Université Paris Cité

Volume Under the Surface: A New Accuracy Evaluation Measure for Time-Series Anomaly Detection

John Paparrizos, Paul Boniol, Themis Palpanas, Ruey S. Tsay, Aaron Elmore, Michael J. Franklin

LIPADE-TR-Nº 7

March 24, 2022

Technical Report

LIPADE
Laboratoire d'Informatique PAris DEscartes

Volume Under the Surface: A New Accuracy Evaluation Measure for Time-Series Anomaly Detection

John Paparrizos
University of Chicago
jopa@uchicago.edu

Paul Boniol
University of Paris
paul.boniol@etu.u-paris.fr

Themis Palpanas
University of Paris; IUF
themis@mi.parisdescartes.fr

Ruey S. Tsay
University of Chicago
ruey.tsay@chicagobooth.edu

Aaron Elmore
University of Chicago
aelmore@uchicago.edu

Michael J. Franklin
University of Chicago
mjfranklin@uchicago.edu

ABSTRACT

Anomaly detection (AD) is a fundamental task for time-series analytics with important implications for the downstream performance of many applications. In contrast to other domains where AD mainly focuses on point-based anomalies (i.e., outliers in standalone observations), AD for time series is also concerned with range-based anomalies (i.e., outliers spanning multiple observations). Nevertheless, it is common to use traditional point-based information retrieval measures, such as Precision, Recall, and F-score, to assess the quality of methods by thresholding the anomaly score to mark each point as an anomaly or not. However, mapping discrete labels into continuous data introduces unavoidable shortcomings, complicating the evaluation of range-based contextual and collective anomalies. Notably, the choice of evaluation measure may significantly bias the experimental outcome. Despite over six decades of attention, there has never been a large-scale systematic quantitative and qualitative analysis of time-series AD evaluation measures. This paper extensively evaluates quality measures for time-series AD to assess their robustness under noise, misalignments, and different anomaly cardinality ratios. Our results indicate that measures producing quality values independently of a threshold (i.e., AUC-ROC and AUC-PR) are more suitable for time-series AD. Motivated by this observation, we first extend the AUC-based measures to account for range-based anomalies. Then, we introduce a new family of parameter-free and threshold-independent measures, VUS (Volume Under the Surface), to evaluate methods while varying parameters. Our findings demonstrate that our four measures are significantly more robust and helpful in assessing and separating the quality of time-series AD methods.

PVLDB Reference Format:
John Paparrizos, Paul Boniol, Themis Palpanas, Ruey S. Tsay, Aaron Elmore, and Michael J. Franklin. Volume Under the Surface: A New Accuracy Evaluation Measure for Time-Series Anomaly Detection. PVLDB, 15(1): XXX-XXX, 2022. doi:XX.XX/XXX.XX

1 INTRODUCTION

Massive collections of time-varying measurements, commonly referred to as *time series* or *data series*, are becoming a reality in virtually every scientific and industrial domain [4, 28]. Notably, there is an increasingly pressing need for developing techniques for efficient and effective analysis of zettabytes of time series produced by millions of Internet-of-Things (IoT) devices [14, 16]. IoT deployments empower diverse data science applications in environmental sciences, astrophysics, neuroscience, and engineering, among others [27, 39], and have revolutionized many industries, including automobile, healthcare, manufacturing, retail, and utilities, among others [25]. Unfortunately, imperfections and inherent complexities in the data generation and measurement pipelines often introduce abnormalities that appear as *anomalies* in time-series databases, impacting the effectiveness of downstream tasks and analytics.

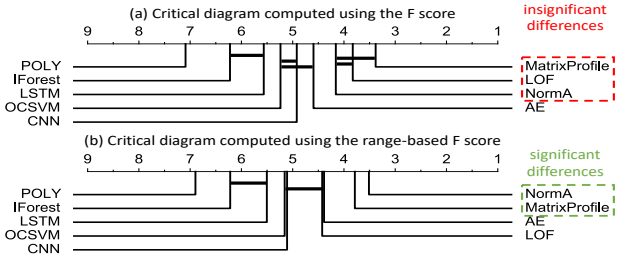

Figure 1: Critical difference diagram computed with the Friedman test followed by a post-hoc Wilcoxon test (with $\alpha = 0.1$) for the (a) F-score and (b) range-based F-score over 250 time series in KDD21 [18]. Bold lines indicate insignificant differences of connected methods.

Consequently, *anomaly detection* (AD) becomes a fundamental problem with broad applications sharing the same goal [5, 31, 37]: analyzing time series to identify observations that do not conform to some notion of expected behavior based on previously observed data. During the past decades, a multitude of AD methods have been proposed and compared [8–10, 21, 37]. Different from other domains that principally focus on *point-based* anomalies (i.e., outliers in standalone observations), AD for time series is also concerned with *range-based* anomalies (i.e., outliers spanning multiple observations). Unfortunately, it has become common practice to use traditional point-based information retrieval (IR) accuracy evaluation measures, such as Precision, Recall, and F-score, to quantify the effectiveness of different anomaly detectors.

In addition, the previously mentioned IR evaluation measures suffer from a significant limitation: a threshold is necessary over the anomaly score produced by AD methods to mark each time-series point as an anomaly or not. The most common approach to set a threshold value is to use the average score plus three times the standard deviation of the anomaly score. However, this popular choice might not suit every AD method, use case, and domain, leading to significant variations in the quality values of the previously mentioned evaluation measures. Therefore, these IR measures are

Proceedings of the VLDB Endowment, Vol. 15, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

difficult to trust and complicate evaluating different AD methods on heterogeneous benchmarks. To eliminate the need to set a threshold, another standard measure for binary classification is used: the receiver operator characteristic (ROC) curve and the Area Under the Curve (AUC), which is the area below the ROC curve (AUC-ROC). The ROC curve is generated by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings (instead of only one threshold used in Precision, Recall, and F-score measures). Another variant, the Precision-Recall (PR) curve, represents the relation between Precision and Recall, and the Area under the PR curve (AUC-PR) is the area below PR [11].

Unfortunately, all previous measures, Precision, Recall, F-Score, AUC-ROC, and AUC-PR, are ideal for point-based anomalies but cannot adequately evaluate ubiquitous range-based contextual and collective anomalies [7]. Remarkably, the mapping of discrete labels into continuous data introduces unavoidable shortcomings (e.g., difficulty in marking precisely the range of the anomalies and handling misalignments between the human labels and the anomaly range produced by thresholding the anomaly score). To address these shortcomings, a range-based definition of Precision and Recall has been proposed by extending the traditional definitions [32]. Range-based Precision, Recall, and F-Score consider several factors: (i) whether a subsequence is detected or not; (ii) how many points in the subsequence are detected; (iii) which part of the subsequence is detected; and (iv) how many fragmented regions correspond to one real subsequence outlier. This definition is detailed and comprehensive; however, several parameters require tuning and, importantly, a threshold over the anomaly score is still required.

Despite over six decades of attention [13, 26, 34], there has never been (to the best of our knowledge) a large-scale systematic quantitative and qualitative analysis of time-series AD evaluation measures. Notably, the choice of evaluation measure may significantly bias the experimental outcome. To understand the implications of choosing an appropriate measure, Figure 1 depicts the critical diagrams of the F-score and range-based F-score computed with the Friedman test followed by a Wilcoxon test [35] over several AD methods (see Section 5 for details) across the 250 time series of the KDD21 dataset [18]. Figure 1 demonstrates that not only the ranking is changing, but also some methods shift from insignificantly to significantly different from one measure to the other.

In this paper, we extensively evaluate quality measures for time-series AD to assess their robustness under noise, misalignments, and different anomaly cardinality ratios. Specifically, our study includes 9 previously proposed quality measures, computed over the anomaly scores of 10 AD methods across 10 previously proposed diverse datasets that contain 900 time series with marked anomalies. Our analysis assess the robustness of quality measures both qualitatively and quantitatively by studying the influence of threshold, lag, noise, and normal-abnormal anomaly ratio to identify robust quality measures that can better separate the accurate from inaccurate methods. Our results indicate that measures producing quality values independently of a threshold (i.e., AUC-ROC and AUC-PR) are more suitable for time-series AD. This is surprising considering that we include the range-based Precision, Recall, and F-score measures, which highlights the strong influence the thresholding of anomaly scores has in assessing the quality of methods.

Motivated by this observation and to address the limitations of existing measures, we propose *four* new accuracy evaluation measures. We first present Range-AUC-ROC and Range-AUC-PR,

Accuracy Measure	# of anomalies	Score Threshold	Sequence-adapted	Parameter-free
Precision@k	✗	✓	✗	✗
Precision	✓	✗	✗	✗
Recall	✓	✗	✗	✗
F-Score	✓	✗	✗	✗
Rprecision	✓	✗	✓	✗
Rrecall	✓	✗	✓	✗
RF-Score	✓	✗	✓	✗
AUC-PR	✓	✓	✗	✓
AUC-ROC	✓	✓	✗	✓
Proposed measures				
R-AUC-PR	✓	✓	✓	✗
R-AUC-ROC	✓	✓	✓	✗
VUS-PR	✓	✓	✓	✓
VUS-ROC	✓	✓	✓	✓

Table 1: Analysis of quality measures based on: (i) independence to the number of annotated anomalies; (ii) independence to the threshold on the anomaly score; (iii) adaptation to continuous sequences; and (iv) independence to setting parameters. Our VUS-based measures match all characteristics while competitors miss one or more.

threshold-independent (for the anomaly score) evaluation measures that use a continuous buffer region in the labels to increase the robustness to potential misalignments with the human labels. Then, we propose the Volume Under the Surface (VUS) family of measures that extend the traditional AUC measures to consider all buffer sizes (in addition to all thresholds). Therefore, VUS-ROC and VUS-PR are parameter-free, threshold-independent, and robust to lags, noise, and anomaly cardinality ratios. Our analysis demonstrates that VUS-ROC and VUS-PR are the most reliable accuracy quality measures for both point-based and range-based anomalies evaluation. Table 1 summarizes the accuracy evaluation measures analysed in this paper based on their independence to four critical characteristics.

Interestingly, even though outside of the scope of this paper, the flexibility of VUS measures in evaluating methods while varying parameters of choice may have profound implications beyond time-series AD. Specifically, VUS measures are applicable across binary classification tasks for evaluating methods with a single quality value while considering different parameter choices (e.g., learning rates, batch sizes, and other critical varying parameters). Similarly, with the rise of new machine learning (ML) techniques for AD and binary classification, the integration of VUS measures into the objective functions may result in substantially more robust models.

We start with a detailed discussion of the relevant background and related work (Section 2). Then, we present our contributions:

- We describe and study the limitations of existing evaluation measures, resulting in a formal definition of the necessary principles of time-series AD quality measures (Section 3).
- We present R-AUC-ROC and R-AUC-PR that rely on a new label transformation to make their score robust and reliable for range-based contextual and collective anomalies (Section 4.1).
- We introduce VUS-ROC and VUS-PR, parameter-free measures that formally extend the mathematical model of AUC-based measures to consider more varying parameters (Section 4.2).
- We extensively evaluate, both qualitatively and quantitatively, 13 quality measures (9 previously proposed and our 4 new measures) across 10 AD methods over 10 diverse datasets containing 900 time series with marked anomalies (Sections 5.2 and 5.3).
- We analyze the separability of the measures by evaluating changes in AD methods' ranks across measures (Section 5.4).

Finally, we conclude with the implications of our work (Section 6).

2 BACKGROUND AND RELATED WORK

We first introduce formal notations useful for the rest of the paper (Section 2.1). Then, we review in detail previously proposed evaluation measures for time-series AD methods (Section 2.2).

2.1 Time-Series and Anomaly Score Notations

We review notations for the time series and anomaly score sequence.

Time Series: A time series $T \in \mathbb{R}^n$ is a sequence of real-valued numbers $T_i \in \mathbb{R}$ $[T_1, T_2, ..., T_n]$, where $n = |T|$ is the length of T, and T_i is the i^{th} point of T. We are typically interested in local regions of the time series, known as subsequences. A subsequence $T_{i,\ell} \in \mathbb{R}^\ell$ of a time series T is a continuous subset of the values of T of length ℓ starting at position i. Formally, $T_{i,\ell} = [T_i, T_{i+1}, ..., T_{i+\ell-1}]$.

Anomaly Score Sequence: For a time series $T \in \mathbb{R}^n$, an AD method A returns an anomaly score sequence S_T. For point-based approaches (i.e., methods that return a score for each point of T), we have $S_T \in \mathbb{R}^n$. For range-based approaches (i.e., methods that return a score for each subsequence of a given length ℓ), we have $S_T \in \mathbb{R}^{n-\ell}$. Overall, for range-based (or subsequence-based) approaches, we define $S_T = [S_{T1}, S_{T2}, ..., S_{Tn-\ell}]$ with $S_{Ti} \in [0, 1]$.

