Transactions on Computational and Scientific Methods | Vol. 4, No. 1, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Deep Learning-Based Multi-Scale Temporal and Structure-Aware Modeling for Metric Anomaly Detection in Microservice Systems

Yue Kang
Carnegie Mellon University, Pittsburgh, USA
rayen.kangyue@gmail.com

Abstract: This study proposes a metric anomaly detection method that integrates multi-scale temporal modeling with structure-aware feature representation to address the operational demands of microservice architectures in modern cloud-native environments. The approach is designed to tackle challenges such as large system scale, complex dependencies, high-dimensional data, and diverse anomaly patterns. Unlike traditional detection techniques that rely on single time-series signals or static features, the proposed method jointly models inter-service invocation relationships and multidimensional temporal features to construct a unified representation space capable of capturing cross-service contextual dependencies, enabling highly sensitive detection of collaborative anomalies and propagation patterns. The model incorporates a multi-scale time-series embedding module to capture the interaction between long-term trends and short-term fluctuations, while the structural modeling component uses a graph convolution mechanism to represent topological dependencies among services. A contextual fusion layer further integrates temporal information and structural semantics dynamically, generating anomaly representations with global consistency and context awareness. During evaluation, comprehensive analyses are conducted on hyperparameter sensitivity, environmental adaptability, and the impact of data quality, demonstrating the model's stability and robustness in scenarios with missing data, noise interference, and workload fluctuations. Experimental results show that the proposed method outperforms representative detection models across multiple performance metrics, significantly improving alert accuracy, detection efficiency, and global separability, and providing a practical solution for automated monitoring and intelligent operations in microservice systems.

Keywords: Multidimensional indicator modeling; anomaly detection; context fusion; microservice monitoring;deep learning

1. Introduction

Microservice architecture has become a fundamental direction in the evolution of modern software systems. It serves as the backbone for internet services, enterprise platforms, and cloud-native applications. Compared with traditional monolithic architectures, microservices significantly improve system maintainability and scalability through high modularity, independent deployment, and elastic scaling. However, this decentralized design also introduces unprecedented operational complexity. Hundreds of service instances interact asynchronously to form a highly dynamic system ecosystem, where system states are influenced by multidimensional metrics and multi-layer dependencies. In this context, the stability of performance metrics is not only crucial for service quality (QoS) and user experience (QoE) but also directly affects system availability, elastic scheduling, and resource optimization. Therefore, achieving efficient metric anomaly

detection in large-scale, dynamic, and complex microservice environments has become a key research topic in the field of AIOps[1].

Metric monitoring in microservice scenarios differs fundamentally from traditional systems, with its primary challenge arising from the diversity and dynamics of the data. Each service instance continuously generates a large volume of metrics related to performance, resources, networks, and call chains. These metrics are collected in a distributed manner and may exhibit inconsistent sampling frequencies, time delays, and diverse anomaly patterns. In addition, complex service interactions mean that the fluctuation of a single metric may originate from a local bottleneck or cascade through inter-service dependency chains. Such multi-source, multidimensional, and multi-granularity time-series characteristics make traditional statistical or rule-based detection methods insufficient, especially under non-stationary distributions, concept drift, or noise interference. Addressing these challenges requires models capable of capturing contextual dependencies among metrics and identifying potential collaborative anomaly patterns, which has become a central difficulty in anomaly detection model design.

Furthermore, the definition and manifestation of anomalies in microservice systems are highly uncertain and diverse. Traditional anomalies are often characterized by abrupt changes, shifts, or trend drifts. In microservice environments, anomalies may present as single-point events such as CPU spikes or sudden increases in response time[2]. They may also manifest as distribution shifts, behavioral changes, or collaborative anomalies across multiple metrics. These complex anomaly patterns often hide in high-dimensional data spaces and long-term dependencies, making them difficult to detect with simple thresholds or univariate analyses. At the same time, normal business workloads are highly dynamic, which further increases the difficulty of anomaly discrimination. This imposes stricter requirements on the robustness, adaptability, and generalization ability of detection models, driving research toward intelligent approaches that can autonomously learn anomaly semantics from complex backgrounds.

