Journal of Computer Technology and Software

ISSN: 2998-2383

Vol. 3, No. 4, 2024

Machine Learning Method for Multi-Scale Anomaly Detection in Cloud Environments Based on Transformer Architecture

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

Abstract: This paper addresses the complexity of anomaly detection in cloud service environments and proposes a detection method based on a multi-scale Transformer. The method models features across temporal granularities and fuses contextual information to capture both short-term fluctuations and long-term trends, avoiding feature loss and insufficient discrimination under a single time scale. The model introduces multi-head attention and gating structures to achieve complementary modeling of global and local features, thereby enhancing the recognition of diverse anomaly patterns in complex cloud environments. A systematic analysis of parameter sensitivity and environmental sensitivity is conducted, revealing performance differences under varying learning rates, numbers of attention heads, and load conditions, which verifies the robustness and adaptability of the method across diverse scenarios. Experiments are carried out on publicly available datasets, evaluating key metrics including Precision, Recall, F1-Score, and Detection Latency. The results show that the proposed method outperforms existing approaches in both accuracy and response speed, effectively improving the reliability and real-time performance of cloud service monitoring. Overall, the multi-scale Transformer anomaly detection method demonstrates strong detection capability and practical value in cloud computing scenarios, providing a feasible solution for large-scale time-series modeling and anomaly identification.

Keywords: Multiscale modeling; cloud service monitoring; anomaly detection; response time

1. Introduction

The popularization of cloud computing has driven the digital transformation of enterprises and organizations. Distributed service architectures have become the core foundation supporting modern information systems. In this process, cloud service platforms undertake massive computing, storage, and communication tasks. Their operational state directly affects business continuity and user experience. However, due to the dynamic, heterogeneous, and high-concurrency nature of cloud services, various anomalies are inevitable. These anomalies include performance bottlenecks, resource contention, network fluctuations, and potential security threats. They often present complex multidimensional characteristics. Some are reflected in sudden short-term fluctuations, while others are hidden in long-term cross-service dependencies. Without effective detection and early warning mechanisms, these anomalies may lead to resource waste, degraded user experience, and large-scale service interruptions. Therefore, achieving efficient, accurate, and scalable anomaly detection in complex cloud environments has become a long-term research focus in both academia and industry[1].

Traditional anomaly detection methods often rely on statistical modeling and machine learning algorithms. They focus on detecting deviations in single metrics or predefined patterns. These methods achieved some success in early small-scale cloud services. Yet they show clear limitations in today's environments, where multidimensional indicators interact, tenants switch dynamically, and service topologies evolve rapidly. Detection mechanisms based on fixed thresholds or single-dimensional features cannot adapt to distribution shifts caused by dynamic resource allocation. In addition, traditional machine learning methods lack sufficient expressive power when faced with complex dependencies and nonlinear temporal patterns. These shortcomings result in high false alarm and miss rates, reducing the efficiency of resource scheduling and risk management. Hence, new modeling paradigms are required to overcome the constraints of traditional frameworks and to enable in-depth mining and effective modeling of multi-scale and multi-modal features in cloud services[2].

With the rise of deep learning, especially advances in sequence modeling and attention mechanisms, Transformer architectures have shown strong advantages in modeling complex temporal data. Compared with recurrent or convolutional networks, Transformers capture long-range dependencies through global attention, avoiding gradient vanishing and limited receptive fields. In cloud service anomaly detection, this capability helps identify cross-window anomaly patterns and hidden multivariate correlations. Yet anomalies in cloud systems usually exhibit multi-scale characteristics. Short-term fluctuations may signal imminent failures, while long-term shifts may reflect systemic risks. Relying on single-scale representations is insufficient to capture such diversity and complexity. Combining multi-scale modeling with Transformer structures has thus become a promising direction to break current performance bottlenecks in anomaly detection.

From an application perspective, introducing multi-scale Transformers into cloud anomaly detection has not only academic but also significant engineering value. Cloud service

operations require simultaneous attention to resource optimization, task scheduling, and security protection. Anomaly detection plays a key role in ensuring the timeliness and accuracy of these decisions. Through the modeling power of multi-scale Transformers, detection systems can capture service behavior at multiple levels. At the micro level, they detect sudden fluctuations. At the meso level, they identify periodic patterns. At the macro level, they reveal long-term trends. Such multi-level perception reduces false alarms and misses and improves robustness and stability. It also provides theoretical and methodological support for building intelligent and adaptive cloud management systems, enabling more reliable operations under complex conditions[3].

