ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888728

Autoregressive Density Estimation Transformers for Multivariate Time Series Anomaly Detection

Mohammed Ayalew Belay*, Adil Rasheed†, Pierluigi Salvo Rossi*‡
* Department of Electronic Systems, Norwegian University of Science and Technology, Trondheim, Norway
† Department of Engineering Cybernetics, Norwegian University of Science and Technology, Trondheim, Norway
‡ Department of Gas Technology, SINTEF Energy Research, Trondheim, Norway
Emails: mohammed.a.belay@ntnu.no, adil.rasheed@ntnu.no, salvorossi@ieee.org

***Abstract*—Anomaly detection in multivariate time series (MTS) from sensor data is critical in many industrial applications. The challenge lies in managing massive unlabeled datasets with complex spatio-temporal correlations, diverse anomalies, and noise. While several unsupervised methods have been proposed, they are often limited to specific applications. In this paper, we introduce a probabilistic self-supervised framework, Autoregressive Density Estimation Transformer (ADET). ADET integrates an efficient transformer for learning spatio-temporal representations with density estimation networks for multi-score anomaly detection, focusing on point-to-point, point-to-distribution, and distribution-to-distribution distances. ADET improves noise resilience using optimal truncated singular value decomposition (OT-SVD) in an end-to-end optimization process. We conducted experiments by employing several encoders and performed an ablation study to examine the effect of OT-SVD.**

***Index Terms*—Anomaly detection, transformer, unsupervised learning, and low-rank approximation**

I. Introduction

In the era of data-driven decision-making, multivariate time series anomaly detection (MTSAD) has emerged as an essential tool for several applications such as industrial monitoring [1], [2], cybersecurity [3], wireless sensor networks [4], [5], healthcare [6], and autonomous vehicles [7]. MTSAD methods aim to discern unexpected temporal patterns in datasets composed of multiple dependent variables sampled at regular or irregular intervals. While several conventional statistical and machine learning methods provide a foundational framework for MTSAD, their efficacy is limited by specific assumptions and scalability issues in high-dimensional, non-linear, and noisy real-world datasets [8]. Recently, deep learning methods enhanced MTSAD significantly in supervised, semi-supervised, and unsupervised learning paradigms [9], [10]. However, supervised and semi-supervised MTSAD algorithms are constrained by their dependence on labeled training data, class imbalance bias, and reduced adaptability for dynamic conditions [11], [12]. Consequently, there is a growing demand for robust self-supervised MTSAD approaches and several aspects are under investigation, including identification of novel anomaly scoring methods and effective modeling of spatial and/or temporal dependencies.

This work was partially supported by the Research Council of Norway under the project DIGITAL TWIN within the PETROMAKS2 framework (project nr. 318899).

Although several deep learning MTSAD methods have been proposed, existing studies overlook the inherent low-rank structure in multivariate time series due to correlated measurements and noise. Additionally, existing frameworks lack versatile anomaly scoring mechanisms. In real-world MTS data, noise and correlations among components are common, leading to higher false positives due to increased rank. This suggests that integrating low-rank approximation (LRA) methods into MTSAD frameworks can better capture patterns for robust anomaly detection. MTS anomalies can be point or sub-sequence anomalies, which present challenges for conventional methods when abnormalities are isolated to specific components. Several MTSAD methods [13]–[16] rely on Euclidean distance-based scoring, using reconstruction or prediction errors, which limits their ability to handle complex multivariate data and diverse anomaly types. To address these limitations, we propose Autoregressive Density Estimation Transformers (ADET), a probabilistic framework for self-supervised MTSAD. ADET combines efficient multi-head transformer encoders for spatio-temporal representation learning with networks for density estimation, enabling comprehensive multi-score anomaly detection. It also supports noise-tolerant detection of point and sub-sequence anomalies using LRA with optimal truncated singular value decomposition (OT-SVD) in an end-to-end optimized framework.

