A Deep Neural Network for Unsupervised Anomaly Detection and Diagnosis in Multivariate Time Series Data

Chuxu Zhang, Dongjin Song, Yuncong Chen, Xinyang Feng, Cristian Lumezanu, Wei Cheng, Jingchao Ni, Bo Zong, Haifeng Chen, and Nitesh V. Chawla

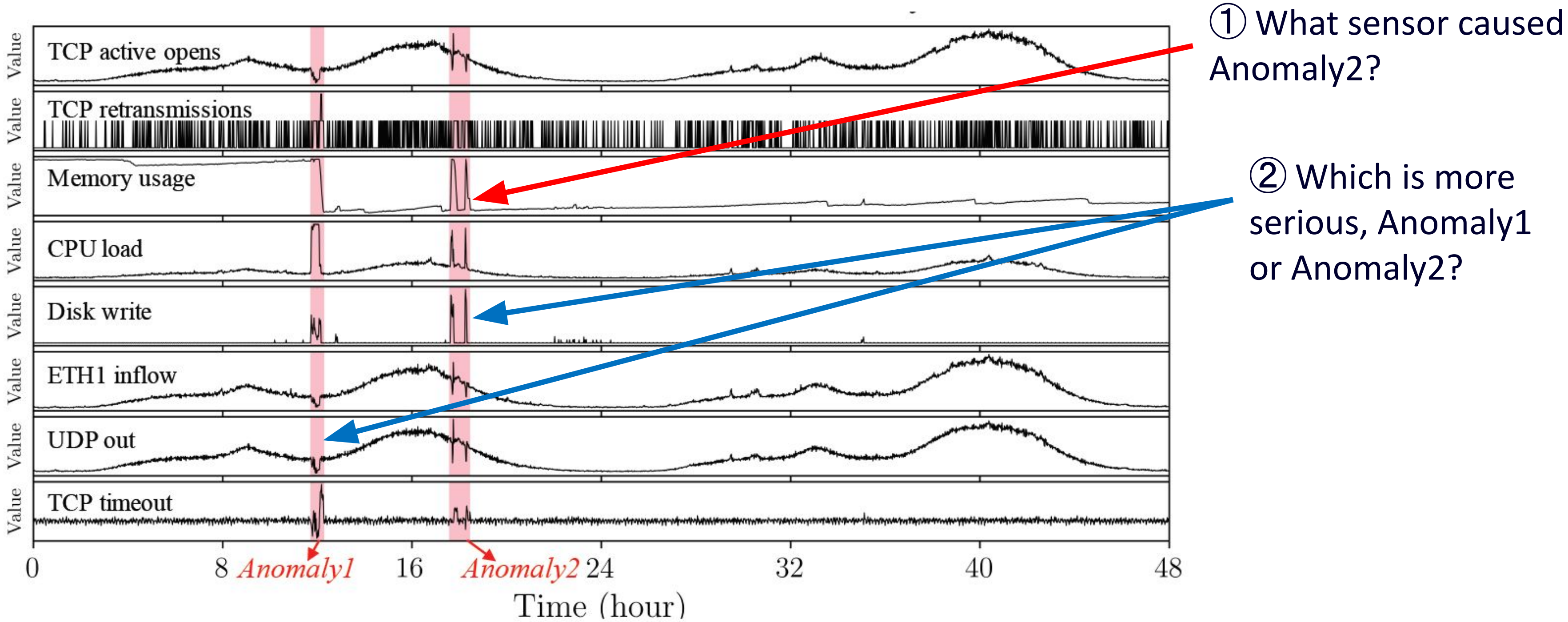
AAAI, 2019

February 12, 2020
Seonyoung Kim

CONTENTS

Introduction

- Anomaly detection and <u>diagnosis</u> in multivariate time series
  - Identifying abnormal status in certain time steps and <u>pinpointing the root causes</u>



Eg. 8-dimensional multivariate time series in 2-day from the server machine dataset

- If we can know about the root causes and the severity of anomalies, it is easy to know what to repair first

Related Work with Issue

1. Distance/clustering methods
   - Grouping a set of data which are similar each other
   - e.g. k-Nearest Neighbor (kNN) [1]

2. Classification methods
   - Learning decision function and classifying data as similar or dissimilar
   - e.g. One-Class SVM (OC-SVM) [2]

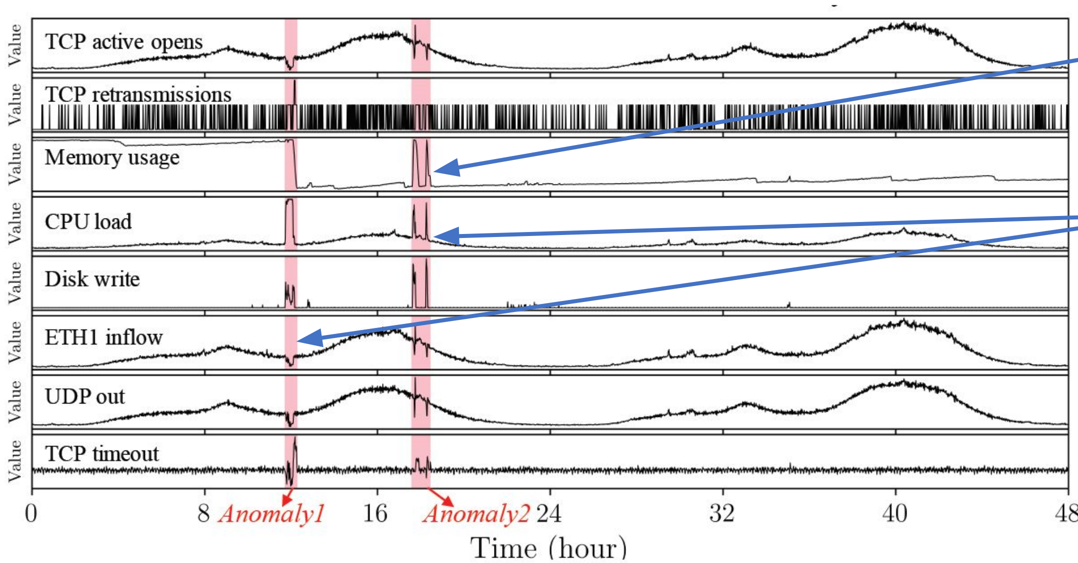
3. Density estimation methods
   - Modeling data density for outlier detection
   - e.g. Deep Autoencoding Gaussian Mixture Model (DAGMM) [3]

