ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747274

STGAT-MAD : SPATIAL-TEMPORAL GRAPH ATTENTION NETWORK FOR MULTIVARIATE TIME SERIES ANOMALY DETECTION

Jun Zhan[1], *Siqi Wang*[1], *Xiandong Ma*[2], *Chengkun Wu*[1], *Canqun Yang*[1,3],*Detian Zeng*[1], *Shilin Wang*[4]

[1]College of Computer Science, National University of Defense Technology, Changsha, P.R. China
[2]Engineering Department, Lancaster University, Lancaster, LA1 4YW, UK
[3]National SuperComputing Center of Tianjin, P.R. China
[4]Beijing Goldwind Smart Energy Technology Co., Ltd. Beijing, P.R. China

ABSTRACT

Anomaly detection in multivariate time series data is challenging due to complex temporal and feature correlations. This paper proposes a novel unsupervised multi-scale stacked **s**patial-**t**emporal **g**raph **at**tention network for **m**ultivariate time series **a**nomaly **d**etection (STGAT-MAD). The core of our framework is to coherently capture the feature and temporal correlations among multivariate time-series data by stackable STGAT networks. Meanwhile, a multi-scale input network is exploited to capture the temporal correlations in different time-scales. Besides, a new dataset derived from a real-world wind farm is built and released for multivariate time series anomaly detection. Experiments on the proprietary dataset and three public datasets show that our method significantly outperforms existing baseline approaches, and provides interpretability for anomaly location.

Index Terms— Multivariate Time Series, Anomaly detection, Spatial-Temporal Graph Attention Network

1. INTRODUCTION

Anomaly detection in multivariate time series is an important research area in data mining and provides a vital basis for intelligent operation and maintenance [1]. Time series data are pervasive in our production facilities and life, such as running data of mechanical equipment and network intrusion data, which reflect the intrinsic activity of a system. Normally, the patterns will not change drastically, and vice versa. Multivariate time series are collected from independent sensors with complex coupling relations. Therefore, each variable depends on its historical and other sensors values, which poses a great challenge to anomaly detection.

For the earlier studies, most researchers solved the problems of anomaly detection in multivariate time series data by using the methods of statistical analysis or autoregressive [2]. However, these methods generally estimate anomalies based on the overall data distribution, which are inadequate for the requirements of increasing data complexity and dimensions. With the development of deep learning-based techniques, many deep models are applied and achieve satisfactory performances [3], which can usually be categorized into *prediction-based* and *reconstruction-based* methods. The former focus on contextual anomalies, and the typical models include long-short term memory (LSTM) [4], convolutional LSTM (ConvLSTM) [5], etc. While the latter focus on overall distribution anomalies, and the frequently-used methods are Autoencoder-based methods (AE) [6] and generative adversarial networks (GANs) [7]. However, most methods lack the ability to explicitly learn the inter-relationship of sensors, increasing instability when dealing with high-dimensional data with abundant potential correlations.

Recently, graph neural network(GCN) has demonstrated its effectiveness in dealing with complex graph structure data. Inspired by these methods, we utilize graph networks to extract complex spatial-temporal correlations from multivariate time series data. This idea faces three challenges: 1) Only normal data are available for training. 2) When temporal features are captured, the correlations among variables should be considered. 3) Multivariate time series data usually show heterogeneity (e.g. concurrent discrete and continuous values), and high-level implicit features are difficult to capture. Therefore, we propose STGAT-MAD framework. Our main contributions are summarized below :

- We for the first time propose to exploit the multi-scale temporal correlations of multivariate time-series input data for anomaly detection.
- A novel stackable STGAT network is designed for coherently capturing the feature and temporal correlations among multivariate time-series data.
- A new wind turbine dataset is built and released for multivariate time series anomaly detection, derived from a real-world wind farm. Our code and dataset are available at: https://github.com/zhanjun717/STGAT.
- Extensive experiments show that the performance of our model is improved by up to 13% on F1 and provides good interpretability, i.e., the underlying correlations among different features and temporal.

