# Supplementary Information: An Evaluation of Anomaly Detection and Diagnosis in Multivariate Time Series

Astha Garg, Wenyu Zhang, Jules Samaran, Ramasamy Savitha, *Senior Member, IEEE,* and Chuan-Sheng Foo

## S1 Taxonomy of algorithms evaluated

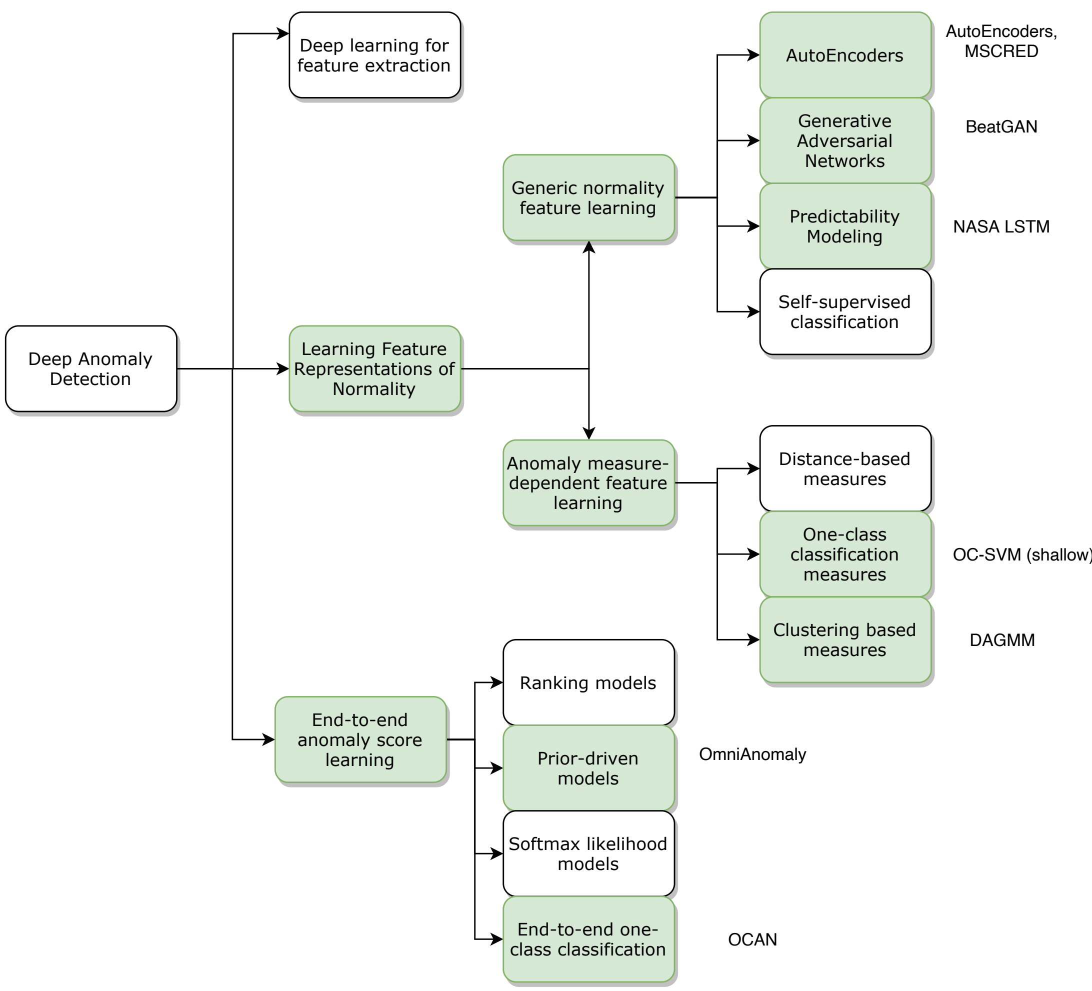



Fig. S1: Anomaly detection algorithms tested in this work (shaded green), based on the taxonomy proposed by [1] for deep anomaly detection. While we do include OC-SVM in the schematic, it is not a deep method. In addition to those shown in the diagram, we also test Raw Signal and PCA, which are not deep methods.

Not all techniques in this taxonomy are suitable for unsupervised or semi-supervised multivariate time-series (MVTS) anomaly detection in a streaming scenario. Based on the discussion on each category in [1]:

- *Deep learning for feature extraction* techniques rely on sophisticated feature extraction techniques which were developed primarily for images, and may not be suitable for MVTS.

- *Self-supervised classification* techniques rely on data augmentation techniques that have been developed primarily for image data. We are not aware of any existing deep algorithms of this type for MVTS.
- *Distance-based measures* are computationally intensive at test-time, and hence not suitable for streaming test setting that we work in.
- *Ranking models* require some form of labelled anomalies which we do not assume.
- *Softmax likelihood models* have been developed primarily for heterogeneous data sources or categorical data. We are not aware of any existing deep algorithms of this type for MVTS.

## S2 Datasets

The datasets contain data from sensors and actuators which interact with each other and the environment in intelligent and stochastic ways. Due to the dynamic nature of the systems we consider, induced anomalies might only become evident after some delay (sometimes several minutes), and the channel behaviour may not return to normal even after the anomaly inducing mechanism has been withdrawn [2].

### S2.1 SWaT

The Secure Water Treatment dataset [2] can be requested on the iTrust website [3]. The dataset has 51 channels, which consist of sensors such as flow meters, level transmitters, conductivity analyzer and actuators such as motorized valves and pumps. The dataset has 14 channels corresponding to signals from various pumps that are constant in train but these are retained as they are allowed to change during testing. The initial 7 days of data consist of normal operation (training set) while 36 attacks were launched in the last 4 days (test set). Each attack compromises one or more channel(s). One of the attacks was much longer (598 mins) than all others (under 30 mins each). This biases the scores of all algorithms depending on whether this event was detected, so we cut this event in the test set to 550 s (the average event length) by discarding anomalous time-points for the rest of the event. Note that since 2 of the attacks were launched right one after the other, we treat it as a single anomalous sequence, and hence consider 35 anomalous events. Of these, root cause labels are available for 33 attacks while the remaining attacks have discrepancies in their root cause labels. For one of the attacks, the start and end points did not match any events listed in the 'List of Attacks' sheet provided on the dataset website [3]. For another attack, the provided root cause does not match any of the available channels.

### S2.2 WADI

The Water Distribution testbed [4] is an operational, scaled-down version of a water distribution network in a city. It is connected to the SWaT plant and takes in a portion of its reverse-osmosis output. The distribution network consists of 3 distinct control processes, namely - primary grid, secondary grid and return water grid - each controlled by its own set of Programmable Logic Controllers (PLCs). The dataset, also hosted by iTrust [3] consists of data from 123 sensors and actuators collected over 14 days for normal operation (training set) and 2 days with 15 attacks (test set). Since 2 attacks were launched at the same time, we consider 14 anomalous events in this paper. We use all 14 events for anomaly detection, but for anomaly diagnosis, we have root cause labels for only 12 of the 14 events. For the remaining 2 events, the labels do not specify exactly which component(s) was compromised.

### S2.3 DMDS

The DAMADICS (Development and Application of Methods for Actuator Diagnosis in Industrial Control Systems) benchmark [5] consists of real process data from the Lublin Sugar Factory as well as a simulator to generate artificial faults. Here we use only the real dataset with induced faults in the industrial system, available publicly [6]. The dataset consists of data for 25 days of operation, from Oct-29 to Nov-22, 2001 of 3 benchmark actuators - one each located upstream and downstream of evaporator station, and the third controlling flow of water to the steam boiler system. Artificial faults were induced on Oct-30, Nov-9, Nov-17 and Nov-20. Unlike SWaT and WADI, the train and test splits for normal and test operation were not provided. We chose train-test splits such that the test is entirely after train as would be expected in a real scenario, the train is continuous, and the train contains no anomalies. Accordingly, we used the data from Nov-3 to Nov-8 (6 days) as the training set and data from Nov-9, Nov-17 and Nov-20 (3 days) as the test set. From this, the first 10800 points from the training set were dropped as the system appeared much more unstable than the rest of the training set, potentially from the anomaly induced earlier on Oct-30. In addition, the initial part of the test set appears quite unstable across multiple channels even though no anomaly is recorded. Therefore we also drop the first 45000 points from the test set.

### S2.4 SKAB

The Skoltech Anomaly Benchmark [7] testbed consists of a water circulation system and its control system, along with a data-processing and storage system. Examples of the type of anomalies induced include partial valve closures, connecting shaft imbalance, reduced motor power, cavitation and flow disturbances. Train and test splits are provided by the authors.

### S2.5 MSL and SMAP

These are expert-labeled datasets from real anomalies encountered during the operation of two spacecraft - Soil Moisture Active Passive (SMAP) satellite and the Mars Science Laboratory (MSL) rover, Curiosity [8]. Unlike the other datasets, each entity in MSL and SMAP consists of only 1 sensor, while all the other channels are one-hot-encoded commands given to that entity. We use all channels as input to the model, but the model error of only the sensor channel is used for anomaly detection, as done by [8]. The total number of variables is 1375 and 1485 for SMAP and MSL respectively, making these much larger than the single-entity datasets in terms of number of variables. The data is however divided into 55 and 27 entities respectively. The authors provide train-test splits so that for the first anomaly encountered in test at time t, the training set is from time t-5 days to t-3 days (if available), and the test set goes from t-3 days to t+2 days. The data is sampled each minute, so the training set is much shorter than other datasets.

### S2.6 SMD

SMD, or Server Machine Dataset was published by [9] on their Github repository https://github.com/NetManAIOps/OmniAnomaly. The data was collected over 5 weeks from a large internet company. It consists of data from 28 entities regularly sampled every minute. The train-test split is 50% each for train and test, suggested by the authors. An interpretation label is provided for each anomaly which we use for anomaly diagnosis.

