anomaly score sequence produced by D_j on T_i, and ground-truth labels
$Acc(\cdot)$	A performance measure (e.g., AUC-PR, F1, VUS)
$\Phi, \Phi^\star$	The model selector function, and the oracle choice
M	The number of all detectors
N	The number of detectors selected for an ensemble
	RAMSAD Core Retrieval & Alignment
f_{ts}, f_{text}	Time-series foundation model (encoder) and Text encoder
$z_{i,t}^{ts}, z_i^{desc}$	Encoded representations of time-series segment and text description
T_q, m	A query time series, and the number of segments extracted from it
k	Number of retrieved neighbors
$\mathcal{A}(T_j)$	Set of retrieved distances associated with segments from series T_j
$base(T_j), freq(T_j)$	Base similarity score and frequency score of candidate series T_j
$score(T_j)$	Final ranking score of candidate series T_j
$\pi(\cdot)$	Function mapping a segment to its originating time series
$top_k(T_q)$	The final top-k retrieved series for query T_q
$\mathbf{V}$	Precomputed performance matrix of detectors over the database
$\bar{V}_D(T_q)$	Aggregated performance of detector D over retrieved neighbors
$\widehat{D}(T_q)$	The final selected detector for query T_q
γ	Similarity threshold for multi-positive matching
τ	Temperature used in the contrastive alignment loss
$\mathbf{P}, \mathcal{L}$	Binary multi-positive matching matrix, and symmetric alignment loss
$\mathcal{R}_i, \mathcal{Q}_j$	Positive sets for time-series segment i and description j in the symmetric alignment loss

remain frozen, and only the projection heads are optimized. Each projection head is implemented as a lightweight multilayer perceptron of the architecture illustrated in Figure 11 with hidden width equal to $2\times$ the projection dimension. The outputs are ℓ_2-normalized before computing similarities in the shared space.

Optimization. Semantic alignment is trained with AdamW using learning rate 10^{-5}, weight decay 10^{-4}, and $\beta = (0.9, 0.999)$. Training runs for 300 epochs with gradient clipping at norm 1.0. We use a warmup-cosine learning-rate schedule, consisting of a linear warmup for the first 10 epochs followed by cosine decay to zero. The default batch size is 8192. The training split is capped at $100{,}000$ samples, and examples are sampled using a stratified-uniform procedure so as to approximately balance the chosen stratification labels.

Validation split and batching. The default train/validation split is $0.8/0.2$, applied independently within each stratification group whenever possible. Stratification is based on the dataset-description column by default, with fallback to raw description text when needed.

Loss configuration. The default alignment objective is the multi-positive loss described in the main paper. We use temperature 0.1 and define positive matches through description-embedding similarity, using a cosine-similarity threshold of 0.95. In addition, we mask same-series pairs so that positives must come from different time series within the batch. The framework also supports alternative loss functions such as contrastive, cosine, and distance-based objectives, though the reported experiments use the multi-positive setting.

Why multi-positive targets. Contrastive alignment is trained in batches, where each segment is paired with exactly one description, its positive. The model computes the similarity between each segment and every description in the batch, normalizes each row into a probability distribution, and minimizes the negative log-probability of the correct entry. Because each row must sum to one, raising the probability of the positive necessarily lowers that of every other entry (i.e. every other description in the batch is implicitly treated as a negative and pushed away). This assumption breaks in our setting. For a segment from dataset A, the description of dataset B is declared a negative even

when A and B have near-synonymous descriptions (e.g., two distinct server-monitoring datasets). The objective then pushes the segment away from a semantically valid description, injecting label noise in proportion to the number of near-duplicate descriptions in the benchmark. We therefore treat pairs of descriptions whose frozen sentence-transformer embeddings have cosine similarity of at least $\gamma = 0.95$ as positives, which replaces the single correct answer per row by a multi-positive target matrix where a segment is no longer penalized for scoring highly against a description whose meaning matches its own. Note that this target matrix is computed from the frozen text embeddings only and it does not depend on the time-series embeddings or on the trainable projections, which only enter the predicted cross-modal similarities that the loss pushes toward the targets.

Use of the trained heads at inference. Only the time-series projection head is used after training: knowledge-base segments and query segments are both encoded by the frozen TSFM, projected by the time-series head, and ℓ_2-normalized, and retrieval compares them by cosine similarity. The text head and the text embeddings are neither stored in the index nor computed for a query. The alignment objective never uses detector identities or detector-performance values.

Symmetric Alignment Loss. The symmetric multi-positive objective can be understood as a batch-wise extension of standard contrastive learning. For each time-series segment i, the denominator $\sum_{j=1}^{B} \exp(s_{ij})$ normalizes the similarities to all descriptions in the batch, thereby defining a softmax distribution over candidate text matches. The numerator $\sum_{j \in \mathcal{R}_i} \exp(s_{ij})$ collects the probability mass assigned to the subset of descriptions regarded as valid positives for that segment. The first summation term therefore encourages each segment embedding to assign high total probability to all compatible descriptions, rather than forcing the mass onto a single one-hot target.

The second term plays the same role in the reverse direction. For each description j, the denominator $\sum_{i=1}^{B} \exp(s_{ij})$ defines a softmax distribution over all time-series segments in the batch, while the numerator $\sum_{i \in \mathcal{Q}_j} \exp(s_{ij})$ aggregates the mass assigned to the set of compatible segments. This enforces alignment both from segment to description and from description to segment, making the objective symmetric across modalities.

This multi-positive formulation is important in our setting because the supervision is inherently many-to-many. Different datasets may have highly similar descriptions, so several descriptions can be valid semantic matches for the same segment. Likewise, one description may be compatible with many segments from related time series. By summing over all positives in the numerator, the objective rewards the model whenever it concentrates probability mass on any semantically valid match, instead of penalizing it for not selecting one arbitrarily designated target. Since the softmax normalization couples positives and negatives, assigning more mass to valid matches necessarily reduces the mass assigned to incompatible pairs, so the loss remains contrastive.

Formally, letting $s_{ij} = (z_i^{ts})^\top z_j^{desc}/\tau$ denote the temperature-scaled similarity between normalized projected embeddings, and letting $\mathcal{R}_i$ and $\mathcal{Q}_j$ denote the positive sets for segment i and description j, respectively, the loss is

$$\mathcal{L} = -\frac{1}{2B}\left[\sum_{i=1}^{B} \log \frac{\sum_{j \in \mathcal{R}_i} \exp(s_{ij})}{\sum_{j=1}^{B} \exp(s_{ij})} + \sum_{j=1}^{B} \log \frac{\sum_{i \in \mathcal{Q}_j} \exp(s_{ij})}{\sum_{i=1}^{B} \exp(s_{ij})}\right].$$

Training loop and model selection. At each epoch, the model is trained on all mini-batches, evaluated on the validation set, and then updated by the scheduler. Validation performance is measured not only by loss but also by retrieval-style top-1 and top-5 accuracy in the shared space. More precisely, projected time-series and text embeddings are compared through matrix similarity, and retrieval is considered correct when the matched item belongs to the valid multi-positive set. Figure 12 reports the evolution of retrieval performance during training. The consistent increase in top-1 and top-5 accuracy across epochs indicates that the semantic alignment objective successfully improves cross-modal matching between time-series segments and dataset descriptions. The checkpoint with the lowest validation loss is stored.

