



# Navigating the metric maze: a taxonomy of evaluation metrics for anomaly detection in time series

Sondre Sørbø<sup>1</sup> · Massimiliano Ruocco<sup>1,2</sup>

Received: 20 February 2023 / Accepted: 24 October 2023 / Published online: 18 November 2023  
© The Author(s) 2023

## Abstract

The field of time series anomaly detection is constantly advancing, with several methods available, making it a challenge to determine the most appropriate method for a specific domain. The evaluation of these methods is facilitated by the use of metrics, which vary widely in their properties. Despite the existence of new evaluation metrics, there is limited agreement on which metrics are best suited for specific scenarios and domains, and the most commonly used metrics have faced criticism in the literature. This paper provides a comprehensive overview of the metrics used for the evaluation of time series anomaly detection methods, and also defines a taxonomy of these based on how they are calculated. By defining a set of properties for evaluation metrics and a set of specific case studies and experiments, twenty metrics are analyzed and discussed in detail, highlighting the unique suitability of each for specific tasks. Through extensive experimentation and analysis, this paper argues that the choice of evaluation metric must be made with care, taking into account the specific requirements of the task at hand.

**Keywords** Time series · Anomaly detection · Evaluation · Taxonomy

---

Responsible editor: Eamonn Keogh.

✉ Sondre Sørbø  
sondre.sorbo@sintef.no

Massimiliano Ruocco  
massimiliano.ruocco@sintef.no

<sup>1</sup> Sintef Digital, Trondheim, Norway

<sup>2</sup> Dept. of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway



For example, Kim et al. (2022b) criticize the point-adjust metric, and show that a detection algorithm outputting random noise is expected to produce very good scores, and capable of outperforming state of the art methods on most of the common benchmark datasets. The same conclusion is reached experimentally by Doshi et al. (2022). Kim et al. (2022a) include a review of several TSAD evaluation metrics from the perspective of industrial control systems, and discuss several properties required for the metrics. Wu et al. (2022) analyse the most commonly used TSAD datasets and find that the majority suffer from flaws such as trivial anomalies, unrealistic anomaly density, mislabelled ground truth, and a high ratio of anomalies at the end of the time series. To address these issues, they introduce a new benchmark dataset, the UCR time series anomaly archive, and also discuss potential issues with the evaluation metrics. Finally, Paparrizos et al. (2022b) point out the lack of consensus regarding the appropriate datasets for benchmarking TSAD algorithms and present a benchmark suite derived from a combination of previous TSAD datasets and transformed classification datasets, which have been subjected to various transformations to increase the complexity and difficulty of the benchmark. They include several evaluation metrics in their work to provide a comprehensive evaluation of the TSAD algorithms.

In this paper, we aim to fill the gap in the literature by providing a comprehensive review of the evaluation metrics used and proposed in the field of time series anomaly detection. To the best of our knowledge, no prior works have offered a thorough overview of all the metrics used in the field. The main contributions of this paper are:

- A comprehensive description of the existing evaluation metrics, highlighting their key properties, both desirable and undesirable.
- A novel and structured taxonomy of the metrics, based on their calculation methods, to facilitate understanding and comparison. To the best of our knowledge, this is the first time a systematic taxonomy for TSAD evaluation metrics is defined.
- An in-depth analysis of the impact of the choice of evaluation metric through a set of hypothetical case studies.
- A clear summary of each metric in terms of a set of defined properties.

In Sect. 2 we define and introduce terms and concepts central to the topic of evaluating TSAD algorithms. We state the scope and limitations of this work in Sect. 3. In Sect. 4 we define 10 different properties distinguishing the metrics, all of which are presented and described briefly in Sect. 5. In Sect. 5 we also present the taxonomy of these metrics. Section 6 presents a series of case studies for testing the properties of the metrics, resulting in a categorization of the metrics in Sect. 7, based on the properties from Sect. 4. Finally, we summarize our findings and draw some conclusions in Sect. 8.

## 2 Background

In this section, we provide an overview of the fundamental concepts necessary to understand the subsequent discussion in this work.

**Time series.** A time series is a sequence of numbers or vectors, indexed by the time. We will refer to each time step as a point. Although not apparent in the definition, the underlying assumption when working with time series, is that the value of the points is dependent on the time variable.

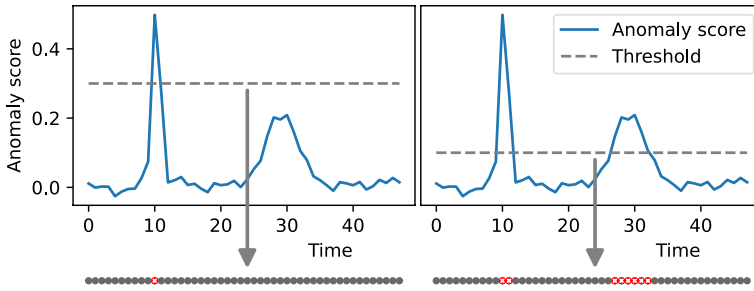
**Time series anomaly.** An anomaly in a time series is defined in various ways (Schmidl et al. 2022), but is in general a point or a subsequence of contiguous points with unexpected or abnormal values. We refer to the subsequence as an anomalous event, and each point in it as an anomalous point - not to be confused with a point anomaly, a term often used for events of length 1. Contrasting anomaly detection in independent data, the abnormality may stem from unsatisfied expectations of the time dependency. That is, a point can have a normal value for the time series in general, but anomalous in the context of its preceding values.<sup>1</sup> Furthermore, what is considered as anomalies depends on the domain and origin of the time series. Finally, it is often unclear just how anomalous an event should be to be considered an anomaly. This lack of an exact definition of time series anomalies is some of the reason it is difficult to come up with reliable evaluation metrics.

**Time series anomaly detection (TSAD).** The goal of TSAD is to identify anomalies in a time series. While a variety of techniques exist for detecting anomalies in time series data, a detailed review of which can be found in Schmidl et al. (2022), ranging from simple to complex and encompassing both machine learning and other approaches, it is not in the scope of this paper to discuss these techniques. Rather, our aim is to provide a comprehensive overview of the metrics used to evaluate these methods and offer a taxonomy of metrics based on their properties. In TSAD, the input data is typically a time series of data points, and the output is a prediction indicating which instances are anomalous. In our work we will refer to the output of the detection algorithm as *prediction*.

**Performance evaluation and analysis.** Using metrics, i.e. quantifying the performance of a particular anomaly prediction on a time series, is useful for two connected but distinct purposes, which we in this paper refer to as performance evaluation and performance analysis. We define *performance evaluation* as the task of assigning a score to each prediction, such that a higher (or lower) score means that the prediction is better, with the purpose of ranking several algorithms and choosing the best one. To be able to easily and objectively sort anomaly detectors in terms of performance, the final score must be a single scalar. On the other hand, we define *performance analysis* as the more general task of using one or several metrics in order to gain insights about the performance. This main focus in this work is performance evaluation, and we will only occasionally mention aspects only relevant for analysis purposes.

---

<sup>1</sup> Several works operate with different classes of time series anomalies (Kovács et al. 2019; Goswami et al. 2022; Lai et al. 2021; Choi et al. 2021), some of which consider if an anomaly is outside the normal values for all points, or just its temporal context.



**Fig. 2** Given an anomaly score, two distinct thresholds yield two different binary predictions. For each threshold, every time step with anomaly score higher than the threshold is considered anomalous. Lowering the threshold increases the number of anomalous points predicted. **x** Anomalous point, **●** Normal point

**Labels.** Evaluation is done by comparing the prediction to a time series of binary labels, that represents the ground truth of which points are anomalous or not. Note that the use of binary labels is a source for several kinds of errors and inaccuracies - when an anomaly starts, ends, and what even should be considered anomalous is a question that rarely has a definite answer, except for synthetical data. Therefore, there are several different labelling strategies, that will lead to quite different labels on the same dataset - e.g. the Numenta labelling strategy discussed in Sect. 5.1.5. Furthermore, when labels are made manually by humans, they will often have inconsistencies.

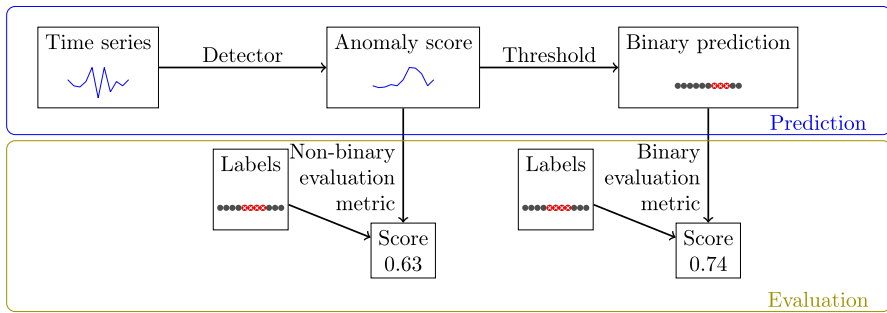
Changes in labels will necessarily affect the evaluation scores, especially if an event is included or excluded, as there are usually very few anomalies. The impact of slight changes in length and position of events, however, highly depend on the metric, and will be discussed and tested later in this article.

Due to high variability in both what is considered as anomalies, and how they are labelled, the relevance of results on data from across domains is not obvious. When selecting a detector for use on a specific TSAD task, one should evaluate detectors on a dataset with both similar time series, anomalies, and a labelling strategy in line with the desired output of the detection algorithm.<sup>2</sup>

### 2.1 Thresholding

An anomaly detector outputs an anomaly score, a time series with scalar values indicating how anomalous each time point is. In order to get a binary prediction, only time steps with anomaly score higher than some threshold are considered anomalous. This is visualized in Fig. 2.

<sup>2</sup> An alternative approach is unsupervised model selection, as described in (Goswami et al. 2022). They present three ways to select the best model based on datasets without labels - by considering prediction/reconstruction error, model centrality and performance on synthetically injected anomalies. The two former methods skip the need for the kind of evaluation metrics presented in this paper altogether.



**Fig. 3** Time series anomaly detection pipeline. The binary predictions are found by applying a threshold to the anomaly score outputted by the detector. The evaluation is done at one of two points, either using the anomaly score, or the binary predictions. In both cases, binary labels are used for comparison

There are several ways of choosing a threshold, some fully automatic, like the non-parametric dynamic thresholding introduced in Hundman et al. (2018), others as simple as just choosing

$$\text{mean} + n \cdot \text{std}$$

for some  $n$  (Geiger et al. 2020; Liu et al. 2022).<sup>3</sup>

Anomaly detections can be evaluated either before or after the thresholding, as shown in Fig. 3. We define *binary evaluation metrics* as metrics evaluating the binary prediction, and *non-binary evaluation metrics* as those evaluating the anomaly score. While the latter class uses the anomaly score as input, thresholding is still done, but as part of the metric. This usually involves calculating a score at several or all thresholds,<sup>4</sup> and either choosing the optimal score or combining the scores.

The difference between the classes may seem subtle but involves a foundational difference in what is evaluated. Binary metrics evaluate the combination of the detector and the thresholding strategy, while non-binary metrics aim at only evaluating the detector. The argument for the latter class is that thresholding is a separate issue, and since any detector can be used with any thresholding strategy, detectors should be compared independently of this choice. Using non-binary metrics ensure that thresholding is done equally for all detectors, which might be fairer. However, as thresholding is indeed a part of the non-binary metrics as well, this class of metrics is not independent of thresholds, but rather a compromise between them - and the metric might focus overly on irrelevant thresholds. Finally, as thresholding is done in practice, it might make more sense to evaluate the whole pipeline in unison, using a binary metric. This also allows for using the thresholding strategies that work well with specific detectors.

<sup>3</sup> As different methods have anomaly scores with different statistics, this may not be fair when comparing different methods. As an example, a method based on reconstruction error will have different outcomes depending on whether it uses MSE or RMSE error.

<sup>4</sup> By all thresholds we mean all thresholds that yield unique sets of anomaly points - at most one more than the number of time points in the time series.

		Prediction	
		⊗	●
Label	⊗	TP	FN
	●	FP	TN

**Fig. 4** The confusion matrix for anomaly detection. Each point can have one of two label values, and one of two prediction values, resulting in four different classes of points. ⊗ Anomalous point, ● Normal point

### 2.2 Traditional evaluation metrics

Before embarking on the time series specific metrics, it is beneficial to understand some of the evaluation metrics used for anomaly detection and classification in general. Common for most of the evaluation metrics is the use of the confusion matrix. The confusion matrix considers the possible combinations of binary prediction and labels, and includes the number of

- True positives (TP): points that are labelled and predicted as anomalies,
- False positives (FP): points that are labelled normal but predicted as anomalous,
- False negatives (FN): points that are labelled anomalies but predicted normal,
- True negatives (TN): points that are labelled and predicted as normal,

as seen in Fig. 4. We refer to these four numbers as counting metrics. They are not used for evaluation directly, but are needed for calculating the following metrics:

**Accuracy** is the fraction of correctly predicted points, i.e.  $\frac{TP+TN}{TP+TN+FN+FP}$ . Although simple, and to the uncritical eye informative, this metric should not be used for classifications with imbalanced classes, which anomaly detection is by definition. Since most points are normal, a prediction of only normal points will get a high accuracy despite not being useful at all.

**Recall**, also known as sensitivity and true positive rate, is the fraction of true anomalies that are correctly classified, i.e.  $\frac{TP}{TP+FN}$ . False positives are not penalized, thus predicting all points as anomalous will get a perfect recall of 1. For this reason, recall is usually not used on its own.

**Precision** is the fraction of anomalous predictions that are actual anomalies, i.e.  $\frac{TP}{TP+FP}$ . Like recall, this is not used on its own, since false negatives are not penalized, and only marking the most obvious anomaly will be the best strategy.

**$f_1$ -score** is the harmonic mean of precision and recall,  $\frac{2PR}{P+R}$ . The prioritization of precision and recall is a trade-off - strict threshold yield few predicted anomalies, thus high precision but low recall, and vice versa. Depending on the situation, a false positive might be highly preferred to a false negative, or vice versa. Thus, a more general definition is  $f_\beta$ -score, defined by  $\frac{(1+\beta^2)PR}{R+\beta^2P}$ . The value of  $\beta$  is the chosen so that the score reflects the relative importance of precision and recall. We will use  $\beta = 1$  in the examples of this paper, as is also common in the literature when comparing methods, but we highlight that an informed choice should be made for this parameter when using this metric for real world problems.

**False positive rate** is the fraction of normally predicted points that are actually anomalies,  $\frac{FP}{FP+TN}$ . Contrary to recall, optimal score is obtained by predicting all the points as normal. This is used for calculating the  $AUC_{ROC}$  score described in Sect. 5.2.3. Note that this metric has an optimal score of 0, and the worst possible score is 1, opposite of the other metrics in this section.

