



Delft University of Technology

PATE

Proximity-Aware Time Series Anomaly Evaluation

Ghorbani, Ramin; Reinders, Marcel J.T.; Tax, David M.J.

DOI

[10.1145/3637528.3671971](https://doi.org/10.1145/3637528.3671971)

Publication date

2024

Document Version

Final published version

Published in

KDD '24

Citation (APA)

Ghorbani, R., Reinders, M. J. T., & Tax, D. M. J. (2024). PATE: Proximity-Aware Time Series Anomaly Evaluation. In *KDD '24: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 872-883). ACM. <https://doi.org/10.1145/3637528.3671971>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.



PATE: Proximity-Aware Time Series Anomaly Evaluation

Ramin Ghorbani*
Delft University of Technology
Delft, Netherlands
r.ghorbani@tudelft.nl

Marcel J.T. Reinders
Delft University of Technology
Delft, Netherlands
m.j.t.reinders@tudelft.nl

David M.J. Tax
Delft University of Technology
Delft, Netherlands
d.m.j.tax@tudelft.nl

ABSTRACT

Evaluating anomaly detection algorithms in time series data is critical as inaccuracies can lead to flawed decision-making in various domains where real-time analytics and data-driven strategies are essential. Traditional performance metrics assume iid data and fail to capture the complex temporal dynamics and specific characteristics of time series anomalies, such as early and delayed detections. We introduce Proximity-Aware Time series anomaly Evaluation (PATE), a novel evaluation metric that incorporates the temporal relationship between prediction and anomaly intervals. PATE uses proximity-based weighting considering buffer zones around anomaly intervals, enabling a more detailed and informed assessment of a detection. Using these weights, PATE computes a weighted version of the area under the Precision and Recall curve. Our experiments with synthetic and real-world datasets show the superiority of PATE in providing more sensible and accurate evaluations than other evaluation metrics. We also tested several state-of-the-art anomaly detectors across various benchmark datasets using the PATE evaluation scheme. The results show that a common metric like Point-Adjusted F1 Score fails to characterize the detection performances well, and that PATE is able to provide a more fair model comparison. By introducing PATE, we redefine the understanding of model efficacy that steers future studies toward developing more effective and accurate detection models.

Public source code: <https://github.com/Raminghorbanii/PATE>

CCS CONCEPTS

- **General and reference** → **Evaluation; Metrics; Measurement;**
- **Mathematics of computing** → Mathematical analysis.

KEYWORDS

Time Series, Anomaly Detection, Evaluation Metrics, Precision, Recall

ACM Reference Format:

Ramin Ghorbani, Marcel J.T. Reinders, and David M.J. Tax. 2024. PATE: Proximity-Aware Time Series Anomaly Evaluation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*, August 25–29, 2024, Barcelona, Spain. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3637528.3671971>

*Corresponding Author

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
KDD '24, August 25–29, 2024, Barcelona, Spain
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0490-1/24/08
<https://doi.org/10.1145/3637528.3671971>

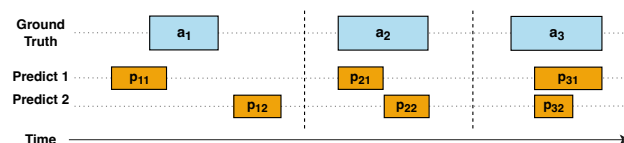


Figure 1: Illustration of anomaly detection in time series data. a_1 – a_3 represent the actual anomalies as ground truth. Predictions are denoted by p . The durations of both events are indicated by the length of the boxes. Overlapping areas between p and a demonstrate where the model has correctly identified anomalies.

1 INTRODUCTION

Anomaly detection in time series (TS) data, the process of identifying unusual patterns that deviate from the expected norm, has become increasingly important across various domains [6, 7]. The rapid advancement of data-driven decision-making and real-time analytics has opened opportunities for developing more accurate anomaly detection methods. Such developments often lead to models competing to claim the status of 'State-of-the-Art' (SOTA). Achieving this status is not just a matter of academic prestige; it often directs the focus of future research, influences industry adoption, and guides the development of practical applications. However, choosing an appropriate evaluation metric is critical to avoid incorrect conclusions about a model's performance. Relying on evaluation metrics that do not accurately reflect the true effectiveness of the models can lead to flawed decisions in real-world applications. This is particularly consequential in critical domains, such as medical diagnostics or financial fraud detection, where relying on a poorly evaluated model can have serious repercussions.

Standard evaluation metrics such as Precision and Recall [2] are effective for point-based anomaly detection as they assess the accuracy of detecting isolated iid events. In this context, each data point is evaluated independently, allowing for straightforward calculation of these metrics. However, in TS data, events and anomalies typically occur in time intervals. This complexity causes several situations: 1) *Early Detection*, when potential anomalies are identified before they fully manifest, based on subtle changes in the data pattern over time. Figure 1 shows an example of early detection where prediction p_{11} detects the anomaly event a_1 earlier than its actual occurrence. Although p_{11} does not align exactly with a_1 , such early detection is valuable for early response actions and should be appropriately appreciated in evaluation metrics. 2) *Delayed Detection*, occurs when an anomaly event is not detected immediately but is identified at a later time, even after its actual occurrence. In Figure 1, the anomaly event a_1 is detected with a delay by prediction event p_{12} . Although p_{12} does not align precisely with a_1 , this type of delayed detection should be accounted for in the evaluation process,

as it reflects the model’s ability to eventually identify anomalies, even after some delay.

Another situation, 3) *Onset Response Time*, refers to how close the detection of an anomaly is to the start of the event. Timely detection is valuable, especially in scenarios where immediate action is required. In Figure 1, anomaly event a_2 is detected by p_{21} and p_{22} . However, p_{21} aligns more closely with the beginning of the anomaly event a_2 , indicating a faster response than p_{22} . Evaluation metrics should reward those that occur promptly after the onset of an anomaly. Finally 4) *Coverage level of Predictions*, refers to the range that a prediction covers an actual anomaly. The effectiveness of a prediction can be measured by how much of the anomaly it successfully captures. In Figure 1, predictions p_{31} and p_{32} both detect anomaly event a_3 , but p_{31} covers a_3 more than p_{32} . This more extensive coverage by p_{31} makes it a more effective prediction for a_3 . Accordingly, evaluation metrics need to consider the coverage range of the predictions over the duration of the anomalies.

Various metrics have been developed that are specifically tailored to the sequential nature of time series data (referred to as *Sequential Adaptability*). For instance, Range-based Precision and Recall metrics, hereafter denoted as *R-based* [18], expand upon traditional metrics by incorporating factors such as existence (detecting the anomaly range with at least one point), size and position (reflecting the number and relative position of correctly detected anomaly ranges), and cardinality (penalizing fragmented predictions for a single anomaly). The Time Series Aware Precision and Recall, hereafter denoted as *TS-Aware* [10], follows a similar approach but omits cardinality and position considerations. This metric requires a prediction to cover a minimum percentage θ of an anomaly for it to be considered a true detection. They also add a buffer zone δ to give some credit for delayed detection in a decreasing manner. An enhanced version, denoted as *ETS-Aware* [11], further refines the evaluation by combining detection and overlap scores for improved accuracy in scoring overlapped detections. Further, the *Affiliation* metric [8], introduces a different perspective by focusing on the distance between prediction and actual anomaly ranges. It assesses the proximity of predicted anomalies to actual ones by measuring the duration between their respective ranges.

Another widely used method is the Point Adjusted F1 Score metric, which we will denote as *PA-F1* [20]. This approach assumes that detecting a single point in an anomaly range is sufficient for human experts to identify the entire range. Thus, it considers all observations within the corresponding anomaly range as correctly detected anomalies. However, it has been criticized for potentially generating optimistic scores. For example, [12] revealed that random anomaly scores from a uniform distribution outperform state-of-the-art methods when evaluated using this metric. To address this, [12] proposed a modified version that requires a portion of $K\%$ of the anomaly range to be detected before making any adjustments.

