Contribution Title

Han, Gao First Author[1][0000-1111-2222-3333] and Second Author[2][1111-2222-3333-4444]

[1] University
[2]

Abstract. Unsupervised anomaly detection of multivariate time series data is of great significance for practical applications. The existing anomaly detection methods mainly adopt a fixed-length sliding window to extract windowed data from the observation data, and then perform deep learning encoding training and anomaly detection on each windowed data individually. However, using a single fixed-length window of data makes it difficult to simultaneously detect anomalies of different scales in the time series, such as small-scale point anomalies and large-scale contextual anomalies. Additionally, the patterns in multivariate time series data may be more complex and diverse.Based on these issues, we propose a multi-scale, multivariate unsupervised time series anomaly detection model(MMTSAD). We use downsampling to obtain coarse-grained and fine-grained sequences from the original data, and design two AutoEoder anomaly detection modules based on Attention mechanism to learn time series patterns at different scales and detect anomalies at different scales respectively. In addition, we introduce GAT module in the coarse-grained AutoEoder to capture the correlation between different variables.In the detection stage, in order to comprehensively consider the anomaly scores obtained from different scale models, we propose an anomaly score fusion method.We conduct experiments on five real-world publicly available datasets and show that MMTSAD outperforms previous methods. We also conduct ablation experiments to verify the effectiveness of the key modules of our proposed model.

Keywords: Time-series, Anomaly Detection.

1 Introduction

Unsupervised anomaly detection methods have received considerable attention in various anomaly detection techniques in recent research. This is because obtaining large-scale labeled anomaly data is extremely difficult or prohibitively costly. Unsupervised methods, on the other hand, can utilize time series data collected under arbitrary conditions as training samples. By learning general patterns and features from the time series data, they can effectively detect anomalies in test time series data that may contain any type of anomaly pattern.

Most unsupervised time series anomaly detection methods choose to compute an anomaly score for each time point and then compare this score with a certain threshold. This method is also adopted in this paper.

In terms of foundational model types, the current deep learning models applied to time series anomaly detection primarily include CNNs[1], RNNs[2,3,4,5,6], GNNs[7,8,9], as well as attention-based models[10,11,12,13,14], and hybrid models[8,9].

Regarding the sources of anomaly scores, the mainstream approaches for time series anomaly detection currently comprise two main methods: reconstruction[4,5,10,11,14] and prediction[2,3,7,9].Reconstruction-based methods involve reconstructing the original data and comparing it with the actual data.Prediction-based methods entail predicting the next one or more values in the time series and comparing them with the actual values. Alternatively, a combination of these two methods can also be employed[8].

In practical applications, anomalies in multivariate time series data are diverse, and their causes are difficult to exhaustively enumerate. However, from the perspective of data manifestation, they can mainly be categorized into point anomalies, contextual anomalies, and inter-variable anomalies.

Point anomalies typically belong to fine-grained anomalies, with short durations. Models only require short contextual sequences around the anomalous points to detect them. Contextual anomalies and inter-variable anomalies, on the other hand, generally belong to coarse-grained anomalies with longer durations. Therefore, models need longer contextual sequences to detect such trend anomalies.

Hence, conventional methods using single fixed-length window data as model input struggle to simultaneously detect anomalies of different scales in practical applications. However, due to the principles of the models themselves, it is not cost-effective to significantly increase the length of input sequences at once. For instance, RNN-based models, owing to their inherent temporal dependencies, rely on computations from previous time steps, making effective parallel computation difficult. Increasing the input sequence length would significantly increase the time and memory consumption for model training and inference.

Transformer-based models with self-attention mechanisms have been widely used for modeling long sequences. However, due to the computational complexity of attention matrix calculations, both time and space complexities grow quadratically with the length L of the input sequence. Therefore, simply increasing the input sequence length would lead to a surge in model parameters, resulting in rapid growth in time consumption and memory usage for model training and inference.

Moreover, with the increasing frequency of time series data collection in real production, there are higher demands on the time consumption of anomaly detection model inference. Our proposed MMTSAD model can significantly increase the receptive field of the model for time series data without additional time consumption constraints and simultaneously handle multi-scale data.