**Precision@k** is the precision of the  $k$  points with highest anomaly score. Although this is just the precision with a specific thresholding strategy, it deserves some extra attention. This is because, since the denominator  $TP + FP = k$  is predetermined, false positives are indeed penalized. Thus, this becomes a valid metric in itself, not needing to be combined with recall. In fact, recall@k is the *same value* as precision@k, except for a predetermined factor  $\frac{k}{TP+FN}$ .<sup>5</sup> Compared to the above metrics, this strategy requires the number of anomalies  $k$  instead of a threshold. This may be a simpler and more intuitive choice - a common practice is to use the number/fraction of anomalies in the dataset. It may also be fairer when comparing methods with differently distributed anomaly score, than many other threshold selection strategies.

The metrics above are often used for time series without adaptation, by regarding every time stamp individually. A large number of the evaluation metrics designed specifically for time series are versions of precision and recall that are redefined to handle events in a different way, either by a redefined confusion matrix, or by redefining precision and recall to not use the counting metrics at all. These are then usually used either to calculate f-score, or an AUC score, which we will discuss in Sect. 5.2.3.

### 3 Method

Several choices were made for the purpose of limiting the scope of this paper and keeping it concise. We did not include metrics from similar domains like time series classification, anomaly detection for non-time series, or change point detection. The latter, although similar to TSAD, only contains point anomalies.

Furthermore, we only consider single scalar metrics aimed at performance evaluation for detector selection, and not supplementary statistics for performance analysis. This means we will not consider the numerous variants of precision and recall as their own metrics, only as part of the  $f_1$ -score or the  $AUC_{PR}$  score described in section 5.2.3. Precision and recall are occasionally used for detector selection in situations where false positives and negatives have very different costs. However, due to the simple optimal strategies described in Sect. 2.2,  $f_\beta$  with a large/small  $\beta$  is a much better alternative. Other interesting statistics excluded by this choice are *early detection* (Buda et al. 2017), *before/after true positives* (Nalepa et al. 2022) and *alert delay* (Xu et al. 2018). Combinations of these statistics with other statistics could result in evaluation metrics with valuable properties. ROC- and PR-curves (see

<sup>5</sup> Note that this is not true for all the redefined versions of precision and recall presented later in this paper.



section 5.2.3) are often used for visualising properties of the anomaly score. We will only consider these for the purpose of calculating the much used single scalar AUC metrics.

There are several ways to vary each metric, by using techniques from one metric on one of the others. Indeed, some of the metrics are modified versions of other metrics, in such a way that all the other metrics could be modified in that same way. Studying all these combinations is not feasible without expanding the work substantially, so we will only study such modifications in their originally proposed, or most used, form. This should give an idea of the effect of the modification. Readers that are interested in a specific metric, either one included here, or that could be made by combining ideas from the ones included, are encouraged to conduct their own experiments.

Finally, for obvious reasons, we only consider metrics that *either* are rigorously defined in their original paper, *or* have open source implementations available.

## 4 Properties

In order to systematically evaluate the various metrics used in TSAD, we have defined several properties that differentiate the metrics. It is important to note that these properties are not inherently positive or negative, but rather the desirability of each property depends on the specific context and scenario. We have organized the properties into two categories:

**Valuation properties:** Properties regarding what qualities in the predictions are valued by the metric, and

**Intrinsic properties:** Other interesting properties apparent directly from the definition of the metric.

### 4.1 Valuation Properties

As time series anomaly detection methods rarely produce perfect prediction, a good metric needs to be able to prefer the best imperfect prediction available, for the situation for which the detector will be used. We listed five properties regarding what kind of prediction are preferred by the metrics.

**Value early detection.** In many situations, both in the literature and in practical scenarios, detection of a possible anomaly should occur as soon as possible (Lavin et al. 2015a), such as when anomaly detection is used in real-time systems where an anomaly indicates there is an issue requiring immediate attention. In these cases, detecting the anomaly at a late stage is of no value since it is too late to rectify the problem. To choose algorithms that detect anomalies in an early stage, such detections should be valued more by the metric than later detections. In other situations, data is analysed offline, or on a much larger time scale, where detection and reaction time is far greater than anomaly length, e.g. for diagnosis based on ECG monitoring (Chuah and Fu 2007; Sivaraks and Ratanamahatana 2015). In these cases, the

differences between early and late detection are of no practical relevance, and a metric without this property is to be preferred.

**Prioritize long anomalies.** Longer anomalies could indicate more serious problems which are also more important to detect, justifying why long anomalies contribute more to the final score than shorter ones, in many metrics. However, long anomalies might just indicate more subtle anomalies which are harder to locate (Kim et al. 2022a). The shortest anomalies might also be the most important ones, e.g. if they indicate serious problems that were fixed quickly, while the less serious ones were ignored and therefore lasted much longer. In most metrics, the contribution of an anomaly to the final score is either proportional to its length, or independent of its length. As many commonly used TSAD datasets have both long anomalies and single point anomalies, this difference has a great impact.

**Favour short predicted events.** Some detectors produce anomaly scores with a short peaks, while other methods, e.g. window-based ones,<sup>6</sup> produce wider areas of high anomaly score. The latter will generally result in longer predicted events. This might not have a big impact on the value of the prediction, but some metrics have a strong preference for short predicted events, independent of the length of the labelled anomaly.

**Prioritize partial detection.** While many metrics focus on predicting each time point correctly, and thereby getting the location and length of the anomalies correct (referred to as "covering"), it is often sufficient, or at least more critical, to detect any subset of it (referred to as "partial detection"). According to Xu et al. (2018), an operator receiving an alert of an anomaly will investigate the data manually, and the manual inspection will be the determining factor going forward, rendering the exact location and duration of the detection less relevant. However, Hwang et al. (2019) note that the operator may not necessarily find the anomaly if it is subtle and of a much longer duration than the detection, which would make the location and duration of the detection significant.

**Temporal tolerance.** The start and end of an anomaly is often unclear (Kim et al. 2022a), and when manually labelled, the labels might not be very reliable (Wu et al. 2022). Furthermore, a predicted event being off by a few time steps might still be very useful. Indeed, window-based detection methods might report the anomaly at either end of the window (Wu et al. 2022). In offline anomaly detection, this should not overly affect the score. For these reasons, detecting an anomaly close to a labelled anomaly should be valued by the metric.

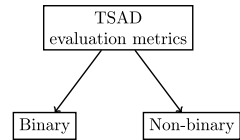
## 4.2 Intrinsic properties

Different metrics use different types of input, utilize different information from the input, and require different degrees of parameter specifications, affecting not only how to use them, but also what kind of situations they are suitable for.

---

<sup>6</sup> Window-based detection methods evaluate the abnormality of windows (contiguous subsequences of a predefined length) of the time series instead of each point separately, and then aggregate the results from all the windows into an anomaly score.

**Fig. 5** We consider two main categories of metrics, defined by their use of binary or non-binary predictions as input



**Binary.** As discussed in Sect. 2.1, metrics may be binary or non-binary, both types having different advantages and uses.

**Chronology aware.** Metrics not made for time series or sequential data do not use the chronology when calculating the score. Awareness of the labels and predictions of surrounding points is necessary for capturing the underlying time dependency specific for time series.

**Insensitivity to true negatives.** Given that anomalies by definition are rare events, a low score should be given when no anomalies are detected, even though the prediction is correct most of the time. Furthermore, it is useful not to be affected by how large the portion of true negative time points is, as this is a rather uninformative part of the data.

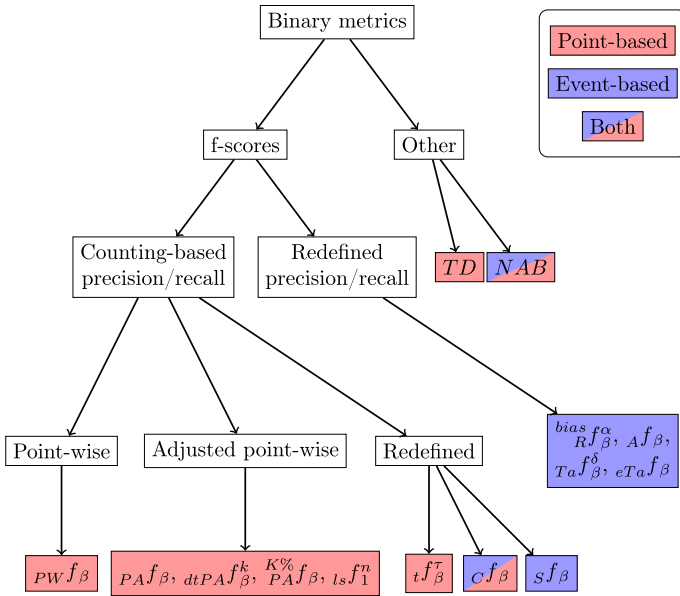
**Number of parameters.** Correctly specifying numerous of parameters to reflect specific needs can be resource demanding (Paparrizos et al. 2022a). Furthermore, it is easier to compare results across research papers when they do not use different parameters. Nevertheless, TSAD tasks vary greatly, and parameters offer flexibility needed for a metric to be useful for most specific cases.

### 4.3 Properties not included

We highlight that there are several desirable properties not included here due to our scope limitations. Such properties can be valuable insights about the performance of the method, e.g. where it performs well or not (Huet et al. 2022), or how early the detections are (Nalepa et al. 2022), or, for multivariate time series, which signals are the most involved in the anomaly. The latter property is often measured using distinct explainability measures (Su et al. 2019; Chen et al. 2021b; Garg et al. 2022; Li et al. 2021c). Furthermore, we have only included properties where more than one of the considered metrics stand out from the others. In cases where only one metric has an interesting property, this will instead be highlighted when presenting the metric in Sect. 5. Finally, we note that other interesting properties could surely be defined within our scope that separate more than one metric. Indeed, we have only considered ones that we have found in the literature, or found to be useful or interesting when analysing the metrics.

## 5 TSAD evaluation metrics: a taxonomy

In this section, a comprehensive examination of the evaluation metrics found through our research is presented. The metrics are divided into two categories, binary metrics in Sect. 5.1 and non-binary metrics in Sect. 5.2, as shown in Fig. 5.



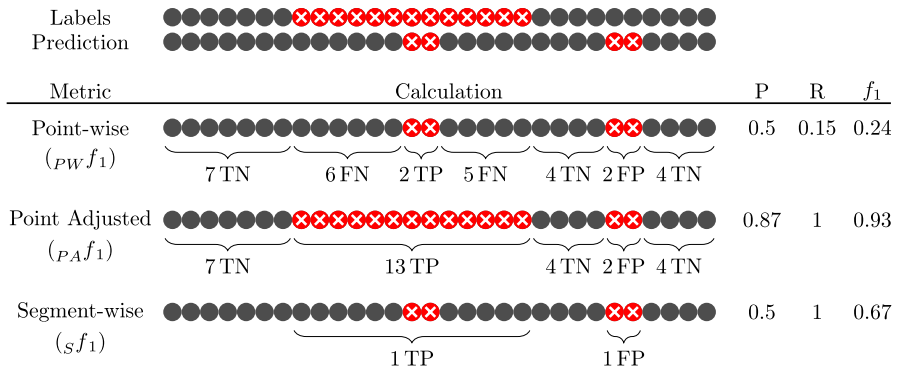
**Fig. 6** A taxonomy of binary evaluation metrics. A large number of these are f-scores based on various definitions of precision and recall. Precision and recall can be defined in many ways. Compared to the original point-wise definition, the difference can be present in the point-wise predictions, the counting metrics (TP, FP, TN, FN) or the formulas for precision and recall. The metrics can also be divided into point-based and event-based metrics, that count respectively individual points or contiguous events when aggregating to the total score.  $cf_\beta$  and  $NAB$  use both of these methods for parts of the total score

For each category, a taxonomy based on their definitions is introduced, followed by a description of each metric including their capabilities and potential limitations in utilization.

A rigorous definition of each metric is not included in this study, as some of them are quite complex, with details not necessary for this work. Readers are referred to the cited literature for further information. Instead, an effort has been made to provide a concise and intuitive understanding of the metrics. In addition, the most noteworthy, distinctive, or potentially problematic characteristics of the metrics are also discussed. This should suffice for understanding the taxonomy, as well as the properties studied in Sect. 6.

### 5.1 Binary evaluation metrics

We define *binary evaluation metrics* as metrics evaluating binary predictions, where each data point is classified as either normal or anomalous, aligning with the binary labelling. Figure 6 shows the proposed taxonomy of binary evaluation metrics, based on how their definitions use counting metrics (TP, TN, FP, FN), precision, recall or f-score. This information is relevant when combining techniques from different metrics, as such techniques may only work on one type of metrics.



**Fig. 7** Counting metrics are found in various ways by the different metrics. The figure shows the counting metrics, precision (P) and recall (R) and finally the  $f_1$ -score, for three of the various definitions of the  $f_1$ -score, for the same binary predictions.  $PWF_1$  considers each time point individually. As does  $PAF_1$ , but only after making an adjustment by expanding partially predicted anomalies.  $SF_1$ , on the other hand, considers events.  $\otimes$  Anomalous point,  $\bullet$  Normal point

Most of the metrics are based on the f-score, with some modification of the definitions. The metric based on *point-wise* counting metrics (Sect. 5.1.1) is the f-score based on counting metrics calculated in each time point. Metrics based on *adjusted point-wise* counting (Sect. 5.1.2) also use counting metrics in each point, but an adjustment is done to the prediction *before* the counting, in order to be more suited for anomalous events. For the metrics based on *redefined* counting metric (Sect. 5.1.3) the counting itself is done in some other way. *Redefined precision/recall* (Sect. 5.1.4) are not based on counting metrics at all, but calculated from some different formulas. They still use the terms precision and recall because the base concepts are the same. Finally, the *other* metrics (Sect. 5.1.5) are not based on f-score at all.

The metrics are also categorized based on their calculation approach, as either *point-based* or *event-based*. All the metrics are computed by aggregating the contributing parts of the time series, but in different ways. The point-based metrics evaluate each time point individually, whereas the event-based metrics evaluate entire events as a single subscore, regardless of the number of time points it comprises. This distinction has significant implications for what is considered a good prediction, as will be demonstrated in Sect. 6. Some metrics calculate part of the score in a point-based way and part event-based, giving parts of the properties from both classes.

### 5.1.1 Point-wise

**Point-wise f-score** ( $PWF_1$ ). One of the most straightforward evaluation metrics involves treating each time point as a single observation and calculating the f-score as outlined in Sect. 2.2. This approach is exemplified in Fig. 7. Although not made for time series, point-wise f-score is widely used in TSAD (Ahmed et al. 2022; Han and Woo 2022; Huang et al. 2022; Feng et al. 2022; Wang et al. 2022; Campos et al.