While all these metrics represent advancements in time series anomaly detection evaluation, they do not fully consider all the critical factors of early and delayed detections, or onset response timing. In addition to these limitations, the aforementioned metrics also require the setting of a threshold, a value where data points with anomaly scores exceeding this value are classified as anomalies. Selecting this threshold adds additional complexity and leads to subjectivity and inconsistency in evaluations. Metrics such as the

Table 1: Comparison of Anomaly Detection Evaluation Metrics. Key features: Sequential Adaptability (SA); Early Detection (ED); Delayed Detection (DD); Onset Response Time (ORT); Coverage Level (CL) and Threshold-Free (TF)

Metric	SA	ED	DD	ORT	CL	TF
Precision/Recall (F1 Score)	-	-	-	-	-	-
R-based	✓	-	-	-	-	-
TS-Aware/ETS-Aware	✓	-	✓	-	-	-
Affiliation	✓	-	-	-	✓	-
PA-F1	✓	-	-	-	-	-
AUC-ROC/PR	-	-	-	-	-	✓
VUS-ROC/PR	✓	-	✓	-	-	✓
PATE	✓	✓	✓	✓	✓	✓

Area Under the Receiver Operating Characteristic curve (AUC-ROC) and the Area Under the Precision-Recall curve (AUC-PR) eliminate the need for thresholding by evaluating the performance of the model across a range of thresholds. However, they fall short in time series contexts due to not considering the order of the data points and the temporal correlation between them. In response to this issue, Volume Under the Surface (VUS) metrics, *VUS-ROC* and *VUS-PR*, are proposed [16]. These metrics acknowledge the need to accommodate close predictions to the true anomaly ranges by adjusting the labels to be between 0 and 1 on a range over both sides of the actual anomaly range. Although the method is threshold-free, it does not pay attention to early and delayed detection, and onset response time. Furthermore, by changing the original labels, the metric gives unrealistic scores, as reaching the maximum detection score of 1 is not possible.

This paper introduces the Proximity-Aware Time series anomaly Evaluation metric, PATE (/pert/). Our novel metric integrates buffer zones around the anomaly events and utilizes a special proximity-based weighting mechanism, enabling a detailed assessment of both early/delayed detections and addressing the onset response time challenge. PATE avoids the subjectivity of threshold-dependent metrics by integrating over the range of thresholds, offering a fair and unbiased evaluation, especially in research settings where expert knowledge might not be available for setting the exact desirable parameters based on the application. Table 1 illustrates a comparison between existing metrics and PATE, highlighting the comprehensive adaptability reconsideration of PATE in evaluating the TS anomaly detection.

2 PROPOSED EVALUATION METRIC - PATE

A time series is denoted as a sequence of observations $X = \{x_t\}_{t=1}^T$, where T represents the length of the time series, and each x_t is the observed data point at time t .

An actual anomaly event (labeled as positive in the ground truth labels) is a subsegment within the time series, denoted as $a_k = (i_k, n_k)$ for points i_k and n_k with $1 \leq i_k \leq n_k \leq T$. The set of all anomaly events in the time series is represented as $A = \{a_k\}_{k=1}^N$, where N is the number of anomaly events present in the time series.

In practice, the detection models output continuous anomaly scores, denoted as $S = \{s_t\}_{t=1}^T$, representing the likelihood of each observation x_t to be anomalous. These scores are then converted into binary predictions by applying a threshold θ , where scores

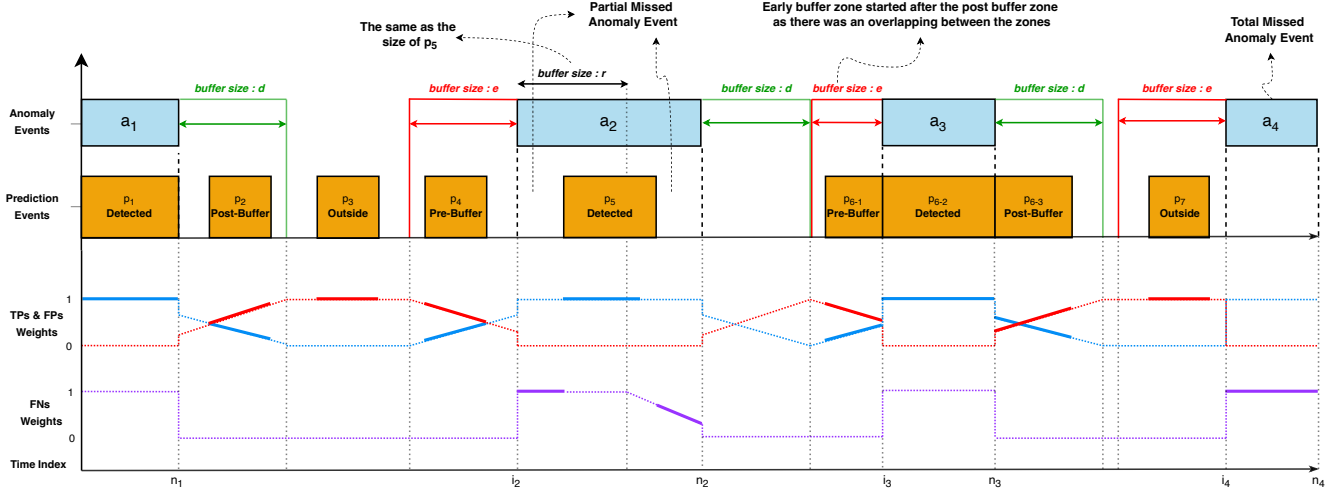


Figure 2: Illustration of the Categorization and Weighting Mechanism in the PATE Method. Prediction events ($p_1 - p_7$) are represented by orange boxes, while anomaly events ($a_1 - a_4$) are depicted by blue boxes. TP weights are illustrated with a blue line —, FP weights with a red line —, and FN weights with a purple line —. Note that the solid segments of the lines, in contrast to the dotted segments, indicate the activated weights for the example scenario depicted in the figure.

equal to or exceeding the threshold are classified as anomalies. We define a prediction event as a subsegment identified by these binary predictions to be anomalous, denoted as $p_l(\theta) = (m_l, j_l)$ for points m_l and j_l with $1 \leq m_l \leq j_l \leq T$. The set of all prediction events is represented as $P = \{p_l(\theta)\}_{l=1}^M$, where M is the number of prediction events identified by the model.

The effectiveness of the anomaly detector is determined by how well these $p_l(\theta)$ events align with the a_k events. PATE distinguishes several categories of matches between ground truth and predictions based on their temporal relationships and assigns proximity-specific weights to each point in each category. These weights are then used to compute a weighted version of Precision and Recall scores. The final measure of PATE is a weighted AUC-PR, which is derived from these weighted Precision and Recall scores. Further details on these computations are provided in the following sections.

2.1 Categorizing the Events

Figure 2 illustrates the different categories of anomaly and prediction events in relation to each other. In assessing each $p_l(\theta)$, we consider its overlap, proximity, or distance (temporal relation) from each a_k . This approach allows for the clear differentiation of the diverse scenarios: complete and partial detection of anomalies, early or delayed detection, and instances where anomalies are either partially or entirely missed. Specifically, we categorize the anomaly and prediction events as follows:

2.1.1 Prediction events categories:

- **True-Detection:** Sub-segments of the prediction event $p_l(\theta)$ that overlap with an anomaly event a_k , indicating anomalies that are accurately identified and not missed. Examples are segments p_1 , p_5 , and p_{6-2} in Figure 2.

- **Post-Buffer Detection:** Sub-segments of the prediction event $p_l(\theta)$ that fall into a buffer zone immediately following an anomaly

event a_k (See segments p_2 and p_{6-3} in Figure 2). This category highlights the capacity of the model for delayed detection. The post-buffer zone size, denoted by d , can be adjusted by experts based on specific application needs. When d is unknown for a specific application, we can consider a range of values for d rather than a fixed one $D = \{0, 1, \dots, d_{\max}\}$. This approach allows for a comprehensive assessment of the model's performance across different scenarios, as each buffer size can provide a different perspective on the performance of the model. Details on how these buffer sizes contribute to the overall PATE score will be discussed in the following sections.

