Transactions on Computational and Scientific Methods | Vo. 5, No. 12, 2025
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Modeling Complex Service Dependencies for Multidimensional Anomaly Detection via Attention Mechanisms

Parker Linton
University of North Texas, Denton, USA
parker8282@gmail.com

Abstract: This study proposes an attention-driven anomaly detection model for complex service dependencies to address the challenges of dynamic coupling among multidimensional metrics, intricate dependency structures, and diverse anomaly patterns in cloud service systems. The model achieves unified representation of temporal dynamics and structural semantics through multi-scale temporal feature encoding and an adaptive dependency modeling mechanism. In the feature extraction stage, a multi-head attention mechanism captures both local fluctuations and long-term trends across multi-granularity time windows, enhancing semantic interaction and contextual awareness among features. In the structural modeling stage, dynamic graph construction and dependency sparsification strategies are employed to characterize latent semantic associations and propagation relationships among services, enabling effective identification of cross-node anomaly patterns. Furthermore, a temporal consistency regularization term is introduced to maintain cross-time smoothness and dependency continuity of latent states, ensuring robustness and generalization under highly dynamic conditions. The model is systematically evaluated on multidimensional cloud resource usage data from perspectives such as hyperparameter sensitivity, environmental perturbation adaptability, and data distribution variation. Experimental results show that the proposed method outperforms several existing models in key metrics, including AUC-ROC, AUPR, F1-score, and detection latency, achieving high-accuracy and low-latency anomaly detection in complex dependency environments. This framework provides a scalable, interpretable, and efficient technical pathway for intelligent operation and dynamic dependency modeling in cloud service systems, offering valuable insights for the field of multidimensional time-series anomaly detection.

Keywords: Cloud service anomaly detection; dynamic dependency modeling; multi-head attention mechanism; temporal consistency constraints

1. Introduction

In modern cloud computing and distributed systems, the collaboration and dependency among services have become increasingly complex. The dynamic changes in system states and multi-layer interactions make anomaly detection one of the core challenges in intelligent operations. With the widespread adoption of microservice architectures, containerized deployment, and multi-tenant scheduling, traditional single-metric monitoring can no longer capture the intricate dependency structures and temporal dynamics within systems. Each service acts both as a caller and a callee, and its performance fluctuations may propagate along the call chain, triggering cascading failures that lead to service degradation or even system crashes. Identifying potential anomalies in high-dimensional, heterogeneous, and non-stationary service data, as well as

understanding their propagation paths within the dependency network, has become a key scientific problem for achieving high reliability and resilience in cloud service systems[1,2].

Complex service dependencies are reflected not only in multi-layer coupled topologies but also in the dynamic evolution of semantic relationships. The interaction intensity, contextual semantics, and resource competition among services may vary across different time windows, forming highly nonlinear and non-stationary dependency patterns. Such dynamic characteristics make traditional static modeling approaches ineffective in representing system behavior over time. In high-concurrency and multi-tenant environments, service metrics are influenced by network fluctuations, resource contention, and scheduling policies, leading to anomalies that are often time-varying and localized. Relying solely on fixed thresholds or single statistical features can easily cause false alarms and missed detections. Therefore, it is essential to develop a modeling mechanism that can adaptively capture dynamic dependencies, focus on critical contextual information, and maintain global consistency in the temporal dimension, ensuring stability and interpretability in cloud service systems[3,4].

With the advancement of deep learning, the attention mechanism has shown remarkable advantages in handling high-dimensional time series data and complex dependency structures. Its core concept lies in learning correlation weights among features, enabling the model to focus on important moments and nodes. This allows it to extract the most informative patterns for anomaly detection amid redundant information and noise[5]. Compared with traditional convolutional or recurrent architectures, the attention mechanism flexibly captures long-range dependencies across time steps and services, exhibiting stronger contextual awareness and feature selection capabilities. This provides new insights for anomaly detection in complex service environments. By introducing attention into dependency modeling, a global dependency view can be constructed in high-dimensional metric spaces, enhancing the model's ability to identify hidden anomaly propagation paths and critical semantic relationships, thus achieving more accurate and robust detection.