2021; Deng et al. 2021; Bashar and Nayak 2020; Niu et al. 2020; Chen et al. 2020; Mamandipoor et al. 2020; Hsieh et al. 2019; Li et al. 2019; Zhang et al. 2018). It is a simple metric, making it easy to implement and the results simple to understand. Also, methods are rewarded for predicting all the points that are labelled as anomalies, and none of the other - exactly what an anomaly detector should do - as opposed to some of the metrics we will describe below. Nevertheless, as we will see in the experiments of Sect. 6, the uneven event weighting and lack of tolerance can be highly problematic.

### 5.1.2 Adjusted point-wise

**Point Adjusted f-score** ( $_{PA}f_{\beta}$ ). The point adjusted metric was first introduced by Xu et al. (2018). They propose that if a single point within a true anomalous segment is accurately detected, a human operator can examine the segment and identify the entire anomaly. As a result, the entire contiguous segment is marked as anomalous in the prediction prior to calculating point-wise precision, recall, and f-score, as shown in Fig. 7. This metric has been widely used in TSAD (Xu et al. 2022; Goswami et al. 2022; Challu et al. 2022; Huang et al. 2022; Tuli et al. 2022; Chen et al. 2022; Li et al. 2021c; Feng et al. 2021; Dai et al. 2021; Chen et al. 2021b; Du et al. 2021; Choi et al. 2021; Zhao et al. 2020; Audibert et al. 2020; Shen et al. 2020; Su et al. 2019).

Previous works (Audibert et al. 2020; Garg et al. 2022; Doshi et al. 2022; Kim et al. 2022b) have shown that this metric can provide overly optimistic scores even if multiple anomalies are missed. In fact, Doshi et al. (2022); Kim et al. (2022b) demonstrate that random guessing outperforms state-of-the-art methods using this metric. The cause of this is a seemingly unintended flaw of the metric, which is illustrated in Fig. 7. Despite the argument that the whole anomaly is detected if an operator receives an alert within the anomaly, which legitimizes a recall of 1, only half of the alerts are correct, so the precision of the prediction should be 0.5. Instead, after adjustment, it is close to perfect. The greater the discrepancy between the duration of labelled and predicted anomalies, the more severe the problem becomes.<sup>7</sup> Calculating precision prior to adjustment would avoid this issue and produce a precision-recall pair that aligns with the reason for the adjustment as well as the meaning of precision and recall. Nevertheless, we instead suggest using the composite f-score (Sect. 5.1.3), a more appropriate metric in cases when a warning during an anomaly is sufficient.

**Delay thresholded Point Adjusted f-score** ( $_{dtPA}f_{\beta}^k$ ). Ren et al. (2019) and (Chen et al. 2021) use an adaptation of the point-adjusted metrics, where a labelled anomaly is only considered detected if an anomaly is predicted within the first  $k$  time steps of the anomaly. If not, all the points in the anomaly are marked as false negatives, even the ones predicted as anomalous. With this metric, precision can still be unreasonably high, but it is much more difficult to achieve this, and the random

<sup>7</sup> Interestingly, the paper of Xu et al. (2018) first using this metric have very short anomalies, compared to some of the datasets used in the papers that adopted this metric.

guessing strategy that prevail for  $P_A f_\beta$  will have a much harder time getting high scores with this metric.

**Point adjusted metrics at  $K\%$  ( $K\%_{PA} f_\beta$ ).** Kim et al. (2022b) suggest altering the point adjusted metric by requiring a portion  $K\%$  of the anomaly to be detected in order to make the adjustment. As with  $diPA f_\beta^k$ , this effectively reduce the effectiveness of random guessing, and short detections in general. Furthermore, as argued by Hwang et al. (2019) and Hwang et al. (2022), an expert receiving a short alert within a much longer anomaly might not be able to see the anomaly, but by requiring a substantial part of the anomaly to be detected, the chance that an expert would actually notice it is much larger.

**Latency and sparsity-aware f-score ( $ls f_\beta^n$ ).** Abdulaal et al. (2021) note that the point adjustment metrics do not value early detections, and changes the algorithm to only adjust the values of an anomalous event after the first TP point. They also note that false positive points require more resources if they are spread out, than in some close proximity (so that it only requires attention once). The prediction is therefore down-sampled by a user-specified factor  $n$ .

This way of awarding earliness reflects situations where the negative effects of an anomaly, which is proportional to its length, is avoided after the point that it is detected.

### 5.1.3 Redefined

**Segment-wise f-score ( $s f_\beta$ ).** Hundman et al. (2018) introduce a segment-wise precision, recall and f-score, where each contiguous segment of anomalous points is considered one event. Here one true positive is recorded for each true anomalous segment with at least one predicted anomalous point, one false negative for each of the rest of the true anomalous segments, and one false positive for any predicted anomalous segment without any true anomalous points. Figure 7 shows an example of this. This metric is used by Geiger et al. (2020); Nalepa et al. (2022); Meng et al. (2020); Flaborea et al. (2022).

A serious problem with this metric is that extending the length of a predicted anomaly will never give worse score, and often better. Thus, it favours detectors with long contiguous events, all the way to the extreme case: Predicting every point in the time series as anomalous will give perfect precision and recall for any time series with at least one anomaly.

**Composite f-score ( $c f_\beta$ ).** Garg et al. (2022) suggest using a combination of point-wise and segment-wise metrics, and propose the composite f-score, defined as the harmonic mean of point-wise precision and segment-wise recall. The point-wise precision ensures that false positive points are discouraged, whereas extra true positive points in an already partially detected anomaly is only awarded through the increased precision.

**Time tolerant f-score ( $t f_\beta^\tau$ ).** Scharwächter and Müller (2020) defines (point-wise) precision and recall with time tolerance  $\tau$ , essentially by counting it as a true positive when a predicted anomaly point is closer than  $\tau$  to a labelled anomaly point. They then show that while the recall and precision of their example prediction

increase drastically with the tolerance, the score of a random prediction increases more, and the statistical significance decreases substantially. Hence reporting results with time tolerance may be less significant than without, despite the scores looking more impressive. It should be noted, however, that their data contain many short anomalies. A tolerance of a few time steps will have a much larger impact on the random prediction score in with such a dataset, than with fewer or larger anomalies. Although these evaluation metrics are not widely used, similar tolerance techniques are - either in the metric (as here), in the labelling of the data (as in *NAB*, explained in Sect. 5.1.5) or in detectors padding their predicted events before outputting them. Such significance tests can be useful when determining how much time tolerance to use.

### 5.1.4 Redefined Precision and Recall

**Range-based f-score** ( ${}_{R}f_{\beta}^{bias}$ ). Tatbul et al. (2018) argue that point-wise precision and recall fail to address many aspects present in time series for anomaly detection, and introduce range-based precision and recall, forming a range-based f-score. The recall is calculated for each *labelled* event, using a formula scoring how well the labelled event is detected. The score is then averaged across all the labelled events. Similarly, the local precision is calculated for each *predicted* event, by scoring how well a predicted event corresponds to the labels, then averaged across all predicted events. The formulas for each event use up to 4 contributing concepts: *Detecting* the anomaly range with at least one anomaly point, while also *covering* as large a portion of the anomaly range as possible. High *cardinality*, i.e. number of predicted events within one labelled anomaly, can be punished, and a function rewarding the *position* of a detected anomaly within a labelled one can be specified. This results in a rather complex and highly customizable metric, with a tunable weight and up to 6 tunable functions to enable aligning the score with the goal of the detection task. Thorough guidelines, defaults and examples are provided in (Tatbul et al. 2018). The metric have been used in Jacob et al. (2021) and Meng et al. (2020).

Although evaluation metrics that consider the relative positions of detection and label are mostly useful for rewarding early detection, this metric can also be set to reward e.g. detections at the middle or at the end of the labelled anomalies, which the authors argue can be useful in certain cases, e.g. as a way of preventing false positive alarms. We have not found the cardinality concept in any other TSAD evaluation metric, and thus we have not considered it a desirable property. This may be more relevant for change point detection (Gensler and Sick 2014).

**Time series aware f-score** ( ${}_{Ta}f_{\beta}^{\delta}$ ). Hwang et al. (2019) propose time-series aware precision and recall. These are similar to range-based precision and recall, but also require that a certain portion  $\theta$  of the labelled anomaly must be correctly predicted for it to be counted as a correct detection. The concepts of cardinality and position are not considered. The authors note that determining the end of a labelled anomaly can be challenging, and therefore include a region of length  $\delta$  following the labelled event, with a positive but decreasing score, to account for this. This reduces the reliance on correct labelling and prediction at the end of and shortly after the event. A



slightly altered version of this metric can be found in Kim et al. (2022a), where the method for determining the length of ambiguous sections is changed.

**Enhanced time series aware f-score** ( $e_{Tad}f_{\beta}$ ). Hwang et al. (2022) highlight that previous evaluation metrics may reward detections that overlap with actual anomalies, even if they are either too long or too short to be useful. To address this issue, they propose an event-based metric similar to  ${}_{R}^{bias}f_{\beta}^{\alpha}$  that considers both a detection score and an overlap score. The metric requires that a certain portion of the actual anomaly to be detected and a certain portion of the detected anomaly to be true. Two parameters can be adjusted to control these portions. The precision calculation includes a weighting function that weights each event by the square root, as a compromise between typical point-based and event-based weighting.

**Affiliation-based f-score** ( $Af_{\beta}$ ). Huet et al. (2022) tackle problems commonly seen in existing metrics and introduces a distance-based metric as a solution. They calculate the average of the local precision and recall for each anomaly event. Local precision is calculated by averaging the distance between each predicted anomaly point and its closest labelled anomaly point, and expressing it as the probability of outperforming a random prediction. Local recall is calculated similarly, using the average distance from each labelled anomaly point to its closest predicted anomaly. By using distance, this metric evaluates the proximity of predicted and labelled anomalies, even if they don't overlap. It also values detection over coverage in a natural way. Finally, by scoring locally, the results are more interpretable, since each anomaly and its impact on the score can be evaluated separately.




### 5.1.5 Other

**NAB score** (*NAB*). The Numenta Anomaly Benchmark (*NAB*), presented by Lavin et al. (2015a), includes a dataset for time series anomaly detection and a novel evaluation metric. The metric penalizes false positive points with a negative value, and rewards true anomalous events with a positive value based on how early the first anomalous point was predicted. The score is normalized by comparing it to a scenario where no anomalies are detected.

Since only one point of the true positive points in an anomalous segment contribute to the score, while every false positive point contributes negatively, the score favours detectors predicting short events - it is almost never beneficial to predict two contiguous points as anomalous.

*NAB* also introduced a different approach to labelling anomalies. This approach allows for rewarding detectors predicting anomalies before they occur,<sup>8</sup> and makes the score less dependent on the individuals who label the anomalies. A simplified explanation of the approach is provided here, see Lavin and Ahmad (2015b) for the full details. The process involves a group of labellers deciding the first anomalous point for each anomalous event. Then, the points on both sides are marked anomalous, such that the original starting point is in the centre of the event, each event has the same duration, and 10% of the dataset is labelled as anomalous.

<sup>8</sup> That is, before they are visible to the human labeller.

Labels:		$Af_{\beta}$	$TD$
Prediction 1:		<b>0.91</b>	<b>14</b>
Prediction 2:		0.9	<b>2</b>

**Fig. 8** We test the affiliation and temporal distance metrics on two predictions of the same label time series. The best score for each metric is shown in bold. The labels include two events, and each prediction is a bit early on one of them. The affiliation metric splits the time series into periods with one event each, and calculates the relative distance of the closest predicted event. In this example, the first anomaly in prediction 1 is seen as closer to a true anomaly than the second anomaly in prediction 2.  $TD$ , on the other hand, uses absolute distance, and prefers prediction 2

This strategy is similar to the time tolerance technique in  $f_{\beta}^{\tau}$ . However, in this case it is part of the labelling strategy, instead of the metric. Thus is it not a part of the implementation used in this paper, and we will not see the effects of this in the experiments in Sect. 6.

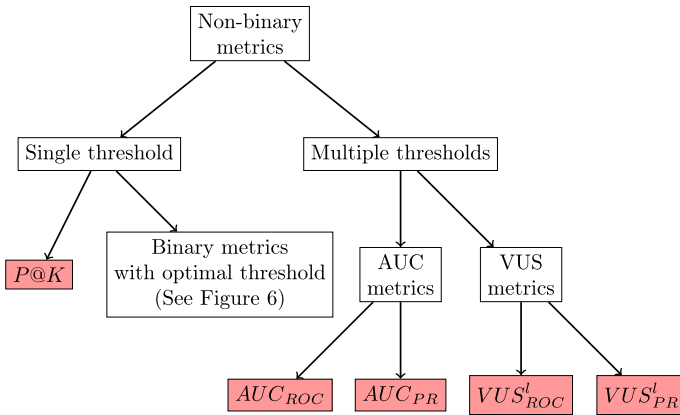
The  $NAB$  score is not widely used,<sup>9</sup> but their datasets are commonly used for benchmarking, using other metrics (Schmidl et al. 2022; Paparrizos et al. 2022b). The labelling strategy of this dataset highlight the importance of not blindly combining arbitrary metrics and datasets. Due to the labelling strategy, at least 50% of the points labelled anomalous were considered normal by the labellers, invalidating metrics counting each point individually, like  $p_w f_{\beta}$ .

**Temporal distance ( $TD$ ).** Temporal distance, presented by Kovács et al. (2019), is a very simple metric - summing the distances from each labelled anomaly point to the closest predicted anomaly point, and from each predicted anomaly point to the closest labelled anomaly point. The lower score the better. This metric prioritizes roughly finding all the correct anomalies over getting the detection exact, since any false positive/negative raises the score by the distance to the closest anomaly. As long FPs and FN are punished roughly proportionally to their length, the metric prioritizes long labelled anomalies, and a method predicting short events has an advantage when predicting FPs. Kovács et al. (2019) present two version of this metric<sup>10</sup>, by summing either absolute or squared distances. Generalizing this, one could use any positive power of the absolute distance. We will consider this exponent a parameter, and use 1 in all the experiments. High values of this parameter punish great distances more than low values.

Temporal distance might seem very similar to the affiliation f-score. However, there are some important differences. Since  $Af_{\beta}$  is calculated locally for every event, it is an event-based score, while  $TD$  is point-based, the effects of which will be clear from the experiments in Sect. 6. It may also lead to some odd situations when two or more anomalies are relatively close, as seen in Fig. 8. While  $TD$  considers the absolute distances, and therefore considers the first event in prediction 1 to be further from the labels than the second event in prediction 2,  $Af_{\beta}$  considers relative distances

<sup>9</sup> Despite very many metrics papers referring and comparing to this metric, we only found one paper using it for evaluation, by the same authors (Ahmad et al. 2017).