- **Pre-Buffer Detection:** Sub-segments of the prediction event $p_l(\theta)$ that fall into a zone that precedes the start of an anomaly event a_k . This category highlights the capacity of the model for early detection, signaling potential anomalies ahead of time. Similar to the post-buffer zone, the size of the pre-buffer zone, denoted by e , varies within the set $E = \{0, 1, \dots, e_{\max}\}$ with the same approach for the assessment. The assignment of points to this category is conditional on not overlapping with the Post-Buffer zone of a preceding anomaly a_{k-1} , ensuring that the model early warning is distinct from a delayed detection of the previous event. In other words, the Post-Buffer category has priority, and therefore, if $i_k - e < n_{k-1} + d$ then the Pre-Buffer zone starts at $n_{k-1} + d + 1$ instead of $i_k - e$. Furthermore, Pre-Buffer detection is dependent on the successful detection of the subsequent anomaly event a_k . In situations where no part of the subsequent event a_k is detected by a True-Detection, this Pre-Buffer detection is considered a false alarm rather than a meaningful early detection. Consequently, this early prediction $p_l(\theta)$ is reclassified as False Positive (the Outside category, which is discussed below). Further details are given in Appendix C. In Figure 2, p_4 and p_{6-1} are the examples of pre-buffer detection category, whereas p_7 is not considered in this category.

- **Outside:** Sub-segments of the prediction event $p_l(\theta)$ located outside the ranges of anomaly event a_k and its buffer zones. These are instances where the model incorrectly flags normal behavior as anomalous (False Positive), like segments p_3 and p_7 in Figure 2.

2.1.2 Anomaly events categories:

- **Total Missed Anomalies:** When an entire anomaly event a_k is not detected by any segments of the prediction event $p_l(\theta)$, that is, all detections are before $i_k - e$ or after $n_k + d$. This category indicates a complete failure (False Negative) of the model to identify the anomaly. See segment a_4 in Figure 2.

- **Partial Missed Anomalies:** This category is assigned when only a part of anomaly event a_k is detected by the prediction events $p_l(\theta)$'s, but there are segments within the anomaly range of a_k that remain undetected. This category not only highlights the model's capability to detect parts of an anomaly but also its inability to identify the anomaly event in its entirety. For instance, segment a_2 in Figure 2, where a part of it is detected by p_5 but before and after p_5 we have partially missed segments.

2.2 Weighting Process

After each individual time point is assigned to its category, we define weights for each of these points to determine their contribution to the True Positive (TP), False Positive (FP), and False Negative (FN) metrics of the detector. It is important to note that time points at which no anomaly is present and no prediction is made, True Negatives (TN), do not actively contribute to the performance metrics and are, therefore, implicitly assigned a weight of zero, reflecting their non-contribution. The bottom half of Figure 2 visually represents the variations in weights across all different categories.

- **True-Detection Weights:** Each point t from the True-Detection category, lying within the range of an anomaly event $[i_k, n_k]$, is considered correctly identified. Thus, such points are assigned the maximum weight of 1 as True Positives:

$$w^{\text{TP}}(t) = 1 \quad \text{for } t \in \text{TrueDetection } p_l(\theta) \quad (1)$$

- **Post-Buffer Detection Weights:** Each point t from the post-buffer category, in the range of $(n_k, n_k + d]$, is evaluated in relation to the anomaly event a_k . These points, while not being true positives in the traditional sense, receive a weight based on their proximity to the a_k , which captures the diminishing influence of an anomaly over time as the distance from the anomaly event increases.

$$w^{\text{TP}}(t) = 1 - \frac{\sum_{y=i_k}^{n_k} |t - y|}{\sum_{y=i_k}^{n_k} |(n_k + d) - y|} \quad \text{for } t \in \text{Post-Buffer } p_l(\theta) \quad (2)$$

Here, the numerator calculates the distance of t from each point within the anomaly event, and the denominator normalizes this against the total potential spread within the buffer zone. With this method, we account for the proximity to the entire anomaly, not just its endpoint. Thus, we address the delayed detection by recognizing that any point within the actual anomaly range might influence

predictions in the buffer zone, not just the most immediate or final points of the anomaly. This also implies that the lengths of the anomalies influence the weights. For smaller anomalies, points in the Post-Buffer zone are closer to the anomaly onset, and will therefore be assigned with higher true positive weights. Further details, regarding the impact of anomaly length on the weights, are given in Appendix B.

In the Post-Buffer zone, as the distance from a_k increases, the likelihood of a detection being a False Positive rises. Thus, the weights assigned to false positives in this zone are calculated as the complement of the TPs weights, acknowledging the reduced significance of detections further from the actual anomaly. Figure 2 visually shows the variations in TP and FP weights across the Post-Buffer categories (p_2 and $p_{6(3)}$).

$$w^{\text{FP}}(t) = 1 - w^{\text{TP}}(t) \quad \text{for } t \in \text{Post-Buffer } p_l(\theta) \quad (3)$$

- **Outside Weights:** Each point t from the Outside category indicates a situation where the model incorrectly identifies normal behavior as anomalous. Given the lack of proximity to any real anomaly, these points are considered FPs with a maximum weight of 1, reflecting a significant deviation from accurate detection.

$$w^{\text{FP}}(t) = 1 \quad \text{for } t \in \text{Outside } p_l(\theta) \quad (4)$$

- **Pre-Buffer Detection Weights:** Each point t in the pre-buffer category, in the range of $[i_k - e, i_k)$, is assessed for potential early detection in relation to the preceding a_k . These points, while not being true positives in the conventional sense, are evaluated for their proximity to the upcoming anomaly:

$$w^{\text{TP}}(t) = 1 - \frac{\sum_{y=i_k}^{n_k} |y - t|}{\sum_{y=i_k}^{n_k} |(i_k - e) - y|} \quad \text{for } t \in \text{Pre-Buffer } p_l(\theta) \quad (5)$$

Here, the numerator represents the distance of t from every point in a_k , capturing how early t occurs relative to the anomaly. The denominator provides normalization against the total potential spread within the pre-buffer zone. This mechanism recognizes that any point within the anomaly event might have an influence on the zone.

Similar to the Post-Buffer zone, the likelihood of a point being a False Positive increases as the distance from the i_k increases. Thus, the weights assigned to FPs are calculated as the complement of the TPs weights, reflecting the reduced relevance of premature detections. Figure 2 shows the variations in weights of the Pre-Buffer categories (p_4 and $p_{6(1)}$).

$$w^{\text{FP}}(t) = 1 - w^{\text{TP}}(t) \quad \text{for } t \in \text{Pre-Buffer } p_l(\theta) \quad (6)$$

- **Total Missed Anomalies Weights:** When the entire range of a_k is undetected, each t within its interval receives a maximum False Negative weight of 1. This assignment underscores the complete failure of the model in detecting the anomaly event. Figure 2 shows the variations in FN weight across a_4 as a total missed event.

$$w^{\text{FN}}(t) = 1 \quad \text{for } t \in \text{Total-Missed } a_k \quad (7)$$

• **Partial Missed Anomaly Weights:** When a_k is only partially detected, the undetected points t within a_k , are evaluated based on their proximity to the start of the anomaly event. The closer the points are to the anomaly onset the higher the FN weight, emphasizing the onset response time in detection. Here for $t \in$ Partial Missed a_k , we have:

$$w^{\text{FN}}(t) = \begin{cases} 1 & \text{if } t \leq i_k + r \\ 1 - \frac{\sum_{y=i_k}^{i_k+r} |t-y|}{\sum_{y=i_k}^{n_k} |n_k-y|} & \text{otherwise} \end{cases} \quad (8)$$

Here, r is the size of the buffer that starts from the onset of the anomaly event. Undetected points in this buffer are penalized with a maximum FN weight of 1. Undetected points outside the buffer received a reduced FN weight, weighted by the distance to the buffer. The rationale behind this design is that more comprehensive coverage of an anomaly by a prediction justifies a more lenient assessment of its exact timing accuracy. In other words, when a prediction successfully captures a larger portion of a_k , the precision of its onset timing becomes less critical. Therefore, r is defined as the fraction of coverage of a_k by its corresponding $p_l(\theta)$. Figure 2 shows the variations in FN weight across the Partial Missed category where some segments of a_2 are missed.