2.2 Accuracy Evaluation Measures for AD

We present previously proposed quality measures for evaluating the accuracy of an AD method given its anomaly score. We first discuss threshold-based and, then, threshold-independent measures.

2.2.1 *Threshold-based AD Evaluation Measures.*

The anomaly score S_T produced by an AD method A highlights the parts of the time series T considered as abnormal. The highest values in the anomaly score correspond to the most abnormal points. Threshold-based measures require to set a threshold to mark each point as an anomaly or not. Usually, this threshold is set to $\mu(S_T) + \alpha * \sigma(S_T)$, with α set to 3 [5], where $\mu(S_T)$ is the mean and $\sigma(S_T)$ is the standard deviation S_T. Given a threshold $Thres$, we compute the $pred \in \{0, 1\}^n$ as follows:

$$\forall i \in [1, |S_T|], pred_i = \begin{cases} 0, & \text{if: } S_{Ti} < Thres \\ 1, & \text{if: } S_{Ti} \geq Thres \end{cases} \quad (1)$$

Threshold-based measures compare $pred$ to $label \in \{0, 1\}^n$, which indicates the true (human provided) labeled anomalies. Given the Identity vector $I = [1, 1, ..., 1]$, the points detected as anomalies or not fall into the following four categories:

- **True Positive (TP)**: Number of points that have been correctly identified as anomalies. Formally: $TP = label^\top \cdot pred$.
- **True Negative (TN)**: Number of points that have been correctly identified as normal. Formally: $TN = (I - label)^\top \cdot (I - pred)$.
- **False Positive (FP)**: Number of points that have been wrongly identified as anomalies. Formally: $FP = (I - label)^\top \cdot pred$.
- **False Negative (FN)**: Number of points that have been wrongly identified as normal. Formally: $FN = label^\top \cdot (I - pred)$.

Given these four categories, several quality measures have been proposed to assess the accuracy of AD methods.

Precision: We define Precision (or positive predictive value) as the number correctly identified anomalies over the total number of points detected as anomalies by the method:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall: We define Recall (or True Positive Rate (TPR), tpr) as the number of correctly identified anomalies over all anomalies:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

False Positive Rate (FPR): A supplemental measure to the Recall is the FPR, fpr, defined as the number of points wrongly identified as anomalies over the total number of normal points:

$$fpr = \frac{FP}{FP + TN} \quad (4)$$

F-Score: Precision and Recall evaluate two different aspects of the AD quality. A measure that combines these two aspects is the harmonic mean F_β, with non-negative real values for β:

$$F_\beta = \frac{(1 + \beta^2) * Precision * Recall}{\beta^2 * Precision + Recall} \quad (5)$$

Usually, β is set to 1, balancing the importance between Precision and Recall. In this paper, F_1 is referred to as F or F-score.

Precision@k: All previous measures require an anomaly score threshold to be computed. An alternative approach is to measure the Precision using a subset of anomalies corresponding to the k highest value in the anomaly score S_T. This is equivalent to setting the threshold such that only the k highest values are retrieved.

To address the shortcomings of the point-based quality measures, a range-based definition was recently proposed, extending the mathematical models of the traditional Precision and Recall [32]. This definition considers several factors: (i) whether a subsequence is detected or not (ExistenceReward or ER); (ii) how many points in the subsequence are detected (OverlapReward or OR); (iii) which part of the subsequence is detected (position-dependent weight function); and (iv) how many fragmented regions correspond to one real subsequence outlier (CardinalityFactor or CF). Formally, we define $R = \{R_1, ...R_{N_r}\}$ as the set of anomaly ranges, with $R_k = \{pos_i, pos_{i+1}, ..., pos_{i+j}\}$ and $\forall pos \in R_k, label_{pos} = 1$, and $P = \{P_1, ...P_{N_p}\}$ as the set of predicted anomaly ranges, with $P_k = \{pos_i, pos_{i+1}, ..., pos_{i+j}\}$ and $\forall pos \in R_k, pred_{pos} = 1$. Then, we define ER, OR, and CF as follows:

$$\begin{aligned} ER(R_i, P) &= \begin{cases} 1, & \text{if } \sum_{j=1}^{N_p} |R_i \cap P_j| \geq 1 \\ 0, & \text{otherwise} \end{cases} \\ CF(R_i, P) &= \begin{cases} 1, & \text{if } \exists P_i \in P, |R_i \cap P_i| \geq 1 \\ \gamma(R_i, P), & \text{otherwise} \end{cases} \\ OR(R_i, P) &= CF(R_i, P) * \sum_{j=1}^{N_p} \omega(R_i, R_i \cap P_j, \delta) \end{aligned} \quad (6)$$

The $\gamma(), \omega()$, and $\delta()$ are tunable functions that capture the cardinality, size, and position of the overlap respectively. The default parameters are set to $\gamma() = 1, \delta() = 1$ and $\omega()$ to the overlap ratio covered by the predicted anomaly range [32].

Rprecision and Rrecall [32]: Based on the above, we define:

$$\begin{aligned} Rprecision(R, P) &= \frac{\sum_{i=1}^{N_p} Rprecision_s(R, P_i)}{N_p} \\ Rprecision_s(R, P_i) &= CF(P_i, R) * \sum_{j=1}^{N_r} \omega(P_i, P_i \cap R_j, \delta) \end{aligned} \quad (7)$$

$$\begin{aligned} Rrecall(R, P) &= \frac{\sum_{i=1}^{N_r} Rrecall_s(R_i, P)}{N_r} \\ Rrecall_s(R_i, P) &= \alpha * ER(R_i, P) + (1 - \alpha) * OR(R_i, P) \end{aligned} \quad (8)$$

The parameter α is user defined. The default value is $\alpha = 0$.

Range F-score (RF) [32]: As described previously, the F-score combines Precision and Recall. Similarly, we define RF_β, with non-negative real values for β as follows:

$$RF_\beta = \frac{(1 + \beta^2) * Rprecision * Rrecall}{\beta^2 * Rprecision + Rrecall} \quad (9)$$

As before, β is set to 1. In this paper, RF_1 is referred to as RF-score.

2.2.2 *Threshold-independent AD Evaluation Measures.*

Until now, we introduced accuracy measures requiring to threshold the produced anomaly score of AD methods. However, the accuracy values vary significantly when the threshold changes. In order to evaluate a method holistically using its corresponding anomaly score, two measures from the AUC family of measures are used.

AUC-ROC [12]: The Area Under the Receiver Operating Characteristics curve (AUC-ROC) is defined as the area under the curve corresponding to TPR on the y-axis and FPR on the x-axis when we vary the anomaly score threshold. The area under the curve is computed using the trapezoidal rule. For that purpose, we define Th as an ordered set of thresholds between 0 and 1. Formally, we have $Th = [Th_0, Th_1, ...Th_N]$ with $0 = Th_0 < Th_1 < ... < Th_N = 1$. Therefore, *AUC-ROC* is defined as follows:

$$AUC\text{-}ROC = \frac{1}{2}\sum_{k=1}^{N} \Delta^k_{TPR} * \Delta^k_{FPR} \quad \text{with:} \begin{cases} \Delta^k_{FPR} & = FPR(Th_k) - FPR(Th_{k-1}) \\ \Delta^k_{TPR} & = TPR(Th_{k-1}) + TPR(Th_k) \end{cases} \tag{10}$$

AUC-PR [11]: The Area Under the Precision-Recall curve (AUC-PR) is defined as the area under the curve corresponding to the Recall on the x-axis and Precision on the y-axis when we vary the anomaly score threshold. As before, the area under the curve is computed using the trapezoidal rule. Thus, we define AUC-PR:

$$AUC\text{-}PR = \frac{1}{2}\sum_{k=1}^{N} \Delta^k_{Precision} * \Delta^k_{Recall} \quad \text{with:} \begin{cases} \Delta^k_{Recall} & = Recall(Th_k) - Recall(Th_{k-1}) \\ \Delta^k_{Precision} & = Precision(Th_{k-1}) + Precision(Th_k) \end{cases} \tag{11}$$

A simpler alternative to approximate the area under the curve is to compute the average Precision of the PR curve:

$$AUC\text{-}PR = \sum_{k=1}^{N} Precision(Th_k) * \Delta^k_{Recall} \tag{12}$$

In this paper, we use the above equation to approximate AUC-PR.

3 PROBLEM MOTIVATION AND LIMITATIONS

Having introduced existing measures to assess the quality of range-based anomalies, we now elaborate on their critical limitations.

3.1 Limitations of Threshold-based Measures

The need to threshold the anomaly score severely impacts the accuracy measures. First, Figure 2(a) depicts an electrocardiogram time series with an arrhythmia in red (Figure 2(a.1)) and the corresponding anomaly score computed with Isolation Forest [22] (Figure 2(a.2)) for one threshold equal to $\mu(score) + \sigma(score)$ and for another threshold $\mu(score) + 0.6 * \sigma(score)$ (Figures 2(a.3) and (a.4)). We compute the different accuracy measures for the first threshold (blue bars in Figure 2(a.5)) and the second threshold (orange bars in Figure 2(a.5)) and their differences (Figure 2(a.6)). We observe that the threshold choice have a strong impact on Precision, Rprecision, F and RF scores. On the contrary, the threshold-independent measures (i.e., measures computing all possible thresholds), namely, AUC-ROC and AUC-PR, show a clear advantage.

Overall, the threshold choice dependents on the application and the type of input time series. Setting the threshold automatically is hard and almost impossible when we compare different categories

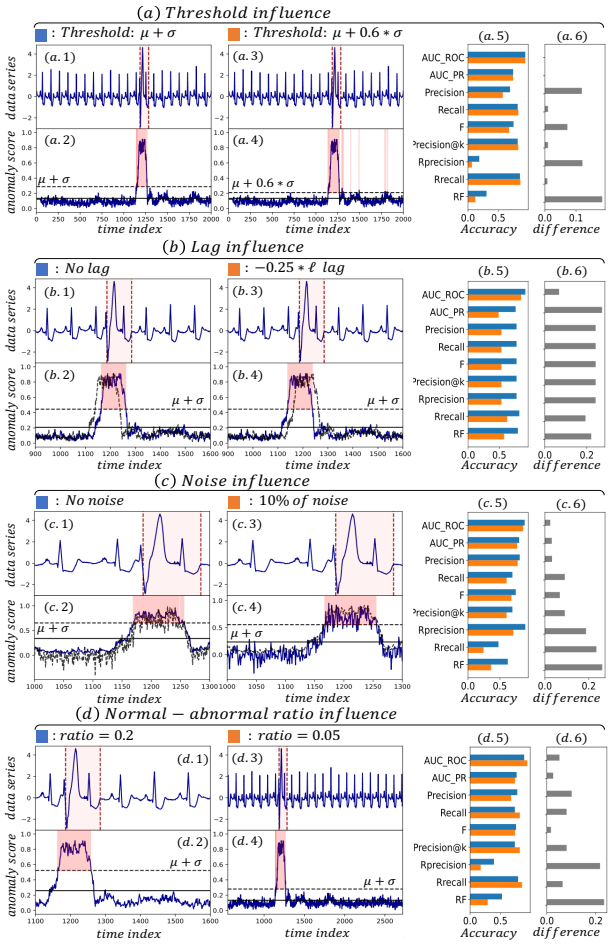

Figure 2: Accuracy evaluation measures when we vary the (a) threshold, (b) lag, (c) noise, and (d) normal/abnormal ratio. Example with Isolation Forest methods over a snippet of the MBA(805) time series.

of AD methods across heterogeneous benchmarks. To illustrate this point, we consider two transformations of the anomaly score that correspond to practical cases we observed (e.g., different methods introduce different lag and noise levels to the anomaly score).

Influence of Noise: Some AD methods applied to some specific time series might result in a noisy anomaly score. In addition, due to manufacturing issues or external causes, a sensor can inject noise on the time series, which then propagates on the anomaly score. Figure 2(c) is depicting two cases: the first corresponds to an anomaly score without any noise (Figure 2(c.2)). The second corresponds to an anomaly score with noise (Figure 2(c.2)). We applied on both cases the same threshold $\mu(score) + \sigma(score)$. We observe in Figure 2(c.6) that most of the threshold-based measures are strongly impacted by noise. This is caused by the fact that the score fluctuates around the threshold, making threshold-based measures less robust to noise. On the contrary, AUC-ROC and AUC-PR are much less influenced by noise, returning approximately the same value.

Influence of Normal/Abnormal Ratio: Depending on the domain and the task, the number of anomalies and consequently the

normal/abnormal ratio changes drastically. A variation on this ratio might cause a variation on the threshold, which leads to variations on threshold-based accuracy measures values. This is explained by the fact that if an anomaly score detects the anomalies correctly, the standard deviation of that score will be higher for a time series with more anomalies. Figure 2(d) depicts two cases: one time series snippet with a 0.2 ratio (Figure 2(d.2)) and one time series snippet with a 0.05 ratio (Figure 2(d.4)). We observe that this change implies a larger variation for several threshold-based measures. Thus, the latter confirms the limitations and the non-robustness of threshold-based measures to the anomaly cardinality ratio.

