Fast or Accurate? Rethinking Time Series Anomaly Detection

Emmanouil Sylligardos
emmanouil.sylligardos@ens.fr
DI ENS, ENS, PSL University, CNRS, Inria
Supervised by: Paul Boniol and Pierre Senellart

ABSTRACT

Anomaly Detection (AD) is a fundamental task for time-series analytics with important implications for the downstream performance of many applications. Despite increasing academic interest and the large number of methods proposed in the literature, recent benchmarks demonstrate that there exists no single best AD method when applied to heterogeneous time series datasets. This Ph.D. work addresses this challenge by studying model selection-based approaches that dynamically select the most suitable anomaly detector based on time series characteristics. Extensive experiments show that model selection methods consistently outperform individual anomaly detectors while maintaining comparable execution times. To support this, ADecimo is introduced, a web application that allows users to interactively compare and analyze our framework's results. In addition, we propose a highly-efficient evaluation measure, built to scale to billions of data points. Together, these contributions form the basis of a robust, and fast framework for next-generation time series anomaly detection.

VLDB Workshop Reference Format:
Emmanouil Sylligardos. *Fast or Accurate?* Rethinking Time Series Anomaly Detection. VLDB 2025 Workshop: VLDB Ph.D. Workshop.

1 INTRODUCTION

Time series data are ubiquitous and continuously generated in virtually every scientific and industrial domain [8]. Notably, the expansion of Internet-of-Things (IoT) devices has brought up the need for efficient and effective analysis of zettabytes of time series data [11]. Among the many analytical tasks for time series, identifying abnormal or rare events is crucial for the effectiveness of downstream tasks. Consequently, *Anomaly Detection* (AD) has received ample academic and industrial interest and finds applications across a wide range of cases.

In recent years, many techniques have been proposed for Time Series Anomaly Detection (TSAD). Multiple surveys and benchmarks summarize and analyze the state-of-the-art proposed methods [7, 10]. Unfortunately, these benchmark and evaluation studies demonstrated that no overall best AD methods exist when applied to highly heterogeneous time series (i.e., coming from very different domains). In practice, we observe that some methods outperform others on specific time series with either specific characteristics (e.g., stationary or non-stationary time series) or anomalies (e.g., point-based or sequence-based anomalies).

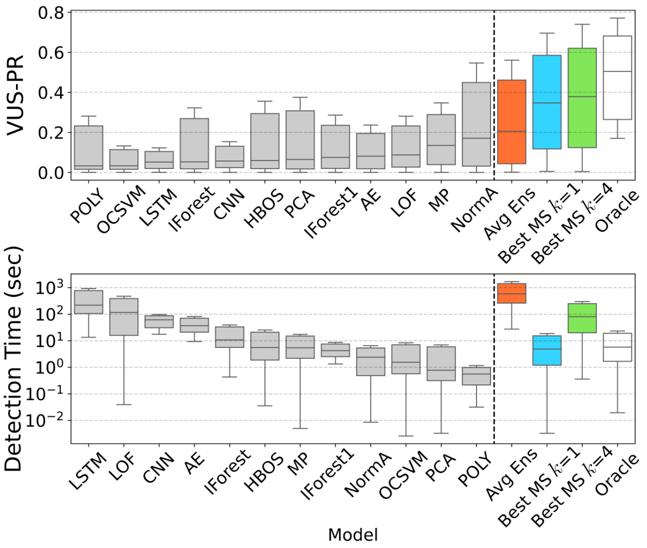

Figure 1: Summary of our evaluation of model selection methods on the TSB-UAD benchmark [10]. The top plot presents an AD accuracy measure, while the bottom part compares execution times (note the logarithmic y-axis). We highlight our model selection methods: best for k=1 (blue) and k=4 (green), compared to 12 AD methods (grey) and the Averaging Ensemble (Avg Ens) baseline (orange).

To overcome the above limitation, ensembling solutions have been proposed [1] that consist of running all existing AD methods and averaging all anomaly scores. Figure 1 shows that this solution (in orange), namely the *Averaging Ensemble* (Avg Ens), outperforms all individual existing techniques in the TSB-UAD benchmark (in grey) [3, 10]. However, as shown in Figure 1, such solutions require running all methods, resulting in an excessive cost that is not feasible in practice.

