Transactions on Computational and Scientific Methods | Vo. 3, No. 2, 2023
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

A Multi-Scale Deep Learning and Uncertainty Estimation Framework for Comprehensive Anomaly Detection in Cloud Environments

Zhimin Qiu
University of Southern California, Los Angeles, USA
zhiminiqiu@gmail.com

Abstract: This study proposes a multi-scale deep learning-based detection method to address the complexity, dynamics, and diversity challenges of anomaly detection in cloud service systems. By introducing multi-scale feature extraction and cross-scale fusion mechanisms, the method effectively characterizes system behavior evolution across different temporal granularities, enabling the capture of both short-term burst anomalies and long-term structural anomalies to improve detection comprehensiveness and accuracy. In terms of model architecture, hierarchical feature modeling and context-aware mechanisms are employed to achieve a deep representation of semantic associations and temporal dependencies among multidimensional metrics. In addition, an uncertainty estimation module is introduced to calibrate boundary samples and low-confidence predictions, which effectively reduces false positives and false negatives and enhances system stability and robustness in highly dynamic environments. The method is also systematically evaluated under various environmental factors, including hyperparameter variations, resource interference, sampling granularity, and data distribution drift. Experimental results show that it outperforms existing methods on multiple key metrics and demonstrates strong adaptability and discriminative power. Overall, the proposed multi-scale detection framework provides reliable technical support for intelligent operation, automated anomaly management, and complex service state monitoring in cloud computing systems, offering an effective solution for ensuring stability in large-scale distributed environments.

Keywords: Multiscale modeling; uncertainty estimation; anomaly detection; environmental sensitivity analysis

1. Introduction

With the rapid development of cloud computing and distributed systems, digital infrastructure has become the core support for modern socio-economic operations. Various internet services, enterprise applications, and data-intensive tasks rely on cloud platforms for computing, storage, and network resources. However, as service scale continues to grow and business logic becomes increasingly complex, system states exhibit high dynamics and uncertainty. Abnormal behaviors occur frequently and can lead to serious consequences. Issues such as service latency, throughput degradation, system crashes, or even data loss often stem from abnormal events such as resource scheduling imbalance, dependency failures, request pattern shifts, or external attacks. Because these anomalies are sudden, hidden, and diverse, failure to detect and respond to them in time can significantly impact business continuity and user experience. Therefore, achieving high-precision, real-time, and robust anomaly detection in complex and evolving cloud environments has become a crucial research topic for ensuring the reliability and stability of cloud services[1].

Traditional anomaly detection methods are mainly based on rule matching, statistical modeling, or simple machine learning techniques. These approaches were effective in earlier systems with simpler architectures and stable data distributions, but their limitations have become increasingly evident as cloud systems have evolved. Static rules and threshold strategies cannot handle data distribution drift caused by dynamic workloads and adaptive scheduling. At the same time, conventional models often rely on fixed features and a single time scale, making them unable to capture complex multidimensional dependencies and cross-temporal anomaly patterns effectively. Moreover, cloud environments feature multi-tenant sharing, heterogeneous resource interactions, and cascading services, where anomalies may manifest not only as single-point deviations but also as structural anomalies, behavioral chain anomalies, or context-related anomalies. These new characteristics impose higher demands on detection methods in terms of generalization, context modeling, and multi-scale perception capabilities[2].

In this context, deep learning offers new solutions for anomaly detection in cloud services. Leveraging the strong representation capabilities of deep neural networks, it is possible to automatically learn complex nonlinear feature representations from massive monitoring data and identify high-dimensional patterns and latent relationships that traditional methods fail to capture. However, anomalies in cloud environments often span multiple time scales. Short-term fluctuations may indicate sudden requests or transient disturbances, while long-term trends may reflect performance degradation or potential risks. Therefore, deep models based on a single time granularity struggle to balance global situational awareness with local detail recognition. Multi-scale deep learning methods have emerged to meet this need by jointly modeling different time windows, feature levels, and semantic abstraction layers[3]. This enables the model to capture rapid changes at the micro level while grasping evolving trends at the macro level, providing a more comprehensive, fine-grained, and interpretable perspective for anomaly detection.