3.2 Limitations of Point-based Measures

In the previous section, we illustrated the limitations of threshold-based measures. By design and because of their independence to the threshold choice, AUC-ROC and AUC-PR measures are robust to those limitations. However, those measures are designed for point-based outliers. Each point is considered independently and the detection of each point has an equivalent contribution to AUC. In contrast, we need to consider two factors, the range detection and the existence detection, for the subsequence AD problem.

The range detection has the same methodology as point detection. We prefer that the algorithm detects every point in the subsequence anomaly. The existence detection is a loose but crucial estimation for the anomaly subsequence detector: detecting a tiny segment of one subsequence outlier is still of great value.

Mismatch between the anomaly score and labels: Compared to point-based AD, the range-based AD encourages accurately capturing each subsequence anomaly, but the existence detection is good enough to be partially rewarded. Two other reasons support the application of this coarse estimation.

First, there is no consistent labeling tradition among different datasets. Some people may label the whole period as an anomaly if this period does not repeat the typical pattern, while others may only mark a partial period. Even if we specify that each period should share the same label, the next question is how to define the starting and end points of one period. Giving accurate starting or end points, it is also challenging to label a small segment in one period. Unlike a point outlier, which appears to be an evident deviation to the trend line of the time series, range-based anomalies may not have outrageous values. This difficulty of labeling is inevitable when we assign the discrete labels to a continuous time series. There may be some transition region between the two statuses, but we have to decide on a discontinuous jumping point between the two statuses artificially.

Secondly, many algorithms, for instance, LOF [10] and iForest [22], would first apply a sliding window to convert a 1-D time series to a set of high-dimensional data points. We denote the original time series as $(T_1, T_2, \ldots, T_n)$, and suppose the length of window is ℓ, then the converted data set is $\{(T_i, \ldots, T_{i+\ell-1}) | i \in \{1, \ldots T-\ell+1\}\}$. The label of point T_k in the time series is defined as the label of high-dimensional point $(T_{k-\ell/2}, \ldots, T_{k+\ell/2-1})$ in the transformed dataset. The conversion from time series to data set has one consequence: every dimension in the high-dimensional point is equally important. So an abnormal value at the middle or end of this point has the same ability to make it an outlier in the high-dimensional space. Usually, if the sliding window covers more anomaly points, a good algorithm should give a higher anomaly score to the converted data point. However, there are some exceptions that one abnormal value at the beginning or the end of sliding windows is enough to make the converted point an outlier. To summarize, an anomaly subsequence $(T_s, \ldots, T_e)$ may induce a high anomaly score for the range $[T_{s-\ell/2}, T_{e+\ell/2}]$. A perfect result is that the peak of the anomaly score is slightly broader than the whole abnormal region. However, the anomaly score is not perfect. A high anomaly score may be assigned at the range $[T_{s-\ell/2}, T_s]$, which fails to reveal the entire range of the subsequence outlier but succeed in indicating their starting region. AUC-based measures will give a low value since there is no overlap between the score peak and the outlier.

Overall Limitations due to Lag: A lag can be injected into the anomaly score depending on the choice of AD method. Overall such a lag may also exist by the approximation made during the labeling phase. As illustrated in Figure 2(b), such a lag (even though small) has a substantial impact on *all* existing evaluation measures. For example, in Figure 2(b) AUC-PR fluctuates between 0.75 and 0.50 for a lag of just 0.25 of the labeled section length. Among all measures, only the AUC-ROC measure demonstrates to be less sensitive to such lag (as well as noise and normal/abnormal ratio).

3.3 Problem Definition

In summary, our goal is to develop a new parameter-free and anomaly score threshold-independent evaluation measure based on the robust principles of AUC. A promising direction is an extension of AUC for the range-based AD with the following desired properties:

- **Robust to Lag**: Two similar anomaly scores with a slight lag difference should return approximately the same accuracy measures. For example, a high anomaly score near the border of the anomaly should be rewarded as close as a high anomaly score in the middle of the range-based anomaly.
- **Robust to Moise**: Two similar anomaly scores with and without noise should return approximately the same accuracy.
- **Robust to the Anomaly Cardinality Ratio**: This ratio should not have an impact on the accuracy measures.
- **High Separability between Accurate and Inaccurate Methods**: While satisfying the previous points, the accuracy measure should well separate accurate from inaccurate methods.

Next, we present new accuracy measures to satisfy these properties.

4 OUR MEASURES: RANGE-AUC AND VUS

We first present new range-based extensions for ROC and PR curves by introducing a new continuous label to enable more flexibility in measuring detected anomaly ranges. We then present the Volume Under the Surface (VUS) for both ROC and PR curves. VUS extents the mathematical model of Range-AUC measures by varying the buffer length, making VUS family of measures truly parameter-free.

4.1 Range-AUC-ROC and Range-AUC-PR

To compute the ROC curve and PR curve for a subsequence, we need to extend to definitions of TPR, FPR, and Precision.

The first step is to add a buffer region at the boundary of outliers. The idea is that there should be a transition region between the normal and abnormal subsequences to accommodate the false tolerance of labeling in the ground truth (as discussed, this is unavoidable due to mapping of discrete data to continuous time series). An extra benefit is that this buffer will give some credit to the high

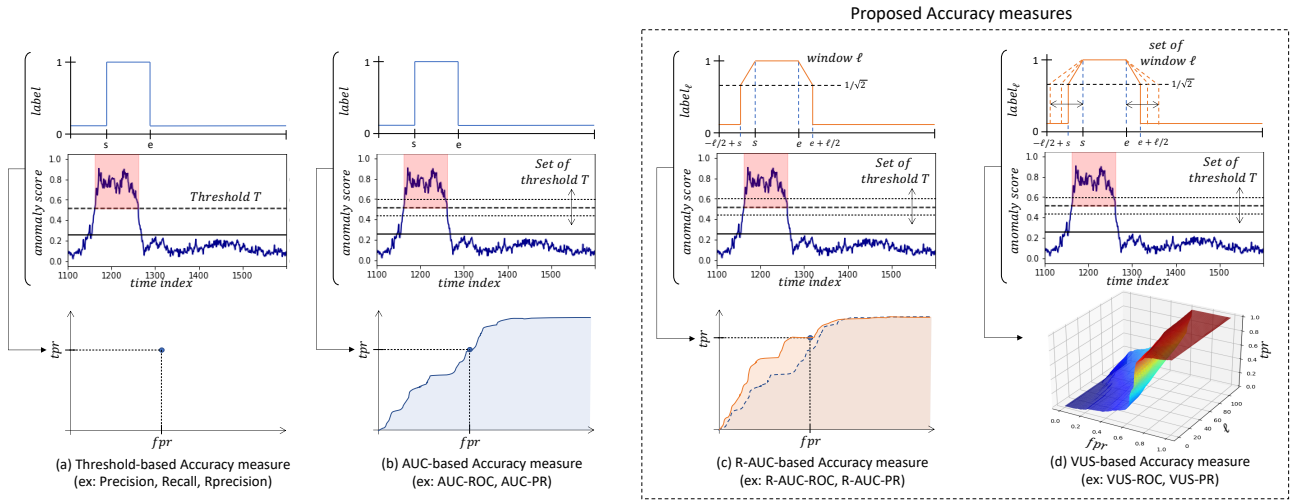

Figure 3: Illustration of previous quality measures compared to our proposed measures. By varying the buffer window from 0 to the period, VUS constructs a surface of TPR, FPR, and window. The volume under the surface is a measure of AUC for various windows.

anomaly score in the vicinity of the outlier boundary, which is what we expected with the application of a sliding window originally.

Figure 3(b) shows the original binary labels (in blue) and Figure 3(c) the new label with buffer region (in orange). By default, the width of the buffer region at each side is half of the period w of the time series (the period is an intrinsic characteristic of time series). Differently, this parameter can be set into the average length of anomaly sizes or can be set to a desired value by the user.

The traditional binary label is extended to a continuous value. Formally, for a given buffer length ℓ, the positions $s, e \in [0, |label|]$ the beginning and end indexes of a labeled anomaly (i.e., sections of continuous 1 in $label$), we define the continuous $label_r$ as follows:

$$\forall i \in [0, |label|], label_{\ell i} = \begin{cases} \left(1 - \frac{|s-i|}{\ell}\right)^{\frac{1}{2}} & \text{if: } s - \frac{\ell}{2} \leq i < s \\ 1 & \text{if: } s \leq i < e \\ \left(1 - \frac{|e-i|}{\ell}\right)^{\frac{1}{2}} & \text{if: } e \leq i < e + \frac{\ell}{2} \\ 0 & \text{if: } i < s \text{ and } e < i \end{cases} \tag{13}$$

When the buffer regions of two discontinuous outliers overlap, the label will be the superposition of these two orange curves with one as the maximum value. Using this new continuous label, one can compute TP/FP/TN/FN similarly as follows:

$$\begin{aligned} TP_\ell &= label_\ell^\top \cdot pred \\ FP_\ell &= (I - label_\ell)^\top \cdot pred \\ TN_\ell &= (I - label_\ell)^\top \cdot (I - pred) \\ FN_\ell &= label_\ell^\top \cdot (I - pred) \end{aligned} \tag{14}$$

The total number of positive points P in this case naively should be $P_{\ell_0} = TP_\ell + FN_\ell = label_\ell^\top \cdot I$. Here, we define it as:

$$\begin{aligned} P_\ell &= (label + label_\ell)^\top \cdot \frac{I}{2} \\ N_\ell &= |label_\ell| - P_\ell \end{aligned} \tag{15}$$

The reason is twofold. When the length of the outlier is several periods, P_{ℓ_0} and P_ℓ are similar because the ratio of buffer region to the whole anomaly region is small. When the length of the outlier is only half-period, the size of the buffer region is nearly two times the original abnormal region. In other words, to pursue false tolerance, the relative change we make to the ground truth is too significant. We use the average of $label$ and $label_\ell$ to limit this change.

We finally generalize the point-based *Recall*, *Precision*, and *FPR* to the range-based variants. Formally, following the definition of R and P as the set of anomalies range and detected predicted anomaly range (see Section 2.2), we define TPR_ℓ, FPR_ℓ, and $Precision_\ell$:

$$\begin{aligned} TPR_\ell = Recall_\ell &= \frac{TP_\ell}{P_\ell} * \sum_{R_i \in R} \frac{ExistenceR(R_i, P)}{|R|} \\ FPR_\ell &= \frac{FP_\ell}{N_\ell} \\ Precision_\ell &= \frac{TP_\ell}{TP_\ell + FP_\ell} \end{aligned} \tag{16}$$

Note that $TPR_r = Recall_r$. Moreover, for the recall computation, we incorporate the idea of Existence Reward [32], which is the ratio of the number of detected subsequence outliers to the total number of subsequence outliers. However, consistent with their work [32], we do not include the Existence ratio in the definition of range-precision. We can then compute R-AUC-ROC and R-AUC-PR using Equation 10 and Equation 11.

Relation between Range-ROC and Range-PR: PR curve is a supplement to the ROC curve. In a highly unbalanced dataset, because the number of positive points is too small, at the same level of FPR, it is easy to have a high TPR (or TPR_ℓ) at the cost of low precision. There are deep connections between ROC and PR [11]. First, ROC and PR have one-to-one mapping for a given dataset because the confusion matrix is uniquely determined given TPR and FPR. This relation is broken for the range method because we include an extra Existence factor for range-TPR. Therefore, the confusion matrix cannot be decided in the range-ROC space. Secondly, for a point-based version, if one ROC curve *dominates* another ROC curve, its corresponding PR curve would also dominate another one. Here, dominate means the curve is always higher or equal to another one. Because of the Existence factor, this rule is also lifted for the range definition. This is true only if both of the methods

have the same existence ratio. However, this is not always guaranteed. Finally, a maximized AUC does not necessarily correspond to a maximized AP. This holds for the range definition. For a robust evaluation, both measures should be used.

4.2 VUS: Volume Under the Surface

Our range-AUC family of measures choose the width of the buffer region to be half of a subsequence length ℓ of the time series. Such buffer length can be either set based on the knowledge of an expert (e.g., the usual size of arrhythmia in an electrocardiogram) or set automatically using the time series's period (which can easily be computed using either techniques based on cross-correlation or the Fourier transform). Since the period is an intrinsic property of the time series, we can compare various algorithms on the same basis. However, a different approach may get a slightly different period. In addition, there are multi-period time series. So other groups may get different range-AUC because of the difference in the period. As a matter of fact, the parameter ℓ, if not well set, can strongly influence range-AUC measures. To eliminate this influence, we introduce two generalizations of range-AUC family of measures.