Additionally, evaluating such models introduces a second major challenge. Traditional, point-based evaluation measures, often adapted from Information Retrieval (IR), fail to capture the temporal structure of anomalies in time series [9]. A recent work [2] introduced Volume Under the Surface (VUS), a next-generation, time series-specific measure designed to overcome these shortcomings. While effective, its computational cost makes it impractical for large-scale benchmarking or real-time applications.

To address the aforementioned challenges, we propose two distinct but complementary contributions: (1) a robust model selection framework for TSAD, and (2) a highly efficient evaluation measure.

Proceedings of the VLDB Endowment. ISSN 2150-8097.

These components work hand-in-hand: the former introduces a new direction for robust time series anomaly detection, while the latter enables fast evaluation at scale. Together, they lay the foundation for building fast, accurate, and scalable TSAD systems. Our past and present work, in the context of my PhD research, includes the following:

- **Choose Wisely:** In our initial work [12], we introduced a framework that reframes the TSAD task as a time series classification problem. This approach enables learning which detector to select for a given time series based on time series characteristics. To evaluate our method, we compare 16 different families of classifiers (with a total of 128 configuration) on the TSB-UAD benchmark [10], and we propose the first extensive experimental evaluation of model selection for TSAD. Our results demonstrate the effectiveness of data-driven model selection.
- **MSAD:** We extended the model selection approach into the MSAD (Model Selection for Anomaly Detection) framework. Unlike its predecessor, MSAD supports combining multiple detectors for a single time series. This not only boosts overall performance (Fig. 1), but also significantly enhances robustness under Out-of-distribution (OOD) scenarios.
- **ADecimo:** To make our framework accessible and interpretable, we developed `ADecimo` [4], a web-based tool that allows users to inspect results on the TSB-UAD benchmark or upload their own time series for analysis.
- **Fast Evaluation at Scale:** We develop a highly optimized version of the time series-specific evaluation measure VUS [2], achieving massive speed-ups through algorithmic redesign and GPU acceleration (Fig. 3). This new version maintains the theoretical guarantees of the original measure, while making it practical for large-scale benchmark evaluations and integration into Deep Learning (DL) pipelines.

2 BACKGROUND AND RELATED WORK

Time Series. A time series $T \in \mathbb{R}^n$ is a sequence of real-valued numbers $[T_1, T_2, \ldots, T_n]$, where $n = |T|$ is the length of T, and T_i is the i^{th} point of T. We focus typically on subsequences, which are continuous subsets of T of length ℓ starting at position i, defined as $T_{i,\ell} = [T_i, T_{i+1}, \ldots, T_{i+\ell-1}]$. A dataset $\mathcal{D}$ is a set of time series, which can vary in length, and its size is denoted as $|\mathcal{D}|$.

Anomaly Score Sequence. For a time series $T \in \mathbb{R}^n$, an AD method (or detector) D returns an anomaly score sequence S_T. In most applications, the anomaly score has to be the same length as the time series.

Anomaly Detection Accuracy. For a time series $T \in \mathbb{R}^n$, an AD method (or detector) D returns an anomaly score sequence $D(T) = S_T$. The labels $L \in [0,1]^n$ indicate with 0 or 1 if the points in T are normal or abnormal, respectively. We define $Acc : \mathbb{R}^n \times \{0,1\}^n \rightarrow [0,1]$ as an accuracy function, for which $Acc(D(T), L)$, namely the accuracy score, indicates how accurate D is.

Anomaly Detection Methods. Over the years, a wide range of AD algorithms have been proposed for diverse types of time series, and applications. They can be grouped into the following families:

- Distance-based methods: These analyze subsequences by comparing distances to a given model to detect anomalies.
- Density-based methods: These detect recurring or isolated behaviors by evaluating the density of points or subsequences into a specific representation space.
- Prediction-based methods: These predict future values, or reconstruct the input, and use the prediction, or reconstructing error, as an anomaly score.

Despite the diversity of approaches, recent benchmark studies [10] have consistently shown that there is no single best detector across heterogeneous datasets.

Ensembling and Model Selection. To address the lack of a universally best detector, ensemble methods have been proposed. One of the most widely used is the *Averaging Ensemble* (Avg Ens in Fig. 1), which assumes access to M detectors. Each detector outputs an anomaly score sequence, and these are aggregated (e.g., by simple averaging) to produce a single score. This strategy has shown strong empirical results across heterogeneous datasets [4, 12], however, ensembling methods come with a high computational cost.

