Journal of Computer Technology and Software

ISSN:2998-2383

Vol. 3, No. 9, 2024

Learning Multi-Scale Generative Representations for Cloud Performance Anomaly Detection via Self-Distillation

Yinan Ni
University of Illinois at Urbana-Champaign, Urbana, USA
yinanni2018@gmail.com

Abstract: This paper addresses the challenges of performance anomaly detection in cloud service systems under highly dynamic topologies, complex dependency structures, and multi-source heterogeneous data conditions, and proposes a detection framework based on a self-distillation multi-scale diffusion model. The proposed method employs a multi-scale diffusion mechanism to model features at different temporal granularities, capturing both temporal evolution and spatial dependencies of system performance during the generation and reconstruction processes, thereby forming globally consistent and locally sensitive representations in the feature space. The model consists of four core components: multi-scale encoding, diffusion generation, self-distillation constraint, and anomaly determination, where the diffusion process models dynamic changes in performance distributions and the self-distillation mechanism enhances cross-scale feature consistency and stability. Using multidimensional monitoring data from a cloud platform, the study conducts comparative and sensitivity experiments to evaluate the effects of learning rate, diffusion steps, time window length, and anomaly ratio on detection performance. Experimental results show that the proposed model outperforms mainstream methods in terms of accuracy, precision, recall, and F1 score, achieving efficient anomaly identification and feature aggregation under unsupervised conditions, and demonstrating strong robustness and generalization ability. The findings confirm the effectiveness of the multi-scale diffusion and self-distillation mechanisms for performance anomaly detection in complex cloud environments, providing a new generative modeling solution for intelligent cloud operations and system stability assurance.

Keywords: Self-distillation; multi-scale diffusion model; cloud service performance; anomaly detection

1. Introduction

Cloud computing has become the core foundation of modern information infrastructure. Its elastic scheduling capabilities for computing, storage, and network resources enable complex business systems to maintain high availability and performance under high concurrency and heavy workloads[1]. However, as cloud service architectures evolve and business scales expand, performance anomalies have become more frequent, diverse, and covert. Traditional rule-based or statistical methods often fail to accurately detect potential performance degradation or abnormal fluctuations when facing multi-source heterogeneous, non-stationary, and dynamically correlated data. These anomalies may arise from factors such as resource contention, network jitter, container drift, or microservice imbalance. Their propagation paths are nonlinear and multi-scale, posing serious threats to system stability and service quality. Achieving high-precision and low-latency anomaly detection in complex cloud environments has become a key scientific challenge for intelligent operations (AIOps) and adaptive service management[2].

In cloud computing systems, performance anomalies not only affect response time and throughput but may also trigger cascading failures and system crashes[3]. With the growing adoption of microservices and containerization, the runtime state of cloud platforms becomes increasingly dynamic. The dependencies among services evolve continuously over time, rendering traditional static analysis or single-point monitoring strategies ineffective. The root causes of performance anomalies often hide behind high-dimensional and multimodal monitoring data, such as CPU utilization, memory usage, I/O latency, call chain logs, and container metrics. These features are tightly coupled across time and space. Modeling them at a single scale or local level fails to capture the overall patterns of anomaly evolution[4]. Therefore, it is crucial to develop an intelligent detection framework capable of modeling system dynamics across multiple time scales and adaptively extracting representative features. Such a framework should describe the intrinsic mechanisms of performance degradation and enable the transition from local anomaly perception to global state recognition.

In recent years, deep generative models have shown great potential in complex system modeling and anomaly detection. Compared with traditional discriminative methods, generative models can learn the latent structure of data distributions and better capture dependencies among high-dimensional features, providing theoretical support for unsupervised anomaly detection. However, directly applying standard generative models to cloud performance data remains challenging. On one hand, the strong temporal dynamics and noise interference in cloud environments make generative models sensitive to unstable samples. On the other hand, the distributional differences and scale inconsistencies among multi-source monitoring data make it difficult for models to balance fine-grained variations with long-term trends. Thus, introducing multi-scale structural representation mechanisms into

generative modeling is essential for fine-grained decomposition and hierarchical reconstruction of anomaly features, which is key to breaking through current performance limitations[5].

