When Model Meets New Normals: Test-Time Adaptation for Unsupervised Time-Series Anomaly Detection

Dongmin Kim, Sunghyun Park, Jaegul Choo

KAIST

tommy.dm.kim@kaist.ac.kr, psh01087@kaist.ac.kr, jchoo@kaist.ac.kr

Abstract

Time-series anomaly detection deals with the problem of detecting anomalous timesteps by learning normality from the sequence of observations. However, the concept of normality evolves over time, leading to a "new normal problem", where the distribution of normality can be changed due to the distribution shifts between training and test data. This paper highlights the prevalence of the new normal problem in unsupervised time-series anomaly detection studies. To tackle this issue, we propose a simple yet effective test-time adaptation strategy based on trend estimation and a self-supervised approach to learning new normalities during inference. Extensive experiments on real-world benchmarks demonstrate that incorporating the proposed strategy into the anomaly detector consistently improves the model's performance compared to the baselines, leading to robustness to the distribution shifts.

Introduction

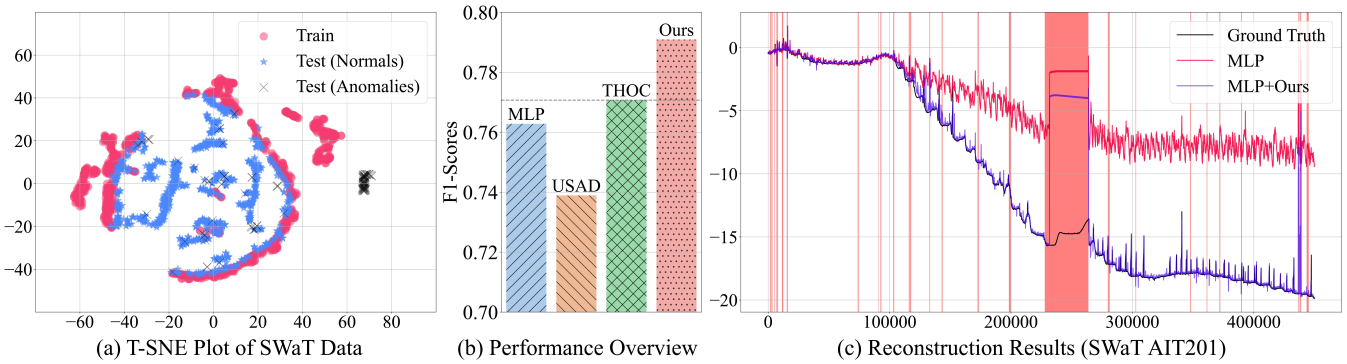
In real-world monitoring systems, the continuous operation of numerous sensors generates substantial real-time measurements. Time-series anomaly detection aims to identify observations that deviate from the concept of normality (Ruff et al. 2021; Pang et al. 2022) within a sequence of observations. Examples of anomalous events include physical attacks on industrial systems (Mathur and Tippenhauer 2016; Han et al. 2021), unpredictable robot behavior (Park, Hoshi, and Kemp 2018), faulty sensors from wide-sensor networks (Wang, Kuang, and Duan 2015; Rassam, Maarof, and Zainal 2018), cybersecurity attacks (Su et al. 2019; Abdulaal, Liu, and Lancewicki 2021), and spacecraft malfunctions (Hundman et al. 2018; Shin et al. 2020; Liu, Liu, and Peng 2016).

However, detecting abnormal timesteps presents significant challenges due to several factors. Firstly, the complex nature of system dynamics, characterized by the coordination of multiple sensors, complicates the task. Secondly, the increasing volume of incoming signals to monitoring systems adds to the difficulty. Lastly, acquiring labels for abnormal behaviors is problematic. To address these challenges, unsupervised time-series anomaly detection models (Xu et al. 2022; Audibert et al. 2020; Su et al. 2019; Park, Hoshi, and Kemp 2018; Malhotra et al. 2016) have emerged, focusing on learning normal patterns from available training datasets.

Nevertheless, the concept of normality can change over time, widely known as a distribution shift (Quinonero-Candela et al. 2008; Kim et al. 2022b; Sun et al. 2020; Gulrajani and Lopez-Paz 2021; Wang et al. 2021, 2022), as can be seen in the Fig. 1-(a). We have observed that off-the-shelf models are susceptible to such shifts, leading to a "new normal problem", where the distribution of normality during test time cannot be fully characterized solely based on training data. Without consideration of distribution shifts, these models tend to rely on past observations and generate false alarms, compromising the consistency of monitoring systems (Dragoi et al. 2022; Cao, Zhu, and Pang 2023).

Recently, test-time adaptation mechanisms (Wang et al. 2021, 2022; Niu et al. 2022) have been proposed to adapt models for alleviating performance degradation due to distribution shifts between training and test datasets, especially in the computer vision field. Test-time adaptation methods update the model parameters to generalize to different data distributions, without relying on either additional supervision from labels or access to training data. Time-series anomaly detection task also shares motivation for applying test-time adaptation strategies; frequent access to past data for adaptation is costly as monitoring systems work in real-time (Abdulaal, Liu, and Lancewicki 2021; Shin et al. 2020; Su et al. 2019) and model update without supervision is desired as acquiring labels is often limited (Geiger et al. 2020; Ruff et al. 2021; Audibert et al. 2020). Motivated by these advancements, we propose a test-time adaptation for unsupervised time-series anomaly detection under distribution shifts.

Our paper highlights the prevalence of the new normal problem in time-series anomaly detection literature. To address this issue properly, we propose a simple yet effective adaptation strategy using trend estimates and model updates with normal instances based on the model's prediction itself. Trend estimate, given by the exponential moving average of the observations, follows the expected value of a time-series with adaptation to changing conditions (Muth 1960) with computational efficiency. After model deployment, we update the model parameters with the normalized input sequence, which is detrended by subtracting the trend estimate, to learn complicated dynamics that cannot be captured solely on the trend estimate. Our proposed method makes the model robust to such distribution shifts, thereby increasing detector performance, as shown in Fig. 1-(b) and Fig. 1-(c).