From a system operations perspective, metric anomaly detection is not only the starting point for problem awareness but also the foundation for ensuring service continuity, optimizing resource allocation, and enabling self-healing mechanisms. Timely and accurate detection can trigger early warnings and interventions before anomalies spread, preventing performance degradation, service interruptions, or resource waste[3]. Analyzing historical anomaly patterns also provides valuable decision support for system evolution, such as optimizing scheduling policies, capacity planning, and automatic scaling strategies. With the growing adoption of cloud-native technologies and DevOps practices, automated and intelligent anomaly detection capabilities are becoming essential for continuous delivery and elastic operations. They enhance system observability and controllability and lay the groundwork for future autonomous and self-evolving systems.

More importantly, the widespread adoption of microservices is shifting anomaly detection from a "single-service perspective" to a "system-wide perspective." In complex scenarios involving multi-tenant sharing, heterogeneous resource collaboration, and cross-regional deployment, single-point metric anomalies are no longer sufficient to represent the overall system health. The research focus is shifting toward understanding the intrinsic relationships among metrics and the mechanisms of anomaly propagation to achieve precise perception and root-cause localization at the system level. This trend places higher demands on the semantic representation and structural modeling capabilities of detection models and opens new directions for building intelligent operations systems. By integrating machine learning, deep representation learning, and graph-based modeling, anomaly detection is evolving from data-driven passive perception to knowledge-driven proactive decision-making, providing a solid foundation for the adaptability and intelligence of large-scale systems[4].

2. Related work

With the widespread adoption of microservice architecture and the rise of large-scale distributed systems, metric anomaly detection has gradually become an important branch of intelligent operations research. Early studies mainly focused on traditional statistical methods and rule-based detection strategies. These methods typically identify anomalies in performance metrics through statistical features such as mean, variance, sliding windows, or control charts. They are simple to implement and highly interpretable, and they show acceptable performance in static or low-dynamic systems. However, they face significant limitations in microservice scenarios. First, system metrics often exhibit nonlinear and non-stationary characteristics, and traditional approaches fail to capture complex temporal dependencies. Second, multidimensional and multi-granular data are intertwined in microservices, and anomalies may not manifest as abrupt changes in a single metric but instead hide within collaborative patterns across multiple variables, leading to degraded detection performance. Moreover, manual configuration of thresholds and rules requires extensive domain knowledge and continuous maintenance, making it difficult to adapt to highly dynamic and rapidly evolving environments[5].

With the advancement of machine learning, researchers have begun exploring data-driven anomaly detection models that automatically learn metric distribution characteristics and anomaly patterns through supervised, semi-supervised, or unsupervised approaches. Traditional machine learning models, such as clustering, support vector machines, and isolation forests, can partially mitigate the limitations of threshold settings and improve the recognition of complex anomaly patterns. However, these approaches typically rely on static feature extraction and have limited capability to capture temporal dependencies and contextual changes. In scenarios with high dimensionality, long sequences, and heterogeneous data sources, feature engineering becomes expensive and often leads to information loss. Furthermore, the scarcity of anomaly samples and class imbalance in microservice systems also limit the performance of supervised models, driving research toward unsupervised and self-supervised methods that are more suitable for unlabeled data[6].

In recent years, the introduction of deep learning has significantly advanced the field of metric anomaly detection. Architectures such as recurrent neural networks, convolutional neural networks, and attention mechanisms have shown strong capabilities in modeling time-series data and complex dependencies. These methods can automatically extract deep semantic features from raw metric data, capture both short-term fluctuations and long-term trends, and represent nonlinear patterns in high-dimensional spaces, thereby improving detection accuracy and robustness. At the same time, the application of generative approaches, including variational autoencoders, autoregressive models, and generative adversarial networks, enables more precise modeling of normal behaviors and identification of abnormal deviations[7]. However, these deep models still face challenges in handling concept drift, distribution shifts, and cross-service dependencies. Their training often requires significant computational resources and time, and their "black-box" nature reduces interpretability and auditability in operational scenarios, limiting large-scale deployment in industrial environments.

In the latest research trends, increasing attention is being paid to incorporating microservice-specific structural information and contextual semantics into anomaly detection frameworks. System dependency graphs, call chain relationships, and joint modeling of multidimensional metrics are now being integrated into detection models. Such approaches often leverage graph neural networks, spatiotemporal attention mechanisms, or multimodal fusion strategies to combine metric evolution with structural topology and build context-aware anomaly detection models. By modeling service interaction relationships, systems can more accurately locate root causes of anomalies, identify cascading fault patterns, and achieve global anomaly awareness across services. Meanwhile, adaptive learning, transfer learning, and online update mechanisms have become key research directions to enhance model generalization and long-term applicability in dynamic environments with evolving data distributions. This evolution marks a shift in metric anomaly detection from "point-level perception" to "system-level cognition" and lays the technical foundation for building more intelligent and autonomous operational systems[8].