2.3 PATE Final Score

The PATE final metric is designed to comprehensively evaluate anomaly detection by considering a full range of combinations of pre-buffer (e) and post-buffer (d) sizes. For each combination of e and d , we apply a range of thresholds (θ) to convert the continuous anomaly scores (S) into binary predictions, capturing the model's performance across different sensitivity levels. Based on these binary predictions, we identify the prediction events P and then categorize all prediction and anomaly events. Based on this categorization, we assign appropriate weights to each observation.

We calculate weighted Precision and Recall across all thresholds in the considered range for each specific combination of e and d . Using these calculations, we construct the Precision-Recall curve for each combination and compute the area under the curve (AUC-PR). Note that the weights $w^{\text{TP}}(t)$, $w^{\text{FP}}(t)$, and $w^{\text{FN}}(t)$ are assigned based on the categorization of each time point t . For time points that do not fall into any specific category, the weights are considered to be 0. Thus, the summation in the formulas for Precision and Recall effectively includes only those time points that have been categorized.

$$\text{Precision}_{e,d}(\theta) = \frac{\sum_{t=1}^T w^{\text{TP}}(t)}{\sum_{t=1}^T w^{\text{TP}}(t) + \sum_{t=1}^T w^{\text{FP}}(t)} \quad (9)$$

$$\text{Recall}_{e,d}(\theta) = \frac{\sum_{t=1}^T w^{\text{TP}}(t)}{\sum_{t=1}^T w^{\text{TP}}(t) + \sum_{t=1}^T w^{\text{FN}}(t)} \quad (10)$$

Finally, the overall PATE score is determined by averaging the computed AUC-PRs across all combinations of e and d :

$$\text{PATE} = \frac{1}{|E| \times |D|} \sum_{e \in E, d \in D} \text{AUC-PR}_{e,d} \quad (11)$$

Here, $|D|$ and $|E|$ represent the number of distinct values for d and e within their respective sets.

3 EXPERIMENTS AND RESULTS

3.1 Synthetic Data Experiments

To highlight the merits of PATE, we first compare PATE with alternative evaluation metrics on a synthetic time series with a binary anomaly detector. The alternative measures can be threshold-dependent or independent. Threshold-independent metrics are inherently evaluated across a range of possible thresholds. For this example, we consider thresholds $\theta = \{0, 1\}$ to distinguish between normal and anomalous predictions. For threshold-dependent metrics, we define the optimal threshold as $\theta = 1$, identifying points predicted as '1' (anomalous) for evaluation.

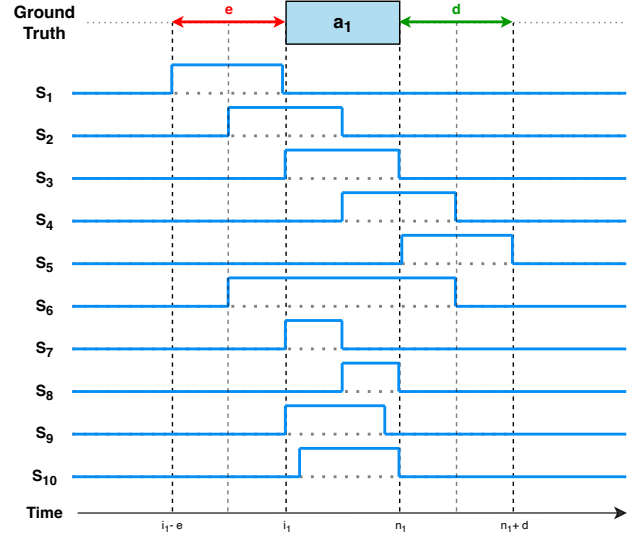


Figure 3: *Illustration of examples with synthetic data.* The figure shows the placement of different anomaly scores S from a binary anomaly detector.

Table 2: *Comparison of evaluation metrics for synthetic data examples depicted in Figure 3.* 'F1' refers to the F1 Score. 'Standard-F1' specifically denotes the conventional F1 Score calculated from standard Precision and Recall.

Scenarios	Threshold-independent Metrics					Threshold-dependent Metrics			
	PATE	VUS-ROC	VUS-PR	AUC-ROC	AUC-PR	Standard-F1	PA-F1	R-based-F1	ETS-Aware-F1
S_1	0.03	0.63	0.37	0.48	0.02	0.00	0.00	0.00	0.00
S_2	0.76	0.79	0.72	0.74	0.51	0.50	0.80	0.60	0.75
S_3	1.00	0.87	0.88	1.00	1.00	1.00	1.00	1.00	1.00
S_4	0.69	0.79	0.70	0.74	0.51	0.50	0.80	0.60	0.75
S_5	0.31	0.63	0.34	0.48	0.02	0.00	0.00	0.00	0.00
S_6	0.87	0.99	0.91	0.98	0.75	0.67	0.67	0.75	0.86
S_7	0.85	0.69	0.71	0.75	0.76	0.67	1.00	0.75	0.86
S_8	0.77	0.69	0.71	0.75	0.76	0.67	1.00	0.75	0.86
S_9	0.95	0.78	0.79	0.88	0.88	0.86	1.00	0.89	0.93
S_{10}	0.88	0.78	0.79	0.88	0.88	0.86	1.00	0.89	0.93

Figure 3 shows anomaly a_1 with its pre and post-buffer zones. Below, ten different detection scenarios are shown, S_1, \dots, S_{10} . Results in Table 2 demonstrate that PATE effectively distinguishes the scenarios based on temporal proximity, duration, coverage level, and response timing. For instance, although S_1 is temporally close to the anomaly event, it fails to detect any part of it. In the context of time series, where past data is crucial for prediction, the inability to detect any part of the anomaly after it starts suggests that the prediction might be a true false alarm rather than a meaningful early detection. A low score for S_1 reflects a metric that appropriately penalizes lucky guesses or irrelevant detections. On the other hand, S_2 gets a higher score as it captures part of the anomaly itself, and then the non-overlapping part can be recognized as relevant early detection, which should be valued. Note that the PATE score of 0.03 for S_1 is not exactly zero because it considers a range of thresholds, including zero. At a threshold of 0, every point is labeled as a potential anomaly, thus increasing both true and false positives. This broad consideration prevents the PATE score from being zero for this specific example.

Meanwhile, S_1 and S_2 should be evaluated differently from delayed detections S_4 and S_5 . Although S_4 's coverage level is the same as that of S_2 , due to response timing, it gets a lower score. Similarly, the evaluation of S_5 is completely different from S_1 as it occurs after the anomaly event. This late detection might indicate that the model is responding to the anomaly, albeit with a significant delay. Hence, it is reasonable to evaluate S_5 higher than S_1 as it could reflect some response to the actual anomaly, even though it is late and fails to detect any part of the anomaly. Other metrics, while effective in certain scenarios, do not distinguish between the finer details of anomaly detection. For instance, these metrics just mirror the results of S_1 and S_2 for S_4 and S_5 without considering the early and delayed context. Moreover, S_3 , as an example of accurate detection, is expected to get the maximum score of 1 by all evaluation metrics, and S_6 is expected to get a lower score than S_3 . However, the VUS-ROC/PR metrics fail to evaluate these scenarios correctly. The scenarios S_7, S_8, S_9 , and S_{10} further exemplify the importance of the coverage level and response timing in detection. In each pair, S_7 and S_9 detect the anomaly right from the start; thus they should get scored higher than S_8 and S_{10} . While other metrics tend to score these pairs similarly, PATE recognizes the earlier detections in S_7 and S_9 and gives them higher scores. Moreover, in scenarios like S_9 and S_{10} , where the anomaly is covered more extensively, PATE assigns less penalties for response timing inaccuracies. This is seen in the smaller score difference between early and late detections in scenarios with greater coverage.