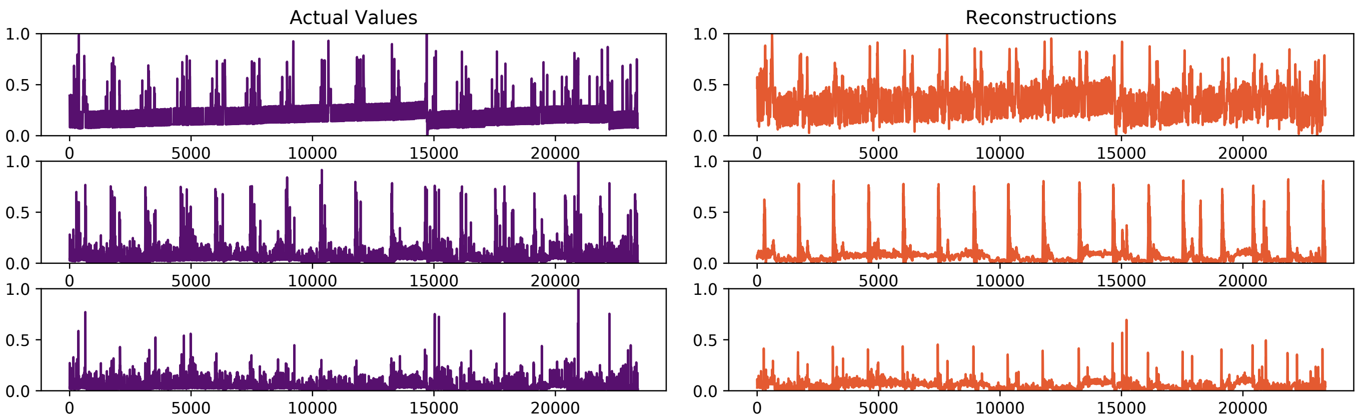



Fig. S2: Left: Signals from three channels from the SMD dataset, showing strong correlations across channels and time. Right: Reconstructions of the three channels by Univariate AutoEncoder, accounting for only temporal correlations.

## S3 Additional details about models

**Univariate Fully-Connected Auto-Encoder (UAE)**: In this method, we train a separate auto-encoder for each channel. Each encoder is a multi-layer perceptron with $l_w$ nodes in the input to $p$ dimensions in the latent space, with the number of dimensions reducing in powers of 2 in each layer, similar to [10]. The decoder is a mirror image of the encoder, and we use tanh activation.

**Fully-Connected Auto-Encoder (FC AE)**: This is similar to UAE, but now we have a single model over all the channels. Thus, the input sample is a flattened subsequence, a vector of size $l_w \times m$.

**Long Short Term Memory Auto-Encoder (LSTM AE)**: We use a single LSTM layer for each of the encoder and decoder. [11] used only the first principal component (PC) of the MVTS as input to LSTM-ED, but since this can result in major information loss, we choose the number of PCs corresponding to 90% explained variance. We set the hidden size to be the same size as the number of PCs.

## S4 Hyperparameters and Implementation

In some algorithms, we set the architecture size as a function of $m$ to accommodate different datasets, eg. the hidden size. All other hyperparameters are kept constant across datasets, obtained by hyperparameter tuning on the SWaT dataset only. All the deep learning methods except OmniAnomaly and OCAN are trained for a maximum of 100 epochs with early stopping using the reconstruction or prediction error on the 25% held-out validation set and patience = 10.

TABLE S1: Online code references

| Model | Adapted from |
|---|---|
| LSTM VAE | https://github.com/TimyadNyda/Variational-Lstm-Autoencoder |
| NASA LSTM | https://github.com/khundman/telemanom |
| DAGMM | https://github.com/danieltan07/dagmm |
| OmniAnomaly | https://github.com/NetManAIOps/OmniAnomaly |
| MSCRED | https://github.com/Zhang-Zhi-Jie/Pytorch-MSCRED, https://github.com/SKvtun/MSCRED-Pytorch |
| OCAN | https://github.com/PanpanZheng/OCAN |
| BeatGAN | https://github.com/Vniex/BeatGAN |

TABLE S2: Hyperparameters are tuned using the minimum validation reconstruction error criterion, based only on the SWaT dataset. We tested 50 random hyperparameter configurations for each of TCN AE, LSTM VAE and OC-SVM, and 20 random configurations for FC AE. The chosen hyperparameter is shown in the 'Value' column. $m$ is the number of channels in each entity of a dataset.

| Model | Hyperparameters | Sampling distribution | Value |
|---|---|---|---|
| FC AE | learning rate | log-unif(log($10^{-4}$), log($10^{-3}$)) | $10^{-4}$ |
| | z-dim | int of $\{\frac{m}{5}, \frac{m}{4}, \frac{m}{3}, \frac{m}{2}, \frac{m}{1}, 2m, 3m, 4m, 5m\}$ | $int(\frac{m}{2})$ |
| TCN AE | learning rate | log-unif(log($10^{-4}$), log($10^{-2}$)) | $1.5 \times 10^{-4}$ |
| | z-dim | int $\in [3, 10]$ | 8 |
| | dropout rate | unif(0.2, 0.5) | 0.42 |
| LSTM VAE | learning rate | log-unif(log($10^{-4}$), log($10^{-2}$)) | $9.5 \times 10^{-3}$ |
| | (z-dim, hidden-dim) | $\{(3, 15), (3, 5)\}$ | (3, 15) |
| | $\lambda_{reg}$ | unif(0,1) | 0.55 |
| | $\lambda_{kulback}$ | unif(0,1) | 0.28 |
| OC-SVM | $\gamma$ | $\left\{\frac{1}{\#features}, \text{uniform}(10^{-5}, 10^{5})\right\}$ | $\frac{1}{\#\text{features}}$ |
| | $\nu$ | unif(0,0.5) | 0.489 |

TABLE S3: Key hyperparameter values used for each models. $m$ is the number of channels in an entity of the dataset, LR is the learning rate, p refers to the hidden size. * Learning rate annealing was used as per [9]

| Model | Epochs | LR | Batch size | Design | Framework |
|---|---|---|---|---|---|
| PCA | - | - | - | $n_{PCA,0.9}$=components for explained variance=0.9 | Scikit-learn |
| OC-SVM | - | - | - | $\gamma$=1/m, $\nu$=0.489 | |
| UAE | 100 | 0.001 | 256 | p=5 | Pytorch |
| FC AE | 100 | 0.0001 | 128 | p=int(m/2) | Pytorch |
| LSTM AE | 100 | 0.001 | 64 | PCA before model with $n_{PCA,0.9}$, hidden-layers=3 | Pytorch |
| TCN AE | 100 | $1.5 \times 10^{-4}$ | 128 | dropout=0.42, p=3, hidden-layers=min(10, int(m/6)), kernel-size=5 | Pytorch |
| LSTM VAE | 100 | $9.5 \times 10^{-3}$ | 128 | Hidden layers=2, hidden-dim=15, z-dim=3, $\lambda_{reg}$=0.55, $\lambda_{kulback}$=0.28 | Tensorflow |
| BeatGAN | 100 | $10^{-4}$ | 128 | z-dim=10, beta1=0.5 | Pytorch |
| MSCRED | 100 | $10^{-4}$ | 128 | As in [12]. For WADI, PCA before model with $n_{PCA,0.9}$ | Pytorch |
| DAGMM | 200 | $10^{-4}$ | 128 | z-dim=m, $GMM_k$=3, $\lambda_{energy}$=0.1, $\lambda_{cov}$=0.005 | Pytorch |
| NASA LSTM NPT | 100 | $10^{-3}$ | 64 | LSTM-units=m, LSTM-layers=2, dropout=0.3, NPT params from [8] | Keras |
| NASA LSTM | 100 | $10^{-3}$ | 64 | LSTM units=m, LSTM layers=2, dropout=0.3 | Keras |
| OmniAnomaly | 20 | 0.001* | 50 | All params as in [9]: z-dim=3, RNN-units=500, dense-units=500, NF-layers=20 | Tensorflow |
| OCAN | 100 | $10^{-4}$ | 128 | As in [13] | Tensorflow |

TABLE S4: Training time for deep learning algorithms in minutes for the single-entity datasets (with early stopping), trained on a single Nvidia GeForce RTX 2080 Ti GPU. UAE, the top performing model in the evaluation, is the third slowest. The slow speed of UAE training is because we trained the channel-wise models sequentially but since each model is independent, it is easily parallelizable. On the other hand, FC AE the second best model, is the fastest to train.

| | DAMADICS | SWaT | WADI |
|---|---|---|---|
| BeatGAN | 9 | 10 | 19 |
| DAGMM | 37 | 45 | 115 |
| FC AE | 6 | 11 | 2 |
| LSTM AE | 23 | 116 | 142 |
| LSTM VAE | 98 | 93 | 77 |
| MSCRED | 221 | 510 | 330 |
| NASA LSTM | 22 | 25 | 14 |
| OCAN | 41 | 42 | 121 |
| OmniAnomaly | 186 | 165 | 262 |
| TCN AE | 20 | 20 | 38 |
| UAE | 65 | 104 | 268 |

## S5 Comparison with published results

TABLE S5: Point-adjusted $F_1$ score for the MSL, SMAP and SMD datasets with the best-f1 threshold, comparing results reported in the literature against UAE Gauss-D and Random Anomaly Detector (discussed in section 6 in the main text).

| Model | MSL | SMAP | SMD |
|---|---|---|---|
| OmniAnoAlgo (reported by authors) [9] | 0.9014 | 0.8535 | 0.9620 |
| OmniAnoAlgo (reproduced by us) | 0.8459 | 0.8678 | 0.9424 |
| USAD (reported by authors) [10] | 0.9109 | 0.8186 | 0.9382 |
| Random Anomaly Detector | 0.8512 | 0.7418 | 0.7585 |
| UAE Gauss-D | **0.9204** | **0.8961** | **0.9723** |

Here we discuss the results in our paper with other published works on the datasets we use.

