A Comparative Study of Oracle's Anomaly Detection Solution and Modern Alternatives in Time Series Prognostics

Matthew Gerdes
Oracle
Austin, Texas, USA
matthew.gerdes@oracle.com

Guang Wang*
Oracle
Austin, Texas, USA
guang.wang@oracle.com

Abstract

Time series anomaly detection is a difficult problem that has been studied in a broader spectrum of research areas due to its diverse applications in different domains. Despite significant progress in this field, including the widespread adoption of modern machine learning algorithms, no single anomaly detection method has proven to generalize effectively across all time series datasets. Nevertheless, the adoption of deep learning techniques—particularly the Long Short Term Memory (LSTM) algorithm—for time series analysis has continued to grow in both academia and among major cloud service providers. The increase in the usage of LSTM is largely driven by the belief that neural networks (NN)—given their success in many other domains—can be generalized for all predictive tasks, along with the widespread availability of open-source implementations. However, there are alternatives to LSTM that may be better suited to address the unique challenges of time series analysis—one such method is MSET (Multivariate State Estimation Technique). In this study, we conducted a comprehensive comparative evaluation of MSET against other state-of-the-art techniques from the literature to better understand its value proposition. A benchmark test bed is developed to evaluate the detection results, reconstruction accuracy, and computational cost of the anomaly detection techniques of being studied. The benchmark datasets consist of synthetic datasets and publicly available datasets. MSET is demonstrated to achieve a higher F1 score, on average, than LSTM in most cases, and deliver an advantage over other competing methods in regards to false alarms, reconstruction accuracy, and computational cost. Lastly, although the explainability cannot be quantified in our study, we showcase it is a key value proposition of MSET favored by the IoT industries targeted by MSET.

Keywords

Anomaly Detection, Unlabeled Training Data, Supervised Learning, Time Series

*Corresponding author

MILETS '25, Toronto, Canada

ACM Reference Format:
Matthew Gerdes and Guang Wang. 2025. A Comparative Study of Oracle's Anomaly Detection Solution and Modern Alternatives in Time Series Prognostics. In *Proceedings of The 11th Mining and Learning from Time Series Workshop: From Classical Methods to LLMs (MILETS '25).* ACM, New York, NY, USA, 9 pages.

1 Introduction

Anomalies are patterns in data that do not conform to a well defined notion of normal behavior. While there are many anomaly detection applications specific to the type of data, this paper focuses on anomaly detection in time series data. Machine learning (ML) techniques are increasingly being used in the area of time series anomaly detection thanks to the ever-growing computational capacity. Conventional attempts to remove anomalies in the time series data are based on simplistic outlier detection methods, such as three standard deviation thresholds [3], which are not capable of detecting "inlier" (anomalies that stay within the normal range of the signal). Unsupervised techniques are often utilized when no prior knowledge of the training dataset is available. However, many unsupervised techniques employ clustering type algorithms that assumes anomalous data resides inside small clusters [19], or are built through linear projection and transformation, which is unable to handle non-linearity and exploit the inter-correlations of multivariate time series [16]. More sophisticated unsupervised methods such as Generative Adversarial Networks (GAN) [6] have proven helpful, but the fact that its training objective often results in saddle point convergence makes the GAN models difficult to train. On the other hand, supervised techniques for time series anomaly detection have been intensively studied and widely adopted for decades, from the conventional approaches such as Support Vector Machines [17] and Artificial Neural Networks [12], to more recent Long Short Term Memory (LSTM) models which have shown strong performance in capturing temporal dependencies[11, 15, 22].

Despite a wide spectrum of anomaly detection approaches being available, they often show promise on specific datasets only. Additionally, there is a dearth of evaluation for these techniques against prognostic functional requirements. Motivated by this, we review the literature to identify state-of-the-art anomaly detection techniques and conduct a benchmark study to evaluate their performance—alongside our Multivariate State Estimation Technique (MSET) [21] across multiple aspects relevant to time series anomaly detection. Specifically, we compare the detection decisions, reconstruction accuracy as well as the computation time between the methods using both synthetic and publicly available datasets. We provide a thorough quantitative assessment about the performance of these techniques using a standard set of benchmark metrics,

alongside a general notion of the competitive differentiating features that MSET possesses over LSTM and other 7 alternatives.

The paper is organized as follows. In Section 2, the multivariate time series anomaly detection models and the datasets used in this work are introduced. In Section 3, we present two detailed case studies using both synthetic and published datasets, followed by an extensive benchmark testing with more published datasets. Section 4 provides deep insight into the performance and value propositions of MSET over the alternative options presented in the study. Finally, Section 5 concludes the paper.

2 Data and Benchmark Setup

2.1 Anomaly Detection Technique: MSET

The anomaly detection technique used in this benchmark study is a nonlinear, nonparametric, multivariate pattern recognition technique, called the Multivariate State Estimation Technique (MSET). It was originally developed by Argonne National Laboratory (ANL) to discover anomalies in time series sensor data in nuclear power applications [7, 10]. Over the years, MSET has been evolved and scaled to the big data prognostic applications commonly seen in safety-critical industries including aerospace, utilities, and computer systems [8, 9].

2.2 Synthetic Data for Benchmark

Public benchmark datasets have been commonly used to benchmark various anomaly detection techniques. However, Wu and Keogh [27] recently conducted a careful evaluation of these datasets and concluded the majority of the faults suffer from one or more categories of flaws. Other recent studies can be found in [14]. Our team has been aware of the difficulties in procuring viable real world datasets [25], and to address the reasonable concerns outlined in [27], we have developed a compendium of realistic types of time series sensor fault signatures. The signatures have been observed in real anomaly detection use cases, across a variety of industries, providing a reliable test bed with "known ground truth" injected faults and "known ground truth" absence of faults, for evaluating the anomaly detection techniques. The compendium of fault types is summarized as follows:

- Ramp: An anomally emulating signal drift, where degradation initiates "inside" the noise band, develops over time, and eventually exceeds signal's normal range.
- Mean Shift: A graduate or abrupt change point in the mean level of a time series (e.g., caused by sensor decalibration bias, sudden changes in the ambient condition, or onset of severe asset degradation).
- Gain Change: A class of "inlier failures" in IoT industries, where the physical transducer diminishes its response to the physical parameter it is sensing, which the instrumentation specialists call "Loss-of-Gain" failures.
- Time Lag: A type of clock-skew error in one or more distributed Data Acquisition modules.
- Signal Dropout: One sensor out of an array of sensors suddenly disappears (data acquisition fault) or goes to a perfectly flat line (transducer "stuck-at" fault), resulting in the loss of its share in subsequent accumulated measurements.

