MOC: MAMBA-BASED MULTI-SCALE ONE-CLASS TIME-SERIES ANOMALY DETECTION

Tiejun Wang[§], *Rui Wang*[§], *Xudong Mou*[§], *Xu Li*[§], *Tianyu Wo*[†] [*], *Shiru Chen*[‡], *Xudong Liu*[§], *Renyu Yang*[†]

[§] School of Computer Science and Engineering, Beihang University, Beijing, China.
[†] School of Software, Beihang University, Beijing, China.
[‡] Shandong Inspur Intelligent Production Technology Co., Ltd, Beijing, China.

ABSTRACT

Time series anomaly detection is essential for applications such as finance and healthcare. However, existing methods often struggle to capture complex multi-scale dependencies or are trapped in capturing low-level sample characteristics. To address the issue, we propose MOC, a novel multi-scale one-class anomaly detection framework that integrates Mamba and CNN to model global dependencies and local dynamics jointly. In addition, MOC introduces a one-class to compact normal patterns into a hypersphere, by a hinge loss to prevent hypersphere collapse, improving robustness against noise. Extensive experiments on three public datasets show that MOC significantly outperforms state-of-the-art baselines, demonstrating its effectiveness for robust and accurate performance.

***Index Terms*—** Time series anomaly detection, multi-scale, one-class classification, state space model

1. INTRODUCTION

Time Series Anomaly Detection (TSAD) plays a pivotal role in domains such as finance, healthcare, and industrial systems [1], where the goal is to identify abnormal patterns or unexpected events hidden within temporal dynamics. However, in practice, anomalies are extremely rare, leading to severe class imbalance. Meanwhile, time series often exhibit complex dependencies across multiple scales, from short-term dynamics to long-term correlations. Together with high levels of noise, this makes labeling the data costly and difficult. Consequently, TSAD is inherently challenged by the scarcity of anomalies and the lack of reliable labels [2].

To tackle these challenges, a prevailing paradigm is the normality assumption [3–5], whose target is to learn a common distribution of normal samples via self-supervised pretext tasks. Samples that disobey the distribution are inferred as anomalies. Typical approaches include reconstruction-based [6, 7], generation-based [8], and one-class classification methods [9]. In particular, one-class classification minimizes the volume of the hyperspace of normal representations, thereby enclosing them in a compact normal class. Since its assumption aligns well with the essence of anomaly detection, where normal samples overwhelmingly outnumber anomalies, it has attracted considerable attention. For example, [10, 11] incorporate contrastive learning and reconstruction assumptions into one-class classification using LSTM networks, aiming to better capture the temporal dependencies of normal samples. However, they depend on sequential computation, struggle with learning long-term temporal dependencies, and overlook the need to model normal patterns at multiple scales. In this context, Transformers seem to be a promising alternative [7]. Still, their quadratic computational complexity creates a bottleneck, limiting anomaly detection to small feature maps and potentially reducing detection performance. In other research domains, Peng et al. [12–14] proposed theories such as clustering structural entropy to enrich the semantic feature space.

Recently, advances in State Space Models (SSMs), such as S4 [15] and Mamba [16], have significantly driven the development of large language models by balancing computational efficiency with long-range sequence modeling. Inspired by Mamba, a growing body of work has attempted to integrate it into AD tasks. For instance, MixMamba [8] combines Mixture-of-Experts with Mamba for time-series forecasting, while MambaAD [17] incorporates Mamba with two CNNs of different scales to capture both global and local features in image-domain AD. In addition, [18] proposed a Mamba-TCN sequence decomposition encoder for TSAD. However, existing methods either resort to a straightforward substitution of the Transformer or LSTM backbone with Mamba, or remain constrained by reconstruction- or prediction-based normality assumptions. As a result, these methods fall into the trap of optimizing reconstruction or generation quality, missing the essence of anomaly detection that lies in high-level semantic representation.