4. Prediction methods
   - Modeling the temporal dependency of data and predicts the value
   - e.g. Autoregressive Moving Average (ARMA) [4]

- ①, ②, ③ → Cannot capture temporal dependencies across different time steps
- ④ → Sensitive to noise

Motivations

- Main challenges in two aspects
  1. **Anomaly Detection**
     - Detect anomalies in certain time steps
     - Consider temporal dependencies and correlations between two time series
  2. **Anomaly Diagnosis**
     - Identify the root causes of anomalies
     - Provide the anomaly severity interpretation



root cause identification
(Which sensor cause anomaly?)

anomaly severity interpretation
(Which anomaly is more serious?)

correlations between two time series
(UDP out and TCP timeout have a relationship?)

temporal dependencies

Proposed method: MSCRED

- Overview
  - Characterizing status with signature matrices
    - ✔ Construct signature matrices, considering the correlation between two time series
  - Convolutional encoder
    - ✔ Encode the spatial patterns of signature matrices
  - Attention based ConvLSTM
    - ✔ Capture the spatial patterns of signature matrices with temporal information
  - Convolutional decoder
    - ✔ Reconstruct the signature matrices with the outputs of attention based ConvLSTM

Proposed method: MSCRED

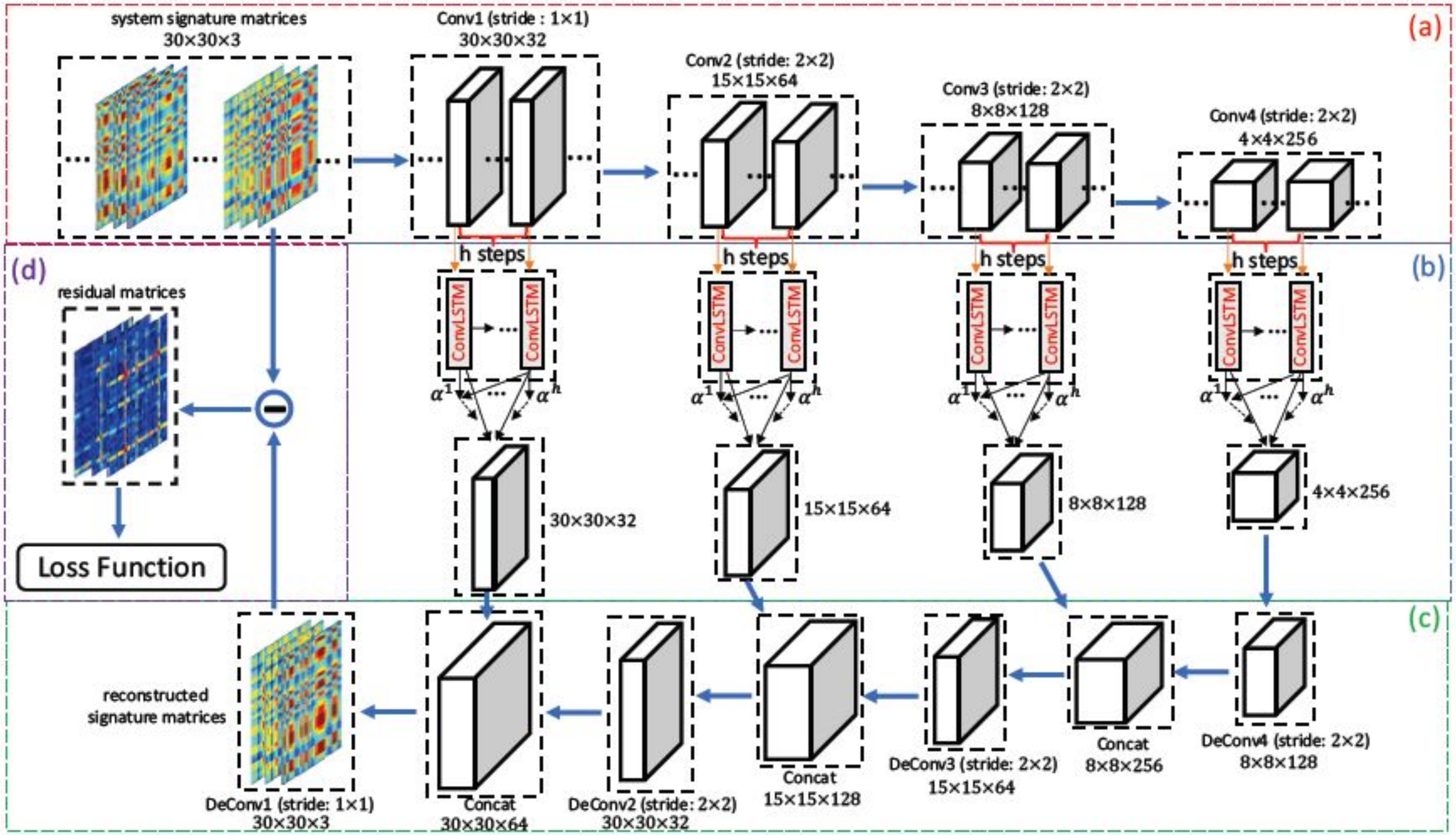


Figure 1.
(a) Characterizing Status with Signature Matrices & Convolutional Encoder
(b) Attention based ConvLSTM
(c) Convolutional Decoder
(d) Loss function

Proposed method: MSCRED

- Characterizing status with s signature matrices M^t (Figure 1.(a))
  - Construct signature matrices, considering the correlation between two time series
    - $\Upsilon^{\omega} = (X_1^{\omega}, \dots, X_n^{\omega}) \in \mathbb{R}^{n \times \omega}$: n time series with length ω
    - $X_i^{\omega} = \left(x_i^{t-\omega}, x_i^{t-\omega-1}, \dots, x_i^{t}\right)$: i-th time series from $t - \omega$ to t
    - ω : window size
    - M^t : $n \times n$ signature matrix
    - $m_{ij}^t \in M^t$: an element of the signature matrix
      - ✓ $m_{ij}^t = \dfrac{\sum_{\delta=0}^{\omega} x_i^{t-\delta} x_j^{t-\delta}}{\kappa}$
      - ✓ $\kappa = \omega$

