Anomaly Perception and Early Fault Prediction in Cloud Services via Graph-Structured Temporal Representation Learning

Cancan Hua *

Posted Date: 3 March 2026

doi: 10.20944/preprints202603.0253.v1

Keywords: cloud service anomaly detection; graph time series modeling; fault prediction; intelligent operation and maintenance

 doi:10.20944/preprints202603.0253.v1

Article

Anomaly Perception and Early Fault Prediction in Cloud Services via Graph-Structured Temporal Representation Learning

Cancan Hua

University of Southern California, Los Angeles, USA; cancanh@alumni.usc.edu

Abstract

This study proposes an anomaly perception and prediction model that integrates graph structure and temporal dependence to address the complexity of anomaly detection and fault prediction in cloud service systems. The model first constructs a service dependency graph to capture the structural topology of the system and then employs a temporal encoding module to model the dynamic evolution of service metrics. In the graph-temporal fusion layer, structural and temporal features are deeply integrated to enhance the model's ability to capture anomaly propagation paths and temporal patterns. Using multidimensional monitoring data as input, the model achieves a global representation of cloud service states through spatial aggregation and temporal dependency modeling, leading to stable performance in anomaly detection and root cause localization. Experimental results show that the proposed model significantly outperforms traditional methods in accuracy, recall, and F1-Score, maintaining robustness and generalization under complex topologies and highly dynamic environments. Further experiments on hyperparameter, environment, and data sensitivity verify the model's adaptability to key factors such as learning rate, communication delay, data missing rate, and anomaly ratio. This method provides an interpretable and structured solution for intelligent cloud operations and offers transferable theoretical and technical support for anomaly detection and risk awareness in large-scale distributed systems.

Keywords: cloud service anomaly detection; graph time series modeling; fault prediction; intelligent operation and maintenance

1. Introduction

The rapid development of cloud computing has driven deep integration across information technology infrastructures, while also making service systems increasingly complex and dynamic. A cloud platform typically consists of a large number of virtualized components, containerized instances, and heterogeneous network nodes, all of which are highly interdependent and interactive [1]. As the scale expands and concurrent workloads increase, the stability and reliability of cloud services face unprecedented challenges. Even minor performance fluctuations, configuration drifts, or dependency anomalies can quickly trigger cascading effects, leading to service degradation or large-scale outages. Therefore, how to achieve early perception and dynamic prediction of potential failures within complex service topologies has become a core issue in ensuring cloud platform stability and service quality. Traditional monitoring mechanisms rely on static thresholds and rule matching, which are difficult to adapt to dynamic and time-varying environments. Moreover, single-metric analysis cannot fully capture the multidimensional coupling relationships that reflect the overall system state [2].

In the multilayer architecture of cloud services, dependencies among components are often implicitly embedded in graph structures. Whether in service call chains, container deployment topologies, or communication networks among microservices, such relationships exhibit the characteristics of complex directed graphs. The anomaly of one node is influenced not only by its own condition but also by propagation through dependency paths, forming hidden cascading failures. This structural dependency limits traditional time-series-based independent analysis

 doi:10.20944/preprints202603.0253.v1

methods, which often overlook upstream and downstream propagation effects as well as cross-layer associations. Meanwhile, various time-series indicators in cloud environments, such as CPU utilization, memory usage, network latency, and I/O throughput, display strong nonlinearity and multi-scale variation. Short-term spikes coexist with long-term drifts, showing evident non-stationarity and complex coupling. A purely static graph model fails to capture such dynamic evolution, while a single-dimensional temporal model cannot express structural dependencies. Hence, a comprehensive framework that integrates both graph structure and temporal dependence is urgently needed [3].

The concept of integrating graph structure and temporal dependence for cloud service anomaly monitoring aims to capture spatial dependencies among multiple nodes through graph representation learning, while employing temporal modeling to achieve continuous perception of dynamic states. This paradigm simultaneously describes "who influences whom" in structure and "when changes occur" in time, creating a unified semantic space for multilevel and fine-grained anomaly cognition. Based on joint modeling of topological constraints and temporal evolution, it becomes possible to identify potential fault propagation paths, reveal the generation mechanisms of anomalies, and shift from mere "anomaly detection" to "anomaly understanding." This approach holds great significance for building interpretable and transferable intelligent operation and maintenance systems, and provides theoretical support for automated root cause analysis, performance optimization, and self-healing mechanisms in AIOps [4].