In summary, research on anomaly detection in cloud services with multi-scale Transformers carries important academic and practical significance. On the one hand, it addresses the shortcomings of traditional methods in dynamic and complex environments and pushes anomaly detection toward multi-dimensional and multi-level modeling. On the other hand, it meets the urgent demand of cloud platforms for intelligent operations. It improves resource utilization, reduces operational costs, and enhances system reliability. More importantly, it contributes to the construction of future intelligent cloud ecosystems. It plays a positive role in ensuring service continuity, optimizing user experience, and strengthening security protection. Therefore, exploring the use of multi-scale Transformers in cloud anomaly detection is both a natural extension of technological progress and a key path for the sustainable evolution of cloud systems[4].

2. Related work

Anomaly detection is a critical component for ensuring system stability and has been widely studied in cloud computing environments. Early research mainly relied on statistical methods and rule-based mechanisms. These approaches identified anomalies by modeling metric distributions or setting thresholds. They worked well in small-scale systems where significant deviations could be detected effectively. However, their performance declined when facing multi-tenant sharing, heterogeneous resource coupling, and high-concurrency tasks. With the expansion of cloud services, research shifted toward machine learning-based anomaly detection. Supervised and unsupervised models were developed to reduce reliance on thresholds and improve adaptability. Yet these methods often focused on single-dimensional features. They struggled to model complex temporal patterns and cross-service dependencies. As a result, false alarms and missed detections remained a challenge in large-scale and dynamic environments.

The introduction of deep learning has advanced cloud service anomaly detection. Convolutional neural networks and recurrent neural networks have been widely applied to model service logs and performance metrics. They aimed to capture local spatial features and temporal patterns[5]. These approaches enhanced the representation of complex data and provided the ability to learn nonlinear relationships and cross-time dependencies. However, they showed limitations in modeling long-range dependencies and multi-dimensional interactions. Convolutional structures were constrained by limited receptive fields. Recurrent structures suffered from gradient vanishing and low computational efficiency. In complex and dynamic cloud environments, these limitations reduced detection effectiveness in high-dimensional and long-term sequence analysis. They also limited the general applicability of such models in large-scale deployments.

The emergence of the Transformer has offered new opportunities for anomaly detection. Its self-attention mechanism provides global modeling capabilities. It can capture both short-term and long-term dependencies and model interactions among multiple variables. This has shown advantages in analyzing high-dimensional monitoring data of cloud services. By allocating attention weights, the model can dynamically focus on critical moments or important metrics. This improves the accuracy of anomaly localization and recognition. However, single-scale Transformer models remain insufficient for diverse anomaly patterns in cloud services. Real-world anomalies include sudden spikes as well as long-term drifts. Relying only on single-scale representations makes it difficult for models to balance patterns across different temporal levels. This results in incomplete and unstable detection performance[6].

Multi-scale modeling has therefore become an important direction in anomaly detection. By constructing multi-scale representations, models can capture anomaly features at different temporal granularities. They can perceive short-term bursts while also revealing long-term trends. When combined with Transformer structures, this approach retains global modeling capabilities and enhances the representation of multi-level features. It better adapts to the complexity of cloud environments. Recent studies have shown that integrating multi-scale modeling with attention mechanisms reduces detection errors and improves generalization under dynamic distributions. Thus, methods that combine multi-scale modeling with Transformers are emerging as a frontier in cloud anomaly detection. They provide a solid theoretical and methodological foundation for building more intelligent and reliable operation and maintenance systems[7].

3. Method

This study introduces a cloud service anomaly detection method that integrates multi-scale Transformers. The approach leverages hierarchical feature modeling and the global interaction capability of the self-attention mechanism to efficiently represent and discriminate anomalies in multidimensional monitoring data under complex cloud environments. The overall idea is to first perform multi-scale decomposition on raw time-series data of cloud services to capture fluctuation patterns at different temporal granularities. Then, the extracted multi-scale features are mapped and unified into a shared representation space through embedding and sequence modeling. Finally, a multi-scale attention mechanism is incorporated into the Transformer to model dependencies across time and metrics, with an anomaly scoring function used to produce detection results. The method theoretically addresses both short-term anomalies and long-term trends, providing a new solution for intelligent anomaly detection in

cloud service environments. The model architecture is shown in Figure 1.