**Comparisons on WADI dataset with the point-wise $F_1$ score**: Recently, [10] propose USAD, an adversarially trained auto-encoder model for MVTS anomaly detection, and tested it on 5 of the 6 datasets that we test here. They report a point-wise $F_1$ of 0.2328 with the best-F-score threshold on the WADI dataset. [14] report $F_1$ score of 0.37 on the WADI dataset with the best-F-score threshold using a Generative Adversarial Network, MAD-GAN. We obtain comparable or better scores in this work, shown in Table S10 using the Gauss-D scoring function and Table S11 with the Gauss-D-K scoring function. The best overall algorithm, UAE, attains an $F_1$ score of 0.4740 (an improvement of 46.5% over MAD-GAN) with the Gauss-D-K scoring function, while the best performing algorithm on WADI in this table is LSTM VAE, with an $F_1$ score of 0.5025.

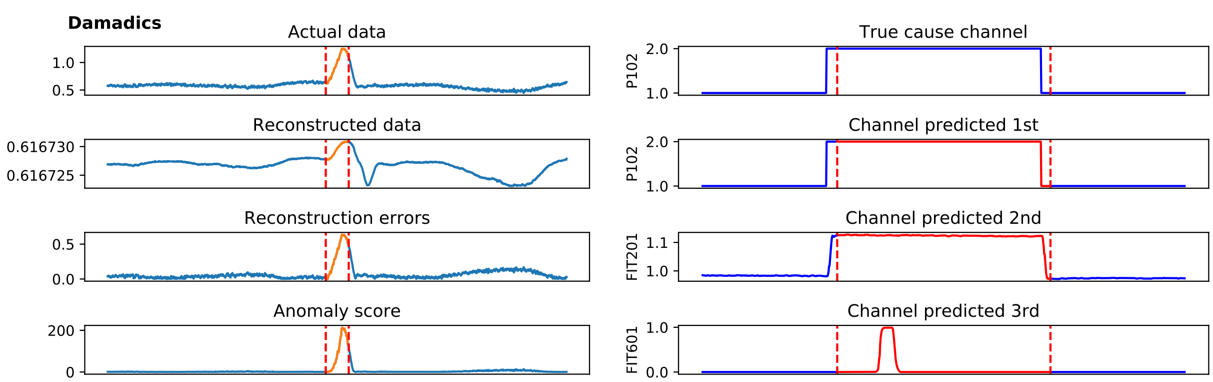
**Comparison with point-adjusted $F_1$ results**: [9] and [10] report results on SMAP, MSL and SMD datasets with the point-adjusted $F_1$ score with the OmniAnomaly and USAD algorithms respectively. Based on the experiments discussed in the main text section 6, we do not use this metric to draw comparisons in this paper. However, for the sake of completeness, we show a comparison of the scores of UAE using Gauss-D threshold vs. the results reported by the authors of OmniAnomaly and USAD in Table S5. Once again, UAE is the top performing algorithm.

Aside from the works discussed above, point-wise $F_1$ score with the best-F-score threshold have been published previously on the SWaT dataset. [14] report a score of 0.77 with the MAD-GAN algorithm, and [10] report a score of 0.79 with the USAD algorithm. However these results are not directly comparable with our results. This is because in our study, we have shortened a long anomaly that biases the results (discussed in section III in the main text) in the SWaT dataset, and as a result our test set is different from these works. We note that some algorithms we tested indeed achieved better $F_1$ score with the best-F-score threshold than the literature methods on SWaT dataset, but these algorithms were not the top-performing by $Fc_1$ score.

## S6 Demonstrations of methods

**Algorithm 1:** MVTS anomaly detection with an auto-encoder model, Gauss-S scoring and tail-p threshold.

**input :** Anomaly-free time-series: $\mathbf{X}_{train} \in \mathbb{R}^{n_1 \times m}$
Test time-series with anomalies: $\mathbf{X}_{test} \in \mathbb{R}^{n_2 \times m}$
Window size $l_w$; step size $l_s$; tail-probability $\epsilon$
**output:** Predicted binary anomaly labels $\hat{y}_t$ for each time point in $n_2$
Channel-wise anomaly scores $\mathbf{A}^1$, ..., $\mathbf{A}^m$ each of size $n_2$
**Step 1: Train FC AE** Hidden size $p$
**for** *epoch in max epochs* **do**
  $Loss \leftarrow 0$
  **for** *Training sub-sequences* $\mathbf{S}_{t,train} \leftarrow \mathbf{X}_{train}[t - l_w + 1, .., t] \in \mathbb{R}^{l_w \times m}$ *with step size* $l_s = 10$ **do**
    Encode sub-sequence to latent space:
    $z^1, .., z^p \leftarrow Encoder(\mathbf{S}_{t,train})$
    Reconstruct original sub-sequence:
    $\hat{\mathbf{S}}_{t,train} \leftarrow Decoder(z^1, .., z^p)$
    $Loss \leftarrow Loss + rms(\mathbf{S}_{t,train} - \hat{\mathbf{S}}_{t,train})$
  **end**
  Update parameters of $Encoder$ and $Decoder$ to minimize $Loss$
**end**
**Step 2: Apply Gauss-S scoring function**
Obtain reconstruction errors from FC AE on train: $\mathbf{E}^i_{t,train} = \mathbf{S}^i_{t,train}[t] - \hat{\mathbf{S}}^i_{t,train}[t]$,for each channel $i \in m$
Fit $\mathbf{E}^i_{train} \sim N(\mu^i, \sigma^i)$ for each channel $i \in m$
Get the reconstruction errors $\mathbf{E}^i_t = \mathbf{S}^i_{t,test}[t] - \hat{\mathbf{S}}^i_{t,test}[t]$ on test data
**for** $t \leftarrow 1$ **to** $n_2$ **do**
  Get probability scores for each channel
  **for** *each channel* $i \leftarrow 1$ **to** $m$ **do**
    $\mathbf{A}^i_t \leftarrow \log(1 - \Phi(\frac{\mathbf{E}^i_t - \mu^i}{\sigma^i}))$, $\Phi$ is the cdf of $N(0, 1)$
  **end**
  $\mathbf{a}_t \leftarrow -\sum_{i=1}^{m} \mathbf{A}^i_t$
**end**
**Step 3:** Apply tail-p threshold
$th_{tail-p} \leftarrow -m \log(\epsilon)$
**prediction** $\leftarrow \mathbf{1}_{(x >= th_{tail-p})}(\mathbf{a})$




(a) Anomaly Detection (b) Anomaly Diagnosis

Fig. S3: Examples of (a) anomaly detection on DAMADICS dataset and (b) anomaly diagnosis on SWaT dataset using UAE model and Gauss-S scoring function.

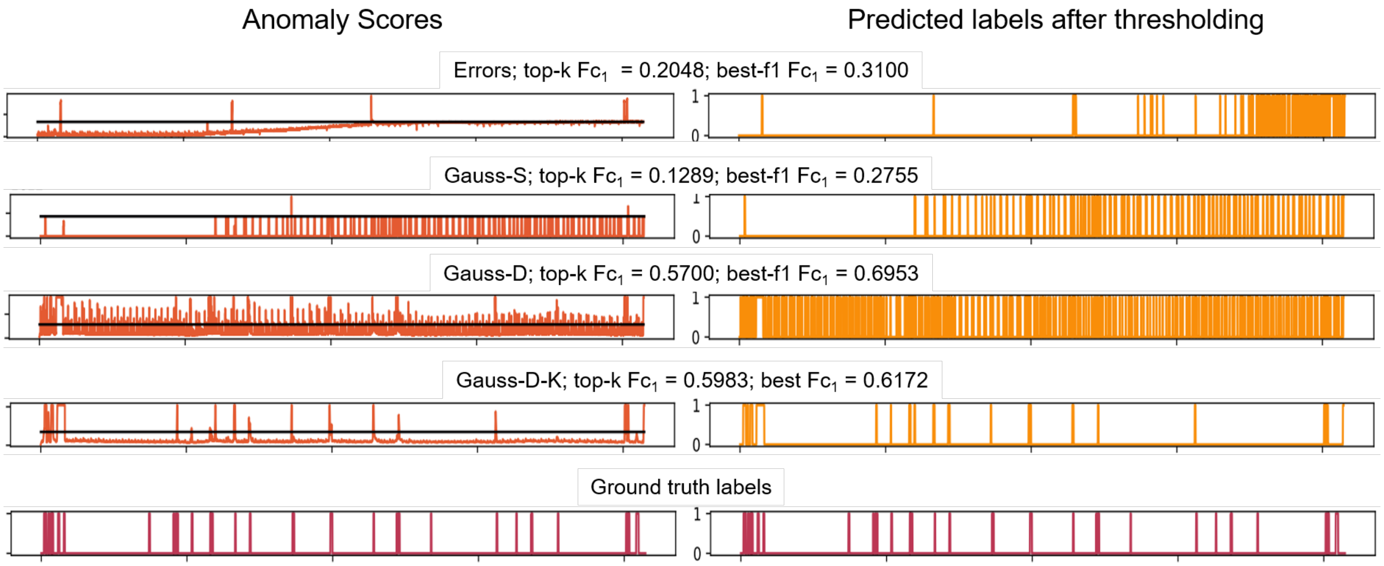



Fig. S4: The effect of scoring function on the anomaly detection performance of UAE for the SWaT dataset. Plots on the left show the scoring function after aggregation across channels, and the solid black line is the top-k threshold. Plots on the right show the anomaly labels for each scoring function. Some anomalies stand out with the Error scoring function but others go undetected. Gauss-D has a much better $Fc_1$ score, but it appears noisy. The Gauss-D-K scoring function does smoothing across time and channels with a Guassian kernel, so the scoring function appears much less noisy.