Incidentally, to create synthetic datasets for analysis, a synthetic signal generator is often used for machine learning tuning and validation. Lai et. al [13] have developed a quality signal generator[1]. We utilize our advanced Telemetry Parameter Synthesis System, which generates synthetic time series telemetry based on real measured signals using Fourier decomposition and reconstruction. The system allows customization of sampling rates, signal-to-noise ratios, serial correlation, amplitudes, mean values, and a user-defined "dispersion factor" to distribute the signals more broadly and arbitrarily across a specified range.

The aforementioned library of fault types is synthesized and injected individually or collectively, into one or more synthetic signals, at the same and/or different times under a wide variety of single-fault, multiple-sequential fault, and multiple concurrent fault scenarios. A total of 15 datasets are generated as the test cases.

2.3 Real World Data for Benchmark

We also conducted an extensive literature review and identified several real-world datasets commonly used in anomaly detection research. We carefully evaluated them, and selected a subset of datasets to serve as the benchmark data in this study. The selected datasets are listed and briefly described below.

2.3.1 Pool Server Metric (PSM). A dataset collected internally from multiple application server nodes at eBay[1], including numerous signals from business performance data, such as user traffic and activity, to infrastructure data, such as application CPU and memory utilization. An anomaly indicates a potential threat to the business operation, for example, cybersecurity attacks or an internal code bug, which could result in service downtime. The dataset consists of 26 variables, and the training set consists of 13 weeks, followed by eight weeks testing data with known anomalies in presence.

2.3.2 Secure Water Treatment System Data (SWAT) & Water distribution testbed (WADI). SWaT[5] is a water treatment testbed for research in the area of cyber security, which was built by SUTD with the aim to publish time series data meant for anomaly detection research. The data contain multivariate time series measurements from a scale model of Singapore's water treatment system and has real anomalies in the testing dataset that are labeled for a ground truth comparison[2]. The data consists of 51 features, and spans over 11 days of continuous operation including 7 days under normal operation and 4 days with attack scenarios.

WaDi[2] is an extension of SWaT. WaDi utilizes portion of SWaT's reverse osmosis permeate and raw water, into its additional water tanks and reservoirs, thus forming a complete and realistic water treatment, storage and distribution network. The combination of these two testbeds allow researchers to witness the cascading effects of cyber attacks on one testbed to another. The WaDi data features 123 sensors and actuators, operating for 16 days. The data collected during the first 14 days are purported to be clean and used for training, while the rest contains anomalies. The test dataset had a total of 15 anomaly segments.

[1] https://github.com/datamllab/tods/tree/benchmark/benchmark/synthetic/Generator
[2] https://itrust.sutd.edu.sg/itrust-labs_datasets/, accessible as of May 2025.

2.4 Benchmark Setup

In the benchmark study, we first compare MSET to LSTM that is the core algorithm of many commercial anomaly detection services (e.g., Azure Anomaly Detection[3], AWS Lookout for Metrics[4]) with the synthetic datasets. Note that the role of ML in timeseries anomaly detection applications is predicting sensor values, and therefore an additional anomaly detector is required to detect and label the anomalies. While MSET leverages a sequential probability ratio test (SPRT)[24] as its anomaly detector, a common approach [26, 30] to pairing an anomaly detector with LSTM is implementing an algorithm that slides a moving window across the residuals between the LSTM model estimates and actual measurements, calculates the MAE in that window, and compares it to a predefined threshold. The threshold and the window size are tuned by iteratively running the LSTM model through ground truth clean data until all false alarms are removed.

To enable a more robust evaluation, we expanded the study to include additional modern anomaly detection techniques that are publicly available in the academic research with the real-world datasets. The selected techniques are listed and briefly described below.

2.4.1 DeepAnT[18]. DeepAnT consists of two modules: time series forecaster and anomaly detector. The time series predictor module uses a convolutional neural network (CNN) to predict the next time stamp on the defined horizon. This module takes a window of time series (used as a context) and attempts to predict the next time stamp. The predicted value is then passed to the anomaly detector module, which is responsible for tagging the corresponding time stamp as normal or abnormal.

2.4.2 RANSynCoders[1]. RANSynCoders is an unsupervised deep learning architecture for real-time anomaly detection and localization within large multivariate time series. It uses spectral analysis on the feature representation to capture frequency domain information in multivaraite time series signals. The method utilizes synchrony-analysis on latent representations for adjusting asynchronous variate fed into an encoder, bootstrap aggregation of decoders, and quantile loss optimization for anomaly detection.

2.4.3 OmniAnomaly[23]. OmniAnomaly is a recurrent neural network (RNN) based method that integrates RNNs and Variational Auto-encoder to account for both temporal dependence and the stochastic nature of MTS, which improves the capability of learning robust representations of multivariate timeseries.

2.4.4 TS2Vec[29]. TS2Vec performs contrastive learning in a hierarchical way over augmented context views, which enables a robust contextual representation for each timestamp.

2.4.5 Temporal Hierarchical One-Class (THOC) network[20]. THOC is a temporal one-class classification model for time series anomaly detection. It captures temporal dynamics in multiple scales by using a dilated RNN with skip connections. Using multiple hyperspheres obtained with a hierarchical clustering process, a one-class objective called Multiscale Vector Data Description is defined. This allows the temporal dynamics to be well captured by a set of multi-resolution temporal clusters.

2.4.6 CARLA (ContrAstive Representation Learning Approach)[4]. CARLA leverages existing generic knowledge about time series anomalies and injects various types of anomalies as negative samples. It learns normal behavior as well as deviations indicating anomalies. It creates similar representations for temporally closed windows and distinct ones for anomalies. Additionally, it leverages the information about the representation's neighbors through a self-supervised approach to classify windows based on their nearest/furthest neighbors in the representation space to further enhance the performance of anomaly detection.