<sup>10</sup> They also present several other metrics, although they do not pass the limitations presented in Sect. 3



**Fig. 9** A taxonomy of non-binary evaluation metrics. Although the input is different from the binary metrics, they are quite similar, and indeed any binary metric can be made non-binary by using the optimal threshold strategy (See section 5.2.2)

within the local surroundings of each event, and therefore considers the distance in the last anomaly in prediction 2 as bigger than the first anomaly in prediction 1.

### 5.2 Non-binary evaluation metrics

The *non-binary evaluation metrics* are those evaluation the anomaly score, as opposed to a binary prediction obtained by using a threshold on the anomaly score. For these metrics, the thresholding step is part of the evaluation.

A taxonomy of non-binary evaluation metrics is proposed in Fig. 9. The primary difference between these metrics lies in the way they handle the threshold. Some metrics, such as  $P@K$  and *binary metrics with optimal threshold*, choose a single threshold, resulting in a single binary prediction. These metrics are still considered non-binary as the threshold selection is part of the metric. The other non-binary metrics evaluate the performance for the full range of possible thresholds,<sup>11</sup> and combine it into a single number score. This is done either by calculating the area under a curve (*AUC metrics*) or the volume under a surface (*VUS metrics*). The choice of non-binary metric will depend on the specific requirements and goals of the evaluation, and the suitability of each metric for the task at hand.

#### 5.2.1 Precision at K ( $P@K$ )

As described in Sect. 2.2,  $P@K$  is the fraction of the  $K$  points with the highest anomaly scores that are labelled anomalous. The point-wise  $P@K$  is occasionally used for TSAD evaluation (Paparrizos et al. 2022a, b). Other definitions of precision than point-wise could in principle be used, e.g. Deng et al. (2022); Zhang et al.

<sup>11</sup> This is done by evaluating for each threshold that yield unique binary predictions, or a representative selection of them.

(2019) use an event-based variant of recall at  $K$  for spatiotemporal anomaly detection, although for precision it would require defining how the number  $K$  of anomalies included in the prediction is counted.

A variant of  $P@1$  is the UCR score used by Rewicki et al. (2022), defined by Wu et al. (2021). The duration of the labelled anomaly is increased in both ends to include some time tolerance, before  $P@1$  is calculated.

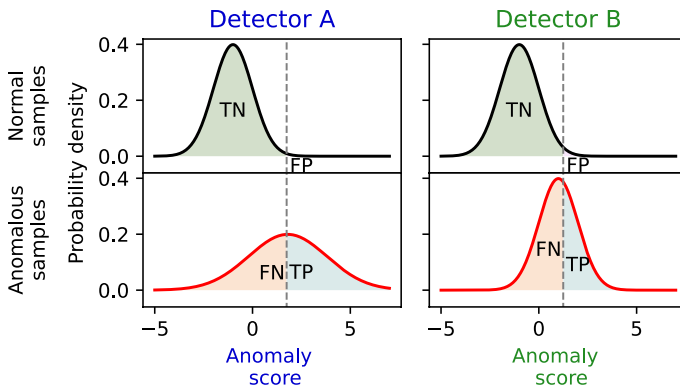
## 5.2.2 Binary metrics with optimal threshold

Binary metrics are typically used with the threshold that yields the best score (Liu et al. 2022; Huang et al. 2022; Campos et al. 2021; Deng et al. 2021; He and Zhao 2019; Lavin et al. 2015a). This can be achieved with any binary evaluation metric. The use of a metric combined with this thresholding strategy requires the input of an anomaly score, resulting in non-binary evaluation. The optimal threshold is determined by using labels, and can only be determined during the evaluation phase, thus providing an upper limit to the score that can be achieved using the binary metric. The relevance of this upper limit depends on the situation and the chosen binary metric.<sup>12</sup> For the sake of brevity, we will only consider the point-wise f-score with the optimal threshold strategy ( $f_{PW,\beta}^{best}$ ) in the remainder of this work.

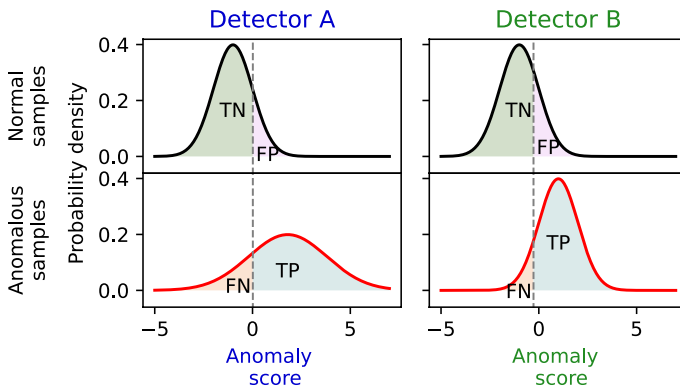
## 5.2.3 Area under the curve ( $AUC_{ROC}$ , $AUC_{PR}$ )

The receiver operator characteristic (ROC) is an evaluation metric commonly used for TSAD, as well as in binary classification in general. For each choice of threshold, the prediction has a specific value of recall and false positive rate. Plotting these against each other result in the ROC-curve. This is often inspected directly, as it visualizes the trade-off between recall and false positives, e.g. how large false positive rate must be allowed for certain levels of recall. In order to get a single scalar evaluation metric from this curve, it is common to integrate the area under the curve (AUC), to get the  $AUC_{ROC}$ . This value summarizes the detection performance across all thresholds, and is widely used in TSAD (Feng et al. 2022; Dai et al. 2022; Schmidl et al. 2022; Campos et al. 2021; Bhatia et al. 2021; Li et al. 2021b; Huang et al. 2020; Goodge et al. 2020; Braei and Wagner 2020; Zhang et al. 2020; Ergen and Kozat 2020; Wang et al. 2019; Kieu et al. 2019; Zhou et al. 2019; Pang et al. 2019; Park et al. 2017). An alternative method to comparing recall and false positive rate is to apply an area under curve approach to precision and recall, resulting in the calculation of the area under the precision-recall curve ( $AUC_{PR}$ ), also known as average precision. This approach too is commonly utilized in TSAD (Li et al. 2022; Campos et al. 2021; Li et al. 2021; He et al. 2020; Huang et al. 2020; Chen et al. 2020; Kieu et al. 2019; Zhou et al. 2019; Pang et al. 2019). In our experiments, we only consider the point-wise precision and recall for the PR curve, as is by far most used, although any other pairs can be used. For instance,  $AUC_{PR}$  with point-adjusted precision and recall is used by Dai et al. (2021), and  $AUC_{PR}$  with the range based precision and recall is

<sup>12</sup> E.g. optimal threshold  $s_{f_{\beta}}$  score is always 1, independent of the anomaly score.



(a) Thresholded optimally for  $PWF_1$

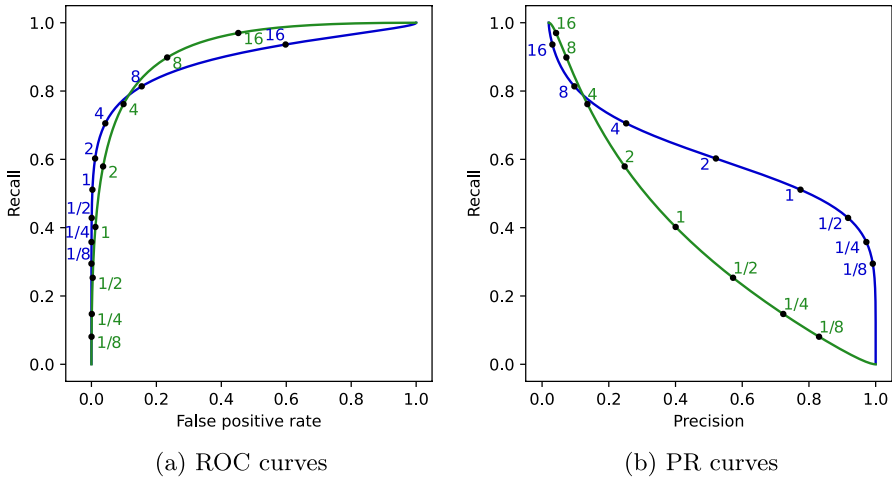


(b) Thresholded optimally for  $PWF_8$

**Fig. 10** Probability density functions for the anomaly scores of the positive (red) and negative (black) samples, for detectors A and B. The stippled lines are the thresholds. Anomalous samples generally give higher anomaly scores, although many normal and anomalous points have similar scores, making it hard to set a threshold. (a, b) show the counting metrics (as the shaded areas) for thresholds optimal for  $PWF_\beta$  for  $\beta = 1$  and  $\beta = 8$  respectively. Keep in mind that only 2% of samples are anomalies, i.e.  $TN + FP = 49(FN + TP)$ , which is not shown in the figures. FP and FN are of comparable sizes in 10a (Color figure online)

used by Schmidl et al. (2022). Variations of the ROC curves can be used as well, but the false positive rate is not defined for the event-based metrics.

The use of  $AUC_{ROC}$  has been criticized for its integration over all thresholds, which can result in a large portion of the score coming from thresholds that may not be relevant for a specific use case (Baker and Pinsky 2001; Lobo et al. 2008; Berrar and Flach 2012). A possible solution can be to only consider parts of the curve, as suggested by Baker and Pinsky (2001), although it can be hard to determine how much of it to use. Another possibility is to use  $AUC_{PR}$  instead. While  $AUC_{PR}$  also



**Fig. 11** The ROC and PR curves for detectors A (blue) and B (green). The marked dots are the points corresponding to the optimal thresholds for  $p_{WF\beta}$  with  $\beta \in 16, 8, 4, 2, 1, 1/2, 1/4, 1/8$ . We observe that detector A is best for higher thresholds, which are optimal for lower values of  $\beta$ , and vice versa. Detector B has the higher  $AUC_{ROC}$ , while detector A has the higher  $AUC_{PR}$  (Color figure online)

integrates over all thresholds, it has been argued that it is more informative than the ROC for imbalanced datasets (Davis et al. 2006; Saito and Rehmsmeier 2015), which by definition is the case for anomaly detection.<sup>13</sup> The reason is that precision and false positive rate respond differently to changes in false positives (FPs). In anomaly detection, the number of true negatives will typically be very large compared to FPs, making the false positive rate low for all relevant choices of threshold. As a result, only a small part of the ROC curve is relevant in such cases.

We visualize this with an example. Assume a very large dataset has 2% anomalies, and that two detectors, named A and B, produce anomaly scores from the normal distributions visualized in Fig. 10. That is, the detectors produce anomaly scores from the black distributions in Fig. 10 for normal points, and from the red one for anomalous points. Note that since  $AUC_{ROC}$  and  $AUC_{PR}$  are independent of the time dimension, time is not included in this example. This results in the ROC-curves in Fig. 11a and PR-curves in Fig. 11b.

From the roc curves in Fig. 11a we see that detector B (green) outperforms detector A (blue) for most values of the false positive rate. This would result in  $AUC_{ROC}$  preferring detector B. By inspecting the graph, we see that for smaller false positive rate, detector A is better. Inspecting the PR curves in Fig. 11b, we see that detector A by far would have the best  $AUC_{PR}$ , but for low precision detector B is better. While the figures really contain the same information (Davis et al. 2006), it is

<sup>13</sup> As pointed out by Wu et al. (2022), not all commonly used datasets for TSAD are particularly imbalanced. Finding the labels in these datasets cannot really be considered anomaly detection, but should rather be regarded as classification or segmentation.

clear that the difference in x-axis is crucial, not only for AUC-values but for inspection of the curves as well.

Figure 10a and b also show the thresholds yielding the optimal f-score at different values of  $\beta$ . The points on the curves of these values, and more, are shown in Fig. 11. We see that for  $\beta \approx 1$ , the recall value has very little impact on the  $AUC_{ROC}$ , compared to the  $AUC_{PR}$ . Indeed, the relevant values of  $\beta$  should be quite high for  $AUC_{ROC}$  to be more informative than  $AUC_{PR}$ . But from Fig. 10, the high  $\beta$  might seem more relevant, due to the large increase in TP, and high values of  $\beta$  make up a relatively small part of the pr curves in Fig. 11b. As always, what is most suitable comes down to the situation. Since the ROC curve uses the fraction of FP to all normal samples, instead of anomalous predictions, the difference between ROC and PR scales with the imbalance of the data - when the anomalies make up an even smaller fraction of the data,  $AUC_{ROC}$  corresponds to even higher values of  $\beta$ .

### 5.2.4 Volume under the surface ( $VUS_{ROC}^l, VUS_{PR}^l$ )

The concept of volume under the surface (VUS) was introduced by Paparrizos et al. (2022a), extending  $AUC_{ROC}$  and  $AUC_{PR}$ . The authors recognize the need for some tolerance for predicted anomalies close to actual anomalies. They address this issue by adjusting the labels, and instead of using binary labels of 0 or 1, they use labels with real values in the range [0, 1]. The original labelled anomalies are still given a value of 1, and normal points that are a certain distance  $l$  away from anomalies are given a value of 0. Labels closer to the original labelled anomalies gradually decrease as the distance from the anomaly increases.<sup>14</sup> The authors refine the point-wise recall by multiplying it with the existence factor used in  $\frac{bias\ f\ \alpha}{R\ \beta}$ . Using the new definitions of recall, precision, and false positive rate, they define range versions of  $AUC_{ROC}$  and  $AUC_{PR}$ . However, since this approach depends heavily on the tolerance threshold,  $l$ , they also introduce the volume under surface metric. Inspired by the way that the AUC metrics integrate away the dependency on the threshold by considering the area under a curve generated from all values of the threshold, the VUS metrics integrate over  $l$  to get the volume under the surface generated by the ROC or PR curve along an axis of values of  $l$ . This way, the final value takes into account multiple tolerance levels. Nevertheless, the metrics still depend on the maximum value for  $l$ .

## 6 Case studies

In this section, we evaluate the presented evaluation metrics on 14 different case studies, to illustrate the different properties of the metrics. It is important to note that the desirability of these properties is highly dependent on the specific domain and use case. Thus, there is no universal "correct" answer for which metrics are best, but for a specific use case there is often one that is most appropriate. By

<sup>14</sup> A similar smoothing strategy is done by Dai et al. (2022) to account for noisy labels, before applying  $AUC_{ROC}$ .