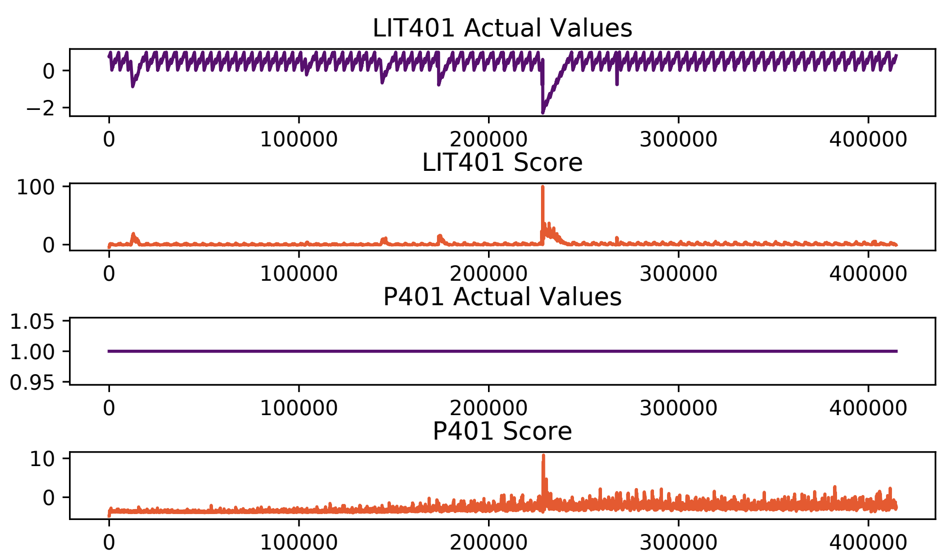



Fig. S5: An example showing spurious correlations learnt by OmniAnomaly on SWaT. Channel P401 does not vary at all during test, but OmniAnomaly's score for channel P401 shows peaks corresponding to an anomaly in LIT 401. The signals are shown from the test set. P401 stays constant in both train and test.

# S7 Statistical Tests

We follow recommendations from [15] for multiple classifier comparisons across multiple datasets. For each comparison, we first use the Friedman test to find whether the overall comparison between k methods on N groups (eg. datasets) is statistically significant. If this comparison is significant, we follow this with post-hoc tests using Hochberg's step-up procedure [16] to compare the best method against all other methods. All tests below are conducted at significance level $\alpha = 0.05$.

## S7.1 Effect of scoring functions on anomaly detection performance

The null hypothesis for Friedman test is that all scoring functions perform the same over N=70 combinations of datasets and models, and k=4 scoring functions. The Friedman-statistic is 57.13 which is larger than the critical value resulting in a p-value of $2.4e-12$, thus we reject the null hypothesis. Post-hoc tests using Hochberg's step-up procedure indicate that the difference between the performance of Gauss-D-K vs. all other methods is statistically significant. Furthermore, the difference between the performance of Gauss-D vs. Gauss-S and Error is statistically significant.

## S7.2 Effect of model on anomaly detection performance

The null hypothesis for Friedman test is that all algorithms in Table IV (main text) perform the same with N=7 datasets and k=13 methods. We find that the Friedman statistic 43.53 is greater than the exact tabled critical value 20.23 [17], resulting in p-value $1.83e-5$ and reject the null hypothesis. Next we compare UAE against other models using post-hoc tests [16] and find that only the comparisons between UAE and the bottom 6 algorithms are statistically significant.

# S8 Additional Results

## S8.1 $Fc_1$ score with top-k threshold

TABLE S6: $Fc_1$ score mean and standard deviation over 5 seeds, with the top-k threshold using the chosen hyperparameters. The scoring function is Gauss-D for all algorithms except those denoted with *. The overall mean is a mean over all the datasets. The performance of Raw Signal and PCA models is deterministic.

| | DMDS | | MSL | | SKAB | | SMAP | | SMD | | SWaT | | WADI | | Mean | Rank |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Algo | Mean | Std | Mean | Std | Mean | Std | Mean | Std | Mean | Std | Mean | Std | Mean | Std | | |
| Raw Signal | 0.4927 | | 0.2453 | | 0.5349 | 0.0000 | 0.2707 | | 0.5151 | | 0.3796 | | 0.4094 | | 0.4068 | 9.3 |
| PCA | 0.5339 | | 0.4067 | | 0.5524 | 0.0000 | 0.3793 | | 0.5344 | | 0.5314 | | 0.3747 | | 0.4733 | 5.6 |
| UAE | **0.6378** | 0.008 | **0.5111** | 0.0085 | **0.5550** | 0.0022 | **0.4793** | 0.0074 | 0.5501 | 0.0046 | **0.5713** | 0.0087 | 0.5105 | 0.01 | **0.5450** | **1.6** |
| FC AE | 0.6047 | 0.005 | 0.4514 | 0.0048 | 0.5408 | 0.0040 | 0.3788 | 0.0056 | 0.5395 | 0.0064 | 0.4478 | 0.0528 | 0.5639 | 0.0193 | 0.5038 | 4.7 |
| LSTM AE | 0.5999 | 0.0141 | 0.4481 | 0.0065 | 0.5418 | 0.0054 | 0.4536 | 0.0109 | 0.5271 | 0.0062 | 0.5163 | 0.0148 | 0.4265 | 0.0057 | 0.5019 | 4.7 |
| TCN AE | 0.5989 | 0.0204 | 0.4354 | 0.0105 | 0.5488 | 0.0041 | 0.3873 | 0.0054 | **0.5800** | 0.0037 | 0.4732 | 0.0114 | 0.5126 | 0.0784 | 0.5052 | 3.9 |
| LSTM VAE | 0.5939 | 0.0085 | 0.3910 | 0.0059 | 0.5439 | 0.0022 | 0.2988 | 0.004 | 0.5427 | 0.0046 | 0.4456 | 0.003 | **0.5758** | 0.0143 | 0.4845 | 6.0 |
| BeatGAN | 0.5391 | 0.1099 | 0.4531 | 0.0075 | 0.5437 | 0.0063 | 0.3732 | 0.0091 | 0.5479 | 0.0099 | 0.4777 | 0.0061 | 0.4908 | 0.0558 | 0.4894 | 5.0 |
| MSCRED | 0.2906 | 0.0129 | 0.3944 | 0.0045 | 0.5526 | 0.0076 | 0.3724 | 0.0062 | 0.4145 | 0.0057 | 0.4315 | 0.0117 | 0.3253 | 0.0033 | 0.3973 | 8.1 |
| NASA LSTM | 0.1284 | 0.0074 | 0.4715 | 0.0124 | 0.5339 | 0.0100 | 0.4280 | 0.0077 | 0.3879 | 0.0036 | 0.1398 | 0.0143 | 0.1058 | 0.0449 | 0.3136 | 8.9 |
| DAGMM* | 0.0000 | 0 | 0.1360 | 0.0188 | 0.0000 | 0.0000 | 0.1681 | 0.0205 | 0.0187 | 0.015 | 0.0000 | 0 | 0.0256 | 0.0573 | 0.0498 | 12.9 |
| OmniAnomaly* | 0.1425 | 0.1189 | 0.4120 | 0.0108 | 0.4561 | 0.0264 | 0.3767 | 0.0094 | 0.5002 | 0.0121 | 0.1466 | 0.0985 | 0.2443 | 0.0202 | 0.3255 | 9.4 |
| OCAN* | 0.2532 | 0.0925 | 0.3009 | 0.0323 | 0.4369 | 0.0271 | 0.2787 | 0.0177 | 0.4614 | 0.0095 | 0.1547 | 0.1502 | 0.0000 | 0 | 0.2694 | 11.0 |

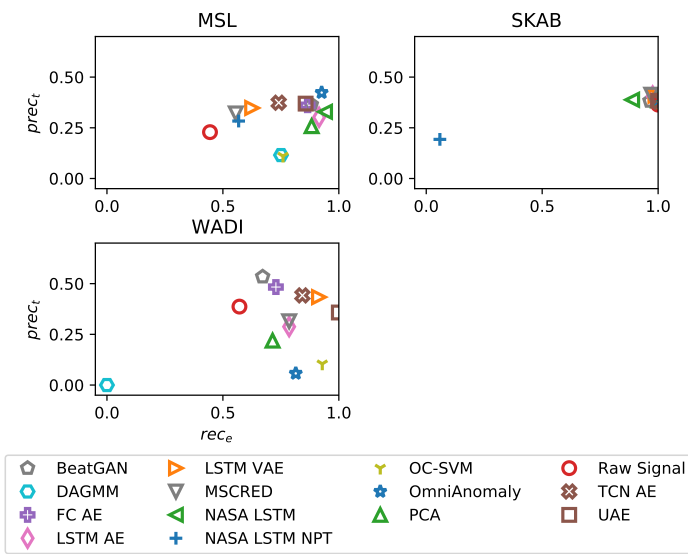



Fig. S6: Plots of $prec_t$ vs. $rec_e$ for algorithms with the top-k threshold and scoring functions as in Table IV in main manuscript. See plots for additional datasets in Fig. 4 in main manuscript.