This paper proposes MOC, a novel **M**amba-based multi-scale **O**ne-**C**lass anomaly detection framework that leverages Mamba to capture long-term dependencies and CNNs for local dynamics, producing rich, high-level multi-scale representations. A classification projection module further clusters normal patterns into a compact hypersphere, with a hinge loss preventing hypersphere collapse. In addition, a soft-boundary loss is introduced to handle varying levels of training data contamination, ensuring robust anomaly discrimination. The contributions are summarized as follows:

- We present MOC, the first high-level semantic TSAD framework based on Mamba, shifting the focus from low-level reconstruction or generation to principled representation learning with a one-class objective.
- MOC integrates Mamba and CNNs to jointly capture long-term dependencies and local variation, yielding multi-scale time series representations.
- Extensive experiments on three public datasets demonstrate that MOC consistently outperforms state-of-the-art methods in time series anomaly detection.

2. PRELIMINARY

2.1. Problem Definition

Given a time series $\mathcal{S} = \{x_1, x_2, ..., x_M\}$ with length $\mathcal{M}$, where $x_t \in \mathbb{R}^d$ is a $d-$dimensional vector collected at time t. $d = 1$ means that the time series is univariate, and $d > 1$ for multivariate. We use sliding windows with length T to process $\mathcal{S}$ into subsequence set $\mathcal{D} = \{X_1, X_2, ..., X_N\}$, where $X_i = \{x_1, x_2, ..., x_T\}$is a subsequence of $\mathcal{S}$ with length T, and $\mathcal{N}$ is the number of the subsequence. In TSAD, the model calculates an anomaly score S_i for each X_i, and the higher S_i, the more likely X_i is an anomalous time series.

[*]Corresponding author (woty@buaa.edu.cn).

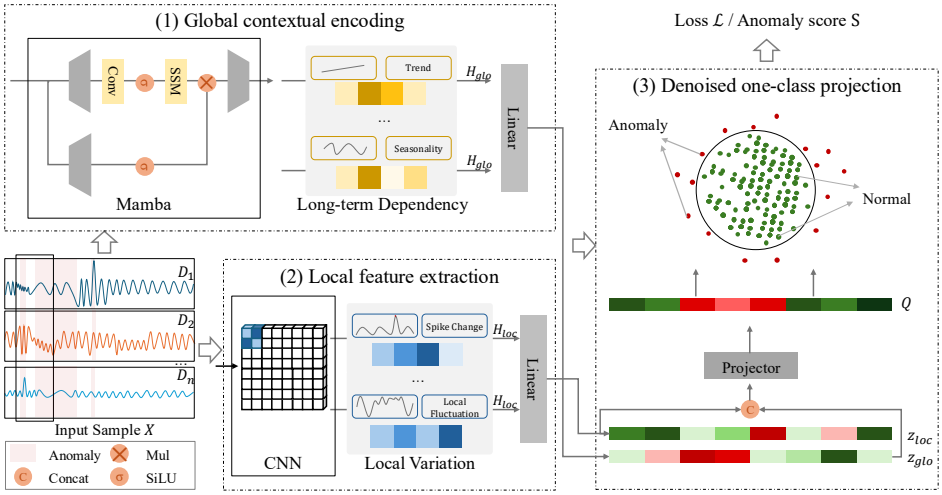


Fig. 1: Overall structure of the proposed MOC framework.

2.2. State Space Models

SSMs [15] map continuous input sequences $x(t)$ to output sequences $y(t)$ via a hidden state $h(t)$, which are typically described using ordinary differential equations as follows:

$$h^{'}(t) = \mathrm{A}h(t) + \mathrm{B}x(t), \quad y(t) = \mathrm{C}h(t), \tag{1}$$

where the state transition matrix $\mathrm{A} \in \mathbb{R}^{N\times N}$, $\mathrm{B} \in \mathbb{R}^{N\times 1}$ and $\mathrm{C} \in \mathbb{R}^{1\times N}$ for a state size N.