The full semantic-alignment training procedure is given in Algorithm 1. For each mini-batch, RAMSAD first encodes time-series segments and their associated textual descriptions using frozen pretrained encoders, and then maps both embeddings into a shared space through trainable projection heads. The algorithm constructs a multi-positive target matrix from in-batch description similarities,

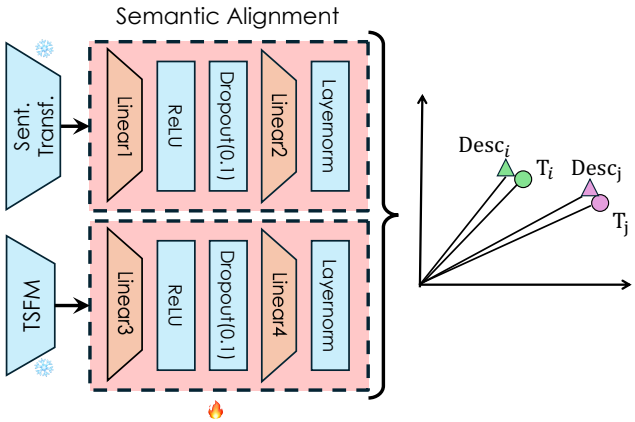
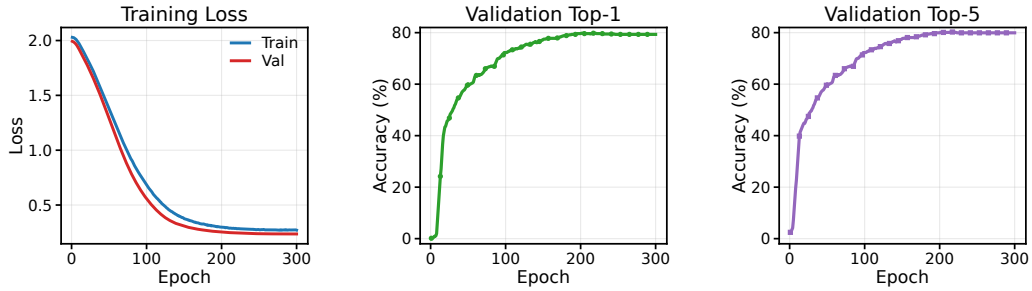


Figure 11: Semantic Alignment Architecture



(a) Train/Validation Loss (b) Top-1 (c) Top-5

Figure 12: As training evolves, the true series–description matches are found more effectively in their batch multi-positive similarity matrices.

allowing semantically similar descriptions to act as valid positives rather than enforcing a single one-to-one match. After masking same-series pairs, the model optimizes the symmetric alignment loss described above, thereby learning a cross-modal space in which time-series segments and compatible descriptions are brought closer together.

B.3 Segment-to-Series Aggregation

Segment-level retrieval returns nearest neighbors for each query segment independently. However, detector selection in RAMSAD is performed at the level of full time series, since detector-performance profiles are stored for each training series rather than for individual segments. We therefore aggregate the segment-level neighbors into a ranked list of candidate series. The aggregation procedure counts how consistently a training series is retrieved across different query segments and combines this frequency information with the average retrieval distance of its matched segments. This favors training series that are both close to the query in representation space and repeatedly retrieved across multiple parts of the query series.

Algorithm 2 describes this aggregation step. For each retrieved segment neighbor, we record its source time series and associated retrieval distance. For every candidate source series, RAMSAD computes a base distance score by averaging all distances assigned to that series, and a frequency score measuring the fraction of query segments for which the series appears among the retrieved neighbors. The final aggregation score rescales the average distance by the inverse of the frequency score, so that series retrieved more consistently across the query receive a better rank. The top-ranked series are then passed to the detector-selection step.

Algorithm 1 Semantic Alignment Training for RAMSAD

Input: dataset $\mathfrak{D} = \{T_1, \ldots, T_{|\mathfrak{D}|}\}$, dataset descriptions $\{c_i\}_{i=1}^{|\mathfrak{D}|}$, time-series encoder f_{ts}, text encoder f_{text}, projection heads $g_{\theta^{\text{ts}}}$ and $g_{\theta^{\text{desc}}}$, segment length ℓ, similarity threshold γ, maximum number of epochs $E_{\max}$, batch size b, learning rate η

Output: learned projection parameters $\theta^{\text{ts}}, \theta^{\text{desc}}$

construct the set of segment–description pairs

$$\mathcal{P} = \left\{ \left(T_i^{(t,\ell)}, c_i \right) \ : \ T_i \in \mathfrak{D},\ T_i^{(t,\ell)} \text{ is a segment extracted from } T_i \right\}$$

for epoch $= 1, \ldots, E_{\max}$ **do**

  for each mini-batch $B = \{(T_{i_u}^{(t_u,\ell)}, c_{i_u})\}_{u=1}^{b} \subset \mathcal{P}$ **do**

    encode time-series segments

$$z_u^{\text{ts}} = f_{\text{ts}}\left(T_{i_u}^{(t_u,\ell)} \right), \qquad u = 1, \ldots, b$$

    encode descriptions using the frozen text encoder

$$z_u^{\text{desc}} = f_{\text{text}}(c_{i_u}), \qquad u = 1, \ldots, b$$

    project both modalities into the shared space

$$\tilde{z}_u^{\text{ts}} = g_{\theta^{\text{ts}}}\left(z_u^{\text{ts}} \right), \qquad \tilde{z}_u^{\text{desc}} = g_{\theta^{\text{desc}}}\left(z_u^{\text{desc}} \right)$$

    compute the in-batch description similarity matrix

$$A_{uv} = \left\langle z_u^{\text{desc}}, z_v^{\text{desc}} \right\rangle$$

    build the multi-positive target matrix $P \in \{0,1\}^{b \times b}$

$$P_{uv} = \mathbf{1}[A_{uv} \geq \gamma]$$

    mask pairs coming from the same source series

    compute the cross-modal similarity matrix

$$S_{uv} = \left\langle \tilde{z}_u^{\text{ts}}, \tilde{z}_v^{\text{desc}} \right\rangle$$

    compute the symmetric cross-modal alignment loss $\mathcal{L}(S, P)$

    update $\theta^{\text{ts}}, \theta^{\text{desc}}$ using $\nabla \mathcal{L}$

return $\theta^{\text{ts}}, \theta^{\text{desc}}$

B.4 RAMSAD Model Selection

After aggregating segment-level retrieval results into a ranked list of similar training series, RAMSAD uses the retrieved neighborhood to select a detector for the query series. The key assumption is that time series that are similar in the chosen representation space are likely to favor the same anomaly detector. Rather than training a separate classifier or regressor to predict the best detector, RAMSAD directly reuses the precomputed detector-performance profiles of the retrieved training series.

Algorithm 3 summarizes the full model-selection procedure. The query series is first segmented, encoded, and used to retrieve nearest-neighbor segments from the knowledge base. These segment-level neighbors are then aggregated into a ranked list of candidate training series using Algorithm 2. For each detector in the candidate pool, RAMSAD averages its stored performance values over the retrieved series and selects the detector with the highest aggregated score. This produces a non-parametric detector recommendation that can adapt to changes in the knowledge base or detector pool without retraining a selector model.

Algorithm 2 Segment-to-Series Aggregation

Input: query series T_q; retrieved segment neighbors $\{T_{j_{u,r}}^{(t_{u,r},\ell)}\}_{r=1}^{k}$ and distances $\{d_{u,r}\}_{r=1}^{k}$ for each query segment $T_q^{(t_u,\ell)}$, $u = 1, \ldots, m$; small constant $\epsilon > 0$; retrieval size k

Output: ranked candidate series set $\mathrm{top}_k(T_q)$

for all query segments $T_q^{(t_u,\ell)}$, $u = 1, \ldots, m$ **do**
  for all retrieved neighbors $T_{j_{u,r}}^{(t_{u,r},\ell)}$, $r = 1, \ldots, k$ **do**
    append $d_{u,r}$ to $\mathcal{A}(T_{j_{u,r}})$
for all candidate series T_j such that $\mathcal{A}(T_j) \neq \emptyset$ **do**
  compute $\mathrm{base}(T_j) = \frac{1}{|\mathcal{A}(T_j)|} \sum_{d \in \mathcal{A}(T_j)} d$
  compute $\mathrm{freq}(T_j) = \frac{1}{m} \Big| \{\, u \in \{1, \ldots, m\} : \exists r \in \{1, \ldots, k\} \text{ s.t. } j_{u,r} = j \,\} \Big|$
  compute $\mathrm{score}(T_j) = \frac{1}{\max(\epsilon, \mathrm{freq}(T_j))} \cdot \mathrm{base}(T_j)$
rank candidate series $T_j \in \mathfrak{D}$ in ascending order of $\mathrm{score}(T_j)$
return the top-k ranked series, denoted by $\mathrm{top}_k(T_q)$

Algorithm 3 RAMSAD Model Selection

Input: query series T_q, retrieval index $\mathcal{I}$, encoder f_{ts}, segment length ℓ, retrieval size k, detector pool $\mathcal{B}$, performance table $\mathbf{V}$

Output: selected detector $\widehat{D}(T_q)$

segment T_q into $\{T_q^{(t_u,\ell)}\}_{u=1}^{m}$
zero-pad the last segment if needed
for all segments $T_q^{(t_u,\ell)}$ **do**
  encode $T_q^{(t_u,\ell)}$ with f_{ts}
  retrieve its top-k nearest database segments and their distances from $\mathcal{I}$
aggregate the retrieved segments into a ranked series list $\mathrm{top}_k(T_q)$ using Algorithm 2
for all $D \in \mathcal{B}$ **do**
  $\bar{V}_D(T_q) = \frac{1}{|\mathrm{top}_k(T_q)|} \sum_{T \in \mathrm{top}_k(T_q)} \mathbf{V}(T, D)$
$\widehat{D}(T_q) = \arg\max_{D \in \mathcal{B}} \bar{V}_D(T_q)$
return $\widehat{D}(T_q)$

