Transactions on Computational and Scientific Methods | Vo. 4, No. 10, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Intelligent Cloud Service Anomaly Monitoring via Uncertainty Estimation and Causal Graph Inference

Feng Liu
Stevens Institute of Technology, Hoboken, USA
genflandrew@gmail.com

Abstract: This paper addresses the challenges of complex dependencies, diverse anomaly patterns, and the coexistence of label scarcity and pseudo-label noise in cloud service environments by proposing an anomaly monitoring method that integrates uncertainty estimation with causal inference. The method models cloud service interactions as dependency graphs, extracts cross-temporal and cross-service contextual features through graph embedding, and applies uncertainty estimation to provide confidence intervals for boundary samples, thereby mitigating prediction instability caused by short-term fluctuations and noise. On this basis, a causal inference mechanism is introduced to suppress spurious correlations, while causal consistency constraints enhance the identification of complex anomalies under cross-tenant coupling and multi-tenant interference. The optimization objective jointly incorporates classification loss, contrastive loss, and uncertainty calibration to balance threshold performance and global ranking stability. Experiments systematically analyze hyperparameter sensitivity, environmental sensitivity, and data sensitivity, including the effects of prediction head depth and width on boundary confidence, the trade-off between false positives and false negatives under varying interference and coupling, and the impact of label scarcity and pseudo-label noise on causal accuracy. Results show that the proposed method outperforms existing public models on metrics such as AUC, F1-Score, Precision, Recall, and AUROC, and maintains robustness and stability under complex interference and high-noise conditions, fully validating its effectiveness and applicability in cloud service anomaly monitoring tasks.

Keywords: Cloud service anomaly monitoring; uncertainty estimation; causal inference; data sensitivity

1. Introduction

In the rapid development of cloud computing environments, service architectures are becoming increasingly complex. Business requests span multiple service modules and resource units, forming dynamically coordinated and highly coupled dependency networks. This complexity brings flexibility and scalability but also introduces potential risks of anomalies. When network jitter, resource bottlenecks, or external load shocks occur, they often spread rapidly through chain reactions, turning local anomalies into global failures. Such events pose serious threats to system stability and service continuity. Therefore, achieving efficient and accurate anomaly monitoring in complex service dependency environments has become a core challenge in ensuring the reliability of cloud computing systems. Traditional statistical methods or threshold-based detection mechanisms often fail in these scenarios, as they cannot adapt to multi-source, high-dimensional, and dynamically changing feature distributions[1].

At the same time, anomalies in cloud service environments are not caused by a single factor but by the interplay of multiple elements. Load imbalance, resource contention, abnormal invocation paths, and external disturbances can all lead to high uncertainty in time-series metrics. Sole reliance on deterministic modeling often fails to capture confidence boundaries, making anomaly judgments prone to distortion in borderline cases. Introducing uncertainty estimation helps quantify the credibility of model predictions and provides confidence intervals for anomaly detection. This reduces misjudgments caused by noise or occasional fluctuations. Especially in cross-tenant and cross-service environments, explicitly indicating prediction uncertainty allows operators to make robust decisions under multi-task and multi-scenario interference, enhancing system robustness and interpretability[2].

However, uncertainty estimation alone is insufficient to fully reveal the roots of complex anomalies. Cloud service dependencies exhibit causal structures. Anomalies often do not occur in isolation but propagate downstream from upstream deviations. The introduction of causal inference provides effective tools to understand the generative mechanisms of anomalies. By constructing causal graph models, it becomes possible to distinguish superficial correlations from true causal relationships. This enables more precise identification of the sources and propagation paths of anomalies[3]. Compared with simple correlation analysis, causal inference avoids the misleading effects of spurious dependencies. Even when faced with intertwined multidimensional metrics and potential confounding factors, the detection system can maintain accurate diagnostic capabilities. This plays a key role in reducing false positives and false negatives in large-scale systems[4].

From the perspective of academic research, integrating uncertainty estimation with causal inference enables the construction of a comprehensive anomaly detection framework that combines confidence quantification with causal interpretation. On one hand, uncertainty estimation provides stability under complex inputs, ensuring that anomaly detection not only outputs results but also indicates their confidence intervals. On the other hand, causal inference assigns interpretable causal mechanisms to the results, allowing detection to go beyond surface-level findings and trace anomalies back to their generative logic. The combination of the two is expected to break through the limitations of existing methods and provide stronger theoretical foundations and methodological pathways for anomaly monitoring in cloud services. This represents not only a technological innovation but also a significant extension of intelligent operations in complex systems[5].