**Table 1** Overview of all the metrics considered in the case studies

	Short name	Long name	Point-based	Event-based	Worst/best values	Section	
Binary	$PWF_{\beta}$	Point-wise f-score	✓		0/1	5.1.1	
	$PA_{\beta}$	Point adjusted f-score	✓		0/1	5.1.2	
	$dtPA_{\beta}$	Delay thresholded point adjusted f-score	✓		0/1	5.1.2	
	$K\%f_{\beta}$	Point adjusted metrics at K%	✓		0/1	5.1.2	
	$PA_{\beta}^n$	Latency and sparsity-aware f-score	✓		0/1	5.1.2	
	$IS_{\beta}^n$	Segment-wise f-score		✓	0/1	5.1.3	
	$Sf_{\beta}$	Composite f-score	✓	✓	0/1	5.1.3	
	$Cf_{\beta}$	Time tolerant f-score	✓		0/1	5.1.3	
	$f_{\beta}^T$	Range based f-score		✓	0/1	5.1.4	
	$bias_{R_{\beta}}^{f_{\beta}}$	Time series aware f-score		✓	0/1	5.1.4	
	$Tid_{\beta}^{SS}$	Enhanced time series aware f-score		✓	0/1	5.1.4	
	$eTid_{\beta}$	Affiliation based f-score		✓	0/1	5.1.4	
	$Af_{\beta}$	NAB score		✓	$-\infty/100$	5.1.5	
	NAB	Temporal distance		✓	$\infty/0$	5.1.5	
	TD	Precision at K		✓	0/1	5.2.1	
	Non-binary	$P@K$	Point-wise f-score with optimal threshold	✓		0/1	5.2.2
		$bcs_{PWF_{\beta}}$	Area under the receiver operator characteristic curve	✓		0/1	5.2.3
$AUC_{ROC}$		Area under the precision-recall curve	✓		0/1	5.2.3	
$AUC_{PR}$		Volume under the receiver operator characteristic surface	✓		0/1	5.2.4	
$VUS_{ROC}^I$		Volume under the precision-recall surface	✓		0/1	5.2.4	



presenting examples and highlighting the properties of the metrics, we aim to provide a clearer understanding of how they can be used effectively in different situations.

To simplify reading the results, the name for each evaluation metric presented, is repeated in Table 1. The range of values is also shown. Note that the temporal distance metric is the only one where the score should be as low as possible, as opposed to as high as possible.

Here we outline the decisions made regarding the implementation of the evaluation metrics, and parameter selection. Most of the metrics have parameters that need to be specified. To maintain consistency in our experiments, we have chosen the same evaluation metric parameters for most of the case studies. However, in some cases, we adjust these parameters to highlight a specific effect.

The  $\beta$  in the  $f_\beta$  is 1 for all f-score based metrics. For  $dtPAf_\beta^k$  we use a delay threshold of  $k = 2$  time points. For  $PAf_\beta^{K\%}$  we require 20% of the anomaly detected for adjustment. The downsampling factor of  $lsf_\beta^n$  is set to 2, and the time tolerance of  $f_\beta^\tau$  to  $\tau = 2$  for most experiments, except for the on in Fig. 18, where we use  $\tau = 10$  to better visualize its effect.

For the range based f-score  $bias_{Rf_\beta}^\alpha$ , we use  $cardinality = 1$ , and specify  $\alpha$  and the positioning bias in the metric name in the table for each experiment. See Tatbul et al. (2018) for the definition of these parameters and functions. We use the same configuration for precision and recall.

For  $Taf_\beta^\delta$  we set  $\alpha = \theta = 0.5$  for all tests. We use  $\delta = 0$  in most cases since this is more in line with the tests. We use  $\delta = 10$  for the graph in Fig. 18 to show the effect of this delta. For  $eTaf_\beta$ , we use  $\theta_p = 0.5$ ,  $\theta_r = 0.1$ , to show the effect of using different values of these parameter. This will effectively ignore any predicted event with less than 0.5 precision, i.e. if less than half of the predicted event overlaps with anomalies. On the other hand, less than 10% of an anomalous event must be detected for not to be counted as detected. Using  $\theta_r = \theta_p = 0.5$  would yield results similar to that of  $Taf_1^{10}$  in most cases.

*NAB* is implemented using the *standard application profile* (Lavin and Ahmad 2015b). As *NAB* is implemented for use with longer anomalies, it does not run in the cases where there are events of length 1 in the labels. We do not include *NAB* in these cases.

*P@K* is the precision of the *K* highest anomaly scores. For *P@K* we set *K* to the number of anomaly points in the labels. Due to many equal anomaly scores in the test cases, a threshold including *K* points will often include  $L > K$  points. In these cases, we report *P@L* instead.

For  $VUS_{ROC}^l$  and  $VUS_{PR}^l$  we use a maximum tolerance of  $l = 4$ .

While we have implemented the simple metrics ourselves, the more complicated ones were taken from open source implementations by the authors of the metrics.  $AUC_{ROC}$  and  $AUC_{PR}$  are from sklearn (Pedregosa et al. 2011). Our implementation of the metrics, along with the code for generating the tables and figures in this paper, is available on Github <sup>15</sup>.

<sup>15</sup> [https://github.com/sondsorb/TSAD\\_eval](https://github.com/sondsorb/TSAD_eval)

## 6.1 Binary cases

To test the valuation properties of the different metrics, we have made a series of simple experiments with one time series of labels, and two imperfect prediction time series that resemble the labels in different ways. We then test which of the two predictions each of the metrics prefer. For each test we refer to a figure showing the time series and scores, with the optimal one for each metric shown in bold.

As the different metrics have different ranges, and different values of scores that could be considered good in each case, comparing numerical values of different metrics is difficult. To keep the discussion short and simple, we will only focus on which prediction is preferred, and not the actual values of the scores. Nevertheless, the scores are shown in the figures for the interested readers. It is worth noting that the discrimination between the predictions varies a lot from case to case and metric to metric. Whether this is good or not, again, entirely depends on the situation.

### 6.1.1 Partial detection vs covering

In anomaly detection, it may be sufficient to detect only a portion of the anomalous event. However, the correct duration of the event is still useful. Figure 12 illustrates the different ways in which these aspects are addressed by various metrics. The point-wise f-score considers each point equally, regardless of whether the event has already been partially detected. In contrast, some metrics give the highest score to methods that detect only one point, providing no incentive to detect the entire event.

### 6.1.2 Effect of anomaly length

Most point-based metrics value each time point equally, while most event-based metrics value each event equally. Other options are  $eTaf_\beta$ , which weights events by the square root of their length,  $cf_\beta$  which counts points and events for precision and recall respectively, and  $NAB$ , counting TP event-wise and FP point-wise. These differences may lead to some unwanted prioritizations. Figure 13 shows a situation with two short anomalies and one longer. For point-based metrics, it is better to predict the long one than both short ones. For datasets with high variance in anomaly length, or a combination of point anomalies and event anomalies, an event-based metric is often more appropriate. On the other hand, event-based evaluation metrics can be sensitive to sets of short anomalies close to each other, as seen in Fig. 14, where the event-based metrics prioritize the cluster of three events over the single long one.

### 6.1.3 Preference for short predicted anomalies

For  $pAf_\beta$  and  $NAB$ , there is no gain in having more than one TP point within an anomaly, while every FP is punished point-wise. This leads to a considerable preference for short predicted anomalies, as they can give high reward with a comparatively low risk. As seen in Fig. 15, if two detection methods find the same

	$PWf_1$	$PAf_1$	$dtPAf_1^2$	$20\%PAf_1^2$	$lsf_1^2$	$sf_1$	$cf_1$	$tf_1^2$	$f^{lat}f_1^{0.2}$	$Rf_1^0$	$Taf_1^0$	$eTaf_1$	$Af_1$	$NAB$	$TD$	
	<b>0.67</b>	0.67	0.67	<b>0.67</b>	0.71	0.67	0.67	<b>0.67</b>	<b>0.67</b>	<b>0.67</b>	0.67	0.67	0.67	0.67	50.0	92
	0.22	<b>1.0</b>	<b>1.0</b>	0.22	<b>1.0</b>	<b>1.0</b>	<b>1.0</b>	0.55	0.46	0.12	<b>0.72</b>	<b>0.72</b>	<b>0.77</b>	<b>100.0</b>	<b>56</b>	

Fig. 12 Partial detection vs covering: Detecting all the anomalies can be more valuable than covering one of them. Some metrics reflect this, but not all. Some do not value covering at all, and give optimal score to the bottom prediction

	$PWf_1$	$PAf_1$	$dtPAf_1^2$	$20\%PAf_1^2$	$lsf_1^2$	$sf_1$	$cf_1$	$tf_1^2$	$f^{lat}f_1^{0.2}$	$Rf_1^0$	$Taf_1^0$	$eTaf_1$	$Af_1$	$NAB$	$TD$	
	<b>0.71</b>	<b>0.71</b>	<b>0.71</b>	<b>0.71</b>	0.6	0.5	0.5	<b>0.71</b>	0.5	0.5	0.5	0.5	0.5	0.5	33.33	46
	0.62	0.62	0.62	0.62	<b>0.73</b>	<b>0.8</b>	<b>0.8</b>	0.62	<b>0.8</b>	<b>0.8</b>	<b>0.8</b>	<b>0.8</b>	<b>0.8</b>	<b>0.8</b>	<b>66.67</b>	<b>45</b>

Fig. 13 Effect of anomaly length: Point-based metrics value anomalies by their length, thus giving higher score to the top prediction than the bottom prediction, even if the latter one discover more of the anomaly events

	$PWf_1$	$PAf_1$	$dtPAf_1^2$	$20\%PAf_1^2$	$lsf_1^2$	$sf_1$	$cf_1$	$tf_1^2$	$f^{lat}f_1^{0.2}$	$Rf_1^0$	$Taf_1^0$	$eTaf_1$	$Af_1$	$NAB$	$TD$
	0.46	0.46	0.46	0.46	0.67	<b>0.86</b>	0.67	0.55	<b>0.71</b>	<b>0.77</b>	<b>0.77</b>	<b>0.77</b>	<b>0.77</b>	<b>0.77</b>	57
	<b>0.86</b>	<b>0.86</b>	<b>0.86</b>	<b>0.86</b>	<b>1.0</b>	0.67	<b>0.67</b>	<b>0.93</b>	0.67	0.67	0.67	0.67	0.67	0.67	<b>6</b>

Fig. 14 Effect of anomaly length: Depending on the labelling strategy, some datasets might have several non-contiguous anomalies in close proximity. The event-based metrics will value detecting the (whole) cluster of anomalies over detecting the later contiguous anomaly

anomalous events, but one of them produce longer predicted anomalies, the score may be very different. This may seem like the precision/recall tradeoff in disguise - these two predictions could come from the same anomaly scores, but using different thresholds. However, some methods indeed predict shorter anomaly events than other methods, independent of the threshold.

### 6.1.4 Score as a function of position of the predicted event

To visualize how the different metrics value predicted events at different positions relative to a labelled event, we made a scenario with a time series of length 100, with one anomalous event from step 40 to 60, and a prediction with one anomalous event of length 5, at variable positions. We calculate the score for each position of the predicted anomaly, and plot this in a graph, as visualized in Fig. 17. Figure 18 visualizes the score for each metric as a function of the position of the predicted event. We include  ${}_{R}^{bias}f_{\beta}^{\alpha}$  with two positioning bias functions to show the different effects they have on the score. As we see, the sensitivity to the position of the prediction varies considerably.  ${}_S f_1$  only has two values in the score, and  ${}_{PA} f_1$  and  ${}_{eTa} f_1$  have almost the same shape, with only slightly reduced score at the edges. Many of the other metrics have more gradually changing scores. As abnormality in reality seldom is a binary concept, gradually changing scores should be fairer in most cases.

**Value early detection.** We see that  ${}_{dtPA} f_1^2$ ,  ${}_{ls} f_1^2$ ,  $NAB$  and  ${}_{R}^{front} f_1^0$  all value earliness, but in different ways, and to varying degree.  $NAB$  only has a slight preference for early detection, while  ${}_{ls} f_1^2$  and  ${}_{R}^{front} f_1^0$  have about linearly decreasing scores.  ${}_{dtPA} f_1^2$  changes very abruptly, and only values very early detections.

**Temporal tolerance.** In cases where ground truth labels are not precise, methods should be rewarded more/punished less for a false positive close to a true anomaly than farther from them. Note that the value of earliness might interfere with this, so balancing these concepts can be difficult. We have not found any one metric considering both of these concepts.  ${}_A f_1$  and  $TD$  stand out as the only ones with tolerance across the whole time series. Along with  ${}_I f_1^{10}$ , these are the only ones valuing detecting anomalies *before* the labelled anomaly, while also  ${}_{Ta} f_1^{10}$  and (barely)  $NAB$  value detection *after* the labelled anomaly.

An effect of temporal tolerance is that the score is less dependent on the labeling strategy. We show this with an example. The labels of a dataset are usually not perfect, and often it is not clear what is an anomaly, and where an anomaly starts or ends. While the score of an anomaly detector always will depend heavily on what is considered a ground truth anomaly and not, the sensitivity to the exact length and location to an anomaly varies. Figure 19 shows a situation where it is not clear where to put the anomaly labels. One possibility is to mark all the high valued points as anomalous. Another strategy is to label only the points around the discontinuities, e.g. as done by Lai et al. (2021). Indeed, there may be nothing anomalous about the points in between these jumps. Yet, if the distance between the jumps is small enough, it makes more sense to view it as a single contiguous anomaly - as noted by Wu et al. (2022), a single normal point between two anomalies is an anomaly in its own right. Thus, at some time scale in between these situations, it should be

unclear how to label this event. Two possible labels corresponding to this time series are shown in Fig. 16, along with scores for predicting the labels from the opposite strategy. The metrics with temporal tolerance are also more tolerant to the labelling strategy, and give good scores in both cases, as opposed to the other metrics.

## 6.2 Non-binary cases

As non-binary metrics use the raw anomaly score as input, the space of possible inputs is much larger, making it more difficult to do extensive examinations of how these metrics react to a representative variation of realistic inputs. Nevertheless, we attempt to visualize some properties of these metrics as well. Before presenting these tests, we emphasize that the results of these metrics are dependent only on the relative anomaly score at each point, and not their actual value. This is shown in Fig. 20, where the anomaly scores are both symmetric, and decreasing in the distance from the middle. This gives the same scores for all the metrics, independent of the labels. For most experiments in this section, we have only a very few possible values of the anomaly scores, and the points that are not visually different, have the same score. The exceptions of this are specified in the captions.

### 6.2.1 Effect of anomaly length

Figure 21 shows that the non-binary metrics mostly favour detecting the long anomalies, as these have more points. However, the VUS metrics can favour detecting the short ones if there are more of them, as the anomaly events are effectively widened by the metric.

### 6.2.2 Preference for short predicted anomalies

Figure 22 shows predictions with short and wide anomalies, similar to the binary case shown in Fig. 15. We see that none of these metrics have the short predicted anomaly preference like  $pAf_{\beta}$  and  $NAB$ .