The solution is to compute ROC and PR curves for different buffer lengths from 0 to the ℓ as shown in Figure 3(d). Therefore, the ROC and the PR curves become a surface in a three-dimensional space. Then, the overall accuracy measure corresponds to the Volume Under the Surface (VUS) for either the ROC surface (VUS-ROC) or PR surface (VUS-PR). As the R-AUC-ROC and R-AUC-PR are anomaly score threshold-independent measures, the VUS-ROC and VUS-PR are independent to both the threshold and buffer length. Formally, given $Th = [Th_0, Th_1, ...Th_N]$ with $0 = Th_0 < Th_1 < ... < Th_N = 1$, and $\mathcal{L} = [\ell_0, \ell_1, ..., \ell_L]$ with $0 = \ell_0 < \ell_1 < ... < \ell_L = \ell$, we define VUS-ROC as:

$$VUS\text{-}ROC = \frac{1}{4}\sum_{w=1}^{L}\sum_{k=1}^{N}\Delta^{(k,w)} * \Delta^{w}$$

$$\text{with: } \begin{cases} \Delta^{(k,w)} & = \Delta^{k}_{TPR_{\ell_w}} * \Delta^{k}_{FPR_{\ell_w}} + \Delta^{k}_{TPR_{\ell_{w-1}}} * \Delta^{k}_{FPR_{\ell_{w-1}}} \\ \Delta^{k}_{FPR_{\ell_w}} & = FPR_{\ell_w}(Th_k) - FPR_{\ell_w}(Th_{k-1}) \\ \Delta^{k}_{TPR_{\ell_w}} & = TPR_{\ell_w}(Th_{k-1}) + TPR_{\ell_w}(Th_k) \\ \Delta^{w} & = |\ell_w - \ell_{w-1}| \end{cases} \quad (17)$$

Similarly, We can compute VUS-PR as follows:

$$VUS\text{-}PR = \frac{1}{4}\sum_{w=1}^{L}\sum_{k=1}^{N}\Delta^{(k,w)} * \Delta^{w}$$

$$\text{with: } \begin{cases} \Delta^{(k,w)} & = \Delta^{k}_{Pr_{\ell_w}} * \Delta^{k}_{Re_{\ell_w}} + \Delta^{k}_{Pr_{\ell_{w-1}}} * \Delta^{k}_{Re_{\ell_{w-1}}} \\ \Delta^{k}_{Re_{\ell_w}} & = Recall_{\ell_w}(Th_k) - Recall_{\ell_w}(Th_{k-1}) \\ \Delta^{k}_{Pr_{\ell_w}} & = Precision_{\ell_w}(Th_{k-1}) + Precision_{\ell_w}(Th_k) \\ \Delta^{w} & = |\ell_w - \ell_{w-1}| \end{cases} \quad (18)$$

From the above equations, we observe that the computation of VUS measures requires $O(N * L)$. In comparison, range-AUC measures require $O(N)$. Thus, the application of VUS versus range-AUC depends on our knowledge of which buffer length to set. If one user knows which would be the most appropriate buffer length, range-AUC-based measures are preferable compared to VUS-based measures. However, if there exists an uncertainty on ℓ, then setting a range and using VUS increases the flexibility of the usage and the robustness of the evaluation. Finally, more parameters than ℓ can be included in VUS-based measures. If, in addition to ℓ, there is a need to define a range for an other parameter (such as the normal model length ℓ_{N_M} of NormA), the two dimensional surface is transformed into a three dimensional hyper-surface. In general, for P parameters, the value is the volume under a $|P| - 1$ hyper-surface.

Dataset	Size	Average Length	Average # Anomalies	Average # Abnormal Points	Average Abnormal Density %
Dodgers [17]	1	50400.0	133.0	5612.0	11.14
ECG [15]	52	230351.9	195.6	15634.0	6.8
IOPS [1]	58	102119.2	46.5	2312.3	2.1
KDD21 [18]	250	77415.06	1	196.5	0.56
MGAB [33]	10	100000.0	10.0	200.0	0.20
NAB [3]	58	6301.7	2.0	575.5	8.8
NASA-MSL [6]	27	2730.7	1.33	286.3	11.97
NASA-SMAP [6]	54	8066.0	1.26	1032.4	12.39
SensorScope [36]	23	27038.4	11.2	6110.4	22.5
YAHOO [19]	367	1561.2	5.9	10.7	0.70

Table 2: Summary characteristics of the public datasets of TSB-UAD.

This flexibility of VUS measures in evaluating methods while varying parameters of choice may have profound implications beyond time-series AD. For example, VUS measures are applicable across binary classification tasks for evaluating methods with a single quality value while considering different parameter choices. Similarly, the integration of VUS measures into the objective functions of machine learning methods may result in substantially more robust models. Interestingly, our current formulation enables easy integration of weights (summing to 1) in case there are auxiliary information about the importance of anomalies (or class labels). Consider for example the case where the downstream task is concerned with time-series forecasting. An anomaly in the beginning of the time series is less likely to influence the forecasting ability. However, an anomaly near the forecasted period may have significant impact to the forecasting model. Therefore, penalizing differently the positions of anomalies may be required. We plan to study such variants and generalizations of VUS in future works.

5 EXPERIMENTAL ANALYSIS

We now describe in detail our experimental analysis. The experimental section is organized as follows:

(1) In Section 5.1, we first start by introducing the datasets and the methods to evaluate the previously defined accuracy measures.
(2) In Section 5.2, we illustrate the limitations of existing measures with some selected qualitative examples.
(3) In Section 5.3, we continue by measuring quantitatively the benefits that our proposed measures bring in terms of robustness to lag, noise, and normality/abnormality ratio. We then evaluate the separability degree of accurate and inaccurate methods using the existing and our proposed approaches.
(4) In Section 5.4, we finally conduct an evaluation of the accuracy measures in which we analyse the variation of ranks that a AD method can have with regards to the accuracy measures used.

5.1 Experimental Setup and Settings

We implemented the experimental scripts in Python 3.8 with the following main dependencies: sklearn 0.23.0, tensorflow 2.3.0, pandas 1.2.5, and networkx 2.6.3. In addition, we used implementations from our TSB-UAD benchmark suite: www.timeseries.org/TSB-UAD. For reproducibility purposes, we make all our datasets, codes, and scripts publicly available.[1]

Datasets: For our evaluation purposes, we use the public datasets identified in our TSB-UAD benchmark. The latter corresponds to 10 datasets proposed in the past decades in the literature containing 900 time series with labeled anomalies. Specifically, each point in

[1] http://chaos.cs.uchicago.edu/tsb-uad/VUS.zip

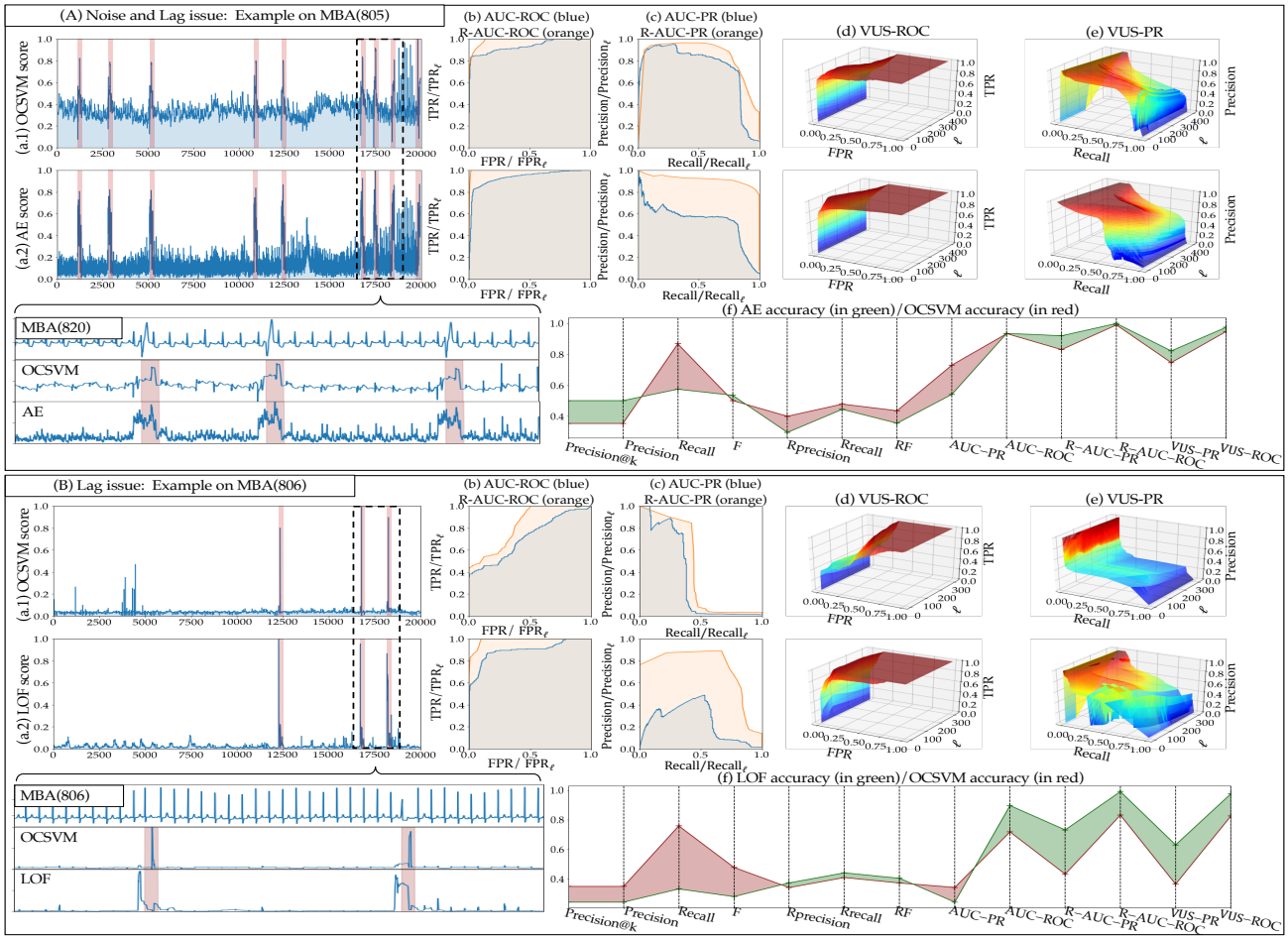

Figure 4: Comparison of evaluation measures (only our proposed measures are illustrated in subplots (b,c,d,e) but all others are summarized in subplots (f)) on two examples ((A)AE and OCSM applied on MBA(805) and (B) LOF and OCSVM applied on MBA(806)) illustrating the limitation when the anomaly score is either noisy or lagged with human labels.

every time series is labeled as normal or abnormal. Table 2 summarizes relevant characteristics of the datasets, including their size, length, and statistics about the anomalies. In more detail:

- **ECG** [15] is a standard electrocardiogram dataset and the anomalies represent ventricular premature contractions. Long series MBA(14046) is split to 47 series.
- **IOPS** [1] is a dataset with performance indicators that reflect the scale, quality of web services, and health status of a machine.
- **KDD21** [18] is a composite dataset released in a recent SIGKDD 2021 competition with 250 time series.
- **MGAB** [33] is composed of Mackey-Glass time series with non-trivial anomalies. Mackey-Glass data series exhibit chaotic behavior that is difficult for the human eye to distinguish.
- **NAB** [3] is composed of labeled real-world and artificial time series including AWS server metrics, online advertisement clicking rates, real time traffic data, and a collection of Twitter mentions of large publicly-traded companies.
- **NASA-SMAP** and **NASA-MSL** [6] are two real spacecraft telemetry data with anomalies from Soil Moisture Active Passive (SMAP) satellite and Curiosity Rover on Mars (MSL).
- **SensorScope** [36] is a collection of environmental data, such as temperature, humidity, and solar radiation, collected from a typical tiered sensor measurement system.
- **Yahoo** [19] is a dataset published by Yahoo labs consisting of real and synthetic time series based on the real production traffic to some of the Yahoo production systems.

Anomaly Detection Methods: For the experimental evaluation, we consider the following AD baselines.

- **Isolation Forest (IForest)** [22] constructs binary trees based on random space splitting. The nodes (subsequences in our specific case) with shorter path lengths to the root (averaged over every random tree) are more likely to be anomalies.
- **The Local Outlier Factor (LOF)** [10] computes the ratio of the neighboring density to the local density.
- **Matrix Profile (MP)** [38] detects as anomaly the subsequence with the most significant 1-NN distance.