B.5 System Complexity and Scalability

Asymptotic retrieval cost. We analyze the computational cost of RAMSAD by comparing exact segment retrieval with an HNSW-based implementation. Let n denote the number of indexed knowledge-base segments, d the embedding dimensionality, b the number of query segments processed jointly, k_{seg} the number of retrieved segment neighbors per query segment, and k the number of final retrieved series [2]. For HNSW, we use deg for the maximum graph degree, ef for the search candidate list size, ef_c for the construction candidate list size, and $L \approx \log_{deg} n$ for the number of graph layers. The dominant cost in RAMSAD comes from segment retrieval. With exact search, the index consists of the segment embedding matrix, requiring $\mathcal{O}(nd)$ memory, together with the detector-performance matrix, which requires $\mathcal{O}(|\mathfrak{D}|\, M)$ memory. Querying a batch of b segments requires computing all pairwise similarities against the indexed segments, followed by top-k_{seg} selection, for a cost dominated by $\mathcal{O}(bnd)$. By contrast, HNSW adds an offline indexing stage and a small graph-memory overhead, but avoids scanning all n indexed segments at query time by performing graph traversal over a bounded candidate set.

Cost of aggregation and recommendation. The remaining stages of RAMSAD are comparatively lightweight. After retrieval, the segment-to-series aggregation step processes at most bk_{seg} retrieved items, maps them to source series, and computes the score introduced in Section 3. Its complexity is therefore linear in the size of the retrieval output and negligible relative to nearest-neighbor search. The final detector recommendation step is smaller still, since it only averages precomputed

[2] $k < k_{seg}$ only if two or more query segments retrieve database segments that correspond to the same source series, otherwise $k = k_{seg}$

Table 3: Asymptotic complexity of exact retrieval and HNSW within RAMSAD. Here, n is the number of indexed training segments, $|\mathfrak{D}|$ the number of indexed training series, M the number of candidate detectors, d the embedding dimensionality, b the number of query segments processed jointly, and k_{seg} the segment-level retrieval size.

Metric		Exact	HNSW
Time	Build	—	$\mathcal{O}(nd(\mathrm{ef}_c + deg\, L))$
	Query	$\mathcal{O}(bnd + bn + bk_{seg} \log k_{seg})$	$\mathcal{O}(b(\mathrm{ef_s}\, d + \mathrm{ef_s} \log \mathrm{ef_s}))$
Space	Persistent	$\mathcal{O}(nd + \lvert\mathfrak{D}\rvert\, M)$	$\mathcal{O}(nd + n\ deg\ (L+1) + \lvert\mathfrak{D}\rvert\, M)$
	Query	$\mathcal{O}(bd + bn + bk_{seg})$	$\mathcal{O}(bd + b\,\mathrm{ef} + bk_{seg})$

detector-performance values over the top-k retrieved series. Overall, the computational complexity of RAMSAD is governed by the retrieval backend rather than by aggregation or detector selection. The theoretical complexities are shown in Table 3.

C Similarity Spaces: Performance and Runtime Analysis

C.1 Parameter Sensitivity with Respect to k

Figure 14 provides the main visual summary of how performance varies with k across the different similarity spaces and TSFM backbones. Overall, the results show that RAMSAD is relatively stable with respect to the choice of k, although the magnitude of this sensitivity differs across backbones and between ID and OOD settings. In the ID setting, performance varies only moderately across k, with a maximum difference of roughly 0.04 across the tested configurations. This variability tends to decrease as the number of selected models increases, indicating that larger ensembles reduce the sensitivity to the retrieval size. A common pattern is that intermediate values of k yield the strongest results, whereas very small or very large values are slightly less effective.

In the OOD setting, the variation across k is somewhat larger, reaching about 0.07 in the single-model case, but this difference also shrinks as the number of selected models increases. In contrast to the ID case, larger values of k tend to be consistently more favorable in OOD, suggesting that broader neighborhoods help stabilize model recommendation under domain shift. This behavior is particularly visible for the aligned representations, which generally exhibit clearer gains in OOD than in ID.

Among the evaluated backbones, `Chronos2` [4] and `Chronos2_aligned` provide the strongest overall behavior, with alignment yielding the clearest gains in the OOD case. The other backbones—`Flowstate` [26], `Mantis` [20], `Moirai` [62], `Moment` [25], `TimesFM` [18], `Tirex` [7], and `Toto` [17]—follow the same broad pattern, but with different levels of sensitivity to k and different benefits from alignment. We therefore conclude that RAMSAD is robust to the choice of k, while still benefiting from moderate-to-larger neighborhoods, especially in OOD settings, and that it consistently remains stronger than the best trainable selector baseline across this parameter range.

C.2 ID and OOD Performance Tables

Tables 4 and 5 report the complete numerical results corresponding to the parameter-sensitivity analysis Figure 14. For each similarity space, we report the mean, median, and standard deviation of the evaluation scores across different values of k. The tables are organized so that raw and aligned variants of the same backbone are grouped together and separated from other backbones by double horizontal lines. This grouping makes it possible to compare, for each backbone, the effect of semantic alignment under the same retrieval setting.

Within each raw/aligned backbone group, bold values indicate the better result for a given statistic and value of k. For the mean and median rows, larger values are better, whereas for the standard-deviation row, smaller values are better. Underlined values indicate the best overall value in the entire table for each statistic: the highest mean, the highest median, and the lowest standard deviation.

Table 4: In-distribution. Bold per k within each raw/aligned pair: higher mean and median, lower std. Underline marks the best value in the full table for each stat (max mean, max median, min std) across all methods and k.