Furthermore, the diversity and heterogeneity of service systems add additional complexity to anomaly detection. Metrics collected from different services differ in distribution, scale, and sampling frequency, and many key dependencies are difficult to represent explicitly in topology graphs[6]. Such "implicit dependencies" are often hidden within collaborative variations of multi-dimensional features and require deep representation learning to uncover. Attention-driven dependency modeling can reconstruct this implicit semantic space adaptively, enabling dynamic perception of potential service interactions. This allows the model to capture weak yet crucial deviation signals in the early stages of anomaly occurrence. Through dual aggregation in both temporal and structural dimensions, the model achieves a holistic understanding of complex system behaviors, providing a solid foundation for subsequent root cause localization and self-healing decision-making[7].

From the perspective of intelligent operations and cloud service management, studying attention-driven anomaly detection under complex service dependencies holds significant theoretical and practical value. On one hand, it drives the evolution of anomaly detection from point-level monitoring to system-level cognition, enabling unified modeling and feature interaction over the global dependency graph. On the other hand, it lays the technical foundation for building adaptive and self-learning intelligent operation systems, allowing cloud environments to maintain stable performance and service continuity even under highly dynamic, concurrent, and uncertain conditions. This research enhances the observability and interpretability of complex systems and provides critical support for the development of future intelligent cloud service platforms, promoting the deep integration of artificial intelligence and cloud computing[8].

2. Method

To address the challenge of anomaly detection in complex service dependencies, this paper proposes an attention-based anomaly detection model designed to jointly capture temporal features and structural semantics under the coupling of multidimensional system metrics and dynamic inter-service dependencies.

Inspired by the principles of adaptive robust control in dynamic system modeling [9], the proposed method integrates attention mechanisms into a multi-level dependency representation learning framework to model the time-varying interaction strengths and contextual correlation patterns among services, while reinforcing temporal consistency in learned representations. The overall framework comprises three key components. First, a temporal feature encoding module embeds multidimensional metrics and extracts multi-scale features to capture both local dynamics and global operational trends of services. Second, an adaptive dependency modeling mechanism constructs a dynamic relational graph among service nodes, enabling contextual fusion of structural semantics and adapting to the temporal evolution of service relationships. Finally, a multi-head attention mechanism, combined with a consistency regularization term, is employed to learn critical dependency paths and anomaly decision boundaries within the global feature space, thereby enhancing the model's robustness and interpretability in complex anomaly scenarios. The model architecture is illustrated in Figure 1, and the detailed mathematical formulation is as follows:

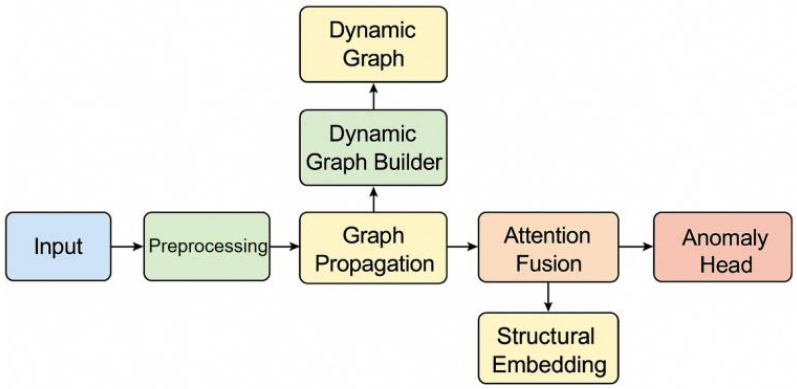

Figure 1. Overall model architecture

In the model input phase, the multi-dimensional monitoring indicators of the system at time step t are represented as matrix $X_t \in R^{N\times D}$, where N is the number of service nodes and D is the feature dimension. First, the basic time series representation is obtained through linear embedding and nonlinear mapping:

$$Ht = \sigma\left(W_e X_t + b_e\right)$$

Where $W_e \in R^{D_h \times D}$ is the feature transformation matrix, b_e is the bias term, and $\sigma(\cdot)$ is the activation function. This step completes the mapping from the original monitoring indicators to the semantic feature space, providing input representation for subsequent dependency modeling.

