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

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

[1]College of Computer Science, National University of Defense Technology, Changsha, China
[2]Engineering Department, Lancaster University, Lancaster, LA1 4YW, UK
[3]National Supercomputing Centre of Tianjin, China
[4]Beijing Goldwind HuiNeng Technology Co. Ltd., Beijing, China

ABSTRACT

Anomaly detection in multivariate time series data is challenging due to complex temporal and feature correlations and heterogeneity. This paper proposes a novel unsupervised multi-scale stacked **s**patial-**t**emporal **g**raph **att**ention 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 with stackable STGAT networks. Meanwhile, a multi-scale input network is exploited to capture the temporal correlations in different time-scales. Experiments on a new wind turbine dataset(built and released by us) and three public datasets show that our method detects anomalies more accurately than baseline approaches and provide interpretability through observing the attention score among multiple sensors and different times.

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 widespread in our production facilities and lives, 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 suddenly, and vice versa. Multivariate time series are collected from independent sensors with complex coupling relations. Therefore, each variable depends on its historical value and other variable values, bringing great challenges to anomaly detection.

For the earlier studies, the researchers solved the problems of anomaly detection in multivariate time series data mainly by using the methods of statistical analysis or autoregressive [2]. These methods estimated anomalies based on the overall data distribution, while the spatial-temporal correlations of data were not considered. With the development of deep learning, many deep learning models are applied and achieved good performances [3] which can usually be categorized into prediction-based and reconstruction-based methods. The prediction-based methods focus on contextual anomalies, and the typical models include long-short term memory (LSTM) [4], convolutional LSTM (ConvLSTM) [5], etc. While the reconstruction-based methods focus on overall distribution anomalies of subsequences, and the most usual methods are based on encoder-decoder (AE) [6] and generative adversarial networks (GANs) [7]. However, these methods lack the considerations of the complex coupling relation between sensors, resulting in instability when dealing with high-dimensional data with lots of potential correlations.

Recently, graph neural network(GCN) [8] is effective in dealing with complex graph structure data. Inspired by this, we utilize graph networks to extract complex spatial-temporal correlations from multivariate time series data. There are three significant challenges to this idea: 1) Abnormal mode is not available during the training, which means the algorithm must depend separately on non-abnormal data. 2) When temporal features are captured, the correlations between variables should be considered. 3) Multivariate time series data usually show heterogeneity, and high-level implicit features are difficult to capture. Therefore, we propose STGAT-MAD framework. Our main contributions are summarized as :

- 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.
- Extensive experiments show that the performance of our model is improved by up to 13% in terms of F1 and provides good interpretability, i.e., the underlying correlations among different features collected from multiple sensors.
- A new wind turbine dataset is built and released for multivariate time series anomaly detection, derived from a real wind farm. Our code and dataset are available at: https://github.com/zhanjun717/STGAT_MAD.

2. RELATED WORK

Traditional anomaly detection methods: Due to high complexity and anomaly uncertainty, it is hard to obtain labels. Hence, the unsupervised method is more suitable for anomaly detection. The traditional statistical models [2], distance-based [9] and clustering-based [10] models have been widely used. These methods do not map input features to a more discriminant feature space, while linear discriminant analysis [11] maps input features to different spaces to discriminate between normal and abnormal features. However, the above methods ignore the importance of temporal correlations. Autoregressive [12] is a method for specially handling time series, but it requires the data must be autocorrelative.

Deep learning-based methods: To extract complex patterns implied in multivariate time series, deep learning methods are widely adopted. Autoencoder-based methods (AE) are the most typical methods, such as VAE-LSTM [6, 3, 13, 14], which recognize abnormal data according to reconstruction error but are usually rough. The introduction of GANs has well-optimized this problem. Dan Li et al. [7] proposes MAD-GAN, using both the discrimination and reconstruction errors for anomaly detection. Julien Audibert [15] proposes USAD combing AE and GAN to improve network stability.

Spatial-Temporal Networks: Spatial-temporal networks have optimal performance in dealing with complicated spatial-temporal data, such as Spatio-Temporal Graph Convolutional Networks(STGCN)[16], attention based spatial-temporal graph convolutional network (ASTGCN) [17]. In a broader perspective view, multivariate time series data is spatial-temporal data in essence. Hence, Graph Convolutional Networks(GCN) [18], MTAD-GAT [19] and GAT [20] have applied graph networks into the correlations extraction between variables in multivariate time series and achieved good performance. Nevertheless, these methods still extract spatial-temporal correlations in a single layer.

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. 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 [14, 21, 15].

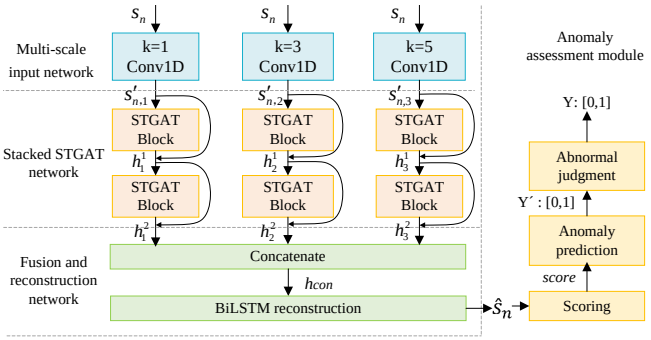


Fig. 1. STGAT-MAD basic framework.