2.4.7 DCdetector[28]. DCdetector is a multi-scale dual attention contrastive representation learning model. It utilizes a dual attention asymmetric design to create the permutated environment and pure contrastive loss to guide the learning process, thus learning a permutation invariant representation with superior discrimination abilities.

Among the seven state-of-the-art anomaly detection techniques, we were able to obtain the source code for RANSync and DeepAnT. To enable a comprehensive benchmark, we implemented these models in our testbed and evaluated their performance against MSET and LSTM using the published datasets introduced in Section 2.3. For the remaining five techniques, we referenced the reported performance metrics from their respective studies directly.

2.5 Benchmark Metrics

To compare different prognostic algorithms, a set of statistical metrics is required to quantify the performance of the algorithms. This paper uses the common benchmark metrics to investigate the accuracy of the model estimates as well as the anomaly decisions, such as Precision–Recall–F1 score, and Root-Mean-Square Error (RMSE), as defined in Eqns. (1 - 2).

Finally, an important quantitative functional requirement—often a key factor in scaling time series prognostics—is computational cost. It is well established that, for time series machine learning algorithms, computational cost increases approximately linearly with the number of observations but non-linearly with the number of input signals. To assess this overhead, compute time (in seconds) was measured and reported as part of the evaluation.

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP+FP}}, \\ \text{Recall} &= \frac{\text{TP}}{\text{TP+FN}}, \\ \text{F1} &= \frac{2\text{TP}}{2\text{TP+FP+FN}}, \end{aligned} \tag{1}$$

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{\sum_{i=0}^{N_{\text{total}}-1}(y_i - \hat{y}_i)^2}{N_{\text{total}}}}, \tag{2}$$

where TP, FP, FN are true positives, false positives, false negatives, y_i and $\widehat{y_i}$ represent the model estimate and actual observation at timestamp i.

[3]https://azure.microsoft.com/en-us/products/ai-services/ai-anomaly-detector
[4]https://aws.amazon.com/lookout-for-metrics/

3 Benchmark Results

3.1 Benchmark using Synthetic Data

As introduced in Section 2.2, We have developed 15 test cases, utilizing a library of faults, that mimic real world anomalies in time series data. These cases are derived from sophisticated sinusoidal composites; each has 5000 timestamps and 20 signals. The first 2500 timestamps are considered training data whereas the rest contains one or more contextual anomalies across one or more signals. The fault location (initial timestamp and signal), length, type, and quantity were all randomly selected. Other features such as the mean and variance were also randomized to increase the complexity of the datasets.

3.1.1 Case Study. One detailed example is analyzed in this section to illustrate the performance of MSET and LSTM. The benchmark study introduced in Section 2.4 is conducted on Case #13, in which the dataset contains three faults located on 3 of 20 signals. The fault types are contextual and within 3 standard deviations from the mean. Two of the three anomalies spanning different lengths are caused by attenuation of the sensor readings of different magnitudes. The third anomaly is caused by a time lag. Fig. 1 illustrates one of the signals that contains a fault to explain the performance of MSET.

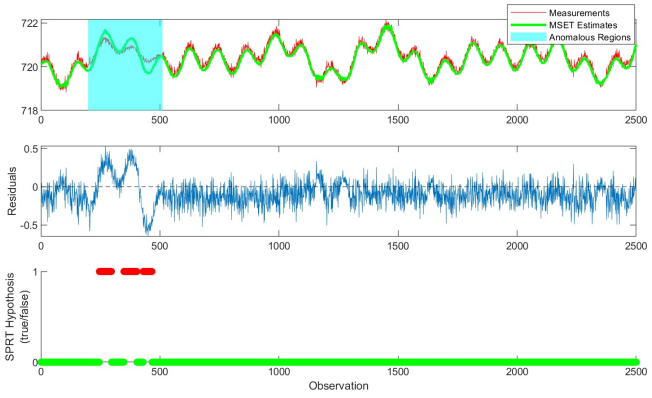


Figure 1: The anomaly detection results of MSET on a faulty signal in Case #13. Top: model estimates (green) vs. the actual measurements (red). Dashed line indicates the range of the training data. Middle: residuals between the MSET estimates and the actual measurements. Bottom: Anomaly decisions classified as either green (normal) or red (anomalous) are made by applying the anomaly detector (SPRT) to the residuals.

In the case, an MSET model is trained using the training part of the dataset (1st half), and then used to produce estimates for testing part of the dataset (2nd half). Pairwise differences between the model estimates and the actual testing observations are calculated. The residuals are analyzed using SPRT to produce anomaly detections. A decision value of 0 is assigned to the residual values deemed as expected behavior (green dots), and 1 to the residual values deemed as anomalous (red dots).

Similarly, LSTM is evaluated on the same use case (Fig. 2), where a LSTM model is trained using the same training data and used to make predictions on the same testing data. The residuals are then analyzed by the MAE dynamic threshold algorithm (Section 2.4) to produce anomaly decisions. A requirement of LSTM is the normalization of the signals before training. Consequently, the test measurements must also be normalized for reasonable results.

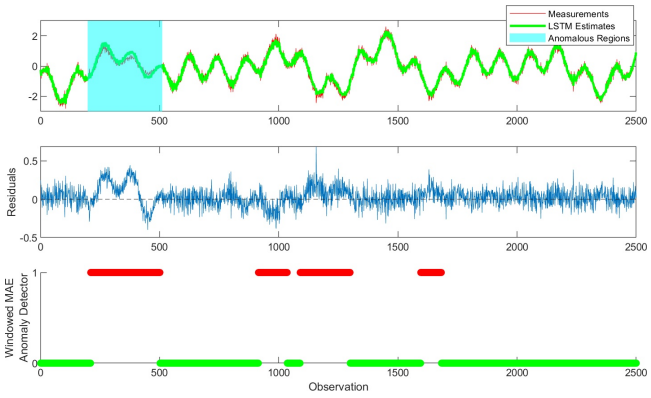


Figure 2: The anomaly detection results of LSTM on the same faulty signal in Case #13, in comparison to the MSET results as shown in Fig. 1.