Next, to characterize the dynamic structural dependencies between services, this study constructs a time-variable adjacency matrix A_t, whose elements reflect the interaction intensity between nodes i and j at time step t. This matrix is generated through the feature similarity projection mechanism:

$$A_t(i,j) = \frac{\exp(sim(H_t^i, H_t^j))}{\sum_{k=1}^{N} \exp(sim(H_t^i, H_t^k))}$$

Where $sim(\cdot)$ represents the similarity function, which can be in the form of dot product or cosine similarity. This adaptive composition process enables the model to dynamically capture the time-varying characteristics of service dependencies and provide structural priors for subsequent attention weighting.

After the graph structure is constructed, the model implements context aggregation and structural embedding updates through the graph propagation mechanism. Let $\widetilde{A}_t$ be the adjacency matrix with self-loops added, and D_t be the corresponding degree matrix. The updated structure is represented as:

$$Z_t = \widetilde{D}_t^{-\frac{1}{2}} \widetilde{A}_t \widetilde{D}_t^{-\frac{1}{2}} H_t W_g$$

Where W_g is the graph convolution weight matrix. This step achieves collaborative updates of features between multiple services through neighborhood normalization aggregation, enabling the model to capture abnormal propagation paths at the structural level.

To further enhance the expression of global dependencies and key patterns at the feature level, this study introduces a multi-head attention mechanism to learn the importance of features from different semantic perspectives in the global feature space. For any set of query, key, and value vectors (Q, K, V), the attention output is defined as:

$$Attn(Q, K, V) = soft\ \max\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where d_k is the dimension of the key vector. By computing the attention distribution of different subspaces in parallel through a multi-head architecture, the model can simultaneously capture short-term dependencies and long-term associations in feature interactions, thereby achieving more fine-grained service relationship modeling capabilities.

Finally, to maintain the continuity and robustness of the model in the time dimension, a time consistency regularization term is introduced to constrain the implicit states of adjacent time steps to suppress the sudden interference caused by non-stationary noise. This constraint is formalized as:

$$L_{temp} = \frac{1}{T-1} \sum_{t=1}^{T-1} \left\| Z_{t+1} - Z_t \right\|_2^2$$

This ensures that the representations generated by the model in the time series dimension are smooth and dynamically consistent, thereby maintaining stable detection performance in the face of high noise and distribution drift.

In summary, the proposed attention-driven anomaly detection model for complex service dependencies achieves precise modeling and representation of abnormal behaviors in complex service environments through a collaborative framework that integrates multidimensional temporal feature extraction and dynamic structural modeling. The model consists of five key components: feature embedding, adaptive dependency modeling, graph semantic propagation, attention-based aggregation, and temporal consistency regularization. This approach provides a structured, interpretable, and generalizable technical solution for anomaly detection in high-dimensional, dynamically coupled system environments.

3. Performance Evaluation

2.1 Dataset

This study employs the Cloud Resource Usage Dataset for Anomaly Detection as the primary dataset for model validation. The dataset records time series data of resource utilization across multiple service nodes in a multi-tenant cloud environment, covering key metrics such as CPU utilization, memory usage, disk I/O, and network throughput. It also includes anomaly labels such as "overuse," which are used to identify hidden abnormal behaviors.

The dataset samples time series data at fixed intervals and synchronously records resource usage across different service dimensions. The anomaly labels do not explicitly specify inter-service call relationships but instead mark instances of resource overuse based on overall system anomalies. This design simulates the complexity of implicit dependencies and anomaly propagation in cloud service systems, requiring the proposed model to infer potential service couplings and anomaly correlations from the coordinated variations of multidimensional metrics without relying on an explicit call graph.

