Attention-Driven Deep Learning Framework for Intelligent Anomaly Detection in ETL Processes

Cong Nie , Haige Wang , Chifu Chiang *

Posted Date: 9 December 2025

doi: 10.20944/preprints202512.0884.v1

Keywords: ETL process; anomaly detection; time series modeling; attention mechanism

Article

Attention-Driven Deep Learning Framework for Intelligent Anomaly Detection in ETL Processes

Cong Nie [1], Haige Wang [2] and Chifu Chiang [3,*]

[1] Washington University in St. Louis, St. Louis, USA
[2] University of Miami, Miami, USA
[3] University of Connecticut, Hartford, USA
* Correspondence: chiang.chifu@gmail.com

Abstract

This paper addresses the problem of anomaly detection in multi-source heterogeneous data within the ETL (Extract-Transform-Load) process and proposes an intelligent detection framework that integrates temporal modeling and attention mechanisms. The method achieves effective dynamic aggregation of multidimensional features and temporal dependency modeling in ETL logs through the coordinated design of feature encoding, gated recurrent modeling, and multi-head attention allocation. At the feature level, the model uses a unified encoding structure to map raw logs, monitoring metrics, and task status information into a high-dimensional latent space, ensuring consistency of feature scales and completeness of information. At the temporal level, a GRU-based time modeling structure is introduced to capture long-term dependencies, enhancing the model's ability to perceive the evolution of anomaly patterns. At the attention level, a multi-head mechanism is applied to weight different time segments and feature dimensions, enabling adaptive focus on key moments and important features. Finally, the model combines anomaly scoring with distribution consistency constraints to achieve accurate identification and discrimination of potential anomalies. Experimental results show that the proposed framework significantly outperforms traditional rule-based detection, statistical methods, and basic deep models across various ETL task scenarios, demonstrating higher detection accuracy, stability, and generalization capability. The findings verify the effectiveness of integrating temporal modeling and attention mechanisms for anomaly detection in complex data streams and provide a feasible solution for building reliable and scalable intelligent ETL monitoring systems.

Keywords: ETL process; anomaly detection; time series modeling; attention mechanism

I. Introduction

In modern data-driven enterprise information systems, the ETL (Extract-Transform-Load) process serves as the core component of data warehouse and analytics operations. It is responsible for extracting raw data from heterogeneous sources, performing transformation and cleaning, and loading the processed data into target databases. However, as data volume grows exponentially and business logic becomes increasingly complex, the stability and reliability of ETL workflows have become major bottlenecks affecting data quality and system efficiency. The emergence of abnormal data not only leads to deviations in analytical results but can also trigger cascading effects such as task failures, delays, and resource waste, ultimately compromising decision accuracy and business continuity [1]. Traditional rule-based or threshold-based monitoring approaches are limited in capturing complex temporal dependencies and contextual relationships within dynamic ETL environments [2]. Therefore, developing an intelligent and generalizable anomaly detection framework has become particularly critical.

Data anomalies in ETL processes exhibit high diversity and strong temporal variability. Anomalies may arise from sudden fluctuations in input data, logical errors, missing values, or format

 doi:10.20944/preprints202512.0884.v1

mismatches during transformation, and can also result from unstructured external factors such as interface delays or network instability [3–5]. This complexity makes anomalies often hidden, nonlinear, and multi-scale in nature, posing challenges to traditional static detection methods in accurately characterizing their evolution. Moreover, ETL tasks typically exhibit strong temporal dependencies. The execution of different stages follows causal and sequential constraints, where small deviations in earlier steps can amplify in later stages. Hence, the key to anomaly identification lies in deep modeling of time-series features, capturing trends, periodicities, and abrupt changes to uncover potential abnormal behavior patterns and provide early warnings and automated recovery cues [6–9].

With the rapid development of artificial intelligence and deep learning, the integration of temporal modeling and attention mechanisms has opened new directions for anomaly detection [10]. Compared with traditional statistical models or sliding-window monitoring methods, deep temporal networks can learn complex time dependencies through nonlinear mappings, while attention mechanisms dynamically assign weights to features, focusing on key moments and critical attributes in vast historical data [11]. This integration not only enhances the model's ability to capture long-term dependencies and multi-dimensional feature interactions but also improves interpretability and localization accuracy of anomalies. In ETL scenarios, such a fusion mechanism enables multi-perspective anomaly perception by jointly modeling task logs, data flow features, and system metrics, thereby overcoming the performance bottlenecks of traditional frameworks under high-dimensional and non-stationary conditions.