From the perspective of application value, anomaly monitoring methods that integrate uncertainty estimation and causal inference carry far-reaching implications for enhancing observability and controllability in cloud computing environments. They can detect potential risks early, reduce the probability of large-scale outages, and help operations teams locate problem sources more efficiently. This shortens troubleshooting time and lowers maintenance costs. In future intelligent and automated operation systems, such methods are expected to become a cornerstone for supporting high service availability and continuous business growth. Furthermore, this research direction has strategic significance in safeguarding the stability of critical infrastructure, improving the rationality of resource scheduling, and promoting the sustainable development of the cloud computing ecosystem[6].

2. Related work

In the study of anomaly detection in cloud services, traditional methods often rely on statistical modeling and threshold setting. They attempt to identify potential risks by monitoring fluctuations in key performance indicators. These methods are simple to implement and suitable for early scenarios of single service or resource monitoring. However, they show clear limitations in complex environments with multiple tenants, multiple tasks, and dynamic loads. Fixed thresholds cannot adapt to the highly dynamic nature of cloud environments, leading to many false alarms and missed detections. Statistical modeling also struggles to capture the diversity of anomaly patterns, making it difficult to handle complex anomalies under cross-dimensional dependencies. These limitations have driven both academia and industry to shift toward machine learning and deep learning methods, which can improve flexibility and adaptability through data-driven feature modeling[7].

With the introduction of machine learning, both supervised and unsupervised approaches have been widely applied to anomaly detection in cloud services. Supervised methods train classifiers using large amounts of labeled data and can achieve high detection accuracy in specific scenarios. However, in real cloud environments, the high cost of labeling and the scarcity of anomaly samples limit their generalization performance. In contrast, unsupervised methods use clustering, reconstruction errors, or latent space modeling to automatically discover anomaly patterns. They are better suited for high-dimensional and low-label environments. These methods improve automation to some extent, but still face challenges such as unstable feature selection and a lack of interpretability of detection results. Under complex dependencies, purely data-driven methods often fail to distinguish correlation from causality, which makes root cause analysis insufficient.

In recent years, the development of deep learning has provided new ideas for anomaly detection. Through temporal modeling, attention mechanisms, and graph neural networks, researchers have tried to capture cross-service dependencies and multi-dimensional interaction features, improving the representation of anomaly patterns[8]. Deep temporal models can learn multi-granularity time dependencies and adapt to short-term fluctuations and long-term trends in cloud environments. Graph-based modeling can leverage topological information between services and extend anomaly detection to the level of global dependency networks. These methods overcome the limitations of traditional approaches in high-dimensional and dynamic environments. However, they also bring challenges such as high computational complexity, large training costs, and insufficient explanation of anomaly causes. In practice, deep learning models provide higher detection accuracy, but their black-box nature limits usability for operators in anomaly tracing and decision-making.

Against this background, more research has started to explore the integration of uncertainty estimation and causal inference to address the shortcomings in confidence quantification and causal interpretation. Uncertainty estimation provides confidence boundaries for model predictions, enabling detection systems to output judgments with credibility in complex scenarios. This improves robustness and reliability in anomaly discovery. Causal inference offers theoretical support for understanding anomaly propagation paths under complex dependencies. It helps distinguish true causal relationships from surface correlations. The combination of the two not only enhances detection performance but also enables interpretable anomaly localization, offering more practical value for intelligent operations. Therefore, how to efficiently integrate these two methods in cloud service environments to build an anomaly monitoring framework with robustness and interpretability has become both a key research trend and a critical direction for future development[9].

3. Method

This study introduces a cloud service anomaly monitoring method that integrates uncertainty estimation with causal inference. The core idea is to jointly model prediction confidence and causal structures in complex service dependencies to achieve robust anomaly detection and interpretable anomaly localization. The method first represents cloud service interactions as multidimensional time-series features and applies probabilistic modeling to quantify uncertainty, allowing the detection phase to output both results and their confidence intervals. It then incorporates a causal inference mechanism, using causal graphs to represent and reason about service dependencies, distinguishing superficial correlations from underlying causalities to improve interpretability and localization accuracy. The overall framework is optimized jointly so that uncertainty estimation and causal structure learning complement each other, enabling anomaly monitoring that is both more robust and more interpretable in dynamic and complex environments. The model architecture is shown in Figure 1.