Therefore, the only scalable and viable solution for solving AD over very different time series collected from various domains is to propose a model selection method that selects, based on time series characteristics, the best AD methods to run. This topic has been tackled in several recent research works related to AutoML (Automated Machine Learning) for the general case of AD [15] and also for time series [6]. Nevertheless, existing AutoML solutions require (i) a universal objective function among models, which is not applicable to AD methods; (ii) a predefined set of features, which is difficult to obtain for time series due to varying lengths and the lack of standardized featurization solutions; (iii) running multiple AD methods several times, which is prohibitively expensive in practice; or (iv) labeled anomalies, which (in contrast to classification tasks) are difficult to obtain. Therefore, more work is needed to make AutoML solutions applicable to TSAD.

Evaluation Measures. Evaluating anomaly detectors over time series is non-trivial, and a variety of measures have been proposed. They can be grouped into the following categories:

- **Threshold-Dependent Measures:** They require converting continuous anomaly scores into binary predictions using a threshold. Popular measures include precision, recall, and F1-score, which assess the overlap between predicted anomalies and ground-truth labels. However, their performance is sensitive to the choice of threshold, and they often fail to capture partial overlaps between predicted and actual anomaly regions [14].
- **Threshold-Independent Measures:** To overcome these limitations, threshold-independent measures have been developed. The most widely used are the Area Under the Precision-Recall Curve (AUC-PR) and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). While these measures provide a more holistic view of a model's performance, they still fail to account for the temporal nature of anomalies.
- **Time series-specific Measures:** To address these shortcomings, time series-specific measures have been proposed, such as Volume Under the Surface (VUS) [2]. This family of measures evaluates the detection performance over a range of tolerance

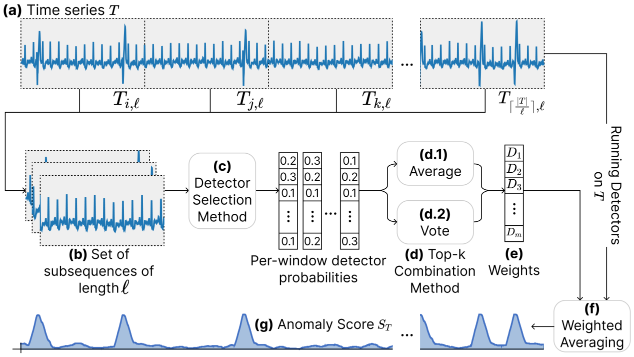

Figure 2: Proposed architecture of the model selection framework MSAD

and latency parameters. However, the original VUS implementation suffers from high computational complexity, making it impractical for large-scale benchmarking. Our work builds upon VUS, proposing a redesigned and highly efficient version that enables fast and scalable evaluation.

3 PAST WORK: MODEL SELECTION FOR TSAD

In this section, we outline our contribution to the problem of model selection for TSAD. We propose a novel framework that reformulates model selection as a time series classification task. This approach enables us to learn which AD method performs best based on the characteristics of the input time series. Our contribution includes: (i) a generic and extensible framework for model selection in TSAD, (ii) an extensive evaluation of a wide range of time series classifiers, and (iii) an interactive web-based application for visualizing and analyzing model selection results.

3.1 Choose Wisely

Time series classification has shown promising results in both cross-domain settings [5] and domain-specific applications such as healthcare [13]. Driven by such examples, we propose an approach to turn TSAD into a time series classification problem that will help identify the most suitable anomaly detector for a given series.

We begin by constructing a classification dataset, where each time series is labeled with the anomaly detection method that performs best on it, selected from a pool of 12 detectors. Classifiers are then trained to predict the optimal detector based solely on the input time series. In our initial work [12], we conducted an extensive evaluation of this strategy, testing over 128 classifier configurations across 16 families, including both feature-based and deep learning models. The results demonstrated that model selection not only improves performance over individual detectors but does so while remaining comparable in execution time.

3.2 MSAD

Model Selection for Anomaly Detection (MSAD) is an extension of the approach mentioned above. While the previous version predicted a single detector for each time series, MSAD can predict multiple good detectors as long as weights on how to combine them with a simple weighted average approach.