(a) T-SNE Plot of SWaT Data (b) Performance Overview (c) Reconstruction Results (SWaT AIT201)

Figure 1: Motivation for learning new normals. (a) T-SNE visualization of the SWaT benchmark (Mathur and Tippenhauer 2016) reveals distinct behavior between the training (red) and test data (blue). (b) Our test-time adaptation strategy surpasses previous state-of-the-art time-series anomaly detection models in terms of F1 score, even with simple baselines such as MLP-based autoencoders. (c) This improvement arises from effectively handling significant distribution shifts in the time-series data. Over time, off-the-shelf models fail to adapt to these new normals, while our approach exhibits robustness to such distribution shifts. Consequently, previous approaches (Audibert et al. 2020; Shen, Li, and Kwok 2020) produce false positive cases due to the model's inability to keep pace with changing dynamics, thereby *"the model is staying in the past while the world is changing."*

Our contributions can be summarized as follows:

- We discover that new normal problems pose a significant challenge in modeling unsupervised time-series anomaly detection under distribution shifts.
- We propose a simple yet effective adaptation strategy following the trend estimate of the time-series data and update the model parameters using a detrended sequence to address these problems.
- Through extensive experiments on various real-world datasets, our method consistently improves the model's performance when facing a severe distribution shift problem between training data and test data.

Related Works

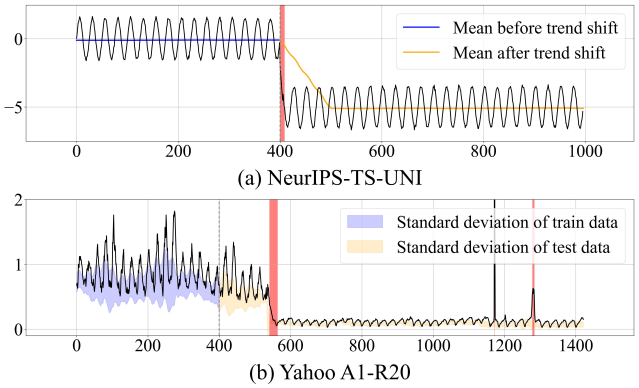
Unsupervised time-series anomaly detection. Unsupervised time-series anomaly detection (Su et al. 2019; Audibert et al. 2020; Xu et al. 2022) aims to detect observations that deviate considerably from normality, assuming the non-existence of the available labels. To the extent of conventional anomaly detection approaches (Breunig et al. 2000; Schölkopf et al. 1999; Tax and Duin 1999) and deep-learning-based anomaly detection approaches (Zong et al. 2018; Ruff et al. 2018), unsupervised time-series anomaly detection models aim to build an architecture that can model the temporal dynamics of the sequence.

The main categories of unsupervised time-series anomaly detection models include reconstruction-based models, clustering-based models, and forecasting-based models. Building upon the assumption of better reconstruction performance of normal instances compared to anomalous instances, reconstruction-based models encompass a range of approaches involving LSTM (Malhotra et al. 2016; Park, Hoshi, and Kemp 2018; Su et al. 2019) and MLP (Audibert et al. 2020) architectures, as well as the integration of GANs (Schlegl et al. 2017; Geiger et al. 2020; Han et al. 2021). Clustering-based methods include the extension of one-class support vector machine approaches (Schölkopf et al. 1999; Tax and Duin 1999), tensor decomposition-based clustering methods for the detection of anomalies (Shin et al. 2020), and the utilization of latent representations for clustering (Ruff et al. 2018; Shen, Li, and Kwok 2020). Forecasting-based methods rely on detecting anomalies by identifying substantial deviations between past sequences and ground truth labels, as exemplified by the use of ARIMA (Pena, de Assis, and Jr. 2013), LSTM (Hundman et al. 2018), and transformer (Xu et al. 2022).

Distribution shift in time-series data. Due to the nature of continually changing temporal dynamics, mitigating distribution shifts emerges as a pivotal concern within the time-series data analysis, notably within tasks such as time-series forecasting (Kim et al. 2022b; Liu et al. 2022) and anomaly detection (Sankararaman et al. 2022; Dragoi et al. 2022).

Online RNN-AD (Saurav et al. 2018) adapts to concept drift with RNN architectures, which update the model with backpropagation of anomaly scores using all stream data. Our work differentiates from this work by introducing detrending modules for model updates and selective learning of a set of normal instances in a self-supervised way. Although recent work (Sankararaman et al. 2022) also presents an adaptable framework for anomaly detection, it hinges on a dynamic window mechanism applied to historical data streams. Notably, our approach diverges from their assumption of accessibility of past sequences; we keep model parameters at hand, process input sequences immediately, and evict after.

Test-time adaptation. To alleviate the performance degradation caused by distribution shift, unsupervised domain adaptation (Ganin et al. 2016; Zou et al. 2018; Yoo, Chung, and Kwak 2022; Liang, Hu, and Feng 2020) methods have been developed in various fields. These methods align with our work from the perspective of addressing the covariate shift problem. In recent times, fully test-time adaptation (TTA) (Wang et al. 2021) methods have emerged to enhance the model performance on test data through real-time adaptation using unlabeled test samples during inference, without



(a) NeurIPS-TS-UNI

(b) Yahoo A1-R20

Figure 2: Illustration of the necessity for estimating trends and test-time adaptation. (a) NeurIPS-TS-UNI shows synthetic data generated based on the previous work (Lai et al. 2021), revealing an abrupt trend shift while preserving underlying dynamics. The objective of the trend estimation module is to adapt to such trend shifts successfully. (b) Solely relying on trend estimation may not be adequate to fully capture the dynamics, as demonstrated by the Yahoo-A1-R20 series. The shaded purple and yellow areas represent the standard deviations of the train and test data, respectively. To model this shift in dynamics, which cannot be fully captured alone with trend estimates, it is necessary to learn distribution shifts through test-time model updates outlined directly.

relying on access to the training data. Most TTA approaches employ entropy minimization (Wang et al. 2021; Niu et al. 2022; Choi et al. 2022) or pseudo labels (Wang et al. 2022) to update the model parameters using unlabeled test samples. However, simply adopting previous TTA methods may not be directly applicable to unsupervised time-series anomaly detection. This is due to the vulnerability of the model when updating the model using all test samples, as abnormal test samples have the potential to disrupt its functionality. Consequently, this work aims to successfully apply the concept of test-time adaptation to the unsupervised time-series anomaly detection task.