S4 [15] and Mamba [16] are discrete SSMs, applying zero-order hold with a timescale parameter Δ to derive the discrete-time parameters $\overline{\mathrm{A}}$ and $\overline{\mathrm{B}}$ from the continuous-time matrices A and B:

$$\overline{\mathrm{A}} = \exp(\Delta\mathrm{A}), \quad \overline{\mathrm{B}} = (\Delta\mathrm{A})^{-1}(\exp(\Delta\mathrm{A}) - \mathrm{I}) \cdot \Delta\mathrm{B}. \tag{2}$$

Then, the discretized model formulation can be represented as:

$$h_t = \overline{\mathrm{A}}h_{t-1} + \overline{\mathrm{B}}x_t, \quad y_t = \mathrm{C}h_t. \tag{3}$$

At last, the model computes the output by a global convolution for training as the following:

$$\overline{\mathrm{K}} = (\mathrm{C}\overline{\mathrm{B}}, \mathrm{C}\overline{\mathrm{A}}\overline{\mathrm{B}}, ..., \mathrm{C}\overline{\mathrm{A}}^{T-1}\overline{\mathrm{B}}), \quad y = x * \overline{\mathrm{K}}, \tag{4}$$

where $*$ represents convolution operation, and $\overline{\mathrm{K}} \in \mathbb{R}^{T}$ is a structured convolutional kernel.

3. METHODOLOGY

3.1. Overall Framework

The primary objective of MOC is one-class time-series classification, where the majority of normal samples are compactly enclosed within a hypersphere, while samples lying outside are deemed anomalous. The key intuition behind this work is that leveraging multi-scale temporal representations makes it easier to cluster normal samples within such a hypersphere. To this end, MOC integrates Mamba and CNN to jointly capture global dependencies and local dynamics, enabling more effective modeling of normal patterns.

Fig. 1 shows the overall architecture of the proposed MOC method, which consists of three main components: local feature extraction (LoFE), global contextual encoding (GCE), and denoised one-class projection (OC). For a time series $\mathcal{S}$, firstly, it is segmented to construct learning samples $\mathcal{X}$. Each sample $\mathcal{X}$ is simultaneously forwarded to the local and global encoders, enabling the extraction of sequence representations at different temporal scales. The multi-scale features are then fused to form a unified representation. The denoising projection module aggregates the variations of normal sequences into a compact hypersphere in the latent space, while simultaneously maintaining a discriminative boundary. Finally, the model not only reduces noise sensitivity but also enforces tighter clustering of normal instances, thereby improving robustness against anomalous patterns.

3.2. Multi-scale One-class Anomaly Detection

According to the characteristics of contextual variations, anomalies can generally be categorized into two types in TSAD: point- and pattern-wise anomalies [19]. Therefore, effective detection methods must not only focus on the long-term global dynamic characteristics of the sequence, but also capture its local temporal characteristics. Based on this consideration, this paper proposes introducing a multi-scale encoder into the framework to model the sequence at different temporal granularities, thereby simultaneously extracting both global and local features of the time series.

Local feature extraction. Local vibration, such as spike change or local fluctuation imbalance, often causes point-wise anomalies and short-term pattern-wise anomalies. This module aims to be sensitive to short-term changes and sudden jitter in time series. CNN demonstrates its strength in capturing local features, so this module introduces CNN to perform fine-grained local modeling of sequences.

Specifically, given a sequence segment $\mathcal{X}$, it first passes through a convolution operation, which makes the model more sensitive to point anomalies or local mutations:

$$h_{loc} = Conv2D_{(w,c)}(\mathcal{X}), \tag{5}$$

where w and c represent the time dimension and channel dimension of the convolution kernel, respectively. Then, the features are mapped and compressed:

$$z_{loc} = Linear(H_{loc}). \tag{6}$$

It is worth noting that the introduction of the CNN structure in this module can not only effectively detect point-wise anomalies, but also has good adaptability to short-term pattern-wise anomalies.

Global contextual encoding. In addition to local changes, many pattern anomalies are often hidden in longer-term contextual dependencies. Mamba [16] supports efficient modeling of long-term dependencies, so we take advantage of its strengths to capture global dynamic changes.