As shown in Figs. 1-2, both methods were able to detect the known anomaly event ranging from the 250th to 500th timestamps. MSET, in conjunction with SPRT, localized the anomaly event with a 3-segment "alarm string" (red dots in Fig. 1), which reveals the anomaly event. The rest of the testing data was correctly labeled as clean. The LSTM method localized the anomaly event with less uncertainty by flagging more anomalous observations, revealing the duration of the anomaly more concisely; however, that comes with a tradeoff. More false positives were produced, which is indicated by the consecutive red dots around the 1000th and 1600th observations in Fig. 2, reducing the overall reliability.

The ability to correctly identify truly anomalous observations is measured by Recall, while the ability to correctly distinguish truly normal observations is measured by Precision. The ideal performance metrics would be full Recall (1) and full Precision (1). In reality, the perfect metrics can never be achieved. There is always a trade-off between Recall and Precision metrics: to label all anomalous observations that bracket an anomaly, the model and detector have to act aggressively and label any unusual behavior, including statistical noise, as anomalous, which inevitably causes false alarms. MSET is able to achieve a balance between aggressive and conservative detection strategies because SPRT functions with defined false alarm and missed alarm probabilities. Illustrated by this case, although MSET captured fewer anomalous points than LSTM, it still performed adequately in localizing the fault, which is equally informative to domain experts in real world applications when it is compared to LSTM. Other than the anomaly region, MSET achieved perfect Precision (0 False Positive), which avoids expensive asset shutdown.

3.1.2 Aggregate Case Statistics. The evaluation of both MSET and LSTM methods is conducted over an extensive set of 15 cases and the case-by-case performance comparison is summarized in Figure

3. Overall, MSET outperforms LSTM in F1 score and Precision, and underperforms in Recall, as the findings in Case #13 are persistent throughout the 15 cases: MSET tends to outperform LSTM in Precision by a wide margin while underperform LSTM in Recall by a small margin over the course of all signals, resulting in an overall better F1 score. Again, it is worth noting that although MSET generally misses more anomalous observations than LSTM from a quantitative perspective, we have confirmed that MSET does not miss any anomalies from a qualitative perspective (meaning it never completely misses the anomalies).

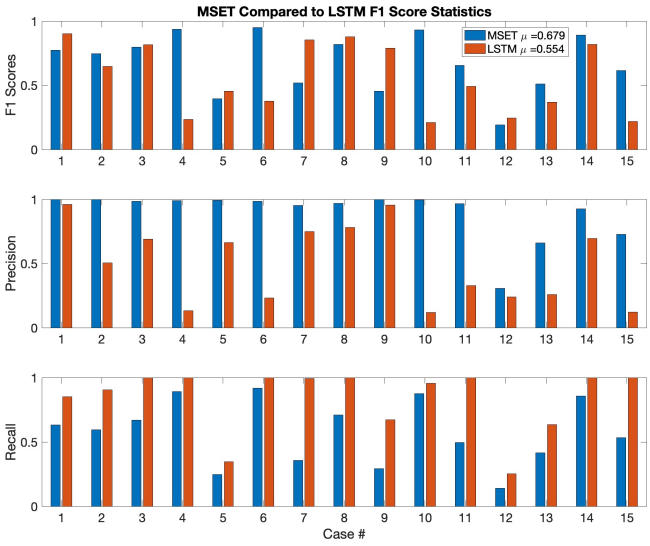


Figure 3: Bar charts of F1 score (top), Precision (middle), and Recall (bottom) for MSET (blue) and LSTM (red) over 15 test cases.

Moreover, the complexity of the dataset varies with the test cases. Cases #14 and #15 particularly are much more complicated than the others as the datasets are composed of groups of signals of different correlations or exhibit a high degree of seasonality. MSET outperforms LSTM on both cases by a wide margin, which demonstrates that MSET may possess advantages over other methods in some circumstances.

3.1.3 Computational Cost. In addition to the performance metrics for anomaly detection, the training time for each model was also recorded. As the complexity of the cases increases, the models must also become more sophisticated to maintain high detection performance. The parameter to increase the model complexity for MSET is the number of memory vectors, which are sampled from the original training data and used for making similarity comparisons in the training subspace—the more training vectors, the more computational cost. The analogous parameter for the LSTM is the number of hidden layers in the network.

The computational cost of the training process for both methods on three specific test cases is compiled in Table 1. These cases represent the least, most, and median levels of complexity among the anomaly detection applications. The values provide a reasonable indication of the overall trends for the computational cost of MSET and LSTM. It is obvious that MSET is much less computationally expensive, which is expected as it is a deterministic regression type algorithm. Finally, the difference in training compute cost is expected to scale with the dimensionality of the problem.

Table 1: Computational cost of training for MSET and LSTM on selected test cases representing the least, most, and median levels of complexity in anomaly detection applications

	Case 1	Case 7	Case 15
MSET	1.41s	1.59s	2.62s
LSTM	22.0s	37.0s	90.0s

3.2 Case Study with Public Datasets

3.2.1 PSM Data. In this section, we set up a testing framework to evaluate the performance of MSET and LSTM using several published benchmark datasets that have become popular to assess multivariate anomaly detection techniques in the literature. As we stated earlier, we additionally deployed RANSync and DeepAnT in the testing framework. The F1 scores and the associated statistics of each methods are reported in Fig. 4.

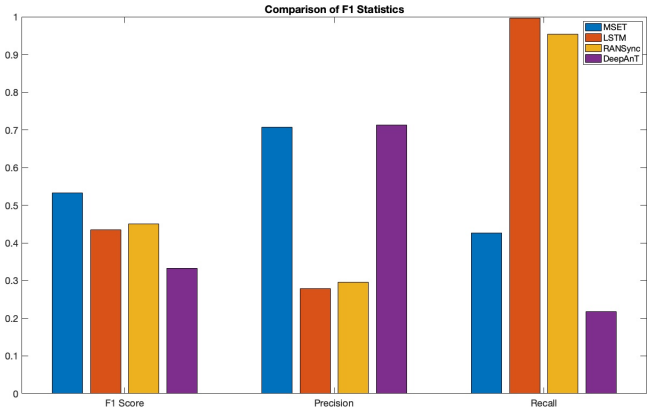


Figure 4: Bar charts of F1 score, Precision, and Recall for MSET (blue), LSTM (red), RANSync (orange), and DeepAnT (purple) models on the PSM dataset.