### 6.2.3 Partial detection versus covering

Similar to for the binary metrics, we test the value of detection compared to covering in Fig. 23. Since all the non-binary metrics considered are point-based, none of them value the detection of the second anomaly over covering the first one.  $P@K$ , however, value them equally in this case, since  $K$  is larger than the number of points with positive anomaly score.

### 6.2.4 Temporal tolerance

By smoothing out the labels, the VUS metrics value predicted anomalies close to the labelled anomalies, as seen in Fig. 24. The other non-binary metrics do not

	$PWf_1$	$PAf_1$	$dtPAf_1^{20\%}$	$PAf_1^{20\%}$	$lsf_1^2$	$sf_1$	$cf_1$	$tf_1^2$	$f_{R_1}^{lat\ f_1^{0.2}}$	$Taf_1^0$	$eTaf_1$	$Af_1$	$NAB$	$TD$
	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●
	0.22	<b>0.93</b>	0.22	<b>0.89</b>	<b>0.67</b>	<b>0.67</b>	0.53	0.39	0.12	0.53	0.68	<b>93.8</b>	<b>28</b>	
	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●
	<b>0.63</b>	0.7	<b>0.7</b>	<b>0.73</b>	<b>0.67</b>	<b>0.67</b>	<b>0.67</b>	<b>0.64</b>	<b>0.65</b>	<b>0.65</b>	<b>0.76</b>	<b>66.3</b>	58	

Fig. 15 Preference for short predicted anomalies: Most metrics value the bottom prediction at least as good as the top prediction, as it has the same precision but detects more of the anomaly.  $PWf_1$ ,  $TD$  and  $NAB$ , on the other hand, have strong preferences for short predicted events

	$PWf_1$	$PAf_1$	$dtPAf_1^{20\%}$	$PAf_1^{20\%}$	$lsf_1^2$	$sf_1$	$cf_1$	$tf_1^2$	$f_{R_1}^{lat\ f_1^{0.2}}$	$Taf_1^0$	$eTaf_1$	$Af_1$	$TD$
	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●
	0.25	0.92	0.92	0.25	0.86	0.67	0.67	0.91	0.4	0.14	0.54	0.91	10
	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●	●●●●●●●●●●●●●●●●
	0.25	0.25	0.25	0.25	0.4	0.67	0.25	0.91	0.4	0.14	0	0.83	10

Fig. 16 Temporal tolerance: Possible predictions for the time series in Fig. 19. For each prediction, we use the other prediction as the labels. This shows how much metrics are affected if the detector and labeller use different labelling strategies. A metric not sensitive to this would give good scores in both these situations. Point-based metrics like  $PWf_1$  and  $PWf_1$  are heavily affected, event-based metrics are clearly the most tolerant for differing labelling strategies

value high anomaly scores close to an anomaly. However, the anomaly scores in Fig. 24 only have two values, which is not realistic. Since anomaly scores often are somewhat smooth, high anomaly scores close to the anomaly can indicate that the anomaly score is also relatively high at the anomaly. This is shown in Fig. 25, where the anomaly scores are bell curves at different locations. When the centres of the bell curves are closer to the anomaly, the anomaly scores of these points are also higher, giving a better score at these points. This kind of temporal tolerance does however fully rely on the form of the anomaly score, which may not necessarily be fair.

### 6.2.5 Effect of class imbalance

Unlike all the other metrics we have considered,  $AUC_{ROC}$  and  $VUS_{ROC}^l$  include the number TNs in their formulas. This means including extra points that would not affect other metrics, will affect these. This is shown in Fig. 26. We see that the scores of  $AUC_{ROC}$  and  $VUS_{ROC}^l$  increase from less than 0.1 to more than 0.9 as the anomaly ratio decreases from 4/8 to 4/64. This means that for low anomaly ratios, precise detections are less important. An example of this changing the required precision is shown in Fig. 27. While  $AUC_{ROC}$  and  $VUS_{ROC}^l$  prefer the short predicted event in the short time series, they prefer the less precise one in the long time series. When working with multiple time series, this makes it challenging to have an idea of what should be considered a good score for each time series.

## 7 Categorization

In this section we present how each metric relates to the properties presented in Sect. 4. Table 2 shows the properties of each metric. We will in the following paragraphs explain how these results were determined. If the result of any test depends on a parameter<sup>16</sup>, we mark it by an asterisk. We also use an asterisk for partially obtained properties. What we mean by this is explained for the relevant properties below.

For *Value early detection*, we consider the results at the vertical marks on the axes shown in Fig. 18 for the binary metrics. If the score at the second mark is higher than the third mark, we consider the metric to value early detection. Since none of the non-binary metrics use the direction of the time series in their calculation, they can not have this property due to symmetry in the time dimension.

The *Prioritize long anomalies* property is based on the result in Figs. 13 and 21 for binary and non-binary metrics respectively - metrics not preferring the prediction in the bottom row are considered to have this property. Similarly, the *Favour short predicted events* property is based on Figs. 15 and 22. Metrics giving *better* score to the prediction in the top row have this property. Metrics that *Prioritize partial detection* are those not preferring the top prediction in Figs. 12 and 23.

<sup>16</sup> For  $\frac{bias}{R^j \beta}$ , where the function parameters could in principle be anything, we have only used the ones suggested in the original paper

The *Temporal tolerance* property for binary metrics is based on Fig. 18. Metrics that have non-zero score at the first *and* last vertical marks are considered to fully have this property, while metrics with non-zero score only the last mark partially have the property and are marked with an asterisk. For non-binary metrics, metrics distinguishing the anomaly scores Fig. 24 are considered to have the property, while metrics only distinguishing the smoother anomaly scores of Fig. 25 are marked with an asterisk. We do not consider this property to depend on parameters for  $f_{\beta}^{\tau}$ ,  $VUS_{ROC}^l$  and  $VUS_{PR}^l$ , since  $\tau = 0$  or  $l = 0$ , the only values where the property is not obtained, would make the metrics identical to  $PWF_{\beta}$ ,  $AUC_{ROC}$  and  $AUC_{PR}$ .

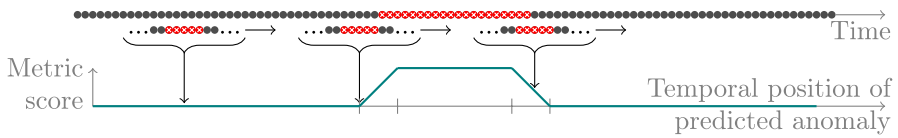
The *Binary* property simply indicates the binary vs non-binary metrics. *Chronology aware* indicates the metrics that consider time dimension adjacency in any way and *Insensitivity to true negatives* are the metrics ignoring the amount of true negatives. *# parameters* indicates the number of parameters for each metric, including  $\beta$  for f-scores, all specifiable functions for  $R_{\beta}^{bias f_{\beta}^{\alpha}}$ , the distance exponent in *TD*, and all the TP, FP and FN weights in *NAB*.

## 8 Conclusion

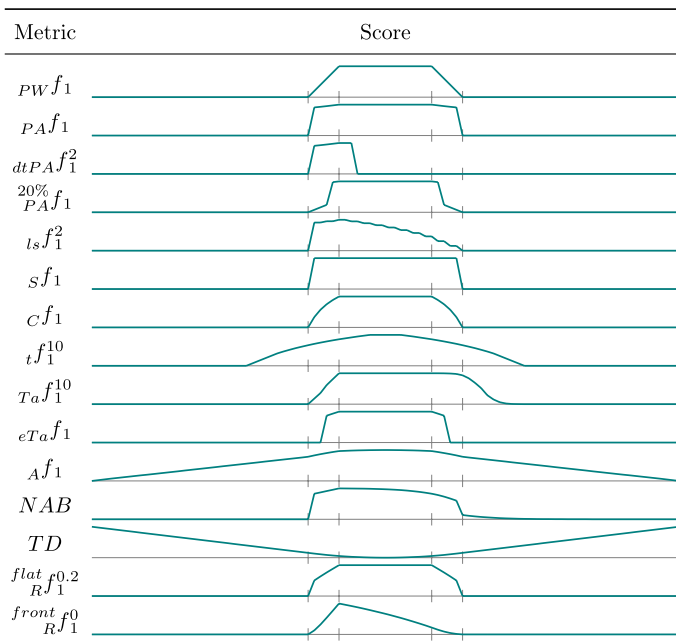
Through an extensive literature review on time series anomaly detection (TSAD), we found several different ways to evaluate algorithms. While a comparison of several of the available metrics can be found in a few papers, some of which strongly disagree with each other on what are important properties of an evaluation metrics, most papers choose metrics that have been repeatedly faulted in the literature, such as the point-adjusted f-score. We have tested 20 TSAD evaluation metrics in several case studies, and categorized them based on 10 different properties. As TSAD is a diverse field, no evaluation metric is appropriate in all cases, and it should be chosen with care in each case. For the same reason, it is difficult to provide detailed guidelines for how to do this. However, we summarize some of the main takeaways from our study:

- The choice of evaluation metric has a large impact on the rankings of TSAD methods, underscoring the need for careful alignment of evaluation metrics with specific problem requirements.
- Some metrics give high scores to certain prediction strategies that do not yield useful predictions. For example, predicting only very long or very short anomalies can result in unreasonably high scores, leading to the selection of inappropriate methods and an overestimation of expected performance.
- Some metrics result in very bad scores for certain types of predictions, even though the predictions are valuable, such as predicting long anomalous events, or predicting anomalies too early or late. This can lead to selecting ineffective methods and underestimating the expected performance.
- Due to the way the labels are compared to the prediction, many metrics are not appropriate for certain kinds of labelling strategies.



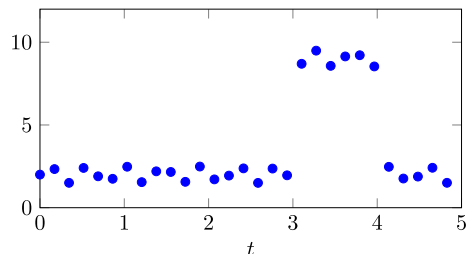


**Fig. 17** Score as a function of position of the predicted event: Each point on the graph is the output of the considered evaluation metric ( $pwf_1$  in this case) for one full prediction, where the position of the predicted anomaly changes. The 4 vertical marks on the axis indicate the start and end of areas where some, but not all, predicted anomaly points are correct. In other words, before the first and after the last mark, the predictions have no TP points, while between the second and third mark, the predictions have no FP points



**Fig. 18** Score as a function of position of the predicted event: For each metric, the graph shows the scores of the detection scenario shown in Fig. 17, as a function of the position in the prediction. All graphs are scaled to the same interval for easy comparison

**Fig. 19** Is there one anomaly at  $t \approx 3$  and another at  $t \approx 4$ , or just one anomalous event from 3 to 4? Without any information about domain and time scale, we may only guess what is an anomaly here, or if there even exist any



	$P@3$	$best_{PW}f_{\beta}$	$AUC_{ROC}$	$AUC_{PR}$	$VUS^4_{ROC}$	$VUS^4_{PR}$
	0.67	0.75	0.96	0.76	0.96	0.83
	0.67	0.75	0.96	0.76	0.96	0.83

**Fig. 20** Ordering the time stamps by value of the anomaly scores yields the same order. Only the order matter for the non-binary metrics, not their value, hence these predictions have the same scores. This would be true for any labels

	$P@11$	$best_{PW}f_{\beta}$	$AUC_{ROC}$	$AUC_{PR}$	$VUS^4_{ROC}$	$VUS^4_{PR}$
	0.41	0.78	0.8	0.78	0.61	0.74
	0.41	0.58	0.6	0.62	0.61	0.69

**Fig. 21** Effect of anomaly length: Importance of anomaly length for non-binary metrics

	$P@7$	$best_{PW}f_{\beta}$	$AUC_{ROC}$	$AUC_{PR}$	$VUS^4_{ROC}$	$VUS^4_{PR}$
	0.32	0.48	0.61	0.4	0.61	0.49
	0.5	0.63	0.76	0.49	0.73	0.63

**Fig. 22** Preference for short predicted anomalies: None of the non-binary metrics prefer the short predicted anomalies

	$P@16$	$best_{PW}f_{\beta}$	$AUC_{ROC}$	$AUC_{PR}$	$VUS^4_{ROC}$	$VUS^4_{PR}$
	0.57	0.73	0.75	0.79	0.65	0.86
	0.57	0.73	0.56	0.62	0.59	0.83

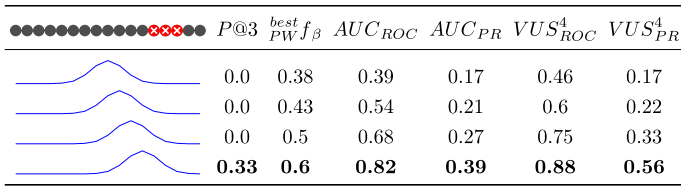
**Fig. 23** Partial detection versus covering: As these metrics are point-based, they do not value detection of new anomalies over fully covering existing ones

	$P@3$	$best_{PW}f_{\beta}$	$AUC_{ROC}$	$AUC_{PR}$	$VUS^4_{ROC}$	$VUS^4_{PR}$
	0.18	0.3	0.46	0.18	0.48	0.11
	0.18	0.3	0.46	0.18	0.48	0.11
	0.18	0.3	0.46	0.18	0.5	0.19
	0.18	0.3	0.46	0.18	0.56	0.4

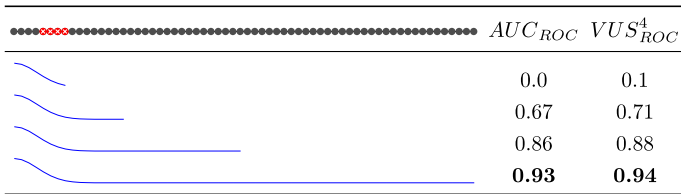
**Fig. 24** Temporal tolerance: Only VUS metrics value proximity of predicted and labelled anomalies. The other metrics still give positive score, since a low enough threshold marks every point as anomalous

Therefore, it is crucial to carefully select the appropriate evaluation metric for a given problem, taking into consideration its preferences for specific types of predictions. Simple case studies such as the ones presented in this work can be helpful for gaining such understanding.

