

Towards a Rigorous Evaluation of Time-Series Anomaly Detection

Siwon Kim¹ Kukjin Choi^{1,2} Hyun-Soo Choi^{3,4} Byunghan Lee^{5*} Sungroh Yoon^{1,6,7*}

¹ Data Science and AI Laboratory, Seoul National University, Korea

² DIT Center, Samsung Electronics, Korea ³ Ziovision, Korea

⁴ Department of CSE and Education Research Team for Medical Big-data Convergence, Kangwon National University, Korea

⁵ Department of Electronic and IT Media Engineering, Seoul National University of Science and Technology, Korea

⁶ Department of ECE and Interdisciplinary Program in AI, Seoul National University, Korea

⁷ AIIS, ASRI, and INMC, Seoul National University, Korea

{tuslkkk, kj21.choi}@snu.ac.kr choi.hyunsoo@kangwon.ac.kr bhlee@seoultech.ac.kr sryoon@snu.ac.kr

Abstract

In recent years, proposed studies on time-series anomaly detection (TAD) report high F1 scores on benchmark TAD datasets, giving the impression of clear improvements in TAD. However, most studies apply a peculiar evaluation protocol called point adjustment (PA) before scoring. In this paper, we theoretically and experimentally reveal that the PA protocol has a great possibility of overestimating the detection performance; even a random anomaly score can easily turn into a state-of-the-art TAD method. Therefore, the comparison of TAD methods after applying the PA protocol can lead to misguided rankings. Furthermore, we question the potential of existing TAD methods by showing that an untrained model obtains comparable detection performance to the existing methods even when PA is forbidden. Based on our findings, we propose a new baseline and an evaluation protocol. We expect that our study will help a rigorous evaluation of TAD and lead to further improvement in future researches.

1 Introduction

As Industry 4.0 accelerates system automation, consequences of system failures can have a significant social impact (Baheti and Gill 2011; Lee 2008; Lee, Bagheri, and Kao 2015). To prevent this failure, the detection of the anomalous state of a system is more important than ever, and it is being studied under the name of anomaly detection (AD). Meanwhile, deep learning has shown its effectiveness in modeling multivariate time-series data collected from numerous sensors and actuators of large systems (Chalapathy and Chawla 2019). Therefore, various time-series AD (TAD) methods have widely adopted deep learning, and each of them demonstrated its own superiority by reporting higher F1 scores than the preceding methods (Choi et al. 2021). For some datasets, the reported F1 scores exceed 0.9, giving an encouraging impression of today’s TAD capabilities.

However, most of the current TAD methods measure the F1 score after applying a peculiar evaluation protocol named point adjustment (PA), proposed by Xu et al. (Audibert et al.

*Correspondence with Byunghan Lee and Sungroh Yoon
Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

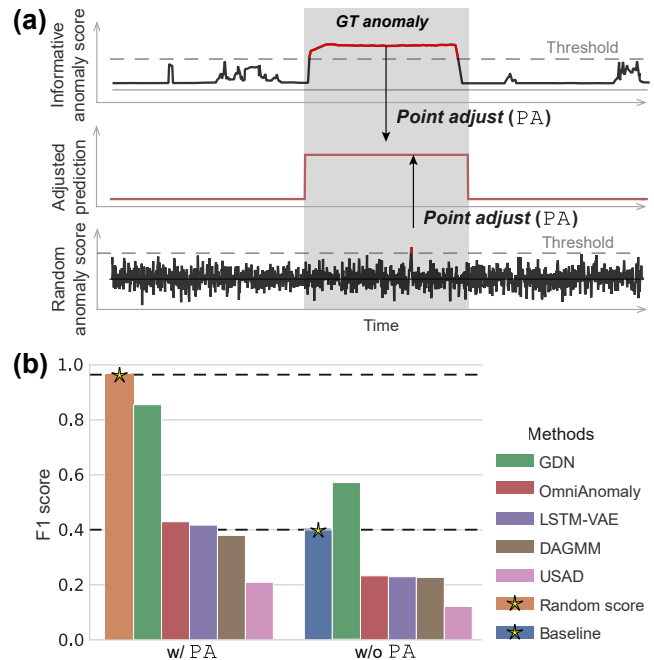


Figure 1: (a) PA makes different anomaly scores indistinguishable. The black lines, gray area, and dashed line indicate the anomaly scores, GT anomaly segment, and TAD threshold, respectively. After applying PA, the predictions for informative and random anomaly scores degenerate to the same adjusted prediction (red). (b) Existing methods fail to exceed $F1_{PA}$ of a randomly generated anomaly score (left) and show no improvement against the newly proposed baseline (right) even when PA is forbidden for the WADI dataset.

2020; Shen, Li, and Kwok 2020; Su et al. 2019). PA works as follows: if at least one moment in a contiguous anomaly segment is detected as an anomaly, the entire segment is then considered to be correctly predicted as anomaly. Typically, F1 score is calculated with the adjusted predictions (hereinafter denoted by $F1_{PA}$). The F1 score computed without PAs denoted as $F1$. The PA protocol was proposed on the

basis that a single alert within an anomaly period is sufficient to take action for system recovery. It has become a fundamental step in TAD evaluation, and some of the following studies reported only $F1_{PA}$ without $F1$ (Chen et al. 2021). A higher $F1_{PA}$ has been indicated better detection capability.