2. RELATED WORK

Unsupervised anomaly detection: Due to the scarcity of anomalies, anomaly detection is usually addressed by unsupervised methods that leverage pure normal data for training. One classical type is the statistical-based models [2]. For instance, box plots [8], and gaussian mixture model (GMM) [9] model the data according to statistical characteristics. Similarly, the clustering-based models, e.g., [10] detect attacks considering the distance to the cluster center. However, these methods may not achieve satisfactory performance since the temporal correlations are captured ambiguously. Autoregressive is a method for specially handling time series, e.g., autoregressive moving average (ARMI) [11], but generally requires data to be autocorrelative.

Deep learning-based methods: To extract complex patterns implied in multivariate time series, deep learning methods are pervasively adopted. Autoencoder-based methods (AE) are the commonly-used techniques, e.g., VAE-LSTM [6, 3, 12, 13, 14], which recognize abnormal data according to reconstruction error. In addition, the GANs-based methods, e.g., MAD-GAN [7], and USAD [15] use both the discrimination and reconstruction errors for anomaly detection, which also improve the stability of networks.

Spatial-Temporal Networks: Recently, graph neural network (GNN) achieves significant progress in dealing with complex spatial-temporal data, e.g., Spatio-Temporal Graph Convolutional Networks(STGCN)[16], attention based spatial-temporal graph convolutional network (ASTGCN) [17]. From a broader perspective, multivariate time series data is spatial-temporal data in essence. Accordingly, Graph Convolutional Networks(GCN) [18], MTAD-GAT [19] and GAT [20] apply GNN into the correlations extraction of multivariate time series. However, the temporal and feature correlations in different time-scales are prone to be ignored.

3. METHODOLOGY

3.1. Problem Formulation

Given a time series $X = \{x_1, x_2, \cdots, x_T\}$ with length T, where $x_t \in \mathbb{R}^d$ is a d–dimensional vector collected at each time t, we first use sliding window with length w to process long sequence into subsequence set $S = \{s_1, s_2, \cdots, s_N\}$, where N is the number of the subsequence. Each subsequence s_n is a subset of X, that is $s_n \subseteq X$. The task of anomaly detection is to reconstruct subsequences. Finally, the deviation of reconstruction value and actual value is quantified as an anomaly score for abnormal discrimination.

3.2. Proposed Framework

As shown in Fig.1, the basic framework of STGAT-MAD contains four core components—**multi-scale input network** to obtain different receptive field features, **stacked STGAT network** to extract high-level implicit feature and temporal correlations, **data fusion and reconstruction network** to reconstruct input data according to the implicit features, **anomaly assessment module** to distinguish the abnormality and provide an explanation. At the former three networks, complex spatial-temporal correlations between multivariate time series is extracted, and finally, signal reconstruction is achieved. At the anomaly assessment module, abnormal data are usually not isolated but form a continuous abnormal segment, hence, we use point-adjustment method widely used in [13, 21, 15].

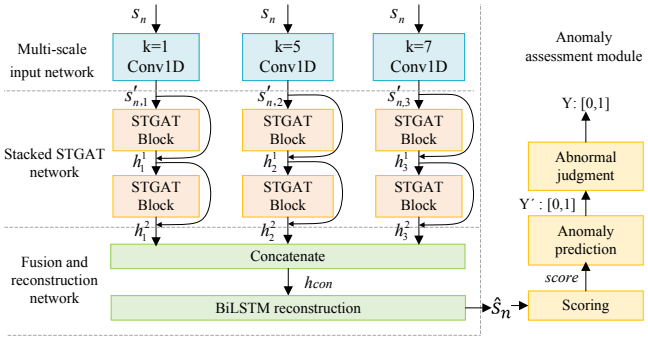


Fig. 1. STGAT-MAD basic framework.

Multi-scale input network: For multivariate time series signals, the data with various time scales contains various information [22]. As shown in Fig.1, we set three input channels composed of convolution units with different kernels. For input subsequence s_n, the multi-scale feature information matrix can be calculated by:

$$s'_n(k, step) = \sigma\left(Conv1D_{k,step}(s_n, W) + b\right) \quad (1)$$

where σ denotes $ReLU$ activation function. $k \in \{1, 5, 7\}$ is the size of convolution kernel, $step = 1$ is convolution step. W and b are weight and bias, respectively.