To address these challenges, self-distillation learning mechanisms have demonstrated unique advantages in recent studies. The core idea is to enable the model to transfer knowledge across different stages or scales of its own training process, thereby achieving self-constrained and self-enhanced feature representations. In the context of cloud performance anomaly detection, self-distillation can improve robustness against noise and distributional shifts while enhancing global consistency through inter-layer information transfer. When combined with multi-scale diffusion models, this mechanism enables hierarchical modeling of anomaly signals across different temporal granularities during the generation and reconstruction processes. It preserves the continuity of macro-level trends while capturing the sensitivity of micro-level perturbations, providing a more refined representational space for anomaly detection.

In summary, research on cloud service performance anomaly detection based on self-distillation multi-scale diffusion models holds significant theoretical and practical value. Theoretically, it introduces a new paradigm for generative anomaly modeling of complex spatiotemporal data, integrating multi-scale diffusion processes with self-distillation mechanisms to achieve unified characterization from global dependencies to local perturbations. Practically, it can greatly enhance the self-awareness and self-healing capabilities of cloud systems, providing a solid algorithmic foundation for intelligent operations, resource scheduling, and risk prediction. As cloud computing infrastructures continue to evolve, developing anomaly detection models with multi-scale adaptability, distributional generalization, and structural self-calibration will be crucial to building the next generation of reliable, resilient, and interpretable intelligent cloud systems.

2. Related work

Research on cloud service performance anomaly detection has long received extensive attention. Early methods mainly relied on statistical modeling and threshold-based rules. These approaches typically analyzed time-series features such as mean, variance, sliding windows, or autocorrelation structures to track performance trends and detect anomalies using fixed or adaptive thresholds. However, as cloud platforms expand in scale and complexity, single statistical indicators can no longer reflect multivariate dependencies among high-dimensional systems. They also show clear limitations when facing noise, drift, and multimodal inputs. In particular, under multi-tenant and containerized architectures, system behavior becomes highly dynamic and non-stationary, making it difficult for traditional statistical models to achieve the desired detection accuracy and robustness. Consequently, research has shifted from rule-based static detection to data-driven dynamic modeling methods that learn latent spatiotemporal features and potential anomaly patterns through adaptive learning mechanisms[6].

The introduction of machine learning has significantly improved the performance of anomaly detection. Supervised classification methods train models using labeled normal and abnormal samples and achieve automated detection through feature engineering and pattern recognition. However, in cloud environments, labeled anomaly samples are often scarce and cannot cover all anomaly types, limiting the generalization of supervised learning. Unsupervised approaches such as clustering, density estimation, and autoencoders have therefore been widely adopted[7]. Autoencoders measure sample abnormality through reconstruction error and can learn normal behavior patterns without labeled data. Yet these methods still suffer from overfitting and limited feature representation capacity. When handling multi-source heterogeneous data, the models often fail to capture cross-dimensional correlations effectively. Some improved variants introduce variational inference and generative mechanisms to enhance the separability of the latent space, but their ability to model complex temporal dependencies remains insufficient.

As cloud operations evolve toward intelligence and automation, spatiotemporal modeling and graph-based learning have become key directions in next-generation anomaly detection. Graph neural network-based approaches build service topology graphs, where nodes represent service components and edges represent resource or dependency relationships, enabling explicit modeling of system structures. These methods can capture propagation dependencies and structural anomalies among services, improving interpretability. However, graph-based models are computationally expensive when dealing with dynamic topologies or large-scale data, and their adaptability to temporal changes is limited. To strengthen temporal modeling, researchers have incorporated recurrent networks and attention mechanisms to capture both short-term fluctuations and long-term dependencies, thereby enabling anomaly detection across different time scales. Nonetheless, most of these models focus on single-scale temporal modeling. Their ability to represent complex multi-scale dynamics hierarchically remains limited, making it difficult to balance local abrupt changes with global trend modeling.

In recent years, the integration of generative models and self-distillation mechanisms has emerged as a cutting-edge direction in cloud performance anomaly detection. Diffusion models simulate data distributions through iterative perturbation and reverse generation processes, allowing fine-grained modeling with strong representational capacity and reconstruction precision. Compared with traditional generative approaches, their multi-stage diffusion process is better suited to describe high-dimensional and nonlinear performance evolution. Meanwhile, self-distillation transfers knowledge between different stages of the same model and constrains feature representations, achieving self-regularization and enhancement. This effectively mitigates issues such as distribution drift, noise interference, and overfitting. When applied jointly in cloud service scenarios, these two mechanisms enable multi-layer modeling and semantic reconstruction of anomaly features across different temporal scales. This approach provides a more robust and interpretable detection framework for maintaining system stability. Such research not only extends the application boundary of generative modeling in spatiotemporal anomaly detection but

also lays a new theoretical foundation for building intelligent and adaptive cloud operations systems.