- Example
  - $\Upsilon^2 = (X_1^2, X_2^2)$ $(\omega = 2, n = 2)$: 2 time series with length 2
  - X_1: temperature, X_2: atmospheric pressure
  - $X_1^2 = (15°\mathrm{C}, 14°\mathrm{C})$
  - $X_2^2 = (3\ \mathrm{atm}, 2\ \mathrm{atm})$

$\frac{15 * 3 + 14 * 2}{2}$

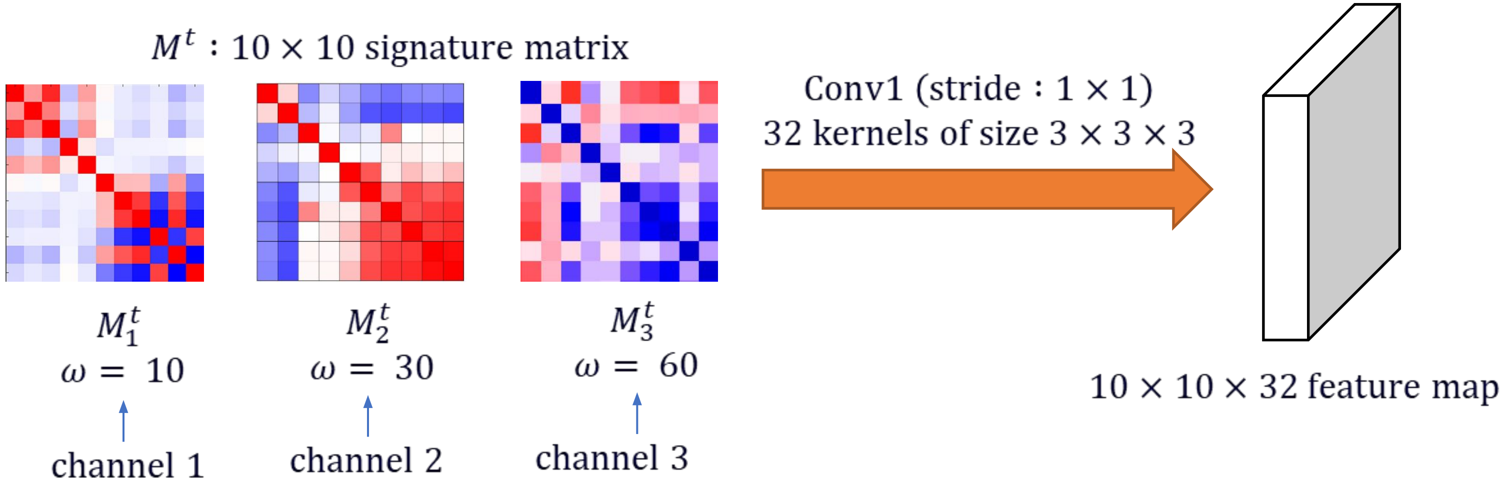
$$\begin{array}{c|c|c|}
 & X_1 & X_2 \\ \hline
X_1 & 210.5 & 36.5 \\ \hline
X_2 & 36.5 & 6.5 \\ \hline
\end{array}$$

M^t

Proposed method: MSCRED

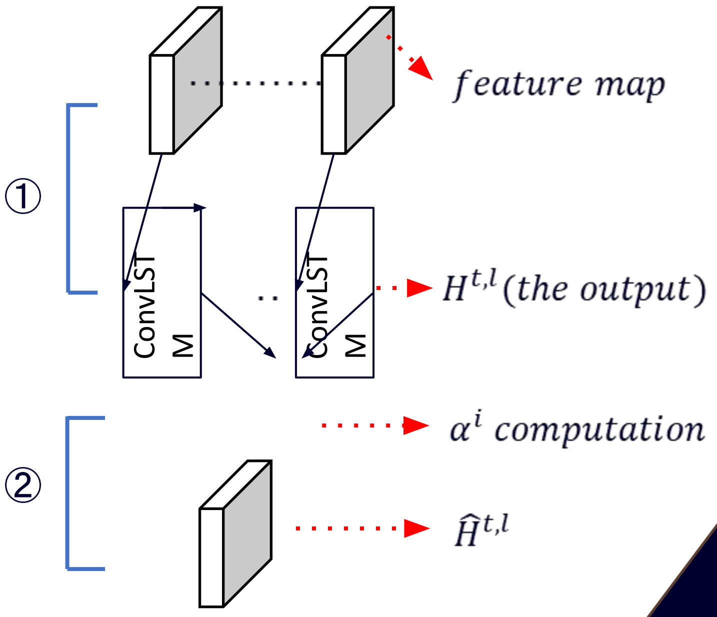
- Convolutional Encoder (Figure 1.(a))
  - Encode the spatial patterns of system signature matrices
    - $\chi^{t,l} \in \mathbb{R}^{n_l \times n_l \times d_l}$: a feature map in the l-th layer
    - d_l : the number of kernels (filters)
      - ✓ $\chi^{t,l} = f(W^l * \chi^{t,l-1} + b^l)$
        - ✓ $*$: the convolutional operation
        - ✓ $f(\cdot)$: the activation function
        - ✓ $W^l \in \mathbb{R}^{k_l \times k_l \times d_{l-1} \times d_l}$: d_l convolutional kernels of size $k_l \times k_l \times d_{l-1}$
        - ✓ $b^l \in \mathbb{R}^{d_l}$: bias