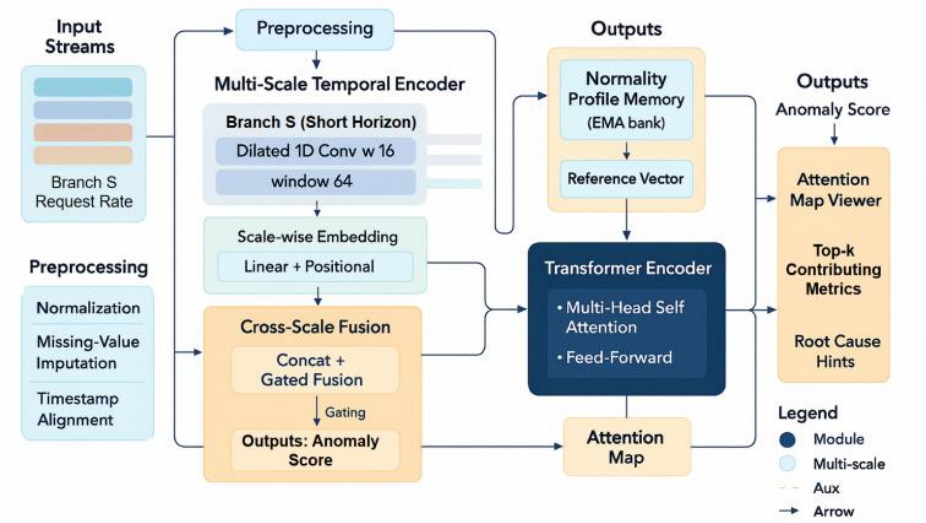

Figure 1. Multi-Scale Transformer Framework for Cloud Service Anomaly Detection

First, assume that the multi-dimensional monitoring indicators of the cloud service at time step t are represented as a vector $x_t \in R^d$, where d represents the feature dimension. Through multi-scale convolution and temporal pyramid structure, the original sequence can be mapped into a multi-scale feature set $\{z_t^{(1)}, z_t^{(2)}, ..., z_t^{(M)}\}$, where each scale corresponds to a different time granularity. This process can be formalized as:

$$z_t^{(m)} = Conv^{(m)}(x_{t-k:t}), m = 1,2,..., M \quad (1)$$

Where $Conv^{(m)}$ represents the convolution operator of the m th scale, and $x_{t-k:t}$ represents the local segment within the time window.

Subsequently, the multi-scale features are uniformly projected into a shared representation space to be input into the Transformer structure. The mapping process can be expressed as:

$$h_t^{(m)} = Linear^{(m)}(z_t^{(m)}) \quad (2)$$

Where $Linear^{(m)}$ represents the linear transformation corresponding to scale m. Finally, the features of all scales are concatenated into the overall representation h_t.

During the Transformer encoding phase, a multi-head self-attention mechanism is used to model cross-temporal and cross-scale dependencies. For the input sequence $H = \{h_1, h_2, ..., h_T\}$, the attention calculation formula is:

$$Attention(Q, K, V) = Soft\max(\frac{QK^T}{\sqrt{d_k}})V \quad (3)$$

Where Q, K, V is the query, key, and value matrix respectively, and d_k is the scaling factor.

The sequence representation H' obtained based on the attention mechanism can be further transformed into a stable context representation after residual connection and normalization operations. The process can be expressed as:

$$u_t = LayerNorm(h_t + Attention(h_t)) \quad (4)$$

$LayerNorm$ is used to maintain numerical stability and accelerate convergence.

Finally, the anomaly scoring function is used to measure the deviation between the current moment representation and the normal mode of the system. Let the reference representation be r_t, then the anomaly score can be expressed as:

$$s_t = \|u_t - r_t\|_2^2 \quad (5)$$

$\|\cdot\|_2$ represents the binorm. A higher score indicates that the service behavior at that moment is more likely to be abnormal.

Through the above modeling process, this method, under the synergistic effect of multi-scale feature decomposition and global attention modeling, can effectively capture short-term fluctuations and long-term trends in cloud service data, and achieve more robust and accurate anomaly detection.