Method

Problem Statement

Unsupervised time-series anomaly detection aims to detect anomalous timesteps during test time without explicit supervision by learning the concept of normality. The concept of normality is defined as the probability distribution $\mathbb{P}$ on data $\mathcal{D}$ that is the ground-truth law of normal behavior in a given task (Ruff et al. 2021). Accordingly, a set of anomalies is defined as data with sufficiently small probability under such distribution, *i.e.*, $p(x) < \epsilon$. New normal problem that we tackle can be formulated as the phenomena of underlying distribution $\mathbb{P}$ is not stationary, *i.e.*, $\mathbb{P}_{train} \neq \mathbb{P}_{test}$.

For observations over N timesteps with F features, time-series data is specified by a sequence $\mathcal{D} = \{X_1, X_2, ..., X_N\}$, where each $X_i \in \mathbb{R}^F$. An anomaly detector aims to map each observation to a class label $y = \{0, 1\}$, where $y = 0$ and $y = 1$ each denote normal and abnormal timesteps. The detector is specified by an anomaly score function $\mathcal{A} : \mathbb{R}^F \rightarrow \mathbb{R}$, along with a decision threshold τ. Concretely, observation X_t is classified as anomalous if $\mathcal{A}(X_t) > \tau$. We denote the set of train-time instances as $\mathcal{D}_{train}$ and the set of test-time instances as $\mathcal{D}_{test}$. Accordingly, test-time normals and anomalies can be defined each as $\{X \in \mathcal{D}_{test} \mid y = 0\}$ and $\{X \in \mathcal{D}_{test} \mid y = 1\}$.

To reflect the temporal context of time-series data to detect anomalous timestep(s), a set of observations $\mathcal{D}$ is preprocessed with a sliding window setting. Specifically, we denote a sequence of w observations until timestep t as $\mathcal{X}_{w,t} = [X_{t-w+1}, X_{t-w+2}, ..., X_{t-1}, X_t]$ and its corresponding class label and prediction of the model as $\mathcal{Y}_{w,t} = [y_{t-w+1}, y_{t-w+2}, ..., y_{t-1}, y_t]$ and $\hat{\mathcal{Y}}_{w,t} = [\hat{y}_{t-w+1}, \hat{y}_{t-w+2}, ..., \hat{y}_{t-1}, \hat{y}_t]$ following conventional approaches of the time-series anomaly detection literatures (Shen, Li, and Kwok 2020; Su et al. 2019).

Input Normalization Using Trend Estimate

A trend estimation module aims to adapt to new normals that significantly differ in trend with preserving the underlying dynamics of the sequence. Accordingly, previous work (Lai et al. 2021) defines trend-outlier as:

$$\Delta(\mathcal{T}(\cdot), \tilde{\mathcal{T}}(\cdot)) > \delta, \tag{1}$$

where Δ is a function that measures the discrepancy between two functions. $\tilde{\mathcal{T}}$ is a function that returns the trend of normal sequences. $\mathcal{T}$ is a trend of an arbitrary sequence to compare to the trend of normal sequences. Fig. 2-(a) illustrates the importance of properly estimating the trend of normalities. Even though sequences before and after the transition shares the same dynamics, observations after the trend shift are classified as anomalies without proper adaptation to trends. To address such a problem, we simply detrend with trend estimates using the exponential moving average statistics. Technically, we estimate the trend as:

$$\tilde{\mathcal{T}}(\cdot) : \mu_t \leftarrow \gamma \mu_{t-w} + (1 - \gamma)\hat{\mu}, \tag{2}$$

where $\hat{\mu} = \frac{1}{w}\Sigma_{i=t-w+1}^{t} X_i$, which is the empirical mean of the stream data, and γ is a hyperparameter that controls an exponentially moving average update rate for tracking the trend of the data stream. This procedure is one form of eliminating nonstationary trend components with mean adjustment (Shumway and Stoffer 2017), allowing models to be updated with numerical stability. Concretely, as shown in Fig. 3, along with reconstruction-based anomaly detection models, the model reconstructs detrended sequence $\mathcal{X}_{w,t} - \mu_t$ instead of $\mathcal{X}_{w,t}$ and denormalize the reconstructed sequence by adding estimated trend for the final output.

Model Update with New Normals

Test-time adaptation with model update aims to learn the underlying dynamics of the time series data, which cannot be fully captured by trend estimation alone, as shown in Fig. 2-(b). Specifically, our approach continuously updates the model parameters with normal sequences during test time in a fully unsupervised manner. Formally, the normal

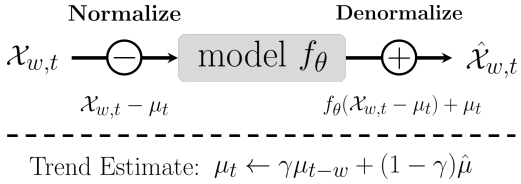


Figure 3: Illustration on detrend module.

instances during test-time observations can be formulated as $\{X \in \mathcal{D}_{test} \mid y = 0\}$. To update the model parameters θ during test-time, the prediction of the model itself, $\hat{\mathcal{Y}}$ acts as selection criteria for filtering normal timesteps. The model is updated based on online gradient descent (Zinkevich 2003) using the following scheme:

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\mathcal{X}_{w,t}, \hat{\mathcal{Y}}_{w,t}, \mu_t, \tau), \tag{3}$$

where η is the test-time learning rate for the model update. τ denotes a threshold for classifying the anomalous timesteps. Specifically, our approach uses autoencoder architectures along with reconstruction loss. Hence, mentioned updating scheme can be further described as:

$$\mathcal{L}(\mathcal{X}_{w,t}, \hat{\mathcal{Y}}_{w,t}, \mu_t, \tau) = (1 - \hat{\mathcal{Y}}_{w,t}^\top)(\hat{\mathcal{X}}_{w,t} - \mathcal{X}_{w,t})^2, \tag{4}$$

where $\hat{\mathcal{X}}_{w,t}$ denotes reconstructed output from the model and $\hat{\mathcal{Y}}_{w,t}$ denotes predicted labels, where 0 and 1 indicate normal and abnormal, respectively.

Although we utilize the entire time-series data for trend estimate, we only incorporate the normal instances to update the model based on the model's predictions. The rationale behind this strategy stems from the assumption that unsupervised anomaly detectors are trained using normal data before model deployment. Consequently, the inclusion of anomaly samples for model updates during test time can potentially have a detrimental impact on the model's performance. In contrast, to enable trend estimation even in scenarios with substantial variations, it is essential to incorporate normal instances that could potentially be predicted as anomalies by the anomaly detector.