- **NormA** [8] identifies the normal patterns based on clustering and calculates each point's effective distance to the normal patterns weighted using statistical criteria of these patterns.
- **Principal Component Analysis (PCA)** [2] projects data to a lower-dimensional hyperplane, and data points with a significant distance from this plane can be identified as outliers.
- **Autoencoder (AE)** [29] projects data to the lower-dimensional latent space and reconstructs the data, and outliers are expected to have more evident reconstruction deviation.
- **LSTM-AD** [23] use an LSTM network that from the current subsequence tries to predict the following value. The error prediction is then used to identify anomalies.
- **Polynomial Approximation (POLY)** [20] fits a polynomial model that tries to predict the values of the data series from the previous subsequences. The outliers are detected by measuring the prediction error.
- **CNN** [24] build, using a convolutional-based deep neural network, a correlation between current and previous subsequences, and the outliers are detected by the deviation between the prediction and the actual value.
- **One-class Support Vector Machines (OCSVM)** [30] is a support vector method that fits the normal training dataset and finds the normal data's boundary.

5.2 Qualitative Analysis

We first evaluate qualitatively the different accuracy evaluation measures. We pick two examples that illustrate well the motivation and the limitation to lag and noise. These two examples are depicted in Figure 4. The first example in Figure 4(A) corresponds to the application of OCSVM and AE on the MBA(805) dataset.

We observe in Figure 4(A)(a.1) and (a.2) that both scores identify most of the anomalies (highlighted in red). However, the OCSVM score points to more false positives (at the end of the time series) and only captures small sections of the anomalies. On the contrary, the AE score points to fewer false positives and captures all abnormal subsequences. Thus we can conclude that, visually, AE should obtain a better accuracy score than OCSVM. Nevertheless, we also observe that the AE score is lagged with the labels and contains more noise. The latter has a significant impact on the accuracy of evaluation measures. First, Figure 4(A)(c) is showing that AUC-PR is better for OCSM (0.73) than for AE (0.57). This is contradictory with what is visually observed from Figure 4(A)(a.1) and (a.2). However, when using our proposed measure R-AUC-PR, OCSVM obtains a lower score (0.83) than AE (0.89). This confirms that a buffer region before the labels helps to capture the true value of an anomaly score. Overall, Figure 4(A)(f) is showing in green and red the evolution of accuracy score for the 13 accuracy measures for AE and OCSVM, respectively. The latter shows that, in addition to Precision@k and Precision, our proposed approach captures the quality order between the two methods well.

We now present a second example illustrated in Figure 4(B). In this case, we demonstrate the anomaly score of OCSVM and LOF (depicted in Figure 4(B)(a.1) and (a.2)) applied on the MBA(806) dataset. In this case, we observe that both methods produce the same level of noise. However, LOF points to fewer false positives and captures more sections of the abnormal subsequences than OCSVM. Nevertheless, the LOF score is slightly lagged with the labels such that the maximum values in the LOF score are slightly outside of the labeled sections. Thus, as illustrated in Figure 4(B)(f), even though

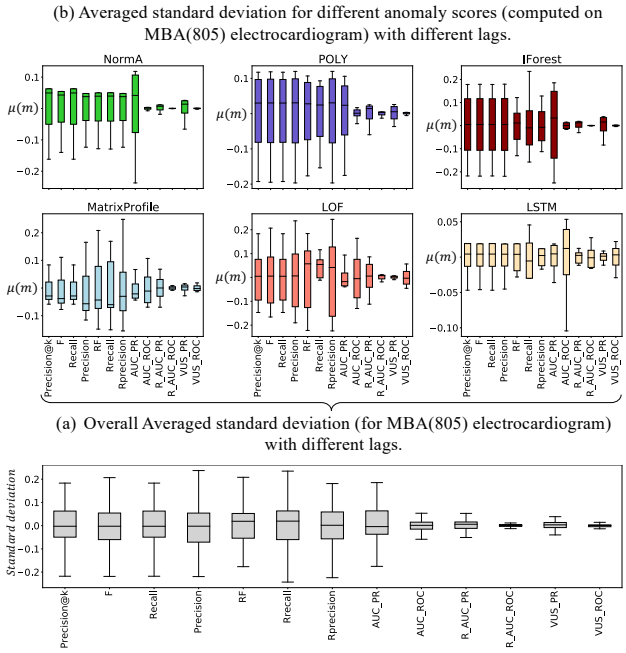

Figure 5: For each methods, we compute 10 times the accuracy measures with random lag $\ell \in [-0.25 * \ell, 0.25 * \ell]$ injected in the anomaly score. We center the accuracy average to 0.

we can visually consider that LOF is performing better than OCSM, all usual measures (Precision, Recall, F, precision@k, and AUC-PR) are judging OCSM better than AE. On the contrary, measures that consider lag (Rprecision, Rrecall, RF) rank the methods correctly. However, due to threshold issues, these measures are very close for the two methods. Overall, only AUC-ROC and our proposed measures are giving a higher score for LOF than for OCSVM.

5.3 Quantitative Analysis

Until now, we illustrated with specific examples several of the limitations of current measures. Due to space restrictions, we omit such examples and, instead, we evaluate statistically and quantitatively the robustness and validity of our proposed measures versus currently used measures. We first evaluate the robustness to noise, lag, and normal versus abnormal points ratio. We then measure their ability to separate accurate and inaccurate methods.

Sensitivity Analysis: We first analyze the sensitivity of different approaches quantitatively to different factors: (i) lag, (ii) noise, and (iii) normal/abnormal ratio. As already mentioned, these factors are realistic. For instance, lag can be either introduced by the anomaly detection methods (such as methods that produce a score per subsequences are only high at the beginning of abnormal subsequences) or by human labeling approximation. Furthermore, even though lag and noises are injected, an optimal evaluation metric should not vary significantly. Therefore, we aim to measure the variance of the different evaluation measures when we vary the lag, noise, and the normal/abnormal ratio. Thus, we proceed as follows:

(1) For each anomaly detection method, we first compute the anomaly score on a given time series.

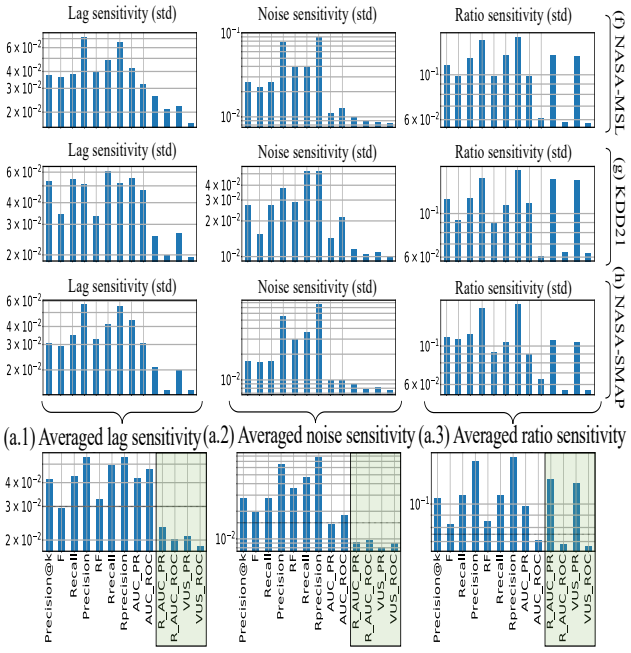

Figure 6: Sensitivity Analysis: Over the entire benchmark, average standard deviation of the accuracy values of the measures under varying (a.1) lag, (a.2) noise, (a.3) normal/abnormal ratio. Sensitivity Analysis for (f) NASA-MSL, (g) KDD21 and (h) NASA-SMAP

(2) We then inject either lag l, noise n or change the normal/abnormal ratio r. For 10 different values of $l \in [-0.25 * \ell, 0.25 * \ell]$, $n \in [-0.05 * (max(S_T) - min(S_T)), 0.05 * (max(S_T) - min(S_T))]$ and $r \in [0.01, 0.2]$, we compute the 13 different evaluation measures.
(3) For each evaluation measure, we compute the standard deviation of the ten different values. Figure 5(b) box plots depict the different lag values for six AD methods applied on the MBA(805) data series in the ECG dataset.
(4) We compute the average standard deviation for the 13 different AD quality measures. For example, figure 5(a) depicts the average standard deviation for ten different lag values over the AD methods applied on the MBA(805) time series.
(5) We compute the average standard deviation for the every time series in each dataset (as illustrated in Figure 6(b,c,d) for three datasets of the benchmark. The results for all datasets of the benchmark are depicted in Figure 12(b,c,d,e,f,g,h,i,j) in the appendix, which we include in the zip file along with the codes and scripts).
(6) We compute the average standard deviation for the every dataset (as illustrated in Figure 6(a.1) for lag, Figure 6(a.2) for noise and Figure 6(a.3) for normal/abnormal ratio).
(7) We finally compute the Wilcoxon signed-rank test [35] and display the critical diagram over the average standard deviation for every data series (as illustrated in Figure 8(a.1) for lag, Figure 8(a.2) for noise and Figure 8(a.3) for normal/abnormal ratio).

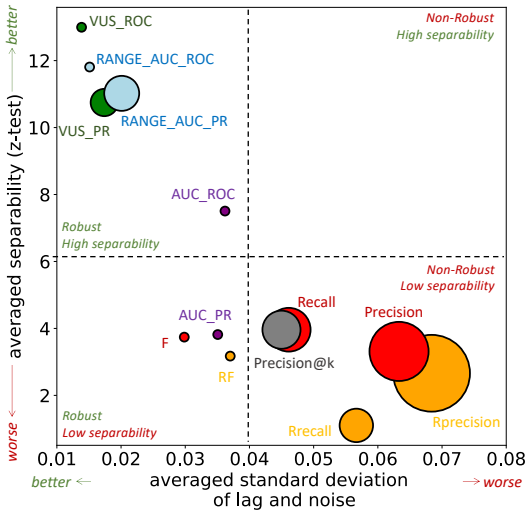

Figure 7: Evaluation of all measures based on: (y-axis) their separability (measured as the averaged z-test between the accuracy values distributions of accurate and inaccurate methods), (x-axis) average standard deviation of the accuracy values when varying lag and noise, (scatter size) average standard deviation of the accuracy values when varying the normal/abnormal ratio.

The methods with the smallest standard deviation can be considered more robust to lag, noise, or normal/abnormal ratio from the above framework. First, as stated in the introduction, we observe that non-threshold-based measures (such as AUC-ROC and AUC-PR) are indeed robust to noise (see Figure 6(a.2)), but not to lag. Figure 8(a.1) demonstrates that our proposed measures VUS-ROC, VUS-PR, R-AUC-ROC, and R-AUC-PR are significantly more robust to lag. Similarly, Figure 8(a.2) confirms that our proposed measures are significantly more robust to noise. However, we observe that, among our proposed measures, only VUS-ROC and R-AUC-ROC are robust to the normal/abnormal ratio and not VUS-PR and R-AUC-PR. This can be explained by the fact that Precision-based measure varies significantly when the ratio of positive versus negative labels changes. This is confirmed by Figure 6(a.3), in which we observe that both Precision and Rprecision have a high standard deviation. Overall, we observe that VUS-ROC is significantly more robust to lag, noise, and normal/abnormal ratio than other measures.

Separability Analysis: We now evaluate the separability capacities of the different evaluation metrics. We thus manually select accurate and inaccurate anomaly detection methods and verify if the accuracy evaluation scores are indeed higher for the accurate than for the inaccurate methods. Figure 9 depicts the latter separability analysis applied on the MBA(805) dataset. The accurate and inaccurate anomaly scores are plotted in green and red, respectively. We then consider 12 different pairs of accurate/inaccurate methods among the eight previously mentioned anomaly scores. We slightly modify each score 50 different times in which we inject lag and noises and compute the accuracy measures. Figure 9(a.4) is separated into four different subplots corresponding to 4 pairs (selected among the twelve different pairs due to lack of space). Each subplot corresponds to two box plots per accuracy measure. The green and red box plots correspond to the 50 accuracy measures on the

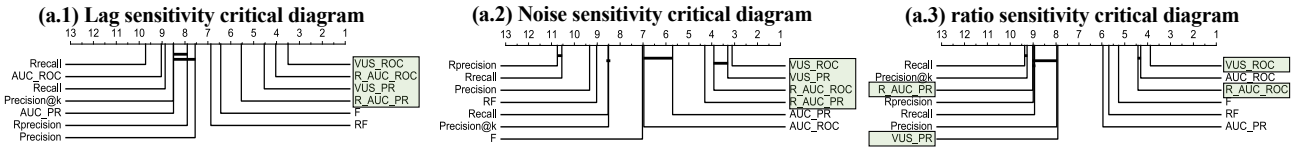

Figure 8: Critical difference diagram computed using the signed-rank Wilkoxon test (with $\alpha = 0.1$) for the sensitivity to (a.1) lag, (a.2) noise and (a.3) normal/abnormal ratio.