From an industrial perspective, cloud computing has become the backbone of digital infrastructure. Fields such as financial transactions, online education, intelligent manufacturing, e-government, and healthcare all rely on highly available and reliable cloud services. When systems experience latency, stalling, or downtime, the consequences go beyond economic loss to affect public trust and societal functioning. With the widespread adoption of multi-cloud and hybrid-cloud architectures, cross-platform deployment, dynamic migration, and elastic scaling have become the norm, making monitoring and prediction even more complex. The modeling approach that combines graph structures with temporal dependencies enables unified anomaly perception in large-scale, heterogeneous, and dynamic environments. It supports end-to-end performance supervision and risk warning. Such research not only has strong theoretical value but also directly relates to the resilience, security, and sustainability of cloud infrastructures.

At the academic level, this direction represents a key evolution in intelligent operation and maintenance, moving from "static detection" toward "structured intelligence." It breaks through the limitations of single-metric monitoring and isolated models by unifying structural properties, temporal dynamics, and semantic dependencies of complex systems within a single modeling space. By incorporating graph structures, the contextual expressiveness of temporal modeling is enhanced, allowing the model to capture latent causal relationships between anomalies more precisely. At the same time, temporal dependence endows structural learning with dynamic adaptability, reflecting how topology changes and workload fluctuations affect overall system health. This integrated paradigm can also be extended to distributed computing, network security, and energy scheduling. It offers new perspectives and methodological foundations for intelligent management of cloud services. Therefore, research on cloud service fault prediction and anomaly perception that integrates graph structure and temporal dependence is of profound importance for advancing the observability, autonomy, and intelligence of cloud computing systems.

2. Related Work

Research on anomaly detection and fault prediction in cloud services has evolved through multiple stages, from traditional statistical modeling to intelligent learning-based approaches. Early studies mainly relied on threshold- or rule-based monitoring methods, where upper and lower limits of performance metrics were manually set to identify abnormal behaviors. These methods are simple to implement and easy to interpret. However, they often suffer from high false alarm and missed detection rates when facing the dynamic variability and high dimensionality of cloud environments. Later, machine learning-based anomaly detection methods emerged, using clustering, classification, and regression models to automatically learn the relationships among metrics, thereby improving

 doi:10.20944/preprints202603.0253.v1

robustness and adaptability. Yet, these models usually assume that data are independent and identically distributed, failing to capture the complex dependency structures among cloud services. As the system scales to thousands of nodes, limitations in feature engineering, computational complexity, and generalization of traditional machine learning models become apparent, making it difficult to meet the requirements of efficient and real-time operation [5].

With the advancement of deep learning, researchers began to employ neural networks to model the multidimensional temporal characteristics of cloud services. Recurrent neural networks, convolutional neural networks, and their variants have achieved significant progress in anomaly detection. They can automatically extract latent patterns from raw monitoring sequences and capture both short-term and long-term temporal dependencies [6]. Some studies further introduced attention mechanisms to emphasize important time points or critical indicators, enhancing the model's sensitivity to anomalies. However, these methods share a common limitation. They primarily focus on the temporal evolution of individual nodes or single metrics while ignoring the topological structure and cross-node dependencies within the system. Cloud service failures are rarely isolated events but are often influenced by anomalies propagated from upstream services. Time series models that lack structural information are unable to identify such cascading effects, leading to prediction results that lack global consistency.

To address these challenges, recent research has introduced graph-based anomaly detection methods that incorporate structural information. By constructing service call graphs, dependency graphs, or communication topologies, these approaches employ graph neural networks to learn relationships and contextual features among nodes, thereby achieving joint modeling of global states. Graph neural networks aggregate features from neighboring nodes during message passing, enabling high-level abstraction of complex dependencies. Such methods have significantly improved spatial awareness in anomaly detection and can identify potential propagation risks in the early stages of system degradation. However, most existing studies still treat structure and time as independent dimensions. Graph models focus on static topological correlations, while temporal models emphasize sequence dynamics. The lack of collaborative modeling makes it difficult to describe the joint spatiotemporal evolution of cloud services. In dynamic topologies and multi-tenant environments, dependencies among nodes also change over time, making static graph-based methods unable to accurately reflect the real operating state [7].

Recent studies have begun to explore integrated modeling that combines graph structures with temporal dependencies. By combining graph neural networks and temporal networks, researchers have proposed spatiotemporal frameworks for anomaly detection. The core idea is to learn both structural dependencies among nodes and the temporal evolution of features, enabling cross-dimensional interaction through dynamic adjacency matrices, adaptive attention, or graph convolutional temporal units. These models can identify local anomalies and infer propagation paths and potential root causes, demonstrating stronger interpretability and predictive power. However, several challenges remain. First, topology update strategies in dynamic environments are often coarse and cannot handle frequent service migrations and reconfigurations in cloud systems. Second, scalability and interpretability still need improvement. Third, there is no unified standard for fusing multimodal monitoring data such as logs, metrics, and traces. Therefore, building an adaptive anomaly perception framework that captures evolving topologies while integrating temporal dynamics and structural dependencies has become a critical direction and future breakthrough in intelligent cloud service operations.