Experiments

Experiment Setups

Datasets. We selected datasets for experiments based on the following criteria: (i) widely used datasets in time-series anomaly detection literature (SWaT), (ii) subsets with significant distribution shifts from commonly utilized datasets (SMD, MSL, SMAP), (iii) datasets including substantial distribution shifts (WADI, Yahoo), (iv) datasets with minimal distribution shifts (CreditCard).

Descriptions for the real-world datasets we utilized are as follows. (1) The SWaT (Mathur and Tippenhauer 2016) and WADI [1] consist of measurements collected from water treatment system testbeds. SWaT dataset covers 11 days of measurement from 51 sensors, while WADI dataset covers

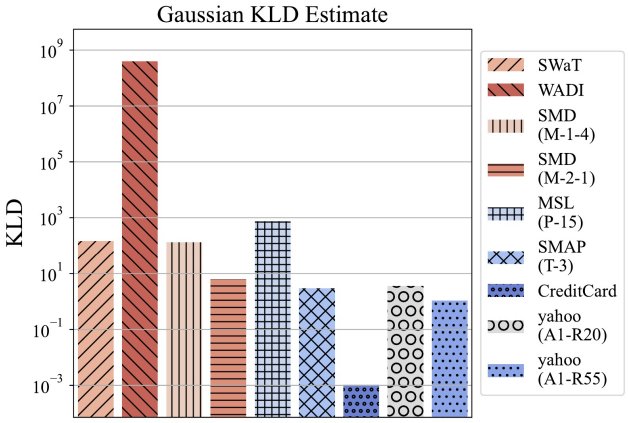


Figure 4: Kullback–Leibler Divergence (KLD) of various datasets. $D_{KL}(\mathcal{D}_{test}||\mathcal{D}_{train})$ is given, which implies how much additional information is needed to fully describe $\mathcal{D}_{test}$, given $\mathcal{D}_{train}$. The measure quantifies the distribution shift problem of the datasets.

16 days of measurement from 123 sensors. (2) The SMD dataset (Su et al. 2019) includes 5 weeks of data from 28 distinct server machines with 38-dimensional sensor inputs. For the experiment, two specific server machines (Machine 1-4 and Machine 2-1) were selected due to their distribution shift problems. (3) The SMAP and MSL (Hundman et al. 2018) datasets are derived from spacecraft monitoring systems. SMAP dataset comprises monitoring data from 28 unique machines with 55 telemetry channels, whereas MSL dataset includes data from 19 unique machines with 27 telemetry channels. Data from two specific machines with distribution shifts, MSL (P-15) and SMAP (T-3), are selected for our experiments. (4) The CreditCard dataset [2] consists of transactional logs spanning two days. It contains 28 PCA-anonymized features along with time and transaction amount information. (5) The Yahoo dataset [3] is a combination of real (A1) and synthetic (A2, A3, A4) datasets. Yahoo-A1 dataset contains 67 univariate real-world datasets, with a specific focus on two datasets (A1-R20 and A1-R55) exhibiting distribution shift problems. Further details and main statistics of the datasets can be found in the supplementary.

Baselines. We compare our methodology with 5 baselines: MLP-based autoencoder (MLP), LSTMEncDec (LSTM) (Malhotra et al. 2016), USAD (Audibert et al. 2020), THOC (Shen, Li, and Kwok 2020) and anomaly transformer (AT) (Xu et al. 2022). LSTM, USAD, and THOC have been re-implemented based on the description of each paper. Official implementation of anomaly transformer[4] is utilized in our experiments. We use hyperparameters and default settings of THOC, USAD, and AT described in their papers. MLP and LSTM use the latent dimension of 128 as default. As all the approaches are fully unsupervised, we trained all the models with the assumption of normality for train datasets. During

[1] iTrust, Centre for Research in Cyber Security, Singapore University of Technology and Design

[2] https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

[3] https://webscope.sandbox.yahoo.com/

[4] https://github.com/thuml/Anomaly-Transformer

test time, our approach gets input of w non-overlapping window, which is the same input as the train-time window size. Details of hyperparameters can be found in supplementary.

Evaluation metrics. We report a metric called F1-PA (Xu et al. 2018), widely utilized in the recent time-series anomaly detection studies (Xu et al. 2022; Shen, Li, and Kwok 2020; Audibert et al. 2020; Su et al. 2019). This metric views the whole successive abnormal segment as correctly detected if any of the timesteps in the segment is classified as an anomaly. Note that F1-PA metric overestimate classifier performance (Kim et al. 2022a), even though this metric has practical justifications (Xu et al. 2018).

Therefore, we consider three additional evaluation metrics, which are F1 score, area under receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC). Different from F1-PA, the F1 score can measure the anomaly detection status for each individual timestep, which directly reflects the performance of the anomaly detector. We also report AUROC and AUPRC over test data anomaly scores, which gives an overall summary of anomaly detector performance for all possible candidates of thresholds τ. AUROC takes into account the performance across all possible decision thresholds, making it less sensitive to the choice of a specific threshold. We measure AUPRC, which is well-suited for imbalanced classification scenarios (Saito and Rehmsmeier 2015; Sørbø and Ruocco 2023).

For brevity, we report these four metrics in the main paper. Other metrics for adjusted and non-adjusted metrics, including accuracy, precision, recall, F1, and confusion matrix (The number of true negatives, false positives, false negatives, and true positives), are provided in the supplementary.

Comparison with Baselines

Main results. To validate the effectiveness of our method, we conducted a comparative analysis between unsupervised time-series anomaly detection models and the MLP model combined with our approach. As presented in Table 1, the results demonstrate that our method consistently improves the performance of the MLP model across various evaluation metrics. Notably, we achieve a significant improvement of up to 13% in the AUROC of the WADI dataset and 51% in the AUPRC of MSL (P-15), which exhibits a distribution shift problem as illustrated in Fig. 4. In the case of the Yahoo A1-R20 dataset, shown in Fig. 2-(b), our method demonstrates the highest performance gain in terms of the F1 score. In contrast to most of the datasets, our method shows only marginal improvement in the CreditCard dataset.