From an application standpoint, anomaly detection in ETL workflows is essential not only for data pipeline stability and security but also as a key component of data governance and intelligent operations. With the proliferation of cloud computing, the Internet of Things, and enterprise data platforms, the number and diversity of data sources have increased dramatically, resulting in unprecedented process complexity. Intelligent and automated anomaly detection can significantly reduce manual monitoring costs, enhance data flow efficiency, and provide reliable foundations for higher-level analytics. In domains such as finance, manufacturing, and healthcare, ETL pipelines handle mission-critical data streams, and the effectiveness of anomaly detection directly impacts risk management, production scheduling, and decision reliability. Thus, designing intelligent anomaly detection frameworks for ETL processes has strong engineering value and promotes the advancement of autonomous and trustworthy data systems.

In summary, the ETL anomaly detection framework that integrates temporal modeling and attention mechanisms aims to model and accurately identify complex abnormal behaviors from both temporal dependency and feature association perspectives. This research direction aligns with the evolution toward intelligent data operations and automated control while providing technical support for the stable operation of large-scale data systems. By establishing unified feature perception and temporal inference mechanisms, early detection and global correlation analysis of anomalies can be achieved across heterogeneous environments. This enhances the adaptability, interpretability, and long-term reliability of ETL processes. The proposed research has profound theoretical and practical significance for advancing intelligent data infrastructure, ensuring robust enterprise data ecosystems, and strengthening the reliability of data-driven decision-making systems.

II. Method

The abnormal data identification framework for ETL processes proposed in this paper consists of a feature encoding layer, a time series modeling layer, an attention interaction layer, and an anomaly discrimination layer, and is designed as a unified architecture rather than a set of isolated components. At the feature level, we adopt an ETL-oriented deep representation paradigm similar to autoencoder-based anomaly detection in enterprise ETL logs, mapping heterogeneous inputs such as raw logs, monitoring metrics, and task status into a unified high-dimensional latent space to ensure consistency of feature scales and completeness of information[12].On top of this representation, the time series modeling layer introduces gated recurrent units to capture long-term temporal

 doi:10.20944/preprints202512.0884.v1

dependencies in ETL workflows[13], while the attention interaction layer integrates a multi-head self-attention mechanism to dynamically weight different time segments and feature dimensions, directly applying attention-based anomaly identification strategies that have been shown effective for complex sequential data to the ETL domain [14].Finally, the anomaly discrimination layer combines anomaly scoring with distribution consistency constraints, forming an end-to-end pipeline that links multi-dimensional feature embedding, temporal dependency modeling, and attention-based interaction to accurately capture abnormal patterns in complex ETL data streams. The model architecture is shown in Figure 1.

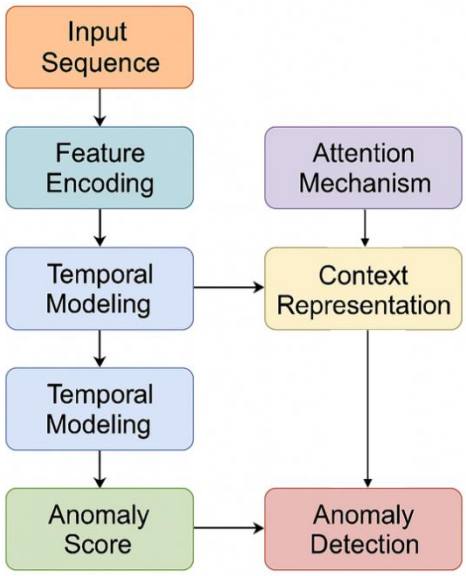

Figure 1. Overall model architecture.

First, a unified feature encoding mechanism was designed to address the multi-source heterogeneity of ETL logs and monitoring metrics. Assume the input sequence is $X = \{x_1, x_2, ..., x_T\}$, where $x_t \in R^d$ represents the multidimensional feature vector at time step t. The encoding layer maps the original features into a high-dimensional latent space using linear mapping and nonlinear activation functions:

$$H_t = \sigma(W_e x_t + b_e) \quad (1)$$

Here, $W_e \in R^{d_h \times d}$ is the learnable weight matrix, b_e is the bias term, and $\sigma(\cdot)$ is the activation function (such as ReLU). This process unifies the structure of the original data and lays the foundation for subsequent temporal relationship modeling.