TABLE S7: $Fc_1$ score of various models with the *Gauss-D-K* scoring function (except the starred algorithms that specify their own scoring functions) with the *top-k* threshold.

| | DMDS | | MSL | | SKAB | | SMAP | | SMD | | SWaT | | WADI | | Mean | Rank |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Algo | Mean | Std | Mean | Std | Mean | Std | Mean | Std | Mean | Std | Mean | Std | Mean | Std | | |
| Raw Signal | 0.4693 | | 0.2974 | | 0.5405 | 0.0000 | 0.3190 | | 0.5591 | | 0.4086 | | 0.4790 | | 0.4390 | 7.9 |
| PCA | 0.5356 | | 0.5217 | | 0.5585 | 0.0000 | 0.4585 | | 0.5557 | | 0.4811 | | 0.3068 | | 0.4883 | 5.4 |
| UAE | **0.6199** | 0.0054 | 0.5193 | 0.021 | 0.5604 | 0.0034 | 0.6116 | 0.0141 | 0.5805 | 0.0041 | **0.6105** | 0.0158 | **0.5561** | 0.0149 | **0.5798** | **2.0** |
| FC AE | 0.6048 | 0.0039 | 0.5060 | 0.0076 | 0.5442 | 0.0042 | 0.5672 | 0.0093 | 0.5651 | 0.0056 | 0.4786 | 0.0475 | 0.5083 | 0.011 | 0.5392 | 4.0 |
| LSTM AE | 0.6029 | 0.0097 | 0.5346 | 0.0103 | 0.5364 | 0.0202 | 0.5605 | 0.0068 | 0.5381 | 0.0032 | 0.4715 | 0.0092 | 0.3505 | 0.005 | 0.5135 | 5.9 |
| TCN AE | 0.6035 | 0.0199 | 0.5231 | 0.0057 | 0.5515 | 0.0059 | 0.5615 | 0.0072 | **0.6034** | 0.0053 | 0.4353 | 0.0158 | 0.4685 | 0.0854 | 0.5353 | 4.0 |
| LSTM VAE | 0.5924 | 0.0075 | 0.4623 | 0.0035 | 0.5433 | 0.0042 | 0.4975 | 0.0046 | 0.5784 | 0.0045 | 0.4444 | 0.001 | 0.5289 | 0.0132 | 0.5210 | 5.3 |
| BeatGAN | 0.5380 | 0.1129 | 0.5329 | 0.0132 | 0.5391 | 0.0095 | 0.5698 | 0.0097 | 0.5550 | 0.0107 | 0.4760 | 0.0266 | 0.4648 | 0.0723 | 0.5251 | 5.3 |
| MSCRED | 0.2960 | 0.0173 | 0.4096 | 0.007 | 0.5502 | 0.0081 | 0.4009 | 0.0073 | 0.4085 | 0.0045 | 0.3769 | 0.0147 | 0.2905 | 0.0216 | 0.3904 | 8.7 |
| NASA LSTM | 0.1276 | 0.008 | **0.5503** | 0.0098 | 0.5338 | 0.0103 | **0.6410** | 0.0131 | 0.3874 | 0.0042 | 0.1348 | 0.0199 | 0.1958 | 0.0446 | 0.3672 | 8.4 |
| DAGMM* | 0.0000 | 0 | 0.1360 | 0.0188 | 0.0000 | 0.0000 | 0.1681 | 0.0205 | 0.0187 | 0.015 | 0.0000 | 0 | 0.0256 | 0.0573 | 0.0498 | 12.9 |
| OmniAnomaly* | 0.1425 | 0.1189 | 0.4120 | 0.0108 | 0.4561 | 0.0264 | 0.3767 | 0.0094 | 0.5002 | 0.0121 | 0.1466 | 0.0985 | 0.2443 | 0.0202 | 0.3255 | 10.1 |
| OCAN* | 0.2532 | 0.0925 | 0.3009 | 0.0323 | 0.4369 | 0.0271 | 0.2787 | 0.0177 | 0.4614 | 0.0095 | 0.1547 | 0.1502 | 0.0000 | 0 | 0.2694 | 11.1 |

## S8.2 $Fc_1$ score with best-F-score threshold

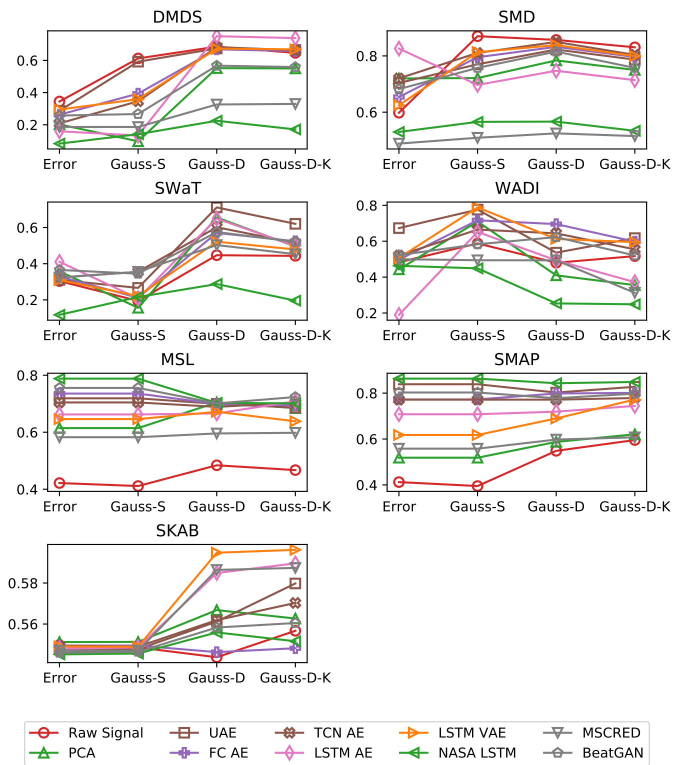



Fig. S7: Effect of scoring functions on the $Fc_1$ score using the best-F-score (best-$Fc_1$) threshold.

TABLE S8: $Fc_1$ score of various models with the *Gauss-D* scoring function (except the starred algorithms that specify their own scoring functions) with the *best-F-score* threshold (best $Fc_1$ in this case).

| Algo | DMDS | | MSL | | SKAB | | SMAP | | SMD | | SWaT | | WADI | | Mean | Rank |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | Std | Mean | Std | Mean | Std | Mean | Std | Mean | Std | Mean | Std | Mean | Std | | |
| Raw Signal | 0.6856 | | 0.4841 | | 0.5438 | 0.0000 | 0.5485 | | **0.8568** | | 0.4469 | | 0.4801 | | 0.5780 | 8.0 |
| PCA | 0.5521 | | 0.7045 | | 0.5668 | 0.0000 | 0.5874 | | 0.7840 | | 0.6556 | | 0.4099 | | 0.6086 | 6.3 |
| UAE | 0.6744 | 0.0083 | 0.6996 | 0.0104 | 0.5612 | 0.0034 | 0.8024 | 0.0154 | 0.8226 | 0.0075 | **0.7112** | 0.0127 | 0.5363 | 0.0185 | 0.6868 | **4.0** |
| FC AE | 0.6687 | 0.0173 | 0.6991 | 0.0078 | 0.5463 | 0.0009 | 0.7984 | 0.0111 | 0.8321 | 0.0026 | 0.5703 | 0.0269 | **0.6955** | 0.037 | **0.6872** | 5.3 |
| LSTM AE | **0.7504** | 0.0246 | 0.6661 | 0.0096 | 0.5849 | 0.0244 | 0.7193 | 0.0108 | 0.7472 | 0.0066 | 0.6493 | 0.0048 | 0.4934 | 0.047 | 0.6587 | 5.4 |
| TCN AE | 0.6824 | 0.0208 | 0.6890 | 0.0127 | 0.5619 | 0.0103 | 0.7716 | 0.0189 | 0.8495 | 0.0038 | 0.6016 | 0.0293 | 0.6438 | 0.1075 | 0.6857 | 4.0 |
| LSTM VAE | 0.6730 | 0.0042 | 0.6718 | 0.025 | **0.5948** | 0.0391 | 0.6897 | 0.0207 | 0.8383 | 0.0096 | 0.5215 | 0.0082 | 0.6128 | 0.0089 | 0.6574 | 5.0 |
| BeatGAN | 0.5673 | 0.1219 | 0.7011 | 0.0189 | 0.5583 | 0.0097 | 0.7778 | 0.0294 | 0.8153 | 0.0078 | 0.5750 | 0.0085 | 0.6230 | 0.0618 | 0.6597 | 5.1 |
| MSCRED | 0.3262 | 0.0053 | 0.5958 | 0.0053 | 0.5864 | 0.0097 | 0.5977 | 0.0068 | 0.5249 | 0.0083 | 0.5030 | 0.0098 | 0.4935 | 0.0666 | 0.5182 | 8.0 |
| NASA LSTM | 0.2249 | 0.041 | 0.7030 | 0.0118 | 0.5559 | 0.0158 | **0.8431** | 0.0103 | 0.5666 | 0.0047 | 0.2874 | 0.0353 | 0.2535 | 0.0324 | 0.4906 | 8.0 |
| DAGMM* | 0.0305 | 0.0027 | 0.2601 | 0.0201 | 0.5449 | 0.0049 | 0.3351 | 0.0198 | 0.1066 | 0.0095 | 0.0891 | 0.0004 | 0.1264 | 0.0255 | 0.2132 | 12.7 |
| OmniAnomaly* | 0.2317 | 0.1833 | **0.7381** | 0.0188 | 0.5491 | 0.0007 | 0.6782 | 0.019 | 0.7904 | 0.0174 | 0.2836 | 0.1009 | 0.4192 | 0.0396 | 0.5272 | 8.0 |
| OCAN* | 0.2488 | 0.0988 | 0.5396 | 0.0457 | 0.5482 | 0.0016 | 0.4902 | 0.0173 | 0.6666 | 0.0139 | 0.2541 | 0.1604 | 0.1245 | 0.0117 | 0.4103 | 11.1 |

### S8.3 $Fc_1$ score with tail-p threshold

TABLE S9: $Fc_1$ score mean and standard deviation over 5 seeds, with the tail-p threshold and Gauss-D scoring function, with the following exceptions - * predefined scoring function and tail-p threshold. † predefined scoring and threshold. NASA LSTM NPT uses Non-Parametric Threshold. OC-SVM scores are thresholded at 0.5. The value of the threshold, $-\log_{10}(\epsilon) \in \{1:5\}$, and the value that gives the best $Fc_1$ is used here. The mean is calculated over all the datasets.