Furthermore, in practical applications, more often, multivariate time series data is encountered, which may introduce complexity and diversity in the temporal patterns. This imposes high demands on the model's ability to capture inter-variable correlations.

The MMTSAD model addresses this by introducing GAT modules in the AutoEncoder to ensure that the model can simultaneously learn both temporal and spatial dependencies.

The contribution of our paper is summarised as follows:

- A parallel multi-scale anomaly detection model is proposed, which can simultaneously learn the long-term and short-term patterns of sequences without additional time overhead, and detect anomalies of different scales.
- A correlation representation module based on coarse-grained sequences is designed to capture dependencies between variables while avoiding interference from temporal lag between variables.
- A anomaly score fusion method is proposed, which can comprehensively consider the anomaly detection results obtained from models at different scales

2 Related Work

As an important practical problem, many methods have been proposed for the unsupervised time series anomaly detection task. If the anomaly is classified according to the judgment criteria, the existing methods mainly include density estimation based methods, clustering based methods, reconstruction based methods and prediction based methods.

For example, density-based methods include mppccad[20] and DAGMM[21], which integrate Gaussian mixture models to estimate the density of representations.

In cluster-based methods, the anomaly score is usually obtained from the distance of the sample to the center of the cluster. And with the wide application of deep learning, some methods combine deep learning models with clustering methods for anomaly detection. Deep SVDD[22] collects representations of normal data into compact clusters. THOC[23] fuses the multi-scale temporal features of the intermediate layers through a hierarchical clustering mechanism and detects anomalies through multi-layer distances.

Reconstruction-based models attempt to detect anomalies by reconstructing the model input data. The simplest reconstruction technique is to train a separate model for each channel in turn using UAE[24]. Park et al.[25] proposed the LSTM-VAE model, which uses an LSTM backbone for temporal modeling and a Variational Autoencoder (VAE) for reconstruction. OmniAnomaly proposed by Su et al.[26] further extends the LSTM-VAE model by adopting a normalization process and using reconstruction probabilities instead of anomaly scores for detection. Li et al. [6] InterFusion updates the backbone to a hierarchical VAE that simultaneously models both interdependencies and internal dependencies between multiple sequences. There are also improvements based on improving the performance of anomaly scores. For example, USAD uses two weighted reconstruction errors [27], OmniAnomaly uses "reconstruction probability" as an alternative anomaly score [26], MTADGAT combines prediction error and reconstruction probability [8], and TranAD uses comprehensive reconstruction error and discriminator loss as anomaly score [11].

3 Methods

3.1 Problem Definition

In unsupervised anomaly detection tasks, the training set comprises unlabeled multidimensional time series data $\mathrm{X}_{\mathrm{train}}$, while the test set consists of labeled data $\mathrm{y} = \{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_T\}, \mathrm{y}_t \in \{0, 1\}$ associated with multidimensional time series data $X_{test} = \{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_T\}$ which $\mathrm{y}_t = 0$ represents a normal state, and $\mathrm{y}_t = 1$ represents an anomalous state. The labels correspond one-to-one with the multidimensional time series data at each timestamp, indicating whether the data is anomalous at that timestamp.

During training, the model learns the data distribution and general characteristics. In the testing phase, the model computes anomaly scores $\mathrm{s} = \{\mathrm{s1}, \mathrm{s2}, \ldots, \mathrm{sT}\}$ for the multidimensional time series data at each timestamp and provides a threshold θ. The detection label $\hat{y}$ can be obtained by the following formula:

$$\hat{y}_t = \begin{cases} 1, & s_t > \theta \\ 0, & s_t \leq \theta \end{cases}, \quad t \in \{1, 2, \ldots, T\} \tag{1}$$

The optimization objective of anomaly detection algorithms is to make the detection labels $\hat{\mathrm{y}} = \{\widehat{y_1}, \widehat{y_2}, \ldots, \widehat{y_T}\}$ as consistent as possible with the true labels y.

3.2 Sequence Downsampling

To extend the model's receptive field for sequences within a certain transformer input length, we employ downsampling techniques to process the original granular data. We control the sampling method and granularity through two parameters: the aggregation function and the sampling factor.