II. The Proposed Method

We consider a multi-sensor system with K components and $x_k[n] \in \mathbb{R}$ denoting the value of the kth univariate component at discrete time n. The *data vector* $\boldsymbol{x}[n] = (x_1[n], x_2[n], \ldots, x_K[n])^T \in \mathbb{R}^K$ collects all the components at time n, and the collection of data vectors related to N discrete times steps is arranged into the *data matrix* $\boldsymbol{X} = (\boldsymbol{x}[1], \boldsymbol{x}[2], \ldots, \boldsymbol{x}[N]) \in \mathbb{R}^{K \times N}$. We assume a training data matrix ($\boldsymbol{X}_{\text{train}}$) is available and represents the MTS observations under normal conditions. The objective of the proposed framework is to construct a model $\mathcal{G}(\cdot)$ that characterizes the MTS behavior under normal conditions and is capable of detecting deviations from that behavior. For model evaluation, we consider a test data matrix ($\boldsymbol{X}_{\text{test}} \in \mathbb{R}^{K \times M}$ with $M \ll N$) which includes MTS from both normal and anomalous conditions. In addition, a *label vector* $\boldsymbol{y} = (y_1, y_2, \cdots, y_M)^T$ paired with the test data matrix is available and represents ground

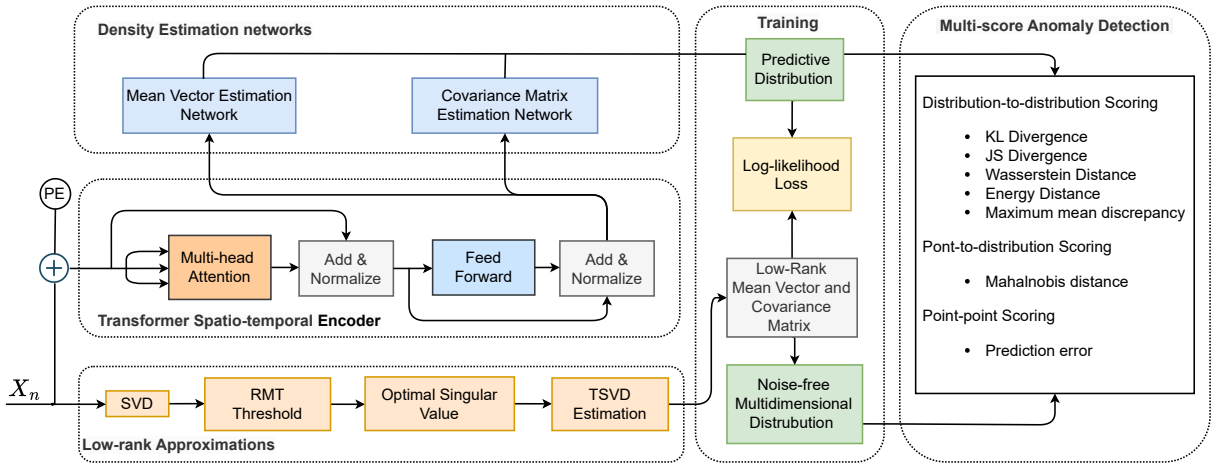

Fig. 1. Autoregressive Density Estimation Transformers (ADET) for self-supervised MTSAD

truth information, with $y_m = 1$ (resp. $y_m = 0$) denoting the presence (resp. absence) of an anomaly at discrete time m.

A. Architecture Overview

The proposed ADET framework is based on a probabilistic architecture utilizing autoregressive density estimation and low-rank approximation (LRA), as illustrated in Fig. 1. ADET consists of three key components: (i) Multi-head transformer encoders for efficient spatio-temporal representation learning; (ii) an LRA module to improve anomaly detection in the presence of correlation and noise; (iii) a density estimation block to generate predictive distributions for detecting both point and sub-sequence anomalies. Let $\boldsymbol{X}[n] = (\boldsymbol{x}[n], \boldsymbol{x}[n-1], \ldots, \boldsymbol{x}[n-L+1])$ represent the input at time n, built using a sliding window with size L. The input matrix $\boldsymbol{X}[n]$ is processed by the transformer encoders to produce a latent representation $\boldsymbol{Z}[n]$. This latent representation is then used for parametric estimation of the predictive probability distribution of the next data point $\boldsymbol{x}[n+1]$ using feed-forward neural networks (FFNNs). The predictive distribution is modeled as a multivariate Gaussian, defined by a mean vector $\boldsymbol{\mu}[n+1]$ and a covariance matrix $\boldsymbol{\Sigma}[n+1]$. The system aims to predict a "noise-free" version of $\boldsymbol{x}[n+1]$. The input matrix $\boldsymbol{X}[n]$ is also processed by an LRA block, using OT-SVD, to produce a matrix $\hat{\boldsymbol{X}}[n]$. This matrix serves as ground truth for system training, using the negative log-likelihood loss function $\mathcal{L}_{\text{NLL}}$.