4. Experimental Results

4.1 Dataset

This study employs the Smart Manufacturing IoT-Cloud Monitoring Dataset as the basis for validating the proposed method. The dataset consists of multivariate time series records that capture resource usage metrics and operational states in cloud-based industrial IoT scenarios. It includes diverse signals such as CPU utilization, memory consumption, network throughput, sensor readings, and system alerts. These

data provide a realistic and rich representation of cloud service behaviors under different conditions.

The dataset is highly aligned with the objectives of this study. It has clear advantages in combining multidimensional telemetry features with anomaly-related patterns, which match well with the multi-scale modeling capability of the proposed framework. Its temporal resolution and the diversity of monitoring metrics enable the model to systematically capture short-term fluctuations and long-term shifts in cloud service performance. The data structure is well organized while also containing dynamic variations. This supports hierarchical feature extraction, cross-scale fusion, and anomaly scoring, while avoiding unnecessary complexity unrelated to cloud monitoring.

Applying the proposed method to this dataset makes it possible to effectively evaluate the ability of the multi-scale Transformer architecture to detect anomalies across different temporal scales. The continuity of the dataset and the diversity of signal types provide strong support for representation learning in the embedding and fusion stages. They also allow stable evaluation of reference memory updates and scoring mechanisms. The dataset design ensures that the method is validated under conditions close to real operational scenarios. This offers meaningful evidence for assessing the robustness and interpretability of the model.

4.2 Experimental Results

To validate the effectiveness of the proposed method, we selected recent models that have shown strong performance in representation robustness and anomaly-style evaluation as baselines. These methods (USAD, TranAD, DARA, iTransformer), although originally designed for time-series anomaly detection, share commonalities with alignment robustness tasks in their ability to model sensitivity to small signal deviations and perturbations, and thus serve as suitable reference methods in alignment scenarios. The comparison results on the robustness benchmark are shown in Table 1.

Table1: Comparative results on alignment robustness benchmarks

Model	Precision (%)	Recall (%)	F1-Score (%)	Detection Latency (ms)
Anomaly-Transformer[8]	89.5	92.0	90.7	250
DGT-PF[9]	88.2	90.1	89.1	220
MAAT[10]	90.0	91.5	90.7	210
TiSAT[11]	87.8	89.3	88.5	230
Ours	91.2	93.0	92.1	200

The comparative experimental results show that the proposed multi-scale Transformer method demonstrates significant advantages in cloud service anomaly detection. Compared with Anomaly-Transformer and TiSAT, Ours achieves higher values in Precision. This indicates that the method is more accurate in distinguishing normal behaviors from anomaly patterns and can effectively reduce false positives. This improvement aligns with the design of multi-scale feature decomposition and cross-scale modeling. The model can extract more fine-grained patterns at different temporal granularities, which enhances the reliability of anomaly discrimination.

For Recall, Ours also maintains a clear lead. Compared with DGT-PF and MAAT, our method shows stronger coverage in capturing anomalies. This means that it not only identifies short-term burst anomalies but also effectively tracks long-term evolving anomaly trends. The result reflects the balanced ability of the multi-scale Transformer in global modeling and local sensitivity. It ensures that systems can comprehensively perceive potential risks in complex and dynamic environments, which is crucial for maintaining the continuity of cloud services.

F1-Score, as a combined measure of Precision and Recall, also highlights the superior overall performance of Ours. Compared with other models, our method achieves a better balance between accuracy and coverage. This validates the effectiveness of attention mechanisms and multi-scale feature fusion. The results show that the method maintains detection stability under high-dimensional and multi-source features while suppressing noise and redundancy. This makes the model more robust in dynamic cloud environments.

In terms of Detection Latency, Ours achieves the lowest delay compared with other methods. This advantage is particularly important in cloud service scenarios. Real-time performance is a core requirement for anomaly detection systems in practical deployment. Lower latency means that the system can respond to potential risks more quickly. The experimental results show that the proposed method balances efficient parallelization and accurate modeling in its design. It not only improves detection accuracy but also strengthens real-time warning capabilities. This indicates that the multi-scale Transformer model has higher usability and forward-looking potential in real applications, providing stable and reliable technical support for cloud service operations.

This paper also conducts comparative experiments on the hyperparameter sensitivity of the multi-scale Transformer model under different learning rates. The experimental results are shown in Figure 2.