It is due to the fact that the dataset has a minimal distribution shift problem, resulting in a limited performance gain. The dataset that exhibits lower F1 performance compared to the off-the-shelf baseline is WADI. This discrepancy is a result of the threshold setting with test anomaly scores. Specifically, the maximum anomaly score for the WADI train data using the USAD model is 0.225, while the threshold that yields the reported F1 score in the table is 585.845, which is significantly higher. Consequently, although USAD and LSTM models exhibit higher scores for F1, the overall classifier performance measured by AUROC is lower.

Dataset	Metrics	MLP	LSTM	USAD	THOC	AT	Ours
SWaT	F1	0.765	0.401	0.557	0.776	0.218	**0.784**
	F1-PA	0.831	0.768	0.655	0.862	**0.962**	0.903
	AUROC	0.832	0.697	0.737	0.838	0.530	**0.892**
	AUPRC	0.722	0.248	0.457	0.744	0.195	**0.780**
WADI	F1	0.131	0.245	**0.260**	0.124	0.109	0.148
	F1-PA	0.175	0.279	0.279	0.153	**0.915**	0.346
	AUROC	0.485	0.525	0.530	0.484	0.501	**0.624**
	AUPRC	0.052	0.195	**0.205**	0.144	0.059	0.081
SMD (M-1-4)	F1	0.273	0.282	0.159	0.379	0.059	**0.463**
	F1-PA	0.544	0.500	0.296	0.521	0.799	**0.874**
	AUROC	0.805	0.818	0.673	**0.869**	0.479	0.845
	AUPRC	0.169	0.151	0.103	0.223	0.034	**0.354**
SMD (M-2-1)	F1	0.236	0.283	**0.308**	0.295	0.094	0.249
	F1-PA	0.814	0.910	0.922	0.705	0.866	**0.974**
	AUROC	0.674	0.727	0.738	0.668	0.498	**0.764**
	AUPRC	0.190	0.251	0.246	0.161	0.052	**0.280**
MSL (P-15)	F1	0.263	0.056	0.060	0.018	0.071	**0.440**
	F1-PA	0.848	0.351	0.097	0.027	0.437	**0.944**
	AUROC	0.645	0.617	0.661	0.332	0.568	**0.801**
	AUPRC	0.061	0.012	0.016	0.005	0.023	**0.575**
SMAP (T-3)	F1	0.095	0.091	0.044	0.154	0.042	**0.218**
	F1-PA	0.992	**0.998**	0.940	0.747	0.772	0.708
	AUROC	0.510	0.515	0.500	0.591	0.490	**0.617**
	AUPRC	0.044	0.050	0.031	0.049	0.017	**0.111**
Credit Card	F1	0.127	0.220	**0.323**	0.138	0.039	0.135
	F1-PA	0.145	0.234	**0.323**	0.148	0.056	0.151
	AUROC	0.943	0.930	0.887	0.770	0.548	**0.943**
	AUPRC	0.055	0.109	**0.234**	0.041	0.007	0.063
Yahoo (A1-R20)	F1	0.067	0.065	0.277	0.106	0.098	**0.678**
	F1-PA	0.259	0.426	0.695	0.106	0.185	**0.895**
	AUROC	0.367	0.394	0.668	0.198	0.525	**0.971**
	AUPRC	0.056	0.057	0.161	0.067	0.048	**0.637**
Yahoo (A1-R55)	F1	0.366	0.446	0.281	0.059	0.010	**0.633**
	F1-PA	0.424	0.446	0.320	0.059	0.010	**0.744**
	AUROC	0.916	0.877	0.867	0.875	0.478	**0.958**
	AUPRC	0.303	0.242	0.177	0.019	0.002	**0.624**

Table 1: Comparison with the existing baselines. All results are based on five independent trials. This table reports the average of five trials for each metrics. Complete results with confidence intervals are reported in the supplementary.

Moreover, we compared our method to the anomaly transformer (AT), one of the state-of-the-art methods. While AT shows comparable performance in terms of F1-PA, it falls short regarding the F1 score, AUROC, and AUPRC. This disparity arises because the anomaly transformer generates positive predictions at certain intervals rather than specifying the exact moments of anomalous points. Details of test-time anomaly scores of baselines are reported in supplementary.

Analysis on ROC and Precision-Recall curves. Our method consistently outperforms previous approaches in terms of AUROC across all datasets except for SMD (M-1-4), and AUPRC across all datasets except for WADI and Creditcard. This indicates that previous off-the-shelf baselines are

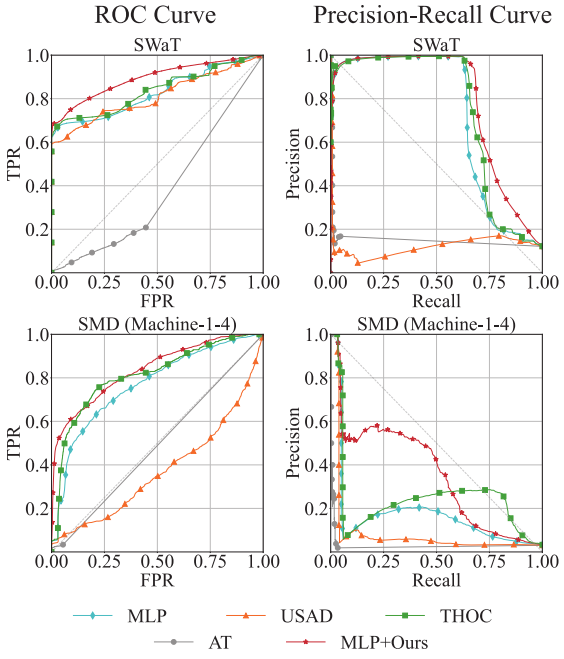


Figure 5: ROC curves (left) Precision-Recall curves (right) visualizations of baselines and MLP+Ours.

sensitive to threshold settings, which poses a challenge for robustness in real-world scenarios where finding an optimal threshold is difficult. Fig. 5 shows a visualization of the receiver operating curve (ROC curve) and precision-recall curve of our approach, along with baselines. Consistently, for both, our approach (red) improves the off-the-shelf classifier results (blue) significantly.

Results on AnoShift Benchmark

The AnoShift benchmark (Dragoi et al. 2022) offers a testbed for the robustness of anomaly detection algorithm under distribution shift problem. The dataset spans a decade, partitioned into a training set covering the period 2006-2010, and two distinct test sets denoted as NEAR (2011-2013) and FAR (2014-2015). Visualized in Fig. 6-(a), the data distribution progressively deviates from the train set as time progresses.