B. Transformer Spatio-Temporal Encoder

The proposed architecture utilizes transformer encoders for learning spatio-temporal correlations. It consists of stacked transformer-encoder modules, denoted as $\mathcal{T}(\cdot)$, each containing a multi-head self-attention mechanism, position-wise feed-forward networks (FFNNs), and normalization layers with residual connections. Unlike recurrent networks, the transformer processes the input sequence in parallel. To maintain sequential information, positional encoding is applied via a positional encoding matrix $\boldsymbol{P} \in \mathbb{R}^{K \times N}$. The multi-head self-attention mechanism computes attention scores using queries, keys, and values derived from the input sequence with positional encoding. The query, key, and value matrices are calculated as $\boldsymbol{Q}_i = (\boldsymbol{X}+\boldsymbol{P})\boldsymbol{W}_{\text{Q},i}$, $\boldsymbol{K}_i = (\boldsymbol{X}+\boldsymbol{P})\boldsymbol{W}_{\text{K},i}$, and $\boldsymbol{V}_i = (\boldsymbol{X}+\boldsymbol{P})\boldsymbol{W}_{\text{V},i}$, where $\boldsymbol{W}_{\text{Q},i} \in \mathbb{R}^{K \times (K/H)}$, $\boldsymbol{W}_{\text{K},i} \in \mathbb{R}^{K \times (K/H)}$, and $\boldsymbol{W}_{\text{V},i} \in \mathbb{R}^{K \times K}$ are learned weights for the ith attention head, and H is the number of attention heads. Scaled-dot attention operations are performed in parallel across all heads, with outputs concatenated and transformed linearly:

$$\boldsymbol{A} = \text{Concat}(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_H)\boldsymbol{W}_{\text{O}} \,, \tag{1}$$

where $\boldsymbol{W}_{\text{O}} \in \mathbb{R}^{HK \times K}$ is the output projection weight matrix. The output of the attention mechanism is added to the input, followed by a normalization layer. A position-wise FFNN is then applied to capture complex non-linear dependencies, with residual connections and normalization layers for stability. This enables the transformer encoders to learn spatio-temporal relationships effectively in the multivariate time series data represented by $\boldsymbol{X}$.

C. Optimal Low-Rank Approximations

The proposed architecture integrates OT-SVD as a LRA for denoising, thus reducing false-positive rates in anomaly detection. The optimal (i.e. minimizing the Frobenius norm) rank-r approximation ($\tilde{\boldsymbol{X}}_{(r)}$) of an input matrix ($\boldsymbol{X}$) is determined by retaining only the first r singular values and their corresponding singular vectors [17]:

$$\tilde{\boldsymbol{X}}_{(r)} = \underset{\hat{\boldsymbol{X}}:\, \text{rank}(\hat{\boldsymbol{X}}) \leq r}{\arg\min} \|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_F^2 = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T \tag{2}$$

where $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$, are the left and right singular vectors of the input matrix, and σ_i the corresponding singular values arranged in descending order (i.e., $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(K,L)}$). The optimal rank r is selected via an information-theoretic approach based on random matrix theory [18]. More specifically, the optimal threshold (τ^*) minimizing the asymptotic

mean square error (MSE) between the original matrix and its low-rank approximation

$$\tau^* = \arg\min_{\tau} \lim_{S\to\infty} \mathbb{E}\left[\|\boldsymbol{X} - \tilde{\boldsymbol{X}}_{(r)}\|_F^2\right] . \tag{3}$$

is computed, assuming additive white Gaussian noise, as $\tau^* = \omega(\rho)\sigma_{\text{med}}$, where σ_{med} is the median of the singular values and $\omega(\rho)$ depends on the matrix dimensions. Finally, the optimal threshold-dependent rank is thus determined by $r(\tau) = \max\{i : \sigma_i > \tau^*\}$. During the training process, the "noise-free" reference signals ($\tilde{\boldsymbol{x}}[n] \in \mathbb{R}^K$) are generated.