MSET performs generally better on the F1 score than the other three methods. Additionally, it achieves approximately 1.5 times higher Precision than LSTM and RANSync, but performs worse in Recall by a similar margin. DeepAnT behaves similarly to MSET in the sense of achieving high Precision at the cost of low Recall, yet it is prone to missing more anomalous observations in the fault region than MSET.

While this may seem like a trade-off but in reality the Recall metric is a "red herring". The F1 scores can be easily propped up by the high but imprecise Recall. A detailed discussion is provided in Section 4.

3.2.2 Reconstruction Accuracy. In addition to evaluating the anomaly detection performance, we also assessed the reconstruction accuracy (i.e., accuracy of estimates) of the 4 methods by tracking

the RMSE of the residuals between the test data and the model estimates. For many industry and utility applications, reconstruction performance and anomaly detection decisions are equally informative to the subject matter experts, because accurate model estimates for the signals being monitored provide insights to the early and subtle signs of the degradation and facilitate the troubleshooting process after the anomalies are identified.

The RMSE metrics for the 4 methods on all signals are reported in Fig. 5. The reconstruction quality of MSET is significantly better (i.e. lowest RMSE values) than the other three methods. Note that DeepAnT performs the worst in all PSM signals, which is disproportionate to its anomaly detection performance indicated in Fig. 4. We have investigated the root cause and found that DeepAnT analyzes the multivariate dataset in a univariate fashion, although it is claimed to be a multivariate anomaly detection solution. The fact that it builds a unique model for each signal means it is not leveraging the correlated observations across the entire signal set, resulting in inaccurate estimates.

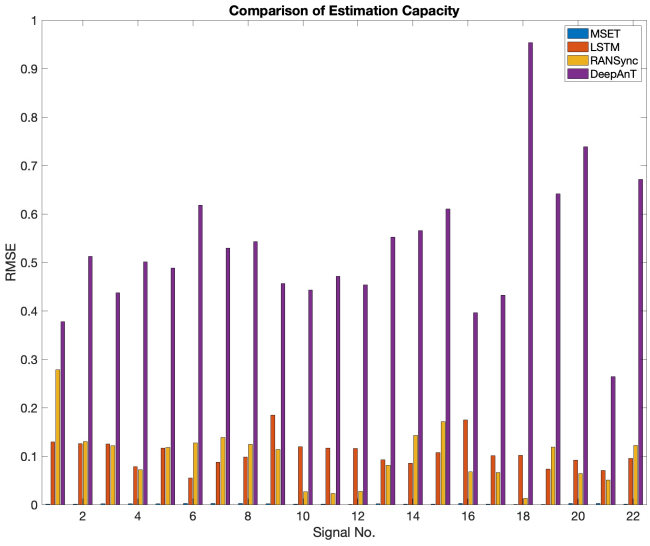


Figure 5: Comparison of the RMSE metrics among MSET (blue), LSTM (red), RANSync (orange), and DeepAnT (purple).

Figure 6 further illustrates how the performance of reconstruction (Fig. 5) correlates with the accuracy of the detection results (Fig. 4). First, as shown in the top subplot in Fig. 6a, the MSET estimates closely align with the actual measurements. Therefore, the mean of the residuals (middle suplot) is near zero and the variance is minimal except for the timestamps where the faults are present (highlighted in cyan), resulting in a low RMSE and accurate signal reconstruction. All faults are correctly identified as they are flagged at the corresponding timestamps (red dots in bottom subplot). The false alarms are present only in the end of the signal (after 8.6e4 where no faults is present). Overall, the false alarm count is low, which boosts the Precision metric.

For LSTM (Fig. 6b), the mean of the residuals starts to deviate from zero and the variance increases. The dynamics of the residuals marginally exhibit the dynamics of the actual measurements (top

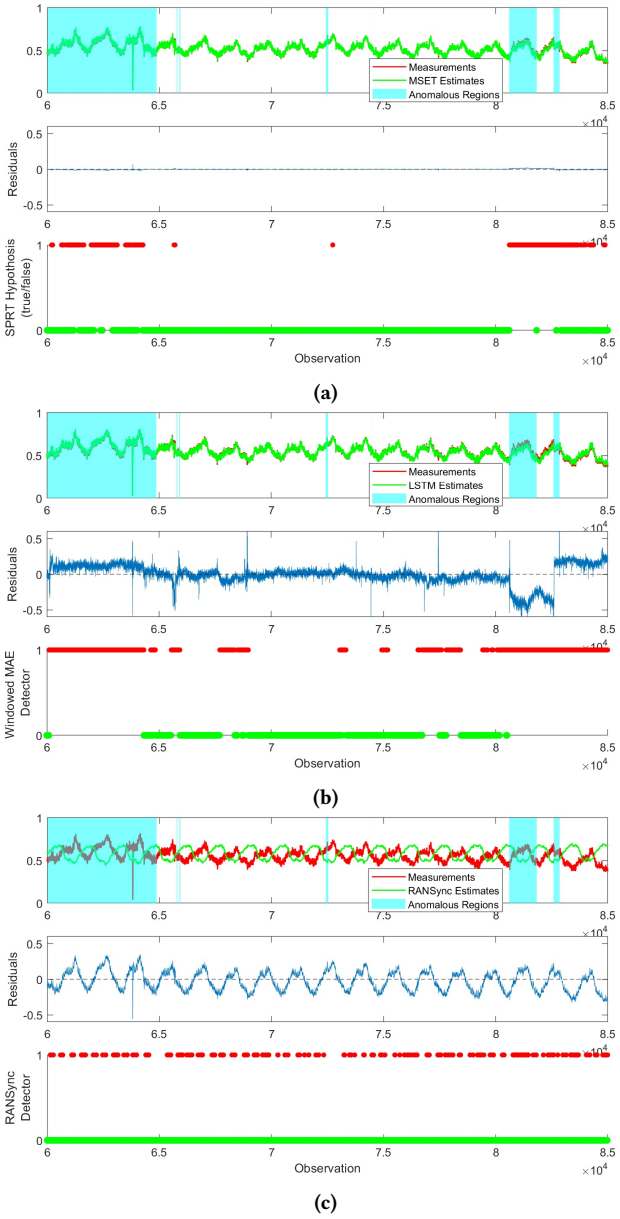


Figure 6: Comparison of anomaly detection results among MSET (a), LSTM (b), and RANSync (c) on the PSM dataset. Cyan highlights show periods of time when anomalies are present. Same figure layout as Fig. 1.

subplot), indicating a less accurate set of estimates from the model. Compared to MSET, LSTM was able to detect all the faults and identified more anomalous observations, as indicated by the denser string of red dots in the bottom subplot. However, it also generated significantly more false alarms, such as around 6.7e4 and 7.6e4, which significantly boots Recall while largely lowers Precision.