The principal objective of evaluation on the AnoShift benchmark is to investigate the effectiveness of our proposed algorithm against such distribution shifts. The evaluation entails three metrics—namely, Area Under the Receiver Operating Characteristic curve (AUROC), Area Under the Precision-Recall Curve with inliers as the positive class (AUPRC-in), and Area Under the Precision-Recall Curve with outliers as the positive class (AUPRC-out), following previous work (Dragoi et al. 2022). The performance of our method is compared to other deep-learning-based baselines, including SO-GAAL (Liu et al. 2020), deepSVDD (Ruff et al. 2018), LUNAR (Goodge et al. 2022), ICL (Shenkar and Wolf 2022), BERT (Devlin et al. 2019) for anomalies.

Table 2 demonstrates a significant improvement in performance when our method is integrated into an MLP-based autoencoder, as evidenced by an increase in AUROC of up to 0.216. Despite its simplicity, our approach markedly augments the baseline MLP performance, which previously showed inferior performance. This improvement is especially significant in FAR splits, which entail a severe distribution shift problem compared to NEAR splits. While our experiments focused on MLP, it's worth noting that our module can be seamlessly added to other baselines.

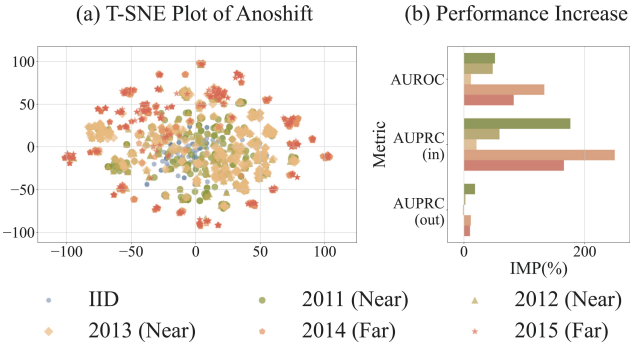


Figure 6: (a) T-SNE plot according to chronological distance and (b) performance increase with respect to three different evaluation metrics: AUROC, AUPRC-in and AUPRC-out.

Method	NEAR			FAR		
	ROC	PRC (in)	PRC (out)	ROC	PRC (in)	PRC (out)
SO-GAAL†	0.545	0.435	0.877	**0.493**	0.107	**0.927**
deepSVDD†	**0.870**	**0.717**	0.942	0.345	0.100	0.823
LUNAR†	0.490	0.294	0.809	0.282	0.093	0.794
ICL†	0.523	0.273	0.819	0.225	0.088	0.775
BERT†	0.861	0.589	**0.960**	0.281	0.082	0.784
MLP†	0.441	0.262	0.730	0.200	0.085	0.757
MLP	0.441	0.207	0.776	0.208	0.085	0.758
MLP+Ours	0.639 (+0.194)	0.404 (+0.197)	0.841 (+0.065)	0.424 (+0.216)	**0.259** (+0.173)	0.838 (+0.081)

Table 2: Performance on Anoshift benchmark. † denotes that metrics are reported from the results in the original paper. AUROC and AUPRC are denoted as ROC and PRC.

Ablation Study

As shown in Table 3, we perform the ablation study on our method to analyze the effectiveness of each component. MLP with detrend module and test-time adaptation with the model update is consistently showing better results, compared to the cases when used alone (MLP+DT, MLP+TTA) and none of them used (MLP). Here, DT and TTA denote a detrend module and test-time adaptation with model updates, respectively. Moreover, Fig. 7 also demonstrates that when the appropri-

DT	TTA	SWaT				SMD (M-2-1)				MSL (P-15)			
		F1	F1-PA	AUROC	AUPRC	F1	F1-PA	AUROC	AUPRC	F1	F1-PA	AUROC	AUPRC
✗	✗	0.765	0.834	0.832	0.722	0.236	0.814	0.674	0.190	0.263	0.848	0.645	0.061
✓	✗	0.762	0.837	0.846	0.738	0.234	0.855	0.749	0.205	0.221	0.703	0.799	0.124
✗	✓	**0.784**	**0.907**	0.888	0.778	0.239	0.881	0.689	0.204	0.019	0.027	0.640	0.060
✓	✓	**0.784**	0.903	**0.892**	**0.780**	**0.249**	**0.974**	**0.764**	**0.280**	**0.440**	**0.944**	**0.801**	**0.575**

Table 3: Ablation study on our proposed method. DT and TTA indicate a detrend module and test-time adaptation, respectively.

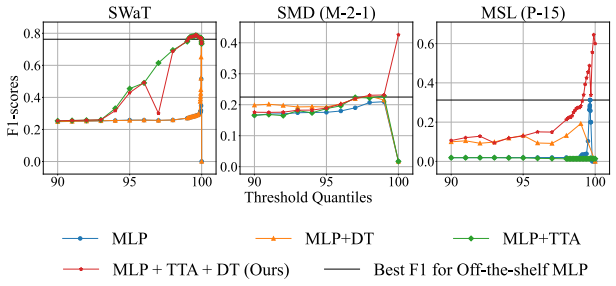
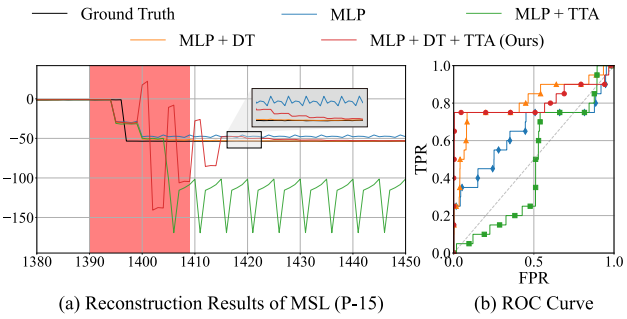


Figure 7: F1 scores according to various thresholds.



(a) Reconstruction Results of MSL (P-15) (b) ROC Curve

Figure 8: Ablation study on the proposed method using the MSL (P-15) dataset.

ate threshold is selected MLP model with our full method consistently outperforms these baselines, including the best performance of the off-the-shelf MLP model.