D. Autoregressive Density Estimation

In autoregressive framework, we employ a likelihood function parameterized by neural network outputs to model the conditional distribution of future time series values based on past observations. The latent representation from transformer encoder is employed to estimate the parameters of predictive Gaussian distribution ($\boldsymbol{\mu}[n+1]$ and $\boldsymbol{\Sigma}[n+1]$) from the latent representation of the input matrix. The conditional distribution is assumed to be a multivariate Gaussian distribution with statistically-independent components, thus represented by its mean vector $\boldsymbol{\mu}[n+1] = (\mu_1[n+1], \mu_2[n+1], \ldots, \mu_K[n+1])^T$ and diagonal covariance matrix $\boldsymbol{\Sigma}[n+1] = \text{diag}(\sigma_1^2[n+1], \sigma_2^2[n+1], \ldots, \sigma_K^2[n+1])$. The FFNN used for mean estimation ($\mathcal{D}_\mu$) is linear, while the FFNN used for covariance estimation ($\mathcal{D}_\sigma$) employs softplus activation.

E. Loss function and optimization

In the probabilistic framework, we predict the parameters of a conditional distribution given past observations. Therefore, ADET training ADET relies on optimizing the network parameters to minimize the negative log-likelihood ($\mathcal{L}_{\text{NLL}}$) of the predictive Gaussian distribution ($\boldsymbol{\mu}[n+1]$ and $\boldsymbol{\Sigma}[n+1]$) evaluated for the next "noise-free" sample $\tilde{\boldsymbol{x}}[n+1]$, i.e.

$$\begin{aligned}\mathcal{L}_{\text{NLL}} &= -\ln p(\tilde{\boldsymbol{x}}[n+1]; \boldsymbol{\mu}[n+1], \boldsymbol{\Sigma}[n+1]) \\ &= \frac{1}{2}\sum_{k=1}^{K}\left[\ln(2\pi\sigma^2[n+1]) + \frac{(\tilde{x}_k[n+1] - \mu_k[n+1])^2}{\sigma^2[n+1]}\right] .\end{aligned} \tag{4}$$

F. Inference and Anomaly Detection

The proposed architecture supports flexible multi-score anomaly detection mechanisms capable of identifying both point and sub-sequence anomalies. Specifically, it combines point-to-point (S_{PP}), point-to-distribution (S_{PD}), and distribution-to-distribution (S_{DD}) scoring mechanisms.

1) Point-to-Point Scoring: Point-to-point scoring is straightforward and computationally efficient, making it suitable for initial anomaly detection. It is based on the ℓ_2-norm between the observed data point and the mean of the predictive distribution:

$$S_{\text{PP}}(\boldsymbol{x}[n]) = \left[\sum_{k=1}^{K}(x_k[n] - \mu_k[n])^2\right]^{\frac{1}{2}} . \tag{5}$$

2) Point-to-Distribution Scoring: Point-to-distribution scoring, particularly useful for correlated measurements, is based on the Mahalanobis distance. It detects anomalies by measuring the deviation of a single data point from the entire predictive distribution:

$$S_{\text{PD}}(\boldsymbol{x}[n]) = \sum_{k=1}^{K}\frac{(x_k[n] - \mu_k[n])^2}{\sigma_k^2[n]} . \tag{6}$$

3) Distribution-to-Distribution Scoring: This method uses the KL divergence to quantify the difference between the predicted distribution and the measured distribution extracted from the current input $\boldsymbol{X}[n]$. The measured distribution is modeled as a multivariate Gaussian with mean vector $\check{\boldsymbol{\mu}}[n] = \boldsymbol{x}[n]$ and diagonal covariance matrix $\check{\boldsymbol{\Sigma}}[n]$, where diagonal terms represent the sample variance from the "noise-free" input data. The distribution-to-distribution score is computed as:

$$\begin{aligned}S_{\text{DD}}(\boldsymbol{x}[n]) = &\sum_{k=1}^{K}\frac{(\mu_k[n] - x_k[n])^2}{\sigma_k^2[n]} + \sum_{k=1}^{K}\frac{\check{\sigma}_k^2[n] - \sigma_k^2[n]}{\sigma_k^2[n]} \\ &+ \sum_{k=1}^{K}\ln\left(\frac{\check{\sigma}_k^2[n]}{\sigma_k^2[n]}\right) .\end{aligned} \tag{7}$$

Other probabilistic measures, such as JS divergence, Wasserstein distance, or total variation distance, can also be considered.