We choose the mean function as the aggregation function because mean sampling can to some extent mitigate the influence of noise in the original granular sequence. The sampling factor depends on the original sequence length and sampling frequency. A larger sampling factor implies a larger receptive field for the model within the same input length. This aids the model in learning long-term trends in the data and identifying coarse-grained contextual anomalies.

3.3 Architecture of the MMTSAD

The structure of MMTSAD is illustrated in Figure(see 错误!未找到引用源。). we use AutoEncoder to reconstruct the input sequence, where Encoder is used to compress the original sequence to get the latent space representation z of input data x, Decoder accepts the latent space representation z as output to reconstruct the original input data x.

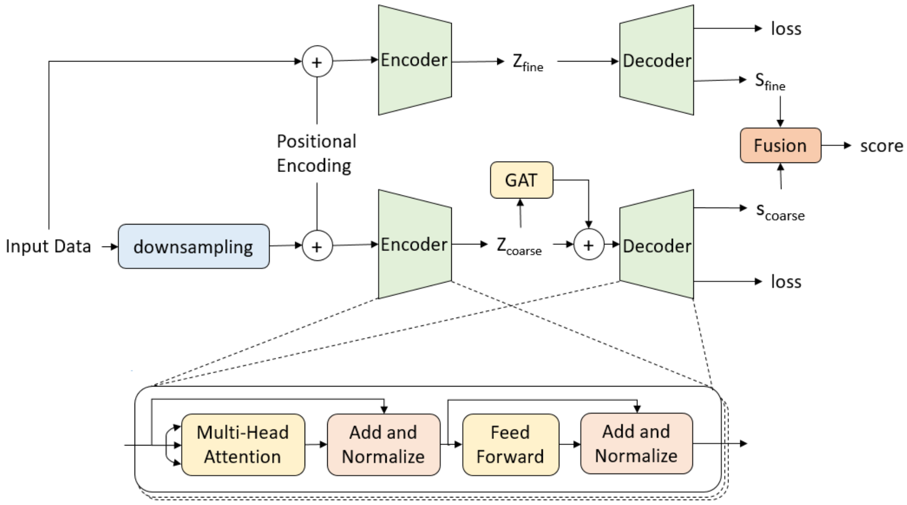

Fig. 1. Architecture of the MMTSAD

In a fine-grained autoencoder, the input sequence of the encoder and the output sequence of the decoder are of length L, which means that the receptive field of the model to the sequence is also L. In a coarse-grained autoencoder, the input sequence of the encoder and the output sequence of the decoder are of length L. That is, the coarse-grained autoencoder does not consume more parameters, memory, and computation time than the fine-grained autoencoder, and the receptive field of the model to the sequence can be greatly increased to the scale of L*Sampling_factor, which is important for the model to learn the long-term trends and dependencies in the data.

In fact, in addition to downsampling, there are some models that can also increase the sequence receptive field exponentially, such as pooling layers in models such as CNNS.However, because the model eventually needs to reconstruct the input data, considering the difficulty of model training optimization and reconstruction accuracy during reconstruction, this paper chooses downsampling to process data.

Due to the successful use of Transformer for sequence modeling, we use its Encoder part as the encoder and Decoder infrastructure for capturing temporal patterns within a window.The self-attention mechanism based on Transformer model is permutation invariant, so it cannot consider the order of time series data. Therefore, this paper uses location encoding to make the model perceive the location information represented by each time step in the input time series data.

We adopted GAT network[28] for auto-encoder intermediate variables to model dependencies between nodes. Given a graph with D nodes, namely$\{\mathrm{h}_1, \mathrm{h}_2, \cdots, \mathrm{h}_D\}$, where h_d is the feature vector for each node and the input to the GAT layer. The dimension L of the vector h_i represents the length of the window time series.The GAT layer computes the output representation for each node as follows:

$e_{i,j}$ represents the importance of the features of node j to node i, specifically to our problem which is the importance or dependence of the features of sequence j to sequence i.

$$e_{i,j} = LeakyReLU(\vec{a}[Wh_i||Wh_j])$$

Where W represents the weight matrix shared by each node, which is used as a learnable linear transformation to transform the input features into more general and high-level features. a is a shared single-layer feedforward neural network.