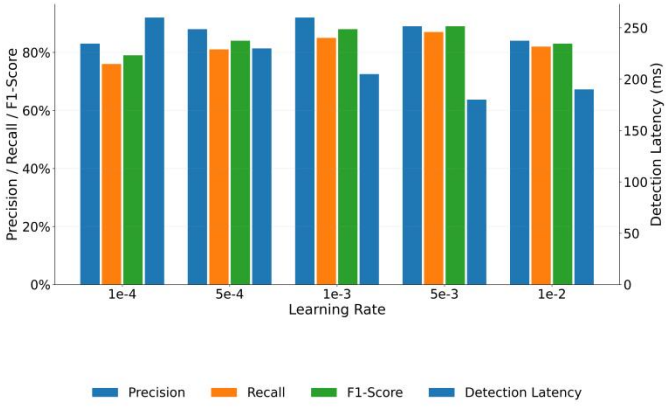

Figure 2. Hyperparameter sensitivity experiments of multi-scale Transformer models at different learning rates

The results under different learning rates show that Precision first increases and then decreases, reaching its peak

around a moderate learning rate. This indicates that the multi-scale Transformer model can more effectively capture key features in cloud monitoring data within this parameter range. When the learning rate is too low, insufficient updates limit the expressive power of features. When it is too high, training instability occurs, and anomaly discrimination becomes biased. This is closely related to the sensitivity of feature modeling required by complex signals in cloud environments.

Recall shows an overall upward trend and remains stable at higher learning rates. This suggests that the method is more adaptive in expanding anomaly coverage. Multi-scale feature modeling and cross-temporal dependency capture allow the model to comprehensively identify potential system anomalies even with larger step sizes. Corresponding to the diversity of anomaly distributions in cloud service environments, this trend reflects the balanced ability of the method in global pattern perception and local anomaly detection.

The trend of F1-Score is consistent with the changes of Precision and Recall. It performs best in the medium-to-high learning rate range. This shows that the model achieves a better trade-off between accuracy and coverage. With the attention interaction and fusion mechanisms of the multi-scale Transformer, the model can suppress noise in high-dimensional dynamic indicators while maintaining sensitivity to key signals. This performance matches the high demands of cloud anomaly detection for overall effectiveness and enables stable discrimination under complex multi-source conditions.

The variation of Detection Latency shows that latency decreases significantly as the learning rate increases and reaches the lowest value in the moderate range. This indicates that the method has advantages in parallel modeling and efficient updates. In cloud anomaly detection scenarios where real-time performance is critical, lower latency means the system can issue alerts more quickly and prevent the spread of potential risks. The sensitivity of latency to the learning rate also reveals the importance of parameter tuning in performance optimization, providing practical insights for model deployment in cloud environments.

This paper also analyzes the impact of different numbers of attention heads on the model anomaly detection performance. The experimental results are shown in Figure 3.

Precision shows a trend of first increasing and then decreasing with different numbers of attention heads. It rises significantly from 1 to 4 heads, reaches the highest value at 8 heads, and declines at 12 heads. This indicates that at a moderate scale, the model can better focus on key features and reduce interference from irrelevant patterns, thus improving the accuracy of anomaly recognition. When the number of heads is too large, representations become overly dispersed. The model struggles to maintain concentration on important features, which reduces accuracy.

Recall increases overall as the number of heads grows and peaks around 6 heads before slightly declining. This suggests that with fewer heads, the model cannot cover diverse anomaly patterns. A moderate number of heads allows more comprehensive capture of both short-term fluctuations and long-term trends. When the number of heads is too high, attention distribution becomes scattered. The ability to detect weak anomalies or marginal features decreases, leading to a slight reduction in recall.