Lastly, for RANSync (Fig. 6c), the reconstruction accuracy becomes much worse with the residuals closely resembling the sinusoidal pattern of the test signal (red in the top subplot). As a result,

more anomalous observations were missed in the fault regions, and more false alarms were present throughout the signal. Note that the red dots are visually dense in the anomaly-free regions suggesting high false alarm count, but they are more sporadic than they appear to be due to the high density of observations in the plot. In fact, the number of false alarms is approximately the same as that of LSTM, as indicated by the density of green dots. Nevertheless, compared to MSET and LSTM, RANSync missed more anomalous observations, resulting in the worst Recall among the models, though its Precision is comparable to that of LSTM.

As demonstrated with Figs. 5 and 6, MSET exhibits superior reconstruction performance when compared to the competing methods, which correlates with higher Precision and better overall anomaly detection performance. Moreover, in this case, the high Recall for LSTM and RANSync benefits from the dispersed faults and frequent flagging. Conversely, an inferior anomaly detection technique that excessively flags alarms can also achieve a great Recall score when faults are sporadically distributed. Therefore, high Recall without high precision is not meaningful in assessing the anomaly detection performance. MSET achieves a favorable balance between Precision and Recall, and ultimately yields the highest F1 score among the four methods evaluated in this use case.

3.2.3 Computational Cost. Lastly, the computational cost of the training process for all 4 methods is tabulated in Table 2. MSET outperforms the other methods by orders of magnitude owing to its deterministic mathematical algorithm.

Table 2: Training computational cost on the PSM dataset for MSET, LSTM, RANSync, and DeepAnT methods

MSET	LSTM	RANSync	DeepAnT
0 min 3 sec	169 min 2 sec	96 min 54 sec	325 min 12 sec

3.3 Benchmarking over More Published Datasets

3.3.1 SWaT and WADI Datasets. We further validate MSET against LSTM and five state-of-the-art anomaly detection algorithms on two published datasets. Table 3 presents the performance metrics for MSET and other 6 competing methods on the SWaT and WaDi datasets. Note that in this table, the algorithms other than MSET and LSTM are proprietary to private firms or public universities, which limited our access to their source code. Thus we only evaluated MSET and LSTM on the two datasets, while the performance metrics for the remaining algorithms were cited from a benchmark study [4] that utilized the same datasets.

Overall, MSET ranks the 4th and 2nd out of 7 in term of the F1 score on the SWaT and WaDi datasets, respectively. We further investigated the difference in the anomaly detection performance on the two datasets. For the SWaT dataset, the testing data contains 41 unique anomalies. But there is one particular anomaly that is a large and sudden mean shift, resembling a long lasting step change (spans thousands of consecutive anomalous timestamps), which constitutes 62.68% of all anomalous timestamps, thereby dominating the remaining 40 anomalies. Because of that, the F1 score can become biased in favor of algorithms that accurately flag anomalous timestamps associated with that dominant anomaly, thereby significantly boosting Precision while incurring only marginal losses in Recall. As demonstrated in Fig. 7, we can improve the F1 score of LSTM by up to 47.92% by fine-tuning its anomaly detector to label as many anomalous observations from the dominant anomaly as possible, at the expense of ignoring many of other subtle and short-lived anomalies (refer to the red dots in Fig. 7b), which is misleading. In fact, this is a well-recognized issue referenced in the anomaly detection domain. The lack of a weighting function in the Recall metric—one that differentiates between the number of distinct anomaly events detected and the total count of anomalous timestamps—limits the effectiveness of the F1 score as a general-purpose performance metric for time series anomaly detection.

Without explicit knowledge of the undisclosed parameter tuning process for these algorithms, it is difficult to determine whether whether the comparison on the SWaT dataset was conducted on an equitable basis. By contrast, the WADI dataset does not contain any dominant anomalies that deviate significantly from the normal statistical operating range. Instead, it contains many sporadic subtle anomalies that are inherently more challenging to detect. Consequently, the performance metrics for all algorithms are notably lower, and in this case, MSET ranks as the second-best performer, while two of the previous top performers THOC and OmniAnomaly struggle to effectively detect the anomalies.

4 Discussions

4.1 Applications for MSET

MSET is a multivariate anomaly detection technique that is designed for correlated signals. The variety of the competing methods presented in this study can generally produce reasonable anomaly detection results on a given multivariate dataset. However, MSET demonstrates advantages over them on low false alarms and accurate reconstruction in the multivariate datasets consisting of correlated signals, and the gap between the performance is expected to expand as the number of correlated signals increases.

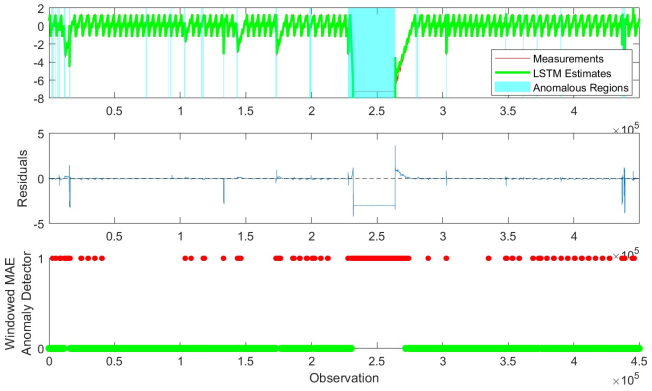
Multivariate datasets comprised of correlated signals are becoming commonly available in the recent years, with the widespread adoption of dense-sensor across multiple IoT industries including utilities, Oil&Gas, manufacturing, commercial aviation, and enterprise IT assets in data centers. These industries are ideal for MSET, as assets in these industries typically have a large number of sensors installed for prognostic monitoring and predictive maintenance so the telemetry signals of these assets are intrinsically correlated. For use cases that instead just contain a collection set of uncorrelated signals, the common deep learning based techniques like LSTM and the more recent techniques introduced in Section 2.4 should be considered, and in some cases, a univariate anomaly detection technique would be a better choice.