The coefficients calculated by the attention mechanism can be expressed as follows:

$$\alpha_{ij} = softmax\left(e_{i,j}\right) = \frac{\exp(e_{i,j})}{\sum_{d=1}^{D} \exp(e_{i,d})}$$

Based on the attention coefficients calculated above, the feature output h_i' aggregated at each node can be obtained:

$$h_i' = \sigma(\sum_{j=1}^{D} \alpha_{ij} W h_j)$$

To stabilize the learning process of self-attention, we also apply multi-head attention.

With the progress of observation equipment and technology, the number of variables in the recorded multivariate time series data is increasing. Some data sets may record the values of hundreds of variables. Obviously, using GAT network on such data will lead to a rapid increase of network parameters, and generally only a few of the variables have significant correlation with each other. Therefore, we use GAT on the intermediate hidden layer representation of the data. Compared with the original data, the intermediate variables of the compressed autoencoder have higher dimensional features and the redundant information is greatly reduced, so that GAT can learn the dependencies between variables with a very low number of parameters.

3.4 Anomaly score Fusion

Our model calculates the anomaly score at each time step based on the reconstruction difference, and the anomaly score is calculated as follows.

$$\Delta x_t^i = \hat{x}_t^i - x_t^i, \qquad s_t = \sum_{i=1}^{D} (\Delta x_t^i)^2$$

Where D represents the number of variables of the multivariate time series.

The anomaly score at each time step is obtained by the sum of squared differences Δx_t^i between the reconstructed values and the actual values for each dimension in that time step. Because the reconstructed values of the model fluctuate around the actual values, under normal conditions, the difference between the reconstructed values and the actual values is relatively small, while under abnormal conditions, the difference is larger. Moreover, normal states account for the vast majority of the entire sequence, while abnormal states are less frequent. Therefore, we can consider the difference between the reconstructed value and the actual value for each dimension at a time step as a random variable approximately following a normal distribution with a mean of 0.

$$\Delta x_t^i \sim N(0, (\sigma)^2)$$

Where σ represents the scale parameter of the sequence reconstruction error.

$$s_t \sim \chi^2(\mathrm{v}), \quad v \leq D$$

Where v represents the degrees of freedom of the multivariate sequence variable.

$$\mathrm{score}_t = s_{fine_t} + \frac{{\sigma_{fine}}^2}{{\sigma_{coars}}^2} s_{coars_t}$$

After transformation, we can add the coarse-grained anomaly score with the fine-grained anomaly score to obtain a unified single-time-step scale anomaly score. Such anomaly scores fairly take into account the possibility of anomalies at different scales at a time step.

4 Experiments

4.1 Datasets

In our experiments, we utilized five publicly available datasets.

- Water Distribution (WADI) dataset[15]: The WADI dataset is acquired from a reduced city water distribution system with 123 sensors and actuators operating for 16 days. The first 14 days contain only normal data, while the remaining two days have 15 anomaly segments.
- Secure Water Treatment (SWaT) dataset[16]: The SWaT dataset is collected over 11 days from a scaleddown water treatment testbed with 51 sensors. During the last 4 days, 41 anomalies were injected using diverse attack methods, while only normal data were generated during the first 7 days.
- Pooled Server Metrics (PSM) dataset[17]: The PSM dataset is collected internally from multiple application server nodes at eBay. There are 13 weeks of training data and 8 weeks of testing data.
- Mars Science Laboratory (MSL) dataset[18]: The MSL dataset is a dataset similar to SMAP but corresponds to the sensor and actuator data for the Mars rover itself .
- Soil Moisture Active Passive (SMAP) dataset[19]: The SMAP dataset is a dataset of soil samples and telemetry information using the Mars rover by NASA.

We summarized the statistical characteristics of these public datasets in Table 1. For datasets with more than one entity, the lengths of the training and test sets were the sum of the lengths of all entity sequences. Additionally, the anomaly rate in the test set was calculated as the ratio of the total number of anomalous points in all entity test sets to the total length of the test set sequences.

Table 1. Dataset Statistics.

Dataset	Train	Test	Entity	Dimensions	Anomalies
WADI	1048571	172801	1	123	5.99%
SWaT	495000	449919	1	51	11.98%
PSM	132481	87841	1	25	27.76%
MSL	58317	73729	27	55	10.48%
SMAP	140825	444035	35	25	12.83%

4.2 Main results

We report the results for F1* on five datasets in Table 2.