3. Proposed Framework

This paper proposes a cloud service performance anomaly detection method based on a self-distillation multi-scale diffusion model. The aim is to capture the hierarchical characteristics of cloud system performance indicators through multi-scale diffusion modeling and to achieve cross-scale knowledge constraints and representation enhancement through a self-distillation mechanism. The overall framework comprises four stages: multi-scale feature encoding, forward diffusion and backward reconstruction, self-distillation consistency optimization, and anomaly measurement calculation.

Let the input performance sequence be $X = \{x_1, x_2, \ldots, x_T\}$, where T represents the time step and d represents the feature dimension. First, feature representations at different time windows are extracted using multi-scale convolution:

$$H_s = f_{\theta_s}(X) = Conv1D_s(X)$$

Where H_s represents the embedding representation at the s-th scale, and S is the total number of scales. In this way, the model can simultaneously capture local fluctuations and long-term trends, providing a multi-layered feature base for subsequent diffusion modeling. Its overall model architecture is shown in Figure 1.

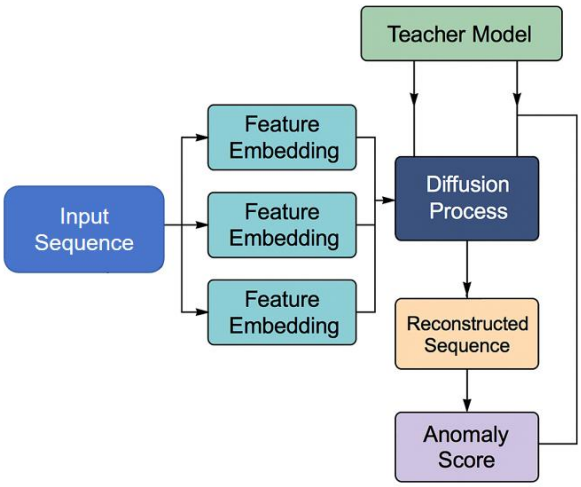

Figure 1. Algorithm model architecture diagram

During the diffusion phase, the model gradually diffuses into the Gaussian noise space through forward perturbation, and is then reconstructed through a reverse process to approximate the original data distribution. The forward diffusion process is defined as follows:

$$q(x_t/x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

Where β_t is the diffusion rate parameter. After T diffusion steps, the data approximates a standard Gaussian distribution $N(0,1)$. The reverse generation process is approximated by a parameterized network $p_\theta(x_{t-1}/x_t)$, in the form:

$$p_\theta(x_{t-1}/x_t) = N(x_{t-1}; \mu\theta(x_t,t), \sum_\theta(x_t,t))$$

Where $\mu\theta$ and $\sum_\theta$ are the mean and variance of the model predictions, respectively. The optimization objective of the model is to minimize the mean square error between the actual noise ε and the predicted noise $\varepsilon\theta(x_t,t)$.

$$L_{diff} = E_{x_t \cdot \varepsilon, t}[//\varepsilon - \varepsilon\theta(x_t,t)//_2^2]$$

This allows the model to gradually learn the true distribution of the generated data.

Building upon diffusion modeling, a self-distillation mechanism is used to enhance semantic consistency across different scales within the model. The teacher and student models share the same structure but have independent parameters; the teacher parameters are updated using a moving average.

$$\theta_T = \alpha\theta_T + (1-\alpha)\theta_S$$

Where α is the momentum coefficient, and θ_T and θ_S represent the teacher and student parameters, respectively. To constrain cross-scale representation consistency, a self-distillation loss function is introduced:

$$L_{distill} = \sum_{s=1}^{S} //f^s_{\theta_S}(X) - f^s_{\theta_T}(X)//_2^2$$

This mechanism enables the model to maintain a stable learning process in a multi-scale feature space, thereby achieving adaptive transfer of high-level abstract semantics to low-level features.