Stacked STGAT network: To extract spatial-temporal correlations in the feature and temporal dimensions, we respectively map data from these two dimensions into graph structure data. As shown in the Fig.2, in the feature dimension, the subsequences s_n at the moment t containing d dimension are expressed as weighted undirected graph $\mathbb{G}_{x_t} = (V, E^f)$, where $V = \{v_d | d \in [1, D]\}$ is node set, and E^f is edge set. Arbitrary two nodes i and j have connection relations. Adjacent matrix is denoted as $A^f_{ij} = 1,\ for\ i; j \in [1, D]$. In the temporal dimension, we denote subsequences s_n as new representations $\mathbb{G}_{x_v} = (W, E^t)$, where $W = \{x_{t-k} | k \in [0, w-1]\}$ is node set, k denotes the time interval between the current and last node x_t, and w is the corresponding window size of input subsequence s_n. E^t represents edge set of the corresponding nodes at different moment t and the nodes at other moments. Discriminated from the feature graph, we consider there is no connection between the current node and the future nodes. Hence, the connection relation be-

tween the moments m and n can be expressed as:

$$A_{mn}^{t} = \begin{cases} 1, m \geq n \\ 0, m < n \end{cases} \quad \text{for } m; n \in [1, w] \tag{2}$$

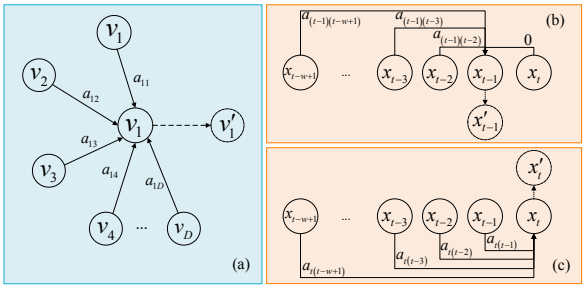


Fig. 2. The spatial-temporal graph structure, where the dotted line denotes the aggregate representation of the node. (a) indicates feature dimension. (b) and (c) represent temporal dimension.

Fig.3 is the basic framework of the STGAT block. To optimize training, the residual connection is added between blocks. Input subsequence after STGAT block processing can coherently explore the correlations of feature dimension and temporal dimension. In the feature dimension, for $V = \{\vec{v}_1, \vec{v}_2, \cdots, \vec{v}_D\}, \vec{v}_i \in \mathbb{R}^w$ in graph $\mathbb{G}_{x_t}$. After the STGAT block processing, attention coefficient can be computed through the following formula:

$$\alpha_{ij} = \frac{\exp\left(\delta\left(\overrightarrow{\mathrm{a}}^T\left[\mathbf{W}\vec{v}_i \| \mathbf{W}\vec{v}_j\right]\right)\right)}{\sum_{k\in \mathrm{N}_i}\exp\left(\delta\left(\overrightarrow{\mathrm{a}}^T\left[\mathbf{W}\vec{v}_i \| \mathbf{W}\vec{v}_k\right]\right)\right)} \tag{3}$$

where δ is $LeakyReLU$ activation function[23]. $\overrightarrow{\mathrm{a}} \in \mathbb{R}''$

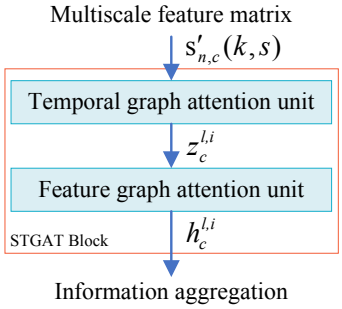


Fig. 3. The basic framework of the STGAT block, which is connected by temporal and feature attention units.

is learnable weight vector, $\mathbf{W}$ is shared weight matrix, and $\|$ denotes the split joint of two nodes information. Finally, the output of each node can be gained by aggregating its neighboring nodes, as shown in $\boldsymbol{z}_c^{l,i}$ of Fig.3.

$$\boldsymbol{z}_c^{l,i} = \sigma\left(\frac{1}{H}\sum_{h=1}^{H}\sum_{\mathrm{j}\in N_i} a_{ij}^k \mathbf{W}^k \vec{v}_j\right) \tag{4}$$

where c and l are the channel number and layer number of STGAT block, respectively. i is node number, H denotes the number of multi-head attention mechanism. To facilitate the training, we adopt average method to aggregate the results of multi-head attention mechanism.