However, PA has a high possibility of overestimating the model performance. A typical TAD model produces an anomaly score that informs about the degree of input abnormality, and predicts an anomaly if this score is higher than a threshold. With PA , the prediction from the randomly generated anomaly score and that of the well-trained model becomes identical, as depicted in Figure 1-(a). The black solid lines show two different anomaly scores; the upper line shows informative scores from a well-trained model, while the lower is randomly generated. The shaded area and dashed line indicate the ground truth (GT) anomaly segment and TAD threshold, respectively. The informative scores (above) are ideal given that they are high only during the GT segment. In contrast, randomly generated anomaly scores (below) cross the threshold only once within the GT segment. Despite their disparity, the predictions after PA become indistinguishable, as indicated by the red line. If random anomaly scores can yield $F1_{PA}$ as high as a proficient detection model, it is difficult to conclude that a model with a higher $F1_{PA}$ performs better than the others. Our experimental results in Section 5 show that random anomaly scores can overturn most state-of-the-art TAD methods (Figure 1-(b)).

Another question that arises is whether PA is the only problem in the evaluation of TAD methods. Until now, only the absolute $F1$ has been reported, without any attempt to establish a baseline and relative comparison against it. If the accuracy of a binary classifier is 50%, it is not much different from random guessing despite being an apparently large number. Similarly, a proper baseline should be discussed for TAD, and future methods should be evaluated based on the improvement compared to the baseline. According to our observations, existing TAD methods do not seem to have obtained a significant improvement over the baseline that this paper proposes. Furthermore, several methods fail to exceed it. Our observations for one of the benchmark dataset are summarized in the right of Figure 1-(b).

In this paper, we raise the question of whether the current TAD methods that claim to bring significant improvements are being properly evaluated, and suggest directions for the rigorous evaluation of TAD for the first time. Our contributions are summarized as follows:

- We show that PA , a peculiar evaluation protocol for TAD, greatly overestimates the detection performance of existing methods.
- We show that, without PA , existing methods exhibit no (or mostly insignificant) improvement over the baseline.
- Based on our findings, we propose a new baseline and an evaluation protocol for rigorous evaluation of TAD.

2 Background

2.1 Types of Anomaly in Time-series Signals

Various types of anomalies exist in TAD dataset (Choi et al. 2021). A *contextual anomaly* represents a signal that has a

different shape from that of the normal signal. A *collective anomaly* indicates a small amount of noise accumulated over a period of time. The *point anomaly* indicates a temporary and significant deviation from the expected range owing to a rapid increase or decrease in the signal value. Point anomaly is the most dominant type in the current TAD datasets.

2.2 Unsupervised TAD

A typical AD setting assumes that only normal data are accessible during the training time. Therefore, an unsupervised method is one of the most appropriate approaches for TAD, which trains a model to learn shared patterns in only normal signals. The final objective is to assign different anomaly scores to inputs depending on the degree of their abnormality, i.e., low and high anomaly scores for normal and abnormal inputs, respectively. Reconstruction-based AD method trains a model to minimize the distance between a normal input and its reconstruction. Anomalous input at the test time results in large distance as it is difficult to reconstruct. The distance, or reconstruction error, serves as an anomaly score. Forecasting-based approach trains a model to predict the signal that will come after the normal input, and takes the distance between the ground truth and predicted signal as an anomaly score. Please refer to Appendix for detailed examples of each category.

2.3 Assessment of TAD Evaluation

There have been several approaches to point out the pitfalls in current TAD evaluation. (Wu and Keogh 2021) proposed limitations of the benchmark TAD datasets and shows that a simple detector, so-called one-liner, is sufficient for some datasets. They also provided several synthetic datasets. (Lai et al. 2021) built a new taxonomy for anomaly types (e.g. point vs. pattern) and introduced new datasets synthesized under new criteria. In contrast, we propose the pitfalls in TAD evaluation: the risk of PA 's overestimation and the absence of baseline and the solutions. If the pitfalls are not resolved, it is impossible to evaluate whether the improvement of a TAD methods is significant even with the better datasets proposed by above papers.

3 Pitfalls of the TAD Evaluation

3.1 Problem Formulation

First, we denote the time-series signal observed from N sensors during time T as $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$, $\mathbf{x}_t \in \mathbb{R}^N$. As conventional approaches, it is normalized and split into a series of windows $\mathbf{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_{T-\tau+1}\}$ with stride 1, where $\mathbf{w}_t = \{\mathbf{x}_t, \dots, \mathbf{x}_{t+\tau-1}\}$ and τ is the window size. The ground truth binary label $y_t \in \{0, 1\}$, indicating whether a signal is an anomaly (1) or not (0), is given only for the test dataset. The goal of TAD is to predict the anomaly label \hat{y}_t for all windows in the test dataset. The labels are obtained by comparing anomaly scores $\mathcal{A}(\mathbf{w}_t)$ with a TAD threshold δ given as follows:

$$\hat{y}_t = \begin{cases} 1, & \text{if } \mathcal{A}(\mathbf{w}_t) > \delta \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

An example of $\mathcal{A}(\mathbf{w}_t)$ is the mean squared error (MSE) between the original input and its reconstructed version, which is defined as follows:

$$\mathcal{A}(\mathbf{w}_t) = \text{MSE}(\mathbf{w}_t, \hat{\mathbf{w}}_t) = \frac{1}{\tau} \|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_2 \quad (2)$$

where $\hat{\mathbf{w}}_t = f_\theta(\mathbf{w}_t)$ denotes the output from a reconstruction model f_θ parameterized with θ . After labeling, the precision (P), recall (R), and F1 score for the evaluation are computed as follows:

$$\begin{aligned} P &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad R = \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{F1 score} &= \frac{2 \cdot P \cdot R}{P + R}, \end{aligned} \quad (3)$$

where TP, FP, and FN denote the number of true positives, false positives and false negatives, respectively.