The framework's architecture, shown in Figure 2, operates in three main stages:

- **Preprocessing:** Each time series is segmented into non-overlapping subsequences of fixed length. These segments are used as inputs to classification models, ensuring compatibility across models that require uniform input sizes.
- **Classification:** A set of classifiers, ranging from feature-based models to deep learning architectures (including convolutional and transformer-based networks), are trained to predict the most suitable anomaly detector. When multiple detectors are selected (top-k setting), the output probabilities serve as weights for combining anomaly scores.
- **Score Aggregation:** Using either a voting or averaging strategy, detector scores are combined to produce a final anomaly score for the input time series. This allows for flexible trade-offs between accuracy and execution time by varying the parameter k, which controls the number of detectors combined.

To evaluate the effectiveness of our framework, we conduct extensive experiments on a large and heterogeneous benchmark comprising 1980 time series and 12 AD methods. We assess performance in both supervised (in-distribution) and unsupervised (out-of-distribution) settings. The key results are as follows:

- Model selection methods consistently outperform individual detectors and even the Averaging Ensemble baseline in terms of accuracy.
- Selecting and combining a small number of detectors ($k > 1$) yields improved performance with significantly reduced execution time compared to ensembling all detectors.
- Deep learning classifiers, particularly convolutional architectures, demonstrate the highest selection accuracy and robustness across domains.
- In out-of-distribution evaluations, model selection approaches maintain strong performance, validating their generalization capabilities.

To support reproducibility and further exploration, we release all experimental data and provide *ADecimo*, a web application that allows users to browse, compare, and analyze the performance of model selection methods across the benchmark.

3.3 ADecimo

To facilitate exploration and application of our model selection approach, we developed `ADecimo` [4], an interactive web application designed to bridge the gap between complex experimental results and practical usability. The system enables three primary modes of interaction. First, users can explore benchmark-wide performance results to identify the best model selection strategies for their domain. Second, they can dive into individual time series and visually compare the selected detector's output against all others, enabling intuitive validation of the selection's quality. Third, users can upload their own time series and test pre-trained model selectors on real or synthetic data.

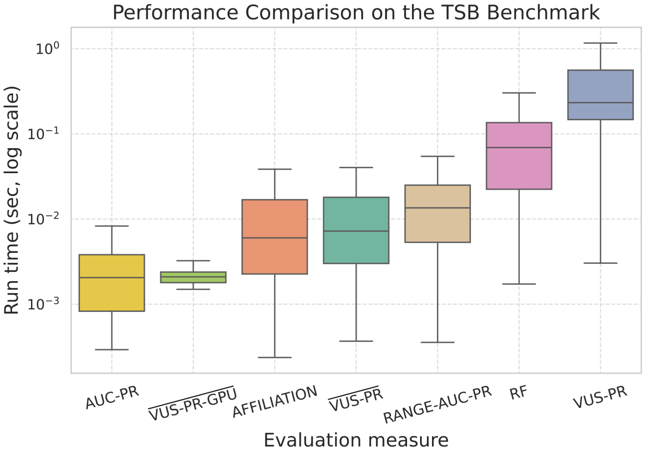

Figure 3: Runtime per time series for various evaluation measures when applied to the entire TSB benchmark.

4 FUTURE WORK

Next-Generation TSAD Evaluation Measures. A core component of TSAD research is the ability to evaluate detectors in a way that accurately captures the temporal nature of anomalies. Existing measures are often threshold-dependent or adapted from domains such as information retrieval, limiting their applicability to TSAD. We are currently developing a highly efficient, next-generation evaluation measure, based on the recently proposed VUS measure [2]. This new measure retains the expressive power of VUS while drastically reducing its computational cost by leveraging algorithmic redesign and hardware acceleration.

In Figure 3, we present preliminary results on the full TSB-UAD benchmark, which includes nearly 1,900 time series of varying lengths (up to 600k points). Our lightweight CPU-based version, denoted $\overline{VUS-PR}$, is already up to two orders of magnitude faster than the current implementation of VUS (note that the y-axis, indicating the runtime per time series, is in logarithmic scale). Moreover, our highly optimized, GPU-based version is consistently up to three orders of magnitude faster and closely matches the runtime performance of AUC-PR while providing the full functionality of VUS. In the continuation of this research, we will provide a detailed analysis of the implementation and the performance of our new evaluation measure, and release them as part of an open-source TSAD evaluation toolkit. Additionally, we plan to make it scalable to billion-point time series. By making time-aware evaluation practical at scale, we aim to pave the way for more ambitious model development and benchmarking.