This behavior can be further described in Fig. 8-(a), illustrating those four options at once. (1) Our approach (red) shows better reconstruction compared to off-the-shelf MLP (blue). The off-the-shelf MLP model is constantly generating reconstruction errors even after the transition of an overall trend, which results in many false positive cases. (2) Also, the detrend module alone fails to detect anomalies, showing less sensitivity compared to our approach, although they share the same EMA parameter γ. This shows model update can contribute to such sensitivity of the anomaly detector, as it keeps updating with recent observations. (3) Without proper update of such trend estimate, test-time adaptation with model updates alone (green) can harm the robustness of the model, as it can be overfitted to sequence before trend shift, with a lack of ability to adapt to newly coming sequences.

Discussion and Limitation

Threshold for Anomaly Detection. Existing unsupervised time-series anomaly detection studies (Audibert et al. 2020; Xu et al. 2022) have a major limitation in that they determine the threshold for normality by inferring the entire test data and selecting it based on the best performance. However, this approach is not practically feasible in real-world scenarios. Therefore, we report AUROC to evaluate overall performance and decide the threshold based on the training data statistics in our experiments. We posit that the performance of the anomaly detector could be further enhanced with an appropriate choice of threshold.

Inconsistent Labeling in Anomaly Detection. In the time-series anomaly detection task, the criteria of anomaly vary for each scenario, making it difficult to establish consistent labels. For this reason, distinguishing whether test samples with significant differences from the normal in train sets are abnormal or normal with distribution shifts is challenging. In our case, based on the assumption that there are more normal instances in test sets, we employ trend estimation and model predictions for test-time adaptation. To improve the adaptation performance, employing active learning (Ren et al. 2021) where human annotators provide labels for a subset of test data can be a valuable research direction.

Conclusion

In this work, we highlighted the distribution shift problem in unsupervised time-series anomaly detection. We have shown that the concept of normality may change over time. This can be a significant challenge for designing robust time-series anomaly detection frameworks, leading to many false positives, which harms the system's consistency. To mitigate this issue, we propose a simple yet effective strategy of incorporating new normals into the model architecture, by following trend estimates along with test-time adaptation. Concretely, our method consistently outperforms standard baselines for real-world benchmarks with such problems.

Acknowledgements

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A2B5B02001913 & No. 2022R1A5A708390812).

References

Abdulaal, A.; Liu, Z.; and Lancewicki, T. 2021. Practical Approach to Asynchronous Multivariate Time Series Anomaly Detection and Localization. In *Proc. the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Audibert, J.; Michiardi, P.; Guyard, F.; Marti, S.; and Zuluaga, M. A. 2020. USAD: UnSupervised Anomaly Detection on Multivariate Time Series. In *Proc. the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Breunig, M. M.; Kriegel, H.; Ng, R. T.; and Sander, J. 2000. LOF: Identifying Density-Based Local Outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA*.

Cao, T.; Zhu, J.; and Pang, G. 2023. Anomaly Detection under Distribution Shift. *CoRR*, abs/2303.13845.

Choi, S.; Yang, S.; Choi, S.; and Yun, S. 2022. Improving test-time adaptation via shift-agnostic weight regularization and nearest source prototypes. In *Proc. of the European Conference on Computer Vision (ECCV)*, 440–458. Springer.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proc. of The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Dragoi, M.; Burceanu, E.; Haller, E.; Manolache, A.; and Brad, F. 2022. AnoShift: A Distribution Shift Benchmark for Unsupervised Anomaly Detection. In *NeurIPS*.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1): 2096–2030.

Geiger, A.; Liu, D.; Alnegheimish, S.; Cuesta-Infante, A.; and Veeramachaneni, K. 2020. TadGAN: Time Series Anomaly Detection Using Generative Adversarial Networks. In *2020 IEEE International Conference on Big Data (IEEE BigData 2020), Atlanta, GA, USA, December 10-13, 2020*, 33–43. IEEE.

Goodge, A.; Hooi, B.; Ng, S.; and Ng, W. S. 2022. LUNAR: Unifying Local Outlier Detection Methods via Graph Neural Networks. In *Proc. the AAAI Conference on Artificial Intelligence (AAAI)*.

Gulrajani, I.; and Lopez-Paz, D. 2021. In Search of Lost Domain Generalization. In *Proc. the International Conference on Learning Representations (ICLR)*.

Han, C.; Rundo, L.; Murao, K.; Noguchi, T.; Shimahara, Y.; Milacski, Z. Á.; Koshino, S.; Sala, E.; Nakayama, H.; and Satoh, S. 2021. MADGAN: unsupervised medical anomaly detection GAN using multiple adjacent brain MRI slice reconstruction. *BMC Bioinform.*, 22-S(2): 31.

Hundman, K.; Constantinou, V.; Laporte, C.; Colwell, I.; and Söderström, T. 2018. Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding. In *Proc. the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Kim, S.; Choi, K.; Choi, H.; Lee, B.; and Yoon, S. 2022a. Towards a Rigorous Evaluation of Time-Series Anomaly Detection. In *Proc. the AAAI Conference on Artificial Intelligence (AAAI)*.

Kim, T.; Kim, J.; Tae, Y.; Park, C.; Choi, J.; and Choo, J. 2022b. Reversible Instance Normalization for Accurate Time-Series Forecasting against Distribution Shift. In *Proc. the International Conference on Learning Representations (ICLR)*.

Lai, K.; Zha, D.; Xu, J.; Zhao, Y.; Wang, G.; and Hu, X. 2021. Revisiting Time Series Outlier Detection: Definitions and Benchmarks. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*.

Liang, J.; Hu, D.; and Feng, J. 2020. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, 6028–6039. PMLR.

Liu, L.; Liu, D.; and Peng, Y. 2016. Detection and identification of sensor anomaly for aerospace applications. In *2016 Annual Reliability and Maintainability Symposium (RAMS)*, 1–6. IEEE.

Liu, Y.; Li, Z.; Zhou, C.; Jiang, Y.; Sun, J.; Wang, M.; and He, X. 2020. Generative Adversarial Active Learning for Unsupervised Outlier Detection. *IEEE Trans. Knowl. Data Eng.*, 32(8): 1517–1528.

Liu, Y.; Wu, H.; Wang, J.; and Long, M. 2022. Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting. In *NeurIPS*.

Malhotra, P.; Ramakrishnan, A.; Anand, G.; Vig, L.; Agarwal, P.; and Shroff, G. 2016. LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection. *CoRR*, abs/1607.00148.

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@CPSWeek 2016, Vienna, Austria, April 11, 2016*.