| dataset | DMDS | | MSL | | SKAB | | SMAP | | SMD | | SWaT | | WADI | | Mean | Avg Rank |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | Std | Mean | Std | Mean | Std | Mean | Std | Mean | Std | Mean | Std | Mean | Std | | |
| Raw Signal | 0.6192 | | 0.2618 | | 0.5340 | 0.0000 | 0.2924 | | 0.6952 | | 0.4266 | | 0.4613 | | 0.4701 | 8.1 |
| PCA | 0.2505 | | 0.3527 | | 0.5559 | 0.0000 | 0.3638 | | 0.4915 | | 0.5979 | | 0.3349 | | 0.4210 | 7.0 |
| UAE | 0.6429 | 0.0139 | 0.4571 | 0.0221 | 0.5545 | 0.0017 | 0.5135 | 0.0181 | 0.5440 | 0.0064 | 0.6467 | 0.011 | 0.5267 | 0.0239 | 0.5551 | 3.1 |
| FC AE | 0.6630 | 0.0144 | 0.4514 | 0.011 | 0.5429 | 0.0027 | 0.4379 | 0.0123 | 0.5209 | 0.007 | 0.5472 | 0.0307 | 0.5802 | 0.0477 | 0.5348 | 4.9 |
| LSTM AE | 0.3679 | 0.0162 | 0.4004 | 0.0102 | 0.5725 | 0.0303 | 0.4276 | 0.0077 | 0.4386 | 0.0023 | 0.5903 | 0.0133 | 0.4212 | 0.0058 | 0.4598 | 6.3 |
| TCN AE | 0.5569 | 0.0414 | 0.4438 | 0.0089 | 0.5484 | 0.0041 | 0.4469 | 0.0117 | 0.5675 | 0.013 | 0.5753 | 0.0346 | 0.5711 | 0.115 | 0.5300 | 4.7 |
| LSTM VAE | 0.6426 | 0.007 | 0.3917 | 0.0232 | 0.5767 | 0.0376 | 0.3700 | 0.0137 | 0.6175 | 0.0138 | 0.5123 | 0.0144 | 0.5878 | 0.0236 | 0.5284 | 4.7 |
| BeatGAN | 0.4171 | 0.141 | 0.4656 | 0.0103 | 0.5469 | 0.0190 | 0.3843 | 0.0063 | 0.4475 | 0.0097 | 0.5446 | 0.0208 | 0.5918 | 0.0659 | 0.4854 | 5.4 |
| MSCRED | 0.1367 | 0.0029 | 0.3629 | 0.0153 | 0.5798 | 0.0105 | 0.3361 | 0.0119 | 0.3085 | 0.0057 | 0.4787 | 0.0088 | 0.4416 | 0.1045 | 0.3778 | 8.1 |
| NASA LSTM | 0.1130 | 0.0096 | 0.4495 | 0.0134 | 0.5396 | 0.0198 | 0.4410 | 0.0181 | 0.3336 | 0.0058 | 0.2088 | 0.0299 | 0.2252 | 0.0514 | 0.3301 | 8.9 |
| DAGMM* | 0.0000 | 0 | 0.1800 | 0.0278 | 0.5430 | 0.0063 | 0.2419 | 0.0103 | 0.0000 | 0 | 0.0000 | 0 | 0.0000 | 0 | 0.1378 | 13.0 |
| OmniAnomaly* | 0.0557 | 0.0012 | 0.5008 | 0.01 | 0.5487 | 0.0006 | 0.4542 | 0.006 | 0.4444 | 0.0108 | 0.1497 | 0.0345 | 0.1072 | 0.0008 | 0.3230 | 7.7 |
| NASA LSTM NPT† | 0.1440 | 0.003 | 0.3403 | 0.0442 | 0.0902 | 0.0000 | 0.5869 | 0.012 | 0.1957 | 0.0028 | 0.0251 | 0.0057 | 0.1333 | 0 | 0.2165 | 10.3 |
| OC-SVM† | 0.0337 | 0 | 0.1717 | 0 | 0.5380 | 0.0000 | 0.1757 | 0 | 0.0893 | 0 | 0.0765 | 0 | 0.1876 | 0 | 0.1818 | 12.7 |

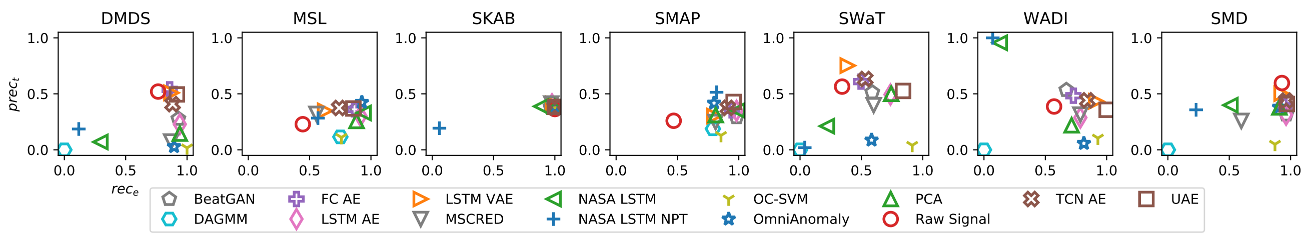



Fig. S8: $prec_t$ vs. $rec_e$ for the tail-p threshold for the algorithms shown in Table S9. For NASA LSTM NPT, we use NPT, i.e. non-parametric threshold, and for OC-SVM we threshold at 0.5, instead of tai-p threshold.

### S8.4 Point-wise $F_1$ score

TABLE S10: Point-wise $F_1$ score (i.e., the $F_1$ score) of various models with the *Gauss-D* scoring function (except the starred algorithms that specify their own scoring functions) with the *best-f1* threshold.

| Model | DMDS | MSL | SKAB | SMAP | SMD | SWaT | WADI | Mean | Avg Rank |
|---|---|---|---|---|---|---|---|---|---|
| Raw Signal | 0.4015 | 0.3059 | 0.5369 | 0.3009 | 0.4106 | 0.3068 | 0.3634 | 0.3751 | 9.6 |
| PCA | 0.3958 | 0.3982 | 0.5377 | 0.3334 | 0.4517 | 0.3954 | 0.2846 | 0.3995 | 6.6 |
| UAE | **0.5165** | **0.4454** | 0.5375 | **0.3937** | 0.4351 | **0.4544** | 0.3503 | **0.4476** | **3.4** |
| FC AE | 0.4902 | 0.4377 | 0.5375 | 0.3692 | 0.4429 | 0.3545 | 0.4056 | 0.4339 | 4.5 |
| LSTM AE | 0.4844 | 0.4088 | 0.5372 | 0.3752 | 0.4295 | 0.3956 | 0.3319 | 0.4232 | 6.1 |
| TCN AE | 0.4654 | 0.4366 | 0.5374 | 0.3729 | **0.4826** | 0.3931 | 0.3805 | 0.4384 | 4.1 |
| LSTM VAE | 0.4724 | 0.4001 | 0.5374 | 0.2998 | 0.4326 | 0.3136 | **0.4081** | 0.4091 | 7.1 |
| BeatGAN | 0.4023 | 0.4193 | 0.5376 | 0.3413 | 0.4529 | 0.3885 | 0.3629 | 0.4150 | 5.3 |
| MSCRED | 0.2624 | 0.4007 | 0.5411 | 0.3468 | 0.4661 | 0.3558 | 0.2316 | 0.3721 | 6.3 |
| NASA LSTM | 0.0971 | 0.4104 | 0.5373 | 0.3838 | 0.4228 | 0.1225 | 0.1284 | 0.3003 | 9.1 |
| DAGMM* | 0.0305 | 0.1938 | 0.5372 | 0.1964 | 0.0662 | 0.0592 | 0.0968 | 0.1686 | 12.8 |
| OmniAnomaly* | 0.1321 | 0.4277 | 0.5400 | 0.3699 | 0.4288 | 0.1726 | 0.2659 | 0.3339 | 7.4 |
| OCAN* | 0.2365 | 0.3155 | **0.5439** | 0.2656 | 0.4437 | 0.1307 | 0.1236 | 0.2942 | 8.9 |

TABLE S11: Point-wise $F_1$ score (i.e., the $F_1$ score) of various models with the *Gauss-D-K* scoring function (except the starred algorithms that specify their own scoring functions) with the *best-f1* threshold.

| Model | DMDS | MSL | SKAB | SMAP | SMD | SWaT | WADI | Mean | Avg Rank |
|---|---|---|---|---|---|---|---|---|---|
| Raw Signal | 0.4119 | 0.3323 | 0.5369 | 0.3593 | 0.4583 | 0.3760 | 0.4132 | 0.4126 | 9.0 |
| PCA | 0.4090 | 0.5323 | 0.5376 | 0.4481 | 0.4860 | 0.4453 | 0.3054 | 0.4520 | 6.4 |
| UAE | 0.5205 | 0.5402 | 0.5382 | 0.5755 | 0.4732 | 0.5799 | 0.4740 | **0.5288** | **2.9** |
| FC AE | 0.4942 | 0.5205 | 0.5387 | 0.5421 | 0.4828 | 0.4506 | 0.4902 | 0.5027 | 3.4 |
| LSTM AE | 0.4923 | 0.5355 | 0.5372 | 0.5317 | 0.4551 | 0.4480 | 0.3247 | 0.4749 | 6.2 |
| TCN AE | 0.4714 | 0.5493 | 0.5377 | 0.5453 | 0.5195 | 0.4256 | 0.4348 | 0.4977 | 4.1 |
| LSTM VAE | 0.4834 | 0.4897 | 0.5377 | 0.4875 | 0.4673 | 0.4158 | 0.5024 | 0.4834 | 5.9 |
| BeatGAN | 0.4055 | 0.5373 | 0.5378 | 0.5254 | 0.4826 | 0.4416 | 0.4155 | 0.4780 | 5.6 |
| MSCRED | 0.2650 | 0.4479 | 0.5412 | 0.3852 | 0.4671 | 0.3652 | 0.2664 | 0.3911 | 7.7 |
| NASA LSTM | 0.0998 | 0.5481 | 0.5381 | 0.5933 | 0.4412 | 0.1309 | 0.2031 | 0.3649 | 7.7 |
| DAGMM* | 0.0305 | 0.1938 | 0.5372 | 0.1964 | 0.0662 | 0.0592 | 0.0968 | 0.1686 | 12.8 |
| OmniAnomaly* | 0.1321 | 0.4277 | 0.5400 | 0.3699 | 0.4288 | 0.1726 | 0.2659 | 0.3339 | 9.4 |
| OCAN* | 0.2365 | 0.3155 | 0.5439 | 0.2656 | 0.4437 | 0.1307 | 0.1236 | 0.2942 | 9.9 |