The TAD test dataset may contain multiple anomaly segments lasting over a few time steps. We denote \mathcal{S} as a set of M anomaly segments; that is, $\mathcal{S} = \{S_1, \dots, S_M\}$, where $S_m = \{t_s^m, \dots, t_e^m\}$; t_s^m and t_e^m denote the start and end times of S_m , respectively. PA adjusts \hat{y}_t to 1 for all $t \in S_m$ if anomaly score is higher than δ at least once in S_m . With PA, the labeling scheme of Eq. 1 changes as follows:

$$\hat{y}_t = \begin{cases} 1, & \text{if } \mathcal{A}(\mathbf{w}_t) > \delta \\ & \text{or } t \in S_m \text{ and } \exists_{t' \in S_m} \mathcal{A}(\mathbf{w}_{t'}) > \delta \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

F1_{PA} denotes the F1 score computed with adjusted labels.

3.2 Random Anomaly Score with High F1_{PA}

In this section, we demonstrate that the PA protocol overestimates the detection capability. We start from the abstract analysis of the P and R of Eq. 3, and we mathematically show that a randomly generated $\mathcal{A}(\mathbf{w}_t)$ can achieve a high F1_{PA} value close to 1. According to Eq. 3, as the F1 score is a harmonic mean of P and R, it also depends on TP, FN, and FP. As shown in Eq. 4, PA increases TP and decreases FN while maintaining FP. Therefore, after the PA, the P, R and consequently F1 score can only increase.

Next, we show that F1_{PA} can easily get close to 1. First, R is restated as a conditional probability as follows:

$$\begin{aligned} R &= \Pr(\hat{y}_t = 1 \mid y_t = 1) \\ &= \Pr(\hat{y}_t = 1 \mid t \in \mathcal{S}) \\ &= 1 - \Pr(\hat{y}_t = 0 \mid t \in \mathcal{S}). \end{aligned} \quad (5)$$

Let's assume that $\mathcal{A}(\mathbf{w}_t)$ is drawn from a uniform distribution $\mathcal{U}(0, 1)$. We use $0 \leq \delta' \leq 1$ to denote a TAD threshold for this assumption. If only one anomaly segment exists, i.e., $\mathcal{S} = \{\{t_s, \dots, t_e\}\}$, R after PA can be expressed as follows,

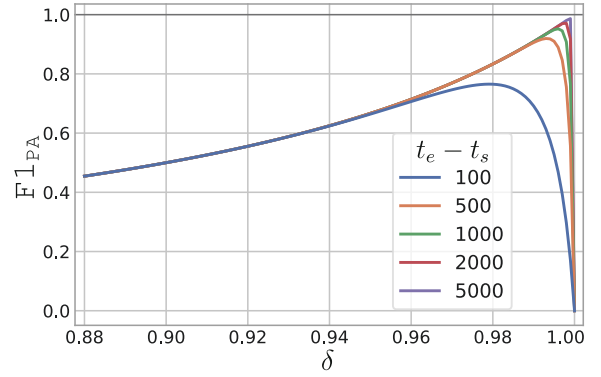


Figure 2: F1_{PA} for the case of uniform random anomaly scores varying with δ for different $t_e - t_s$. If an anomaly segment is considerably long, that is, if $t_e - t_s$ is sufficiently large, F1_{PA} approaches 1 as δ increases.

referring to Eq. 4:

$$\begin{aligned} R &= 1 - \Pr(\forall t' \in \mathcal{S}, \mathcal{A}(\mathbf{w}_{t'}) < \delta' \mid t \in \mathcal{S}) \\ &= 1 - \prod_{t' \in \mathcal{S}} \Pr(\mathcal{A}(\mathbf{w}_{t'}) < \delta' \mid t \in \mathcal{S}) \\ &= 1 - \frac{1}{\gamma} \cdot \prod_{t' \in \mathcal{S}} \Pr(\mathcal{A}(\mathbf{w}_{t'}) < \delta') \\ &= 1 - \frac{1}{\gamma} \cdot \delta'^{(t_e - t_s)}, \end{aligned} \quad (6)$$

where $\gamma = \Pr(t \in \mathcal{S})$ is a test dataset anomaly ratio and $\Pr(\mathcal{A}(\mathbf{w}_{t'}) < \delta') = \int_0^{\delta'} 1 = \delta'$.

$$\begin{aligned} P &= \Pr(y_t = 1 \mid \hat{y}_t = 1) = R \cdot \frac{\Pr(y_t = 1)}{\Pr(\hat{y}_t = 1)} \\ &= R \cdot \frac{\gamma}{\Pr(\hat{y}_t = 1, y_t = 1) + \Pr(\hat{y}_t = 1, y_t = 0)} \\ &= \frac{\gamma - \delta'^{(t_e - t_s)}}{(\gamma - \delta'^{(t_e - t_s)}) + (1 - \delta')}. \end{aligned} \quad (7)$$

For a more generalized proof, please refer to the Appendix. The anomaly ratio γ for a dataset is mostly given between 0 and 0.2; $t_e - t_s$ is also determined by the dataset and generally ranges from 100 to 5,000 in the benchmark datasets. Figure 2 depicts F1_{PA} varying with δ' under different $t_e - t_s$ when γ is fixed to 0.05. As shown in the figure, we can always obtain the F1_{PA} close to 1 by changing δ' , except for the case when the length of the anomaly segment is short.