Method	Stat	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
Chronos2	Mean	0.7112	0.7117	0.7182	0.7329	**0.7350**	**0.7463**	**0.7422**	0.7253	**0.7216**	0.7147
	Median	0.8565	0.8552	0.8599	0.8854	**0.8780**	**0.8925**	0.8925	**0.8891**	**0.8768**	0.8565
	Std	0.3315	0.3329	0.3291	0.3198	**0.3188**	**0.3105**	**0.3177**	0.3295	**0.3274**	0.3270
Chronos2_aligned	Mean	**0.7234**	**0.7233**	**0.7431**	**0.7457**	0.7338	0.7384	0.7374	**0.7260**	0.7200	**0.7285**
	Median	**0.8854**	**0.8959**	**0.9024**	**0.8925**	0.8750	0.8854	**0.8947**	0.8867	0.8750	**0.8867**
	Std	**0.3300**	**0.3324**	**0.3178**	**0.3126**	0.3199	0.3173	0.3206	**0.3278**	0.3294	**0.3269**
Flowstate	Mean	0.6263	0.6551	0.6724	0.6916	0.6911	0.6961	0.7032	0.6981	0.6983	0.6953
	Median	0.7368	0.8043	0.8359	0.8334	0.8270	0.8410	**0.8509**	0.8453	0.8472	0.8410
	Std	0.3640	0.3659	0.3612	0.3411	0.3369	0.3404	0.3404	0.3417	0.3405	0.3382
Flowstate_aligned	Mean	**0.6408**	**0.6725**	**0.6888**	**0.7049**	**0.7099**	**0.7129**	**0.7118**	**0.7125**	**0.7093**	**0.7083**
	Median	**0.7905**	**0.8174**	**0.8442**	**0.8445**	**0.8442**	**0.8496**	**0.8509**	**0.8580**	**0.8555**	**0.8552**
	Std	**0.3630**	**0.3536**	**0.3507**	**0.3325**	**0.3307**	**0.3327**	**0.3315**	**0.3362**	**0.3333**	**0.3321**
Mantis	Mean	0.6445	**0.6718**	**0.6817**	**0.6896**	**0.6606**	**0.6806**	0.6779	**0.6759**	**0.6669**	**0.6700**
	Median	0.7982	0.8096	**0.8335**	**0.8212**	**0.8017**	**0.8239**	**0.8270**	**0.8270**	**0.8169**	**0.8174**
	Std	0.3677	**0.3481**	**0.3456**	0.3393	0.3539	0.3433	0.3464	0.3507	0.3539	**0.3485**
Mantis_aligned	Mean	**0.6601**	0.6701	0.6732	0.6809	0.6592	0.6780	**0.6824**	0.6749	0.6655	0.6602
	Median	**0.8121**	**0.8136**	0.8212	0.8092	0.8007	0.8169	0.8212	**0.8270**	**0.8169**	0.8096
	Std	**0.3632**	0.3498	0.3475	**0.3371**	**0.3473**	**0.3378**	**0.3397**	**0.3489**	**0.3528**	0.3521
Moirai	Mean	**0.6588**	**0.6845**	**0.6960**	**0.7116**	0.6895	**0.6867**	0.6743	0.6697	0.6769	**0.6834**
	Median	**0.8092**	**0.8466**	**0.8552**	**0.8528**	**0.8466**	**0.8419**	0.8419	0.8227	0.8320	**0.8410**
	Std	**0.3616**	0.3509	0.3481	**0.3288**	0.3458	**0.3393**	0.3499	0.3525	0.3460	**0.3442**
Moirai_aligned	Mean	0.6486	0.6701	0.6925	0.6932	**0.6945**	0.6823	**0.6961**	**0.6823**	**0.6827**	0.6833
	Median	0.7930	0.8043	0.8453	0.8419	0.8453	0.8418	**0.8559**	**0.8438**	**0.8419**	0.8320
	Std	0.3655	**0.3469**	**0.3452**	0.3419	**0.3409**	0.3441	**0.3419**	**0.3478**	**0.3448**	0.3457
Moment	Mean	**0.7157**	0.7180	0.7084	**0.7290**	**0.7171**	**0.7101**	**0.7103**	**0.7152**	**0.7087**	**0.7100**
	Median	**0.8590**	0.8584	**0.8768**	**0.8765**	0.8589	**0.8582**	**0.8589**	**0.8637**	**0.8584**	**0.8580**
	Std	**0.3401**	0.3345	0.3492	0.3309	**0.3298**	**0.3329**	**0.3373**	**0.3344**	**0.3405**	**0.3381**
Moment_aligned	Mean	0.7064	**0.7315**	**0.7160**	0.7247	0.7109	0.7033	0.7042	0.7079	0.7050	0.6976
	Median	0.8543	**0.8637**	0.8750	0.8590	**0.8590**	0.8543	0.8543	0.8565	0.8555	0.8509
	Std	0.3467	**0.3276**	**0.3392**	**0.3282**	0.3385	0.3378	0.3387	0.3412	0.3407	0.3437
TimesFM	Mean	0.6809	0.6846	0.6904	0.6907	0.7084	0.7118	0.7143	0.6983	0.6855	0.6969
	Median	0.8142	0.8142	0.8420	0.8295	0.8466	**0.8496**	**0.8565**	0.8453	**0.8453**	**0.8466**
	Std	0.3456	0.3424	0.3403	0.3380	0.3272	0.3232	0.3277	0.3354	0.3423	0.3355
TimesFM_aligned	Mean	**0.6961**	**0.7053**	**0.7075**	**0.7096**	**0.7118**	**0.7124**	**0.7146**	**0.7166**	**0.7083**	**0.7047**
	Median	**0.8420**	**0.8322**	**0.8555**	**0.8431**	**0.8496**	0.8466	0.8496	**0.8496**	0.8438	0.8463
	Std	**0.3438**	**0.3293**	**0.3331**	**0.3253**	**0.3252**	**0.3223**	**0.3218**	**0.3208**	**0.3245**	**0.3287**
Tirex	Mean	**0.6512**	0.6588	**0.6681**	0.6753	0.6714	**0.6652**	**0.6586**	0.6573	0.6649	0.6701
	Median	**0.8095**	**0.8136**	**0.8295**	0.8270	0.8212	0.8001	**0.8001**	0.8017	0.8249	**0.8332**
	Std	**0.3652**	**0.3564**	**0.3563**	**0.3472**	**0.3433**	**0.3439**	**0.3508**	0.3541	0.3497	0.3494
Tirex_aligned	Mean	0.6337	**0.6600**	0.6453	**0.6782**	**0.6812**	0.6621	0.6518	**0.6616**	**0.6770**	**0.6707**
	Median	0.7989	0.8092	0.8016	**0.8295**	**0.8442**	**0.8286**	0.7989	**0.8174**	**0.8322**	0.8249
	Std	0.3726	0.3576	0.3633	0.3479	0.3531	0.3566	0.3547	**0.3513**	**0.3467**	**0.3481**
Toto	Mean	**0.6786**	**0.7139**	**0.7087**	**0.7266**	0.7159	0.7233	0.7130	0.7067	0.6973	0.6908
	Median	**0.8420**	**0.8555**	**0.8709**	**0.8637**	0.8709	**0.8719**	0.8709	0.8584	0.8528	0.8528
	Std	**0.3550**	0.3362	0.3391	0.3225	0.3299	**0.3190**	0.3277	0.3351	0.3383	0.3433
Toto_aligned	Mean	0.6719	0.7116	0.7050	0.7252	**0.7282**	**0.7301**	**0.7271**	**0.7108**	**0.6996**	**0.7039**
	Median	0.8401	0.8466	0.8528	0.8599	**0.8854**	**0.8719**	**0.8854**	**0.8599**	**0.8552**	**0.8552**
	Std	0.3639	**0.3295**	**0.3382**	**0.3221**	**0.3253**	0.3194	**0.3246**	**0.3324**	**0.3380**	**0.3351**
Original (Euc.-seg.)	Mean	0.6579	0.6694	0.6554	0.6565	0.6752	0.6532	0.6460	0.6123	0.6313	0.6148
	Median	0.8092	**0.8096**	0.7903	0.7991	0.8282	0.7905	0.7733	0.7361	0.7502	0.7003
	Std	0.3598	0.3501	0.3517	0.3467	**0.3423**	**0.3474**	0.3472	0.3659	0.3536	0.3542
Original (Euc. - full)	Mean	0.5853	0.5913	0.5937	0.5775	0.5500	0.5582	0.5547	0.5528	0.5454	0.5173
	Median	0.6936	0.7427	0.7148	0.6573	0.6055	0.6185	0.6064	0.6055	0.6161	0.5615
	Std	0.3808	0.3853	0.3823	0.3780	0.3831	0.3849	0.3829	0.3902	0.3855	0.3865
Original (DTW-MESSI)	Mean	**0.6765**	**0.6880**	**0.6884**	**0.6861**	**0.6812**	**0.6775**	**0.6750**	**0.6643**	**0.6626**	**0.6506**
	Median	**0.8333**	0.8092	**0.8333**	**0.8333**	**0.8373**	**0.8351**	**0.8286**	**0.8212**	**0.8286**	**0.8096**
	Std	**0.3554**	**0.3335**	**0.3350**	**0.3368**	0.3440	0.3481	**0.3435**	**0.3480**	**0.3532**	**0.3539**

Table 5: Out-of-distribution. Bold per k within each raw/aligned pair: higher mean and median, lower std. Underline marks the best value in the full table for each stat (max mean, max median, min std) across all methods and k.