3. Proposed Framework

3.1. General Introduction

This paper proposes a cloud service fault prediction and anomaly detection model that integrates graph structure and temporal dependencies. The core idea is to simultaneously model the spatial relationships and multi-dimensional temporal dynamic changes of service topology within a unified framework. The model consists of three key parts: a service dependency graph construction module, a temporal feature encoding module, and a graph-temporal joint perception module. First, a service dependency graph is constructed using cloud platform monitoring logs and call chain

 doi:10.20944/preprints202603.0253.v1

information, with nodes representing services and edges representing call or resource dependencies, thereby capturing potential propagation paths. Second, a temporal coding network is used to embed the indicator sequences of each service node to extract temporal evolution features. Finally, temporal information is embedded into the graph neural structure to achieve cross-node feature aggregation and global state modeling. The entire system structurally implements a "space aggregation first, time evolution later" modeling approach, enabling the model to simultaneously identify local anomalies and global evolution trends in a dynamic environment. Its overall model architecture is shown in Figure 1.

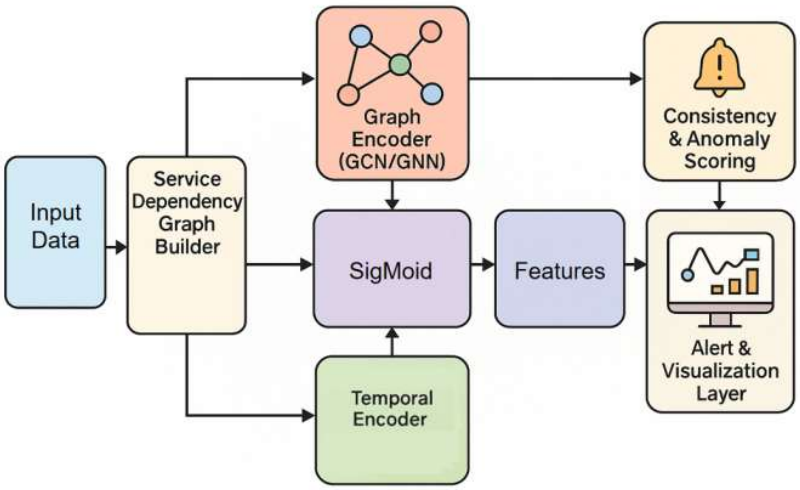

Figure 1. Overall model architecture.

3.2. Service Dependency Graph Modeling

In a cloud environment, suppose the cloud platform contains N service nodes, and let $G = (V, E)$ represent the service dependency graph, where $V = \{v_1, v_2, \ldots, v_N\}$ is the set of nodes and $E \subseteq V \times V$ is the set of dependencies. The time-state t of each node is represented by a monitoring feature vector $x_i^t \in R^d$. To capture dependency strength, an adjacency matrix $A \in R^{N \times N}$ is constructed, whose elements are defined as:

$$A_{ij} = \begin{cases} exp(-||f_i - f_j||_2^2/\sigma^2) \\ 0 \end{cases} \quad (1)$$

Where f_i and f_j are the structural feature vectors of the service nodes, and σ is the smoothing factor. To further enhance the structural representation capability, a spatial feature propagation mechanism based on graph convolution is introduced:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \quad (2)$$

Where $\tilde{A} = A + I$ is its degree matrix, $\tilde{D}$ is the nonlinear activation function, and $\sigma(\cdot)$ is the weight matrix of the l-th layer. This layer realizes the joint update of node and neighbor information, thereby capturing the potential anomaly propagation characteristics in the topology.

3.3. Time-Dependent Modeling

After capturing spatial dependencies, it is necessary to model the temporal dynamics of each service node. Let the input time series of node i be $\{x_i^1, x_i^2, \ldots, x_i^T\}$, and a time gating mechanism be used to remember long-term dependencies and respond to short-term mutations. The time encoding can be defined as:

$$h_i^t = GRU(x_i^t, h_i^{t-1}) \quad (3)$$

Where h_i^t represents the hidden state at time step t. To further model the weight relationships at different time scales, a self-attention mechanism is introduced to calculate the time weights:

$$\alpha_i^t = \frac{exp(q^T ta \quad (W_h h_i^t))}{\sum_{k=1}^T exp(q^T tanh(W_h h_i^k))} \quad (4)$$

And obtain a weighted temporal context representation:

 doi:10.20944/preprints202603.0253.v1

$$z_i = \sum_{t=1}^{T} \alpha_i^t h_i^t \quad (5)$$

This mechanism can adaptively focus on key time segments where anomalies occur, enabling fine-grained modeling of temporal features.