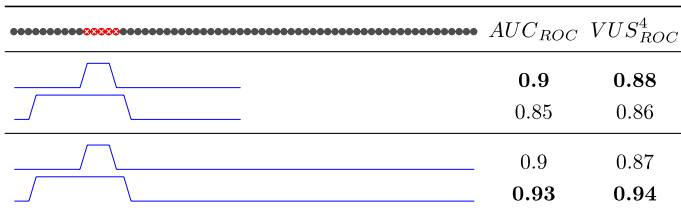
There are several directions of future research based on this study. First of all, there is room for defining novel evaluation metrics. For example, valuing early



**Fig. 25** Temporal tolerance: Non-binary metrics value FPs close to true anomalies indirectly, due to the anomaly score often being somewhat smooth. The scores are gaussian function with different shifts, meaning the anomaly scores at the anomalous points increase as the centre moves towards the anomaly



**Fig. 26** Effect of class imbalance:  $AUC_{ROC}$  and  $VUS_{ROC}^4$  scores are heavily affected by the amount of TNs. In the shorter predictions, only the predicted part of the time series is evaluated. The anomaly score is strictly decreasing, ensuring that none of the added points are more anomalous than the previous ones. As only  $AUC_{ROC}$  and  $VUS_{ROC}^4$  are affected by this, the other metrics are not included



**Fig. 27** Effect of class imbalance: The bottom two anomaly scores are extensions of the top two. The extra TNs change which anomaly score  $AUC_{ROC}$  and  $VUS_{ROC}^4$  prefer

detection and temporal tolerance are two very useful traits, but none of the metrics we could find include both. Furthermore, much more investigation can be done of existing metrics that did not meet the limitations of this work, e.g. supplementary performance analysis metrics, or combinations of techniques of the included metrics. Finally, when publishing results in TSAD research in general, we suggest including results from multiple metrics, as well as making both the code and the anomaly scores available, to enable easy comparison with any evaluation metric.

**Table 2** The properties of all the metrics. ✓ = has property, ✗ = does not have property, \* = partially / parameter dependent

Metric	Valuation properties					Intrinsic properties				# parameters
	Value early detection	Prioritize long anomalies	Favour short predicted events	Prioritize partial detection	Temporal tolerance	Binary	Chronology aware	Insensitivity to true negatives		
$PWf_{\beta}$	✗	✓	✗	✗	✗	✓	✗	✓	✓	1
$PAf_{\beta}$	✗	✓	✓	✓	✗	✓	✓	✓	✓	1
$APf_{\beta}$	*	✓	*	✓	✗	✓	✓	✓	✓	2
$K\%f_{\beta}$	✗	✓	*	*	✗	✓	✓	✓	✓	2
$PAf_{\beta}^m$	*	*	*	✓	*	✓	✓	✓	✓	2
$ISf_{\beta}$	✗	✗	✗	✓	✗	✓	✓	✓	✓	1
$Sf_{\beta}$	✗	✗	✗	✓	✗	✓	✓	✓	✓	1
$Cf_{\beta}$	✗	✗	✗	*	✓	✓	✓	✓	✓	2
$d_{\beta}^T$	✗	✗	✗	*	✗	✓	✓	✓	✓	8
$bias_{Rf_{\beta}}^{\alpha}$	*	✗	✗	*	*	✓	✓	✓	✓	4
$TIdf_{\beta}^s$	✗	✗	✗	*	*	✓	✓	✓	✓	4
$cTIdf_{\beta}$	✗	✗	✗	*	✗	✓	✓	✓	✓	3
$Mf_{\beta}$	✗	✗	✗	✓	✓	✓	✓	✓	✓	1
$NAB$	✓	✗	✓	✓	*	✓	✓	✓	✓	3
$TD$	✗	*	✓	*	✓	✓	✓	✓	✓	1
$P@K$	✗	✓	✗	✗	*	✗	✗	✓	✓	1
$best_f$	✗	✓	✗	✗	*	✗	✗	✓	✓	1
$PWf_{\beta}$	✗	✓	✗	✗	*	✗	✗	✓	✓	1
$AUC_{ROC}$	✗	✓	✗	✗	*	✗	✗	✗	✗	0
$AUC_{PR}$	✗	✓	✗	✗	*	✗	✗	✓	✓	0
$VUS_{ROC}^l$	✗	*	✗	✗	✓	✗	✓	✗	✗	1
$VUS_{PR}^l$	✗	✓	✗	✗	✓	✗	✓	✓	✓	1

**Acknowledgements** This research was carried out with the support of the ML4ITS project (312062), funded by the Norwegian Research Council (NFR).

**Funding** Open access funding provided by SINTEF.

## Declarations

**Competing interests** The authors have no relevant financial or non-financial interests to disclose.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Abdulaal A, Liu Z, Lancewicki T (2021) Practical approach to asynchronous multivariate time series anomaly detection and localization. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery and data mining. Association for computing machinery, New York, NY, USA, KDD '21, p 2485–2494, <https://doi.org/10.1145/3447548.3467174>,
- Ahmad S, Lavin A, Purdy S et al (2017) Unsupervised real-time anomaly detection for streaming data. *Neurocomputing* 262:134–147 <https://doi.org/10.1016/j.neucom.2017.04.070>, [www.sciencedirect.com/science/article/pii/S0925231217309864](http://www.sciencedirect.com/science/article/pii/S0925231217309864), online Real-Time Learning Strategies for Data Streams
- Ahmed AH, Riegler MA, Hicks SA, et al. (2022) Rcad: Real-time collaborative anomaly detection system for mobile broadband networks. In: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining. Association for computing machinery, New York. KDD '22, p 2682–2691, <https://doi.org/10.1145/3534678.3539097>,
- Audibert J, Michiardi P, Guyard F, et al. (2020) Usad: Unsupervised anomaly detection on multivariate time series. In: Proceedings of the 26th ACM SIGKDD International conference on knowledge discovery and data mining. Association for computing machinery, New York. KDD '20, p 3395–3404, <https://doi.org/10.1145/3394486.3403392>,
- Baireddy S, Desai SR, Mathieson JL, et al. (2021) Spacecraft time-series anomaly detection using transfer learning. In: 2021 IEEE/CVF Conference on computer vision and pattern recognition workshops (CVPRW), pp 1951–1960, <https://doi.org/10.1109/CVPRW53098.2021.00223>
- Baker SG, Pinsky PF (2001) A proposed design and analysis for comparing digital and analog mammography. *J Am Stat Assoc* 96(454):421–428. <https://doi.org/10.1198/016214501753168136>
- Bashar MA, Nayak R (2020) Tanogan: Time series anomaly detection with generative adversarial networks. In: 2020 IEEE symposium series on computational intelligence, SSCI 2020, Canberra, December 1–4, 2020. IEEE, pp 1778–1785, <https://doi.org/10.1109/SSCI47803.2020.9308512>
- Berrai DP, Flach PA (2012) Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). *Brief Bioinform* 13(1):83–97. <https://doi.org/10.1093/bib/bbr008>
- Bhatia S, Jain A, Li P, et al. (2021) Mstream: Fast anomaly detection in multi-aspect streams. In: Proceedings of the web conference 2021. Association for computing machinery, New York. WWW '21, p 3371–3382, <https://doi.org/10.1145/3442381.3450023>,
- Braei M, Wagner S (2020) Anomaly detection in univariate time-series: a survey on the state-of-the-art. *CoRR abs/2004.00433*. <https://doi.org/10.48550/arXiv.2004.00433>, [arXiv:2004.00433](https://arxiv.org/abs/2004.00433)

- Buda TS, Assem H, Xu L (2017) ADE: an ensemble approach for early anomaly detection. In: 2017 IFIP/IEEE symposium on integrated network and service management (IM), Lisbon, May 8-12, 2017. IEEE, pp 442–448, <https://doi.org/10.23919/INM.2017.7987310>,
- Campos D, Kieu T, Guo C, et al. (2021) Unsupervised time series outlier detection with diversity-driven convolutional ensembles. *Proc VLDB Endow* 15(3):611–623. <https://doi.org/10.14778/3494124.3494142>, <http://www.vldb.org/pvldb/vol15/p611-chaves.pdf>
- Challu C, Jiang P, Wu YN, et al. (2022) Deep generative model with hierarchical latent factors for time series anomaly detection. In: International conference on artificial intelligence and statistics <https://doi.org/10.48550/arXiv.2202.07586>
- Chen R, Shi G, Zhao W et al (2021) A joint model for IT operation series prediction and anomaly detection. *Neurocomputing* 448:130–139. <https://doi.org/10.1016/j.neucom.2021.03.062>
- Chen Z, Chen D, Yuan Z et al (2021) Learning graph structures with transformer for multivariate time-series anomaly detection in IOT. *IEEE Internet Things J* 9:9179–9189. <https://doi.org/10.1109/JIOT.2021.3100509>
- Chen Z, Chen D, Zhang X et al (2022) Learning graph structures with transformer for multivariate time-series anomaly detection in iot. *IEEE Internet Things J* 9(12):9179–9189. <https://doi.org/10.1109/JIOT.2021.3100509>
- Chen X, Deng L, Huang F, et al. (2021b) DAEMON: unsupervised anomaly detection and interpretation for multivariate time series. In: 37th IEEE international conference on data engineering, ICDE 2021, Chania, April 19-22, 2021. IEEE, pp 2225–2230, <https://doi.org/10.1109/ICDE51399.2021.00228>,
- Chen T, Liu X, Xia B, et al. (2020) Unsupervised anomaly detection of industrial robots using sliding-window convolutional variational autoencoder. *IEEE Access* 8:47,072–47,081. <https://doi.org/10.1109/ACCESS.2020.2977892>,
- Choi K, Yi J, Park C, et al. (2021) Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines. *IEEE Access* 9:120,043–120,065. <https://doi.org/10.1109/ACCESS.2021.3107975>
- Chuah MC, Fu F (2007) ECG anomaly detection via time series analysis. In: Thulasiraman P, He X, Xu TL, et al. (eds) *Frontiers of high performance computing and networking ISPA 2007 workshops, ISPA 2007 international workshops SSDSN, UPWN, WISH, SGC, ParDMCom, HiP-CoMB, and IST-AWSN Niagara Falls*. August 28 - September 1, 2007, Proceedings, Lecture Notes in Computer Science, vol 4743. Springer, pp 123–135, [https://doi.org/10.1007/978-3-540-74767-3\\_14](https://doi.org/10.1007/978-3-540-74767-3_14),
- Dai E, Chen J (2022) Graph-augmented normalizing flows for anomaly detection of multiple time series. *ArXiv abs/2202.07857*. <https://doi.org/10.48550/arXiv.2202.07857>
- Dai L, Lin T, Liu C, et al. (2021) Sdfvae: Static and dynamic factorized vae for anomaly detection of multivariate cdn kpis. In: *Proceedings of the web conference 2021*. Association for computing machinery, New York, WWW '21, p 3076–3086, <https://doi.org/10.1145/3442381.3450013>,
- Davis J, Goadrich M (2006) The relationship between precision-recall and ROC curves. In: Cohen WW, Moore AW (eds) *Machine learning. Proceedings of the twenty-third international conference (ICML 2006)*. Pittsburgh, Pennsylvania, USA, June 25-29, 2006, ACM international conference proceeding series, vol 148. ACM, pp 233–240, <https://doi.org/10.1145/1143844.1143874>,
- Deng L, Lian D, Huang Z et al (2022) Graph convolutional adversarial networks for spatiotemporal anomaly detection. *IEEE Trans Neural Netw Learn Syst* 33(6):2416–2428. <https://doi.org/10.1109/TNNLS.2021.3136171>
- Deng A, Hooi B (2021) Graph neural network-based anomaly detection in multivariate time series. In: *Thirty-Fifth AAAI conference on artificial intelligence, AAAI 2021, Thirty-third conference on innovative applications of artificial intelligence, IAAI 2021, The eleventh symposium on educational advances in artificial intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, pp 4027–4035, <https://ojs.aaai.org/index.php/AAAI/article/view/16523>
- Doshi K, Abudalou S, Yilmaz Y (2022) Reward once, penalize once: Rectifying time series anomaly detection. In: *International joint conference on neural networks, IJCNN 2022, Padua, July 18-23, 2022*. IEEE, pp 1–8, <https://doi.org/10.1109/IJCNN55064.2022.9891913>,
- Du B, Sun X, Ye J et al (2021) Gan-based anomaly detection for multivariate time series using polluted training set. *IEEE Trans Knowl Data Eng* 5:1–1. <https://doi.org/10.1109/TKDE.2021.3128667>
- Ergen T, Kozat SS (2020) Unsupervised anomaly detection with LSTM neural networks. *IEEE Trans Neural Netw Learn Syst* 31(8):3127–3141. <https://doi.org/10.1109/TNNLS.2019.2935975>

- Feng Y, Liu Z, Chen J et al (2022) Unsupervised multimodal anomaly detection with missing sources for liquid rocket engine. *IEEE Trans Neural Netw Learn Syst* 9:1–15. <https://doi.org/10.1109/TNNLS.2022.3162949>
- Feng C, Tian P (2021) Time series anomaly detection for cyber-physical systems via neural system identification and bayesian filtering. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery and data mining. Association for computing machinery, New York. KDD '21, p 2858–2867, <https://doi.org/10.1145/3447548.3467137>,
- Flaborea A, Prenkaj B, Munjal B, et al. (2022) Are we certain it's anomalous? ArXiv abs/2211.09224. <https://doi.org/10.48550/arXiv.2211.09224>
- Garg A, Zhang W, Samaran J et al (2022) An evaluation of anomaly detection and diagnosis in multivariate time series. *IEEE Trans Neural Netw Learn Syst* 33(6):2508–2517. <https://doi.org/10.1109/TNNLS.2021.3105827>
- Geiger A, Liu D, Alnegheimish S, et al. (2020) Tadgan: Time series anomaly detection using generative adversarial networks. In: Wu X, Jermaine C, Xiong L, et al. (eds) 2020 IEEE international conference on big data (IEEE BigData 2020), Atlanta, GA, USA, December 10–13, 2020. IEEE, pp 33–43, <https://doi.org/10.1109/BigData50022.2020.9378139>,
- Gensler A, Sick B (2014) Novel criteria to measure performance of time series segmentation techniques. In: Seidl T, Hassani M, Beecks C (eds) Proceedings of the 16th LWA Workshops: KDML, IR and FGWM, Aachen, Germany, September 8–10, 2014, CEUR workshop proceedings, vol 1226. CEUR-WS.org, pp 193–204, <http://ceur-ws.org/Vol-1226/paper31.pdf>
- Goode A, Hooi B, Ng S, et al. (2020) Robustness of autoencoders for anomaly detection under adversarial impact. In: Bessiere C (ed) Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI 2020. ijcai.org, pp 1244–1250, <https://doi.org/10.24963/ijcai.2020/173>,
- Goswami M, Challu C, Callot L, et al. (2022) Unsupervised model selection for time-series anomaly detection. ArXiv abs/2210.01078. <https://doi.org/10.48550/arXiv.2210.01078>
- Han S, Woo SS (2022) Learning sparse latent graph representations for anomaly detection in multivariate time series. In: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining. Association for computing machinery, New York. KDD '22, p 2977–2986, <https://doi.org/10.1145/3534678.3539117>,
- He Y, Zhao J (2019) Temporal convolutional networks for anomaly detection in time series. *J Phys Conf Ser* 4:1213. <https://doi.org/10.1088/1742-6596/1213/4/042050>
- He Z, Chen P, Li X et al (2020) A spatiotemporal deep learning approach for unsupervised anomaly detection in cloud systems. *IEEE Trans Neural Netw Learn Syst* 12:3027736. <https://doi.org/10.1109/TNNLS.2020.3027736>
- Hsieh RJ, Chou J, Ho CH (2019) Unsupervised online anomaly detection on multivariate sensing time series data for smart manufacturing. 2019 IEEE 12th conference on service-oriented computing and applications (SOCA) pp 90–97. <https://doi.org/10.1109/SOCA.2019.00021>
- Huang T, Chen P, Li R (2022) A semi-supervised vae based active anomaly detection framework in multivariate time series for online systems. In: Proceedings of the ACM web conference 2022. Association for computing machinery. New York. WWW '22, p 1797–1806, <https://doi.org/10.1145/3485447.3511984>,
- Huang X, Lee J, Kwon YW, et al. (2020) Crowdquake: A networked system of low-cost sensors for earthquake detection via deep learning. Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery and data mining <https://doi.org/10.1145/3394486.3403378>
- Huet A, Navarro JM, Rossi D (2022) Local evaluation of time series anomaly detection algorithms. In: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining. Association for computing machinery. New York. KDD '22, p 635–645, <https://doi.org/10.1145/3534678.3539339>
- Hundman K, Constantinou V, Laporte C, et al. (2018) Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In: Guo Y, Farooq F (eds) Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery and data mining, KDD 2018. London. August 19–23, 2018. ACM, pp 387–395, <https://doi.org/10.1145/3219819.3219845>
- Hwang WS, Yun JH, Kim J, et al. (2022) "do you know existing accuracy metrics overrate time-series anomaly detections?". In: Proceedings of the 37th ACM/SIGAPP symposium on applied computing. Association for computing machinery. New York, SAC '22, p 403–412, <https://doi.org/10.1145/3477314.3507024>,
- Hwang W, Yun J, Kim J, et al. (2019) Time-series aware precision and recall for anomaly detection: Considering variety of detection result and addressing ambiguous labeling. In: Zhu W, Tao D, Cheng X,