### S8.5 AU-ROC score

TABLE S12: Area under the receiver operator characteristic curve (AU-ROC) of various models with the *Gauss-D* scoring function (except the starred algorithms that specify their own scoring functions).

| Model | DMDS | MSL | SKAB | SMAP | SMD | SWaT | WADI | Mean | Avg Rank |
|---|---|---|---|---|---|---|---|---|---|
| Raw Signal | 0.7073 | 0.4379 | 0.4855 | 0.4953 | 0.7475 | 0.7273 | 0.7733 | 0.6249 | 10.0 |
| PCA | 0.7413 | 0.6175 | 0.5037 | 0.5684 | 0.7900 | 0.8249 | 0.6583 | 0.6720 | 7.3 |
| UAE | 0.8852 | **0.6712** | 0.5088 | 0.6256 | **0.8298** | **0.8279** | 0.7405 | 0.7270 | **2.6** |
| FC AE | **0.9280** | 0.6602 | 0.4914 | 0.6272 | 0.8028 | 0.8085 | **0.7912** | **0.7299** | 4.0 |
| LSTM AE | 0.8404 | 0.6205 | 0.5040 | 0.6137 | 0.8196 | 0.8164 | 0.6808 | 0.6993 | 5.4 |
| TCN AE | 0.8109 | 0.6184 | 0.5002 | 0.6262 | 0.8282 | 0.7596 | 0.7109 | 0.6935 | 5.5 |
| LSTM VAE | 0.9146 | 0.5986 | 0.5002 | 0.5628 | 0.7587 | 0.7570 | 0.7892 | 0.6973 | 7.2 |
| BeatGAN | 0.7960 | 0.6517 | 0.4941 | 0.5993 | 0.8269 | 0.7889 | 0.7086 | 0.6951 | 6.1 |
| MSCRED | 0.6866 | 0.6436 | **0.5270** | 0.6020 | 0.8157 | 0.7069 | 0.6951 | 0.6681 | 6.4 |
| NASA LSTM | 0.6743 | 0.6580 | 0.4885 | 0.6406 | 0.7538 | 0.5489 | 0.4633 | 0.6039 | 9.1 |
| DAGMM* | 0.4287 | 0.4761 | 0.5080 | 0.5459 | 0.4767 | 0.4663 | 0.5025 | 0.4863 | 11.1 |
| OmniAnomaly* | 0.6804 | 0.6523 | 0.5075 | **0.6588** | 0.7972 | 0.5900 | 0.5275 | 0.6305 | 6.9 |
| OCAN* | 0.6948 | 0.5803 | 0.5192 | 0.5656 | 0.7736 | 0.5327 | 0.4708 | 0.5910 | 9.3 |

TABLE S13: Area under the receiver operator characteristic curve (AU-ROC) of various models with the *Gauss-D-K* scoring function (except the starred algorithms that specify their own scoring functions).

| Model | DMDS | MSL | SKAB | SMAP | SMD | SWaT | WADI | Mean | Avg Rank |
|---|---|---|---|---|---|---|---|---|---|
| Raw Signal | 0.7159 | 0.4815 | 0.4814 | 0.4878 | 0.7917 | 0.7618 | 0.7704 | 0.6415 | 9.7 |
| PCA | 0.7486 | 0.7448 | 0.4983 | 0.6483 | 0.8275 | 0.8440 | 0.7084 | 0.7171 | 6.4 |
| UAE | 0.8890 | **0.7679** | 0.5079 | **0.7643** | 0.8599 | **0.8637** | **0.8178** | **0.7815** | **1.9** |
| FC AE | **0.9317** | 0.7594 | 0.4874 | 0.7396 | 0.8375 | 0.8249 | 0.8166 | 0.7710 | 4.3 |
| LSTM AE | 0.8421 | 0.7548 | 0.4984 | 0.7306 | 0.8410 | 0.8350 | 0.7273 | 0.7470 | 4.7 |
| TCN AE | 0.8117 | 0.7160 | 0.4977 | 0.7239 | **0.8601** | 0.7696 | 0.7514 | 0.7329 | 5.4 |
| LSTM VAE | 0.9186 | 0.6812 | 0.4967 | 0.6479 | 0.7961 | 0.7766 | 0.8116 | 0.7327 | 6.6 |
| BeatGAN | 0.7967 | 0.7591 | 0.4886 | 0.7109 | 0.8523 | 0.8066 | 0.7381 | 0.7360 | 5.7 |
| MSCRED | 0.6898 | 0.6723 | 0.5267 | 0.6374 | 0.8171 | 0.7212 | 0.7233 | 0.6840 | 7.7 |
| NASA LSTM | 0.6830 | 0.7670 | 0.4876 | 0.7577 | 0.7690 | 0.5663 | 0.5091 | 0.6485 | 8.6 |
| DAGMM* | 0.4287 | 0.4761 | 0.5080 | 0.5459 | 0.4767 | 0.4663 | 0.5025 | 0.4863 | 11.3 |
| OmniAnomaly* | 0.6804 | 0.6523 | 0.5075 | 0.6588 | 0.7972 | 0.5900 | 0.5275 | 0.6305 | 8.9 |
| OCAN* | 0.6948 | 0.5803 | **0.5192** | 0.5656 | 0.7736 | 0.5327 | 0.4708 | 0.5910 | 9.9 |

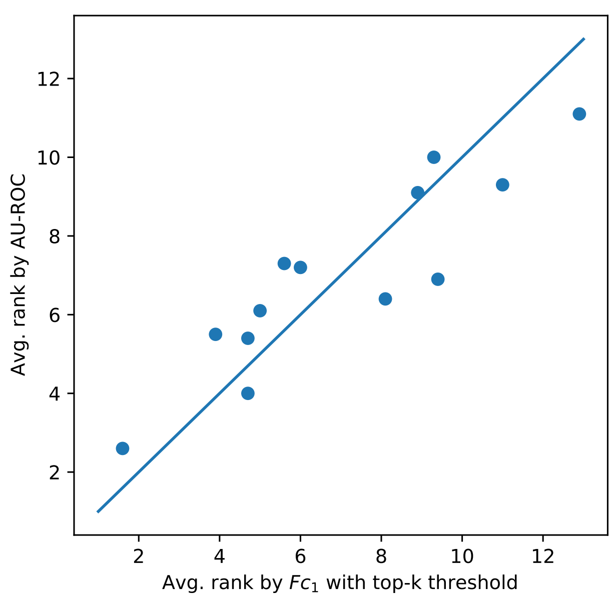



Fig. S9: Comparison of average ranks of algorithms from $Fc_1$ score (Table S6) vs. AU-ROC score (Table S12). The two metrics are generally in agreement, as evidenced by the closeness of scatter points to the x=y line.

### S8.6 AU-PRC score (Average Precision)

TABLE S14: Area under the precision recall curve (AU-PRC), calculated through average precision (AP), of various models with the *Gauss-D* scoring function (except the starred algorithms that specify their own scoring functions).

| Model | DMDS | MSL | SKAB | SMAP | SMD | SWaT | WADI | Mean | Avg Rank |
|---|---|---|---|---|---|---|---|---|---|
| Raw Signal | 0.2946 | 0.2092 | 0.3549 | 0.2121 | 0.3492 | 0.2348 | 0.2525 | 0.2725 | 10.0 |
| PCA | 0.3144 | 0.2659 | 0.3694 | 0.2618 | 0.3859 | 0.3784 | 0.1520 | 0.3040 | 7.0 |
| UAE | 0.4134 | 0.3459 | 0.3699 | 0.3220 | 0.3835 | 0.3829 | 0.2186 | 0.3480 | 3.3 |
| FC AE | 0.4147 | 0.3314 | 0.3535 | 0.2819 | 0.3932 | 0.3298 | 0.3125 | 0.3453 | 4.9 |
| LSTM AE | 0.3972 | 0.3003 | 0.3683 | 0.2978 | 0.3644 | 0.3704 | 0.1756 | 0.3249 | 6.3 |
| TCN AE | 0.3776 | 0.3228 | 0.3654 | 0.2965 | 0.4285 | 0.3425 | 0.2761 | 0.3442 | 4.6 |
| LSTM VAE | 0.3845 | 0.3216 | 0.3666 | 0.2337 | 0.3812 | 0.2843 | 0.2886 | 0.3229 | 6.4 |
| BeatGAN | 0.3273 | 0.3171 | 0.3628 | 0.2592 | 0.4034 | 0.3388 | 0.2780 | 0.3267 | 6.1 |
| MSCRED | 0.1298 | 0.3182 | 0.3870 | 0.2624 | 0.4104 | 0.3208 | 0.1521 | 0.2830 | 6.0 |
| NASA LSTM | 0.0569 | 0.3138 | 0.3616 | 0.3052 | 0.3599 | 0.1011 | 0.0802 | 0.2255 | 9.7 |
| DAGMM* | 0.0129 | 0.1274 | 0.3709 | 0.1808 | 0.0441 | 0.0434 | 0.0597 | 0.1199 | 11.6 |
| OmniAnomaly* | 0.0844 | 0.3299 | 0.3724 | 0.3090 | 0.3802 | 0.1118 | 0.2081 | 0.2565 | 6.6 |
| OCAN* | 0.2067 | 0.2224 | 0.3756 | 0.2232 | 0.3958 | 0.1174 | 0.0513 | 0.2275 | 8.6 |

TABLE S15: Area under the precision recall curve (AU-PRC), calculated through average precision (AP), of various models with the *Gauss-D-K* scoring function (except the starred algorithms that specify their own scoring functions).