Muth, J. F. 1960. Optimal properties of exponentially weighted forecasts. *Journal of the american statistical association*, 55(290): 299–306.

Niu, S.; Wu, J.; Zhang, Y.; Chen, Y.; Zheng, S.; Zhao, P.; and Tan, M. 2022. Efficient test-time model adaptation without forgetting. In *Proc. the International Conference on Machine Learning (ICML)*, 16888–16905. PMLR.

Pang, G.; Shen, C.; Cao, L.; and van den Hengel, A. 2022. Deep Learning for Anomaly Detection: A Review. *ACM Comput. Surv.*, 54(2): 38:1–38:38.

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 Autom. Lett.*

Pena, E. H. M.; de Assis, M. V. O.; and Jr., M. L. P. 2013. Anomaly Detection Using Forecasting Methods ARIMA and HWDS. In *32nd International Conference of the Chilean Computer Science Society, SCCC 2013, Temuco, Cautin, Chile, November 11-15, 2013*, 63–66. IEEE Computer Society.

Quinonero-Candela, J.; Sugiyama, M.; Schwaighofer, A.; and Lawrence, N. D. 2008. *Dataset shift in machine learning*. Mit Press.

Rassam, M. A.; Maarof, M. A.; and Zainal, A. 2018. A distributed anomaly detection model for wireless sensor networks based on the one-class principal component classifier. *Int. J. Sens. Networks*, 27(3): 200–214.

Ren, P.; Xiao, Y.; Chang, X.; Huang, P.-Y.; Li, Z.; Gupta, B. B.; Chen, X.; and Wang, X. 2021. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9): 1–40.

Ruff, L.; Görnitz, N.; Deecke, L.; Siddiqui, S. A.; Vandermeulen, R. A.; Binder, A.; Müller, E.; and Kloft, M. 2018. Deep One-Class Classification. In *Proc. the International Conference on Machine Learning (ICML)*.

Ruff, L.; Kauffmann, J. R.; Vandermeulen, R. A.; Montavon, G.; Samek, W.; Kloft, M.; Dietterich, T. G.; and Müller, K. 2021. A Unifying Review of Deep and Shallow Anomaly Detection. *Proc. IEEE*, 109(5): 756–795.

Saito, T.; and Rehmsmeier, M. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3): e0118432.

Sankararaman, A.; Narayanaswamy, B.; Singh, V. Y.; and Song, Z. 2022. FITNESS: (Fine Tune on New and Similar Samples) to detect anomalies in streams with drift and outliers. In *Proc. the International Conference on Machine Learning (ICML)*.

Saurav, S.; Malhotra, P.; TV, V.; Gugulothu, N.; Vig, L.; Agarwal, P.; and Shroff, G. 2018. Online anomaly detection with concept drift adaptation using recurrent neural networks. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, COMAD/CODS 2018, Goa, India, January 11-13, 2018*.

Schlegl, T.; Seeböck, P.; Waldstein, S. M.; Schmidt-Erfurth, U.; and Langs, G. 2017. Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. In *Information Processing in Medical Imaging - 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings*, volume 10265 of *Lecture Notes in Computer Science*, 146–157. Springer.

Schölkopf, B.; Williamson, R. C.; Smola, A. J.; Shawe-Taylor, J.; and Platt, J. C. 1999. Support Vector Method for Novelty Detection. In Solla, S. A.; Leen, T. K.; and Müller, K., eds., *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*.

Shen, L.; Li, Z.; and Kwok, J. T. 2020. Timeseries Anomaly Detection using Temporal Hierarchical One-Class Network. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*.

Shenkar, T.; and Wolf, L. 2022. Anomaly Detection for Tabular Data with Internal Contrastive Learning. In *Proc. the International Conference on Learning Representations (ICLR)*.

Shin, Y.; Lee, S.; Tariq, S.; Lee, M. S.; Jung, O.; Chung, D.; and Woo, S. S. 2020. ITAD: Integrative Tensor-based Anomaly Detection System for Reducing False Positives of Satellite Systems. In *Proc. the ACM Conference on Information and Knowledge Management (CIKM)*.

Shumway, R. H.; and Stoffer, D. S. 2017. *Time series analysis and its applications: With R examples*. Springer.

Sørbø, S.; and Ruocco, M. 2023. Navigating the Metric Maze: A Taxonomy of Evaluation Metrics for Anomaly Detection in Time Series. *CoRR*, abs/2303.01272.

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 *Proc. the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A. A.; and Hardt, M. 2020. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In *Proc. the International Conference on Machine Learning (ICML)*.

Tax, D. M. J.; and Duin, R. P. W. 1999. Data domain description using support vectors. In *7th European Symposium on Artificial Neural Networks, ESANN 1999, Bruges, Belgium, April 21-23, 1999, Proceedings*.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B. A.; and Darrell, T. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *Proc. the International Conference on Learning Representations (ICLR)*.

Wang, J.; Kuang, Q.; and Duan, S. 2015. A new online anomaly learning and detection for large-scale service of Internet of Thing. *Pers. Ubiquitous Comput.*, 19(7): 1021–1031.

Wang, Q.; Fink, O.; Gool, L. V.; and Dai, D. 2022. Continual Test-Time Domain Adaptation. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Xu, H.; Chen, W.; Zhao, N.; Li, Z.; Bu, J.; Li, Z.; Liu, Y.; Zhao, Y.; Pei, D.; Feng, Y.; Chen, J.; Wang, Z.; and Qiao, H. 2018. Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications. In *Proc. the International Conference on World Wide Web (WWW)*.

Xu, J.; Wu, H.; Wang, J.; and Long, M. 2022. Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy. In *Proc. the International Conference on Learning Representations (ICLR)*.

Yoo, J.; Chung, I.; and Kwak, N. 2022. Unsupervised Domain Adaptation for One-Stage Object Detector Using Offsets to Bounding Box. In *Proc. of the European Conference on Computer Vision (ECCV)*, 691–708. Springer.

Zinkevich, M. 2003. Online Convex Programming and Generalized Infinitesimal Gradient Ascent. In *Proc. the International Conference on Machine Learning (ICML)*.

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 *Proc. the International Conference on Learning Representations (ICLR)*.

Zou, Y.; Yu, Z.; Kumar, B.; and Wang, J. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proc. of the European Conference on Computer Vision (ECCV)*, 289–305.