3. Method

This study proposes a metric anomaly detection approach for microservice architectures, with the core idea of jointly modeling multidimensional performance metrics, contextual dependencies, and structural relationships to accurately identify anomaly patterns in highly dynamic and complexly coupled system environments. The method first performs multi-scale feature representation on raw time-series metrics to capture the evolutionary relationships between short-term fluctuations and long-term trends. It then employs a contextual modeling mechanism to characterize cross-service dependency structures and map the potential semantic relationships among metrics into a unified representation space. Finally, an anomaly scoring function is introduced to quantify the deviation between the predicted distribution and the observed data, enabling anomaly detection and localization. This approach balances representation capability, discriminative power, and scalability, providing a solid technical foundation for intelligent operations in large-scale microservice systems. The model architecture is shown in Figure 1.

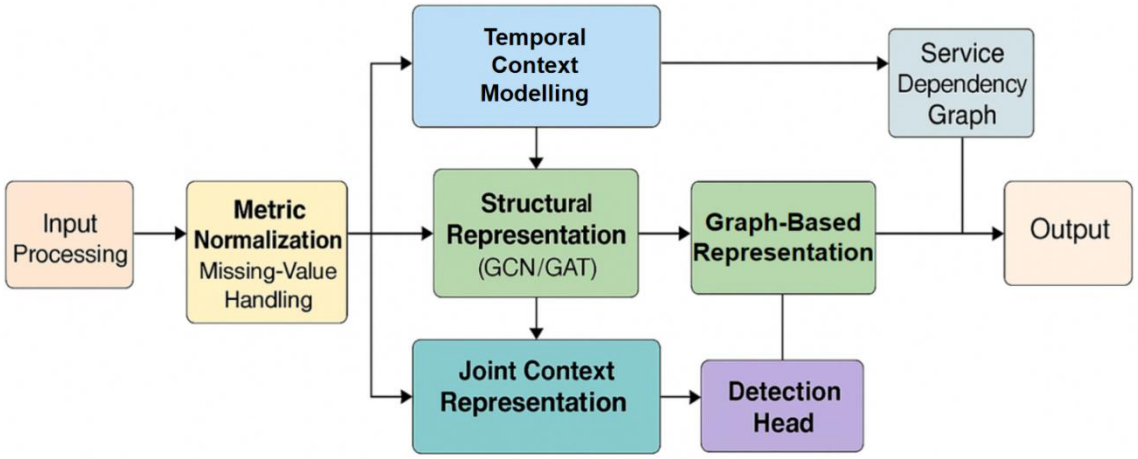


Figure 1. Overall Model Architecture

In the feature modeling phase, the system first represents the input multidimensional time series indicators as a time series signal sequence $\{x_t\}_{t=1}^{T}$ of length T, where the indicator vector at each moment is recorded as $x_t \in R^d$. To capture its dynamic evolution characteristics, a multi-scale time series embedding function is introduced to map the original input into a high-dimensional representation space:

$$h_t = Embed(x_t) = W_e x_t + b_e$$

Among them, $W_e \in R^{d_h \times d}$ and $b_e \in R^{d_h}$ are the learnable weight matrix and bias term, respectively, and $h_t \in R^{d_h}$ represents the moment representation in the embedding space.

To further model dependencies in the time dimension, a context aggregation function is introduced to perform nonlinear fusion of representations at different time steps. By weightedly combining temporal representations through the self-attention mechanism, short-term and long-term dynamic dependencies can be captured:

$$z_t = Attention(H) = \sum_{i=1}^{T} a_{ti} h_i$$

The attention weight a_{ti} is calculated by a mapping function that takes into account the query, key, and value:

$$a_{ti} = \frac{\exp\left(\left(W_q h_t\right)^T \left(W_k h_i\right)\right)}{\sum_{j=1}^{T} \exp\left(\left(W_q h_t\right)^T \left(W_k h_j\right)\right)}$$

Here, W_q and $W_k \in R^{d_h \times d_h}$ are learnable projection matrices used to calculate the similarity weights within the sequence, ensuring that the model can dynamically focus on key time segments and indicator changes.