Table 2. Performance of our models and baselines.

Algorithm \Dataset	WADI	SWaT	PSM	MSL	SMAP
PCA	0.156	0.676	0.535	0.187	0.213
DeepSVDD	0.322	0.724	0.601	0.212	0.203
LSTM-VAE	0.227	0.776	0.455	0.212	0.235
DAGMM	0.121	0.750	0.483	0.199	0.333
MSCRED	0.046	0.757	0.556	0.25	0.170
OmniAnomaly	0.223	0.782	0.452	0.207	0.227
MTAD-GAT	0.437	0.784	0.571	0.275	0.296
GDN	0.570	0.810	0.552	0.217	0.252
TranAD	0.415	0.669	0.649	0.251	0.247
AnomalyTransformer	0.108	0.220	0.434	0.191	0.227
D3R	0.613	0.781	**0.760**	0.510	0.473
NPSR	<u>0.642</u>	<u>0.839</u>	0.648	**0.551**	<u>0.505</u>
Ours	**0.657**	**0.850**	0.635	<u>0.542</u>	**0.521**

MMTSAD achieves leading performance on most datasets. For the multi-entity dataset, we train and test each entity separately, and finally summarize the performance of all entities in the dataset on the test set. The F1* score is calculated according to the test set time points of all entities. Because the optimal F1 score based on point adjustment has limitations, we adopted the results without point adjustment.

Ablation Study

We conducted ablation experiments on five publicly available datasets, as shown in Table 3. Ablation experiments are conducted to study different scale modules in MMTSAD, and it can be seen that on all datasets, multi-scale MMTSAD achieves better F1* scores than single scale, which illustrates the excellent results when using multi-scale anomaly detection.

It is found that in the case of using a single scale, the fine-grained module is better than the coarse-grained module on the WADI, MSL and SMAP datasets, and the coarse-grained module is better than the fine-grained module on the SWaT and PSM datasets.This shows that anomalies from different datasets may tend to be of different scales and hence the need to use a multi-scale module.

Table 3. Ablation experiments on multiple scales.

Model	WADI	SWaT	PSM	MSL	SMAP
MMTSAD	0.657	0.850	0.635	0.542	0.521
w/o coarsness	0.650	0.828	0.625	0.539	0.519
w/o fine	0.642	0.849	0.628	0.532	0.502

5 Conclusion

We propose a multi-scale parallel unsupervised multivariate time series anomaly detection model (MMTSAD), which can perform multi-scale anomaly detection and diagnosis on multivariate time series data.Our results show that MMTSAD exhibits excellent anomaly detection effect, strong applicability, and has a relatively simple training and detection process.Our model can make fault monitoring faster and more automated, and is suitable for both batch and stream processing detection requirements.

References

1. He, Y., & Zhao, J. (2019, June). Temporal convolutional networks for anomaly detection in time series. In Journal of Physics: Conference Series (Vol. 1213, No. 4, p. 042050). IOP Publishing.
2. Wu, W., He, L., Lin, W., Su, Y., Cui, Y., Maple, C., & Jarvis, S. (2020). Developing an unsupervised real-time anomaly detection scheme for time series with multi-seasonality. IEEE Transactions on Knowledge and Data Engineering, 34(9), 4147-4160.
3. Shen, L., Li, Z., & Kwok, J. (2020). Timeseries anomaly detection using temporal hierarchical one-class network. Advances in Neural Information Processing Systems, 33, 13016-13026.
4. Abdulaal, A., Liu, Z., & Lancewicki, T. (2021, August). Practical approach to asynchronous multivariate time series anomaly detection and localization. In Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining (pp. 2485-2494).
5. Zhang, Y., Wang, J., Chen, Y., Yu, H., & Qin, T. (2022). Adaptive memory networks with self-supervised learning for unsupervised anomaly detection. IEEE Transactions on Knowledge and Data Engineering.
6. Li, Z., Zhao, Y., Han, J., Su, Y., Jiao, R., Wen, X., & Pei, D. (2021, August). Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding. In Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining (pp. 3220-3230).
7. Deng, A., & Hooi, B. (2021, May). Graph neural network-based anomaly detection in multivariate time series. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 5, pp. 4027-4035).
8. Zhao, H., Wang, Y., Duan, J., Huang, C., Cao, D., Tong, Y., ... & Zhang, Q. (2020, November). Multivariate time-series anomaly detection via graph attention network. In 2020 IEEE International Conference on Data Mining (ICDM) (pp. 841-850). IEEE.
9. Chen, Z., Chen, D., Zhang, X., Yuan, Z., & Cheng, X. (2021). Learning graph structures with transformer for multivariate time-series anomaly detection in IoT. IEEE Internet of Things Journal, 9(12), 9179-9189.
10. Xu, J., Wu, H., Wang, J., & Long, M. (2021). Anomaly transformer: Time series anomaly detection with association discrepancy. arXiv preprint arXiv:2110.02642.
11. Tuli, S., Casale, G., & Jennings, N. R. (2022). Tranad: Deep transformer networks for anomaly detection in multivariate time series data. arXiv preprint arXiv:2201.07284.
12. Li, Y., Peng, X., Zhang, J., Li, Z., & Wen, M. (2021). Dct-gan: Dilated convolutional transformer-based gan for time series anomaly detection. IEEE Transactions on Knowledge and Data Engineering.