During the time series modeling phase, a dynamic encoding structure based on a gated recurrent unit (GRU) was introduced to capture the temporal dependencies of ETL data flows. For any time step t, the model adaptively controls the transfer and forgetting of historical information through update gates and reset gates:

$$z_t = \sigma(W_z H_t + U_z h_{t-1}) \quad (2)$$

$$r_t = \sigma(W_r H_t + U_r h_{t-1}) \quad (3)$$

$$h_t = \tanh(W_h H_t + U_h(r_t \otimes h_{t-1})) \quad (4)$$

 doi:10.20944/preprints202512.0884.v1

$$h_t = (1 - z_t) \otimes h_{t-1} + z_t \otimes \tilde{h}_t \qquad (5)$$

Where $\otimes$ represents element-wise multiplication and h_t represents the hidden state representation at time step t. This structure can maintain gradient stability in long sequence tasks, thereby capturing the temporal evolution characteristics of abnormal patterns.

To further strengthen the model's focus on key moments and feature dimensions, the framework introduces a multi-head attention mechanism after time series modeling. The attention weight is obtained through a weighted calculation of the query, key, and value:

$$Attention(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V \qquad (6)$$

Here, $Q = HW_Q, K = HW_K, V = HW_V, W_Q, W_K, W_V$ is a learnable matrix, and d_k represents the normalization factor for the feature dimension. Through this multi-head architecture, the model can learn global dependencies between features from different subspaces, thereby improving its sensitivity to complex abnormal behaviors. In ETL scenarios, this mechanism can focus on time series mutation points and changes in key indicators, enabling dynamic perception of hidden anomalies.

Finally, the anomaly discrimination layer constructs an anomaly score function based on the temporal context and attention embedding results to characterize the degree of deviation of the data point. Let the fused context be represented as c_t, then the anomaly score can be defined as:

$$s_t = \| h_t - c_t \|_2^2 \qquad (7)$$

This score reflects the difference between the predicted distribution and the actual observation. A larger value indicates a higher probability of anomaly. To achieve end-to-end optimization, the overall objective function takes into account both the reconstruction error and the temporal consistency constraint:

$$L = \sum_{t=1}^{T} (\| \hat{x}_t - x_t \|_2^2 + \lambda \| h_t - h_{t-1} \|_1) \qquad (8)$$

Here, $\hat{x}_t$ represents the model reconstruction result, and λ represents the balance coefficient. This loss function ensures the distinguishability of anomalies while constraining the smoothness of the time series representation, thereby improving the robustness and generalization of detection. This holistic approach implements full-link modeling, from feature representation to time series dependency and anomaly scoring, providing a systematic solution for intelligent anomaly identification in ETL processes.

III. Performance Evaluation

A. Dataset

This study employs the Kaggle ETL Logs Anomaly Dataset as the primary data source. The dataset is derived from real enterprise-level ETL task execution logs and covers several typical ETL process stages, including data extraction, transformation, loading, and scheduling monitoring. It contains more than 500,000 time-series records, and each record includes feature fields such as task status, execution duration, CPU and memory usage, I/O throughput, latency indicators, error codes, and upstream and downstream dependency information. The dataset records ETL process dynamics at a minute-level time resolution, providing a comprehensive view of task behavior characteristics

 doi:10.20944/preprints202512.0884.v1

and potential anomaly patterns. It is well-suited for studies on various types of anomaly detection and modeling.

During data preprocessing, the raw log data is first uniformly formatted and standardized to eliminate scale differences and unit inconsistencies between tasks. Missing and abnormal values are repaired using median imputation and sliding window smoothing to ensure the continuity and statistical stability of the feature sequences. Then, a time-window segmentation mechanism divides long log sequences into fixed-length subsequences. Each subsequence contains time-dependent information across multiple metrics and serves as input for temporal modeling and feature alignment. The labels are generated based on system operation records and task status indicators. Anomalous samples include multiple categories such as task interruption, timeout, data mismatch, and performance degradation.