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:

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

where σ denotes $ReLU$ activation function. $k \in \{1, 5, 7\}$ is the size of convolution kernel, $s = 1$ 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 $\mathbb{G}_{x_t} = \left(V, E^f\right)$, 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 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 between the moments m and n can be expressed as:

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

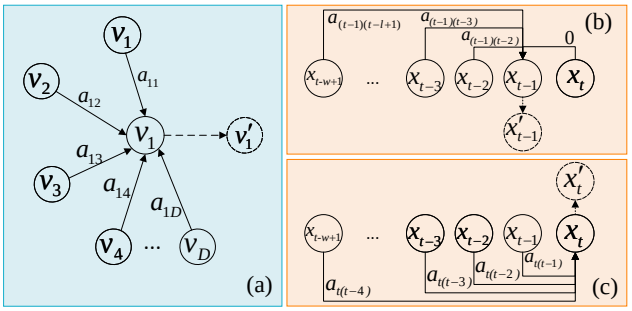


Fig. 2. The spatial-temporal graph structure, where the dotted line denotes the aggregate representation of the node. (a) indicates feature dimension. (b) and (c) represent temporal dimension.

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]. $\overrightarrow{\mathrm{a}} \in \mathbb{R}''$

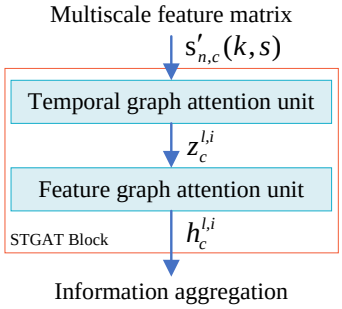


Fig. 3. The basic framework of the STGAT block, which is connected by temporal and feature attention units.

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 new private real-world 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 [13], 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, compared with GDN and

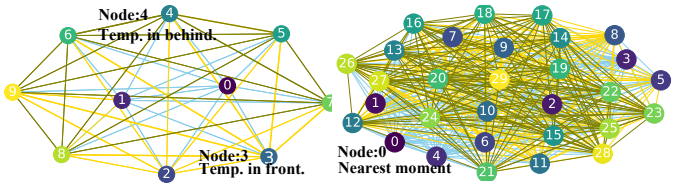
Table 1. Performance comparison of different methods and datasets

	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 [13]	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	**0.965**	0.841	**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**

MTAD-GAT models, the feature extraction layer of STGAT-MAD can extract deeper implicit spatial-temporal features from multivariate data, resulting in the data reconstruction layer being better to restore data distribution contextual information. Meanwhile, STGAT-MAD introduces multi-scale input to obtain the features of different receptive fields on the same layer, 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 effectively improves interpretability for abnormal detection. The correspondence between nodes and sensors is shown in Fig.5. 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 base bearing is a rotating part and the temperatures in front and behind the base 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.

In what follows, 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 in the labeled abnormal period from time points 4000 to 10000. Moreover, the curves of Node:3 and Node:4 which represent the temperatures in front and behind the base 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 prewarning is extremely beneficial to prevent further worsening of the anomalies of wind turbines. However, in the locations

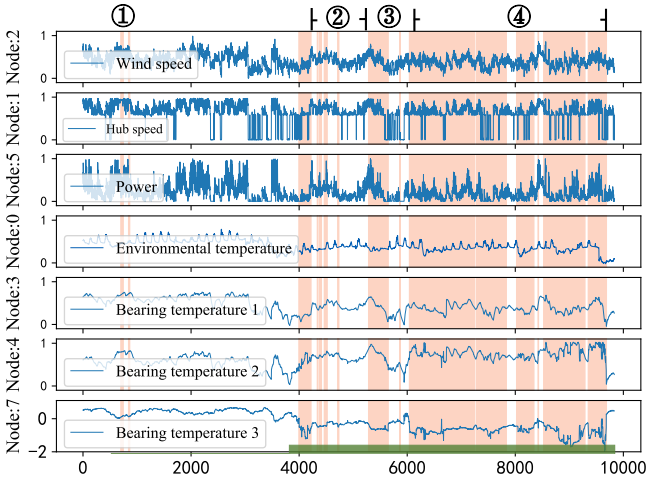


Fig. 5. Case analysis for anomaly detection on WTD dataset.

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.

5. CONCLUSION

This paper proposes an unsupervised anomaly detection framework based on deep SAGAT 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

This work is jointly funded by the National Science Foundation of China (U1811462), National Key R&D project by Ministry of Science and Technology of China(2018YFB1003203), and open fund from the State Key Laboratory of High Performance Computing (201901-11).

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] Thomas N. Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," 2017.

[9] H. Ghorbani, "Mahalanobis distance and its application for detecting multivariate outliers," *Facta Universitatis Series Mathematics and Informatics*, vol. 34, no. 3, pp. 583, 2019.

[10] G. Pu, L. Wang, J. Shen, and F. Dong, "A hybrid unsupervised clustering-based anomaly detection method," *Tsinghua Science and Technology*, vol. v.26, no. 02, pp. 14–21, 2021.

[11] W. Sheng, J. Lu, X. Gu, H. Du, and J. Yang, "Semi-supervised linear discriminant analysis for dimension reduction and classification," *Pattern Recognition*, vol. 57, no. C, pp. 179–189, 2016.

[12] Asrul H. Yaacob, Ian K.T. Tan, Su Fong Chien, and Hon Khi Tan, *ARIMA Based Network Anomaly Detection*, ARIMA based network anomaly detection, 2010.

[13] 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.

[14] 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.

[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] Petar Velikovi, 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.