4) Multi-Score Anomaly Detection: We propose a generic anomaly scoring approach by combining the point-to-point, point-to-distribution, and distribution-to-distribution measures. The total anomaly score (S_n) at a given time is given by:

$$S_n = \alpha S_{\text{PP}} + \beta S_{\text{PD}} + \gamma S_{\text{DD}} \tag{8}$$

where α, β, and γ determine the contribution of each score. For unsupervised anomaly detection, a thresholding function is applied based on the training data. The anomaly score (S_n) is calculated for each time step, and the threshold is determined by:

$$\lambda^* = \frac{1}{N}\sum_{i=1}^{N} S_n + \left[\frac{z^2}{N}\sum_{i=1}^{N}\left(S_n - \frac{1}{N}\sum_{i=1}^{N} S_n\right)^2\right]^{\frac{1}{2}} \tag{9}$$

where S_n is given by equation (8) and z is a scale factor.

III. Experiments and Results

A. Datasets and Pre-processing

To evaluate the performance of our proposed framework, we used two real-world multivariate time series datasets. 1) **SWaT** [19], [20] 2) **WADI** [21] We employ two key preprocessing steps (downsampling and feature normalization) on the input data prior to utilizing it in the framework. Downsampling was performed using a median filter with a 1-minute window size and no overlap for both training and test datasets. Labels for the downsampled test data were assigned based on the presence of anomalies within the corresponding window. For feature normalization, we employed min-max scaling to ensure stable training.

B. Baselines and Implementation Details

To evaluate the performance of our proposed architecture, we compare it with four other encoders: CNN, RNN, GRU, and LSTM. These encoders represent a range of spatio-temporal modeling techniques commonly used for multivariate time series anomaly detection. The CNN encoder consists of two 1D convolutional layers with a kernel size of 5 and ReLU activations. For the RNN, GRU, and LSTM encoders, we used two stacked layers with Tanh activations, with the output size equal to the time series dimension. All baselines were implemented in TensorFlow and trained under similar conditions. Each model was trained using the Adam optimizer with a learning rate of 10^{-4} and a batch size of 64. We also considered some features to avoid an underflow problem. All models were trained in the Google Colaboratory Pro environment using NVIDIA T4 GPUs. We evaluate performance on labeled test datasets, using the F1 score and area under the Precision-Recall curve (AUPR) as the key metrics due to data imbalance.

C. Results and Discussions

1) Anomaly Scoring Performance: We evaluate the ADET framework performance on the SWaT and WADI datasets using different configurations of the α, β, and γ coefficients, which weigh the anomaly detection scoring methods. As shown in Table I, we use the F1 score, AUC, and AUPR to assess performance. On the SWaT dataset, the highest AUC (0.8831) and AUPR (0.7824) are achieved when $\gamma = 1.0$ and both α and β are set to 0, indicating that point-to-distribution scoring is most effective, suggesting the presence of point anomalies. In contrast, the WADI dataset shows more variability across configurations, highlighting that different datasets benefit from different scoring strategies. The results demonstrate that different datasets benefit from different anomaly detection strategies.

TABLE I
ADET PERFORMANCE FOR DIFFERENT α, β, AND γ COEFFICIENTS (SWAT AND WADI).

α	β	γ	SWaT			WADI		
			F1	AUC	AUPR	F1	AUC	AUPR
0.0	0.0	1.0	0.7726	0.8831	0.7824	0.2644	0.6100	0.1681
0.0	1.0	0.0	0.7835	0.8733	0.7713	0.3362	0.7897	0.2203
1.0	0.0	0.0	0.7431	0.8272	0.7135	0.4000	0.8064	0.2736
0.5	0.5	0.0	0.7828	0.8719	0.7712	0.3390	0.7914	0.2209
0.5	0.0	0.5	0.7726	0.8831	0.7824	0.2644	0.6100	0.1681
0.0	0.5	0.5	0.7726	0.8831	0.7824	0.2667	0.6109	0.1687
0.33	0.33	0.34	0.7726	0.8831	0.7824	0.2667	0.6109	0.1687

2) Spatio-temporal Encoder Performance: Table II shows the performance of various spatio-temporal encoders across multiple metrics (F1 score, AUC, and AUPR) on the SWaT and WADI datasets. We evaluate five encoders using three scoring mechanisms: point-to-point (S_{PP}), point-to-distribution (S_{PD}), and distribution-to-distribution (S_{DD}). The Transformer encoder consistently outperforms others on the SWaT dataset and is competitive on the WADI dataset. For SWaT, the Transformer achieves the best results across all scoring mechanisms (S_{PP}, S_{PD}, and S_{DD}). The GRU also performs well, particularly in the WADI dataset. These findings highlight the Transformer's superiority in capturing complex patterns and dependencies in multivariate time series data, making it an essential model for anomaly detection in multi-sensor data.