Furthermore, multi-scale deep learning not only excels at feature extraction but also introduces structured temporal modeling and contextual correlation modeling mechanisms into anomaly detection. Through multi-level feature fusion and attention allocation, the model can identify critical anomaly signals within complex metric systems and understand causal chains and interaction relationships among metrics. At the same time, multi-scale modeling provides richer semantic support for anomaly localization and root cause analysis, shifting detection systems from simple "whether an anomaly exists" judgments to a deeper understanding of "where and why anomalies occur." This transformation is significant for building automated operations, intelligent scheduling, and self-healing systems. It also lays a solid foundation for the intelligent and autonomous management of cloud services[4].

In summary, anomaly detection for cloud environments is evolving from static thresholds and shallow learning toward multi-scale deep learning approaches. This research direction not only addresses the challenges of increasingly complex anomaly behaviors, high-dimensional data structures, and diverse temporal characteristics in modern cloud systems but also holds strategic significance for advancing reliability engineering and intelligent operations. By introducing multi-scale deep learning algorithms, future anomaly detection systems are expected to achieve higher detection accuracy, stronger adaptability to different scenarios, and more comprehensive knowledge extraction capabilities. These advancements will provide robust support for the secure operation of large-scale cloud infrastructures and drive cloud platforms toward greater intelligence, adaptability, and autonomy[5].

2. Related work

The research on anomaly detection in cloud services originated from the need to ensure the stability and availability of large-scale distributed systems. Traditional approaches mainly relied on rule-based strategies and statistical models. These methods typically identify data points that deviate from normal behavior by setting predefined thresholds, patterns, or distribution parameters. They are simple to implement and offer strong interpretability. However, as system structures become more complex, workloads become more dynamic, and anomaly types become more diverse, static rules struggle to adapt to rapidly changing service environments and often lead to missed detections or false alarms. Statistical modeling methods are limited when dealing with high-dimensional and nonlinear features. They also require extensive prior knowledge and manual intervention, making it difficult to capture potential contextual relationships. These limitations make traditional approaches inadequate for handling the complex anomaly patterns in modern cloud services, laying the groundwork for introducing deep learning methods with stronger representation capabilities[6].

In recent years, deep learning-based approaches have gradually become the mainstream direction. Through models such as convolutional neural networks, recurrent neural networks, and self-attention architectures, researchers have explored how to automatically extract high-dimensional representations from raw monitoring data for end-to-end anomaly detection. These methods can adapt to nonlinear, non-stationary, and high-dimensional time series data distributions and demonstrate significant advantages in detection accuracy and generalization ability[7]. At the same time, deep learning models enable cross-dimensional feature modeling and multimodal data fusion. As a result, anomaly detection is no longer limited to a single metric but can jointly analyze multiple dimensions, such as request load, resource utilization, and call chain relationships. However, traditional deep models often focus on fixed time granularity and have limited awareness of feature evolution across different time scales. This makes it difficult to meet the combined detection requirements for short-term fluctuations and long-term trends.

To address this challenge, multi-scale deep learning methods have emerged in the field of cloud service anomaly detection. These methods extract features across different time windows, frequency decomposition levels, or semantic abstraction layers to build multi-scale representations with hierarchical awareness. This enables the model to capture both local burst anomalies and global evolution patterns. Typical multi-scale architectures often combine convolution and attention mechanisms, as well as short-term modeling and long-term dependency modeling[8]. They improve detection sensitivity and enhance the model's adaptability to complex anomalies. In addition, multi-scale feature fusion strategies allow the model to achieve interaction and collaboration across different information levels, further improving anomaly localization and pattern recognition accuracy. Compared with traditional deep models, this approach is better suited for handling the multi-dimensional evolution of heterogeneous data in complex cloud environments and demonstrates stronger robustness and transferability in highly dynamic scenarios[9].

At the same time, recent research has focused on integrating anomaly detection with system structure modeling and contextual reasoning to improve interpretability and operational applicability. Some studies have introduced graph neural networks and causal inference mechanisms to incorporate service call dependencies, resource competition, and topology evolution into the modeling process. This allows anomaly propagation paths and potential root causes to be described from a structural perspective. Other research has attempted to enhance model representation capabilities through multi-task learning, self-supervised pretraining, and contrastive learning, enabling stable performance under conditions of data scarcity and distribution shifts. These advances show that cloud service anomaly detection is evolving from single-metric detection toward structured, context-aware, and knowledge-driven approaches. As the core technical pathway of this transformation, multi-scale deep learning not only improves detection performance but also provides a solid algorithmic foundation for system autonomy and intelligent operations[10].