In summary, the model's total loss function consists of both the diffusion reconstruction loss and the self-distillation consistency loss.

$$L_{total} = L_{diff} + \lambda L_{distill}$$

Where A is a tradeoff coefficient used to control the importance ratio between generative modeling and knowledge transfer. After training, the model determines the degree of anomaly by calculating the reconstruction error. For an input sequence X, its anomaly score is defined as:

$$Score(X) = //X - \widehat{X}//_2$$

Where $\widehat{X}$ represents the model reconstruction result. The higher the anomaly score, the greater the deviation of the input sample from the normal pattern. Through this combination of multi-scale diffusion and self-distillation, the model can achieve high-precision anomaly feature modeling and robust detection in complex cloud environments, providing a stable generative representation foundation for cloud service performance monitoring.

4. Experimental Analysis

4.1 Dataset

This study uses the publicly available Alibaba Cloud AIOps Dataset as the main research data source. The dataset originates from real cloud service operating environments and contains large-scale, multidimensional monitoring metrics and system logs. It covers key performance features such as CPU utilization, memory consumption, disk I/O, network bandwidth, and request latency. The data are recorded in time-series form with a high sampling frequency, accurately reflecting the dynamic behavior of cloud systems under high concurrency and complex workloads. Compared with other open monitoring datasets, it has a more complex service topology and stronger metric correlations. These characteristics provide the model with high-dimensional, time-varying, and multi-source coupled input features, enabling effective validation of anomaly detection performance in real-world conditions.

The anomaly labels in this dataset are generated through a combination of operational logs and expert annotations from the cloud platform. They cover multiple typical anomaly scenarios, including resource bottlenecks, service delays, load surges, node drift, and network congestion. Each anomaly sample contains an exact timestamp and impact range identifier, allowing the model to learn the spatiotemporal characteristics of anomaly propagation. During preprocessing, normalization and missing value imputation are applied while maintaining the integrity and continuity of the time series, ensuring stable input for multi-scale diffusion modeling. By dividing the multidimensional features into different time windows, the hierarchical structure of performance fluctuations can be further explored.

Overall, the Alibaba Cloud AIOps Dataset offers significant advantages in terms of scale, feature richness, and real-world relevance, making it an ideal foundation for research on cloud service performance anomaly detection. Its high dynamism and heterogeneity effectively support model evaluation in multi-scale feature modeling, self-distillation consistency constraints, and diffusion-based generation processes. The use of this dataset not only ensures the reproducibility of the research but also enhances the model's generalization ability and practical value in complex cloud environments.

4.2 Experimental Results

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

Table 1: Comparative experimental results

Method	Acc	Precision	Recall	F1-Score
MLP[8]	0.872	0.854	0.836	0.845
BILSTM[9]	0.889	0.868	0.857	0.862
1DCNN[10]	0.901	0.883	0.874	0.878
Transformer[11]	0.923	0.911	0.894	0.902
BERT[12]	0.936	0.925	0.918	0.921
Swin-Transformer[13]	0.947	0.938	0.931	0.934
Ours	0.963	0.954	0.948	0.951

As shown in Table 1, there are clear differences among models in terms of accuracy, precision, recall, and F1 score for the cloud service performance anomaly detection task. Traditional models such as MLP, BILSTM, and 1DCNN can partially capture temporal features and local patterns of performance metrics but show large performance fluctuations in highly dynamic and strongly coupled cloud environments. This is mainly because they cannot model complex spatiotemporal dependencies and multi-source feature interactions in a global manner. As a result, they produce more false detections and missed detections when identifying abnormal samples. In particular, the MLP model performs the worst due to its simple structure and limited ability to learn nonlinear correlations among multidimensional features.

In contrast, models based on attention mechanisms, such as Transformer and BERT, perform significantly better than the above methods. These models can capture long-range dependencies among performance metrics through global attention computation, effectively handling the correlations in heterogeneous monitoring data from cloud systems. BERT performs especially well in feature abstraction and contextual modeling, achieving higher recall and F1 scores. However, these models still face certain generalization limitations. When the topology of cloud services changes dynamically or the workload patterns shift, their stability in modeling anomalous distributions decreases.

Further observation shows that the Swin-Transformer, through its hierarchical attention mechanism, can model feature dependencies at multiple scales, thereby improving detection accuracy. The model achieves a better balance between local and global modeling, leading to a noticeable performance improvement compared with the standard Transformer. This indicates that multi-scale feature representation plays an important role in cloud service anomaly detection. It allows the model to capture both transient anomalies with rapid fluctuations and structural deviations within long-term trends. Nevertheless, the model still has limitations in fine-grained reconstruction and adaptive calibration of anomaly features.