TABLE II
SPATIO-TEMPORAL ENCODER PERFORMANCE ACROSS METRICS

Score	Encoder	SWaT			WADI		
		F1	AUC	AUPR	F1	AUC	AUPR
S_{PP}	CNN	0.7322	0.8145	0.7041	0.3357	0.7760	0.2385
	RNN	0.6667	0.8407	0.4595	0.2708	0.7665	0.1718
	GRU	0.7150	**0.8704**	0.5410	0.3077	0.7818	0.2680
	LSTM	0.6081	0.8215	0.3671	0.3215	0.7931	0.2527
	Transformer	**0.7431**	0.8272	**0.7135**	**0.4000**	**0.8064**	**0.2736**
S_{PD}	CNN	0.7348	0.8390	0.7220	0.3226	0.5521	0.2434
	RNN	0.2964	0.4556	0.1433	0.2795	0.7389	0.1765
	GRU	0.6572	0.8642	0.6947	**0.3621**	0.7542	**0.3176**
	LSTM	0.2817	0.4477	0.1724	0.3430	0.5697	0.2384
	Transformer	**0.7835**	**0.8733**	**0.7713**	0.3362	**0.7897**	0.2203
S_{DD}	CNN	0.6983	0.8467	0.7237	0.1667	0.4559	0.1164
	RNN	0.2355	0.4066	0.1058	0.1739	0.5896	0.1089
	GRU	0.7144	0.8561	0.7208	**0.3223**	**0.6379**	**0.2752**
	LSTM	0.2352	0.2834	0.1174	0.1532	0.4583	0.1104
	Transformer	**0.7726**	**0.8831**	**0.7824**	0.2644	0.6100	0.1681

3) Ablation Study: Table III presents the performance of the ADET framework without OT-SVD denoising for different combinations of the α, β, and γ coefficients across the SWaT and WADI datasets. This serves as an ablation study to compare how the anomaly detection framework performs without the proposed OT-SVD denoising step, versus the complete model that includes it (as shown in Table I). The ablation study shows that OT-SVD denoising can be integrated to improve performance of the ADET framework. While the model without denoising still performs reasonably well, particularly in some configurations, the inclusion of denoising leads to consistently better results across all metrics.

TABLE III
WITHOUT: ADET PERFORMANCE FOR DIFFERENT α, β, AND γ COEFFICIENTS (SWAT AND WADI).

α	β	γ	SWaT			WADI		
			F1	AUC	AUPR	F1	AUC	AUPR
0.0	0.0	1.0	0.7538	0.8581	0.7483	0.2875	0.5840	0.2100
0.0	1.0	0.0	0.7830	0.8675	0.7652	0.3448	0.7858	0.2511
1.0	0.0	0.0	0.7629	0.8785	0.7872	0.4121	0.8022	0.2780

IV. CONCLUSIONS

In this paper, we propose a robust, multi-score, and adaptive transformer-based framework for anomaly detection in multivariate time series. The transformer encoder consistently outperformed other baseline models, especially in the SWaT dataset, highlighting its ability to capture complex spatio-temporal dependencies. In future work, we will further experiment on weight optimization methods and evaluate the framework on more datasets.

ACKNOWLEDGMENT

This work was partially supported by the Research Council of Norway under the project DIGITAL TWIN within the PETROMAKS2 framework (project nr. 318899).

REFERENCES

[1] S. Yin, S. X. Ding, X. Xie, and H. Luo, "A review on basic data-driven approaches for industrial process monitoring," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 11, pp. 6418–6428, 2014.

[2] M. A. Belay, A. Rasheed, and P. S. Rossi, "Self-Supervised Modular Architecture for Multi-Sensor Anomaly Detection and Localization," in *2024 IEEE Conference on Artificial Intelligence (CAI)*, 2024, pp. 1278–1283.