3. Method

This study proposes a cloud service anomaly detection method based on multi-scale deep feature modeling. By integrating hierarchical time series modeling with contextual dependency analysis, the method enables precise identification and representation of abnormal behaviors in complex and dynamic environments. Its core idea is to model both short-term fluctuations and long-term trends simultaneously. Through multi-scale feature extraction, temporal dependency aggregation, and anomaly representation generation, it builds a deep detection framework with both global awareness and local sensitivity. Compared with traditional approaches, the proposed method can not only capture fine-grained local anomaly signals but also understand the evolutionary logic of system behaviors from a global perspective, thereby improving detection stability and robustness. The model architecture is shown in Figure 1.

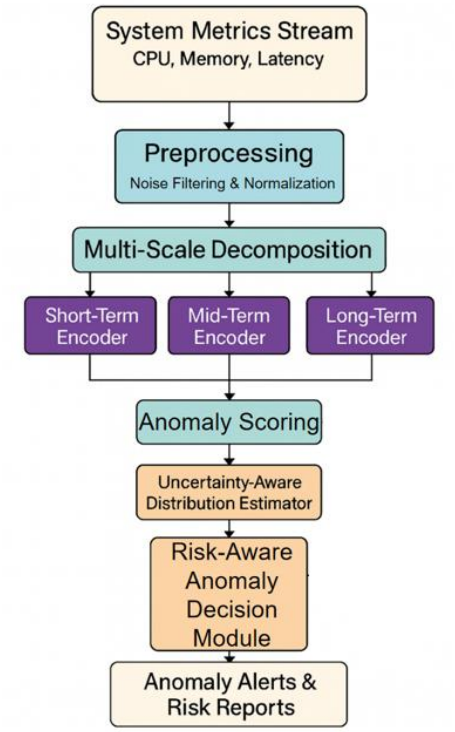
System Metrics Stream
CPU, Memory, Latency
Preprocessing
Noise Filtering & Normalization
Multi-Scale Decomposition
Short-Term Encoder
Mid-Term Encoder
Long-Term Encoder
Anomaly Scoring
Uncertainty-Aware Distribution Estimator
Risk-Aware Anomaly Decision Module
Anomaly Alerts & Risk Reports

Figure 1. Overall model architecture

First, the cloud service system's monitoring sequence in the time dimension is formalized as a multivariate time series $X = \{x_1, x_2, ..., x_T\}$ of length T, where the data vector $x_t \in R^d$ at each time step represents the system's multidimensional indicator state at time t. To capture the evolutionary characteristics at different time scales, we introduce a multiscale decomposition operator $F(\cdot)$ to decompose the original sequence into a set of subsequences $\{X^{(1)}, X^{(2)}, ..., X^{(K)}\}$ at K scales, where each subsequence retains the dynamic characteristics at a specific time granularity. This process can be formalized as:

$$X(k) = Fk(X) \quad , \quad k = 1,2,\ldots,K$$

After obtaining multi-scale subsequences, we construct a deep feature extractor for each scale to capture temporal dependencies and contextual information. Let the sequence input of scale k be $X^{(k)}$, and the deep temporal feature $H^{(k)}$ is extracted through the parameterized mapping function $\Phi_k(\cdot)$, which is defined as follows:

$$H^{(k)} = \Phi_k(X^{(k)}; \theta_k)$$

Where θ_k represents the learnable parameters of the k -th scale model. This feature extractor is usually combined with convolution, attention, or recursive structures to adapt to the dependency structure and pattern complexity at different scales.

To further fuse multi-scale features and achieve global context modeling, we designed a weighted aggregation mechanism to uniformly map representations of all scales into a shared semantic space. Specifically, let the fusion weight be a_k, then the global representation Z can be expressed as:

$$Z = \sum_{k=1}^{K} a_k H^{(k)} \ , \ \sum_{k=1}^{K} a_k = 1$$

This aggregation strategy can dynamically allocate attention among features at different temporal levels, enabling the model to maintain adaptive feature selection capabilities in complex scenarios.