- et al. (eds) Proceedings of the 28th ACM international conference on information and knowledge management, CIKM 2019. Beijing, China, November 3-7, 2019. ACM, pp 2241–2244, <https://doi.org/10.1145/3357384.3358118>,
- Jacob V, Song F, Stiegler A, et al. (2021) Exathlon: A benchmark for explainable anomaly detection over time series. *Proc VLDB Endow* 14(11), 2613–2626. <https://doi.org/10.14778/3476249.3476307>
- Keogh EJ, Lin J, Fu AWC et al (2006) Finding unusual medical time-series subsequences: algorithms and applications. *IEEE Trans Inf Technol Biomed* 10:429–439. <https://doi.org/10.1109/TITB.2005.863870>
- Kieu T, Yang B, Guo C, et al. (2019) Outlier detection for time series with recurrent autoencoder ensembles. In: International joint conference on artificial intelligence, <https://doi.org/10.24963/ijcai.2019/378>
- Kim GY, Lim SM, Euom IC (2022) A study on performance metrics for anomaly detection based on industrial control system operation data. *Electronics* 11(8):1108213. <https://doi.org/10.3390/electronics11081213>
- Kim S, Choi K, Choi H, et al. (2022b) Towards a rigorous evaluation of time-series anomaly detection. In: Thirty-sixth AAAI conference on artificial intelligence, AAAI 2022, Thirty-fourth conference on innovative applications of artificial intelligence, IAAI 2022, The twelfth symposium on educational advances in artificial intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022. AAAI Press, pp 7194–7201, <https://ojs.aaai.org/index.php/AAAI/article/view/20680>
- Kovács G, Sebestyen G, Hangan A (2019) Evaluation metrics for anomaly detection algorithms in time-series. *Acta Univ Sapientiae Inf* 11:113–130. <https://doi.org/10.2478/ausi-2019-0008>
- Lai K, Zha D, Xu J, et al. (2021) Revisiting time series outlier detection: Definitions and benchmarks. In: Vanschoren J, Yeung S (eds) Proceedings of the neural information processing systems track on datasets and benchmarks 1, NeurIPS datasets and benchmarks 2021, December 2021, virtual, <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/ec5decca5ed3d6b8079e2e7ebacc9f2-Abstract-round1.html>
- Lavin A, Ahmad S (2015a) Evaluating real-time anomaly detection algorithms - the numenta anomaly benchmark. In: Li T, Kurgan LA, Palade V, et al. (eds) 14th IEEE international conference on machine learning and applications, ICMLA 2015, Miami. December 9-11, 2015. IEEE, pp 38–44, <https://doi.org/10.1109/ICMLA.2015.141>,
- Lavin A, Ahmad S (2015b) The numenta anomaly benchmark [White paper]. Redwood City, CA: Numenta, Available: <https://github.com/numenta/NAB/wiki>
- Li L, Yan J, Wang H et al (2021) Anomaly detection of time series with smoothness-inducing sequential variational auto-encoder. *IEEE Trans Neural Netw Learn Syst* 32(3):1177–1191. <https://doi.org/10.1109/TNNLS.2020.2980749>
- Li Y, Peng X, Zhang J et al (2021) Dct-gan: dilated convolutional transformer-based gan for time series anomaly detection. *IEEE Trans Knowl Data Eng* 23:1–1. <https://doi.org/10.1109/TKDE.2021.3130234>
- Li L, Yan J, Wen Q et al (2022) Learning robust deep state space for unsupervised anomaly detection in contaminated time-series. *IEEE Trans Knowl Data Eng* 23:1–1. <https://doi.org/10.1109/TKDE.2022.3171562>
- Li D, Chen D, Shi L, et al. (2019) Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks. In: International conference on artificial neural networks [https://doi.org/10.1007/978-3-030-30490-4\\_56](https://doi.org/10.1007/978-3-030-30490-4_56)
- Liu S, Zhou B, Ding QX et al (2022) Time series anomaly detection with adversarial reconstruction networks. *IEEE Trans Knowl Data Eng*. <https://doi.org/10.1109/tkde.2021.3140058>
- Li Z, Zhao Y, Han J, et al. (2021c) Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery and data mining. association for computing machinery, New York. KDD '21, p 3220–3230, <https://doi.org/10.1145/3447548.3467075>,
- Lobo JM, Jiménez-Valverde A, Real R (2008) Auc: a misleading measure of the performance of predictive distribution models. *Glob Ecol Biogeogr* 17:145–151. <https://doi.org/10.1111/J.1466-8238.2007.00358.X>
- Mamandipoor B, Majd M, Sheikhalishahi S et al (2020) Monitoring and detecting faults in wastewater treatment plants using deep learning. *Environ Monitor Assess* 192:1–12. <https://doi.org/10.1007/s10661-020-8064-1>



- Ma M, Zhang S, Chen J, et al. (2021) Jump-starting multivariate time series anomaly detection for online service systems. In: USENIX annual technical conference, <https://www.usenix.org/conference/atc21/presentation/ma>
- Meng H, Zhang Y, Li Y, et al. (2020) Spacecraft anomaly detection via transformer reconstruction error. In: Jing Z (ed) Proceedings of the international conference on aerospace system science and engineering 2019. Springer, Singapore, pp 351–362, [https://doi.org/10.1007/978-981-15-1773-0\\_28](https://doi.org/10.1007/978-981-15-1773-0_28)
- Nalepa J, Myller M, Andrzejewski J et al (2022) Evaluating algorithms for anomaly detection in satellite telemetry data. *Acta Astronautica* 198:689–701 <https://doi.org/10.1016/j.actaastro.2022.06.026>, [www.sciencedirect.com/science/article/pii/S0094576522003162](http://www.sciencedirect.com/science/article/pii/S0094576522003162)
- Niu Z, Yu K, Wu X (2020) Lstm-based vae-gan for time-series anomaly detection. *Sens Basel Switz* 20:3738. <https://doi.org/10.3390/s20133738>
- Pang G, Shen C, van den Hengel A (2019) Deep anomaly detection with deviation networks. Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining <https://doi.org/10.1145/3292500.3330871>
- Paparrizos J, Boniol P, Palpanas T, et al. (2022a) Volume under the surface: A new accuracy evaluation measure for time-series anomaly detection. *Proc VLDB Endow* 15:2774–2787. <https://doi.org/10.14778/3551793.3551830>
- Paparrizos J, Kang Y, Boniol P, et al. (2022b) Tsb-uad: An end-to-end benchmark suite for univariate time-series anomaly detection. *Proc VLDB Endow* 15(8):1697–1711. <https://doi.org/10.14778/3529337.3529354>
- Park D, Hoshi Y, Kemp CC (2017) A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder. *IEEE Robot Autom Lett* 3:1544–1551. <https://doi.org/10.1109/LRA.2018.2801475>
- Pedregosa F, Varoquaux G, Gramfort A, et al. (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12:2825–2830. <https://doi.org/10.48550/arXiv.1201.0490>
- Ren H, Xu B, Wang Y, et al. (2019) Time-series anomaly detection service at microsoft. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining. Association for computing machinery, New York. KDD '19, p 3009–3017, <https://doi.org/10.1145/3292500.3330680>,
- Rewicki F, Denzler J, Niebling J (2022) Is it worth it? an experimental comparison of six deep- and classical machine learning methods for unsupervised anomaly detection in time series. *ArXiv abs/2212.11080*. <https://doi.org/10.48550/arXiv.2212.11080>
- Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* 10:118432. <https://doi.org/10.1371/journal.pone.0118432>
- Scharwächter E, Müller E (2020) Statistical Evaluation of Anomaly Detectors for Sequences. In: 6th ACM SIGKDD workshop on mining and learning from time series (KDD MiLeTS 2020), <https://doi.org/10.48550/arXiv.2008.05788>
- Schmidl S, Wenig P, Papenbrock T (2022) Anomaly detection in time series: a comprehensive evaluation. *Proc VLDB Endow* 15(9):1779–1797. <https://doi.org/10.14778/3538598.3538602>,
- Shen L, Li Z, Kwok J (2020) Timeseries anomaly detection using temporal hierarchical one-class network. In: Larochelle H, Ranzato M, Hadsell R, et al. (eds) Advances in neural information processing systems, vol 33. curran associates, Inc., pp 13,016–13,026, <https://proceedings.neurips.cc/paper/2020/file/97e401a02082021fd24957f852e0e475-Paper.pdf>
- Sivaraks H, Ratanamahatana C (2015) Robust and accurate anomaly detection in ecg artifacts using time series motif discovery. *Comput Math Methods Med* 2015:45314. <https://doi.org/10.1155/2015/453214>
- Su Y, Zhao Y, Niu C, et al. (2019) Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining. Association for computing machinery, New York. KDD '19, p 2828–2837, <https://doi.org/10.1145/3292500.3330672>,
- Tatbul N, Lee TJ, Zdonik S, et al. (2018) Precision and recall for time series. In: Bengio S, Wallach HM, Larochelle H, et al. (eds) Advances in neural information processing systems 31: annual conference on neural information processing systems 2018. NeurIPS 2018, December 3–8, 2018, Montréal, Canada, pp 1924–1934, <https://proceedings.neurips.cc/paper/2018/hash/8f468c873a32bb0619eab2050ba45d1-Abstract.html>

- Tuli S, Casale G, Jennings NR (2022) Tranad: deep transformer networks for anomaly detection in multivariate time series data. *Proc VLDB Endow* 15:1201–1214. <https://doi.org/10.48550/arXiv.2201.07284>
- Wang Y, Han L, Liu W et al (2019) Study on wavelet neural network based anomaly detection in ocean observing data series. *Ocean Eng*. <https://doi.org/10.1016/j.oceaneng.2019.106129>
- Wang X, Pi D, Zhang X et al (2022) Variational transformer-based anomaly detection approach for multivariate time series. *Measurement*. <https://doi.org/10.1016/j.measurement.2022.110791>
- Wang Y, Du X, Lu Z et al (2022) Improved lstm-based time-series anomaly detection in rail transit operation environments. *IEEE Trans Indust Inform* 18:9027–9036. <https://doi.org/10.1109/TII.2022.3164087>
- Wu R, Keogh EJ (2021) Ucr\_anomalydatasets.pptx, supplemental material to the ucr anomaly archive. [https://www.cs.ucr.edu/%7Eeamonn/time\\_series\\_data\\_2018/UCR\\_TimeSeriesAnomalyDatasets2021.zip](https://www.cs.ucr.edu/%7Eeamonn/time_series_data_2018/UCR_TimeSeriesAnomalyDatasets2021.zip), accessed: 2022-11-15
- Wu R, Keogh EJ (2022) Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress (extended abstract). In: 2022 IEEE 38th international conference on data engineering (ICDE), pp 1479–1480. <https://doi.org/10.1109/ICDE53745.2022.00116>
- Xu H, Chen W, Zhao N, et al. (2018) Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. In: Proceedings of the 2018 world wide web conference. International world wide web conferences steering committee, republic and canton of Geneva. CHE, WWW '18, p 187–196. <https://doi.org/10.1145/3178876.3185996>,
- Xu H, Wang Y, Jian S, et al. (2022) Calibrated one-class classification for unsupervised time series anomaly detection. *CoRR* abs/2207.12201. <https://doi.org/10.48550/arXiv.2207.12201>,
- Zhang CK, Li SZ, Zhang H, et al. (2020) Velc: A new variational autoencoder based model for time series anomaly detection. *arXiv:1907.01702*
- Zhang M, Li T, Shi H, et al. (2019) A decomposition approach for urban anomaly detection across spatiotemporal data. In: Kraus S (ed) Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI 2019, Macao. August 10–16, 2019. ijcai.org, pp 6043–6049. <https://doi.org/10.24963/ijcai.2019/837>,
- Zhang C, Song D, Chen Y, et al. (2018) A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. *ArXiv* abs/1811.08055. <https://doi.org/10.1609/aaai.v33i01.33011409>
- Zhang J, Wu D, Boulet B (2021) Time series anomaly detection for smart grids: A survey. 2021 IEEE electrical power and energy conference (EPEC) pp 125–130. <https://doi.org/10.1109/EPEC52095.2021.9621752>
- Zhao H, Wang Y, Duan J, et al. (2020) Multivariate time-series anomaly detection via graph attention network. In: 2020 IEEE international conference on data mining (ICDM), pp 841–850. <https://doi.org/10.1109/ICDM50108.2020.00093>
- Zhou B, Liu S, Hooi B, et al. (2019) Beatgan: Anomalous rhythm detection using adversarially generated time series. In: International joint conference on artificial intelligence. <https://doi.org/10.24963/ijcai.2019/616>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.