For the input sequence $\mathcal{X}$, we first normalize it to eliminate the bias of scale difference. The normalized sequence $\tilde{\mathcal{X}}$ is input into

a global feature encoder consisting of multiple mamba blocks and linear layers. Among them, Mamba models the long-term trends and seasonal cyclical patterns of the sequence.

$$h_{glo} = SSM(Conv(Linear(\tilde{\mathcal{X}}))) \odot \sigma(Linear(\tilde{\mathcal{X}})) \quad (7)$$

Subsequently, we map the global context features to the same latent feature space as the local features.

$$z_{glo} = Linear(H_{glo}). \quad (8)$$

Denoised one-class projection. Local and global features represent sequence characteristics at different scales and complement each other. Therefore, inspired by [20, 21], we concatenate them and input them into the fusion module, ensuring they are in the same space during the final detection.

$$Q = Projector(Concat(z_{loc}, z_{glo})). \quad (9)$$

To eliminate the influence of noise, the variation patterns of normal sequences are aggregated into a compact hypersphere in the latent space, ensuring that normal instances are closely clustered.

3.3. The MOC Objective

In the training phase, the objective function is constructed as follows:

$$\mathcal{L} = \alpha \cdot d + \beta \cdot v(Q), \quad (10)$$

where α and β are hyper-parameters controlling the contribution. d is the consistency term that aims to stabilize relative positions among sequences within the same category in the feature space:

$$d = \frac{1}{N}\sum_{i=1}^{N}[1 - sim(q_i, c) + ||q_i - c||_2^2], \quad (11)$$

where c is the one-class center calculated by $c = \frac{1}{N}\sum_{i=1}^{N} q_i$.

For training datasets containing a small number of anomalies, we adopt a hinge loss function defined as:

$$d_{soft} = L + \frac{1}{vN}\sum_{i=1}^{N}\max\{0, S_i - L\}, \quad (12)$$

where L is the $(1-v)$-quantile of S, hyper-parameter $v \in (0, 1]$ adjusts the amount of sequence allowed to be mapped outside the boundary. The anomaly score S_i is defined as $S_i = 1 - sim(q_i, c) + ||q_i - c||_2^2$. Inspired by [11], $v(Q)$ is defined as a hinge function on the standard deviation to prevent all outputs "collapse" to a constant:

$$v(Q) = \frac{1}{k}\sum_{i=1}^{k}\max\{0, \gamma - \sqrt{\mathrm{Var}(q_i + \epsilon)}\}, \quad (13)$$

where γ is a constant for standard deviation, k represents the dimension of Q, and ϵ as a small scalar is set to 10^{-4}.

During the test phase, whether $\mathcal{X}_i$ is classified as anomalous is determined by the formula below:

$$X_i = \begin{cases} anomaly, & S_i > \tau \\ normal, & S_i \leq \tau, \end{cases} \quad (14)$$

where τ is a predefined threshold. The comprehensive algorithm is described in the code repository.

4. EXPERIMENTS

4.1. Experimental Setup

Datasets. We conduct a comprehensive experimental evaluation on 3 datasets from diverse domains, overcoming the limitation of using a single dataset with domain-specific features: (1) **AIOps** [27]: encompasses well-maintained business cloud KPIs from prominent Internet companies. (2) **SWaT** [28]: contains sensors data from a scaled-down water treatment testbed. (3) **WADI** [29]: include sensors and actuators data from a reduced city water distribution system. The data statistics are summarized in Table 1.

Table 1: Summary of time series anomaly detection datasets.

	AIOps	SWaT	WADI
Number of sub-datasets	29	1	1
Variables	1	51	127
Domain	Cloud KPIs	Waterworks	Waterworks
Window size	32	32	32
Time step	32	16	16
Training	187741	29699	49035
Validation	36493	5624	2160
Testing	182414	28118	10799
Anomaly Rate	2.92%	5.96%	1.06%

Metrics. Currently, there is no unified standard for the selection of evaluation metrics for time series anomaly detection performance. To evaluate the fairness of each metric, [5] conducts analysis and discussion for widely used metrics. As a result, we choose Revised Point Adjusted (RPA) Precision, Recall, and F1-score [30] to conduct a more comprehensive comparative analysis.