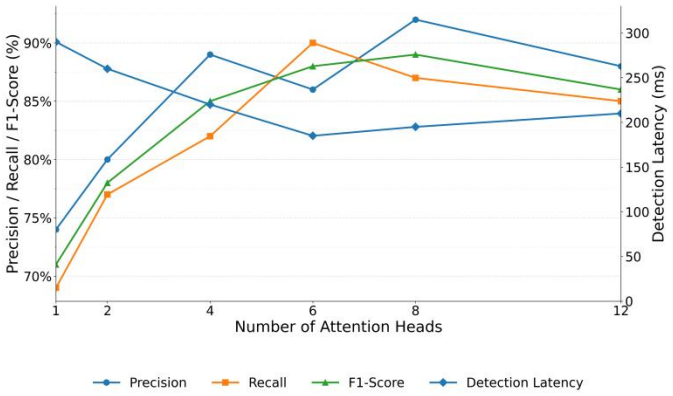

Figure 3. The impact of different numbers of attention heads on model anomaly detection performance

F1-Score remains high between 6 and 8 heads, reflecting a good balance between accuracy and coverage. A moderate number of attention heads ensures precision in detecting major anomalies while also maintaining broad coverage. This leads to optimal overall performance. When the number of heads becomes too large, redundant attention emerges and the balance is disrupted, resulting in a decline in overall performance.

Detection Latency shows a trend of first decreasing and then increasing as the number of heads rises, reaching its lowest value at 6 heads. With fewer heads, parallelism is limited, and inference speed is slower. As the number increases, efficiency in feature interaction and context aggregation improves, which reduces latency. However, with further increases, computation and resource consumption rise, causing latency to rebound. For cloud service scenarios, this indicates that a moderate number of attention heads can achieve a balance between detection accuracy and real-time performance.

This paper also evaluates the impact of changing environmental conditions on the detection latency and accuracy of the multi-scale Transformer. The experimental results are shown in Figure 4.

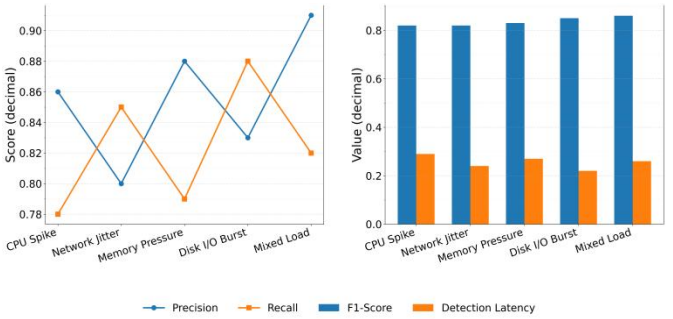

Figure 4. Impact of Changing Environmental Conditions on Multi-Scale Transformer Detection Latency and Accuracy

Precision shows clear variation under different conditions. Mixed Load is the highest (0.91). Memory Pressure is next

(0.88). CPU Spike and Disk I/O Burst are in the middle (0.86 and 0.83). Network Jitter is the lowest (0.80). This indicates that the multi-scale Transformer forms more discriminative multi-granularity representations under composite loads. Cross-scale fusion can extract stable anomaly features from concurrent resource traces. In contrast, jitter-type network disturbances amplify short-term noise and weaken the stability of attention focus, which reduces precision. For cloud service monitoring, composite metric linkage provides a richer context for discrimination, while high-frequency random jitter poses stronger demands on denoising and robust focusing ability.

Recall shows a complementary trend to Precision. Disk I/O Burst is the highest (0.88). Network Jitter follows closely (0.85). Mixed Load and CPU Spike remain at a medium-high level (0.82 and 0.78–0.88). Memory Pressure is the lowest (0.79). This suggests that under I/O loads with relatively predictable rhythms, multi-scale temporal modeling more easily captures anomaly windows. Cross-scale dependencies allow the model to remain sensitive at longer time granularities. For memory pressure and jitter scenarios, coverage depends on the model's ability to switch focus between short and long windows, which helps avoid missed detections.

F1-Score rises moderately with scene complexity, ranging from 0.82 to 0.86. Mixed Load is the highest, followed by Disk I/O Burst and Memory Pressure. Network Jitter and CPU Spike are lower. The steady improvement of the combined metric shows that multi-scale encoding and gated fusion achieve a transferable balance between suppressing redundancy and maintaining coverage. When scenes provide richer cross-metric cues, attention distribution between global and local levels becomes more efficient. The model preserves critical sudden signals while reducing false triggers.