3.4. Graph-Time Joint Sensing and Anomaly Scoring Calculation

In the fusion layer, spatial structure representation and temporal feature representation are fused across modalities to obtain the comprehensive state embedding of nodes:

$$s_i = \phi(H_i^{(L)} || z_i) \quad (6)$$

Where $\phi(\cdot)$ is the fusion function and $||$ represents the concatenation operation. To achieve anomaly detection and fault prediction, a joint estimation mechanism based on reconstruction error is introduced:

$$x_i = \psi(s_i), L_{rec} = \frac{1}{N}\sum_{i=1}^{N} ||x_i - x_i||_2^2 \quad (7)$$

At the same time, the anomaly scoring function for nodes is defined as follows:

$$\Omega_i = \lambda_1 ||x_i - x_i||_2 + \lambda_2 (1 - cos(H_i^{(L)}, z_i)) \quad (8)$$

The first term reflects the reconstruction error, and the second term represents the consistency deviation between the spatial and temporal representations. Through this joint mechanism, the model can adaptively identify structural anomalies, temporal mutations, and their propagation chains, thereby enabling early detection and prediction of potential faults in complex cloud service systems.

4. Experimental Analysis

4.1. Dataset

This study uses the Microsoft Azure Telemetry Dataset as the primary data source. The dataset consists of multidimensional operational logs and performance monitoring metrics collected from a cloud service platform. It covers various components such as virtual machines, container services, load balancers, and database instances. The data include key monitoring indicators such as CPU utilization, memory usage, disk I/O, network latency, throughput, and request error rate. These metrics are continuously sampled at one-minute intervals, allowing a detailed observation of the dynamic behavior of the cloud platform. Unlike traditional static logs, this dataset ensures temporal synchronization across different components through a unified timestamp alignment mechanism, providing a reliable foundation for modeling cross-node dependencies and analyzing anomaly propagation.

Structurally, the dataset also contains service call topologies and node dependency relationships, which are used to construct dynamic service dependency graphs from call-chain logs. This graph structure reveals the interaction patterns and dependency directions among service modules, helping the model capture anomaly propagation paths in the spatial dimension. For example, when an upstream database or cache node experiences performance bottlenecks, downstream services may show request timeouts or response delays. Traditional single-node monitoring cannot directly reflect such cascading effects. By incorporating graph structural information, the model can learn both local node features and global topological correlations during training, thus supporting multilevel anomaly detection and fault prediction.

The dataset is characterized by high dimensionality, long temporal span, and multi-source heterogeneity. It can be widely applied to cloud operations, AIOps, and anomaly diagnosis tasks. It contains millions of monitoring records and thousands of node entities, with rich annotations of anomaly events such as service interruptions, performance degradation, network congestion, and resource contention. Due to its authenticity and complexity, this dataset effectively validates the robustness and generalization ability of models under high-noise conditions. It provides a solid data foundation and experimental basis for subsequent joint modeling of graph and temporal dependencies.

 doi:10.20944/preprints202603.0253.v1

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.

Method	Acc	Precision	Recall	F1-Score
MLP [8]	0.8421	0.8165	0.7894	0.8027
BILSTM [9]	0.8643	0.8312	0.8236	0.8273
1DCNN [10]	0.8715	0.8428	0.8351	0.8389
Transformer [11]	0.8862	0.8563	0.8487	0.8525
GNN [12]	0.8947	0.8639	0.8571	0.8605
GAT [13]	0.9028	0.8716	0.8642	0.8678
Ours	0.9215	0.8927	0.8834	0.8879

From the overall results, the proposed cloud service fault prediction model that integrates graph structure and temporal dependence outperforms all comparison methods across multiple metrics, especially in accuracy and F1-Score. The traditional MLP model lacks temporal modeling ability and structural dependency awareness, and can only learn linear relationships among static features. Therefore, its detection capability is limited in complex and dynamic environments. As network architectures evolve, BiLSTM and 1D-CNN introduce temporal modeling and local feature extraction mechanisms, which enhance the model's sensitivity to temporal patterns of anomalies. Their performance shows significant improvement over baseline methods, demonstrating the importance of temporal dependence modeling for recognizing abnormal behaviors in cloud platforms.

The Transformer model further validates the effectiveness of global dependency modeling. Through the self-attention mechanism, it captures long-range dependencies in the temporal dimension, allowing stable detection under complex workload fluctuations and multi-scale variations. However, this approach still focuses mainly on single-node sequential features and does not fully model the topological relationships and fault propagation paths between different service nodes. As a result, when cross-node anomaly propagation occurs, time-only modeling struggles to achieve high precision, and the model's recall and F1 remain limited.