The dataset has significant advantages in terms of its high-dimensional and heterogeneous nature, as well as its strong temporal characteristics. It enables comprehensive validation of the effectiveness and robustness of temporal modeling and attention mechanisms in complex ETL scenarios. The data source is authentic and diverse, covering ETL tasks of varying scales and structures, ensuring strong representativeness and generalization potential. Furthermore, its structured features and rich contextual information provide a solid foundation for multimodal fusion and anomaly interpretation, offering a reproducible research platform for intelligent ETL monitoring and automated anomaly identification.

B. Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table 1. Comparative experimental results.

Method	Precision	Recall	F1-score	AUC	ACC
DCdetector [15]	0.871	0.842	0.856	0.892	0.878
Transformer[16]	0.896	0.863	0.879	0.911	0.887
1DCNN[17]	0.882	0.854	0.867	0.903	0.881
MLP[18]	0.861	0.831	0.845	0.888	0.872
GAT[19]	0.908	0.877	0.892	0.923	0.901
Ours	0.935	0.913	0.924	0.948	0.931

The proposed ETL anomaly detection framework, which integrates temporal modeling and attention mechanisms, outperforms all comparison methods across every metric, achieving a Precision of 0.935 and a Recall of 0.913 while maintaining a strong balance between false positives and false negatives. Static-feature models such as MLP and 1DCNN perform poorly in ETL scenarios due to their inability to capture contextual and long-range temporal dependencies, falling behind sequence-aware structures like Transformer and GAT. Compared with these stronger baselines, the proposed method further improves AUC by about 0.025 and F1-score by about 0.032, reflecting enhanced robustness and generalization through its combination of gated recurrent units for short-term dependencies and multi-head attention for key temporal feature extraction. This joint design enables adaptive weighting of non-stationary patterns in ETL logs and strengthens anomaly sensitivity. Superior AUC and ACC results also show that the model maintains stable discrimination by constructing a multi-level representation space that suppresses noise and false signals. Overall, the findings confirm the practical value of the proposed temporal-attention fusion framework, with the sensitivity of window length (L) to F1-score illustrated in Figure 2.

 doi:10.20944/preprints202512.0884.v1

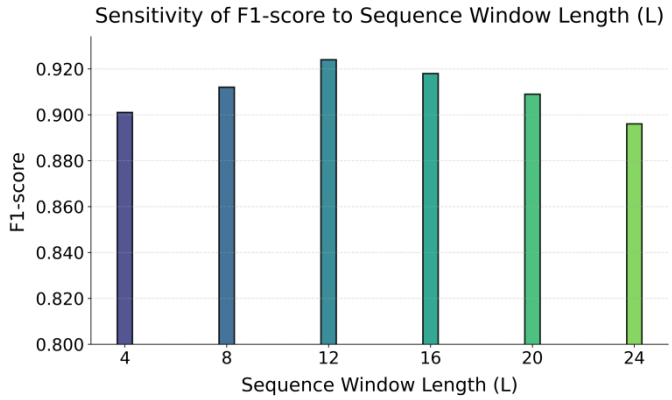

Figure 2. Sensitivity experiment of sequence window length L to F1-score.

The experimental results indicate that the sequence window length L strongly influences the model's F1-score: small windows (e.g., $L = 4$) restrict temporal perception and hinder the capture of cross-stage dependencies, while moderate windows (around $L = 12$) provide the best balance between local and global context, yielding the highest F1-score of 0.924. Larger windows beyond $L = 16$ introduce redundant information, cause feature dilution, and lead to "memory interference," slightly reducing performance. These findings show that window length critically affects temporal representation and anomaly discrimination, and that selecting an appropriate L enhances contextual fusion, hierarchical feature extraction, and anomaly localization. Overall, the results verify the temporal-attention fusion framework's sensitivity to time-dependent patterns and highlight the importance of dynamically optimizing the time-window parameter. Additionally, the sensitivity experiment on the number of attention heads $\hbar$ and its effect on ACC is illustrated in Figure 3.

The experimental results show that the variation in the number of attention heads h has a significant impact on the overall recognition accuracy (ACC) of the model. When the number of heads is small, such as h=1 or h=2, the model's ability to focus on relevant features is limited. It cannot fully capture the dependencies among multidimensional features in ETL logs, leading to insufficient information extraction. As a result, the ACC remains at a relatively low level. With the increase in the number of attention heads, the model can perform parallel computations in higher-dimensional subspaces, enhancing its multi-perspective perception of anomaly patterns. The accuracy gradually improves, indicating that the multi-head mechanism provides clear advantages for modeling complex dependencies in ETL scenarios.