3.2 Real-World Data Experiments

To validate the practicality and effectiveness of PATE in real-world applications, we extracted some examples from the publicly available and widely used datasets, UCR-KDD21 [19] and MIT-BIH Arrhythmia (MBA) ECG [15]. The goal is to evaluate how well PATE, alongside other evaluation metrics, distinguishes between various detection models. To ensure a fair comparison, we compare PATE with threshold-independent evaluation metrics, guaranteeing an unbiased comparison of metrics performances.

We analyzed the anomaly scores generated by 1) a Perfect Model, which serves as the benchmark by perfectly identifying anomalies; 2) established models like MultiVariate Normal distribution (MVN) [5], Autoencoder (AE)[13], and Local Outlier Factor (LOF)[4]; 3) a baseline Random Score that assigns scores uniformly at random from a $[0, 1]$ distribution. This selection covers a spectrum from theoretically ideal to practically random, offering a comprehensive view of the metrics' potential evaluation range. Detailed implementation of the models is available in our public code repository.

Figure 4 showcases two real-world examples: (a) Weather Temperature data from UCR-KDD21 and (b) ECG data. The top row of each example shows the time series data with actual anomalies

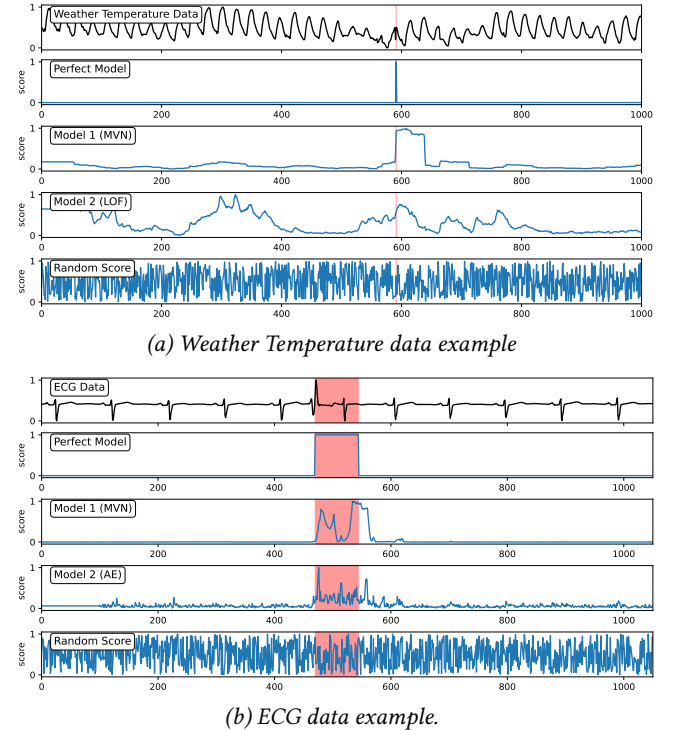


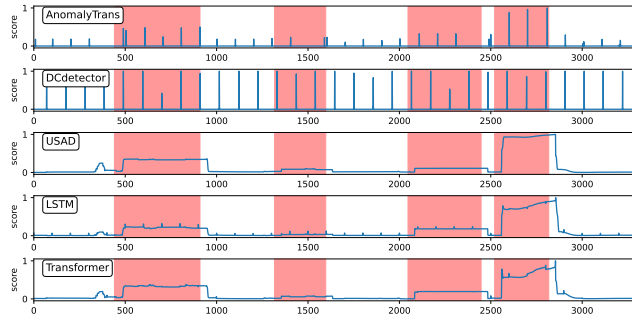
Figure 4: Real-World Datasets and Anomaly Scores of Different Models. The anomalous segment and its corresponding region (labeled by an expert), against which the models' predictions are compared, is highlighted in red

Table 3: Quantitative Evaluation of Anomaly Detection Models. Evaluation score for different anomaly detection models in detecting the anomalous region in examples of Figure 4.

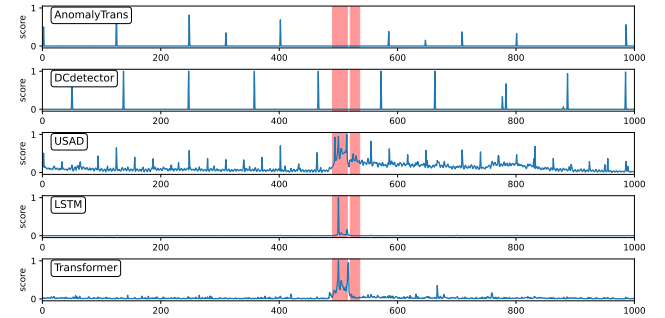
Datasets	Weather Temperature					ECG				
	PATE	VUS-ROC	VUS-PR	AUC-ROC	AUC-PR	PATE	VUS-ROC	VUS-PR	AUC-ROC	AUC-PR
Scenarios										
Perfect Model	1.00	0.55	0.57	1.00	1.00	1.00	0.90	0.91	1.00	1.00
Model 1	0.88	0.98	0.71	0.98	0.02	0.83	0.99	0.89	0.98	0.69
Model 2	0.07	0.86	0.14	0.83	0.01	0.79	0.98	0.81	0.97	0.69
Random Score	0.02	0.67	0.08	0.66	0.01	0.07	0.56	0.11	0.43	0.06

Table 4: Comparison of SOTA anomaly detection model using different evaluation metrics across various benchmark datasets.

Datasets	SMD						MSL						SWaT						PSM					
	PATE	PA-F1	Standard-F1	PA-AUC-ROC	AUC-ROC	VUS-ROC	PATE	PA-F1	Standard-F1	PA-AUC-ROC	AUC-ROC	VUS-ROC	PATE	PA-F1	Standard-F1	PA-AUC-ROC	AUC-ROC	VUS-ROC	PATE	PA-F1	Standard-F1	PA-AUC-ROC	AUC-ROC	VUS-ROC
AnomalyTrans	0.06	0.91	0.03	0.96	0.49	0.50	0.13	0.94	0.02	0.97	0.49	0.52	0.19	0.94	0.02	0.97	0.53	0.54	0.33	0.98	0.02	0.99	0.51	0.52
DCdetector	0.07	0.87	0.01	0.94	0.50	0.51	0.14	0.97	0.02	0.98	0.50	0.58	0.12	0.96	0.02	0.99	0.49	0.50	0.32	0.98	0.02	0.99	0.50	0.52
USAD	0.16	0.94	0.13	0.91	0.63	0.72	0.17	0.91	0.06	0.92	0.53	0.58	0.73	0.85	0.25	0.83	0.82	0.61	0.45	0.89	0.07	0.91	0.60	0.61
LSTM	0.25	0.80	0.14	0.87	0.76	0.81	0.19	0.82	0.08	0.87	0.57	0.64	0.71	0.82	0.03	0.85	0.82	0.60	0.55	0.93	0.15	0.94	0.73	0.73
Transformer	0.27	0.75	0.14	0.84	0.74	0.80	0.20	0.40	0.07	0.63	0.60	0.66	0.72	0.82	0.03	0.85	0.82	0.57	0.56	0.91	0.14	0.92	0.72	0.72



(a) Anomaly Scores of SOTA models for SWaT dataset.



(b) Anomaly Scores of SOTA models for SMD dataset.

Figure 5: Segments of anomaly scores of SOTA models for SWaT and SMD dataset. The highlighted regions in red indicate the true anomaly periods (labeled by an expert).

highlighted in red. The next rows illustrate the the output of the Perfect Model, and Models 1 and 2 (represented by MVN, LOF, or AE), demonstrating their respective detection scores. The final row displays a random score for baseline comparison. Table 3 quantitatively compares various metrics. PATE consistently rates the Perfect Model highest and the Random Score lowest, showing its capability to recognize optimal detection and effectively penalize poor performance. In contrast, VUS-ROC/PR and AUC-ROC metrics seem less capable of such differentiation with the baselines.