After obtaining the fused representation Z, we introduce the anomaly representation function $\psi(\cdot)$ to map the context information into an anomaly score vector s. The process can be formalized as:

$$s = \psi(Z;\phi)$$

Where ϕ represents the parameters of the anomaly discrimination module, and each element of s corresponds to the anomaly intensity of the input sequence at different time steps. This module is designed to extract deviation patterns from the fused representation, thereby providing a separable feature basis for anomaly decision making.

Finally, to improve the model's discriminative ability in complex dynamic environments, this study introduces a context-constrained distribution modeling mechanism to model the probabilistic characteristics of anomaly scores to characterize the uncertainty of the system state. Assuming that the anomaly score follows a parameterized distribution $p(s \mid Z)$, the optimization objective can be expressed as maximizing the conditional likelihood:

$$\max_{\phi} \log p(s|Z)$$

This distribution modeling process enables the model to not only identify anomalies but also quantify their confidence levels, providing a theoretical basis for subsequent decision support and risk assessment.

In summary, the multi-scale deep learning method proposed in this study builds an end-to-end anomaly detection framework through multi-level mechanisms, including feature decomposition, temporal modeling, contextual fusion, and probabilistic characterization. This framework combines global structural awareness with local dynamic responsiveness, effectively addressing the challenges of complex anomaly types, diverse time scales, and multidimensional dependency structures in cloud service systems. It provides a scalable solution for ensuring stability and enabling intelligent operations in large-scale cloud computing environments.

4. Experimental Results

4.1 Dataset

This study uses the "Cloud Resource Usage Dataset for Anomaly Detection" as the dataset for algorithm validation. The dataset is derived from real resource usage monitoring in cloud environments and focuses on anomaly patterns caused by resource misuse or abnormal behaviors in multi-tenant settings. It contains multidimensional time-series metrics such as CPU utilization, memory usage, disk I/O, and network throughput. Anomalous samples are constructed based on resource overload scenarios, which allows for the evaluation of anomaly detection algorithms in terms of applicability and robustness in cloud service scenarios. The dataset is publicly available on Kaggle and has become a suitable benchmark in cloud service anomaly detection research due to its practicality and representativeness.

The dataset records system states using periodic sampling and provides continuous timestamp indexing with multivariate feature structures. It includes complete monitoring logs during both normal operation and anomaly injection phases, which support comparative analysis, segmented validation, and model training under a unified standard. In addition, the dataset is specifically designed with resource overload conditions and anomaly transition phases, simulating potential load fluctuations, sudden request spikes, and anomaly patterns triggered by service failures in real cloud systems. This design aligns well with multi-scale time series analysis methods and provides rich scenario support for evaluating a model's detection capability across different time granularities.

Using this dataset for model validation clearly demonstrates the advantages and potential of multi-scale deep learning in cloud anomaly detection. Training and validating the model on this dataset allows for the assessment of its ability to identify short-term burst anomalies, long-term trend anomalies, and hybrid anomalies. Experiments based on this dataset can showcase the model's performance in handling dynamic workloads, heterogeneous resource interactions, and complex dependency structures in cloud service systems. This provides valuable references for improving the reliability of cloud computing platforms.

4.2 Experimental Results

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

Table1: Comparative experimental results

Model	**Precision**	**Recall**	**F1-Score**	**AUC**
MEMTO[11]	0.842	0.801	0.821	0.912
DDMT[12]	0.865	0.790	0.826	0.918
MIXAD[13]	0.831	0.815	0.823	0.907
LATAD[14]	0.853	0.798	0.824	0.914
Ours	0.883	0.833	0.857	0.935

The experimental results show that the proposed multi-scale deep learning method demonstrates significant performance advantages in cloud service anomaly detection tasks. Compared with MEMTO, which is based on memory enhancement, and DDMT, which is based on diffusion modeling, the proposed method achieves higher Precision and Recall scores, reaching 0.883 and 0.833, respectively. This indicates that the model can maintain high detection accuracy while improving coverage of anomalous samples. As a result, it shows stronger discriminative power when facing sudden, non-stationary, and long-tail anomaly patterns in cloud systems. The multi-scale feature extraction and contextual fusion structure enable the model to perceive both short-term fluctuations and long-term trends, resulting in more comprehensive anomaly detection in complex and dynamic environments.