3.3 Untrained Model with Comparably High F1

This section shows that the anomaly scores obtained from an untrained model are informative to a certain extent. A deep neural network is generally initialized with random weights drawn from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$, where σ is often much smaller than 1. Without training, the outputs

of the model are close to zero because they also follow a zero-mean Gaussian distribution. The anomaly score of a reconstruction-based or forecasting-based method is typically defined as the Euclidean distance between the input and output, which in the above case is proportional to the value of the input window:

$$\mathcal{A}(\mathbf{w}_t) = \|\mathbf{w}_t - \eta\|_2 \simeq \|\mathbf{w}_t\|_2, \quad (8)$$

where $\eta = f_\theta(\mathbf{w}_t)$ and $\theta \sim \mathcal{N}(0, \sigma^2)$.

In the case of a point anomaly, the specific sensor values increase abruptly. This leads to a larger magnitude of $\|\mathbf{w}_t\|_2$ than normal windows, which is connected directly to a high $\mathcal{A}(\mathbf{w}_t)$ for GT anomalies. The experimental results in Section 5 reveal that F1 calculated from $\mathcal{A}(\mathbf{w}_t)$ of Eq. 8 is comparable to that of current TAD methods. It is also shown that F1 increases even more when the window size gets longer.

4 Towards a Rigorous Evaluation of TAD

4.1 New Baseline for TAD

For a classification task, the baseline accuracy is often defined as that of a random guess. It can be said that there is an improvement only when the classification accuracy exceeds this baseline. Similarly, TAD needs to be compared not only with the existing methods but also with the baseline detection performance. Therefore, based on the findings of Section 3.3, we suggest establishing a new baseline with the F1 measured from the prediction of a randomly initialized reconstruction model with simple architecture, such as an untrained autoencoder comprising a single-layer LSTM. Alternatively, the anomaly score can be defined as the input itself, which is the extreme case of Eq. 8 when the model consistently outputs zero regardless of the input. If the performance of the new TAD model does not exceed this baseline, the effectiveness of the model should be reexamined.

4.2 New Evaluation Protocol PA%K

In the previous section, we demonstrated that PA has a great possibility of overestimating detection performance. F1 without PA can settle the overestimation immediately. In this case, it is recommended to set a baseline as introduced in Section 4.1. However, depending on the test data distribution, F1 can unexpectedly underestimate the detection capability. In fact, due to the incomplete test set labeling, some signals labeled as anomalies share more statistics with normal signals. Even if anomalies are inserted intermittently over a period of time, $y_t = 1$ for all t in that period.

We further investigated this problem using t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton 2008), as depicted in Figure 3. The t-SNE is generated by the test dataset of secure water treatment (SWaT) (Goh et al. 2016). Blue and orange colors indicate the normal and abnormal samples, respectively. The majority of the anomalies form a distinctive cluster located far from the normal data distribution. However, some abnormal windows are closer to the normal data than anomalies. The visualization of signals corresponding to the green and red points is depicted in (b) and (c), respectively. Although both

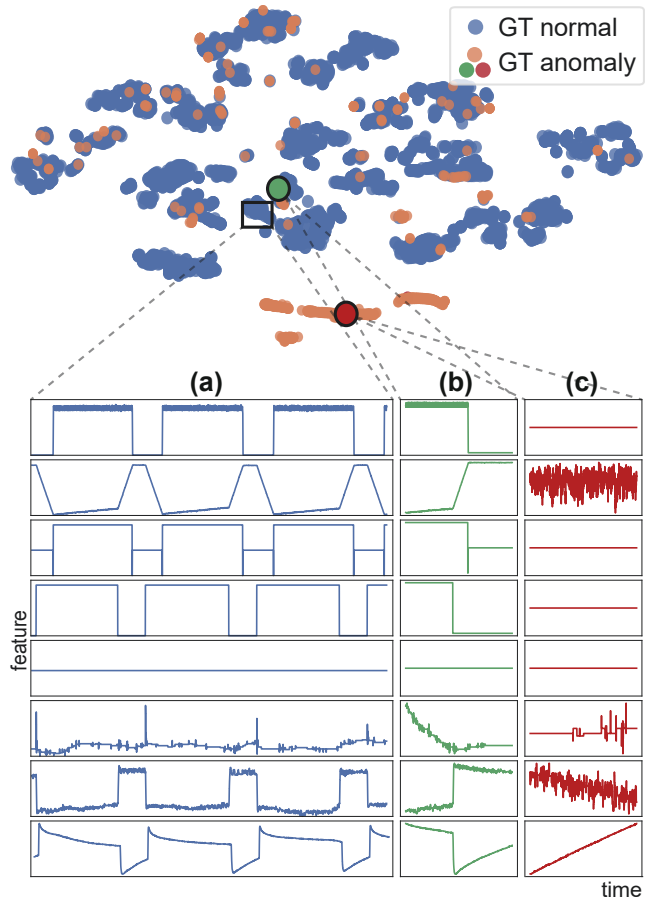


Figure 3: t-SNE of the input windows of the SWaT test dataset and visualization of corresponding signals. Blue color indicates ground truth (GT) normal while orange, green and red color indicate GT anomaly. Even though (b) is GT anomaly, it shares patterns more with (a), GT normal signal, than (c) abnormal signal.

samples were annotated as GT anomalies, (b) shared more patterns with normal data of (a) than (c). Concluding that the model’s performance is deficient only because it cannot detect signals such as (b) can lead to an underestimation of the detection capability.