4.2 Compute Cost

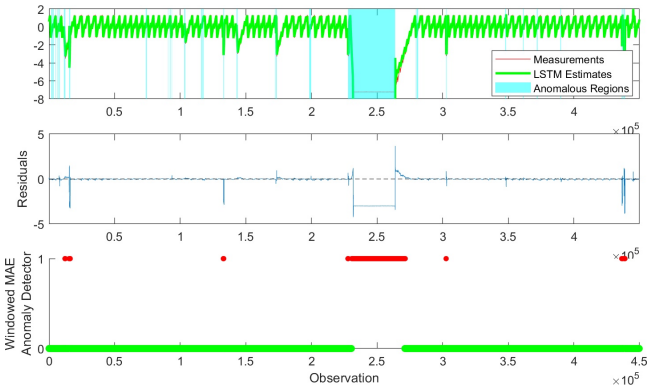
As we discover in this benchmark study, MSET outperforms all the competitors in computational cost by order of magnitudes, thanks to its deterministic mathematical structure, and also its pattern recognition methodology requires much less training data to characterize the behavior of the correlated signals.

Table 3: Performance metrics of MSET and other competing methods on SWaT and WADI Datasets

Dataset	Performance Metric	MSET	LSTM	OmniAnomaly	THOC	TS2Vec	CARLA	DC Detector
SWaT	Precision	0.5024	0.3689	0.9068	0.5453	0.1535	0.9886	0.1214
	Recall	0.7990	0.7647	0.6582	0.7688	0.8742	0.5673	0.9999
	F1	0.6169	0.4977	0.7628	0.6380	0.2611	0.7209	0.2166
WADI	Precision	0.1616	0.1411	0.1315	0.1017	0.0662	0.1850	0.1417
	Recall	0.5750	0.3782	0.8675	0.3507	0.9287	0.7316	0.9684
	F1	0.2523	0.2055	0.2284	0.1577	0.1237	0.2953	0.2472



(a)



(b)

Figure 7: Comparison of anomaly detection results for LSTM before (a) and after (b) tuning the anomaly detector to focus primarily on the dominant anomaly. Despite a 48% increase in the F1 score following the tuning, the number of distinct detected anomalous events significantly decreases.

One related computational cost issue for LSTM and other NN based algorithms is that, it is often too computational expensive for them to scale to a moderate use case in real world applications. Major Cloud providers that deploy LSTM in their anomaly detection offerings typically limit the number of variables to 300[5]. By contrast, MSET is capable of handling thousands of sensor signals or more. In this benchmark study, we intentionally kept the dataset size small to favor NN based algorithms, allowing their anomaly detection performance to be evaluated within a reasonable time frame. If the problem scale were increased to a typical industrial or utility use case—often involving 500 or more signals—iterative comparison and tuning of such models would become impractical.

[5]e.g., as per Azure's service doc for their Anomaly Detection product, https://learn.microsoft.com/en-us/azure/ai-services/anomaly-detector/concepts/best-practices-multivariate

4.3 Value Proposition of MSET

The value proposition of MSET deserves further discussion. Besides the detection performance with correlated signals and minimal computational cost discussed in the previous sections, a key advantage that MSET possesses over other NN based methods is the explainability, which is favored in many safety critical industries. Specifically, MSET excels in discerning the faults on an individual signal basis, which allows it to disambiguate between sensor failure and asset failure. As shown in the example in Section 2.2 (Figs. 1 and 2) where there are three faults appearing in different sensors, MSET is able to generate the anomaly alerts on the faulty sensors only, while LSTM tends to "overreact" by producing excessive alarms during the same time period across multiple signals that do not contain faults. Such false alarms can misleadingly suggest an asset-wide failure, leading to incorrect conclusions regarding the root cause. Furthermore, MSET leverages the similarity among sensor readings, and its prediction process is inherently reversible—reapplying the training observations yields identical anomaly decisions. This deterministic behavior is particularly valuable for root cause analysis, as it enables precise traceback in the context of asset failures. LSTM and other DNN algorithms do not always provide identical anomaly detection decisions despite identical input due to the stochastic optimization mechanism, and also because the neuron coverage depends on the training input and the initial weight configurations, making their behavior less deterministic and less suitable for reproducible root cause analysis. To achieve the same level of definiteness as MSET, those methods would need extraordinarily large amount of training data for the models to converge to the same point, which is often computationally prohibitive. Furthermore, depending on the number of neurons required to accurately capture complex signals dynamics, backwards identification of the neural pathways can become an intractable process, often resistant to effective analysis or interpretation. Overall, although the explainability is not a quantifiable performance metric, it constitutes a value proposition of MSET.

4.4 Future Research Work

With our the findings in this study, we have identified several directions for future research on MSET. Compared to the MAE windowing approach, SPRT appears to be less sensitive to faults

in scenarios where the variance of the normal timestamps is small, resulting in reduced Recall. The future effort will focus on enhancing the sensitivity of SPRT to faults under these circumstances by incorporating temporal changes in L1 norm (MAE) or L2 norm (RMSE) over time into the SPRT decision making process, which is expected to improve Recall and overall detection performance.

Other avenues of research are related to the selection of memory vectors used to train MSET. This parameter determines the similarity metric that ultimately determines the accuracy of the reconstruction. The current method for selecting the training vector values is optimized for finding the topological boundaries of the data but not necessarily for capturing the most representative samples of the training data. Thus, changes in memory vector selection do not exhibit a directly proportional relationship with anomaly detection performance. We will leverage eigenvalue decomposition to identify the most representative values in the dataset and use the spectral radius to determine an approximate eigenvalue equivalence between the memory vector subspace and the training data. Similar processes have been used in finite element analysis and graph theory. A more robust memory vector selection process is expected to handle the complex training data more efficiently although it will likely increase the computation cost.