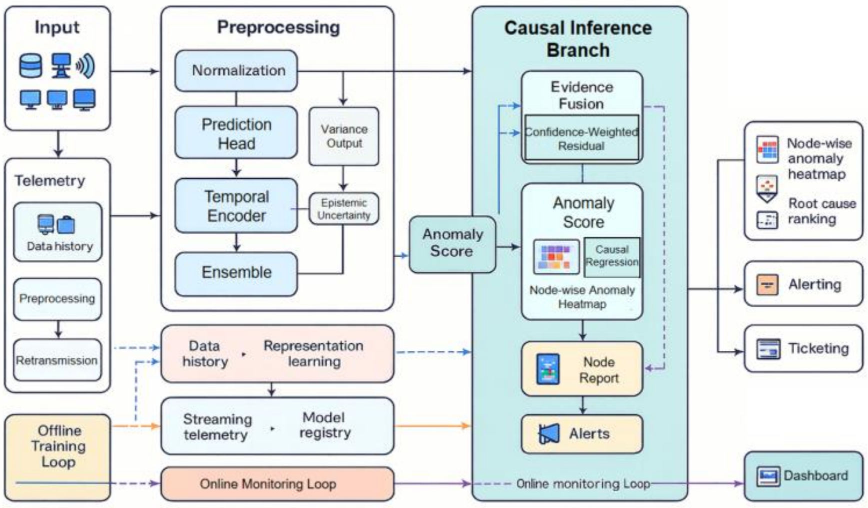

Figure 1. Uncertainty-Causal Anomaly Monitoring Framework

In the uncertainty modeling part, the method first represents the temporal input features of the service as a vector sequence, which is recorded as:

$$x_t = \{x_t^1, x_t^2, \ldots, x_t^d\}$$

Where d represents the feature dimension. To introduce confidence modeling into the prediction results, this study uses a probability distribution to characterize the potential output. Assuming that the model prediction result is y_t, the conditional probability distribution is defined as:

$$p(y_t \mid x_t, \theta) = \int p(y_t \mid z_t, \theta) p(z_t \mid x_t, \theta) dz_t$$

Where θ is the model parameter and z_t is the latent variable. Through this distributed modeling approach, the expected value and variance can be output in the prediction stage, thereby achieving uncertainty estimation. The predicted mean and variance are expressed as:

$$\mu_{\mathrm{t}} = E[y_t \mid x_t], \sigma_t^2 = V[y_t \mid x_t]$$

μ_t represents the prediction result, and σ_t^2 measures the confidence interval of the result. A large variance indicates high uncertainty in the prediction, prompting the model to be cautious about the anomaly detection results at that moment.

In the causal inference modeling part, the method abstracts service dependencies into a directed graph structure $G = (V, E)$, where the node set V represents services and the edge set E represents dependency relationships. For any two services v_i and $v_j \in V$, if there is a dependency relationship, the conditional probability is defined as:

$$P\left(v_j \mid do(v_i)\right) \neq P(v_j \mid v_i)$$

This formula embodies the difference between causal inference and correlation analysis. By comparing the intervention distribution with the conditional distribution, we can identify true causal dependencies and effectively eliminate the interference of spurious correlation features on anomaly detection.

To combine uncertainty modeling and causal inference, this study further defines an anomaly scoring function, which is in the form of:

$$S_t = \alpha \cdot \frac{|y_t - \mu_t|}{\sigma_t} + \beta \cdot \sum_{(v_i, v_j \in E)} \Delta_{ij}$$

α and β are weight coefficients. The first term reflects the confidence deviation between the prediction and the observation, and the second term measures the propagation offset Δ_{ij} under the dependency relationship through causal inference results. A larger anomaly score indicates a higher anomaly risk at that moment or node.

Through the above modeling process, the proposed method achieves dual advantages in dynamic cloud service environments. On the one hand, uncertainty estimation enhances the robustness and credibility of detection results. On the other hand, causal inference enables tracing and explaining the causes of anomalies. This integrated mechanism not only improves detection accuracy but also provides clear diagnostic evidence for operators, thereby effectively supporting the stable operation of large-scale cloud computing systems.

4. Experimental Results

4.1 Dataset

This study selects the Cloud Resource Usage Dataset for Anomaly Detection as the fundamental dataset for method validation. The dataset records time-series observations of multidimensional resource usage metrics in a multi-tenant cloud environment, including CPU, memory, and disk I/O. It covers typical scenarios of resource overload or abnormal usage. Each record consists of a vector of resource metrics, with several embedded anomaly points to simulate the behavioral characteristics of cloud resources under abnormal conditions.