The introduction of graph neural structures, such as GNN and GAT, leads to further improvements across all metrics, indicating the significant value of structural information for anomaly detection. GNN enables feature transmission and aggregation among nodes, capturing potential dependencies and interactions between services. GAT adaptively adjusts neighbor weights through an attention mechanism, giving the model stronger representational power in heterogeneous topologies. This structural awareness in the spatial dimension allows the model to identify anomaly propagation chains, enhancing system-level detection capability and robustness. It highlights the adaptability and generalization of graph structures in cloud service scenarios.

Finally, the proposed model integrating graph structure and temporal dependence achieves the best performance across all metrics, with an accuracy of 0.9215 and an F1-Score of 0.8879. These results show that joint spatiotemporal modeling effectively combines inter-node dependencies with temporal evolution features, overcoming the limitations of one-dimensional modeling. The model can identify both local transient anomalies and cross-node propagation anomalies. In complex cloud service environments, this fusion modeling strategy not only improves fault prediction accuracy but also enhances stability and interpretability under dynamic topologies. It provides a more robust solution for intelligent operation and anomaly prediction in cloud platforms.

This paper also presents an experiment on the effect of learning rate on F1-Score, and the experimental results are shown in Figure 2.

 doi:10.20944/preprints202603.0253.v1

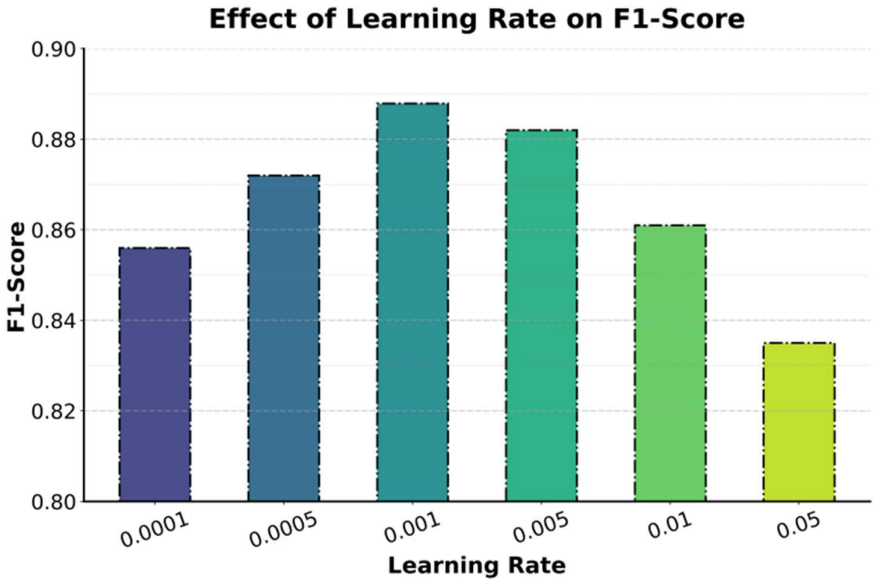

Figure 2. Experiment on the effect of learning rate on F1-Score.

The experimental results show that the learning rate has a significant impact on the model's F1-Score, presenting a typical pattern of "optimal in the middle and declining at both ends." When the learning rate is low (for example, 0.0001 or 0.0005), the parameter update speed is slow, leading to a gradual convergence process. The model struggles to fully capture the complex correlations among cloud service anomaly features, resulting in a relatively low F1-Score. Although the model is stable under this condition, its response to sudden anomalies is not sensitive enough, making it difficult to identify rapidly evolving service faults effectively.

When the learning rate increases to 0.001, the model achieves its best performance with an F1-Score of about 0.888. This indicates that the parameter update speed and gradient variation are well balanced at this stage. The graph structural information and temporal dependency features are effectively integrated, enabling precise anomaly perception and fault prediction in complex cloud service topologies. The model shows optimal gradient convergence, can quickly capture system state changes, and avoids excessive oscillations. This demonstrates the stable learning characteristics of spatiotemporal joint modeling under dynamic cloud environments.

However, when the learning rate continues to increase (for example, from 0.005 to 0.05), the model performance begins to decline. An excessively high learning rate causes violent gradient updates, which may overshoot the optimal solution region and lead to unstable oscillations in parameter space. This instability is more apparent in cloud environments characterized by high noise and non-stationarity. Such irregular updates disrupt the continuity of temporal features and the consistency of topological constraints, resulting in blurred anomaly detection boundaries. Therefore, choosing an appropriate learning rate is essential to ensure stable convergence and high-accuracy prediction in complex and dynamic topologies.