Therefore, we propose an alternative evaluation protocol PA%K, which can mitigate the overestimation effect of $F1_{PA}$ and the possibility of underestimation of F1. Please note that it is not proposed to replace existing TAD metrics, but rather to be used along with them. The idea of PA%K is to apply PA to S_m only if the ratio of the number of correctly detected anomalies in S_m to its length exceeds the PA%K threshold K . PA%K modifies Eq. 4 as follows:

$$\hat{y}_t = \begin{cases} 1, & \text{if } \mathcal{A}(\mathbf{w}_t) > \delta \text{ or} \\ & t \in S_m \text{ and } \frac{|\{t' \mid t' \in S_m, \mathcal{A}(\mathbf{w}_{t'}) > \delta\}|}{|S_m|} > K \\ 0, & \text{otherwise} \end{cases}$$

where $|\cdot|$ denotes the size of S_m (i.e., $t_e^m - t_s^m$) and K can be

selected manually between 0 and 100 based on prior knowledge. For example, if the test set labels are reliable, a larger \mathbb{K} is allowable. If a user wants to remove the dependency on \mathbb{K} , it is recommended to measure the area under the curve of $F1_{PA\&\mathbb{K}}$ obtained by increasing \mathbb{K} from 0 to 100.

5 Experimental Results

5.1 Benchmark TAD Datasets

In this section, we introduce a list of the five most widely used TAD benchmark datasets as follows:

Secure water treatment (SWaT) (Goh et al. 2016): the SWaT dataset was collected over 11 days from a scaled-down water treatment testbed comprising 51 sensors (Mathur and Tippenhauer 2016). In the last 4 days, 41 anomalies were injected using diverse attack methods, while only normal data were generated during the first 7 days.

Water distribution testbed (WADI) (Ahmed, Palleti, and Mathur 2017): the WADI dataset was acquired from a reduced city water distribution system with 123 sensors and actuators operating for 16 days. The last two days contained anomalies. The test dataset included 15 anomaly segments.

Server Machine Dataset (SMD) (Su et al. 2019): the SMD dataset was collected from 28 server machines with 38 sensors for 10 days; only normal data appeared for the first 5 days, and anomalies were intermittently injected for the last 5 days. The reported results for the SMD dataset are the averaged values from 28 different models for each machine.

Mars Science Laboratory (MSL) and **Soil Moisture Active Passive (SMAP)** (Hundman et al. 2018): the MSL and SMAP dataset is a real-world dataset collected from a spacecraft of NASA. These are the anomaly data from an incident surprise anomaly (ISA) report for a spacecraft monitoring system. Unlike other datasets, unlabeled anomalies are contained in the training data, which makes training difficult. The statistics are summarized in Table 1.

5.2 Evaluated Methods

Below, we present the descriptions of 7 representative TAD methods recently proposed and the 3 cases investigated in Section 3.

USAD (Audibert et al. 2020) stands for unsupervised anomaly detection, which trains two autoencoders consisting of one shared encoder and two separate decoders, under a two-phase training scheme: an autoencoder training phase and an adversarial training phase.

DAGMM (Zong et al. 2018) represents deep autoencoding Gaussian mixture model that adopts an autoencoder to yield a representation vector and feed it to a Gaussian mixture model. It uses the estimated sample energy as a reconstruction error; high energy indicates high abnormality.

LSTM-VAE (Park, Hoshi, and Kemp 2018) represents an LSTM-based variational autoencoder that adopts variational inference for reconstruction.

OmniAnomaly (Su et al. 2019) applied a VAE to model the time-series signal into a stochastic representation, which would predict an anomaly if the reconstruction likelihood of a given input is lower than a threshold value. It also defined

| Dataset | Train | Test (anomaly%) | N |
|-------------|---------|------------------|-----|
| SWaT | 495,000 | 449,919 (12.33%) | 51 |
| WADI | 784,537 | 172,801 (5.77%) | 123 |
| SMD | 25,300 | 25,300 (4.21%) | 38 |
| MSL | 58,317 | 73,729 (10.5%) | 55 |
| SMAP | 135,183 | 427,617 (12.8%) | 25 |

Table 1: Statistics of benchmark TAD datasets. N denotes the dimension of input features.

the reconstruction probabilities of individual features as attribution scores and quantified their interpretability.

MSCRED (Zhang et al. 2019) represents a multi-scale convolutional recurrent encoder-decoder comprising convolutional LSTMs to reconstruct the input matrices that characterize multiple system levels, rather than the input itself.

THOC (Shen, Li, and Kwok 2020) represents a temporal hierarchical one-class network, which is a multi-layer dilated recurrent neural network and a hierarchical deep support vector data description.

GDN (Deng and Hooi 2021) represents a graph deviation network that learns a sensor relationship graph to detect deviations of anomalies from the learned pattern.

Case 1. Random anomaly score corresponds to the case described in Section 3.2. The F1 score is measured with a randomly generated anomaly score drawn from a uniform distribution \mathcal{U} , i.e., $\mathcal{A}(w_t) \sim \mathcal{U}(0, 1)$.

Case 2. Input itself as an anomaly score denotes the case assuming $f_\theta(w_t) = 0$ regardless of w_t . This is equal to an extreme case of Eq. 8. Therefore, $\mathcal{A}(w_t) = \|w_t\|_2$.

Case 3. Anomaly score from the randomized model corresponds to Eq. 8, where η denotes a small output from a randomized model. The parameters were fixed after being initialized from a Gaussian distribution $\mathcal{N}(0, 0.02)$.

5.3 Correlation between $F1_{PA}$ and $F1$

$F1$ is the most conservative indicator of detection performance. Therefore, if $F1_{PA}$ reliably represents the detection capability, it should have at least some correlation with $F1$. Figure 4 plots $F1_{PA}$ and $F1$ for SWaT and WADI, as reported by the original studies on USAD, DAGMM, LSTM-VAE, OmniAnomaly, and GDN. The figure also includes the results of **Case 1–3**. It is noteworthy that given that only a subset of the datasets and methods reported $F1_{PA}$ and $F1$ together, we plotted them only. For SWaT, the Pearson correlation coefficient (PCC) and Kendall rank correlation coefficient (KRC) were -0.59 and 0.07, respectively. For WADI, the PCC and KRC were 0.41 and 0.43, respectively. However, these numbers are insufficient to assure the existence of correlation and confirm that comparing the superiority of the methods using only $F1_{PA}$ may have the risk of improper evaluation of the detection performance.