The improvement in the F1-Score further verifies the superiority of the proposed method in terms of overall detection capability. Compared with methods such as MIXAD and LATAD, the F1-Score of the proposed method improves to 0.857. This demonstrates that multi-scale modeling not only achieves breakthroughs in single metrics but also maintains a balance between precision and recall. Cloud service anomaly detection tasks are often affected by multidimensional interactions, complex dependency structures, and contextual drift. Traditional models usually fail to balance global trend awareness and local detail recognition. The proposed method addresses this problem by introducing cross-scale feature fusion and temporal attention modulation, significantly improving adaptability to diverse anomaly behaviors.

The improvement in the AUC metric indicates that the model has stronger robustness and generalization in terms of overall discrimination capability. Compared with recent methods such as DDMT and LATAD, the proposed model achieves an AUC of 0.935, showing that it maintains a high level of separation between normal and anomalous states across different thresholds. This performance benefits from the introduction of uncertainty modeling and distribution estimation modules, which characterize boundary samples probabilistically and calibrate their confidence. This reduces the risk of false positives and false negatives and improves the reliability and interpretability of the detection system. This capability is particularly important for handling soft anomalies, cascading anomalies, and dependency chain anomalies in cloud environments.

Overall, the experimental results fully validate the effectiveness and advancement of the proposed multi-scale deep learning method in cloud service anomaly detection. It not only outperforms representative recent methods across multiple key metrics but also demonstrates strong adaptability to complex distributions, multi-granularity temporal features, and various anomaly types. These advantages enable the model to provide stable and accurate anomaly detection support in dynamic multi-tenant cloud environments, laying a solid technical foundation for automated operations and intelligent resource scheduling.

This paper also conducted a comparative experiment on the hyperparameter sensitivity of multi-scale window length and decomposition layer number to anomaly capture ability. The experimental results are shown in Figure 2.

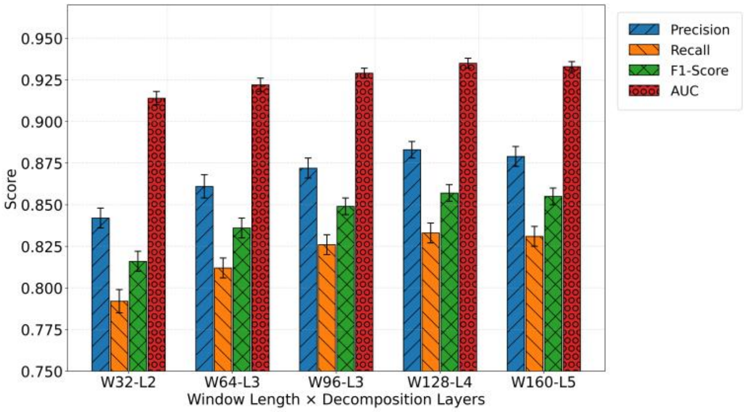

Figure 2. Hyperparameter sensitivity experiment of multi-scale window length and decomposition layer number on anomaly capture ability

The results show that as the window length and decomposition depth increase from W32-L2 to W128-L4, all four metrics improve steadily. Precision rises from 0.842 to 0.883, Recall from 0.792 to 0.833, F1-Score from 0.816 to 0.857, and AUC from 0.914 to 0.935. This trend indicates that a longer temporal receptive field and deeper scale decomposition can more effectively integrate short-term fluctuations with medium- and long-term evolution patterns. As a result, contextual relationships across different metrics are strengthened in a unified representation space, improving the separability of various types of anomalies in cloud services, including bursty, gradual, and context-dependent patterns.

When the configuration is further expanded to W160-L5, Precision and F1-Score show a slight decline to 0.879 and 0.855, while Recall and AUC remain mostly stable at 0.831 and 0.933. This suggests that longer windows and deeper decompositions introduce more redundancy and noise, causing diminishing returns and a mild over-smoothing effect. The ability to distinguish weak signals and boundary samples slightly decreases, but the overall threshold-independent discriminative capability remains stable. For cascading and soft anomalies in cloud environments, this reflects a structural trade-off between broader coverage and sharper discrimination.

From the perspective of metric synergy, the leading Precision combined with the steady rise of Recall results in continuous improvement of the F1-Score. This demonstrates the dual contribution of cross-scale fusion in reducing false positives and increasing recall. The monotonic increase of AUC up to W128-L4 shows that the ranking quality and distribution separation across all thresholds are continuously optimized. This aligns with representation calibration under uncertainty constraints. Multi-scale features achieve more consistent decision boundaries during the fusion stage, and anomaly score distributions become more concentrated in high-confidence regions, which stabilizes the global advantage in ROC space.