In the cross-service dependency modeling phase, considering the complex call relationships and structural coupling between microservices, this study considers the system as a directed graph $G = (V, E)$, where V represents the set of service nodes and E represents the set of dependency edges. Based on the graph convolution mechanism, information within the structural neighborhood can be aggregated to obtain a service-level contextual representation:

$$u_v = \sigma\left(\sum_{u \in N(v)} \frac{1}{c_{vu}} W_g h_u\right)$$

Where $N(v)$ is the neighbor set of nodes v, c_{vu} is the normalization coefficient, W_g is the graph convolution weight, and $\sigma(\cdot)$ is the nonlinear activation function. This step can effectively integrate service interaction relationships and provide structured contextual information for anomaly detection.

Finally, to achieve anomaly determination, an anomaly scoring function based on reconstruction error is constructed to evaluate the degree of anomaly by measuring the degree of deviation between the predicted representation $\hat{x}_t$ and the true observation x_t:

$$S_t = \left\| x_t - \hat{x}_t \right\|_2^2$$

When S_t exceeds a preset threshold, the system determines that abnormal behavior exists at that moment. This scoring mechanism not only quantifies single-point anomalies but can also be extended to detect multi-indicator collaborative anomalies, providing an interpretable numerical basis for anomaly identification in complex systems.

Overall, the proposed method constructs a complete detection pipeline through four key steps: multi-scale feature extraction, temporal context modeling, structural dependency awareness, and anomaly quantification. It fully leverages the dynamic, relational, and structural characteristics of metrics in microservice architectures, achieving end-to-end modeling from raw data to anomaly detection. This approach provides a scalable, robust, and semantically aware solution for intelligent monitoring and reliable operations in large-scale systems.

4. Experimental Results

4.1 Dataset

This study uses the Kubernetes_Resource & PerformanceMetricsAllocation dataset as the data foundation for model validation. The dataset records various resource usage and performance metrics from Kubernetes clusters, including CPU utilization, memory usage, network throughput, and I/O read and write rates. It contains multidimensional time-series features collected in a multi-tenant environment that simulates resource competition and workload variations, effectively reproducing the dynamic behavior of resources and performance in microservice systems. This dataset is representative in the research fields of cloud computing performance optimization and anomaly detection.

During data usage, we first divide the dataset into multiple consecutive time windows, with each segment containing resource metric sequences from several services or containers within that period. We then preprocess each time-series sample through operations such as missing value imputation, normalization, and smoothing filters to prepare the data for model input. Sample labels can be constructed or mapped according to the anomaly detection tasks, such as resource spikes, performance degradation, or contention anomalies, to support model training and evaluation.

This dataset aligns well with the objectives of this study in several ways. First, its high-dimensional and multidimensional resource metrics capture the complexity of microservice systems under resource competition and performance fluctuations. Second, because the data originates from a real Kubernetes cluster, it provides a realistic system context and is closer to real-world environments. Finally, the dataset offers a suitable sample size and rich feature dimensions, meeting the evaluation needs for model generalization, robustness, and anomaly discrimination. Experiments based on this dataset effectively validate the applicability and effectiveness of the proposed model in microservice architecture scenarios.

4.2 Experimental Results

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

Table 1: Comparative experimental results

Model	F1	AUROC	Precision	Detection Latency (ms)
MTAD-GAT [9]	0.81	0.94	0.79	62.0
GDN [10]	0.84	0.95	0.82	58.0
TranAD [11]	0.88	0.97	0.86	49.0
Ours	0.91	0.98	0.90	42.0

From the overall trend, classical methods based on graph attention and temporal modeling form a clear performance hierarchy in microservice metric anomaly detection. The F1 scores of MTAD-GAT and GDN are 0.81 and 0.84, respectively. TranAD further improves to 0.88, while the proposed method reaches 0.91. This result indicates that in scenarios with both multidimensional metrics and cross-service dependencies, pure temporal or graph-based modeling still suffers from fragmented information. By unifying them into a joint contextual representation, the model can more effectively capture the collaborative anomaly patterns in both temporal and structural dimensions, thus improving overall discriminative ability (F1) without increasing the false positive rate.

From the perspective of the threshold-independent AUROC metric, the baseline gradually improves from 0.94 (MTAD-GAT) to 0.97 (TranAD), showing that attention mechanisms bring consistent gains in modeling long-range dependencies. Our method achieves an AUROC of 0.98, indicating better separability between normal and abnormal samples across different thresholds. This is highly relevant to the non-stationary loads and concept drift in microservice environments. When anomalies manifest as interval shifts or cross-service propagations rather than single-point spikes, joint contextual modeling provides more evidence for decision boundaries, reducing sensitivity to thresholds and improving robustness across scenarios and time periods.