Method	Stat	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
Chronos2	Mean	0.4819	0.4847	0.4900	0.5178	0.5427	0.5568	0.5721	0.5695	0.5917	**0.5927**
	Median	0.4453	0.4477	0.4583	0.5837	0.6434	0.6435	0.6764	0.6764	**0.7220**	0.6852
	Std	0.3844	0.3702	0.3834	0.3884	0.3809	0.3756	0.3683	0.3721	0.3704	0.3655
Chronos2_aligned	Mean	**0.5390**	**0.5162**	**0.5745**	**0.5768**	**0.5649**	**0.5822**	**0.5792**	**0.5899**	**0.5934**	0.5906
	Median	**0.6073**	**0.5291**	**0.6794**	**0.6578**	**0.6578**	**0.7220**	**0.6852**	**0.7124**	**0.7220**	**0.6932**
	Std	0.3819	0.3771	0.3817	0.3784	0.3844	0.3805	0.3803	0.3764	0.3750	0.3757
Flowstate	Mean	0.4932	0.4876	**0.5237**	**0.5088**	**0.5311**	**0.5260**	**0.5380**	0.5160	0.5217	**0.5383**
	Median	0.5228	0.5307	**0.6032**	**0.5707**	**0.6064**	**0.6185**	**0.6435**	0.5850	0.6000	**0.6360**
	Std	0.3872	0.3860	0.3918	0.3874	0.3852	0.3861	0.3879	0.3846	0.3865	0.3835
Flowstate_aligned	Mean	**0.4962**	**0.4975**	0.5058	0.4961	0.5157	0.5138	0.5335	**0.5360**	**0.5386**	0.5374
	Median	**0.5392**	**0.5409**	0.5409	0.5300	0.5956	0.5850	0.6236	**0.6366**	**0.6452**	0.6236
	Std	0.3856	0.3922	0.3932	0.3873	0.3865	0.3914	0.3930	0.3922	0.3892	0.3848
Mantis	Mean	0.3751	0.4001	**0.4208**	**0.4058**	0.3880	**0.3901**	0.3960	**0.4131**	**0.4046**	0.3846
	Median	0.2032	0.2948	**0.3559**	**0.3245**	**0.2889**	**0.2889**	0.2962	**0.3308**	**0.3005**	0.2671
	Std	0.3701	0.3709	0.3652	0.3610	0.3544	0.3514	0.3537	0.3579	0.3569	0.3542
Mantis_aligned	Mean	**0.3779**	**0.4171**	0.3990	0.3940	**0.3886**	0.3855	**0.4111**	0.3903	0.3985	**0.4041**
	Median	**0.2808**	**0.3333**	0.2953	0.3053	0.2779	0.2835	**0.3360**	0.2892	0.2936	**0.3005**
	Std	0.3612	0.3683	0.3610	0.3571	0.3554	0.3513	0.3537	0.3499	0.3566	0.3561
Moirai	Mean	0.4676	0.4923	0.5070	0.5136	**0.5257**	0.5363	**0.5383**	0.5256	0.4938	0.4956
	Median	0.4091	0.4302	0.5092	0.5466	**0.6158**	0.6258	**0.6258**	0.5956	0.4765	0.5092
	Std	0.3798	0.3800	0.3843	0.3900	0.3895	0.3880	0.3872	0.3935	0.3973	0.3951
Moirai_aligned	Mean	**0.4952**	**0.4975**	**0.5132**	**0.5161**	0.5176	**0.5377**	0.5300	**0.5305**	**0.5060**	**0.5198**
	Median	**0.4477**	**0.4453**	**0.5334**	**0.5545**	0.5697	**0.6353**	0.5956	**0.6158**	**0.5412**	**0.5778**
	Std	0.3861	0.3788	0.3823	0.3928	0.3901	0.3888	0.3884	0.3885	0.3911	0.3859
Moment	Mean	**0.4711**	**0.4578**	0.4434	0.4504	0.4419	0.4443	0.4529	0.4537	0.4592	0.4453
	Median	**0.4399**	**0.4358**	0.4042	**0.4091**	0.3687	0.3757	0.4001	**0.4042**	**0.4215**	0.3761
	Std	0.3760	0.3653	0.3635	0.3551	0.3549	0.3491	0.3539	0.3516	0.3509	0.3489
Moment_aligned	Mean	0.4581	0.4575	**0.4678**	**0.4523**	**0.4520**	**0.4559**	**0.4581**	**0.4611**	**0.4602**	**0.4582**
	Median	0.4023	0.3946	**0.4073**	0.3983	**0.4007**	**0.4023**	**0.4027**	0.4027	0.4027	**0.4027**
	Std	0.3752	0.3725	0.3617	0.3525	0.3527	0.3515	0.3511	0.3513	0.3531	0.3498
TimesFM	Mean	**0.4772**	0.4541	0.4738	0.4660	0.4731	0.4652	0.4693	0.4654	0.4830	0.4916
	Median	0.3757	0.3687	0.3977	0.3755	0.3946	0.3777	0.3910	0.3910	0.4341	0.4219
	Std	0.3737	0.3697	0.3634	0.3714	0.3602	0.3555	0.3551	0.3592	0.3614	0.3609
TimesFM_aligned	Mean	0.4525	**0.4636**	**0.4876**	**0.4898**	**0.4922**	**0.4879**	**0.4854**	**0.4883**	**0.4994**	**0.4965**
	Median	**0.3761**	**0.4041**	**0.4437**	**0.5092**	**0.5092**	**0.4903**	**0.5063**	**0.5256**	**0.5307**	**0.5269**
	Std	0.3702	0.3556	0.3676	0.3662	0.3608	0.3691	0.3631	0.3656	0.3665	0.3723
Tirex	Mean	**0.4985**	**0.4961**	0.4705	**0.4874**	**0.4941**	**0.4960**	**0.4871**	**0.4921**	**0.4966**	**0.5068**
	Median	**0.4852**	**0.4904**	0.4023	0.4558	0.4551	**0.5198**	0.4154	**0.4659**	**0.4854**	**0.5258**
	Std	0.3731	0.3699	0.3678	0.3705	0.3728	0.3719	0.3740	0.3761	0.3702	0.3671
Tirex_aligned	Mean	0.4698	0.4705	**0.4794**	0.4784	0.4812	0.4906	0.4815	0.4751	0.4790	0.4826
	Median	0.4011	0.3983	**0.4109**	**0.4568**	**0.4568**	0.4663	**0.4188**	0.4109	0.4390	0.4185
	Std	0.3776	0.3813	0.3805	0.3835	0.3828	0.3814	0.3863	0.3864	0.3836	0.3835
Toto	Mean	**0.4447**	0.4455	**0.4459**	0.4360	0.4196	**0.4456**	0.4495	**0.4527**	**0.4578**	**0.4545**
	Median	0.3376	0.3826	**0.4023**	0.3477	0.3103	**0.3777**	**0.3790**	**0.3826**	**0.4207**	**0.3983**
	Std	0.3866	0.3768	0.3743	0.3768	0.3725	0.3702	0.3742	0.3747	0.3713	0.3778
Toto_aligned	Mean	0.4383	**0.4542**	0.4331	**0.4464**	**0.4647**	0.4332	**0.4498**	0.4429	0.4548	0.4517
	Median	**0.3440**	**0.4104**	0.3512	**0.3738**	**0.4041**	0.3168	0.3533	0.3402	0.3738	0.3738
	Std	0.3791	0.3759	0.3772	0.3710	0.3792	0.3821	0.3841	0.3769	0.3821	0.3848
Original (Euc.-seg.)	Mean	0.3862	0.3952	0.3811	0.3913	0.3960	0.4098	0.4141	0.4133	0.4232	0.4218
	Median	0.2744	0.2953	0.2840	0.2943	0.3168	0.3303	**0.3333**	0.3400	**0.3694**	0.3482
	Std	**0.3637**	**0.3523**	**0.3473**	**0.3452**	**0.3474**	**0.3511**	**0.3534**	**0.3538**	**0.3602**	0.3610
Original (Euc. - full)	Mean	0.3707	0.4392	0.4178	0.4151	0.3990	**0.4256**	0.4140	0.3971	0.4039	0.3941
	Median	0.2046	0.3510	0.3006	0.2983	0.2892	**0.3561**	**0.3333**	0.3115	0.3245	0.3035
	Std	0.3645	0.3789	0.3705	0.3739	0.3625	0.3693	0.3629	0.3567	0.3628	**0.3594**
Original (DTW-MESSI)	Mean	**0.4229**	**0.4396**	**0.4276**	**0.4304**	**0.4220**	0.4197	**0.4215**	**0.4295**	**0.4365**	**0.4354**
	Median	**0.3482**	**0.3579**	**0.3606**	**0.4023**	**0.3385**	0.3128	0.3192	**0.3482**	0.3588	**0.3651**
	Std	0.3696	0.3776	0.3668	0.3670	0.3702	0.3780	0.3762	0.3769	0.3782	0.3786

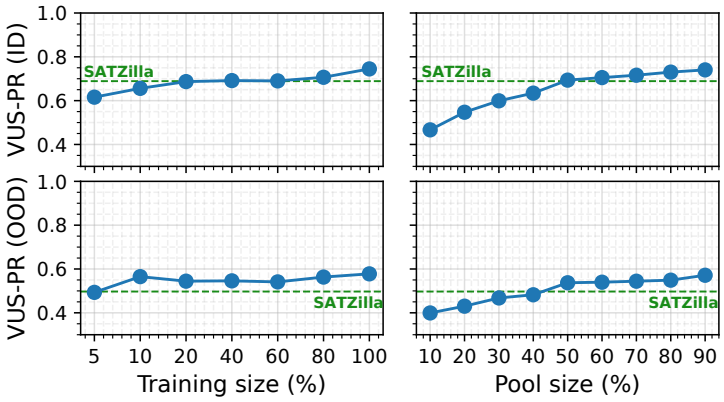


Figure 13: Average VUS-PR achieved by RAMSAD when randomly reducing the size of the training set (left) and detector pool (right) in ID (top) and OOD (bottom) scenarios.