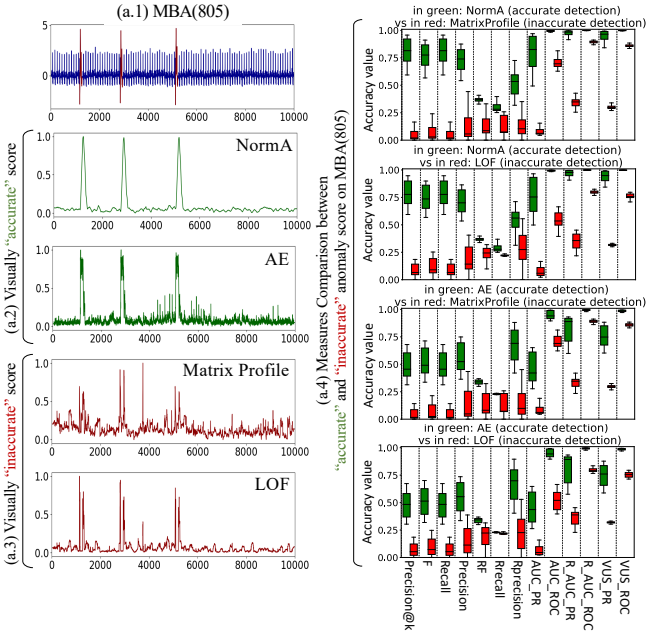

Figure 9: Separability analysis applied on 4 pairs of accurate (in green) and inaccurate (in red) methods on MBA(805) data series.

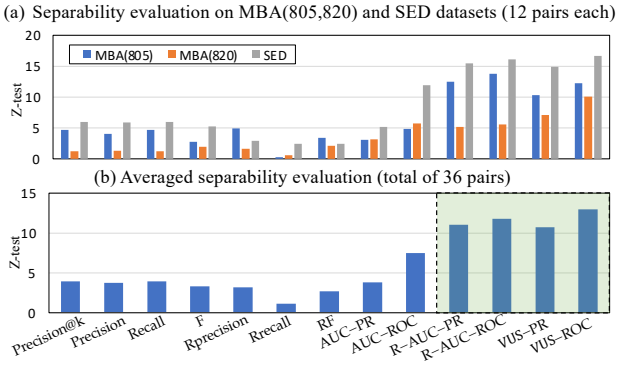

Figure 10: Overall separability analysis (averaged z-test between the accuracy values distributions of accurate and inaccurate methods) applied on 36 pairs on 3 datasets.

accurate and inaccurate methods. If the red and green box plots are well separated, we can conclude that the corresponding accuracy measures are separating the accurate and inaccurate methods well. We thus observe that some accuracy measures (such as VUS-ROC) are more separable than others (such as RF). We thus measure the separability of the two pox plot by computing the Z-test.

We now aggregate all the results and compute the average Z-test for all pairs of accurate/inaccurate datasets (examples can be found in Figure 9(a.2) for accurate and (a.3) for inaccurate anomaly score). Next, we perform the same operation over three different data series: MBA (805), MBA(820), and SED. Then, we depict the average Z-test for these three datasets in Figure 10(a). Finally, we show the average Z-test for all datasets in Figure 10(b).

We observe that our proposed VUS-based and Range-based measures are significantly more separable than other current accuracy measures (up to two times for than AUC-ROC, the best measures of all current ones). Furthermore, when analyzed in detail in Figure 9 and Figure 10, we confirm that VUS-based and Range-based are more separable over all three datasets.

Global Analysis: Overall, we observe that VUS-ROC appears to be the most robust (see Figure 8) and separable (see Figure 10) measures. On the contrary, Precision and Rprecision are non-robust and non-separable. In order to visualize the global statistical analysis, we merge the sensitivity and the separability analysis into one plot. Figure 7 depicts one scatter point per accuracy measure. The x-axis represents the averaged standard deviation to lag and noise (averaged values from Figure 6(a.1) and (a.2)). The y-axis corresponds to the averaged Z-test (averaged value from Figure 10). Finally, the size of the points corresponds to the sensitivity to normal/abnormal ratio (values from Figure 6(a.3)). Figure 7 demonstrates that our proposed measures (that are at the top left section of the plot) are both the most robust and separable. Among all current accuracy measures, only AUC-ROC is on the top left section of the plot. Usual measures such as F, RF, AUC-ROC, AUC-PR are on the bottom left section of the plot. The latter underlines that these usual measures are robust but non-separable. Nevertheless, we observe that VUS-PR and Range-AUC-PR are sensitive to the normal/abnormal ratio, even though separable and robust to lag and noise. This indicates that these measures should be used with caution when applied to unknown data series with very different normal/abnormal ratios. Overall, Figure 7 confirms the interest and the superiority of our proposed measures, especially for the VUS-ROC measure.

5.4 Accuracy Evaluation

In this section, we analyze the accuracy of the anomaly detection methods provided by the 13 accuracy measures. The objective is to observe the changes in the global ranking of the anomaly detection methods. For that purpose, we formulate the following assumptions. First, we assume that the data series in each benchmark dataset

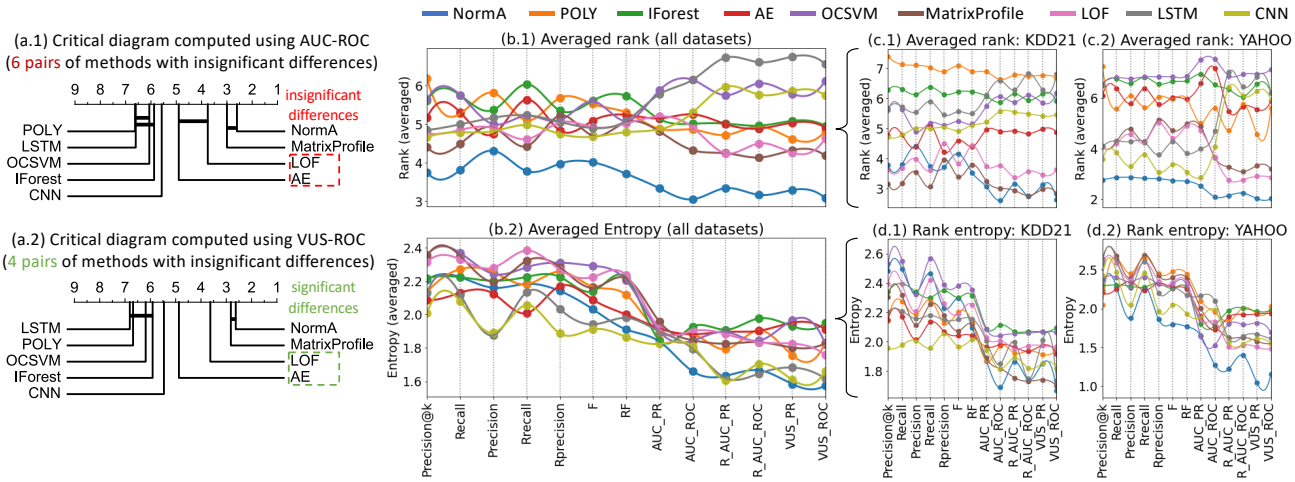

Figure 11: Accuracy evaluation of the anomaly detection methods. Analysis of the (b.1) averaged rank and (b.2) averaged rank entropy for each methods and each accuracy measures over the entire benchmark. Critical difference diagram computed using the signed-rank Wilkoxon test (with $\alpha = 0.1$) when using (a.1) AUC-ROC and (a.2) VUS-ROC.

are similar (i.e., from the same domain and sharing some common characteristics). As a matter of fact, we can assume that an anomaly detection method should perform similarly on these data series of a given dataset. This is confirmed when observing that the best anomaly detection methods are not the same based on which dataset was analyzed. Thus the ranking of the anomaly detection methods should be different for different datasets but similar for every data series in each dataset. Therefore, for a given method A and a given dataset D containing data series of the same type and domain, we assume that a good accuracy measure results in a low entropy for the different ranks for the method A across the dataset D.

We now compute the accuracy measures for the nine different methods (we compute the anomaly scores ten different times, and we use the average accuracy). Figure 11(c) reports the average ranking of the anomaly detection methods obtained on two of the nine datasets. The x-axis corresponds to the different accuracy measures. We first observe that the rankings are more separated using Range-AUC and VUS measures for these two datasets. Figure 11(b.1) depicts the average ranking over the entire benchmark. The latter confirms the previous observation that VUS measures provide more separated rankings than threshold-based and AUC-based measures. We also observe an interesting ranking evolution for the YAHOO dataset illustrated in Figure 11(c.2). We notice that both LOF and MatrixProfile (brown and pink curve) have a low rank (between 4 and 5) using threshold and AUC-based measures. However, we observe that their ranks increase significantly for range-based and VUS-based measures (between 2.5 and 3). As we noticed by looking at specific examples (see Figure 5.2), LOF and MatrixProfile can suffer from a lag issue even though the anomalies are well identified. Therefore, the range-based and VUS-based measures better evaluate these two methods' detection capability.

Overall, we observe from the ranking curves that the ranks seem more chaotic for threshold-based than AUC-based, Range-AUC-based, and VUS-based measures. We now quantify it statistically. For that matter, we compute the Shannon Entropy of the ranks of each anomaly detection method. In practice, we extract the ranks of methods across one dataset and compute Shannon's Entropy of the different ranks. Figure 11(d) depicts the Entropy of each of the nine methods for two out of 9 datasets. Figure 11(b.2) illustrates the averaged Entropy for all datasets in the benchmark. We observe that both for the general case (Figure 11(b.2)) and some specific cases (Figure 11(d)), the Entropy is reducing when using AUC-based, Range-AUC based, and VUS-based measures. We report the lowest Entropy for VUS-based measures. More significantly, we notice a significant drop between threshold-based and AUC-based. This confirms that the ranks provided by AUC and VUS-based measures are consistent for data series belonging to one specific dataset.

Therefore, based on the assumption formulated at the beginning of the section, we can thus conclude that AUC, range-AUC, and VUS-based measures are providing more consistent rankings. Finally, as illustrated in Figure 11, we also observe that VUS-based measures result in the most ordered and similar rankings for data series from the same type and domain.

6 CONCLUSIONS

Time-series AD is a challenging problem, and an active area of research. Given the multitude of solutions proposed in the literature, it is important to be able to properly evaluate them. In this paper, we demonstrate the limitations of threshold-based accuracy measures for the time-series AD methods. Even though, AUC-based measures solve the threshold issues, we show that they cannot handle lag and noise. Overall, we experimentally show that the proposed VUS-based measures are more robust, and better separate accurate methods from inaccurate ones.

REFERENCES

[1] [n.d.]. http://iops.ai/dataset_detail/?id=10.
[2] Charu C. Aggarwal. 2017. *Outlier Analysis* (2 ed.). Springer International Publishing. https://doi.org/10.1007/978-3-319-47578-3
[3] Subutai Ahmad, Alexander Lavin, Scott Purdy, and Zuha Agha. 2017. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing* 262 (2017), 134–147. https://doi.org/10.1016/j.neucom.2017.04.070
[4] Anthony Bagnall, Richard L. Cole, Themis Palpanas, and Kostas Zoumpatianos. 2019. Data Series Management (Dagstuhl Seminar 19282). *Dagstuhl Reports* 9, 7

(2019), 24–39. https://doi.org/10.4230/DagRep.9.7.24

[5] V. Barnet and T. Lewis. 1994. *Outliers in Statistical Data*. John Wiley and Sons, Inc.

[6] Pawel Benecki, Szymon Piechaczek, Daniel Kostrzewa, and Jakub Nalepa. 2021. Detecting Anomalies in Spacecraft Telemetry Using Evolutionary Thresholding and LSTMs. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion* (Lille, France) *(GECCO '21)*. Association for Computing Machinery, New York, NY, USA, 143–144. https://doi.org/10.1145/3449726.3459411

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

[8] Paul Boniol, Michele Linardi, Federico Roncallo, Themis Palpanas, Mohammed Meftah, and Emmanuel Remy. 2021. Unsupervised and scalable subsequence anomaly detection in large data series. *The VLDB Journal* (March 2021). https://doi.org/10.1007/s00778-021-00655-8

[9] Paul Boniol and Themis Palpanas. 2020. Series2Graph: Graph-Based Subsequence Anomaly Detection for Time Series. *Proc. VLDB Endow.* 13, 12 (July 2020), 1821–1834. https://doi.org/10.14778/3407790.3407792

[10] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. *ACM SIGMOD Record* 29, 2 (May 2000), 93–104. https://doi.org/10.1145/335191.335388

[11] Jesse Davis and Mark Goadrich. 2006. The Relationship between Precision-Recall and ROC Curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*. Association for Computing Machinery, New York, NY, USA, 233–240. https://doi.org/10.1145/1143844.1143874

[12] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27, 8 (2006), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010 ROC Analysis in Pattern Recognition.

[13] Anthony J Fox. 1972. Outliers in time series. *Journal of the Royal Statistical Society: Series B (Methodological)* 34, 3 (1972), 350–363.