The comparison of precision and latency further demonstrates the practical engineering value. Precision increases step by step from 0.79 (MTAD-GAT) to 0.86 (TranAD) and finally to 0.90 (ours), indicating that false positives on boundary samples are effectively suppressed in highly dynamic environments. At the same time, detection latency decreases from 62 ms to 42 ms, showing that our method achieves a faster alert loop through optimized feature extraction and detection head design. This "high precision and low latency" combination is particularly important for microservices. When anomalies propagate quickly along

dependency chains, shorter detection latency can significantly reduce the risks of cascading degradation and resource waste.

A comprehensive analysis of all four metrics reveals the key reasons why our method outperforms three representative baselines. Multi-scale temporal embedding captures the combined effects of short-term fluctuations and long-term trends. Structural representation introduced by the dependency graph explicitly aligns causal propagation paths across services. Joint contextual representation mitigates the mismatch between temporal and structural information. The detection head enhances boundary separation for collaborative anomalies in the scoring space. For microservice operations, this translates into more reliable early warning, more stable threshold strategies, and more controllable response windows, providing a stronger signal foundation for elastic scaling, capacity planning, and self-healing mechanisms.

This paper also conducts comparative experiments on the hyperparameter sensitivity of embedding dimension and number of graph convolution layers to joint context representation. The experimental results are shown in Figure 2.

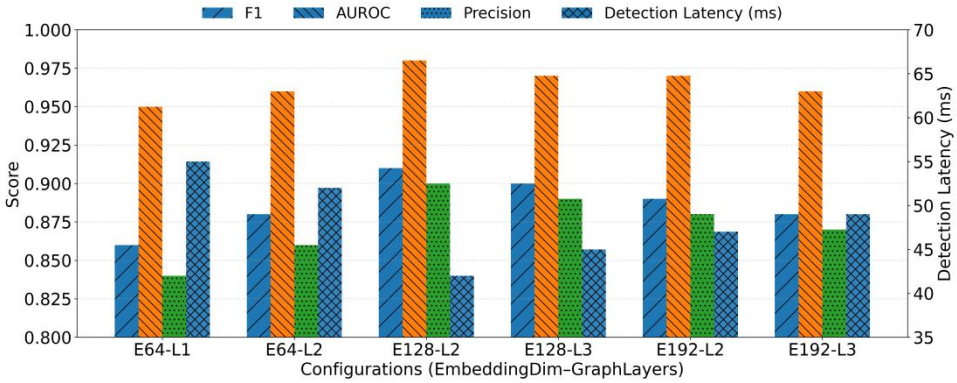


Figure 2. Hyperparameter sensitivity analysis of embedding dimension and number of graph convolution layers on joint context representation

From the overall trend, increasing the embedding dimension from 64 to 128 while maintaining two graph convolution layers (E128-L2) yields the most significant performance gain. Both F1 and Precision improve simultaneously, and AUROC reaches its peak, indicating that joint contextual representation at this capacity best captures both short-term fluctuations and cross-service dependencies. For microservice metrics, this means that the correlations between multidimensional indicators and invocation topology are fully unfolded in a moderately sized embedding space, which enhances the separability of anomalous segments in the representation space and reduces sensitivity to threshold settings.

When the number of graph convolution layers is further increased under the same embedding dimension (E128-L3), classification-related metrics show a slight decline, and detection latency increases from 42 ms to 45 ms. This suggests that overly deep structural propagation introduces feature over-smoothing and additional computational overhead. Given that microservice dependency graphs are often sparse and heterogeneous, excessive neighborhood aggregation can dilute the discriminative power of key dependency edges, smoothing out the "sharp boundaries" of cross-service anomaly propagation and weakening the discriminative strength of the joint representation on boundary samples.

Further increasing the embedding dimension to 192 (E192-L2/L3) does not bring continued benefits. F1, Precision, and AUROC all decrease slightly compared with E128-L2, while latency continues to increase. This indicates that under the constraints of signal-to-noise ratio and sample size in microservice metrics, an excessively large representation space introduces redundant degrees of freedom, causing the model to overfit local fluctuations and noise patterns. Combined with deeper graph layers, this effect intensifies over-smoothing and overfitting, resulting in reduced global separability and inference efficiency.