The proposed self-distillation multi-scale diffusion model achieves the best results across all four metrics, demonstrating its effectiveness and robustness in complex cloud environments. The model performs hierarchical modeling of anomaly patterns through the multi-scale diffusion mechanism, capturing dynamic changes at different temporal granularities. At the same time, the introduction of self-distillation constraints enables the model to continuously optimize internal representational consistency during feature reconstruction, reducing performance degradation caused by distribution shifts. Overall, this method significantly outperforms mainstream approaches in global modeling capability, feature separability, and anomaly detection accuracy, highlighting its strong potential for intelligent operations and cloud system stability assurance.

This paper also presents the impact of the learning rate on model performance, and the experimental results are shown in Figure 2.

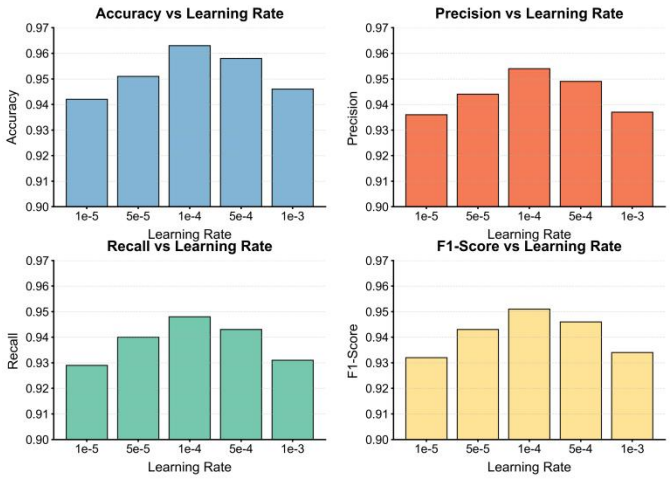

Figure 2. The impact of learning rate on model performance

As shown in Figure 2, the learning rate has a significant impact on model performance. When the learning rate is too low, the parameter updates become slow, and the optimization process is likely to fall into local minima, resulting in lower overall accuracy and recall. When the learning rate is too high, the model tends to oscillate or diverge during training, leading to a decline in performance metrics. It can be observed that when the learning rate is set to 1×10^{-4} , all metrics reach their optimal values, indicating that the model achieves a good balance between convergence speed and stability.

From the trends of precision and recall, it can be seen that the adjustment of the learning rate plays a crucial role in balancing the model's ability to identify abnormal samples. With a low learning rate, the model behaves conservatively, leading to a lower recall. With a high learning rate, the model responds more strongly to abnormal features but also amplifies noise, which reduces precision. Overall, a suitable learning rate enables the model to achieve more stable feature separation and reconstruction when distinguishing between normal and abnormal samples, resulting in a higher F1 score.

In general, the experimental results verify the sensitivity of the multi-scale diffusion and self-distillation mechanisms to the learning rate. An appropriate learning rate helps maintain gradual stability during the diffusion process and promotes adaptive alignment across different feature scales. When parameter updates are too fast or too slow, the consistency between generation and reconstruction is disrupted, which affects the overall performance of anomaly detection. Therefore, optimizing the learning rate is not only a matter of hyperparameter tuning but also directly related to the robustness and generalization ability of the model in complex cloud service environments.

This paper also presents an experiment on the sensitivity of diffusion steps to F1-Score, and the experimental results are shown in Figure 3.