- Example



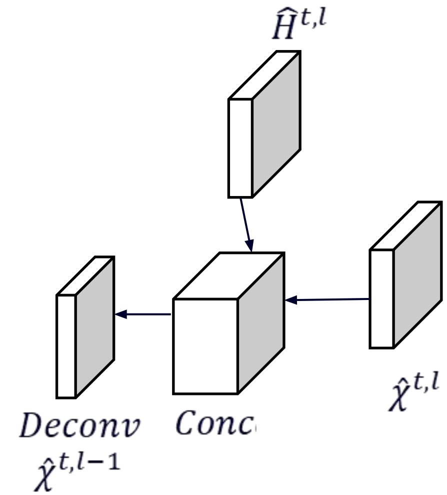
Proposed method: MSCRED

- Attention based ConvLSTM (Figure 1.(b))
  - ① Train model to capture the spatial patterns of signature matrices with temporal information
    - $H^{t,l} = ConvLSTM(\chi^{t,l}, H^{t-1,l})$
      - ✓ $H^{t,l} \in \mathbb{R}^{n_l \times n_l \times d_l}$: the hidden state of $l-$ th layer at time t
  - ② To select the steps that are relevant to current step, adopt a temporal attention mechanism
    - $\widehat{H}^{t,l} = \sum_{i \in (t-h,t)} \alpha^i H^{i,l}$
      - ✓ $\widehat{H}^{t,l}$: the refined output of feature maps
      - ✓ α^i : the importance weights
    - $\alpha^i = \dfrac{exp\left\{\frac{Vec(H^{t,l})^T Vec(H^{i,l})}{x}\right\}}{\sum_{i \in (t-h,t)} \exp\left\{\frac{Vec(H^{t,l})^T Vec(H^{i,l})}{x}\right\}}$
      - ✓ $x(= 0.5)$: rescale factor



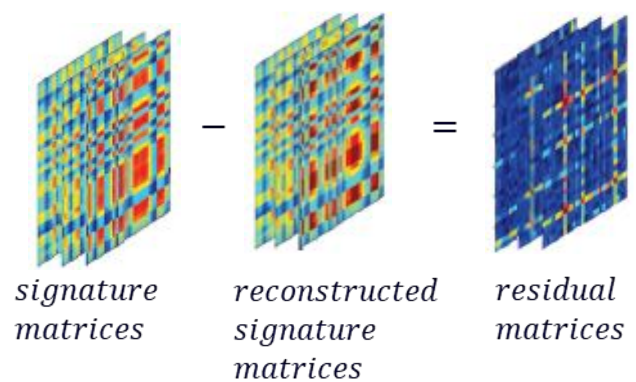
Proposed method: MSCRED

- Convolutional Decoder (Figure 1.(c))
  - With trained model, reconstruct the signature matrices
    - $\hat{\chi}^{t,l-1} = \begin{cases} f(\widehat{W}^{t,l} \circledast \widehat{H}^{t,l} + \hat{b}^{t,l}) & l = 4 \\ f(\widehat{W}^{t,l} \circledast [\widehat{H}^{t,l} \oplus \hat{\chi}^{t,l}] + \hat{b}^{t,l}) & l = 3,2,1 \end{cases}$
      - ✓ $\hat{\chi}^{t,l-1}$: the output feature map
      - ✓ $\circledast$: the deconvolution operation
      - ✓ $\oplus$: the concatenation operation
      - ✓ $f(\cdot)$: the activation function
      - ✓ $\widehat{W}^{l}$: the kernel of $l-$th deconvolutional layer
      - ✓ $\hat{b}^{l}$: bias of $l-$th deconvolutional layer



  - $\hat{\chi}^{t,l-1}$ is concatenated with the output of previous ConvLSTM layer, making the decoder process stacked

- Loss function (Figure 1.(d))
  - $\mathcal{L}_{MSCRED} = \sum_t \sum_{c=1}^{s} \| \chi_{:,:,c}^{t,0} - \hat{\chi}_{:,:,c}^{t,0} \|_F^2$



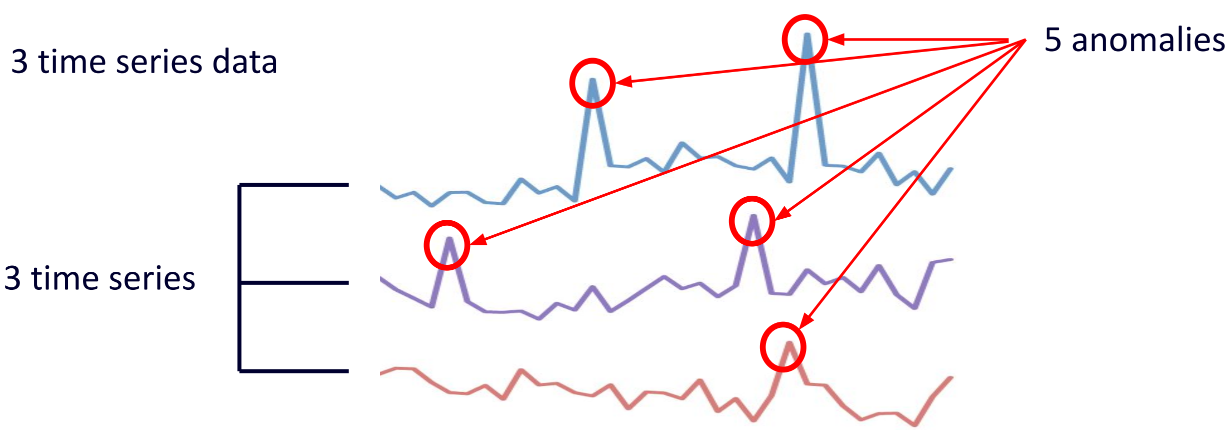
Experiments

- Dataset
  - Synthetic data
    - Generate time series data, using the following formula
    - $S(t) = \begin{cases} sin[(t - t_0)/w] + \lambda \cdot \epsilon, & s_{rand} = 0 \\ cos[(t - t_0)/w] + \lambda \cdot \epsilon, & s_{rand} = 1 \end{cases}$
      - ✓ $t_0 \in [50, 100]$: time delay
      - ✓ $w \in [40, 50]$: frquency
      - ✓ $\epsilon \sim \mathrm{N}(0,1)$: random gaussian noise
      - ✓ $\lambda = 0.3$: scale factor
  - Power plant data
    - Collected on a real power plant
    - 1 anomaly + injected 4 additional anomalies

Experiments

Statistics	Synthetic	Power Plant
# time series	30	36
# points	20,000	23,040
# anomalies	5	5
# root cause	3	3
Train period	0 ~ 8000	0 ~ 10,080
Valid period	8001 ~ 10,000	10,081 ~ 18,720
Test period	10,001 ~ 20,000	18,721 ~ 23,040