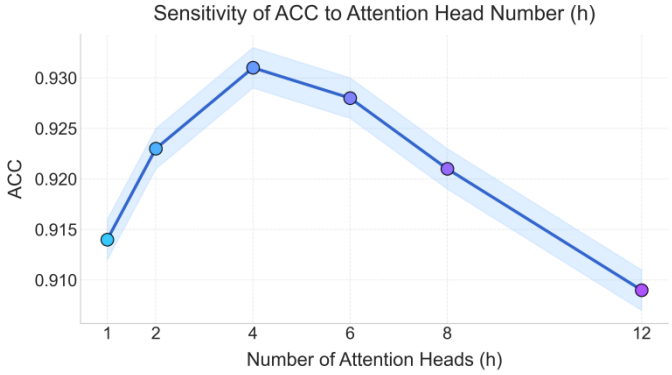

Figure 3. Experiment on the sensitivity of the number of attention heads h to ACC.

When h=4, the model achieves the highest ACC of 0.931, showing that the attention allocation at this setting reaches the optimal balance between feature differentiation and information aggregation. A moderate number of attention heads allows the model to accurately weight key time points and

 doi:10.20944/preprints202512.0884.v1

feature dimensions without losing global information. This helps maintain strong stability and generalization in anomaly recognition. The performance improvement at this stage also highlights the central role of multi-head attention in temporal modeling tasks, as it enables efficient abstraction of complex contextual relationships through hierarchical feature focusing.

However, when the number of attention heads continues to increase, such as h>6, the model performance shows a slight decline. This indicates that too many attention channels may introduce redundant information and cause excessive dispersion in the representation space. The features learned across different attention subspaces can interfere with one another, increasing optimization difficulty and reducing convergence efficiency. Moreover, a large number of heads leads to higher computational cost and less training stability, weakening the model's robustness to minor noise or feature shifts.

In summary, this experiment verifies that the proposed temporal-attention fusion framework is highly sensitive to attention structure parameters. A properly chosen number of attention heads can effectively enhance the model's global feature perception and focus on critical moments in ETL anomaly detection. It ensures stable detection performance in complex, multi-source, and non-stationary data environments. The results indicate that the model architecture should maintain moderate complexity to balance representational capacity and training stability, providing optimal parameter configuration guidance for intelligent ETL anomaly recognition.

IV. Conclusions

This paper focuses on the automated recognition of complex anomaly data in ETL processes and proposes an intelligent detection framework that integrates temporal modeling and attention mechanisms. The method achieves dynamic aggregation of multidimensional heterogeneous log features and temporal dependency modeling through the coordinated design of feature encoding, gated recurrent modeling, and multi-head attention allocation. Experimental results demonstrate that the proposed framework exhibits excellent anomaly detection performance under various complex data scenarios, especially when dealing with long-term dependencies, sudden anomalies, and high-dimensional feature interference. This study not only overcomes the limitations of traditional ETL monitoring systems that rely on fixed rules and static thresholds but also provides a new intelligent approach for automated anomaly discovery and adaptive decision-making in data operations.

From a theoretical perspective, this work further validates the synergistic potential of deep temporal structures and attention mechanisms in multi-source dynamic data modeling. By introducing multi-scale time windows and feature-level attention aggregation, the model captures global information while focusing on critical features and time points, enabling high-precision representation of anomaly patterns. The design concept of this structure is not limited to ETL log analysis but can also be extended to other scenarios requiring multi-temporal signal interaction modeling, such as real-time system monitoring, complex event processing, industrial process optimization, and distributed task scheduling. The proposed framework demonstrates strong scalability in cross-modal and multi-stage information fusion, providing a transferable modeling paradigm for future research.