Table 6: Retrieval-pipeline wall-clock time (ms) across TSFM-based similarity spaces. All methods are evaluated on the same query and database pool ($N_{\text{test}} = 1383$, $N_{\text{train}} = 3264$, $N_{\text{pairs}} = 4{,}514{,}112$). Aligned variants apply the learned 768-dimensional semantic projection before retrieval.

Method	**Load**	**Proj.**	**Search**	**Agg.**	**Total**
Mantis	10.0	–	44.6	56.2	110.8
Tirex	12.3	–	47.4	40.2	99.9
Flowstate	14.1	–	44.8	60.8	119.7
Moirai	14.6	–	56.0	53.0	123.6
Toto	14.8	–	61.0	50.7	126.5
Moment	17.7	–	53.9	54.1	125.7
Chronos-2	16.4	–	69.0	52.6	138.0
TimesFM	25.2	–	64.3	55.4	144.9
Mantis_aligned	10.0	18.7	44.6	56.2	129.5
Tirex_aligned	12.3	18.1	50.3	57.6	138.3
Flowstate_aligned	14.1	17.0	51.5	56.4	139.0
Moirai_aligned	14.6	21.3	49.5	56.2	141.6
Toto_aligned	16.8	20.5	51.6	55.9	144.8
Moment_aligned	17.7	23.2	56.8	54.0	151.7
Chronos-2_aligned	14.8	25.3	48.9	40.1	129.1
TimesFM_aligned	25.6	23.2	49.8	55.9	154.5

C.3 Retrieval Runtime Across TSFM Similarity Spaces

Tables 6 and 7 compare the wall-clock time of the RAMSAD retrieval pipeline across the similarity spaces considered in this work. The goal here is to isolate the computational cost induced by the choice of representation. All methods are therefore evaluated on a common query–database pool, and the reported times should be interpreted as single-run retrieval costs.

Table 6 reports the runtime decomposition for TSFM-based and aligned TSFM-based similarity spaces. A substantial part of the variation across backbones is explained by the dimensionality of the embeddings they produce, since both loading and nearest-neighbor search scale with the size of the representation. In particular, backbones with smaller native embedding dimensions tend to be faster: `Tirex` is the fastest (99.9 ms), followed by `Mantis` (110.8 ms), which uses the lowest-dimensional representation among the tested TSFMs, whereas `TimesFM`, which produces the largest embeddings, is the slowest (144.9 ms). The aligned variants introduce an additional projection step to a shared 768-dimensional space, which costs between 17 and 25 ms depending on the backbone. Despite this overhead, all aligned variants remain well below one second, showing that semantic alignment improves the representation space at only moderate additional cost. In particular, `Chronos-2_aligned`, which provides the strongest retrieval quality in our experiments, requires 129.1 ms and therefore remains practical.

Table 7: Retrieval-pipeline wall-clock time of original similarity spaces on the same query and database pool.

Method	Total time
Segment Euclidean + aggregation	549.2 ms
Full-series Euclidean (truncated)	1.20 s
DTW via MESSI ($w = 10\%$)	11 h 43 m

Table 8: OOD retrieval wall-clock time per domain for the three raw-distance similarity spaces.

Domain	Seg. Eucl.	Full-series Eucl.	DTW (MESSI)
Environment	189.1 ms	106.0 ms	30 m 46 s
Facility	575.0 ms	1.19 s	14 h 42 m
Finance	189.2 ms	126.3 ms	37 m 18 s
Human Activity	292.0 ms	557.0 ms	5 h 29 m
Medical	616.6 ms	1.24 s	12 h 29 m
Sensor	211.9 ms	235.2 ms	2 h 20 m
Synthetic	218.2 ms	467.1 ms	2 h 12 m
Traffic	184.0 ms	24.7 ms	1 m 56 s
Web Service	425.3 ms	1.05 s	6 h 08 m
Total	**2.90 s**	**5.00 s**	**44 h 32 m**

C.4 Retrieval Runtime of Original Similarity Space

Table 7 places these results in the context of the original similarity spaces. Segment-level Euclidean retrieval with aggregation requires 549.2 ms, while full-series Euclidean retrieval requires 1.20 s. By contrast, DTW retrieval via MESSI [47, 19][3] using a warping window of 10% is several orders of magnitude slower, requiring 11 h 43 m on the same pool. This confirms that, although DTW can be competitive in accuracy among original-space distances, its computational cost makes it impractical as a retrieval backend for RAMSAD.

Overall, these results reinforce the main conclusion of this section: embedding-based similarity spaces, and especially aligned TSFM spaces, provide the best practical accuracy–efficiency trade-off for RAMSAD.

D Experimental Setup and Benchmark Details

D.1 Dataset Details

For the evaluation of RAMSAD, we considered the TSB-AD [33] benchmark, which includes a comprehensive set of 870 curated univariate time series from 9 domains, which are summarized in Table 9. To ensure a fair comparison, we used the same train/test splits as recent model selection benchmarks [34]. The splits are the same for both ID and OOD scenarios. However, in the latter, for a given time series, automated solutions can only consider training series that do not belong to its domain. For example, for a test series belonging to the *Medical* category, automated solutions can only consider training series that belongs to the other 8 domains (e.g., a meta-learning approach can be trained only on this reduced training subset).

D.2 List of Candidate Detectors

Table 10 lists all the 32 anomaly detection approaches considered in our experimental evaluation. As it can be seen, the pool of available models covers all the different categories described in the literature review, and aligns with the one used in recent model selection benchmarks [34].

[3] `https://github.com/MChatzakis/DaiSy`

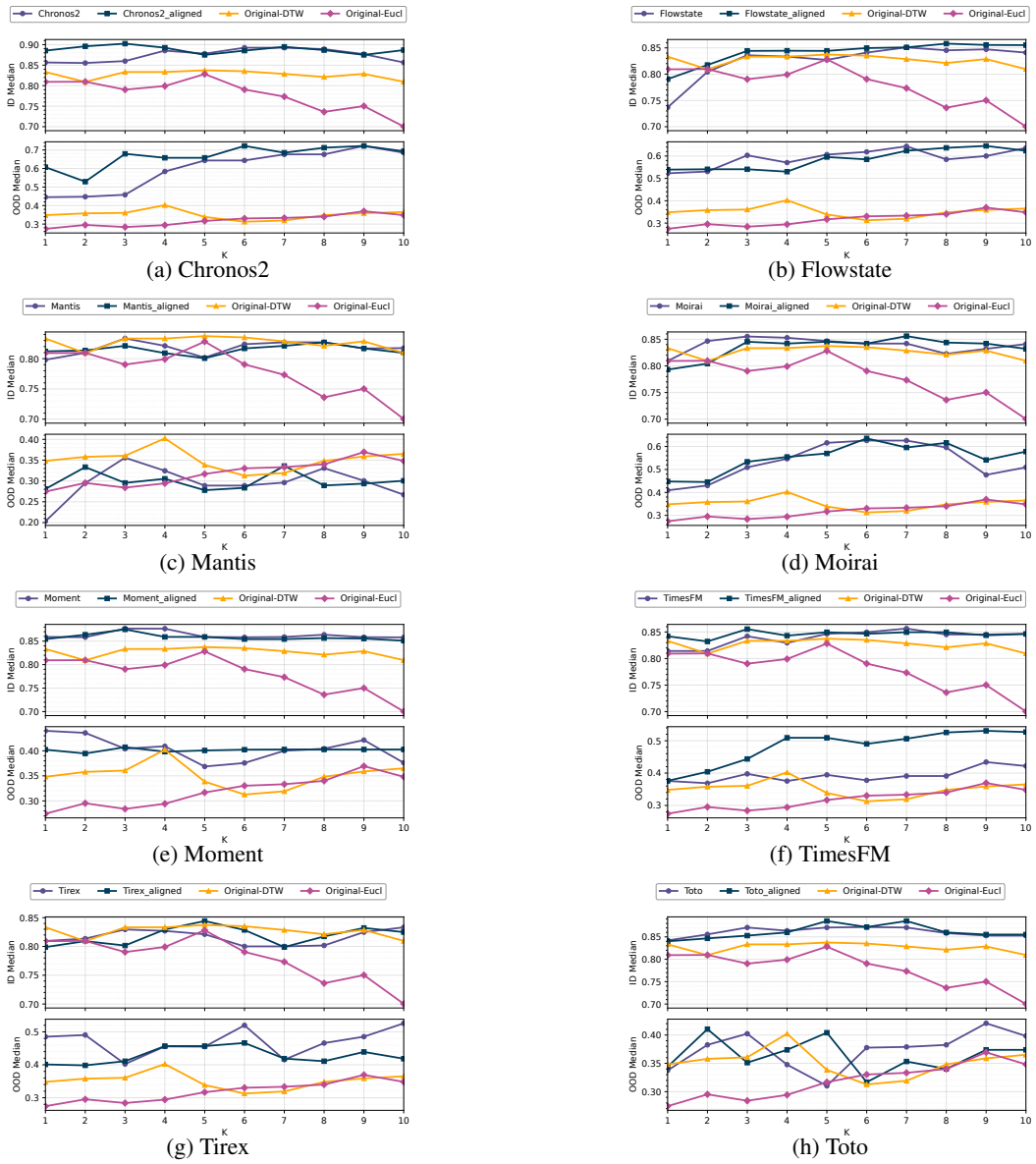


Figure 14: Sensitivity analysis with respect to parameter k across eight forecasting models. Each panel reports the model-specific response under identical evaluation conditions, allowing direct visual comparison of trend, variability, and stability.