In the temporal dimension, we adopt the same method to capture temporal correlation in different time periods. The computational methods of attention coefficient and aggregate expression are similar to formulas (3) and (4). Therefore, we can obtain attention coefficient of temporal dimension β_{ij} and output of each node $\boldsymbol{h}_c^{l,i}$.

Fusion and reconstruction network: After the spatial-temporal feature vectors in multi-channel are concatenated, they are reconstructed by the BiSLTM-AE reconstruction network, which is expressed as f_{recon} according to the shape of original input s_N. The reconstructed vector is :

$$\hat{S}_n = f_{\text{recon}}\left(\|_{c=1}^{C}\left(h_c^L\right)\right) \tag{5}$$

here L is the largest layer number of the stack, C is the number of input channels. In this paper, $C = 3$.

Anomaly assessment module: By comparing the reconstruction sequence and the original sequence of input, we calculate the abnormal score of each sample as:

$$score = \frac{1}{D}\left\|x_t - \hat{x}_t\right\|_2 \tag{6}$$

where x_t and $\hat{x}_t$ correspond to the data at the latest moment in S_n and $\hat{S}_n$, respectively. we discriminate the data whose scores are larger than the threshold as anomalies, and the abnormal result is denoted by Y'. The final result Y can be obtained after point-adjustment [15]. Because the selection of threshold involves complicated expert knowledge, best evaluation results are reported in this paper.

4. EXPERIMENTAL STUDIES

4.1. Performance Evaluation

We evaluate our method on three public datasets (Secure Water Treatment (SWat) Dataset [24], Water Distribution (WADI) Dataset [25] and Server Machine Dataset (SMD) [21], and a private Wind Turbine Dataset (WTD). Precision, recall, F1 and AUC are chosen as evaluation indexes. The results are compared with the present advanced methods including LSTM-NDT [24], GDN [20], LSTM-VAE [12], USAD [15] and MTAD-GAT [19]. In the experiment, we set the STGAT block layer number $l = 2$, the sliding window size $w = 5$ under SWAT, $w = 60$ under WTD, and $w = 100$ under other datasets.

The results in Table 1 indicate that the STGAT-MAD method obtains the optimal F1 and AUC values on almost all datasets. In particular, F1 achieves a 13% improvement on the WTD dataset. In addition, our approach is superior to GDN and MTAD-GAT in extracting deeper implicit spatial-temporal features from multivariate data, resulting in the data reconstruction layer being better to reconstruct contextual information of data. Meanwhile, STGAT-MAD introduces

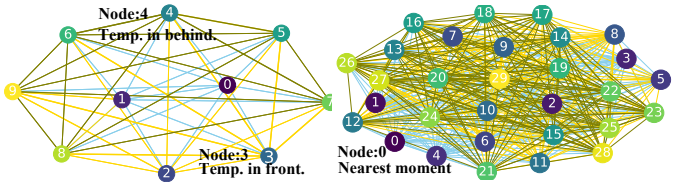
Table 1. Performance comparison of different methods and datasets

	SWat				WADI				SMD				WTD			
Models	Rec	Pre	F1	AUC	Rec	Pre	F1	AUC	Rec	Pre	F1	AUC	Rec	Pre	F1	AUC
LSTM-NDT[24]	0.707	0.990	0.825	0.884	0.906	0.602	0.724	0.615	0.990	0.661	0.796	0.890	0.755	0.530	0.623	0.784
GDN[20]	0.746	0.942	0.833	0.879	0.915	0.409	0.570	0.748	0.990	0.490	0.658	0.931	0.990	0.725	0.821	0.757
LSTM-VAE [12]	0.766	**0.979**	0.860	0.878	0.910	0.603	0.720	0.800	0.990	0.669	0.802	0.849	0.990	0.667	0.790	0.720
USAD[15]	**0.960**	0.347	0.510	0.755	0.834	0.159	0.267	0.647	0.979	0.447	0.615	0.908	**0.990**	0.484	0.652	0.565
MTAD-GAT[19]	0.821	0.903	0.860	0.855	0.518	0.720	0.602	0.687	0.944	0.875	0.908	**0.990**	0.720	0.829	0.771	0.908
STGAT-MAD	0.841	0.965	**0.900**	**0.903**	**0.910**	**0.797**	**0.849**	**0.804**	**0.990**	**0.964**	**0.982**	0.943	0.904	**0.959**	**0.931**	**0.977**