Future work can be further extended in several directions. One direction is to explore lightweight and efficient architectural designs to meet the real-time monitoring requirements of cloud and edge computing environments. Another direction is to incorporate self-supervised pretraining and online learning mechanisms, allowing the model to continuously adapt and optimize in dynamic environments. In addition, causal inference and anomaly explanation modules can be introduced to enhance interpretability and decision transparency in complex ETL ecosystems. Overall, this study lays the foundation for building reliable and scalable intelligent data operation systems and holds significant practical value and application potential for advancing the intelligence and autonomy of large-scale data infrastructures.

 doi:10.20944/preprints202512.0884.v1

References

1. S. Tuli, G. Casale and N. R. Jennings, "Tranad: Deep transformer networks for anomaly detection in multivariate time series data," arXiv preprint arXiv:2201.07284, 2022.
2. Y. Xing, Y. Deng, H. Liu, M. Wang, Y. Zi and X. Sun, "Contrastive Learning-Based Dependency Modeling for Anomaly Detection in Cloud Services," arXiv preprint arXiv:2510.13368, 2025.
3. X. Yan, Y. Jiang, W. Liu, D. Yi and J. Wei, "Transforming Multidimensional Time Series into Interpretable Event Sequences for Advanced Data Mining," 2024 5th International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI), pp. 126-130, 2024.
4. N. Lyu, Y. Wang, Z. Cheng, Q. Zhang and F. Chen, "Multi-Objective Adaptive Rate Limiting in Microservices Using Deep Reinforcement Learning," arXiv preprint arXiv:2511.03279, 2025.
5. Y. Sun, R. Zhang, R. Meng, L. Lian, H. Wang and X. Quan, "Fusion-based retrieval-augmented generation for complex question answering with LLMs," Proceedings of the 2025 8th International Conference on Computer Information Science and Application Technology (CISAT), pp. 116-120, July 2025.
6. X. Yan, J. Du, L. Wang, Y. Liang, J. Hu and B. Wang, "The Synergistic Role of Deep Learning and Neural Architecture Search in Advancing Artificial Intelligence", Proceedings of the 2024 International Conference on Electronics and Devices, Computational Science (ICEDCS), pp. 452-456, Sep. 2024.
7. R. Ying, J. Lyu, J. Li, C. Nie and C. Chiang, "Dynamic Portfolio Optimization with Data-Aware Multi-Agent Reinforcement Learning and Adaptive Risk Control," 2025.
8. P. Xue and Y. Yi, "Sparse Retrieval and Deep Language Modeling for Robust Fact Verification in Financial Texts," Transactions on Computational and Scientific Methods, vol. 5, no. 10, 2025.
9. Z. Xu, J. Xia, Y. Yi, M. Chang and Z. Liu, "Discrimination of Financial Fraud in Transaction Data via Improved Mamba-Based Sequence Modeling," 2025.
10. G. Yao, H. Liu and L. Dai, "Multi-agent reinforcement learning for adaptive resource orchestration in cloud-native clusters," arXiv preprint arXiv:2508.10253, 2025.
11. Y. Lin, "Graph neural network framework for default risk identification in enterprise credit relationship networks," Transactions on Computational and Scientific Methods, vol. 4, no. 4, 2024.
12. X. Chen, S. U. Gadgil, K. Gao, Y. Hu and C. Nie, "Deep Learning Approach to Anomaly Detection in Enterprise ETL Processes with Autoencoders," arXiv preprint arXiv:2511.00462, 2025.
13. Q. R. Xu, W. Xu, X. Su, K. Ma, W. Sun and Y. Qin, "Enhancing Systemic Risk Forecasting with Deep Attention Models in Financial Time Series," 2025.
14. Y. Wang, R. Fang, A. Xie, H. Feng and J. Lai, "Dynamic Anomaly Identification in Accounting Transactions via Multi-Head Self-Attention Networks," arXiv preprint arXiv:2511.12122, 2025.
15. Y. Yang, C. Zhang, T. Zhou et al., "Dcdetector: Dual attention contrastive representation learning for time series anomaly detection," Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 3033-3045, 2023.
16. J. Xu, H. Wu, J. Wang et al., "Anomaly transformer: Time series anomaly detection with association discrepancy," arXiv preprint arXiv:2110.02642, 2021.
17. P. Boniol, M. Meftah, E. Remy et al., "dCNN/dCAM: anomaly precursors discovery in multivariate time series with deep convolutional neural networks," Data-Centric Engineering, vol. 4, Article e30, 2023.
18. Y. Sun, N. Zhang, C. Zhang et al., "DBAD: A Dual-Branch Time Series Anomaly Detection Method Based on Transformer and MLP," Proceedings of the 2025 International Conference on Artificial Intelligence and Computational Intelligence, pp. 440-446, 2025.
19. Z. Liu, X. Huang, J. Zhang et al., "Multivariate time-series anomaly detection based on enhancing graph attention networks with topological analysis," Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, pp. 1555-1564, 2024.