Moreover, PATE accurately takes into account the time series context and delayed detection effect, offering a more realistic and conservative assessment compared to VUS-ROC and AUC-ROC metrics, which appear to overestimate the performance of Models 1 and 2. This overestimation is evident in the Weather Temperature data, where Model 2 is inaccurately scored high by VUS-ROC and AUC-ROC despite its poor detection. Additionally, AUC-PR is also not sensitive in evaluation. For instance, in the Weather Temperature data, Model 1’s delayed yet successful detection is incorrectly evaluated with a very low score, similar to the detection of Model 2. Similarly, in the ECG data, PATE’s evaluation reflects the inconsistent anomaly detection pattern of Model 2 (AE) compared to Model 1 (MVN). However, AUC-ROC/PR and VUS-ROC do not effectively consider this difference. Overall, PATE’s assessments across both examples underscore its effectiveness in real-world applications.

3.3 Impact Analysis: SOTA Models

We re-evaluated several recent SOTA anomaly detection methods to not only assess their true performance but also to examine the

stability of their ranking across various benchmark datasets when evaluated with different metrics, including PATE. Our comparative analysis includes models such as DCdetector [22], AnomalyTrans [21], and USAD [3], all of which have been recognized for their high performance in recent studies, alongside a Transformer and LSTM model, as simpler reconstruction-based anomaly detector baselines. These models are tested across the benchmark datasets of SMD [17], MSL [9], SWaT [14], and PSM [1], used in previous works. Implementation details are available in our public code repository.

In the literature on SOTA models, the PA-F1 is the most frequently used and widely accepted metric. Additionally, in some cases, the standard F1 Score and Point-Adjusted variant of AUC-ROC (PA-AUC-ROC) are also employed. For a comprehensive comparison, we included these metrics in our comparative analysis. Results, shown in Table 4, highlight a significant discrepancy between PATE scores and those obtained from other metrics like PA-F1, Standard F1 Score, and PA-AUC-ROC. Notably, models that performed exceptionally well under PA-F1 and PA-AUC-ROC, such as AnomalyTrans and DCdetector, exhibit markedly lower scores when evaluated with PATE. For instance, for the SMD dataset, AnomalyTrans achieves a PA-F1 score of 0.91, showcasing high performance, yet its PATE score is only 0.06, indicating a substantial reduction in performance. To visually illustrate the differences in detection quality, Figure 5 shows a portion of the anomaly scores for the SWaT and SMD. The figures show that AnomalyTrans and DCdetector models struggle with consistent detection. In particular, for the SWaT, the peaky detections by these models hardly align

with the expert-labeled anomaly intervals, and the high values reported for PA-F1 and PA-AUC-ROC do not reflect this detection pattern. This suggests that these metrics may overestimate model effectiveness.

Next, Table 4 shows that the Standard F1 Score, AUC-ROC, and VUS-ROC, do not exhibit such overestimations. However, they lack sensitivity to the finer aspects of detection as discussed in section 2.1. For instance, on the SWaT dataset, the Standard F1 Score is not able to distinguish between the good performing LSTM and Transformer and the poorly performing AnomalyTrans and DCdetector, see also Figure 5 (a). Furthermore, AUC-ROC does not reflect the small differences between USAD, LSTM, or Transformer. The scores of this metric suggest that all models have an identical performance, that does not match the reality of their output. Moreover, while VUS-ROC offers a slightly better distinction among models than AUC-ROC, its limited scoring range (e.g., 0.54 for AnomalyTrans and 0.57 for Transformer) makes it challenging to clearly identify models that perform exceptionally well from those that do not. Meanwhile, PATE offers a more consistent and transparent assessment. It can be seen that PATE gives a relatively higher score to USAD (0.73), Transformer (0.72), and LSTM (0.71) according to their better detection pattern. PATE even slightly prefers USAD over LSTM, although the difference is small.

We also explored the average rankings of the models for all metrics across all four benchmark datasets. Figure 6 presents these rankings, highlighting noticeable differences in the standings of the models when using different metrics. The average rankings based on the PA-F1 metric place DCdetector at the forefront with an average rank of 1.62, followed by AnomalyTrans (1.88), USAD (3.00), LSTM (3.88), and Transformer (4.62). However, when evaluated with PATE, a significant shift occurs: Transformer and LSTM emerge as the top-performing models with ranks of 1.38 and 1.12, respectively, while AnomalyTrans and DCdetector drop to the bottom ranks of 4.50 each. This variance underscores the critical impact of the chosen evaluation metric and the importance of selecting a proper metric such as PATE.

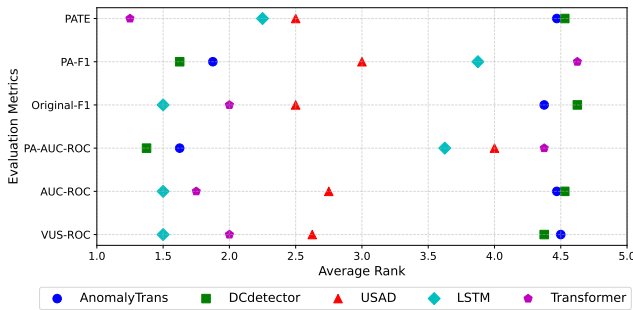


Figure 6: Average rankings of different models for various evaluation metrics across all benchmark datasets.

4 ABLATION ANALYSIS: BUFFER SIZES

The adaptability of PATE to accommodate different buffer sizes is one of its key strengths. This flexibility allows for an expert-driven and context-specific approach to model evaluation, ensuring that the unique characteristics of each dataset are appropriately

considered. Figure 7 illustrates the mean performance of DCdetector, AnomalyTrans, USAD, LSTM, and Transformer across all four benchmark datasets using PATE. Results show that PATE consistently ranks models such as Transformer and LSTM the highest across different buffer sizes. This consistency in model rankings, irrespective of buffer size, highlights PATE’s robustness as an evaluation metric, and showcases PATE’s reliability for diverse applications, ensuring a consistent and dependable assessment for anomaly detection models.

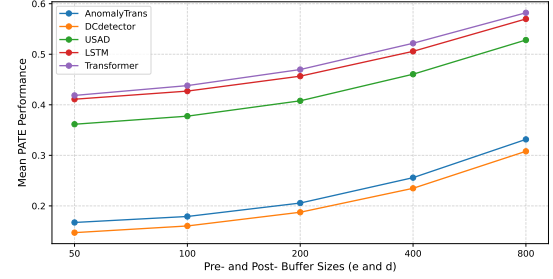


Figure 7: Mean PATE performance of all models across all datasets for different Pre and Post-Buffer sizes ($e = d$).

5 DISCUSSION AND CONCLUSION

We proposed PATE, a novel approach to evaluate anomaly detection models in time series data. PATE addresses the limitations of existing evaluation metrics by categorizing the anomaly and prediction events and assigning proximity-based weighting, considering different buffer zones around the anomaly event. PATE computes the area under the Precision-Recall curve, where the Precision and Recall are computed from weighted versions of True Positive, False Positive, and False Negative performances.

Our experiments with both synthetic and real-world data demonstrate that PATE effectively differentiates between models based on their actual performance, considering early and delayed detection, onset response time, coverage level of the anomaly event, and consistency in detection. The re-evaluation of SOTA anomaly detection methods using PATE reveals notable differences in performance assessments compared to other metrics. For instance, point-adjusted metrics often overestimate the performance of models. However, in practice, metrics such as ROC-AUC and VUS-ROC offer more reasonable estimates for SOTA models, though they might overlook subtle detection errors and sometimes lack discriminability between models. This analysis not only questions the true performance of current SOTA models but also indicates a shift in their rankings, challenging the prevailing understanding of the superiority of these models. PATE’s ability to provide a more matching, context-sensitive, and transparent assessment highlights its potential as a more appropriate metric that can set a new standard for evaluating advancements in anomaly detection. Additionally, PATE’s adaptability to various buffer sizes without compromising consistency and fairness in model evaluation further highlights its robustness and applicability across diverse applications.