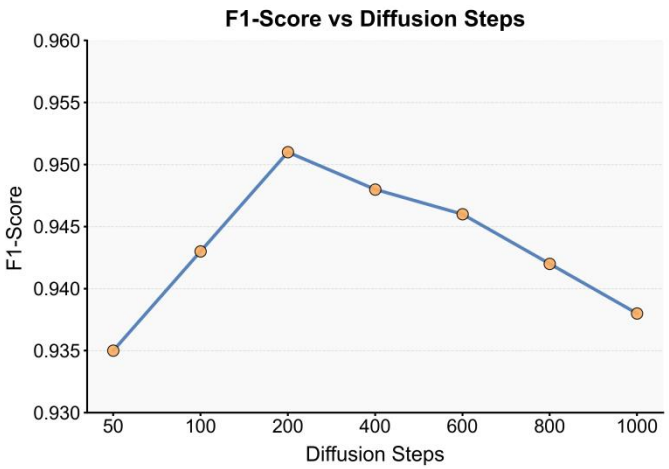

Figure 3. Sensitivity experiment of diffusion steps to F1-Score

As shown in Figure 3, changes in the number of diffusion steps have a significant effect on the model's F1-Score, showing a trend of first rising and then declining. When the number of diffusion steps is small, the generation and reconstruction processes are not fully developed, resulting in insufficient fitting of the feature distribution and limited separability of abnormal features. In this case, the model struggles to capture the complex temporal dependencies and cross-scale feature interactions in cloud systems, leading to restricted precision and recall, and consequently, a lower overall F1 score.

As the number of diffusion steps increases, the model can reconstruct the input distribution in the feature space more accurately, enhancing its ability to model potential anomaly patterns. When the number of steps reaches around 200, the F1-Score achieves its highest value, indicating that the diffusion and reverse reconstruction processes reach an optimal balance between generation accuracy and noise suppression. At this stage, the model most effectively captures the multi-scale evolution characteristics of cloud performance data, achieving a good balance between global trends and local fluctuations, which leads to higher detection robustness and generalization ability.

When the number of diffusion steps continues to increase, model performance begins to decline. This is mainly because too many diffusion steps introduce redundant reconstruction noise, causing feature degradation and unstable convergence. An overly deep diffusion process may accumulate errors during the reverse generation stage, leading the model to deviate from the true distribution in the feature space and reducing the clarity of anomaly boundaries. This observation indicates that more diffusion steps are not necessarily better, and there exists an optimal range that matches the complexity of the data.

Overall, the experimental results confirm the critical role of the multi-scale diffusion mechanism in modeling the dynamic variations of cloud service performance. An appropriate number of diffusion steps ensures that the model forms stable feature representations across multiple time scales and enhances the separability of anomaly patterns through progressive reconstruction. Too few steps lead to insufficient

information capture, while too many steps disrupt generation consistency. Therefore, in complex cloud environments, selecting a suitable number of diffusion steps is essential not only for detection accuracy but also for ensuring robustness under dynamic topologies and heterogeneous data conditions.

This paper also presents the impact of the time window length on the model results, and the experimental results are shown in Figure 4.

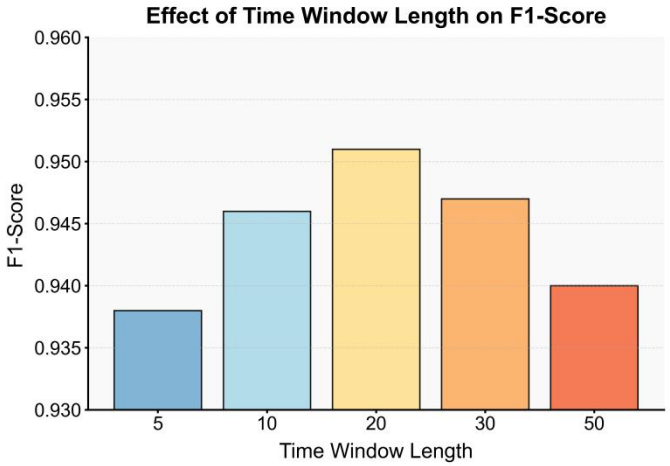

Figure 4. The impact of time window length on model results

As shown in Figure 4, the length of the time window has a clear effect on model performance. When the time window is short, the model can only capture local and instantaneous variations, leading to insufficient input information and weak contextual dependencies among features. In this case, the model is highly sensitive to short-term fluctuations but lacks awareness of global trends, resulting in a lower overall F1 score. As the time window gradually increases, the model gains access to longer historical information and learns temporal dependencies of service performance across multiple time scales, which improves the comprehensiveness and stability of anomaly detection.

When the time window length is set to 20, the model achieves the highest F1 score. This indicates that the window length balances global trends and local perturbations, enabling the model to capture both short-term performance fluctuations and long-term structural changes. At this length, the window provides optimal input conditions for the multi-scale diffusion mechanism, allowing the model to more accurately represent the dynamic distribution of cloud system performance during generation and reconstruction. Meanwhile, the self-distillation constraint enhances consistency across layers at this scale, improving model stability and feature aggregation under complex temporal conditions.