Reinforcement learning for TSAD. While deep learning has transformed fields like computer vision and NLP, its impact on TSAD has been more limited. Recently, DL-based model selection frameworks have unlocked new directions of research in the domain of TSAD. While this has shown promising results, it is inherently a proxy task as the classifier learns to predict the best detector(s) rather than directly optimizing anomaly detection quality. To overcome this limitation, we plan to explore reinforcement learning frameworks and design reward functions that use feedback from evaluation measures, enabling learning directly from actual anomaly detection performance. In this context, RL offers the flexibility to incorporate additional objectives, such as computational efficiency, into the reward function. This could enable model selection systems to jointly optimize for both detection accuracy and detector runtime, making them more practical for real-world deployment.

Building a knowledge base for TSAD. TSAD has seen a lot of improvement in recent years, and model selection with DL approaches aim to push even more the performance barrier. However, little attention has been given to explaining the anomaly detection process in time series. Towards this direction, we intend to construct a data-driven repository that maps time series and anomaly characteristics to detector performance, offering insight into *why* certain methods work better on specific data types.

Together, these next steps aim to move TSAD closer to robust systems that can scale to real-world applications.

REFERENCES

[1] Charu C. Aggarwal and Saket Sathe. 2015. Theoretical Foundations and Algorithms for Outlier Ensembles. *SIGKDD Explor. Newsl.* 17, 1 (sep 2015), 24–47. https://doi.org/10.1145/2830544.2830549

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

[3] Paul Boniol, John Paparrizos, Yuhao Kang, Themis Palpanas, Ruey S Tsay, Aaron J Elmore, and Michael J Franklin. 2022. Theseus: navigating the labyrinth of time-series anomaly detection. *Proceedings of the VLDB Endowment* 15, 12 (2022), 3702–3705.

[4] Paul Boniol, Emmanouil Sylligardos, John Paparrizos, Panos Trahanias, and Themis Palpanas. 2024. Adecimo: Model selection for time series anomaly detection. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 5441–5444.

[5] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2019. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery* 33 (2019), 917–963.

[6] Mononito Goswami, Cristian Challu, Laurent Callot, Lenon Minorics, and Andrey Kan. 2022. Unsupervised Model Selection for Time-series Anomaly Detection. https://doi.org/10.48550/ARXIV.2210.01078

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

[8] Themis Palpanas and Volker Beckmann. 2019. Report on the First and Second Interdisciplinary Time Series Analysis Workshop (ITISA). *SIGMOD Rec.* 48, 3 (Dec. 2019), 36–40. https://doi.org/10.1145/3377391.3377400

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

[10] John Paparrizos, Yuhao Kang, Paul Boniol, Ruey S. Tsay, Themis Palpanas, and Michael J. Franklin. 2022. TSB-UAD: An End-to-End Benchmark Suite for Univariate Time-Series Anomaly Detection. *Proc. VLDB Endow.* 15, 8 (2022).

[11] John Paparrizos, Chunwei Liu, Bruno Barbarioli, Johnny Hwang, Ikraduya Edian, Aaron J Elmore, Michael J Franklin, and Sanjay Krishnan. 2021. VergeDB: A Database for IoT Analytics on Edge Devices.. In *CIDR*.

[12] Emmanouil Sylligardos, Paul Boniol, John Paparrizos, Panos Trahanias, and Themis Palpanas. 2023. Choose wisely: An extensive evaluation of model selection for anomaly detection in time series. *Proceedings of the VLDB Endowment* 16, 11 (2023), 3418–3432.

[13] Emmanouil Sylligardos, Markos Sigalas, Stella Soundoulounaki, Katerina Vaporidi, and Panos Trahanias. 2022. A Deep Learning Approach to Detect Ventilatory Over-Assistance. In *International Conference on Pattern Recognition and Artificial Intelligence*. Springer, 504–515.

[14] Nesime Tatbul, Tae Jun Lee, Stan Zdonik, Mejbah Alam, and Justin Gottschlich. 2018. Precision and recall for time series. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 1924–1934.

[15] Yue Zhao, Ryan Rossi, and Leman Akoglu. 2021. Automatic Unsupervised Outlier Model Selection. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 4489–4502.