This paper also presents an experiment on the impact of time window length on prediction recall performance, and the experimental results are shown in Figure 3.

 doi:10.20944/preprints202603.0253.v1

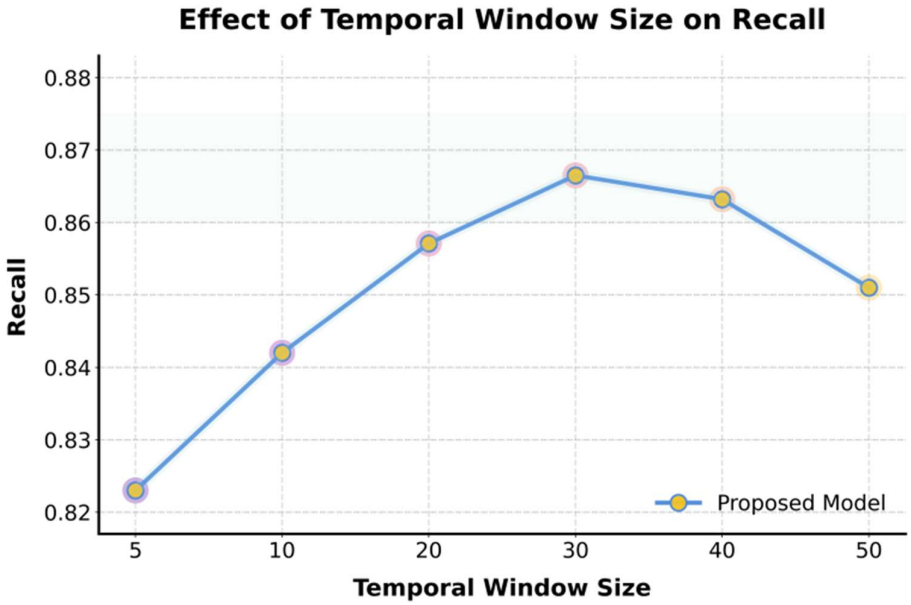

Figure 3. Experiment on the impact of time window length on prediction recall performance.

The experimental results show that the length of the time window has a clear nonlinear effect on the model's recall performance. When the time window is short (for example, 5 or 10), the model has limited ability to capture temporal variation patterns, leading to incomplete recognition of anomalies. A short window cannot fully utilize the historical information of service states and fails to extract long-term dependency features of anomaly evolution. As a result, the model's perception range for sudden faults is restricted, which is reflected in a lower recall value.

As the time window increases to 20 or 30, the model's recall improves significantly and reaches its peak at a window length of 30. This indicates that within this range, the model achieves a good balance between short-term dynamics and long-term trend modeling, effectively capturing cross-temporal dependencies in cloud service systems. A longer temporal context helps the model understand the periodic variations and latent delay effects of service performance, making anomaly detection more comprehensive and stable. These findings confirm the critical role of temporal features in dynamic cloud environments and highlight the advantage of the proposed method in modeling long-term dependencies.

When the time window continues to increase (for example, 40 or 50), the recall slightly decreases. This suggests that overly long sequences introduce redundant or outdated information, which increases noise interference and computational complexity. In such cases, the model's attention may become dispersed, reducing its focus on key anomalous segments and weakening detection sensitivity. Therefore, an appropriate time window design is crucial for cloud service fault prediction. It should cover sufficient temporal information to reflect evolutionary patterns while avoiding the performance degradation caused by redundancy and dynamic drift.

This paper also presents the impact of the missing data ratio on the experimental results of the model, and the experimental results are shown in Figure 4.

 doi:10.20944/preprints202603.0253.v1

Figure 4. The impact of missing data ratio on model experiment results.

The experimental results show that as the data missing rate increases, the model exhibits a gradual decline across all evaluation metrics, with the most significant drops observed in accuracy and F1-Score. When the missing rate is low (0%–10%), the model still maintains a high level of performance. This indicates that the proposed graph-temporal fusion structure has certain robustness and fault-tolerance capabilities. At this stage, the model alleviates the information loss caused by partial data absence through graph structure modeling and temporal context compensation, ensuring stable and reliable predictions under complex cloud service environments.

When the missing rate reaches 20% to 30%, the model performance decreases significantly. Missing samples lead to incomplete node feature representations and disrupt the continuity of temporal dependencies. This interference affects the model's ability to identify anomaly propagation paths and capture dynamic features. Feature loss also causes insufficient information transmission within the graph structure, leading to imbalanced weight updates for some critical nodes and a noticeable decline in recall. In multi-node dependency scenarios of cloud services, missing data amplify local errors and propagate them through the topology, resulting in cumulative effects on global perception.

When the missing rate exceeds 40%, the model's predictive performance becomes nearly unstable, indicating that large-scale data loss surpasses the model's adaptive compensation capacity. Although the model retains some degree of fault tolerance, the spatiotemporal fusion module fails to maintain effective representations under highly sparse inputs, reducing the discriminative power of anomaly features. These results demonstrate that data completeness is crucial for cloud service anomaly perception. The proposed method, however, can still maintain high detection accuracy under low to moderate missing rates, showing its adaptability and robustness to common data missing issues in real-world cloud platforms.