Baselines. We evaluated the effectiveness of MOC for anomaly detection with established baselines, including traditional models: One-Class SVM (OC-SVM) [22], Isolation Forest (IF) [23], and Spectral Residual (SR) [24]. Inspired by [5, 31], we design a simple baseline, the Randomized Anomaly Score (RAS); deep learning models: LSTM Encoder-decoder (LSTM-ED) [6], Deep oneclass (Deep SVDD) [9], TCC [25, 26], AOC [10], COCA [11], and TranAD [7]. In addition, MambaAD [17] and MemMambaAD [18] are excluded as the former targets the image domain and the latter lacks an open-score implementation, so we selected MixMamba [8].

Implementation. We adopt a learning rate from $1e$ - 4 to $5e$ - 4, weight decay of $5e$ - 4, β_1 = 0.9, and β_2 = 0.99 in an Adam optimizer. All the models are built with Pytorch 2.6 and Merlion 2.0.0 [32], and trained on an NVIDIA Tesla V100 GPU. The code and all parameters are available at https://github.com/lottieW/MOC.

4.2. Experimental Results and Analysis

Main Results. The overall performance of the aforementioned methods is shown in Table 2. MOC outperforms the baselines(in terms of F1 score) for all datasets, and even surpasses the others by a large margin on AIOps and WADI. Specifically, compared to the best of the baselines, the proposed method improves F1 from 43.74% to 53.66%(+9.92%) on AIOps, and from 9.77% to 13.15%(+3.38%) on WADI. We observe two key phenomena: firstly, compared with other one-class methods [9–11], our approach achieves superior performance by capturing multi-scale features. Secondly, in contrast to existing multi-scale methods [8], our approach introduces the one-class, which enables the model to focus more effectively on distinguishing anomalies and thereby enhances detection performance. Therefore, by exploiting all inherent dependencies in sequence variability, MOC is more sensitive to normal and anomaly points.

Ablation Study. As shown in Table 3, the results reveal some insights into the effectiveness of our proposed method's modules. For AIOps, which mainly contains point-wise anomalies, removing LoFE leads to a drastic drop in the F1 score. In contrast, SWaT and WADI include long-term pattern-wise anomalies, with performance substantially degraded in the absence of GCE. These results indicate the importance of multi-scale feature modeling. Meanwhile, the performance on all datasets nearly collapses without OC, confirming the pivotal role of OC in denoising for TSAD in real-world

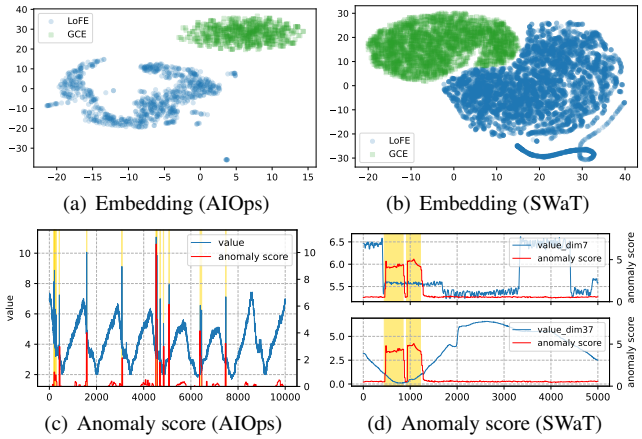
Table 2: Average RPA Precision (%), Recall (%), and F1-score(%) with standard deviation for baselines, our method on different datasets over 10 runs. The terms highlighted in bold indicate the optimal results, while the second-best results are underlined. SR does not support multivariate time series anomaly detection.