Using this dataset enables an empirical evaluation of the attention-driven model's ability to reconstruct dependencies and identify anomalies. By applying the proposed approach to this high-dimensional, heterogeneous, and labeled resource usage dataset, the model's capacity to capture implicit structures, interpret latent interaction paths, and perform fine-grained anomaly recognition can be effectively assessed. Moreover, the dataset's complexity and realism provide a solid foundation for extending the proposed method to real-world production environments.

2.2 Experimental Results

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

Table1: Comparative experimental results

Method	AUC-ROC (%)	AUPR (%)	F1-score (%)	Detection Delay (ms)
CFLOW-AD[10]	92.8	88.5	81.2	45.0
FastFlow[11]	93.5	89.7	82.4	42.0
Maat[12]	90.3	85.1	78.6	50.0
Ours	94.6	90.8	84.3	40.0

Overall, the proposed attention-driven anomaly detection model for complex service dependencies outperforms existing methods across all evaluation metrics, demonstrating its significant advantages in complex cloud service environments. Compared with traditional flow-based anomaly detection models, the proposed method (Ours) achieves AUC-ROC and AUPR scores of 94.6% and 90.8%, respectively, both higher than those of representative models such as CFLOW-AD and FastFlow. This result indicates that the proposed architecture possesses stronger robustness and representational capacity in global feature modeling and anomaly discrimination. The attention mechanism effectively enhances the model's sensitivity to potential anomaly patterns by capturing dynamic dependencies and temporal contextual correlations among services.

From the perspective of detection accuracy, the proposed model achieves an F1-score of 84.3%, significantly surpassing other models. This improvement is primarily attributed to the multi-head attention aggregation mechanism incorporated in dependency modeling, which enhances feature consistency and discriminability across temporal and structural dimensions. In contrast, although FastFlow and CFLOW-AD demonstrate certain advantages in local temporal modeling, they lack global dependency constraints across services, resulting in weaker performance in capturing anomaly propagation paths. This finding highlights the critical role of explicit dependency modeling in identifying anomalies within multidimensional system metrics and shows that the attention mechanism can adaptively focus on anomaly-related contextual information in high-dimensional heterogeneous feature spaces.

From a system response perspective, the proposed method achieves the lowest detection delay of 40 ms, significantly outperforming other approaches and validating its efficiency in real-time scenarios. This low-latency performance benefits from the structured dependency reconstruction and lightweight attention computation strategy, enabling rapid feature aggregation and anomaly judgment in high-frequency monitoring streams of multi-tenant cloud systems. For cloud service infrastructures requiring continuous

monitoring and rapid response, this efficiency translates into higher availability and reduced false-positive and false-negative risks, providing reliable technical support for intelligent operations.

In summary, the proposed approach achieves an effective balance between accuracy and real-time performance. Its superior results not only validate the effectiveness of attention-driven dependency modeling but also demonstrate that jointly considering structural information and temporal semantics is crucial for enhancing anomaly detection in complex environments with multidimensional time series and dynamic dependencies. The leading performance across multiple core metrics confirms the model's ability to capture implicit service correlations, strengthen anomaly semantic separability, and provide strong support for stable operation and intelligent management of cloud systems.

This paper also evaluates the sensitivity of prediction window length and number of attention heads to anomaly recognition performance. The experimental results are shown in Figure 2.