5 Conclusions

Using machine learning for detecting anomalies in time-series data has received extensive attention. Although a wide range of anomaly detection techniques are available, no single algorithm has been shown to generalize effectively across all time-series datasets. In this paper, a quantitative assessment of the anomaly detection algorithm MSET and other competing methods in both academia and industry, is provided using a variety of benchmark datasets including synthetic datasets and real world datasets. The benchmark results have shown that MSET possesses advantages to the other algorithms on the multivariate applications that involve correlated signals and large-scale datasets.

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] Chuadhry Mujeeb Ahmed, Venkata Reddy Palleti, and Aditya P Mathur. 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*. 25–28.

[3] Mohammad Braei and Sebastian Wagner. 2020. Anomaly detection in univariate time-series: A survey on the state-of-the-art. *arXiv preprint arXiv:2004.00433* (2020).

[4] Zahra Zamanzadeh Darban, Geoffrey I Webb, Shirui Pan, and Mahsa Salehi. 2023. CARLA: A Self-supervised Contrastive Representation Learning Approach for Time Series Anomaly Detection. *arXiv preprint arXiv:2308.09296* (2023).

[5] Jonathan Goh, Sridhar Adepu, Khurum Nazir Junejo, and Aditya Mathur. 2016. A dataset to support research in the design of secure water treatment systems. In *International conference on critical information infrastructures security*. Springer, 88–99.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[7] KC Gross, RM Singer, SW Wegerich, JP Herzog, R VanAlstine, and F Bockhorst. 1997. *Application of a model-based fault detection system to nuclear plant signals*. Technical Report. Argonne National Lab., IL (United States).

[8] Kenny Gross and Guang Chao Wang. 2019. AI Decision Support Prognostics for IoT Asset Health Monitoring, Failure Prediction, Time to Failure. In *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 244–248.

[9] Kenny C Gross and Mengying Li. 2017. Method for Improved IoT Prognostics and Improved Prognostic Cyber Security for Enterprise Computing Systems. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*. The Steering Committee of The World Congress in Computer Science, Computer …, 328–334.

[10] Kenny C Gross and Wendy Lu. 2002. Early Detection of Signal and Process Anomalies in Enterprise Computing Systems.. In *ICMLA*. 204–210.

[11] 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.

[12] R Kozma, M Kitamura, M Sakuma, and Y Yokoyama. 1994. Anomaly detection by neural network models and statistical time series analysis. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, Vol. 5. IEEE, 3207–3210.

[13] Kwei-Harng Lai, Daochen Zha, Guanchu Wang, Junjie Xu, Yue Zhao, Devesh Kumar, Yile Chen, Purav Zumkhawaka, Minyang Wan, Diego Martinez, and Xia Hu. 2020. TODS: An Automated Time Series Outlier Detection System. arXiv:2009.09822 [cs.DB]

[14] Kwei-Herng Lai, Daochen Zha, Junjie Xu, Yue Zhao, Guanchu Wang, and Xia Hu. 2021. Revisiting Time Series Outlier Detection: Definitions and Benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. https://openreview.net/forum?id=r8IvOsnHchr

[15] Ming-Chang Lee, Jia-Chun Lin, and Ernst Gunnar Gran. 2020. RePAD: real-time proactive anomaly detection for time series. *arXiv preprint arXiv:2001.08922* (2020).

[16] Dan Li, Dacheng Chen, Baihong Jin, Lei Shi, Jonathan Goh, and See-Kiong Ng. 2019. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks. In *International Conference on Artificial Neural Networks*. Springer, 703–716.

[17] Junshui Ma and Simon Perkins. 2003. Time-series novelty detection using one-class support vector machines. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, Vol. 3. IEEE, 1741–1745.

[18] Mohsin Munir, Shoaib Ahmed Siddiqui, Andreas Dengel, and Sheraz Ahmed. 2019. DeepAnT: A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series. *IEEE Access* 7 (2019), 1991–2005. doi:10.1109/ACCESS.2018.2886457

[19] Kamran Shaukat, Talha Mahboob Alam, Suhuai Luo, Shakir Shabbir, Ibrahim A Hameed, Jiaming Li, Syed Konain Abbas, and Umair Javed. 2021. A review of time-series anomaly detection techniques: A step to future perspectives. In *Future of Information and Communication Conference*. Springer, 865–877.

[20] Lifeng Shen, Zhuocong Li, and James Kwok. 2020. Timeseries anomaly detection using temporal hierarchical one-class network. *Advances in Neural Information Processing Systems* 33 (2020), 13016–13026.

[21] Ralph M Singer, Kenny C Gross, James P Herzog, Ronald W King, and Stephen Wegerich. 1997. *Model-based nuclear power plant monitoring and fault detection: Theoretical foundations*. Technical Report. Argonne National Lab., IL (United States).

[22] Akash Singh. 2017. Anomaly detection for temporal data using long short-term memory (lstm).

[23] 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.

[24] Abraham Wald. 2004. *Sequential analysis*. Courier Corporation.

[25] GC Wang and KC Gross. 2018. Telemetry parameter synthesis system for enhanced tuning and validation of machine learning algorithmics. In *IEEE 2018 Intn'l Symposium on Internet of Things & Internet of Everything (CSCI-ISOT)*.

[26] Yuanyuan Wei, Julian Jang-Jaccard, Fariza Sabrina, Wen Xu, Seyit Camtepe, and Aeryn Dunmore. 2023. Reconstruction-based LSTM-Autoencoder for Anomaly-based DDoS Attack Detection over Multivariate Time-Series Data. arXiv:2305.09475 [cs.CR]

[27] 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).

[28] 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).

[29] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. 2022. TS2Vec: Towards Universal Representation of Time Series. arXiv:2106.10466 [cs.LG]

[30] Zhenyu Zhong, Qiliang Fan, Jiacheng Zhang, Minghua Ma, Shenglin Zhang, Yongqian Sun, Qingwei Lin, Yuzhi Zhang, and Dan Pei. 2023. A Survey of Time Series Anomaly Detection Methods in the AIOps Domain. arXiv:2308.00393 [cs.LG]