[14] Sam George. 2019 (accessed August 15, 2020). *IoT Signals report: IoT's promise will be unlocked by addressing skills shortage, complexity and security.* https://blogs.microsoft.com/blog/2019/07/30/.

[15] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 101, 23 (June 2000), E215–220. https://doi.org/10.1161/01.cir.101.23.e215

[16] Mark Hung. 2017. Leading the iot, gartner insights on how to lead in a connected world. *Gartner Research* (2017), 1–29.

[17] Alexander Ihler, Jon Hutchins, and Padhraic Smyth. 2006. Adaptive Event Detection with Time-Varying Poisson Processes. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Philadelphia, PA, USA) *(KDD '06)*. Association for Computing Machinery, New York, NY, USA, 207–216. https://doi.org/10.1145/1150402.1150428

[18] E. Keogh, T. Dutta Roy, U. Naik, and A Agrawal. [n.d.]. Multi-dataset Time-Series Anomaly Detection Competition 2021, https://compete.hexagon-ml.com/practice/competition/39/.

[19] N. Laptev, S. Amizadeh, and Y. Billawala. 2015. *S5 - A Labeled Anomaly Detection Dataset, version 1.0(16M).* https://webscope.sandbox.yahoo.com/catalog.php?datatype=s&did=70

[20] Zhi Li, Hong Ma, and Yongbing Mei. 2007. A Unifying Method for Outlier and Change Detection from Data Streams Based on Local Polynomial Fitting. In *Advances in Knowledge Discovery and Data Mining (Lecture Notes in Computer Science)*, Zhi-Hua Zhou, Hang Li, and Qiang Yang (Eds.). Springer, Berlin, Heidelberg, 150–161. https://doi.org/10.1007/978-3-540-71701-0_17

[21] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation Forest. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (ICDM '08)*. IEEE Computer Society, Washington, DC, USA, 413–422. https://doi.org/10.1109/ICDM.2008.17

[22] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*. 413–422. https://doi.org/10.1109/ICDM.2008.17 ISSN: 2374-8486.

[23] Pankaj Malhotra, L. Vig, Gautam M. Shroff, and Puneet Agarwal. 2015. Long Short Term Memory Networks for Anomaly Detection in Time Series. In *ESANN*.

[24] M. Munir, S. A. Siddiqui, A. Dengel, and S. Ahmed. 2019. DeepAnT: A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series. *IEEE Access* 7 (2019), 1991–2005. https://doi.org/10.1109/ACCESS.2018.2886457

[25] Irene CL Ng and Susan YL Wakenshaw. 2017. The Internet-of-Things: Review and research directions. *International Journal of Research in Marketing* 34, 1 (2017), 3–21.

[26] ES Page. 1957. On problems in which a change in a parameter occurs at an unknown point. *Biometrika* 44, 1/2 (1957), 248–252.

[27] Themis Palpanas. 2015. Data Series Management: The Road to Big Sequence Analytics. *SIGMOD Rec.* 44, 2 (Aug. 2015), 47–52. https://doi.org/10.1145/2814710.2814719

[28] 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

[29] Mayu Sakurada and Takehisa Yairi. 2014. Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis* (Gold Coast, Australia QLD, Australia) *(MLSDA'14)*. Association for Computing Machinery, New York, NY, USA, 4–11. https://doi.org/10.1145/2689746.2689747

[30] Bernhard Schölkopf, Robert Williamson, Alex Smola, John Shawe-Taylor, and John Platt. 1999. Support vector method for novelty detection. In *Proceedings of the 12th International Conference on Neural Information Processing Systems (NIPS'99)*. MIT Press, Cambridge, MA, USA, 582–588.

[31] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos. 2006. Online Outlier Detection in Sensor Data Using Non-Parametric Models. In *Proceedings of the 32nd International Conference on Very Large Data Bases* (Seoul, Korea) *(VLDB '06)*. VLDB Endowment, 187–198.

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

[33] Markus Thill, Wolfgang Konen, and Thomas Bäck. 2020. *MarkusThill/MGAB: The Mackey-Glass Anomaly Benchmark, https://doi.org/10.5281/zenodo.3762385*. https://doi.org/10.5281/zenodo.3762385

[34] Ruey S Tsay. 1988. Outliers, level shifts, and variance changes in time series. *Journal of forecasting* 7, 1 (1988), 1–20.

[35] Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1, 6 (1945), 80–83. http://www.jstor.org/stable/3001968

[36] Yuan Yao, Abhishek Sharma, Leana Golubchik, and Ramesh Govindan. 2010. Online anomaly detection for sensor systems: A simple and efficient approach. *Performance Evaluation* 67, 11 (2010), 1059–1075. https://doi.org/10.1016/j.peva.2010.08.018 Performance 2010.

[37] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. 2016. Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. 1317–1322. https://doi.org/10.1109/ICDM.2016.0179

[38] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Zachary Zimmerman, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. 2018. Time series joins, motifs, discords and shapelets: a unifying view that exploits the matrix profile. *Data Mining and Knowledge Discovery* 32, 1 (Jan. 2018), 83–123. https://doi.org/10.1007/s10618-017-0519-9

[39] Kostas Zoumpatianos and Themis Palpanas. 2018. Data Series Management: Fulfilling the Need for Big Sequence Analytics. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. 1677–1678. https://doi.org/10.1109/ICDE.2018.00211

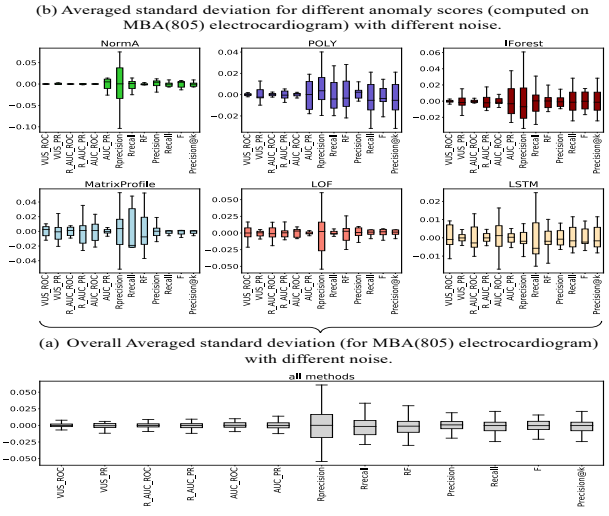

Figure 13: For each method, we compute 10 times the accuracy measures with random noise (noise ratio $n \in [-0.05 * (max(S_T) - min(S_T)), 0.05 * (max(S_T) - min(S_T))]$) injected in the anomaly score. We center the accuracy average to 0.

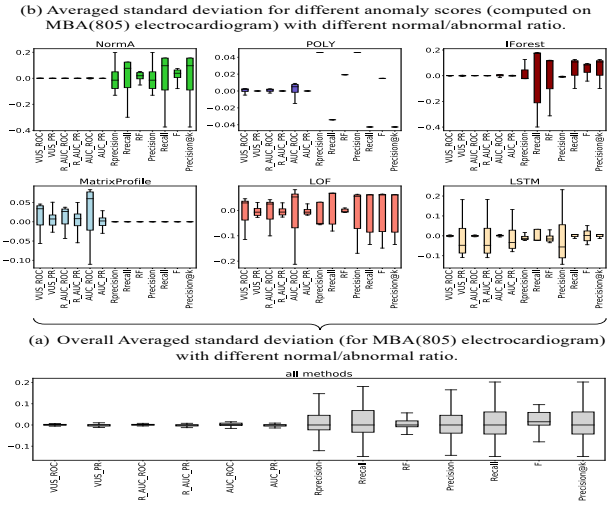

Figure 14: For each method, we compute 10 times the accuracy measures when we vary the normal abnormal ratio ($r \in [0.01, 0.2]$). We center the accuracy average to 0.

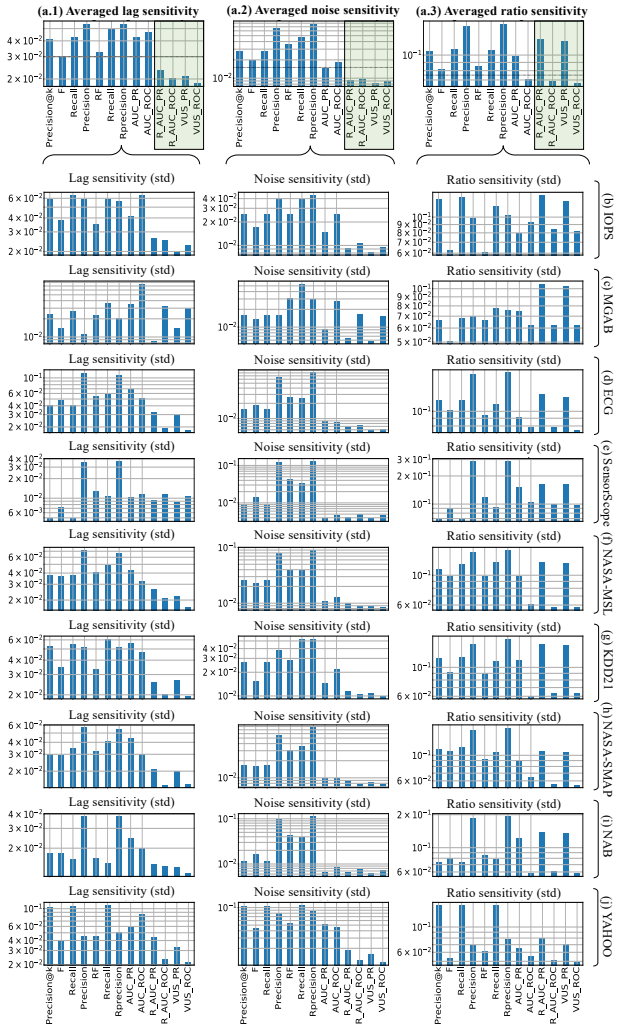

Figure 12: Sensitivity (lag, noise and normal/abnormal ratio) Analysis per dataset.

Appendix A SENSITIVITY EXPERIMENTAL EVALUATION RESOURCES

Appendix B SEPARABILITY EXPERIMENTAL EVALUATION RESOURCES

Appendix C ACCURACY EXPERIMENTAL EVALUATION RESOURCES

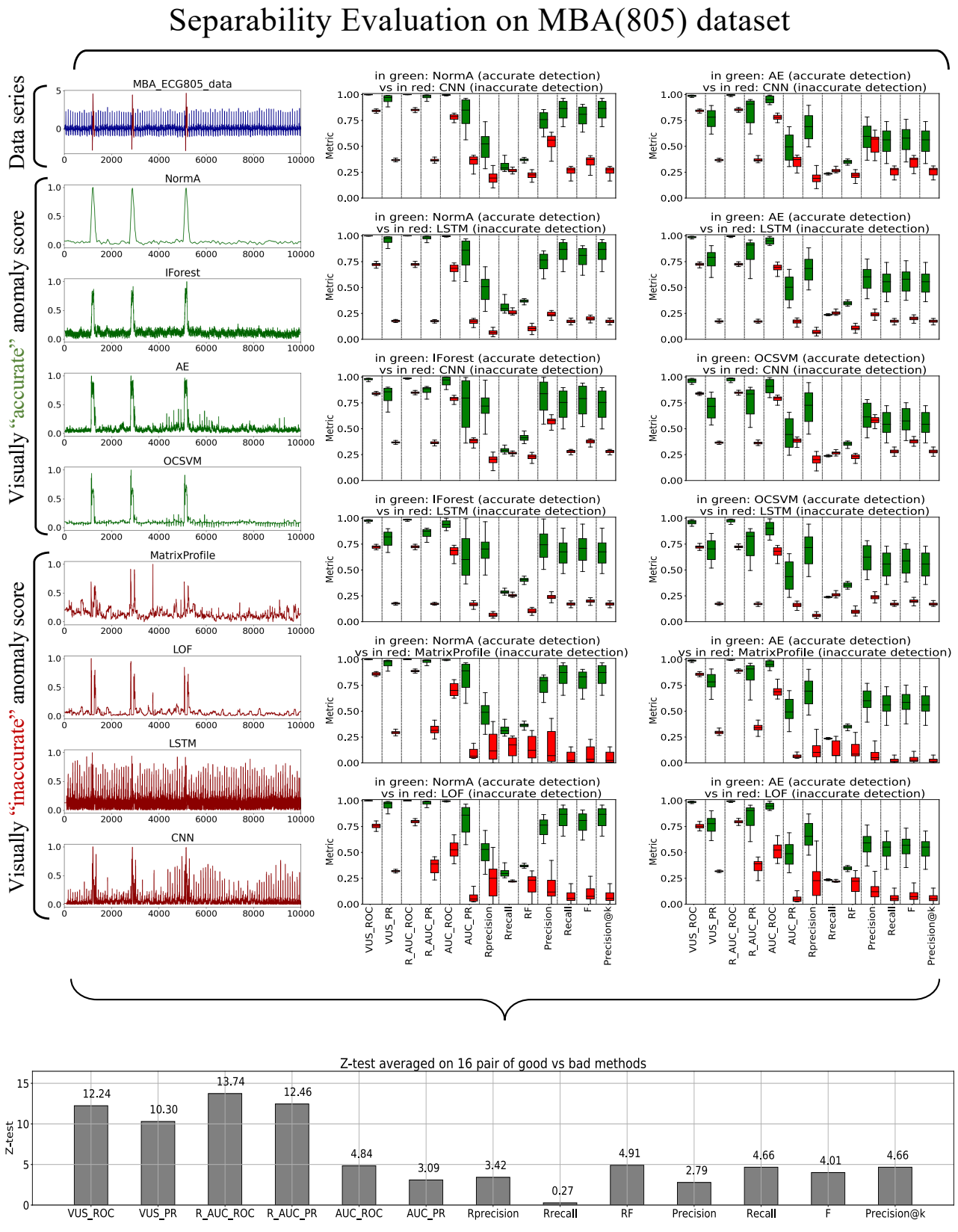

Figure 15: Separability analysis applied on 12 pairs of accurate/inaccurate methods on MBA(805) data series.