Considering the temporal variability of cloud service metrics and multi-tenant interference, W128-L4 can be regarded as the effective capacity limit of the receptive field and decomposition depth. It covers the key time windows of business periodicity and resource fluctuations while avoiding signal dilution caused by deeper decomposition. The performance peak at this configuration reflects strong structural alignment of cross-scale attention in capturing complex patterns such as request surges, link congestion, and resource jitter. The slight performance decline at W160-L5 suggests the need for adaptive calibration and regularization to suppress redundancy accumulation caused by excessive decomposition. This helps multi-scale representations maintain discriminative sharpness and generalization robustness in complex cloud scenarios.

This paper also evaluates the environmental sensitivity under resource interference (CPU jitter, memory pressure, I/O congestion) scenarios. The experimental results are shown in Figure 3.

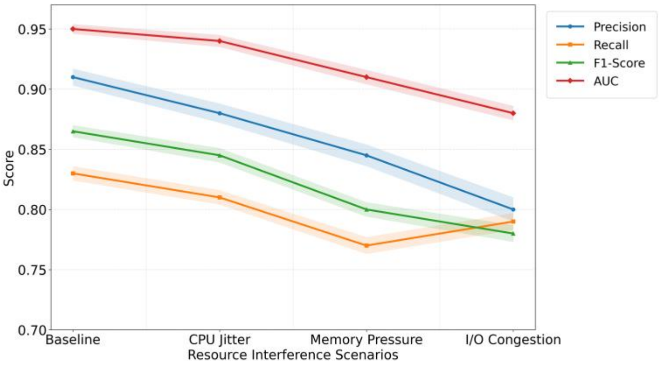

Figure 3. Environmental sensitivity assessment in resource interference scenarios (CPU jitter, memory pressure, I/O congestion)

As the intensity of resource interference gradually increases, precision drops from 0.910 to 0.800, showing an almost linear decline. This indicates that when CPU jitter, memory pressure, and I/O congestion overlap, the "sharpness" of the anomaly score distribution decreases, making false positive control more difficult. For a detector based on multi-scale representations, this trend suggests that local decisions in short and medium windows are more easily polluted by transient noise, reducing cross-scale consistency and weakening high-confidence decision boundaries. In particular, under I/O congestion, queue buildup, and tail latency diffusion broaden local peaks, causing the most significant loss of separability in the high-threshold region.

Recall decreases from 0.830 to 0.770 and then rebounds to 0.790 in the I/O scenario, showing a non-monotonic "drop-then-rise" trend. This reflects that under memory pressure, detecting weak, context-dependent anomalies is the most challenging. However, the persistent queuing and stability degradation caused by I/O congestion generate long-term structural signals that are easier to detect. For the proposed method, this indicates that cross-scale fusion requires stronger short-window enhancement and noise suppression when dealing with "sparse and short-lived" memory disturbances. In contrast, for "persistent and expanding" I/O anomalies, the method should increase the weight of long windows and enhance contextual aggregation along latency chains to maintain recall elasticity.

F1-Score decreases from 0.865 to 0.780, following the combined effect of Precision and Recall. The most significant decline occurs during the memory pressure phase, indicating that phase misalignment and inter-metric asynchrony caused by multidimensional resource contention most severely disrupt representation consistency. This suggests that multi-scale attention needs to adaptively adjust channel weights and temporal receptive fields under resource constraints, while uncertainty constraints should stabilize boundary samples. Otherwise, with fixed decomposition depth and window length, the fusion layer will amplify phase shifts across metrics, directly reducing overall discriminative capability.

AUC drops from 0.950 to 0.880, showing a clear monotonic degradation. This indicates that ranking quality and separability across all thresholds continuously weaken under the three types of interference, with I/O congestion having the strongest impact. Combined with the previous metrics, it can be inferred that distribution-level confidence calibration undergoes systematic shifts under resource interference. CPU jitter introduces high-frequency noise, memory pressure intensifies intermittent congestion, and I/O congestion increases tail latency. Together, these factors compress the margin between anomalous and normal states. Therefore, it is necessary to introduce adaptive calibration feedback at the methodological level to dynamically reallocate cross-scale weights and threshold ranges, ensuring that ranking boundaries maintain sufficient margin and stability under different interference mechanisms.