| Model | DMDS | MSL | SKAB | SMAP | SMD | SWaT | WADI | Mean | Avg Rank |
|---|---|---|---|---|---|---|---|---|---|
| Raw Signal | 0.2968 | 0.2528 | 0.3541 | 0.2874 | 0.4060 | 0.2731 | 0.3519 | 0.3174 | 9.1 |
| PCA | 0.3167 | 0.4582 | 0.3669 | 0.3855 | 0.4358 | 0.3943 | 0.1665 | 0.3606 | 6.0 |
| UAE | 0.4130 | 0.4543 | 0.3710 | 0.5323 | 0.4334 | 0.4608 | 0.3814 | 0.4352 | 3.3 |
| FC AE | 0.4150 | 0.4422 | 0.3502 | 0.4871 | 0.4451 | 0.3900 | 0.4315 | 0.4230 | 4.6 |
| LSTM AE | 0.3976 | 0.4568 | 0.3673 | 0.4773 | 0.4037 | 0.3864 | 0.2395 | 0.3898 | 5.7 |
| TCN AE | 0.3778 | 0.4753 | 0.3648 | 0.5009 | 0.4802 | 0.3513 | 0.3439 | 0.4135 | 4.3 |
| LSTM VAE | 0.3822 | 0.4226 | 0.3646 | 0.4318 | 0.4278 | 0.3411 | 0.4073 | 0.3968 | 6.1 |
| BeatGAN | 0.3268 | 0.4595 | 0.3592 | 0.4710 | 0.4478 | 0.3604 | 0.3577 | 0.3975 | 5.1 |
| MSCRED | 0.1314 | 0.3779 | 0.3874 | 0.3288 | 0.4115 | 0.3063 | 0.2514 | 0.3135 | 7.3 |
| NASA LSTM | 0.0585 | 0.4682 | 0.3587 | 0.5573 | 0.3839 | 0.1076 | 0.1538 | 0.2983 | 8.6 |
| DAGMM* | 0.0129 | 0.1274 | 0.3709 | 0.1808 | 0.0441 | 0.0434 | 0.0597 | 0.1199 | 11.7 |
| OmniAnomaly* | 0.0844 | 0.3299 | 0.3724 | 0.3090 | 0.3802 | 0.1118 | 0.2081 | 0.2565 | 9.4 |
| OCAN* | 0.2067 | 0.2224 | 0.3756 | 0.2232 | 0.3958 | 0.1174 | 0.0513 | 0.2275 | 9.7 |

### S8.7 Root cause results

TABLE S16: RC-Top3-all metric for all datasets with root cause labels using the Gauss-D scoring function except starred.

| Algo | DMDS Mean | DMDS Std | SMD Mean | SMD Std | SWaT Mean | SWaT Std | WADI Mean | WADI Std | Overall mean | Avg Rank |
|---|---|---|---|---|---|---|---|---|---|---|
| Raw Signal | 0.7059 | | **0.9635** | | 0.3143 | | 0.5000 | | 0.6209 | 6.5 |
| PCA | 0.8235 | | 0.7831 | | 0.4857 | | 0.5000 | | 0.6481 | 7.2 |
| UAE | **0.9647** | 0.032 | 0.9498 | 0.006 | **0.6286** | 0.020 | **0.5428** | 0.039 | **0.7715** | **1.8** |
| FC AE | 0.9412 | 0.0000 | 0.9522 | 0.0053 | 0.6000 | 0.0286 | 0.4428 | 0.0783 | 0.7341 | 3.8 |
| LSTM AE | 0.8117 | 0.0263 | 0.9360 | 0.0032 | 0.5771 | 0.0128 | 0.5143 | 0.0319 | 0.7098 | 5.2 |
| TCN AE | 0.9177 | 0.0526 | 0.9448 | 0.0064 | 0.5143 | 0.0452 | 0.5286 | 0.1083 | 0.7263 | 4.5 |
| LSTM VAE | 0.9177 | 0.0322 | 0.9501 | 0.0019 | 0.4571 | 0.0000 | 0.5000 | 0.0000 | 0.7062 | 4.8 |
| BeatGAN | 0.9176 | 0.0526 | 0.9424 | 0.0089 | 0.5543 | 0.0256 | 0.4571 | 0.0391 | 0.7178 | 5.8 |
| MSCRED | 0.7765 | 0.0263 | 0.8526 | 0.0101 | 0.5200 | 0.0424 | 0.0714 | 0 | 0.5551 | 8.2 |
| OmniAnomaly* | 0.9177 | 0.0671 | 0.9272 | 0.0067 | 0.3772 | 0.0619 | 0.4857 | 0.0319 | 0.6770 | 7.2 |

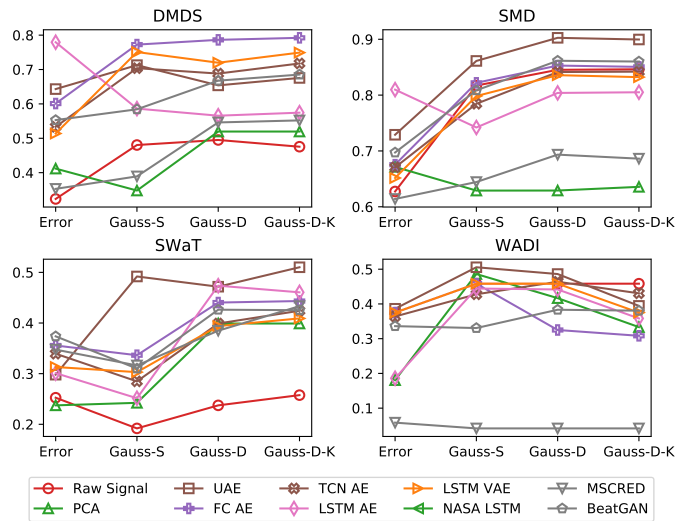



Fig. S10: Effect of scoring functions on the HitRate@150 performance for various algorithms and datasets.

TABLE S17: Hitrate@150 for independent anomaly diagnosis with the Gauss-D scoring function except starred.

| Algo | DMDS Mean | DMDS Std | SMD Mean | SMD Std | SWaT Mean | SWaT Std | WADI Mean | WADI Std | Overall mean | Avg Rank |
|---|---|---|---|---|---|---|---|---|---|---|
| Raw Signal | 0.4951 | 0.0000 | 0.8453 | 0.0000 | 0.2374 | 0.0000 | 0.4583 | 0.0000 | 0.5090 | 6.6 |
| PCA | 0.5196 | 0.0000 | 0.6291 | 0.0000 | 0.3990 | 0.0000 | 0.4167 | 0.0000 | 0.4911 | 7.6 |
| UAE | 0.6540 | 0.011 | **0.9026** | 0.003 | 0.4717 | 0.013 | **0.4861** | 0.031 | **0.6286** | **2.2** |
| FC AE | **0.7863** | 0.0096 | 0.8529 | 0.0025 | 0.4404 | 0.0083 | 0.3250 | 0.0745 | 0.6012 | 4.0 |
| LSTM AE | 0.5657 | 0.0281 | 0.8039 | 0.0089 | **0.4737** | 0.0206 | 0.4417 | 0.0410 | 0.5712 | 5.2 |
| TCN AE | 0.6883 | 0.0356 | 0.8413 | 0.0031 | 0.3990 | 0.0279 | 0.4639 | 0.0999 | 0.5981 | 3.9 |
| LSTM VAE | 0.7226 | 0.0172 | 0.8357 | 0.0009 | 0.3939 | 0.0000 | 0.4583 | 0.0000 | 0.6026 | 4.6 |
| BeatGAN | 0.6676 | 0.05 | 0.8615 | 0.0045 | 0.4263 | 0.0347 | 0.3833 | 0.0712 | 0.5847 | 4.2 |
| MSCRED | 0.5461 | 0.0602 | 0.6934 | 0.0056 | 0.3848 | 0.0263 | 0.0417 | 0 | 0.4165 | 8.8 |
| OmniAnomaly* | 0.6451 | 0.1127 | 0.8294 | 0.0071 | 0.2232 | 0.0322 | 0.3528 | 0.0076 | 0.5126 | 7.8 |

## References


[1] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection," *ACM Computing Surveys*, vol. 54, no. 2, p. 1–38, Mar 2021. [Online]. Available: http://dx.doi.org/10.1145/3439950

[2] J. Goh, S. Adepu, K. N. Junejo, and A. Mathur, "A dataset to support research in the design of secure water treatment systems," in *International Conference on Critical Information Infrastructures Security*. Springer, 2016, pp. 88–99.

[3] (2020) itust website. [Online]. Available: "https://itrust.sutd.edu.sg/itrust-labs_datasets/"

[4] C. M. Ahmed, V. R. Palleti, and A. P. Mathur, "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*, 2017, pp. 25–28.

[5] M. Bartyś, R. Patton, M. Syfert, S. de las Heras, and J. Quevedo, "Introduction to the damadics actuator fdi benchmark study," *Control engineering practice*, vol. 14, no. 6, pp. 577–596, 2006.

[6] (2020) Damadics benchmark website. [Online]. Available: "http://diag.mchtr.pw.edu.pl/damadics/"

[7] I. D. Katser and V. O. Kozitsin, "Skoltech anomaly benchmark (skab)," 2020. [Online]. Available: https://www.kaggle.com/dsv/1693952
[8] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, "Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 387–395.
[9] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei, "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*, 2019, pp. 2828–2837.
[10] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, "Usad: Unsupervised anomaly detection on multivariate time series," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 3395–3404.
[11] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "Lstm-based encoder-decoder for multi-sensor anomaly detection," *arXiv preprint arXiv:1607.00148*, 2016.
[12] C. Zhang, D. Song, Y. Chen, X. Feng, C. Lumezanu, W. Cheng, J. Ni, B. Zong, H. Chen, and N. V. Chawla, "A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 1409–1416.
[13] P. Zheng, S. Yuan, X. Wu, J. Li, and A. Lu, "One-class adversarial nets for fraud detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 1286–1293.
[14] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, and S.-K. Ng, "Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks," in *International Conference on Artificial Neural Networks*. Springer, 2019, pp. 703–716.
[15] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.
[16] Y. Hochberg, "A sharper bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 75, no. 4, pp. 800–802, 1988.
[17] C. López-Vázquez and E. Hochsztain, "Extended and updated tables for the friedman rank test," *Communications in Statistics-Theory and Methods*, vol. 48, no. 2, pp. 268–281, 2019.