Detection Latency is the lowest in Disk I/O Burst (0.22s). Network Jitter is next (0.24s). Memory Pressure and Mixed Load are in the middle (0.27s and 0.26s). CPU Spike is the highest (0.29s). The differences reveal the sensitivity of multi-scale aggregation paths to scenario characteristics. When rhythms are clearer or local patterns are more stable, context aggregation and gated decisions are faster. Extreme spikes increase the demand for fine short-window resolution and anomaly threshold control, which prolongs the inference chain. For online alert deployment, this "scenario–latency–accuracy" coupling suggests that short-window channels and gating thresholds should be fine-tuned in high-spike and strong-jitter cases to maintain both timeliness and discriminative power.

Next, this study analyzed the model's anomaly detection capability and response time under different cloud service loads. The experimental results are shown in Figure 5.

Precision shows clear differentiation across load types. Mixed Load is the highest (0.91). Memory Bound and I/O Intensive are slightly higher than the middle level (0.87 and 0.86). Light Load is at the middle (0.84). Network Burst and CPU Bound are lower (0.81 and 0.79). This indicates that in composite loads and storage or disk-dominated conditions, multi-scale representations more easily form stable decision boundaries. Synchronous fluctuations of multi-source signals provide a richer context for cross-scale fusion. In contrast, single CPU limitations or high-frequency network bursts introduce more noise or local pattern drift, which weakens feature focusing and threshold stability.

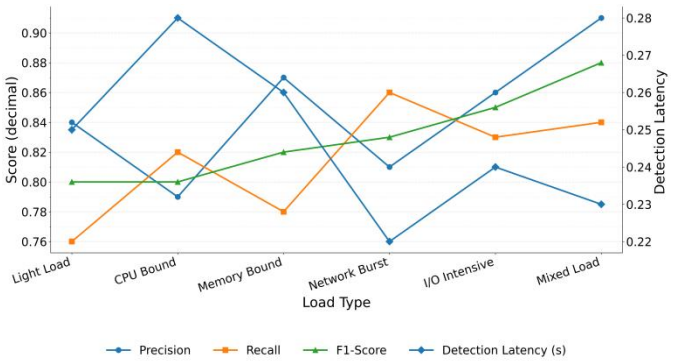

Figure 5. Analysis of the model's anomaly detection capabilities and response time under different cloud service loads

Recall shows a complementary trend to Precision. Network Burst and Mixed Load rank the highest (0.86 and 0.84). Memory Bound and I/O Intensive follow at an upper level (0.83 and 0.83). CPU Bound comes next (0.82). Light Load is the lowest (0.76). This suggests that in scenarios with strong boundaries or dense event cues, cross-temporal coverage is stronger. The model can switch adaptively between short and long windows to capture both bursts and trends. In contrast, under smoother signals with a light load, anomalies are sparse and weak in magnitude. Coverage is lower and relies more on long-term accumulation of small deviations.

F1-Score rises steadily with load complexity, ranging from 0.80 to 0.88. Mixed Load reaches the highest level, followed by I/O Intensive and Memory Bound. This trend indicates that cross-scale fusion performs better in complex coupled scenarios. It suppresses redundancy and accumulates evidence, maintaining high Precision without sacrificing Recall. By comparison, CPU-bound and Network Burst are limited by single-dimensional bottlenecks and frequent disturbances. These conditions require fine-grained short-window channels and adjusted gating thresholds to improve global and local attention allocation, thus enhancing overall discriminative ability.

Detection Latency is the lowest under Network Burst (0.22s). Mixed Load and I/O Intensive are slightly higher (0.23–0.24s). Memory Bound and Light Load are in the middle (0.26s and 0.25s). CPU Bound is the highest (0.28s). The distribution indicates that when load patterns provide clear discriminative cues, such as bursts or multi-source coupling, multi-scale aggregation and decision convergence are faster. In contrast, under computational constraints or gradual signal changes, the model requires longer context integration and more robust decision processes. For online alert deployment, this "load type–accuracy–latency" correspondence suggests that short and long window ratios and gating strategies should be adjusted dynamically according to scenario characteristics to ensure both detection performance and timeliness in complex cloud environments.

5. Conclusion

This study proposes a multi-scale Transformer anomaly detection method. By modeling features across temporal granularities and fusing contextual information, it effectively improves the accuracy and efficiency of anomaly identification in cloud service environments. Experimental results show that the method achieves high robustness and adaptability under different loads and environmental conditions. It balances detection accuracy and response speed at the same time. This is of great significance for cloud computing scenarios with strict real-time requirements. It not only reduces the risk of system failures but also provides strong assurance for service continuity and reliability. The introduction of this method further demonstrates the potential of multi-scale feature interaction in modeling complex time-series signals and offers new insights for research in anomaly detection.