Method \ Dataset	AIOps			SWaT			WADI		
	P	R	F1	P	R	F1	P	R	F1
OC-SVM [22]	3.71	66.33	7.02	0.01	**97.14**	0.02	0.02	**100.00**	0.03
IF [23]	1.98±0.00	69.49±0.83	3.86±0.07	**57.22±38.46**	8.00±1.14	12.79±1.94	0.44±0.08	42.86±9.04	0.87±0.16
SR [24]	4.47	**91.02**	8.52	-	-	-	-	-	-
RAS	3.12±0.57	23.60±4.85	5.46±0.83	7.58±2.73	22.29±10.67	10.30±2.00	6.14±2.84	32.86±12.04	9.77±3.55
LSTM-ED [6]	7.85±0.33	67.26±0.94	14.05±0.52	3.39±0.03	14.29±0.00	5.48±0.04	2.24±0.35	14.29±0.00	3.86±0.52
Deep SVDD [9]	23.02±8.53	32.83±8.08	25.53±7.38	8.62±12.08	30.57±8.09	8.88±6.97	3.54±1.41	32.14±14.37	6.14±2.09
TCC [25, 26]	1.67±0.42	18.57±3.09	3.03±0.66	1.17±0.44	17.14±14.85	2.06±0.66	12.40±29.26	45.71±39.02	5.02±3.79
AOC [10]	33.28±3.73	55.27±3.42	41.40±3.28	34.78±0.01	22.86±0.01	27.59±0.02	0.18±0.01	57.14±0.02	0.36±0.01
COCA [11]	39.95±8.42	49.40±7.27	43.74±6.90	16.05±13.40	29.52±7.13	15.83±6.81	2.47±1.31	37.50±32.09	4.27±2.04
TranAD [7]	5.05±0.02	28.69±0.19	8.59±0.04	27.66±1.96	21.14±1.48	23.86±0.21	2.22±0.00	28.57±0.00	4.12±0.00
MixMamba [8]	9.33±0.07	79.98±0.15	16.72±0.12	1.23±0.03	50.00±2.02	2.39±0.07	2.51±0.22	33.57±5.88	4.66±0.39
MOC	**53.86±5.97**	54.04±5.00	**53.66±3.76**	35.88±5.02	23.43±3.57	**27.90±2.31**	**40.89±34.12**	15.00±12.14	**13.15±9.01**

Table 3: Ablation study results (RPA F1) for variants of MOC.

Method \ Dataset	AIOps	SWaT	WADI
w/o LoFE	1.48±0.00	10.13±2.84	12.67±4.36
w/o GCE	45.57±6.69	18.34±12.27	6.41±5.52
w/o OC	2.92±0.07	0.55±0.00	0.28±0.00
MOC	**53.66±3.76**	**27.90±2.31**	**13.15±9.01**

scenarios. Overall, this validates the adaptability and robustness of MOC across diverse anomaly types.



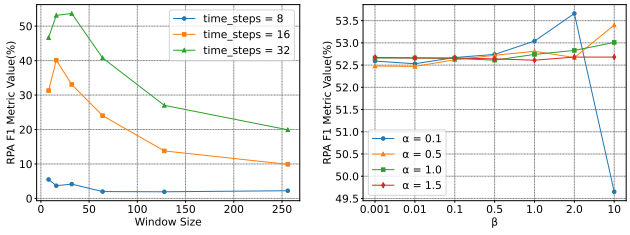
(a) Embedding (AIOps) (b) Embedding (SWaT)

(c) Anomaly score (AIOps) (d) Anomaly score (SWaT)

Fig. 2: The visualization of MOC on AIOps and SWaT datasets.

Visual Analysis. As shown in Fig. 2 (a, b), LOFE and GCE yield embeddings with distinguishable structures in Umap [33] space, indicating that the two modules capture complementary aspects of sequence dynamics. In (c, d), the red curves exhibit higher anomaly scores around the yellow anomaly regions, which is consistent with the ground truth. The visualization demonstrates that our method, by effectively capturing multi-scale dependencies, achieves high accuracy and robustness when detecting diverse types of anomalies.