D.3 Automated TSAD Baselines

Table 11 lists the 25 automated solutions that we compared as baselines. More specifically, we include 2 theoretical baselines, i.e., selecting always the best detector (Oracle) and selecting the best detector basing on a labeled portion of the target series. We also include 3 simple baselines (i.e., Random selection, Global best, and Foundation model). In addition to these baselines, we included a comprehensive set of state-of-the-art automated solutions, including 8 model selection approaches, 4 internal selection methods, 5 ensembles strategies and 3 model generation approaches.

D.4 Hardware Details

All experiments were performed on a machine running Linux kernel 5.4.0-212-generic. The node is equipped with two Intel(R) Xeon(R) Gold 6242R CPUs, providing 80 logical CPU cores across

Table 9: Dataset Domains.

Domain	Total	Training	Test
Environment	20	14	6
Facility	143	102	41
Finance	20	14	6
Human Activity	58	41	17
Medical	147	105	42
Sensor	44	32	12
Synthetic	122	87	35
Traffic	6	4	2
WebService	310	220	90
Total	**870**	**619**	**251**

2 sockets with 2 threads per core. The system has 503 GiB of main memory and an 8 GiB swap partition. GPU-accelerated experiments used a single NVIDIA Quadro RTX 6000 with 23 GiB of onboard memory.

E Additional Ablation Studies

E.1 Effect of Reducing the Training Size

In this part, we evaluate the performance of RAMSAD when reducing the size of the training set. Figure 13 shows the average VUS-PR achieved by RAMSAD with training size varying from 5% to 100% of the original set, comparing it with the performance of the best model selector (i.e., SATZilla) in ID and OOD scenarios. By considering the former scenario, we can see that RAMSAD can outperform SATZilla even with just the 80% of the training set, and that it achieves comparable scores even with sizes of 20%, 40%, and 60%. These results underscore the ability of RAMSAD of retrieving meaningful examples from the knowledge base, enabling effective model selection. This aspect is even more evident in the OOD scenario. As we can see, RAMSAD outperforms SATZilla even when using just the 10% of the original training set (i.e.,62 series).

E.2 Effect of Reducing the Pool Size of Detectors

Figure 13 also reports the performance of RAMSAD when randomly varying the pool of available detectors in ID and OOD scenarios. Each value reported represents the average across 10 repetitions to take into account the variability due to the random sampling. In both scenarios, we can see that RAMSAD is able to match or outperform SATZilla even when considering only half of the original pool (i.e., 16 detectors); note that SATZilla is shown with its performance on the full detector pool, so this comparison is conservative for RAMSAD, highlighting that, by retrieving meaningful examples, it allows to select an effective detector even in this more challenging scenario.

F Additional Discussions

F.1 Limitations

Although RAMSAD achieves strong performance for time-series anomaly detector selection, it still has several limitations.

Segmentation-induced information loss. To take advantage of TSFM representations, RAMSAD partitions each time series into fixed-length segments before retrieval. While this makes heterogeneous series compatible with the encoder, it may also discard information carried by longer-range temporal dependencies and by the global structure of the original series. As a result, similarity computed from segment-level representations may not fully reflect similarity at the full-series level. More generally, the segmentation strategy itself is fixed rather than optimized, and we do not yet study how to choose

Table 10: List of anomaly detection approaches included in our experimental evaluation.

Approach	Category	Description
Sub-LOF [13] LOF [13]	Distance-based Distance-based	Identifies density-based local outliers.
MatrixProfile [72]	Distance-based	Computes the distance between all subsequences.
KShapeAD [44]	Distance-based	Identifies the normal pattern based on the k-Shape clustering algorithm and computes distance-based anomaly scores.
SAND [10]	Distance-based	Incrementally updates a summary of normal patterns using subsequences from batches and evaluates anomaly scores based on distance.
Series2Graph [9]	Density-based	Transforms subsequences into nodes and their transition patterns into edges. Anomalies are identified based on irregular transition patterns.
SR [51]	Distance-based	Computes anomaly scores as the relative difference between saliency map values and their moving averages.
Sub-KNN [48]	Distance-based	Leverages the distance to the k-th nearest neighbor.
KMeansAD [70]	Distance-based	Measures the distance to the centroid of assigned cluster.
Sub-IForest [32] IForest [32]	Density-based Density-based	Isolates anomalies using random partitioning of the data space.
Sub-PCA [1]	Density-based	Projects data to a lower-dimensional hyperplane, with significant deviation from this plane indicating anomalies.
Sub-HBOS [23]	Density-based	Constructs a histogram for the data and uses the inverse of the height of the bin as the anomaly score of the data point.
Sub-OCSVM [56]	Density-based	Fits the dataset to find the normal data's boundary by maximizing the margin between the origin and the normal samples.
Sub-MCD [52]	Density-based	Leverages minimum covariance determinant
POLY [29]	Density-based	Uses local polynomial approximation.
AutoEncoder [53]	Prediction-based	Leverages reconstruction errors through encoding-decoding.
CNN [40]	Prediction-based	Leverages prediction errors through CNN.
LSTMAD [39]	Prediction-based	Leverages prediction errors through LSTM.
TranAD [60]	Prediction-based	Leverages self-conditioning and adversarial training.
AnomalyTransformer [67]	Prediction-based	Leverages Anomaly-Attention mechanism to compute association discrepancies.
OmniAnomaly [57]	Prediction-based	Leverages a Stochastic recurrent neural network.
USAD [6]	Prediction-based	Leverages adversely trained autoencoders.
Donut [66]	Prediction-based	Based on Variational Auto-Encoders (VAE).
TimesNet [63]	Prediction-based	Based on temporal 2D-variation modeling.
FITS [69]	Prediction-based	Interpolates time series in the frequency domain.
TimesFM [18]	Prediction-based	Pre-trains a decoder-only attention model with input patching.
MOMENT (ZS) [25]	Prediction-based	Uses a Zero Shot, pre-trained transformer model.
MOMENT (FT) [25]	Prediction-based	Uses a Fine-Tuned, pre-trained transformer model.
OFA [74]	Foundation model	Based on fine-tuned GPT-2 model.
Lag-Llama [49]	Foundation model	Leverages a decoder-only transformer architecture.
Chronos [3]	Foundation model	Tokenizes time series values using scaling and quantization.

segment boundaries or segment lengths so as to minimize retrieval distortion. Adaptive segmentation and encoders with longer contexts are natural directions to reduce this loss. Relatedly, we do not resample series, and therefore make no claim of invariance to the sampling rate.

Remaining gap to the oracle. RAMSAD provides a strong non-parametric baseline, but it does not fully close the gap to the theoretical oracle in either the ID or the OOD setting. As shown in Figure 1, there is still room for improvement in how retrieved neighbors are identified and how detector recommendations are transferred from them. Reducing this gap remains an important direction for future work.

Table 11: List of automated solutions compared with RAMSAD.