A cross-metric comparison reveals that AUROC responds more smoothly to capacity changes, reflecting its threshold-independent robustness. Precision and F1 are more sensitive to marginal changes in structural depth and embedding size, directly reflecting the trade-off between false positives and false negatives. Latency increases monotonically with capacity and layer depth, forming a practical engineering constraint. Therefore, in microservice anomaly detection scenarios, it is preferable to adopt a "moderate embedding and shallow layers" configuration, such as E128-L2. This configuration ensures sufficient modeling of cross-service dependencies while keeping the detection loop within a low-latency range, thereby narrowing the propagation window of cascading failures.

This paper also analyzes the hyperparameter sensitivity of the anomaly score threshold and temperature coefficient to the precision-recall trade-off. The experimental results are shown in Figure 3.

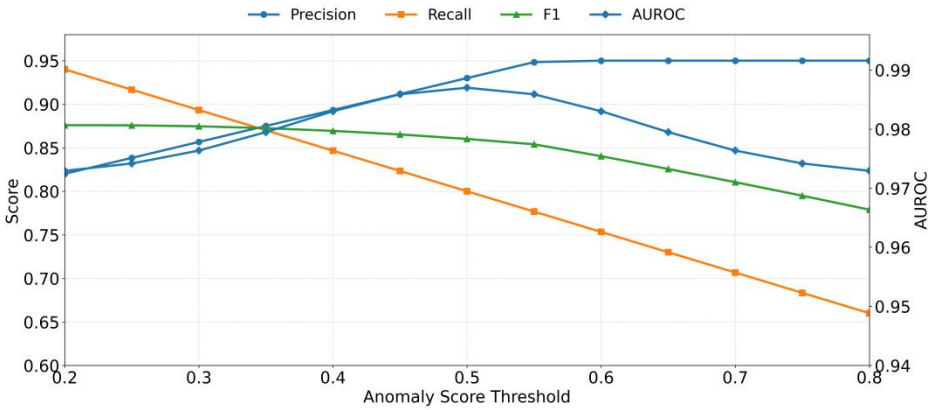


Figure 3. Hyperparameter Sensitivity Evaluation of Anomaly Scoring Threshold and Temperature Coefficient on Precision-Recall Tradeoff

From the overall trend, as the anomaly scoring threshold gradually increases, Precision shows a steady upward trajectory, while Recall continuously decreases, reflecting the inherent trade-off between false positives and false negatives in detection models. When the threshold is low, the model is more sensitive to potential anomalies and can capture more abnormal instances, but this comes at the cost of a higher false positive rate. When the threshold increases, the decision criterion for anomalies becomes stricter, resulting in a clear improvement in Precision. However, some boundary samples are misclassified as normal, leading to a decline in Recall. This asymmetric pattern indicates that the proposed method can adaptively optimize for different detection objectives in microservice environments through threshold adjustment, thus balancing the requirements of alert accuracy and coverage in operational scenarios.

The F1 curve reaches its peak in the middle threshold range, indicating that Precision and Recall achieve an optimal balance around this point, representing the best overall performance of the model. For microservice systems with complex multidimensional metric interactions and diverse anomaly propagation mechanisms, this result is significant. The location of the F1 peak reveals the capability boundary of the joint contextual representation in modeling cross-service dependencies and multidimensional anomaly patterns. It also provides a reference for subsequent automatic threshold search and dynamic alert strategies. Particularly when the system experiences workload fluctuations or concept drift, this optimal threshold point can maintain stable detection performance, reducing the risks of false positive accumulation and false negative propagation.

From the perspective of the AUROC curve, it remains consistently high and shows only minor fluctuations across different threshold ranges, indicating that the model's discriminative power is not sensitive to specific threshold settings. This demonstrates that through multi-scale contextual fusion and structure-aware feature modeling, the model can stably distinguish between normal and abnormal states at a global level, with strong generalization and robustness. In highly dynamic microservice scenarios, such threshold-independent

separability is especially crucial, ensuring that even when anomaly patterns shift or data distributions change, the model maintains strong detection capability.

Further analysis of the impact of the temperature coefficient shows that temperature adjustment slightly affects the variation of Precision and Recall, but does not change the overall trend. A lower temperature coefficient enhances the model's ability to identify high-confidence anomalies, slightly increasing Precision. A higher temperature is more beneficial for recalling sparse or weak-signal anomalies. For microservice metric anomaly detection tasks, this indicates that the model's output distribution is highly tunable. It allows the detection sensitivity and risk tolerance to be dynamically adjusted according to operational policies, providing a more flexible alert mechanism for different business priorities.