This paper also presents the impact of the anomaly ratio on the experimental results of the model, and the experimental results are shown in Figure 5.

 doi:10.20944/preprints202603.0253.v1

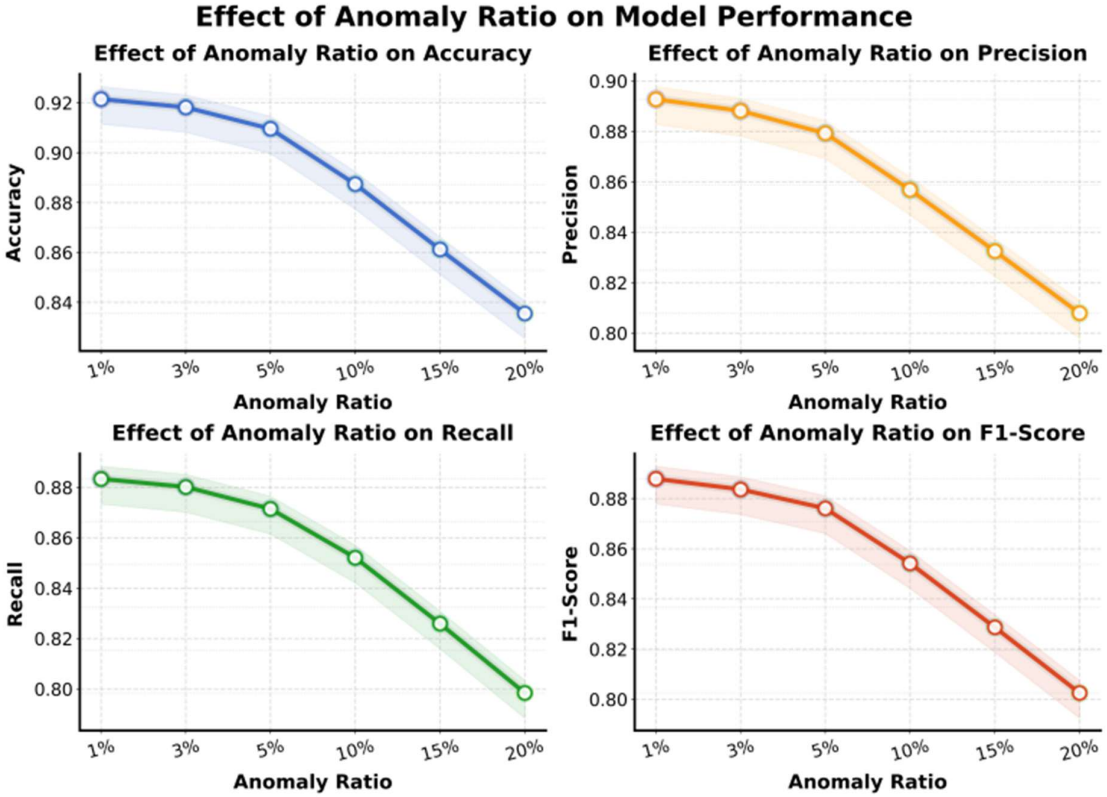

Figure 5. The impact of the anomaly ratio on model experiment results.

The experimental results show that as the proportion of anomalous samples increases, the model exhibits a gradual decline in performance across all metrics, with the most significant decreases observed in accuracy and F1-Score. When the anomaly proportion is low (1%–5%), the model maintains stable overall performance. This indicates that the proposed graph-temporal fusion structure can effectively recognize anomalies when the data distribution is balanced. At this stage, the model fully utilizes structural dependencies and temporal correlations to detect subtle fluctuations and potential risks in cloud service systems, demonstrating robustness and sensitivity in low-anomaly environments.

When the anomaly proportion rises to around 10%, the model performance begins to decline noticeably. The main reason is that the increased proportion of anomalous samples in the training set causes a shift in the decision boundary, which interferes with the learning of normal patterns. Since anomalies in cloud services often exhibit nonlinear propagation and complex cross-node dependencies, a higher anomaly proportion leads to feature noise accumulation when aggregating graph structural information, weakening the temporal consistency of the model. In addition, some anomaly patterns overlap semantically, further increasing classification difficulty and causing simultaneous drops in recall and precision.