To address the specific scenarios where either an expert has predetermined the threshold or models inherently output binary

labels, we have developed *PATE-F1* as an essential extension of the original PATE framework. The methodology and experimental insights on *PATE-F1* are detailed in Appendix D. *PATE-F1* effectively distinguishes between different scenarios based on temporal proximity, duration, coverage level, and response timing, setting it apart from other metrics that face limitations in capturing these aspects in evaluation. Additionally, our findings indicate that the original PATE framework, through strategic threshold application, naturally extends to effectively evaluate binary outputs. However, employing *PATE-F1* in such scenarios offers a more direct and simplified approach. This adaptation ensures PATE's methodology remains a versatile and applicable measure across a broader spectrum of anomaly detection approaches and contexts.

In conclusion, PATE represents a significant advancement in the evaluation of time series anomaly detection methods which has the potential to guide future research, influence industry adoption, and enhance the development of practical applications in critical domains such as healthcare and finance.

ACKNOWLEDGMENTS

Funding: This work was supported by the Dutch Research Council (NWO) [grant numbers 628.011.214].

REFERENCES

- [1] Ahmed Abdulaal, Zhuanghua Liu, and Tomer Lancewicki. 2021. Practical approach to asynchronous multivariate time series anomaly detection and localization. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 2485–2494.
- [2] Charu C Aggarwal and Charu C Aggarwal. 2017. *An introduction to outlier analysis*. Springer.
- [3] Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A Zuluaga. 2020. Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 3395–3404.
- [4] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. 93–104.
- [5] Sounak Chakraborty. 2011. An Intermediate Course in Probability.
- [6] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)* 41, 3 (2009), 1–58.
- [7] Ramin Ghorbani, Marcel JT Reinders, and David MJ Tax. 2024. Personalized anomaly detection in PPG data using representation learning and biometric identification. *Biomedical Signal Processing and Control* 94 (2024), 106216.
- [8] Alexis Huet, Jose Manuel Navarro, and Dario Rossi. 2022. Local evaluation of time series anomaly detection algorithms. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 635–645.
- [9] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. 2018. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 387–395.
- [10] Won-Seok Hwang, Jeong-Han Yun, Jonguk Kim, and Hyoung Chun Kim. 2019. Time-series aware precision and recall for anomaly detection: considering variety of detection result and addressing ambiguous labeling. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2241–2244.
- [11] Won-Seok Hwang, Jeong-Han Yun, Jonguk Kim, and Byung Gil Min. 2022. Do you know existing accuracy metrics overrate time-series anomaly detections?. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. 403–412.
- [12] Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. 2022. Towards a rigorous evaluation of time-series anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 7194–7201.
- [13] Mark A Kramer. 1991. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal* 37, 2 (1991), 233–243.
- [14] Aditya P Mathur and Nils Ole Tippenhauer. 2016. SWaT: A water treatment testbed for research and training on ICS security. In *2016 international workshop on cyber-physical systems for smart water networks (CySWater)*. IEEE, 31–36.
- [15] George B Moody and Roger G Mark. 2001. The impact of the MIT-BIH arrhythmia database. *IEEE engineering in medicine and biology magazine* 20, 3 (2001), 45–50.
- [16] John Paparrizos, Paul Boniol, Themis Palpanas, Ruey S Tsay, Aaron Elmore, and Michael J Franklin. 2022. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. *Proceedings of the VLDB Endowment* 15, 11 (2022), 2774–2787.
- [17] Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. 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 & data mining*. 2828–2837.
- [18] Nesime Tatbul, Tae Jun Lee, Stan Zdonik, Mejbah Alam, and Justin Gottschlich. 2018. Precision and recall for time series. *Advances in neural information processing systems* 31 (2018).
- [19] Renjie Wu and Eamonn Keogh. 2021. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE Transactions on Knowledge and Data Engineering* (2021).
- [20] Haowen Xu, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, Dan Pei, Yang Feng, 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*. 187–196.
- [21] Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. 2021. Anomaly transformer: Time series anomaly detection with association discrepancy. *arXiv preprint arXiv:2110.02642* (2021).
- [22] Yiyuan Yang, Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. 2023. DCdetector: Dual Attention Contrastive Representation Learning for Time Series Anomaly Detection. *arXiv preprint arXiv:2306.10347* (2023).

A REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, the source code, along with comprehensive documentation, is publicly available at: <https://github.com/Raminghorbanii/PATE>.

This repository includes detailed instructions for using PATE, including how to set the buffer size, and complete descriptions of all models implemented for our experiments, covering configuration settings, training procedures, and experimental details to ensure accurate replication. Researchers seeking additional information are encouraged to contact the corresponding author.

B EFFECT OF ANOMALY LENGTH ON BUFFER WEIGHTS

To explore the effect of anomaly length on the assignment of weights within the PATE framework, we consider three distinct anomaly events with varying durations: a_1 , a_2 , and a_3 , with a_1 being the longest and a_3 the shortest. Each was followed by a post-buffer zone of fixed size d . Figure 8 depicts the potential True Positive (TP) weights along the timeline, capturing the period before the anomaly, within its range, and throughout the post-buffer zone. The analysis of this figure indicates that TP weights for detections in the post-buffer zone are higher for a_3 , the shortest anomaly, and progressively lower for a_1 and a_2 , the longer anomalies. This observation underscores the direct correlation between the duration of an anomaly and the corresponding TP weights assigned to post-buffer detections. Higher TP weights for detections following shorter anomalies signify the critical nature of these detections, as they are in closer proximity to the anomaly onset. The PATE weighting mechanism accommodates this by adjusting the weights based on the distance from detections to the entire anomaly. This phenomenon also extends to the pre-buffer zone, where early detections are similarly influenced by the length of the forthcoming anomaly.

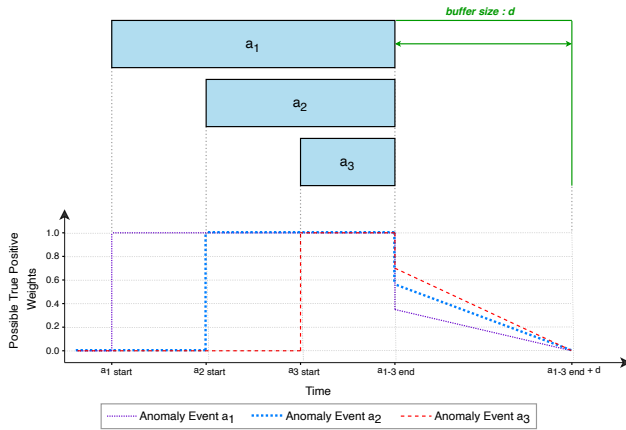


Figure 8: Potential True Positive (TP) weights relative to the anomaly events with varying lengths. The graph illustrates the higher TP weights for detections following the shortest anomaly event a_3 , and the progressively lower weights for the longer events a_1 and a_2 .

C CLARIFICATION ON EARLY AND DELAYED DETECTIONS

To understand the distinct approaches PATE takes toward Early Detection (in the pre-buffer zone) and Delayed Detection (in the post-buffer zone), it is essential to consider the foundational goal of this evaluation metric.

For an anomaly detector, the ability to learn from past data and accurately predict future anomalies is essential. An early prediction that fails to correspond with an actual, subsequent anomaly suggests a fundamental modeling failure of the data’s underlying structure—like sounding an alarm for an event that never happens. Ideally, if a model detects early signs of an impending anomaly, it should also identify the anomaly when it occurs. The early signs—small changes or patterns of deterioration—lead to a larger and more evident departure from the norm. If the model has correctly identified these early signs, it should also recognize the anomaly itself, given the now more noticeable deviation. When the early detection is successfully followed by a true detection of the anomaly, the early detection is not considered just a lucky guess. It supports the model’s predictive power and consistency.

In contrast, the context for delayed detection significantly differs as it showcases the capability of the model to identify anomalies post hoc. The model is apparently able to detect some deviation in the input, albeit a bit late. Such late detections still allow for the identification of the anomaly. Failing to have True Positive detections in the anomaly event is therefore not considered fatal for the Delayed Detection.