This paper also analyzes the environmental sensitivity of changes in observation granularity and sampling frequency to the ability to discriminate time series. The experimental results are shown in Figure 4.

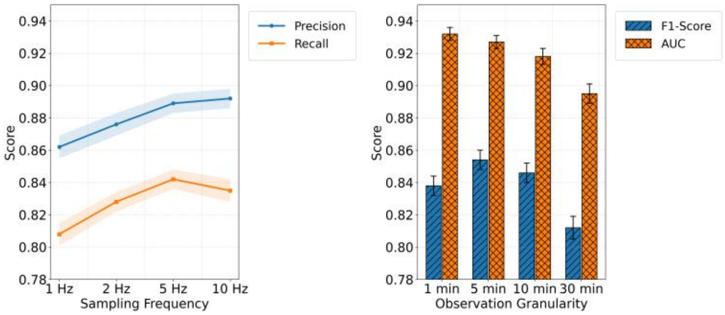

Figure 4. Experiment on the environmental sensitivity of changes in observation granularity and sampling frequency to temporal discrimination ability

In terms of sampling frequency, Precision increases steadily from 1 Hz to 10 Hz and approaches saturation (0.862 → 0.892). This indicates that higher temporal resolution allows multi-scale representations to capture finer-grained burst fluctuations and boundary transitions, forming sharper decision boundaries in high-threshold regions and reducing the probability of false positives. As the frequency increases, the short-window branch becomes more responsive, and its contribution to cross-scale fusion increases. This concentration of high-confidence anomaly samples leads to improved Precision.

Recall shows an inverted U-shaped trend, peaking at 5 Hz and slightly decreasing at 10 Hz (0.808 → 0.842 → 0.835). This suggests that in the mid-to-high frequency range, the model achieves the most comprehensive coverage of weak and context-dependent anomalies. However, when the frequency increases further, the proportion of transient noise and micro-disturbances rises. The short-window branch becomes more sensitive to noise, and if the fusion layer lacks sufficient denoising and uncertainty suppression, low-intensity anomalies may become dispersed or be masked by short-term pseudo-peaks, leading to a slight decrease in Recall. This indicates the need for adaptive calibration of channel weights and threshold ranges under high-frequency inputs to ensure that coverage is not dominated by noise.

In terms of observation granularity, the F1-Score reaches its highest value at medium granularity (5 min) with 0.854, then drops to 0.846 and 0.812 at coarser granularities (10 min and 30 min). This shows that excessive temporal aggregation weakens short-term structures and phase differences, which degrades the combined performance of precision and recall. Moderate aggregation can mitigate the jitter caused by high-frequency noise and isolated anomaly points, providing more stable contextual support for cross-scale attention in medium and long windows, which improves overall classification consistency. However, when the window is further extended, local anomaly patterns become overly smoothed. Although cross-scale consistency increases, it comes at the cost of reduced separability.

AUC decreases monotonically as granularity becomes coarser (0.932 → 0.895), indicating that ranking quality across all thresholds is highly sensitive to observation granularity. Coarser aggregation causes information blending and distortion, compressing the distribution gap between weak or short-term anomalies and normal states after aggregation, which reduces global separability in the ROC space. For cascading anomalies and tail latency diffusion scenarios in cloud services, this means that finer observation granularity should be prioritized to preserve edge structures. Cross-scale fusion at the medium-window level should be used to impose stability constraints. When storage or data collection costs require coarser granularity, uncertainty calibration and de-mixing modeling should be introduced to compensate for the loss of separability caused by temporal aggregation.

5. Conclusion

This study addresses the complexity of anomaly detection in cloud service environments and proposes a multi-scale deep learning-based detection method to tackle challenges such as strong system dynamics, diverse anomaly patterns, and complex contextual dependencies. By introducing multi-scale temporal feature modeling, cross-scale semantic fusion, and uncertainty estimation mechanisms, this work builds an end-to-end detection framework capable of capturing both local variations and global trends. Compared with traditional approaches, the proposed method shows significant advantages in anomaly pattern recognition, boundary sample discrimination, and global robustness, providing a new technical pathway for enhancing the stability and reliability of cloud computing systems. The findings demonstrate that the multi-scale architecture not only improves the discriminative accuracy of the model but also enhances its ability to perceive the evolution of anomaly contexts, offering a more detailed feature representation foundation for intelligent monitoring in complex scenarios.