Hyperparameter Analysis. We conducted sensitivity analysis on the AIOps to study the impact of hyperparameters, including the window size Ws, time steps Ts, and the weight parameter controlling α, β, as shown in Fig. 3. As shown in Fig. 3(a), it is evident that within a certain range, increasing the degree of the window size



(a) Ws, Ts (AIOps) (b) α, β (AIOps)

Fig. 3: Sensitivity analysis of MOC on AIOps dataset.

Ws and time steps Ts enhances the ability of MOC to capture multi-scale features, leading to improved performance. However, beyond a certain point, it is not conducive to MOC learning complex features in time series. We observe that Ws and Ts $= 32$ perform the best. Fig. 3(b) shows that the stabilization term $v(Q)$ plays an important role in the optimization function Eq. 10. The model demonstrates superior performance when the weight β is set to 2.0.

5. CONCLUSION

We propose MOC, the first Mamba-based high-level semantic framework that unifies multi-scale representation learning with one-class classification for principled TSAD. By jointly leveraging Mamba for global dependency modeling and CNN for local dynamics, MOC is effective in handling complex dependencies from short to long terms. Meanwhile, introducing a one-class projection module further clusters normal patterns into a compact hypersphere, by a hinge loss to prevent hypersphere collapse, MOC outperforms state-of-the-art methods. Moreover, a soft-boundary loss is introduced to handle varying levels of training data contamination, ensuring robust anomaly discrimination. In future work, we aim to further extend our framework to broader real-world scenarios and integrate domain-specific prior knowledge to further enhance practicality.

6. ACKNOWLEDGMENT

This work is supported in part by National Key R&D Program of China (No. 2024YFB4505901), NSFC Grant (No. 62402024), Beijing Natural Science Foundation Grant (No. L241050), and Fundamental Research Funds for the Central Universities.

7. REFERENCES

[1] Rui Wang, Xudong Mou, Tianyu Wo, Mingyang Zhang, Yuxin Liu, Tiejun Wang, Pin Liu, Jihong Yan, and Xudong Liu, "Acbot: an iiot platform for industrial robots," *Frontiers of Computer Science*, vol. 19, no. 4, pp. 194203, 2025.

[2] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel, "Deep learning for anomaly detection: A review," *ACM computing surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.

[3] Mengyuan Ma, Tiejun Wang, Rui Wang, Xudong Mou, Tianyu Wo, and Xudong Liu, "Foca: Foundation-model-based one-class anomaly detection for time series," in *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2025, pp. 1–8.

[4] Xudong Mou, Rui Wang, Bo Li, Tianyu Wo, Jie Sun, Hui Wang, and Xudong Liu, "Roca: Robust contrastive one-class time series anomaly detection with contaminated data," *arXiv preprint arXiv:2503.18385*, 2025.

[5] Rui Wang, Xudong Mou, Renyu Yang, Kai Gao, Pin Liu, Chongwei Liu, Tianyu Wo, and Xudong Liu, "Cutaddpaste: Time series anomaly detection by exploiting abnormal knowledge," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 3176–3187.

[6] Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff, "Lstm-based encoder-decoder for multi-sensor anomaly detection," *arXiv preprint arXiv:1607.00148*, 2016.

[7] Shreshth Tuli, Giuliano Casale, and Nicholas R Jennings, "Tranad: Deep transformer networks for anomaly detection in multivariate time series data," *arXiv preprint arXiv:2201.07284*, 2022.

[8] Khaled Alkilane, Yihang He, and Der-Horng Lee, "Mixmamba: Time series modeling with adaptive expertise," *Information Fusion*, vol. 112, pp. 102589, 2024.

[9] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft, "Deep one-class classification," in *International conference on machine learning*. PMLR, 2018, pp. 4393–4402.