multi-scale input to obtain the features of different receptive fields, thus showing better performances on all datasets.

4.2. Case Study

This section provides a case study of WTD dataset abnormal detection to study how STGAT improves interpretability. We select seven sensors related to the abnormality of the main bearing for analysis, including wind speed, hub speed, power, environment temperature, and three main bearing temperatures at different locations. The attention scores are shown in Fig.4, where lines with different widths represent connection relationships. From Fig.4(a), we can see that nodes 3 and 4 have a strong correlation. It is reasonable because the main bearing is a rotating part and the temperatures in front and behind the main bearing change with wind speed and rotation rate due to friction. Meanwhile, Fig.4(b) shows that the attention score at the nearest moment is the highest, indicating the close relation of the value at the current moment and values at its neighboring moments.



(a) Feature attention score (b) Temporal attention score

Fig. 4. Attention score in WTD case.

The detection results are shown in Fig.5. Green line on the bottom of the figure represents the labeled abnormal data in testing set, while the red shadows represent detected abnormal data. As shown in the location ④, our method recognizes most of the anomalies from time points 4000 to 10000. Moreover, the curves of Node:3 and Node:4 which represent the temperatures in front and behind the main bearing indicate that over the abnormal time period, their patterns change a lot, which is consistent with the results in Fig.4. In the location ①, our algorithm detects the expert unmarked anomalies. This pre-warning is extremely beneficial to prevent further worsening of the anomalies of wind turbines. However, in the locations ② and ③, the algorithm still presents missing detection. Taking into account these false negative and false positive rates, we need to combine more domain knowledge for analysis, which is one of the important works in the future.

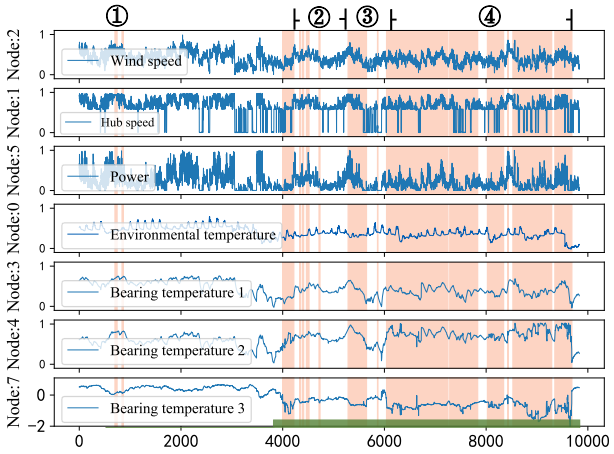


Fig. 5. Case analysis for anomaly detection on WTD dataset.

5. CONCLUSION

This paper proposes an unsupervised anomaly detection framework based on deep STGAT network. By learning complex feature and temporal correlations and combining them with a multiscale input strategy, we achieves state-of-the-art results on four different datasets consistently. Furthermore, our model demonstrates better abnormal detection capability and interpretability for anomalies, enabling users to rapidly find and position the anomalies when dealing with actual anomaly detection. Future work can further combine with domain knowledge to improve the accuracy and consider extra architecture to optimize the model's training and improve the practicability of the method.

Acknowledgment

The work is supported by National Natural Science Foundation of China (62006236), NUDT Research Project(ZK20-10), National Key Research and Development Program of China (2020YFA0709803), Hunan Provincial Natural Science Foundation (2020JJ5673), National Science Foundation of China (U1811462), National Key R&D project by Ministry of Science and Technology of China(2018YFB1003203), and Autonomous Project of HPCL (201901-11, 202101-15). Siqi Wang and Chengkun Wu are corresponding authors.