5.4 Comparison Results

Here, we compare the results of the AD methods with **Case 1–3**. It should be noted that the anomaly score is directly generated without model inference for **Case 1** and **2**. For

| | SWaT | | WADI | | MSL | | SMAP | | SMD | |
|---------------|--------------|--------------------|--------------|--------------------|--------------|--------------------|--------------------|-------------------------------|--------------------|-------------------------------|
| | $F1_{PA}$ | F1 | $F1_{PA}$ | F1 | $F1_{PA}$ | F1 | $F1_{PA}$ | F1 | $F1_{PA}$ | F1 |
| USAD | 0.846 | <u>0.791</u> (†) | 0.429 | 0.232 | <u>0.927</u> | 0.211 [†] | <u>0.818</u> | 0.228 [†] | <u>0.938</u> (†) | 0.426 [†] |
| DAGMM | 0.853 | 0.550 | 0.209 | 0.121 | 0.701 | 0.199 [†] | 0.712 | 0.333 [†] (†) | 0.723 | 0.238 [†] |
| LSTM-VAE | 0.805 | 0.775 | 0.380 | 0.227 | 0.678 | 0.212 [†] | 0.756 | 0.235 [†] (†) | 0.808 (†) | 0.435 [†] |
| OmniAnomaly | 0.866 | 0.782 | 0.417 | 0.223 | 0.899 | 0.207 [†] | 0.805 | 0.227 [†] | 0.944 (†) | 0.474 [†] |
| MSCRED | 0.868 | 0.662 [†] | 0.346 | 0.087 [†] | 0.775 | 0.199 [†] | 0.942 [†] | 0.232 [†] (†) | 0.389 [†] | 0.097 [†] |
| THOC | 0.880 | 0.612 [†] | 0.506 | 0.130 [†] | 0.891 | 0.190 [†] | 0.781 [†] | 0.240 [†] (†) | 0.541 [†] | 0.168 [†] |
| GDN | <u>0.935</u> | 0.81 (†) | <u>0.855</u> | 0.57 (†) | 0.903 | 0.217 [†] | 0.708 [†] | <u>0.252</u> [†] (†) | 0.716 [†] | 0.529 [†] (†) |
| Case 1 | 0.969 | 0.216 | 0.965 | 0.109 | 0.931 | 0.190 | 0.961 | 0.227 | 0.804 | 0.080 |
| Case 2 | 0.873 | 0.781 | 0.694 | <u>0.353</u> | 0.812 | <u>0.239</u> | 0.675 | 0.229 | 0.896 | <u>0.494</u> |
| Case 3 | 0.869 | 0.789 | 0.695 | 0.331 | 0.427 | 0.236 | 0.699 | 0.229 | 0.893 | 0.466 |

Table 2: F1 score for various methods. † indicates the reproduced results. Bottom three rows represent the followings: Case 1. Random anomaly score, Case 2. Input itself as a anomaly score, Case 3. Anomaly score from a randomized model. Please refer to the manuscript for the detailed explanations. Bold and underlined cases indicate the best and the second best, respectively. † is marked in the following cases: (1) $F1_{PA}$ is higher than Case 1, (2) F1 is higher than Case 2 or 3.

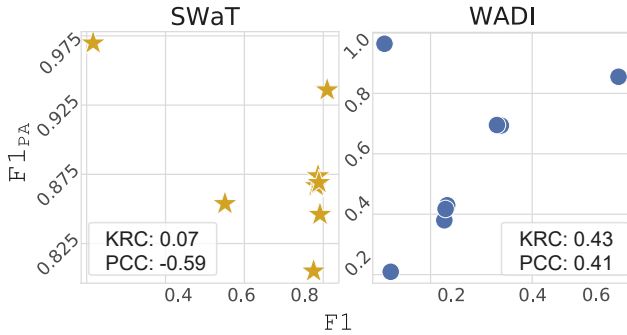


Figure 4: Correlation between $F1_{PA}$ and F1 of the existing methods on SWaT and WADI dataset. The Kendall rank correlation (KRC) and Pearson correlation coefficient (PCC) are indicated in the figure.

Case 3, we adopted the simplest encoder-decoder architecture with LSTM layers. The window size τ for **Case 2** and **3** was set to 120. For experiments that included randomness, such as **Case 1** and **3**, we repeated them with five different seeds and reported the average values. For the existing methods, we used the best numbers reported in the original papers and officially reproduced results (Choi et al. 2021); if there were no available scores, we reproduced them referring to the officially provided codes. Please note that we did not apply any preprocessing such as early time steps removal or downsampling. The F1 for MSL, SMAP, and SMD have not been provided in previous papers; thus they are all reproduced. It is worth noting that we searched for optimal hyperparameters within the suggested range in the papers and we did not apply down-sampling. All thresholds were obtained from those that yielded the best score. Further details of the implementation are provided in the Appendix. The results are shown in Table 2. The reproduced results are marked as †. Bold and underlined numbers indicate the best and second-best results, respectively. The up

arrow (†) is displayed with the result for the following cases: (1) $F1_{PA}$ is higher than **Case 1**, (2) F1 is higher than **Case 2** or **3**, whichever is greater.