From a technical perspective, the proposed method addresses the limitations of traditional approaches in single-scale feature extraction and discrimination capability by leveraging multi-granularity time windows and deep semantic modeling. The cross-scale feature fusion mechanism enables complementary integration of information from different temporal granularities, allowing the model to adapt to a wide range of anomaly behaviors, from short-term fluctuations to long-term evolutions. Additionally, the introduction of uncertainty estimation strategies enables probabilistic modeling of boundary samples and weak anomalies, reducing both false positive and false negative rates. The combination of multi-scale representation and confidence constraints provides important insights for building more intelligent and interpretable anomaly detection systems and offers a scalable theoretical framework for future research.

From an application perspective, this study holds strong practical significance. As the scale of cloud computing platforms continues to grow and business complexity increases, the types and impacts of system anomalies are expanding exponentially, making traditional static detection approaches insufficient for automation and service continuity requirements. The proposed multi-scale detection method maintains stable performance under different resource states, workload patterns, and service topologies, providing a unified modeling foundation for anomaly detection, root cause analysis, and fault prediction tasks. This capability not only delivers direct value to core domains such as data center operations, automated scheduling, and cloud resource management but can also be extended to applications such as financial risk monitoring, industrial IoT anomaly diagnosis, and intelligent manufacturing quality control, offering strong support for intelligent operations across industries.

Looking forward, the multi-scale deep detection framework proposed in this study can be further extended in several directions. On one hand, integrating federated learning and privacy-preserving mechanisms can enable collaborative anomaly detection across organizations and platforms, providing more secure and trustworthy operational solutions for multi-tenant cloud environments. On the other hand, future research can explore the integration of large models and knowledge graphs into anomaly semantic modeling, enhancing interpretability and scalability through structured knowledge and contextual reasoning. Moreover, combining this framework with adaptive resource scheduling, proactive fault tolerance mechanisms, and intelligent decision-making systems may enable a shift from "passive detection" to "proactive defense," laying a crucial technological foundation for cloud computing platforms and broader distributed intelligent systems.

References

[1] Ho, J., Jain, A. and Abbeel, P., "Denoising diffusion probabilistic models", Advances in Neural Information Processing Systems, vol. 33, pp. 6840-6851, 2020.

[2] Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W. and Pei, D., "Robust anomaly detection for multivariate time series through stochastic recurrent neural network", Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2828-2837, July 2019.

[3] Audibert, J., Michiardi, P., Guyard, F., Marti, S. and Zuluaga, M. A., "USAD: Unsupervised anomaly detection on multivariate time series", Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 3395-3404, August 2020.

[4] Liu Y, Pagliardini M, Chavdarova T, et al. The peril of popular deep learning uncertainty estimation methods[J]. arXiv preprint arXiv:2112.05000, 2021.

[5] Rahaman R. Uncertainty quantification and deep ensembles[J]. Advances in neural information processing systems, 2021, 34: 20063-20075.

[6] Ulmer D, Hardmeier C, Frellsen J. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation[J]. arXiv preprint arXiv:2110.03051, 2021.

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

[8] Bergmann, P., Fauser, M., Sattlegger, D. and Steger, C., "MVTec AD--A comprehensive real-world dataset for unsupervised anomaly detection", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9592-9600, 2019.

[9] Sohn, H., "Effects of environmental and operational variability on structural health monitoring", Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 365, no. 1851, pp. 539-560, 2007.

[10] Sohn, K., Lee, H. and Yan, X., "Learning structured output representation using deep conditional generative models", Advances in Neural Information Processing Systems, vol. 28, 2015.

[11] 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.

[12] Yang C, Wang T, Yan X. Ddmt: Denoising diffusion mask transformer models for multivariate time series anomaly detection[J]. arXiv preprint arXiv:2310.08800, 2023.

[13] Jacob, V., Song, F., Stiegler, A., Rad, B., Diao, Y. and Tatbul, N., "Exathlon: A benchmark for explainable anomaly detection over time series", arXiv preprint arXiv:2010.05073, 2020.

[14] Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X. and Xu, H., "Time series data augmentation for deep learning: A survey", arXiv preprint arXiv:2002.12478, 2020.