Approach	Category	Description
Oracle	Theoretical baseline	Selects the single best anomaly detector for each series
Supervised Selection (SS)	Theoretical baseline	Oracle-style upper bound: selects the best detector using labels of a portion of the target series, unavailable to the other methods
Global Best (GB)	Baseline	Global best: selects the detector with the best overall accuracy on the entire benchmark.
Random	Baseline	Randomly selects one of the detectors for each series.
Foundation model (FM) [18]	Baseline	Foundation model pre-trained on large scale datasets.
SATZilla [68]	Model Selection	Builds per-instance algorithm portfolios using empirical hardness models to predict performance and select the best SAT solver for each instance.
ARGOSMART [43]	Model Selection	Uses 1NN on dataset meta-features to select the model that performed best on the most similar training dataset.
ISAC [27]	Model Selection	Clusters datasets via meta-features and selects the best-performing model from the nearest cluster to the new dataset.
MetaOD [73]	Model Selection	Leverages collaborative filtering and matrix factorization on dataset meta-features to predict and rank candidate models for a new dataset.
MSAD [58]	Model Selection	Uses classification to map datasets to the best-performing anomaly detection model among a set of candidates.
UReg [41]	Model Selection	Learns a regression from dataset meta-features to latent SVD factors to predict model performance for novel datasets.
CFact [41]	Model Selection	Clusters configurations in latent space and trains cluster-specific regressors to predict performance and select the best model for known datasets.
Model Centrality (MC) [30]	Model Selection	Computes model centrality via Kendall's distances between anomaly score rankings to select models closest to their nearest neighbors.
Clustering Quality (CQ) [42]	Internal	Applies cluster validation metrics on thresholded anomaly scores to select models with better separation between normal/anomalous traces.
Unsupervised evaluation curves (UEC) [22]	Internal	Evaluates anomaly detectors using unsupervised mass-volume and excess-mass curves instead of labeled ROC/PR-based metrics.
Synthetic [15]	Internal	Injects artificial anomalies into simulated time series and selects the model with the best detection performance.
TSAdams [24]	Internal	Aggregates surrogate metric rankings of anomaly detectors using Kemeny/Borda methods to produce a more robust final model ranking.
AutoTSAD [55]	Ensemble	Generates synthetic anomalous time series to optimize detector configurations and ensemble their ranked anomaly scores into a final prediction.
Outlier Ensemble (OE) [2]	Ensemble	Uses an ensemble of all available detectors.
SELECT [50]	Ensemble	Applies unsupervised ensemble selection to choose a subset of detectors using vertical and horizontal consensus strategies for final aggregation.
Iterative Outlier Ensemble (IOE) [38]	Ensemble	Iteratively builds a pseudo–ground truth by selecting and refining anomaly score consensus to produce a final aggregated ensemble score.
HITS [38]	Ensemble	Recursively computes hub and centralities over anomaly score relations to rank and select more reliable models.
AutoOD-A [14]	Model Generation	Iteratively refines pseudo-labels via voting and learns to prune weak detectors using logistic regression and multi-view updates.
AutoOD-C [14]	Model Generation	Iteratively cleans noisy pseudo-labels by tracking early training losses to remove likely mislabeled instances and refine anomaly detection.
UADB [71]	Model Generation	Applies knowledge distillation to transfer multiple anomaly detectors' knowledge into a booster model that adapts and improves detection across datasets.

Restriction to univariate time series. Our current study focuses only on univariate time series. Although the overall retrieval-based framework could be extended to multivariate settings, such an extension is not explored here. Adapting RAMSAD to multivariate anomaly detection, including the design of suitable segment representations and similarity spaces, is therefore left for future work.

Temporal concept drift and streaming. Our experiments concern domain shift between static series: the query comes from a domain that is absent from the knowledge base. We do not evaluate temporal concept drift within a domain or a stream, nor non-stationarity within a single long series, such as an anomalous pattern that becomes normal over time. Nothing in the method precludes these cases, since RAMSAD retrieves per segment and aggregates segment matches into a series ranking, so later segments of an evolving series are embedded and retrieved afresh and the recommendation can change as they accumulate. We present this as a direction for future work, together with a streaming evaluation protocol, and not as a claim supported by our experiments.

Labeled knowledge base and queries far from every covered domain. RAMSAD requires a labeled knowledge base on which every candidate detector has been evaluated, the same supervision used by trained meta-learning selectors, and it does not apply where no labels exist at all. Our leave-one-domain-out protocol evaluates entirely unseen domains: no series, performance profile or description from the target domain enters the knowledge base or the alignment training. Missing descriptions are not a failure mode either, since a query never supplies one. What our experiments do not isolate is a query that is far from every indexed domain in representation space. Since what matters for selection is proximity in that space rather than shared domain membership, such a query may retrieve weakly related neighbors, and characterizing when the knowledge-base coverage is insufficient, and how to build a knowledge base that avoids it, is left for future work.

F.2 Broader Impacts

Improved model selection for time-series anomaly detection can support more reliable monitoring systems in domains such as healthcare, manufacturing, finance, energy, cybersecurity, and scientific monitoring. By reducing the need to run large detector pools, RAMSAD may also improve the efficiency and accessibility of anomaly-detection pipelines. At the same time, anomaly-detection systems can produce false alarms or missed detections, which may have negative consequences in high-stakes deployments if used without appropriate human oversight. The method could also be applied in monitoring or surveillance contexts, where privacy and governance considerations depend on the deployment setting. We therefore view RAMSAD as a decision-support component rather than a replacement for domain-specific validation and responsible deployment practices.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the scope of the paper: RAMSAD is a retrieval-based framework for time-series anomaly-detector selection that queries a knowledge base of previously observed instances, retrieves the most similar series, and transfers detector recommendations from their performance profiles. The stated claims are consistent with the paper's contributions, including the retrieval formulation, segment-to-series aggregation procedure, semantic alignment of time-series and textual representations, and empirical evaluation in both ID and OOD settings, as presented in Sections 3 and 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the main practical limitations relevant to the setting, including OOD generalization, dependence on the similarity space, sensitivity to retrieval and ensemble sizes, and retrieval complexity in the results summary in Section 4. We also have a dedicated section in Appendix F.1 that discussed the main limitation and perspectives of our work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not introduce formal theoretical results requiring proofs. The mathematical content is used to define the problem, the retrieval procedure, the semantic alignment objective, and the computational complexity of RAMSAD.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies the benchmark, ID/OOD evaluation protocol, candidate detector pool, baselines, evaluation metric, retrieval settings, and additional implementation details in the appendix. These details are sufficient to reproduce the main experimental comparisons at the level of the reported methodology.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same

dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is released in an open repository at `https://github.com/Hendrix8/RAMSAD`, together with instructions to reproduce the main experimental results. The experiments rely on existing benchmark assets, and the repository documents the required setup, data access, and execution steps.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

Justification: The experimental setting is described in Section 4 and detailed in Appendix D. This includes the description of the TSB-AD benchmark, train/test splits, ID/OOD evaluation protocol, detector pool, competing baselines, evaluation metric, retrieval size, similarity spaces, and parameter-sensitivity analyses.

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports variability across test series through boxplots and compares methods using critical difference diagrams. The text also explicitly discusses when RAMSAD is or is not statistically different from competing approaches.

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports computational complexity, selection latency, and detector runtime for RAMSAD, and the appendix specifies the hardware configuration, wall-clock execution time, and total compute used for the reported experiments. These details are sufficient to reproduce the experiments under comparable compute resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses existing time-series anomaly-detection benchmarks and automated model-selection experiments. We do not collect human-subject data, release high-risk generative models, or introduce procedures that deviate from the NeurIPS Code of Ethics.

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix F.2 discusses positive impacts through more reliable and efficient anomaly-detector selection for time-series monitoring applications, including healthcare, manufacturing, finance, energy, cybersecurity, and scientific monitoring. It also discusses potential negative impacts, including false alarms or missed detections in high-stakes settings, over-reliance on automated monitoring, and possible use in surveillance contexts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a high-risk dataset, pretrained generative model, or other asset with a high risk of misuse. RAMSAD is a model-selection framework evaluated on existing time-series anomaly-detection benchmarks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper uses existing benchmark assets from TSB-UAD [46], which are cited in the experimental section and provide references, dataset sources, and license information for the constituent datasets. We follow the benchmark protocol and terms of use, and do not redistribute or repackage the original datasets beyond their documented usage.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper releases the RAMSAD code as a new asset in an open repository at `https://github.com/Hendrix8/RAMSAD`. The repository documents the implementation, experimental setup, data access, and reproduction instructions for the main results.

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. No participant instructions, screenshots, or compensation details are applicable.

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject research. IRB approval or equivalent review is therefore not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The proposed RAMSAD method is not an LLM-based method and does not rely on an LLM for its core retrieval-based detector selection pipeline. However, the paper includes a non-standard auxiliary experiment on LLM-based model selection. In Section 4, we evaluate a high-capacity LLM model selector under multiple time-series representations, both with and without RAMSAD retrieval support. The LLM is given retrieved training examples and their detector-performance profiles, and the resulting performance is reported separately from the main RAMSAD results.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.