Clearly, the randomly generated anomaly score (**Case 1**) is not able to detect anomalies because it reflects nothing about the abnormality in an input. Correspondingly, F1 was quite low, which clearly revealed a deficient detection capability. However, when applying the PA protocol, **Case 1** appears to yield the state-of-the-art $F1_{PA}$ far beyond the existing methods, except for SMD. If the result is provided only with PA, as in the case of the MSL, SMAP, and SMD, distinguishing whether the method successfully detects anomalies or whether it merely outputs a random anomaly score irrelevant to the input is impossible. In particular, F1 of the MSL and SMAP is quite low; this implies difficulty in modeling them, originating from the fact that they are both real-world datasets, and the training data contain anomalies. However, $F1_{PA}$ appears considerably high, creating an illusion that the anomalies are being detected well for those datasets.

The F1 of **Case 1** of SMD is lower than that in other datasets, and there are previous methods surpassing it. This may be attributed to the composition of the SMD test dataset. According to Eqs. 6 and 7, $F1_{PA}$ varies with three parameters: the ratio of anomalies in the test dataset (γ), length of anomaly segments ($t_e - t_s$), and TAD threshold (δ). Unlike the other datasets, the anomaly ratio of SMD was quite low, as shown in Table 1. Moreover, the lengths of the anomaly segments are relatively short; the average length of 28 machines is 90, unlike other datasets ranging from hundreds to thousands. This is similar to the lowest case in Figure 2, which shows that the maximum achievable $F1_{PA}$ in this case is only approximately 0.8. Therefore, we can conclude that the overestimation effect of PA depends on the test dataset distribution, and its effect becomes less conspicuous with shorter anomaly segments.

Across all datasets, the F1 for the existing methods is mostly inferior to **Case 2** and **3**, implying that the currently proposed methods may have obtained marginal or even no

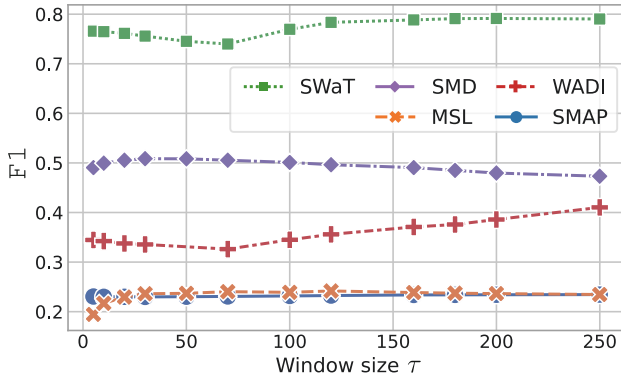


Figure 5: F1 for various window sizes (τ). As τ increases, F1 mostly increases after a short decrease.

advancement against the baselines. Only the GDN consistently exceeded the baselines for all datasets. The F1 of **Case 2** and **3** depend on the length of the input window. With a longer window, the F1 baseline becomes even larger. We experimented with various window lengths ranging from 1 to 250 in **Case 2** and depicted the results in Figure 5. For SWaT, WADI, and SMAP, F1 begins to increase after a short decrease as τ increases. This increase occurs because a longer window is more likely to contain more point anomalies, resulting in high anomaly score for the window. If τ becomes too large, F1 saturates or degrades, possibly because the windows that used to contain only normal signals unexpectedly contain anomalies in it.

5.5 Effect of $\text{PA}\%K$ Protocol

To examine how $\text{PA}\%K$ alleviates the overestimation effect of PA and underestimation tendency of F1, we observed $F1_{\text{PA}\%K}$ varying with different $\text{PA}\%K$ thresholds K . Figure 6 shows the $F1_{\text{PA}\%K}$ for SWaT from **Case 1** and the fully trained encoder-decoder when K changes in increments of 10 from 0 to 100. The $F1_{\text{PA}\%K}$ values of $K = 0$ and $K = 100$ are equal to the original $F1_{\text{PA}}$ and F1, respectively. The $F1_{\text{PA}\%K}$ of a well-trained model is expected to show constant results regardless of the value of K . Correspondingly, the $F1_{\text{PA}\%K}$ of the trained encoder-decoder (orange) shows consistently high $F1_{\text{PA}\%K}$. In contrast, the $F1_{\text{PA}\%K}$ of **Case 1** (blue) rapidly decreased when K increased. We also proposed measuring the area under the curve (AUC) to reduce the dependence on K . In this case, the AUC were 0.88 and 0.41 for the trained encoder-decoder and **Case 1**, respectively; this demonstrates that $\text{PA}\%K$ clearly distinguishes the former from the latter regardless of K .

6 Discussion

Throughout this paper, we have demonstrated that the current evaluation of TAD has pitfalls in two respects: (1) since PA overestimate the detection performance, we cannot ensure that a method with higher $F1_{\text{PA}}$ has indeed a better detection capability; (2) the results have been compared only with existing methods, not against the baseline. A better anomaly detector can be developed when the current

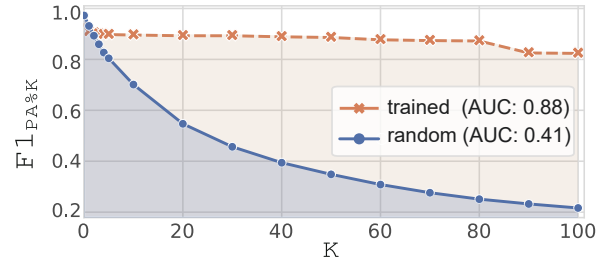


Figure 6: F1 score with $\text{PA}\%K$ with varying K . If $K = 0$, it is equal to the $F1_{\text{PA}}$ and if $K = 100$, it is equal to the F1.

achievements are properly assessed. In this section, we suggest several directions for future TAD evaluations.