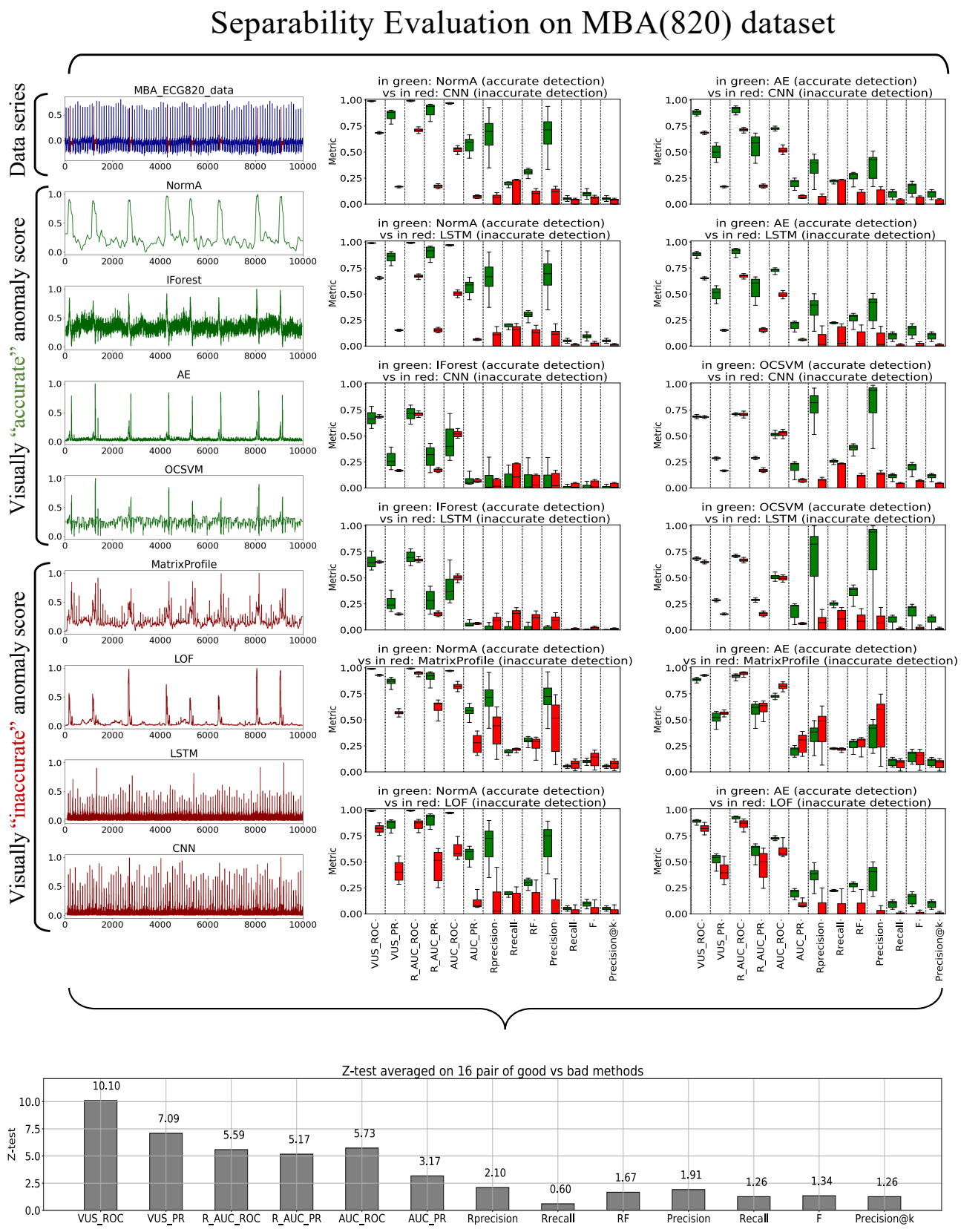

Figure 16: Separability analysis applied on 12 pairs of accurate/inaccurate methods on MBA(820) data series.

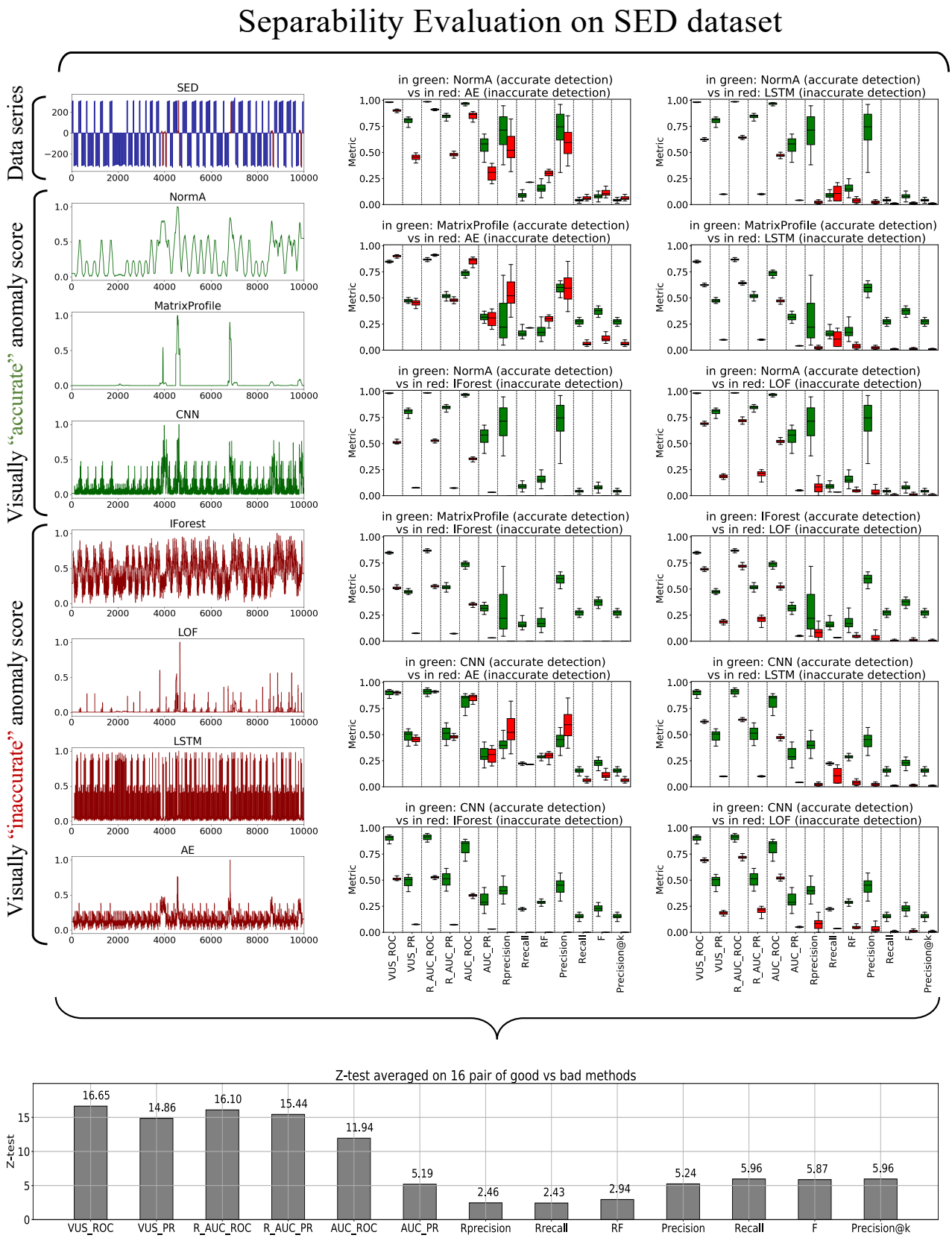

Figure 17: Separability analysis applied on 12 pairs of accurate/inaccurate methods on SED data series.

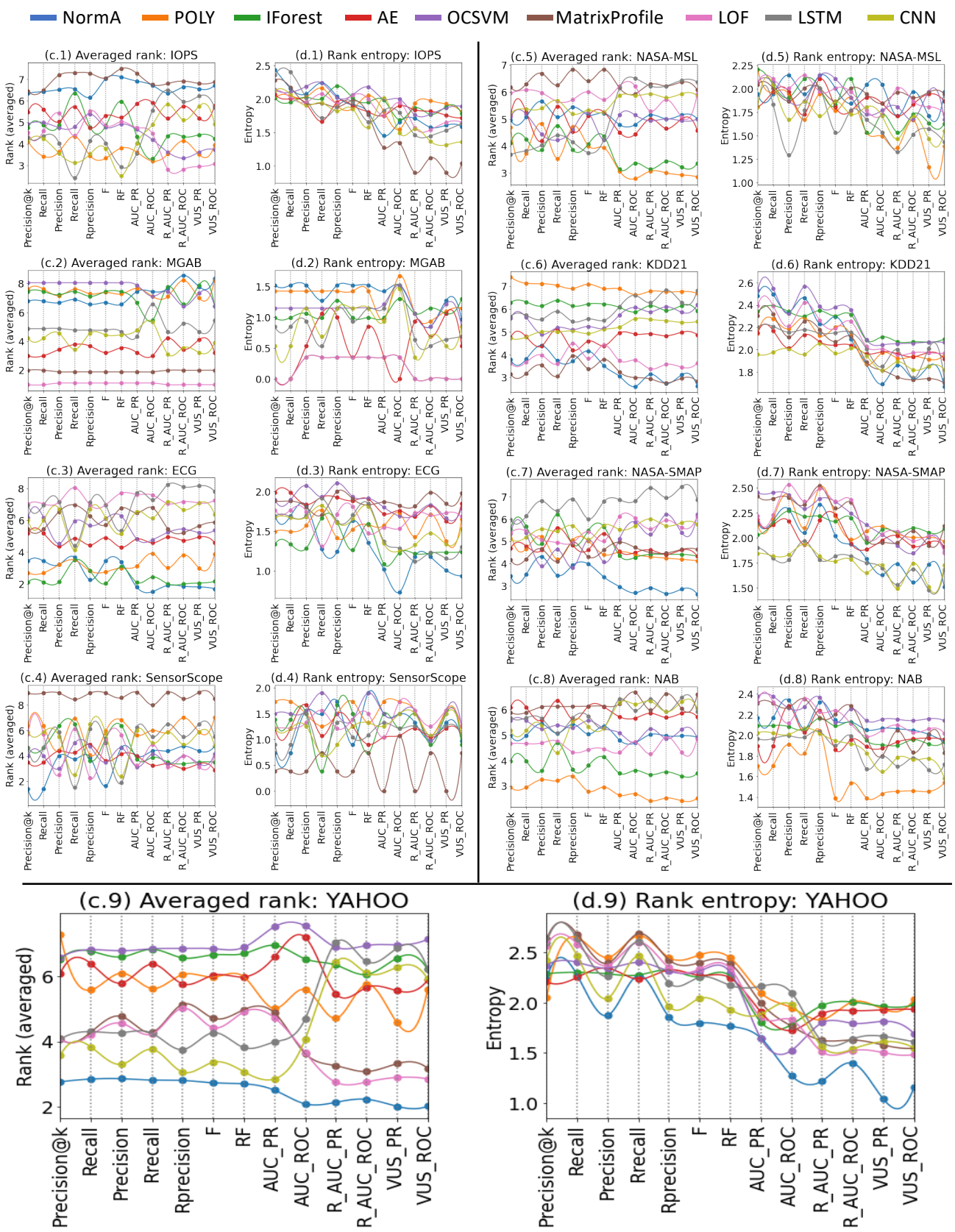

Figure 18: Anomaly detection rank and Entropy for these ranks for nine datasets of the benchmark.