5 anomalies in 3 time series data



Experiments

- Compare MSCRED with 8 baseline methods
  - Classification model
    - ✓ One-Class SVM model (OC-SVM)
  - Density estimation model
    - ✓ Deep Autoencoding Gaussian Mixture model (DAGMM)
    - ✓ The anomaly score : the energy score (Zong et al. 2018)
  - Prediction model
    - ✓ History Average (HA)
    - ✓ Auto Regression Moving Average (ARMA)
    - ✓ LSTM encoder-decoder (LSTM-ED)
    - ✓ The anomaly score : the average prediction error over all time series
  - MSCRED variants
    - ✓ $CNN_{ConvLSTM}^{ED(4)}$: MSCRED with attention module + first three ConvLSTM layers been removed
    - ✓ $CNN_{ConvLSTM}^{ED(3,4)}$: MSCRED with attention module + first two ConvLSTM layers been removed
    - ✓ $CNN_{ConvLSTM}^{ED}$: MSCRED with attention module been removed

Experiments

- The anomaly score
  - the residual matrix : $\chi - \hat{\chi}$
  - x : an element in the residual signature matrix of test data
  - If $x >$ threshold τ, $x \rightarrow$ anomaly
    - threshold $\tau = \beta \cdot max\{s(t)_{valid}\}$
      - ✓ $s(t)_{valid}$: the anomaly scores over the validation period
      - ✓ $\beta \in [1,2]$: set to maximize the F1 score over the validation period

- Use 3 metrics
  - Precision, recall and F1 score
  - Recall and precision scores over the test period are computed based on the threshold

- Experiments on both datasets are repeated 5 times and the average results are reported for comparison

Experiments

Table 1 : Anomaly detection results on two datasets

Method	Synthetic Data			Power Plant Data		
	Pre	Rec	F_1	Pre	Rec	F_1
OC-SVM	0.14	0.44	0.22	0.11	0.28	0.16
DAGMM	0.33	0.20	0.25	0.26	0.20	0.23
HA	0.71	0.52	0.60	0.48	0.52	0.50
ARMA	0.91	0.52	0.66	0.58	0.60	0.59
LSTM-ED	<u>1.00</u>	<u>0.56</u>	<u>0.72</u>	<u>0.75</u>	<u>0.68</u>	<u>0.71</u>
$\mathrm{CNN}^{ED(4)}_{ConvLSTM}$	0.37	0.24	0.29	0.67	0.56	0.61
$\mathrm{CNN}^{ED(3,4)}_{ConvLSTM}$	0.63	0.56	0.59	0.80	0.72	0.76
$\mathrm{CNN}^{ED}_{ConvLSTM}$	0.80	0.76	0.78	0.85	0.72	0.78
MSCRED	**1.00**	**0.80**	**0.89**	**0.85**	**0.80**	**0.82**
Gain (%)	–	30.0	23.8	13.3	19.4	15.5

The best Baseline method

The best score

The improvement (%) of MSCRED over the best baseline method

Observation

- Performance Evaluation
  1. **Anomaly detection**
     - ✔ (RQ1) Whether MSCRED can outperform baseline methods for anomaly detection in multivariate time series?
     - ✔ (RQ2) How does each component of MSCRED affect its performance?
  2. **Anomaly diagnosis**
     - ✔ (RQ3) Whether MSCRED can perform root cause identification and (RQ4) anomaly severity (duration) interpretation effectively?
  3. **Robustness to noise**
     - ✔ (RQ5) Compared with baseline methods, whether MSCRED is more robust to input noise?