13. Song, J., Kim, K., Oh, J., & Cho, S. (2024). Memto: Memory-guided transformer for multivariate time series anomaly detection. Advances in Neural Information Processing Systems, 36.
14. Lai, C. Y. A., Sun, F. K., Gao, Z., Lang, J. H., & Boning, D. (2024). Nominality score conditioned time series anomaly detection by point/sequential reconstruction. Advances in Neural Information Processing Systems, 36.
15. Ahmed, C. M., Palleti, V. R., & Mathur, A. P. (2017, April). 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 (pp. 25-28).
16. Goh, J., Adepu, S., Junejo, K. N., & Mathur, A. (2017). A dataset to support research in the design of secure water treatment systems. In Critical Information Infrastructures Security: 11th International Conference, CRITIS 2016, Paris, France, October 10–12, 2016, Revised Selected Papers 11 (pp. 88-99). Springer International Publishing.
17. Abdulaal, A., Liu, Z., & Lancewicki, T. (2021, August). Practical approach to asynchronous multivariate time series anomaly detection and localization. In Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining (pp. 2485-2494).
18. Entekhabi, D., Njoku, E. G., O'neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. N., ... & Van Zyl, J. (2010). The soil moisture active passive (SMAP) mission. Proceedings of the IEEE, 98(5), 704-716.
19. Hundman, K., Constantinou, V., Laporte, C., Colwell, I., & Soderstrom, T. (2018, July). Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 387-395).
20. Yairi, T., Takeishi, N., Oda, T., Nakajima, Y., Nishimura, N., & Takata, N. (2017). A data-driven health monitoring method for satellite housekeeping data based on probabilistic clustering and dimensionality reduction. IEEE Transactions on Aerospace and Electronic Systems, 53(3), 1384-1401.
21. Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., & Chen, H. (2018, February). Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In International conference on learning representations.
22. Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., ... & Kloft, M. (2018, July). Deep one-class classification. In International conference on machine learning (pp. 4393-4402). PMLR.
23. Shen, L., Li, Z., & Kwok, J. (2020). Timeseries anomaly detection using temporal hierarchical one-class network. Advances in Neural Information Processing Systems, 33, 13016-13026.
24. Garg, A., Zhang, W., Samaran, J., Savitha, R., & Foo, C. S. (2021). An evaluation of anomaly detection and diagnosis in multivariate time series. IEEE Transactions on Neural Networks and Learning Systems, 33(6), 2508-2517.
25. Park, D., Hoshi, Y., & Kemp, C. C. (2018). A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder. IEEE Robotics and Automation Letters, 3(3), 1544-1551.
26. Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., & Pei, D. (2019, July). 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 (pp. 2828-2837).
27. Audibert, J., Michiardi, P., Guyard, F., Marti, S., & Zuluaga, M. A. (2020, August). Usad: Unsupervised anomaly detection on multivariate time series. In Proceedings of the 26th

ACM SIGKDD international conference on knowledge discovery & data mining (pp. 3395-3404).

28. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. stat, 1050(20), 10-48550.

29.