Finally, this study evaluated the data sensitivity of the missing rate and noise ratio to the quality of multidimensional indicator representation. The experimental results are shown in Figure 4.

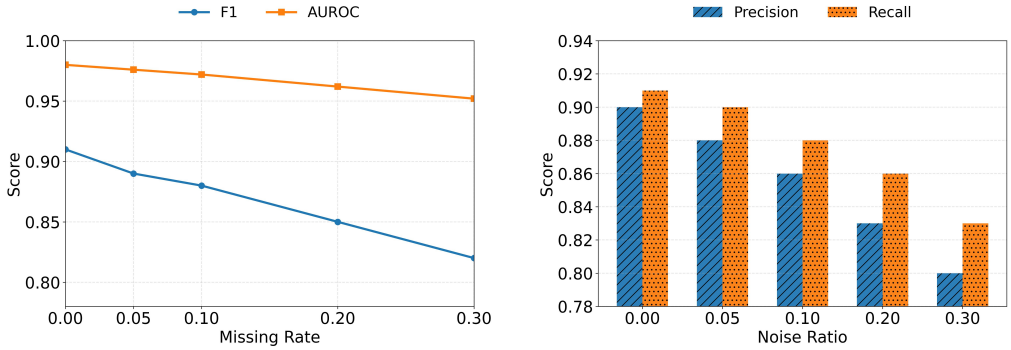


Figure 4. Data sensitivity assessment of missing rate and noise ratio on the quality of multidimensional indicators

As the missing rate increases from 0.00 to 0.30, F1 drops from 0.91 to 0.82, while AUROC decreases more slowly from 0.98 to 0.952, showing a pattern where threshold-dependent metrics decline faster while threshold-independent metrics remain more stable. This indicates that even when data integrity is compromised, the joint contextual representation can still maintain high global separability. However, missing data breaks the continuity of temporal dependencies and structural paths, making boundary samples more susceptible to threshold effects and misclassification. For microservice metrics, missing values weaken the continuous evidence of anomalies along the call chain. The insufficient accumulation of temporal-structural evidence directly manifests as a faster decline in F1.

The nonlinear effect of missing data shows that metric changes remain relatively mild below a 0.10 missing rate, but F1 declines more sharply beyond 0.20, suggesting the existence of a "collapse threshold" for evidence. When gaps between observations grow large enough to span critical dependency edges or key time segments, the trajectory of anomaly propagation is interrupted. This compression of the effective receptive field for temporal attention and graph convolution makes it difficult for the joint representation to reconstruct continuous anomaly chains. This phenomenon aligns with the hierarchical dependencies of microservices. Once missing nodes in the chain exceed the system's redundancy, cross-service collaborative anomaly patterns can no longer form a closed loop, causing the decision boundary to become fragile.

When the noise ratio increases from 0.00 to 0.30, Precision decreases from 0.90 to 0.80, a sharper drop than Recall, which decreases from 0.91 to 0.83. This indicates that label noise primarily undermines the calibration of positive prediction confidence rather than recall capacity. In microservice alerting scenarios, pseudo-labels or observational errors can misrepresent normal fluctuations as anomalous patterns, leading to more false positives in high-confidence prediction segments. The decline in Recall is relatively moderate, suggesting that multi-scale temporal modeling and structural priors still maintain a certain level of coverage. However, in high-noise conditions, the ability to detect weak anomalies gradually deteriorates.

The engineering implications of the "missing-noise" coupling risk are significant. Missing data primarily affects the continuity of the evidence chain, which directly impacts F1, while noise primarily affects confidence calibration, reducing Precision first. The former requires a focus on temporal alignment and graph structure completion, such as neighborhood-consistency-based masked reconstruction and cross-edge imputation. The latter requires a focus on uncertainty scheduling and temperature scaling, such as suppressing low-confidence scores and implementing dynamic threshold fallback. In highly dynamic microservice environments, anomaly detection frameworks based on joint contextual representation should address missing data and noise separately at the data input stage. This ensures stability in evidence continuity and confidence calibration, allowing the system to maintain separability and alert quality under complex operational conditions.