Observation

False positive
False negative

The second best baseline method

The best baseline method

MSCRED

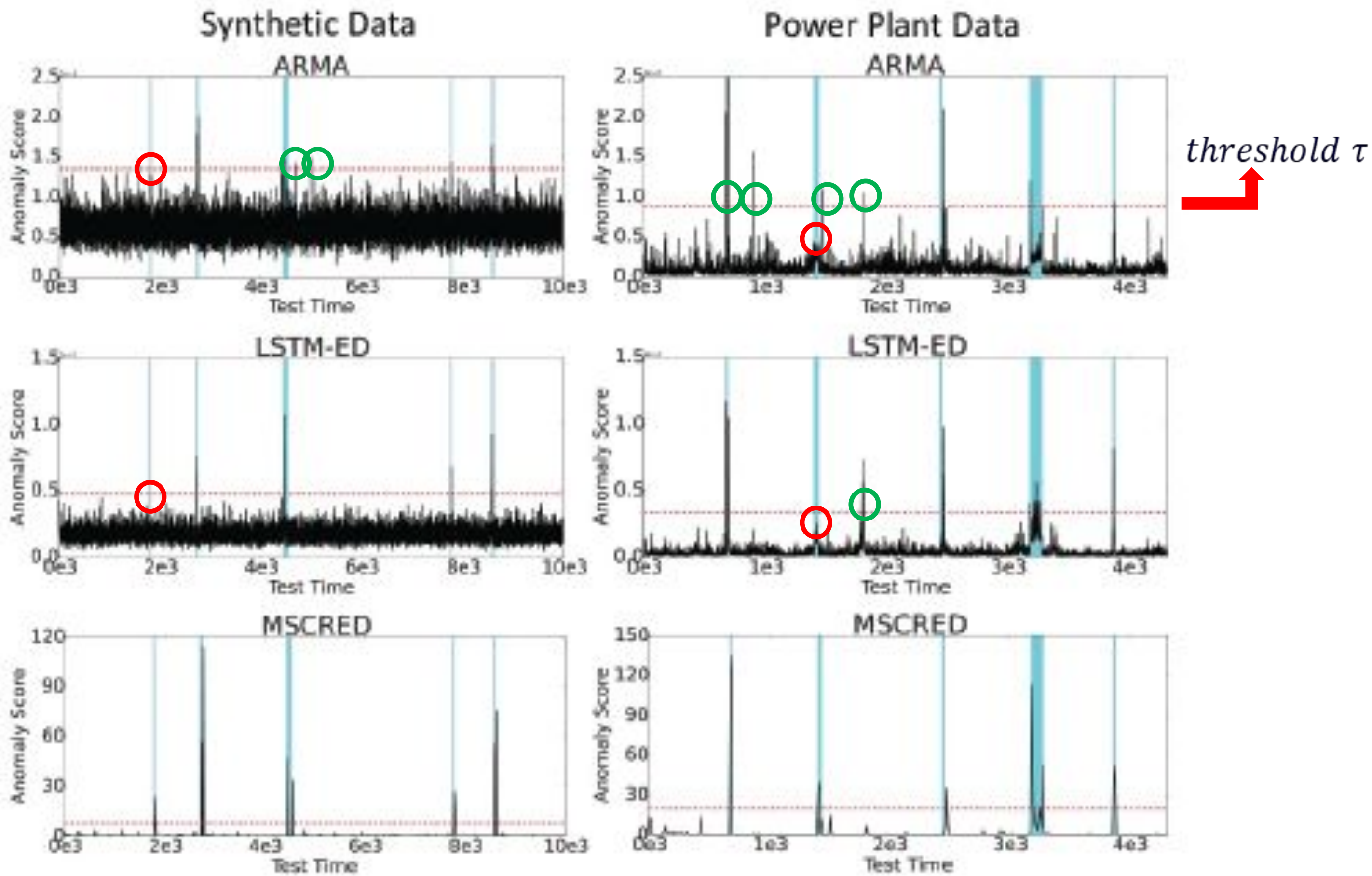


Figure 2 : Case study of anomaly detection. The shaded regions represent anomaly periods. The red dash line is the cutting threshold of anomaly

Observation

- (RQ1) comparison with baseline
  - In Table 1 (p.16)
    - ✔ MSCRED performs best on all settings
    - ✔ The improvements over the best baseline : 13.3% ~ 30.0%
  - In Figure 2 (p.18)
    - ARMA & LSTM-ED
      - ✔ The anomaly score is not stable
      - ✔ Many false positives and false negatives
    - MSCRED
      - ✔ The anomaly score is stable
      - ✔ Detect all anomalies without any false positive and false negative

Observation

- (RQ2) comparison with model variants
  - In Table 1 (p.16), by increasing the number of ConvLSTM layers,
    - $CNN^{ED(3,4)}_{ConvLSTM}$ outperforms $CNN^{ED(4)}_{ConvLSTM}$
    - $CNN^{ED}_{ConvLSTM}$ outperforms $CNN^{ED(3,4)}_{ConvLSTM}$

    → The effectiveness of ConvLSTM layers & stacked decoding process for model refinement

    - $CNN^{ED}_{ConvLSTM}$ is worse than MSCRED

    → The effectiveness of attention based ConvLSTM

Observation

- (RQ3) Root cause identification result
  - Rank all time series by their anomaly scores
  - Then, identify the top-k series as the root causes
  - MSCRED outperforms LSTM-ED by a margin of 25.9% and 32.4% in the synthetic and power plant data

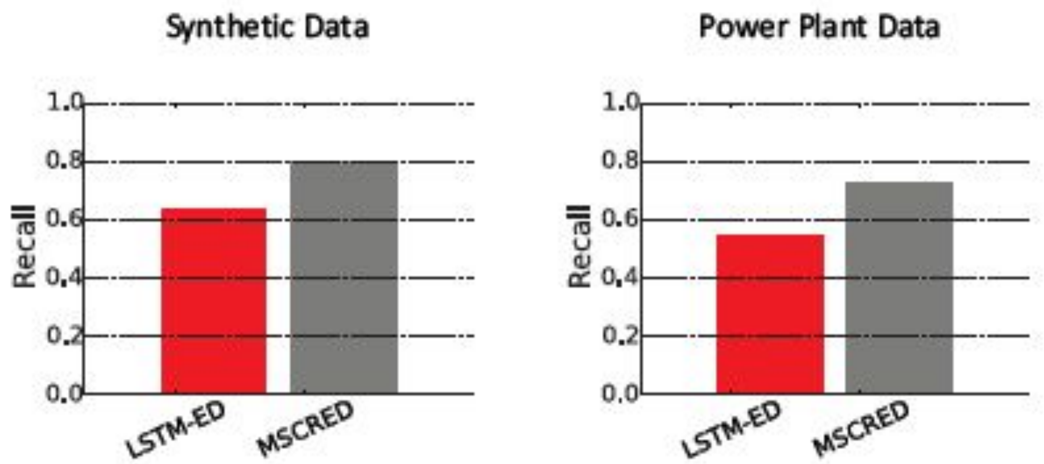


Figure 3 : Performance of root cause identification

Observation

- (RQ4) Anomaly severity (duration) interpretation
  - Compute anomaly scores based on the residual signature matrices of three channels
    - ✓ $\omega = 10$ (channel 1) – MSCRED(S)
    - ✓ $\omega = 30$ (channel 2) – MSCRED(M)
    - ✓ $\omega = 60$ (channel 3) – MSCRED(L)
  - Evaluate their performances on three types of anomalies
    - ✓ The duration of 10 = short anomaly
    - ✓ The duration of 30 = medium anomaly
    - ✓ The duration of 60 = long anomaly