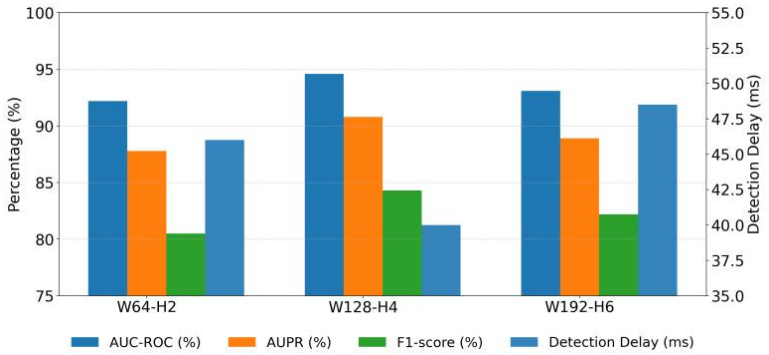

Figure 2. Hyperparameter sensitivity of prediction window length and number of attention heads to anomaly recognition performance

From the overall trend, as the prediction window length and the number of attention heads increase, the model exhibits nonlinear variations across performance metrics. When the window size expands from W64 to W128 and the number of attention heads increases from 2 to 4, both AUC-ROC and AUPR improve significantly, reaching 94.6% and 90.8%, respectively. This indicates that the model, under this configuration, can more effectively capture multi-scale temporal dependencies and inter-service contextual information, achieving stable differentiation of anomaly patterns. However, when the window further increases to W192 and the attention heads rise to 6, the model's discriminative capability slightly decreases. This suggests that an excessively large temporal receptive field may introduce redundant features and noise interference, thereby weakening the focus effect of attention weights on key dependencies.

From the perspective of detection accuracy, the F1-score peaks at 84.3% under the W128-H4 configuration. This result shows that in complex service dependency modeling, a moderate level of attention parallelism and a medium-sized time window achieve the best balance between global consistency and local sensitivity. Smaller windows fail to capture long-term dependency relations, while too many attention heads lead to feature dispersion and diluted attention distributions, reducing the separability of aggregated anomaly features. This finding validates the sensitivity of the proposed multi-head attention aggregation mechanism in dynamic dependency modeling and demonstrates that a synergistic constraint exists between the attention dimension and the temporal receptive field.

From the perspective of time response, the detection delay is minimized under moderate configurations, reaching only 40 ms, which highlights the high efficiency of the proposed method in real-time scenarios. When the window is too small, the model remains computationally lightweight but lacks sufficient feature information, leading to unstable predictions. Conversely, with larger windows and more attention heads, the complexity of graph construction and feature aggregation increases rapidly, resulting in longer latency. These results indicate that the proposed model effectively controls structural complexity through adaptive attention

fusion and sparse dependency modeling, maintaining low runtime latency while preserving strong discriminative power-making it suitable for online monitoring tasks in highly dynamic cloud service environments.

In addition, the variation patterns of different metrics reveal the model's sensitivity to multi-scale structural features. As the time window and attention head count increase simultaneously, AUC-ROC and AUPR remain highly consistent, while F1-score and latency exhibit opposite fluctuations. This divergence reflects the model's dynamic trade-off between feature fusion and anomaly boundary discrimination-balancing the need to capture broader dependency patterns against the risks of noise accumulation and computational overhead. The results demonstrate that the proposed architecture possesses strong adaptability in complex dependency modeling and temporal consistency maintenance, maintaining stable detection performance across multiple dimensions.

4. Conclusion

This study addresses the problem of multidimensional anomaly detection in complex cloud service environments and proposes an attention-based service dependency modeling approach to achieve joint representation of temporal dynamics and structural semantics. The proposed method tackles the limitations of traditional anomaly detection models in handling high-dimensional coupled features and dynamic dependencies through innovative designs across three key aspects: feature extraction, dependency reconstruction, and attention aggregation. A unified representation framework is constructed to ensure both global consistency and local sensitivity. Experimental results demonstrate that the proposed model outperforms mainstream baselines across multiple performance metrics, effectively capturing implicit inter-service dependencies and cross-temporal anomaly propagation patterns, thereby enabling high-precision anomaly identification in complex, dynamic, and non-stationary cloud systems.