When the anomaly proportion exceeds 15%, the model performance deteriorates rapidly. This indicates that in high-anomaly scenarios, the extreme distribution shift severely affects the model's discriminative capability. The local smoothness assumption of the graph structure is no longer valid, resulting in inaccurate feature propagation among nodes and reduced temporal prediction ability. These results suggest that although the proposed method demonstrates strong structural modeling and temporal learning capabilities, additional strategies such as data resampling or anomaly-weight balancing are needed in high-anomaly environments. Such strategies can help maintain model robustness and generalization performance, enabling better adaptation to the imbalanced data distributions commonly found in complex cloud platforms.

 doi:10.20944/preprints202603.0253.v1

5. Conclusions

This paper addresses the challenge of accurately identifying complex topological dependencies, dynamic temporal features, and anomaly propagation in cloud service systems. A fault prediction and anomaly perception model that integrates graph structure and temporal dependence is proposed. By introducing structural awareness and temporal modeling within a unified framework, the model achieves comprehensive representation from local feature recognition to global anomaly propagation. It effectively captures the dynamic interaction patterns and evolution processes among service nodes. Experimental results verify that the proposed method outperforms traditional models across multiple metrics, demonstrating stronger robustness and generalization under diverse anomaly conditions in complex cloud environments. This provides a new perspective for developing reliable and interpretable intelligent operation and maintenance systems.

The significance of this research lies not only in performance improvement but also in offering a structured intelligent analysis framework for proactive monitoring and risk management in cloud systems. Through graph-based modeling, the system can automatically uncover dependency patterns among services and assist engineers in understanding causal chains of anomalies. Through temporal modeling, the approach captures the evolving trends of performance metrics and enables early warning of potential failures. This method can lay an intelligent foundation for the AIOps framework of cloud platforms, equipping systems with self-perception, self-diagnosis, and self-repair capabilities. It can significantly reduce failure recovery time and maintenance costs. Moreover, the theoretical framework is also applicable to other distributed and time-varying systems such as industrial IoT, edge computing, and financial risk control, showing strong transferability and practical potential.

Future work can focus on three directions. First, adaptive graph updating and continual learning mechanisms can be introduced under dynamic topologies to enable the model to adjust to changing service structures in real time. Second, multimodal data fusion, including logs, metrics, and call traces, can enhance the precision and interpretability of anomaly recognition. Third, integrating causal inference and reinforcement learning can help build a decision-capable anomaly response system, achieving an evolution from anomaly detection to proactive intervention. Continuous exploration of these directions will promote the development of higher-level autonomous intelligence in cloud systems and provide theoretical and practical support for the stable operation and risk defense of large-scale distributed services in the future.

References

1. Huang J, Yang Y, Yu H, et al. Twin graph-based anomaly detection via attentive multi-modal learning for microservice system[C]//2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE). IEEE, 2023: 66-78.
2. Li J, Pang G, Chen L, et al. HRGCN: Heterogeneous graph-level anomaly detection with hierarchical relation-augmented graph neural networks[C]//2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2023: 1-10.
3. Zhang Z, Zhu Z, Xu C, et al. Towards accurate anomaly detection for cloud system via graph-enhanced contrastive learning[J]. Complex & Intelligent Systems, 2025, 11(1): 23.
4. Yu M, Zhang X. Anomaly Detection for Cloud Systems with Dynamic Spatiotemporal Learning[J]. Intelligent Automation & Soft Computing, 2023, 37(2).
5. J. Yang, J. Chen, Z. Huang, C. Xu, C. Zhang and S. Li, "Cost-TrustFL: Cost-Aware Hierarchical Federated Learning with Lightweight Reputation Evaluation across Multi-Cloud," arXiv preprint arXiv:2512.20218, 2025.
6. Zhao Z, Xiao Z, Tao J. MSDG: Multi-scale dynamic graph neural network for industrial time series anomaly detection[J]. Sensors, 2024, 24(22): 7218.
7. Jin M, Koh H Y, Wen Q, et al. A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

 doi:10.20944/preprints202603.0253.v1

8. Y. Ni, X. Yang, Y. Tang, Z. Qiu, C. Wang and T. Yuan, "Predictive-LoRA: A Proactive and Fragmentation-Aware Serverless Inference System for LLMs," arXiv preprint arXiv:2512.20210, 2025.
9. N. Lyu, F. Chen, C. Zhang, C. Shao and J. Jiang, "Deep Temporal Convolutional Neural Networks with Attention Mechanisms for Resource Contention Classification in Cloud Computing," 2025.
10. B. Chen, "FlashServe: Cost-Efficient Serverless Inference Scheduling for Large Language Models via Tiered Memory Management and Predictive Autoscaling," 2025.
11. Xu J, Wu H, Wang J, et al. Anomaly transformer: Time series anomaly detection with association discrepancy[J]. arXiv preprint arXiv:2110.02642, 2021.
12. Chen J, Liu F, Jiang J, et al. TraceGra: A trace-based anomaly detection for microservice using graph deep learning[J]. Computer Communications, 2023, 204: 109-117.
13. Ding C, Sun S, Zhao J. MST-GAT: A multimodal spatial–temporal graph attention network for time series anomaly detection[J]. Information Fusion, 2023, 89: 527-536.