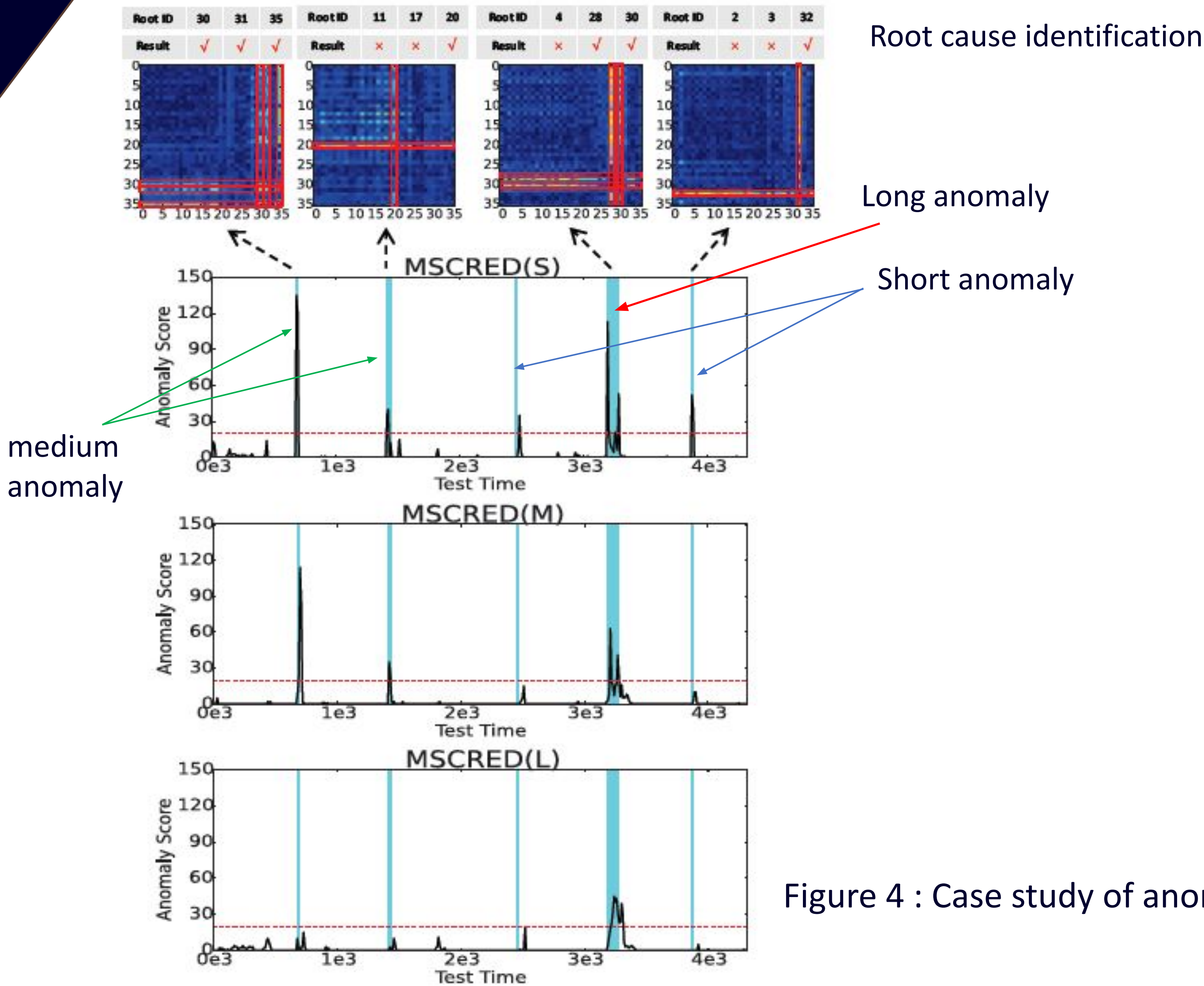


Figure 4 : Case study of anomaly diagnosis

Observation

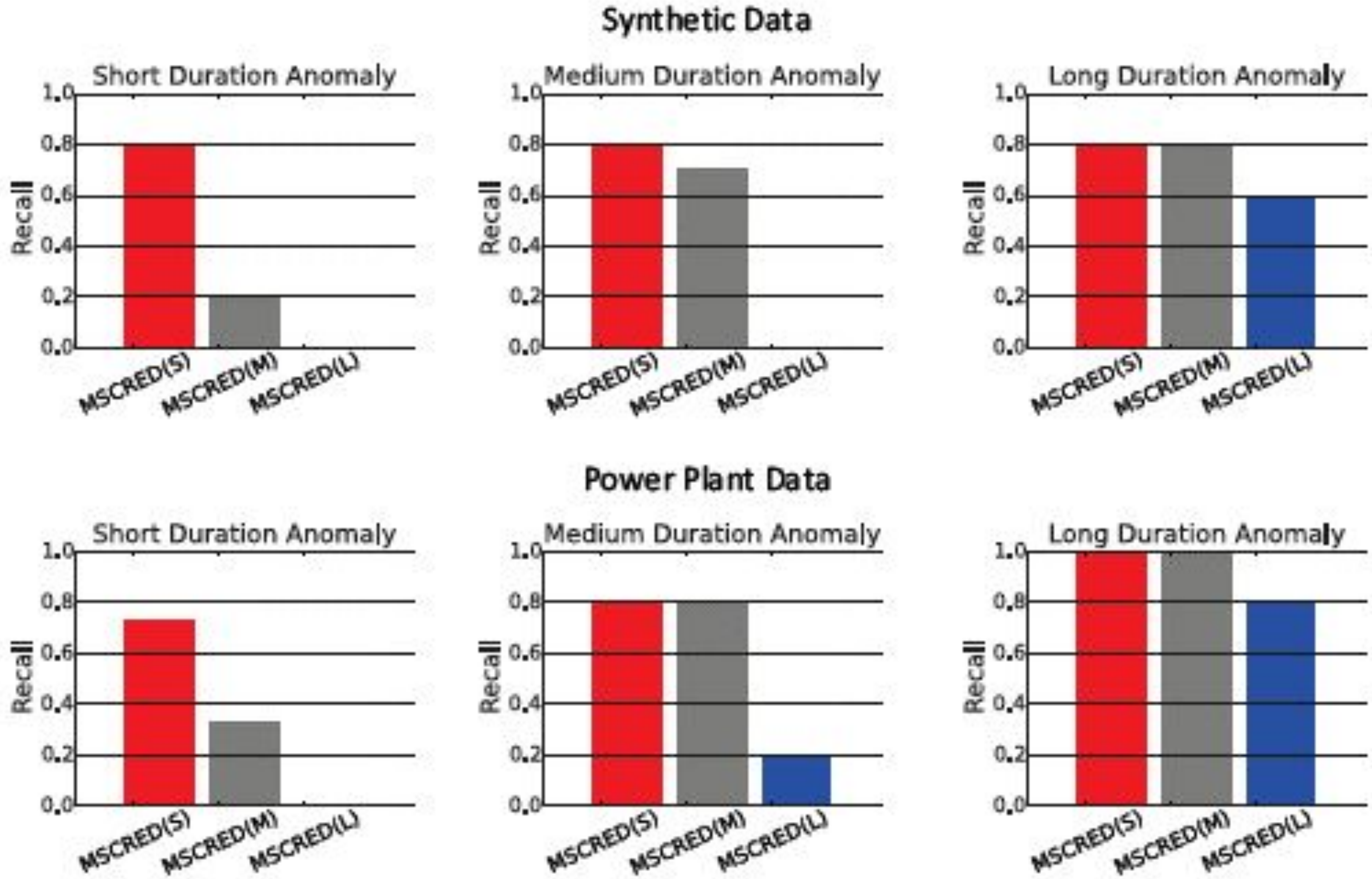
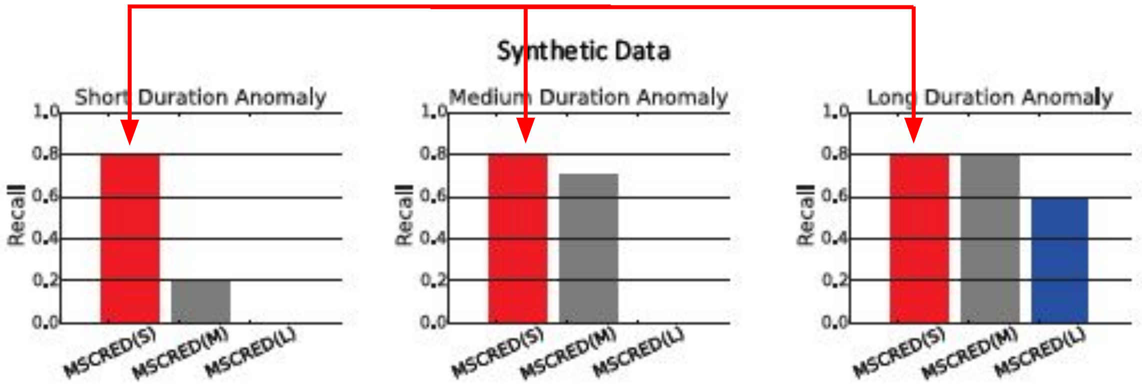