[10] Xudong Mou, Rui Wang, Tiejun Wang, Jie Sun, Bo Li, Tianyu Wo, and Xudong Liu, "Deep autoencoding one-class time series anomaly detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[11] Rui Wang, Chongwei Liu, Xudong Mou, Kai Gao, Xiaohui Guo, Pin Liu, Tianyu Wo, and Xudong Liu, "Deep contrastive one-class time series anomaly detection," in *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*. SIAM, 2023, pp. 694–702.

[12] Hao Peng, Xiang Huang, Shuo Sun, Ruitong Zhang, and Xizhao Wang, "Adaptive and robust dbscan with multi-agent reinforcement learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[13] Xianghua Zeng, Hao Peng, and Angsheng Li, "Proactive bot detection based on structural information principles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[14] Jingyun Zhang, Hao Peng, Mingdai Yang, and Philip S Yu, "Enhanced pre-training for recommendation via hypergraph structural entropy," *ACM Transactions on Information Systems*, vol. 44, no. 2, pp. 1–48, 2025.

[15] Albert Gu, Karan Goel, and Christopher Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[16] Albert Gu and Tri Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[17] Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie, "Mambaad: Exploring state space models for multi-class unsupervised anomaly detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 71162–71187, 2024.

[18] Gang Li, Mingchao Ge, Jin Wan, Delong Han, Min Li, and Mingle Zhou, "Memmambaad: Memory-augmented state space model for multivariate time series anomaly detection," *Engineering Applications of Artificial Intelligence*, vol. 158, pp. 111308, 2025.

[19] Kwei-Herng Lai, Daochen Zha, Junjie Xu, Yue Zhao, Guanchu Wang, and Xia Hu, "Revisiting time series outlier detection: Definitions and benchmarks," in *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 1)*, 2021.

[20] Guangjie Zeng, Hao Peng, Angsheng Li, Jia Wu, Chunyang Liu, and Philip S Yu, "Scalable semi-supervised clustering via structural entropy with different constraints," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[21] Qitong Liu, Hao Peng, Xiang Huang, Zhifeng Hao, Qingyun Sun, Zhengtao Yu, and Philip S Yu, "Hierarchical text classification optimization via structural entropy and singular smoothing," *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[22] Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt, "Support vector method for novelty detection," *Advances in neural information processing systems*, vol. 12, 1999.

[23] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou, "Isolation-based anomaly detection," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 6, no. 1, pp. 1–39, 2012.

[24] Hansheng Ren, Bixiong Xu, Yujing Wang, Chao Yi, Congrui Huang, Xiaoyu Kou, Tony Xing, Mao Yang, Jie Tong, and Qi Zhang, "Time-series anomaly detection service at microsoft," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 3009–3017.

[25] Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minho Jin, and Tomas Pfister, "Learning and evaluating representations for deep one-class classification," *arXiv preprint arXiv:2011.02578*, 2020.

[26] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan, "Time-series representation learning via temporal and contextual contrasting," *arXiv preprint arXiv:2106.14112*, 2021.

[27] "Aiops challenge. the 1st match for aiops.," `https://github.com/NetManAIOps/KPI-Anomaly-Detection`, 2018, Accessed: 2024-12-04.

[28] Aditya P Mathur and Nils Ole Tippenhauer, "Swat: A water treatment testbed for research and training on ics security," in *2016 international workshop on cyber-physical systems for smart water networks (CySWater)*. IEEE, 2016, pp. 31–36.

[29] Chuadhry Mujeeb Ahmed, Venkata Reddy Palleti, and Aditya P Mathur, "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*, 2017, pp. 25–28.

[30] Kyle Hundman, Valentino Constantinou, et al., "Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 387–395.

[31] Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon, "Towards a rigorous evaluation of time-series anomaly detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 7194–7201.

[32] Aadyot Bhatnagar, Paul Kassianik, Chenghao Liu, Tian Lan, Wenzhuo Yang, Rowan Cassius, Doyen Sahoo, Devansh Arpit, Sri Subramanian, Gerald Woo, et al., "Merlion: A machine learning library for time series. 2021," *URL https://arxiv. org/abs/2109.09265*, 2021.

[33] Leland McInnes, John Healy, and James Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.