5. Conclusion

This study proposes a joint contextual anomaly detection method that integrates multi-scale temporal modeling with structure-aware representation to address the complexity, dynamics, and high dimensionality of metric anomaly detection in microservice architectures. By jointly modeling temporal dependencies, cross-service relationships, and contextual interactions, the proposed approach effectively overcomes the limitations of traditional detection techniques in capturing collaborative anomalies and complex dependency propagation. It achieves high-precision representation of multidimensional metrics and sensitive detection of anomaly patterns. Experimental results show that the method outperforms representative baseline models across multiple key metrics, validating its reliability and applicability in real-world microservice environments. This design not only improves detection accuracy and alert timeliness but also provides a solid technical foundation for automated operations and intelligent decision-making in large-scale systems.

The proposed research has significant engineering implications. In the context of cloud-native and large-scale distributed systems, complex dependencies, dynamic topologies, and non-stationary data patterns among microservice instances make anomaly detection a critical component of operational systems. The proposed joint representation framework maintains stable and robust detection performance under multi-source, multidimensional, and multi-temporal conditions. It also provides valuable contextual information for root cause localization, elastic scheduling, and self-healing mechanisms. This capability greatly enhances system observability and automation, enabling a shift from passive alerting to proactive perception. It offers a practical solution for building highly available and low-risk cloud service platforms.

From both academic and practical perspectives, this work provides a scalable approach to anomaly detection by deeply integrating temporal behaviors and structural dependencies through a unified contextual representation mechanism. This idea is not only applicable to microservice metric monitoring but can also be extended to other multidimensional and complex scenarios, such as large-scale IoT monitoring, edge system health diagnosis, and cross-domain distributed resource scheduling. More importantly, the proposed framework serves as a design paradigm for future anomaly detection technologies and lays a theoretical foundation for the evolution from data-driven to knowledge-driven intelligent operations.

Future research can be further extended in several directions. On one hand, causal inference and reinforcement learning techniques can be incorporated to proactively model anomaly propagation chains and system behavior evolution, enhancing the system's adaptability to unknown patterns. On the other hand, integration with federated learning and privacy-preserving computation can be explored to enable secure and efficient operation in multi-tenant and collaborative environments. Furthermore, with the advancement of generative AI and large-scale models, future anomaly detection frameworks are expected to possess autonomous interpretation and policy generation capabilities, providing end-to-end intelligent support from detection to intervention and driving cloud-native operations toward a more autonomous and intelligent era.

References

[1] A. Ikram, S. Chakraborty, S. Mitra, S. Saini, S. Bagchi and M. Kocaoglu, "Root cause analysis of failures in microservices through causal discovery," Advances in Neural Information Processing Systems, vol. 35, pp. 31158-31170, 2022.

[2] A. Chevrot, A. Vernotte and B. Legeard, "CAE: Contextual auto-encoder for multivariate time-series anomaly detection in air transportation," Computers & Security, vol. 116, p. 102652, 2022.

[3] Nobre J, Pires E J S, Reis A. Anomaly detection in microservice-based systems[J]. Applied Sciences, 2023, 13(13): 7891.

[4] Chen N, Tu H, Zeng H, et al. Anomaly detection for key performance indicators by fusing self-supervised spatio-temporal graph attention networks[J]. Knowledge-Based Systems, 2024, 300: 112167.

[5] Y. Wang, J. Peng, J. Zhang, R. Yi, Y. Wang and C. Wang, "Multimodal industrial anomaly detection via hybrid fusion," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8032-8041, 2023.

[6] C. Ding, S. Sun and J. Zhao, "MST-GAT: A multimodal spatial−temporal graph attention network for time series anomaly detection," Information Fusion, vol. 89, pp. 527-536, 2023.

[7] Wang P, Zhang X, Cao Z, et al. MADMM: microservice system anomaly detection via multi-modal data and multi-feature extraction[J]. Neural Computing and Applications, 2024, 36(25): 15739-15757.

[8] Liu X, Liu Y, Wei M, et al. LMGD: Log-Metric Combined Microservice Anomaly Detection Through Graph-Based Deep Learning[J]. IEEE Access, 2024.

[9] Zhao H, Wang Y, Duan J, et al. Multivariate time-series anomaly detection via graph attention network[C]//2020 IEEE international conference on data mining (ICDM). IEEE, 2020: 841-850.

[10] Deng A, Hooi B. Graph neural network-based anomaly detection in multivariate time series[C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(5): 4027-4035.

[11] Tuli S, Casale G, Jennings N R. Tranad: Deep transformer networks for anomaly detection in multivariate time series data[J]. arXiv preprint arXiv:2201.07284, 2022.