Figure 9 shows the detection responses by three different models to an anomalous event, shown by the shaded area in red. Model 1 (top panel) reveals an early detection followed by True Positive detections, indicated by peaks aligning with the anomaly window. This pattern exemplifies an acceptable detection where the model preemptively and accurately identifies an anomaly. Model 2 (middle panel), however, demonstrates early detection without subsequent TPs during the actual anomaly, missing the critical deviation. This outcome might suggest a misinterpretation of the anomaly pattern by Model 2, potentially leading to a false alarm scenario. Conversely, Model 3 (bottom panel) shows a peak that arises post the onset of the anomaly, exemplifying a delayed detection. This detection is valued as it demonstrates the capacity of the model for retrospective analysis, acknowledging and learning from the anomaly event after its occurrence.

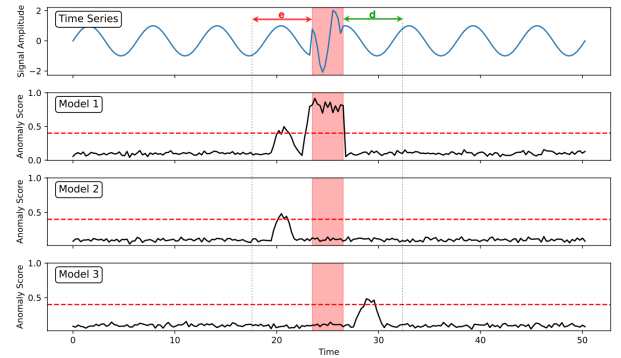


Figure 9: Comparative evaluation of model responses to an anomalous event in time series data.

D PATE-F1 - ADJUSTED FOR BINARY SCORES

• **Methodology:** To enhance the applicability of PATE in scenarios where models use predetermined thresholds or where expert knowledge informs threshold determination, we propose an adapted version, *PATE-F1*. This adaptation leverages the core principles of PATE by assigning proximity-specific weights to categorized points and calculating weighted Precision and Recall. Unlike the original PATE, which evaluates a range of thresholds (θ), *PATE-F1* is tailored for binary scenarios, without the variation of thresholds but rather different combinations of buffer zones (e and d). For each combination, weighted Precision and Recall are calculated using equations 9 and 10 as detailed in Section 2.3. Subsequently, the F1 score for each combination is determined as follows:

$$\text{F1-Score}_{e,d} = 2 \times \frac{\text{Precision}_{e,d} \times \text{Recall}_{e,d}}{\text{Precision}_{e,d} + \text{Recall}_{e,d}} \quad (12)$$

The overall PATE-F1 score is then computed as the average of these F1 scores across all buffer zone combinations:

$$\text{PATE-F1} = \frac{1}{|E| \times |D|} \sum_{e \in E, d \in D} \text{F1-Score}_{e,d} \quad (13)$$

Here, $|E|$ and $|D|$ represent the number of distinct pre-buffer (e) and post-buffer (d) sizes, respectively.

• **Experimental Results:** We extend our analysis to *PATE-F1* by comparing the evaluations against threshold-dependent metrics, tailored for binary score predictions. Figure 10 shows 10 different detection scenarios shown by prediction events p_1, \dots, p_{10} . Table 5 shows that similar to the original PATE, *PATE-F1* effectively differentiates between scenarios based on temporal proximity, duration, coverage level, and response timing. This alignment with PATE’s evaluation logic underlines the adaptability of our methodology to binary score scenarios without compromising the depth of analysis provided by the range of thresholds in the original framework.

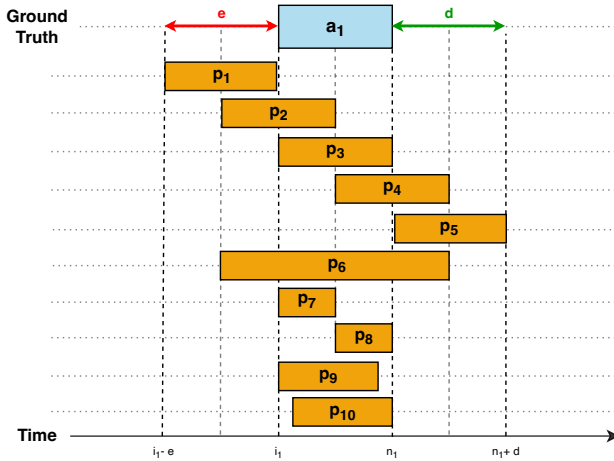


Figure 10: Examples with synthetic prediction events (binary scores). The figure shows the placement of different prediction events $p_i(\theta)$ from a binary anomaly detector.

Table 5: Comparison of evaluation metrics for synthetic prediction event examples depicted in Figure 10. 'F1' refers to the F1 Score.

Scenarios	Metrics						
	PATE	PATE-F1	Standard-F1	PA-F1	R-based-F1	ETS-Aware-F1	Affiliation-F1
p_1	0.03	0.00	0.00	0.00	0.00	0.00	0.94
p_2	0.76	0.75	0.50	0.80	0.60	0.75	0.98
p_3	1.00	1.00	1.00	1.00	1.00	1.00	1.00
p_4	0.69	0.66	0.50	0.80	0.60	0.75	0.98
p_5	0.31	0.28	0.00	0.00	0.00	0.00	0.94
p_6	0.87	0.85	0.67	0.67	0.75	0.86	0.98
p_7	0.85	0.81	0.67	1.00	0.75	0.86	0.99
p_8	0.77	0.67	0.67	1.00	0.75	0.86	0.99
p_9	0.95	0.95	0.86	1.00	0.89	0.93	1.00
p_{10}	0.88	0.86	0.86	1.00	0.89	0.93	1.00

E COMPLEXITY TIME ANALYSIS

We evaluated the computational efficiency of the PATE algorithm against established metrics like AUC-PR and VUS-PR through experiments on synthetic and real benchmark datasets when using a perfect anomaly detector. These experiments were conducted on a standard MacBook with a 2 GHz Quad-Core Intel Core i5 processor, Intel Iris Plus Graphics 1536 MB, and 16 GB RAM, reflecting the performance on commonly available hardware. Although PATE supports parallel execution to potentially decrease computation time, especially on High-Performance Computing (HPC) systems, we used a serial computation approach for consistent comparisons with other metrics.

• **Synthetic Data Experiments:** We generated synthetic time series data ranging from 1,000 to 100,000 points with anomaly ratios of 2%, 5%, and 10% to reflect various common scenarios. As shown in Figure 11, PATE’s computation time increases linearly with data length and varies slightly with different anomaly ratios. Despite this, computation times remained under one second across all conditions, highlighting PATE’s efficiency without parallel processing.

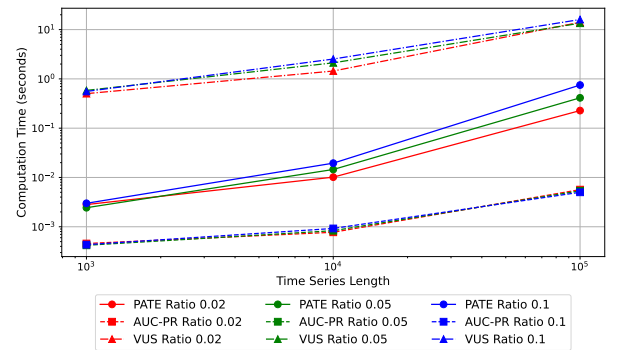


Figure 11: Computation time of PATE on synthetic data with varying lengths and anomaly ratios.

• **Benchmark Dataset Experiments:** We validated PATE on all standard benchmark datasets used in this study. As shown in Table 6, PATE’s computation times are comparable to those of the AUC-PR metric and significantly faster than the VUS metric, remaining under one second for smaller datasets and under two seconds for larger ones. Note that further speed enhancements could be achieved on HPC systems or with parallel processing.

Table 6: Computation times (in seconds) for evaluation metrics across benchmark datasets.

Datasets	Time series Size	Anomaly Ratio	Evaluation Metrics		
			AUC-PR	VUS-PR	PATE
MSL	73700	10%	0.007	42.315	0.278
PSM	87800	4%	0.013	51.683	0.634
SWaT	449900	12%	0.267	249.573	1.895
SMD	708400	4%	0.064	462.252	1.796