Figure 5 : Performance of 3 channels of MSCRED over different types of anomalies

Observation

- (RQ4) Anomaly severity (duration) interpretation
  - MSCRED(S)
    - Detect all types of anomalies

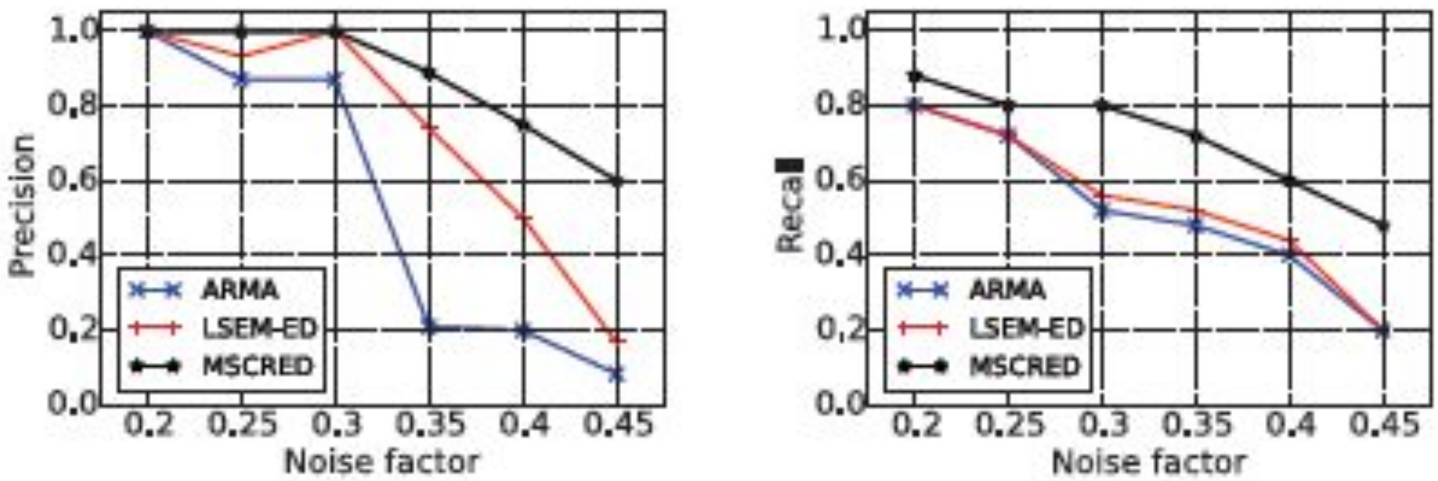


  - MSCRED(M)
    - Detect both medium and long duration anomalies
  - MSCRED(L)
    - Detect the long duration anomaly

  - For anomaly severity interpretation, we consider the three anomaly scores
    - eg) if the anomaly is detected in three channels, the anomaly is likely to be long duration

Observation

- (RQ5) Robustness to Noise
  - MSCRED consistently outperforms ARMA and LSTM-ED



The scale of noise varies from 0.2 to 0.45

Figure 6 : Impact of data noise on anomaly detection

Conclusion

- This paper proposed MSCRED for **anomaly detection and diagnosis in multivariate time series**
- MSCRED not only detects anomalies in certain time steps, but also identifies the root causes of anomalies and provides the interpretations of anomaly severity
- MSCRED employs multi-scale signature matrices and a deep encoder-decoder framework to reconstruct the signature matrices
- MSCRED is able to model both inter-sensor correlations and temporal dependencies in multivariate time series
- However, MSCRED cannot **model correlations between 3 or more elements** since the signature matrices calculate the correlation values between 2 elements
- It needs further study to solve this issue

References

[1] Hautamaki, V.; ¨ Ka¨rkka¨ınen, I.; and Franti, P. 2004. Outlier detection using ¨ k-nearest neighbour graph. In ICPR, 430–433.

[2] Manevitz, L. M., and Yousef, M. 2001. One-class svms for document classification. J. Mach. Learn. Res. 2(Dec):139–154.

[3] Zhou, C., and Paffenroth, R. C. 2017. Anomaly detection with robust deep autoencoders. In KDD, 665– 674.

[4] Gunnemann, N.; ¨ Gunnemann, S.; and Faloutsos, C. 2014. Robust multivariate ¨ autoregression for anomaly detection in dynamic product ratings. In WWW, 361–372.

[5] Hallac, D.; Vare, S.; Boyd, S.; and Leskovec, J. 2017. Toeplitz inverse covariance-based clustering of

[6] Qin, Y.; Song, D.; Chen, H.; Cheng, W.; Jiang, G.; and Cottrell, G. 2017. A dual-stage attention-based recurrent neural network for time series prediction. In *IJCAI*.

[7] Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Thank you