6. REFERENCES

[1] Charu C. Aggarwal, *Outlier Analysis*, Springer, 2015.

[2] M. Markou and S. Singh, "Novelty detection: A review—part 1: Statistical approaches," *Signal Processing*, vol. 83, no. 12, pp. 2481–2497, 2003.

[3] G. Pang, C. Shen, L. Cao, and Avd Hengel, "Deep learning for anomaly detection: A review," 2020.

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

[5] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional lstm for anomaly detection," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 2017, pp. 439–444.

[6] S. Lin, R. Clark, R. Birke, S. Schonborn, and S. Roberts, "Anomaly detection for time series using vae-lstm hybrid model," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[7] D. Li, D. Chen, L. Shi, B. Jin, J. Goh, and SK Ng, "Madgan: Multivariate anomaly detection for time series data with generative adversarial networks," in *Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series*, 2019, pp. 703–716.

[8] J. Laurikkala, M. Juhola, and E. Kentala, "Informal identification of outliers in medical data," *Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology*, 2000.

[9] L. K. Hansen, S. Sigurdsson, T. Kolenda, FÅ Nielsen, U. Kjems, and J. Larsen, "Modeling text with generalizable gaussian mixtures," *CiteSeer*, 1999.

[10] X. Liu, X. Zhu, M. Li, L. Wang, E. Zhu, T. Liu, M. Kloft, D. Shen, J. Yin, and W. Gao, "Multiple kernel kk-means with incomplete kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 5, pp. 1191–1204, 2020.

[11] B. Pincombe, "Anomaly detection in time series of graphs using arma processes," *Asor Bulletin*, 2005.

[12] D. Park, Y. Hoshi, and Charles C. Kemp, "A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder," *IEEE Robotics and Automation Letters*, vol. PP, no. 99, 2017.

[13] H. Xu, Y. Feng, J. Chen, Z. Wang, H. Qiao, W. Chen, N. Zhao, Z. Li, J. Bu, and Z. Li, "Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications," pp. 187–196, 2018.

[14] Q. Wang, J. Xu, R. Li, P. Qiao, K. Yang, S. Li, and Y. Dou, "Deep image clustering using convolutional autoencoder embedding with inception-like block," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 2356–2360.

[15] Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A. Zuluaga, "Usad: Unsupervised anomaly detection on multivariate time series," in *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2020.

[16] C. Song, Y. Lin, S. Guo, and H. Wan, "Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 1, pp. 914–921, 2020.

[17] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 922–929, 2019.

[18] Z. Wu, S. Pan, G. Long, J. Jiang, X. Chang, and C. Zhang, "Connecting the dots: Multivariate time series forecasting with graph neural networks," in *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2020.

[19] H. Zhao, Y. Wang, J. Duan, C. Huang, and Q. Zhang, "Multivariate time-series anomaly detection via graph attention network," 2020.

[20] A. Deng and B. Hooi, "Graph neural network-based anomaly detection in multivariate time series," 2021.

[21] Y. Su, Y. Zhao, C. Niu, R. Liu, and D. Pei, "Robust anomaly detection for multivariate time series through stochastic recurrent neural network," in *the 25th ACM SIGKDD International Conference*, 2019.

[22] J. Wu, L. Guan, M. Bao, Y. Xu, and W. Ye, "Vibration events recognition of optical fiber based on multi-scale 1-d cnn," *Opto-Electronic Engineering*, 2019.

[23] V. Petar, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," 2017.

[24] J. Goh, S. Adepu, M. Tan, and S. L. Zi, "Anomaly detection in cyber physical systems using recurrent neural networks," in *IEEE International Symposium on High Assurance Systems Engineering*, 2017.

[25] Chuadhry M. Ahmed, Venkata R. Palleti, and Aditya P. Mathur, "Wadi: a water distribution testbed for research in the design of secure cyber physical systems," in *the 3rd International Workshop*, 2017.