[3] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network anomaly detection: Methods, systems and tools," *IEEE Communications Surveys and Tutorials*, vol. 16, no. 1, pp. 303–336, 6 2014.

[4] M. Xie, S. Han, B. Tian, and S. Parvin, "Anomaly detection in wireless sensor networks: A survey," *Journal of Network and Computer Applications*, vol. 34, no. 4, pp. 1302–1325, 7 2011.

[5] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier Detection Techniques for Wireless Sensor Networks: A Survey," *IEEE Communications Surveys and Tutorials*, vol. 12, no. 2, pp. 159–170, 6 2008.

[6] A. Ukil, S. Bandyoapdhyay, C. Puri, and A. Pal, "IoT healthcare analytics: The importance of anomaly detection," *International Conference on Advanced Information Networking and Applications, AINA*, pp. 994–997, 6 2016.

[7] F. V. Wyk, Y. Wang, A. Khojandi, and N. Masoud, "Real-time sensor anomaly detection and identification in automated vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 1264–1276, 6 2020.

[8] M. A. Belay, A. Rasheed, and P. S. Rossi, "Multivariate Time Series Anomaly Detection via Low-Rank and Sparse Decomposition," *IEEE Sensors Journal*, vol. 24, no. 21, pp. 34 942–34 952, 2024.

[9] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep Learning for Anomaly Detection: A Review," *ACM Computing Surveys*, vol. 54, no. 2, 6 2021.

[10] M. A. Belay, A. Rasheed, and P. Salvo Rossi, "MTAD: Multiobjective Transformer Network for Unsupervised Multisensor Anomaly Detection," *IEEE Sensors Journal*, vol. 24, no. 12, pp. 20 254–20 265, 6 2024.

[11] M. A. Belay, S. S. Blakseth, A. Rasheed, and P. Salvo Rossi, "Unsupervised Anomaly Detection for IoT-Based Multivariate Time Series: Existing Solutions, Performance Analysis and Future Directions," *Sensors*, vol. 23, no. 5, p. 2844, 6 2023.

[12] M. A. Belay, A. Rasheed, and P. S. Rossi, "Sparse Non-Linear Vector Autoregressive Networks for Multivariate Time Series Anomaly Detection," *IEEE Signal Processing Letters*, vol. 32, pp. 331–335, 2025.

[13] S. Tuli, G. Casale, and N. R. Jennings, "TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data," *VLDB Endowment*, vol. 15, no. 6, pp. 1201–1214, 1 2022.

[14] H. Zhao, Y. Wang, J. Duan, C. Huang, D. Cao, Y. Tong, B. Xu, J. Bai, J. Tong, and Q. Zhang, "Multivariate Time-series Anomaly Detection via Graph Attention Network," *Industrial Conference on Data Mining*, vol. 2020-November, pp. 841–850, 11 2020.

[15] A. Julien, M. Pietro, G. Frédéric, M. Sébastien, and A. Z. Maria, "Usad: Unsupervised anomaly detection on multivariate time series," *dl.acm.org*, vol. 20, pp. 3395–3404, 8 2020.

[16] A. Deng and B. Hooi, "Graph Neural Network-Based Anomaly Detection in Multivariate Time Series," *AAAI Conference on Artificial Intelligence*, vol. 5A, pp. 4027–4035, 2021.

[17] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 9 1936.

[18] M. Gavish and D. L. Donoho, "Optimal Shrinkage of Singular Values," *IEEE Transactions on Information Theory*, vol. 63, no. 4, pp. 2137–2152, 4 2017.

[19] J. Goh, S. Adepu, K. N. Junejo, and A. Mathur, "A dataset to support research in the design of secure water treatment systems," *Critical Information Infrastructures Security*, vol. 10242 LNCS, pp. 88–99, 2017.

[20] P. M. Aditya and O. T. Nils, "SWaT: A water treatment testbed for research and training on ICS security," *International Workshop on Cyber-physical Systems for Smart Water Networks (CySWater)*, 2016.

[21] C. M. Ahmed, V. R. Palleti, and A. P. Mathur, "WADI: A water distribution testbed for research in the design of secure cyber physical systems," *Proceedings of International Workshop on Cyber-Physical Systems for Smart Water Networks, CySWATER 2017*, pp. 25–28, 4 2017.