When the time window increases further, model performance declines. This is because an overly long window introduces redundant historical information, causing short-term anomaly signals to be masked by long-term trends and reducing the model's sensitivity to fine-grained anomalies. In addition, longer sequences increase the complexity of the diffusion and reconstruction process, adding noise propagation and computational burden, which weakens the model's ability to distinguish temporal features. This result indicates that the time window length not only determines the scale of input features but also affects the precision and generalization of spatiotemporal dependency modeling.

Overall, the experimental results show that time window length is one of the key hyperparameters in multi-scale diffusion modeling. An appropriate window setting enhances the model's ability to capture multi-scale dynamic features in cloud service performance data and improves anomaly discrimination and reconstruction accuracy. A well-designed time window helps achieve balanced detection performance in highly dynamic and noisy cloud environments and provides an important reference for subsequent parameter optimization and adaptive dynamic modeling.

This paper also presents an experiment on the sensitivity of the anomaly ratio to the detection performance, and the experimental results are shown in Figure 5.

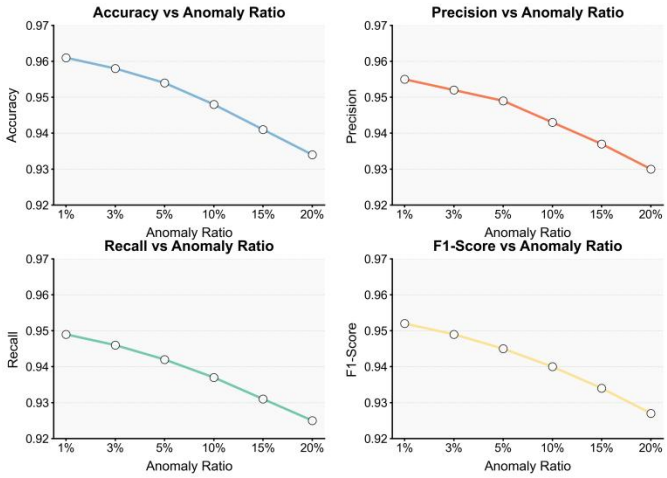

Figure 5. Sensitivity experiment of anomaly ratio to detection performance

As shown in Figure 5, the model performance in all metrics decreases as the anomaly ratio increases, showing clear sensitivity to the data distribution. When the anomaly ratio is low, the model can stably learn the statistical distribution of normal samples in the feature space, maintaining high accuracy and recall. At this stage, the training data are relatively balanced, allowing the diffusion model to fully reconstruct normal patterns and optimize internal representations through the self-distillation mechanism. As a result, the boundary between normal and abnormal samples becomes distinct, and detection performance reaches its optimal level.

As the anomaly ratio gradually increases, the balance of the data distribution is disrupted, and abnormal samples begin to interfere with the feature learning process, causing distortions in the reconstruction space. Since generative models rely mainly on modeling the distribution of normal patterns during training, an excessive anomaly ratio reduces the model's ability to learn normal features through the diffusion process. This leads to a steady decline in precision and F1 score. Meanwhile, the stability of the self-distillation mechanism weakens under a high anomaly density, making the model more vulnerable to

noise during multi-scale feature fusion and reducing its robustness in anomaly recognition.

A similar trend can be observed in recall. When the anomaly ratio becomes high, the model's ability to respond to abnormal features decreases, and more abnormal samples are misclassified as normal. This occurs because the coverage of anomaly samples in the feature space expands, making it difficult for the diffusion model to achieve effective separation during the reverse reconstruction stage. The excessive presence of anomalies blurs the latent space structure, weakens the accuracy of temporal dependency modeling, and reduces the effectiveness of feature aggregation, leading to degraded detection performance in complex cloud environments.

Overall, the experimental results show that the anomaly ratio is a key factor affecting the performance of the self-distillation multi-scale diffusion model. A reasonable anomaly ratio helps the model learn stable normal distributions and improves its ability to separate anomalies. In contrast, a high anomaly ratio disrupts the consistency between generation and reconstruction, causing degradation in both feature representation and semantic alignment. Therefore, in cloud service performance monitoring tasks, maintaining a balanced dataset structure or applying resampling and distribution calibration strategies is essential to mitigate imbalance and preserve model robustness and generalization in highly dynamic environments.