The motivation of PA, i.e., the source of the first pitfall, originates from the incompleteness of the test dataset labeling process, as observed in Section 4.2. An exterminatory solution is to develop a new benchmark dataset annotated in a more fine-grained manner, so that the time-step-wise labels become reliable. As it is often not feasible because fine-grained annotation requires tremendous resources, $F1_{\text{PA}\%K}$ can be a good alternative that can alleviate overestimation without any additional dataset modification. Please note that $\text{PA}\%K$ is an evaluation protocol that can be applied to various metrics other than the F1 score. For the second issue, it is important to set a baseline as the performance of the untrained model as **Case 2** and **3** and measure the relative improvement against it. The window size should be carefully determined by considering its effect on the baselines, as described in Section 5.4.

Furthermore, pre-defining the TAD threshold without any access to the test dataset is often impractical in the real world. Correspondingly, many AD methods in the vision field evaluate themselves using the area under the receiver operating characteristic (AUROC) curve (Yi and Yoon 2020). In contrast, existing TAD methods set the threshold after investigating the test dataset or simply use the optimal threshold that yields the best F1. Thus, the detection result depends significantly on threshold selection. Additional metrics with the reduced dependency such as AUROC or area under precision-recall (AUPR) curve will help in rigorous evaluation. Even in this case, the proposed baseline selection method is valid. Since $\text{PA}\%K$ is a protocol, it also can be used to above metrics.

7 Conclusion

In this paper, we showed for the first time that applying PA can severely overestimate a TAD model's capability, which may not reflect the true modeling performance. We also proposed a new baseline for TAD and showed that only a few methods have achieved significant advancement in this regard. To mitigate overestimation of PA, we proposed a new $\text{PA}\%K$ protocol that can be applied with existing metrics. Finally, we suggest several directions for rigorous evaluation of TAD methods, including baseline selection. We expect that our research help clarify the potential of current TAD methods and lead the improvement of TAD in the future.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [No.2021-0-02068, Artificial Intelligence Innovation Hub, NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)], the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) [2018R1A2B3001628, 2019R1G1A1003253], the BK FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2021, AIR Lab (AI Research Lab) in Hyundai Motor Company through HMC-SNU AI Consortium Fund, and Samsung Electronics.

References

- Ahmed, C. M.; Palleti, V. R.; and Mathur, A. P. 2017. WADI: A Water Distribution Testbed for Research in the Design of Secure Cyber Physical Systems. In *Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks*, CySWATER '17, 25–28. New York, NY, USA: Association for Computing Machinery. ISBN 9781450349758.
- Audibert, J.; Michiardi, P.; Guyard, F.; Marti, S.; and Zuluaga, M. A. 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.
- Baheti, R.; and Gill, H. 2011. Cyber-physical systems. *The impact of control technology*, 12(1): 161–166.
- Chalapathy, R.; and Chawla, S. 2019. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*.
- Chen, Z.; Chen, D.; Zhang, X.; Yuan, Z.; and Cheng, X. 2021. Learning Graph Structures with Transformer for Multivariate Time Series Anomaly Detection in IoT. *IEEE Internet of Things Journal*.
- Choi, K.; Yi, J.; Park, C.; and Yoon, S. 2021. Deep Learning for Anomaly Detection in Time-Series Data: Review, Analysis, and Guidelines. *IEEE Access*.
- Deng, A.; and Hooi, B. 2021. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4027–4035.
- Goh, J.; Adepu, S.; Junejo, K. N.; and Mathur, A. 2016. A dataset to support research in the design of secure water treatment systems. In *International conference on critical information infrastructures security*, 88–99. Springer.
- Hundman, K.; Constantinou, V.; Laporte, C.; Colwell, I.; and Soderstrom, T. 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.
- Lai, K.-H.; Zha, D.; Xu, J.; Zhao, Y.; Wang, G.; and Hu, X. 2021. Revisiting time series outlier detection: Definitions and benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.
- Lee, E. A. 2008. Cyber physical systems: Design challenges. In *2008 11th IEEE international symposium on object and component-oriented real-time distributed computing (ISORC)*, 363–369. IEEE.
- Lee, J.; Bagheri, B.; and Kao, H.-A. 2015. A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manufacturing letters*, 3: 18–23.
- Mathur, A. P.; and Tippenhauer, N. O. 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)*, 31–36. IEEE.
- Park, D.; Hoshi, Y.; and Kemp, C. C. 2018. A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder. *IEEE Robotics and Automation Letters*, 3(3): 1544–1551.
- Shen, L.; Li, Z.; and Kwok, J. 2020. Timeseries anomaly detection using temporal hierarchical one-class network. *Advances in Neural Information Processing Systems*, 33: 13016–13026.
- Su, Y.; Zhao, Y.; Niu, C.; Liu, R.; Sun, W.; and Pei, D. 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.
- Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).
- Wu, R.; and Keogh, E. 2021. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE Transactions on Knowledge and Data Engineering*.
- Xu, H.; Feng, Y.; Chen, J.; Wang, Z.; Qiao, H.; Chen, W.; Zhao, N.; Li, Z.; Bu, J.; Li, Z.; and et al. 2018. Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*.
- Yi, J.; and Yoon, S. 2020. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the Asian Conference on Computer Vision*.
- Zhang, C.; Song, D.; Chen, Y.; Feng, X.; Lumezanu, C.; Cheng, W.; Ni, J.; Zong, B.; Chen, H.; and Chawla, N. V. 2019. A Deep Neural Network for Unsupervised Anomaly Detection and Diagnosis in Multivariate Time Series Data. In *AAAI*.
- Zong, B.; Song, Q.; Min, M. R.; Cheng, W.; Lumezanu, C.; Cho, D.; and Chen, H. 2018. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*.