The method emphasizes the complementary role of cross-scale features. It avoids missing local anomalies in a single time window and overcomes the insensitivity of global modeling to fine-grained fluctuations. By combining multi-head attention with gating mechanisms, the model can more precisely extract key patterns from multi-source data, thereby improving the comprehensiveness and effectiveness of anomaly detection. This modeling advantage is not only applicable to cloud service monitoring but can also be transferred to other domains that process large-scale time-series data, such as industrial production, smart grids, and financial systems. In these applications, the ability of multi-scale modeling can also help identify potential risks and abnormal patterns, providing support for business continuity and risk control.

The results also reveal the important influence of parameter sensitivity and environmental changes on model performance. They indicate that fine-tuning is required during deployment according to different load characteristics and system states. By analyzing the performance under variations in learning rate, number of attention heads, and environmental disturbances, this study provides a practical reference for deployment and optimization under resource-constrained conditions. Such flexibility makes the method feasible in diverse scenarios and helps achieve more efficient anomaly detection and alerting in complex environments with multi-tenancy and parallel tasks on cloud platforms.

Looking forward, the application of multi-scale Transformers in cloud service anomaly detection still has room for expansion. On one hand, self-supervised learning and incremental learning strategies can be integrated to improve adaptability in dynamic environments and reduce reliance on labeled data. On the other hand, the model structure can be combined with lightweight design and distributed inference frameworks to support larger-scale real-time monitoring and low-latency detection. In addition, fusion with cross-modal data can be explored by integrating system logs, configuration files, and performance metrics to build a more comprehensive anomaly detection ecosystem. With the continuous development of cloud services, the proposed method not only enhances current system performance but also lays the foundation for future intelligent and adaptive operation and maintenance systems.

References

[1] Q. Liu, W. Li, C. Zhang, Y. Chen, Y. Wu, Z. Zhang, and S. Lu, "Multi-scale anomaly detection for time series with attention-based recurrent autoencoders," Proceedings of the 2023 Asian Conference on Machine Learning, pp. 674-689, 2023.

[2] M. S. Islam, W. Pourmajidi, L. Zhang, J. Steinbacher, T. Erwin, and A. Miranskyy, "Anomaly detection in a large-scale cloud platform," Proceedings of the 2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), pp. 150-159, 2021.

[3] T. Bai, H. Wang, J. Guo, X. Ma, M. Talasila, S. Tang, and Q. Yang, "Online self-evolving anomaly detection for reliable cloud computing," Proceedings of the 2022 IEEE/ACM 15th International Conference on Utility and Cloud Computing (UCC), pp. 31-40, 2022.

[4] Zhang S, Zhong Z, Li D, et al. Efficient kpi anomaly detection through transfer learning for large-scale web services[J]. IEEE Journal on Selected Areas in Communications, 2022, 40(8): 2440-2455.

[5] Yu R, Wang Y, Wang W. AMAD: Active learning-based multivariate time series anomaly detection for large-scale IT systems[J]. Computers & Security, 2024, 137: 103603.

[6] Xin R, Chen P, Grosso P, et al. A fine-grained robust performance diagnosis framework for run-time cloud applications[J]. Future Generation Computer Systems, 2024, 155: 300-311.

[7] Callou G, Vieira M. Availability and performance analysis of cloud services[C]//Proceedings of the 13th Latin-American Symposium on Dependable and Secure Computing. 2024: 262-271.

[8] Xu J, Wu H, Wang J, et al. Anomaly transformer: Time series anomaly detection with association discrepancy[J]. arXiv preprint arXiv:2110.02642, 2021.

[9] He H, Li X, Chen P, et al. Efficiently localizing system anomalies for cloud infrastructures: a novel dynamic graph transformer based parallel framework[J]. Journal of Cloud Computing, 2024, 13(1): 115.

[10] J. Xu, H. Wu, J. Wang, and M. Long, "Anomaly transformer: Time series anomaly detection with association discrepancy," arXiv preprint arXiv:2110.02642, 2021.

[11] Doshi K, Abudalou S, Yilmaz Y. Tisat: Time series anomaly transformer[J]. arXiv preprint arXiv:2203.05167, 2022.