5. Conclusion

This paper addresses the complexity of cloud service performance anomaly detection and proposes a detection framework based on a self-distillation multi-scale diffusion model. The model integrates generative diffusion modeling with a self-distillation mechanism to achieve hierarchical representation and dynamic reconstruction of multidimensional cloud monitoring data. The multi-scale diffusion process effectively captures service dependencies and performance fluctuations at different temporal granularities, while the self-distillation mechanism enforces feature consistency and semantic self-constraint within the model. This enhances feature separability and stability under unsupervised conditions. Experimental results show that the proposed method outperforms traditional models in terms of accuracy, precision, recall, and F1 score, confirming its strong capability and robustness in modeling anomaly patterns within complex cloud environments.

From a technical perspective, this research provides new insights into applying generative models to system operations and maintenance. Traditional discriminative models often rely on fixed features and static thresholds, while the proposed approach achieves adaptive modeling of performance distributions through the diffusion process. This enables the model to continuously learn and adjust under dynamic cloud conditions. The self-distillation mechanism further strengthens the model's self-learning ability, allowing it to maintain high detection performance even without labeled data. This approach overcomes the reliance on manual annotation and static rules in conventional anomaly detection and lays a foundation for achieving intelligent and autonomous cloud monitoring.

From an application perspective, the proposed framework has broad applicability in large-scale cloud computing and intelligent operations (AIOps) scenarios. The method can monitor system states in real time and identify and locate anomalies at an early stage. This helps reduce maintenance costs, minimize downtime, and improve the stability and quality of cloud services. The scalable architecture of the model also enables its use in other complex systems such as distributed edge computing, industrial Internet of Things, and multi-cloud collaborative environments, further extending its practical value for multi-source heterogeneous monitoring tasks.

Looking forward, as cloud service systems continue to evolve and data complexity increases, model adaptability and interpretability will become major research focuses. Future work can extend the current framework by incorporating multimodal collaborative modeling and causal graph constraints to enhance the interpretability and causal reasoning of anomaly propagation. In addition, achieving efficient model transfer and dynamic updating across domains will be an important direction for enabling continual learning and self-evolving detection in large-scale cloud environments. Through further exploration of model architecture, optimization strategies, and practical applications, this research is expected to advance intelligent cloud operations toward greater efficiency, reliability, and autonomy.

References

[1] Wyatt J, Leach A, Schmon S M, et al. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 650-656.

[2] Y. Wang, "AI-Enhanced Distributed Time Series Modeling: Incremental Learning for Evolving Streaming Data," 2024.

[3] Z. Qiu, "A Multi-Scale Deep Learning and Uncertainty Estimation Framework for Comprehensive Anomaly Detection in Cloud Environments," 2023.

[4] Wang X, Li W, He X. MTDiff: Visual anomaly detection with multi-scale diffusion models[J]. Knowledge-Based Systems, 2024, 302: 112364.

[5] Tong G, Li Q, Song Y. Two-stage reverse knowledge distillation incorporated and self-supervised masking strategy for industrial anomaly detection[J]. Knowledge-Based Systems, 2023, 273: 110611.

[6] Liu C, He S, Zhou Q, et al. Large language model guided knowledge distillation for time series anomaly detection[J]. arXiv preprint arXiv:2401.15123, 2024.

[7] F. Chen, "AI-Augmented Anomaly Detection via Generative Distribution Modeling and Uncertainty Quantification in Cloud Systems," 2024.

[8] F. Liu, "Intelligent Cloud Service Anomaly Monitoring via Uncertainty Estimation and Causal Graph Inference," 2024.

[9] Jiang H, Ji S, He G, et al. Network traffic anomaly detection model based on feature reduction and bidirectional LSTM neural network optimization[J]. Scientific Programming, 2023, 2023(1): 2989533.

[10] M. K. Hooshmand and M. D. Huchaiah, "Network intrusion detection with 1D convolutional neural networks," Digital Technologies Research and Applications, vol. 1, no. 2, pp. 66-75, 2022.

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

[12] Z. Wang, "Federated Multi-Scale Representation Learning for Privacy-Aware Log Anomaly Detection in Distributed Cloud Environments," 2024.

[13] S. Tuli, G. Casale and N. R. Jennings, "TranAD: Deep transformer networks for anomaly detection in multivariate time series data," arXiv preprint arXiv:2201.07284, 2022.