By incorporating a multi-head attention mechanism and temporal consistency constraints, this study achieves a breakthrough in integrating temporal modeling with dependency learning. The attention mechanism allows the model to adaptively focus on anomaly-relevant regions in the feature space, effectively handling multi-source noise and asynchronous fluctuations. Meanwhile, the combination of dynamic graph construction and dependency sparsification strategies significantly enhances the model's ability to capture potential structural relationships among services. This structured representation not only improves the model's discriminability and interpretability but also provides theoretical and technical support for real-time monitoring, root cause analysis, and performance optimization in intelligent operations (AIOps), laying a solid foundation for self-learning and self-evolution in complex systems.

From an application perspective, the significance of this research extends beyond improving detection accuracy and response speed. It also promotes the practical integration of cloud computing and artificial intelligence. As enterprise systems increasingly evolve toward microservice-based and distributed architectures, the multidimensional coupling and dynamic dependencies of system operations become more intricate. The proposed attention-driven dependency modeling framework aligns with this trend, offering a transferable paradigm for cloud-native operation platforms, automated resource scheduling, and intelligent load balancing. Furthermore, its effectiveness in anomaly detection can be extended to multiple domains, such as edge computing resource management, network traffic anomaly identification, and industrial IoT monitoring, providing valuable insights for developing cross-domain intelligent monitoring systems.

Looking ahead, there remains significant room for further exploration. On one hand, future work could incorporate generative contrastive learning and uncertainty modeling mechanisms to enhance the model's generalization and adaptability to unseen anomaly patterns. On the other hand, integrating federated learning and privacy-preserving computation could enable a collaborative anomaly detection framework for multi-tenant environments, supporting secure cross-domain data sharing and joint modeling. Additionally, combining large-scale pre-trained models with graph-enhanced mechanisms may further strengthen

contextual understanding and reasoning capabilities in anomaly detection systems, paving the way toward a truly intelligent and self-evolving paradigm for cloud service operations. The findings of this study provide a solid theoretical and experimental foundation for this direction, holding substantial academic and industrial significance.

References

[1] 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.

[2] Marbel R, Cohen Y, Dubin R, et al. Cloudy with a Chance of Anomalies: Dynamic Graph Neural Network for Early Detection of Cloud Services' User Anomalies[J]. arXiv preprint arXiv:2409.12726, 2024.

[3] Gao C, Ma H, Pei Q, et al. Dynamic graph-based graph attention network for anomaly detection in industrial multivariate time series data[J]. Applied Intelligence, 2025, 55(6): 517.

[4] Ge D, Dong Z, Cheng Y, et al. An enhanced spatio-temporal constraints network for anomaly detection in multivariate time series[J]. Knowledge-Based Systems, 2024, 283: 111169.

[5] Kang H, Kang P. Transformer-based multivariate time series anomaly detection using inter-variable attention mechanism[J]. Knowledge-Based Systems, 2024, 290: 111507.

[6] Song J, Kim K, Oh J, et al. Memto: Memory-guided transformer for multivariate time series anomaly detection[J]. Advances in Neural Information Processing Systems, 2023, 36: 57947-57963.

[7] Ma R, Ma Y, Liu X. Time series anomaly detection via temporal relationship graphs and adaptive smoothing[J]. Applied Soft Computing, 2025: 113298.

[8] Somma M. Hybrid Temporal Differential Consistency Autoencoder for Efficient and Sustainable Anomaly Detection in Cyber-Physical Systems[J]. arXiv preprint arXiv:2504.06320, 2025.

[9] X. - T. Li, X. - P. Zhang, D. - P. Mao and J. - H. Sun, "Adaptive robust control over high - performance VCM - FSM," Optics and Precision Engineering, vol. 25, no. 9, pp. 2428 – 2436, 2017.

[10] Gudovskiy D, Ishizaka S, Kozuka K. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022: 98-107.

[11] Yu J, Zheng Y, Wang X, et al. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows[J]. arXiv preprint arXiv:2111.07677, 2021.

[12] Lee C, Yang T, Chen Z, et al. Maat: Performance metric anomaly anticipation for cloud services with conditional diffusion[C]//2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE). IEEE, 2023: 116-128.