The dataset exhibits strong temporal and multidimensional properties, which align with the typical requirements of cloud service monitoring. Its multidimensional resource metrics can reflect cross-service and cross-instance resource contention, jitter, or bottleneck effects, with the potential to simulate anomaly propagation in complex service dependency environments. Using this dataset allows examination of the model's ability to detect anomalies in heterogeneous metric spaces, as well as its performance in uncertainty estimation and causal inference within high-dimensional dependency structures.

Validation on this dataset demonstrates the applicability and robustness of the proposed method in cloud service environments. With its broad coverage of resource metrics and well-designed anomaly points, the dataset provides realistic testing scenarios for the uncertainty-causality integrated model. This strengthens the adaptability and generalization ability of the method to complex anomaly patterns in cloud computing systems.

4.2 Experimental Results

To systematically validate the effectiveness of the proposed cloud service anomaly monitoring method that integrates uncertainty estimation and causal inference, this study conducts comparative experiments based on representative publicly available models. The selected models include MTAD-GAT, which applies graph attention to multivariate time-series anomaly detection, LGAT, which combines graph structures with long-sequence modeling, DGINet, which captures dynamic graph interactions, and MADGA, which incorporates dependency alignment strategies. These models represent mainstream research directions in recent years, covering both the capture of complex temporal dependencies and the modeling of multi-dimensional interaction patterns. By comparing with these models, the advantages of the proposed method in capturing high-dimensional service dependencies, quantifying uncertainty, and performing causal inference can be more clearly demonstrated. The specific results are shown in Table 1.

Table1: Comparative results on alignment robustness benchmarks

Model	AUC	F1-Score	Precision	Recall
MTAD-GAT[10]	0.871	0.754	0.702	0.817
LGAT[11]	0.889	0.768	0.735	0.803
DGINet[12]	0.901	0.781	0.748	0.820
MADGA[13]	0.905	0.790	0.755	0.828
Ours	0.924	0.812	0.779	0.848

From the perspective of overall separability, the proposed method achieves an AUC of 0.924, which is higher than MADGA (0.905), DGINet (0.901), LGAT (0.889), and MTAD-GAT (0.871). This indicates the strongest discrimination between normal and anomalous instances across the full threshold range. This improvement can be attributed to the joint introduction of uncertainty estimation and causal structure constraints in the detection framework. The former explicitly models predictive variance and confidence boundaries to reduce ranking instability caused by short-term fluctuations. The latter distinguishes causal paths from mere correlations, mitigating spurious associations in cross-service topologies and pushing the ROC curve upward.

In terms of threshold-level performance, F1-Score increases step by step with model capability: MTAD-GAT achieves 0.754, LGAT 0.768, DGINet 0.781, and MADGA 0.790, while the proposed method further improves it to 0.812. This gain reflects the method's ability to simultaneously enhance both precision and recall at boundary samples. Compared with baselines that focus on correlation modeling, confidence-weighted residuals suppress false positives triggered by weak evidence, enabling a better trade-off for F1 under the same threshold setting.

From the balance between precision and recall, the proposed method achieves a precision of 0.779 and a recall of 0.848, outperforming MADGA (0.755 and 0.828) and DGINet (0.748 and 0.820) with simultaneous improvements in "high recall and stable precision." This suggests that causal inference effectively filters out noise from non-causal edges in complex dependency networks, allowing recall to improve without significant sacrifice in precision. At the same time, uncertainty calibration tightens decision boundaries when anomaly confidence is low, preventing a sharp drop in precision caused by excessive sensitivity and stabilizing the trade-off between false positives and false negatives.

A longitudinal comparison of the baselines shows a gradual improvement in AUC and F1 from MTAD-GAT and LGAT, which focus on graph attention and structural learning, to DGINet and MADGA, which incorporate dynamic graphs and alignment mechanisms. However, these methods still fall short of the proposed one. The gap indicates that relying solely on graph structures or sequential representations is insufficient to maintain robustness under multi-tenant interference and cross-tenant coupling. When uncertainty estimation is incorporated into the scoring function and causal consistency constraints are used to guide propagation paths, the model becomes more reliable in distinguishing boundary samples and weak anomalies. This is reflected in the continuous rise of AUC and the joint improvement of precision and recall.

This paper also conducts comparative experiments on the hyperparameter sensitivity of prediction head depth and width to the confidence interval of boundary samples. The experimental results are shown in Figure 2.

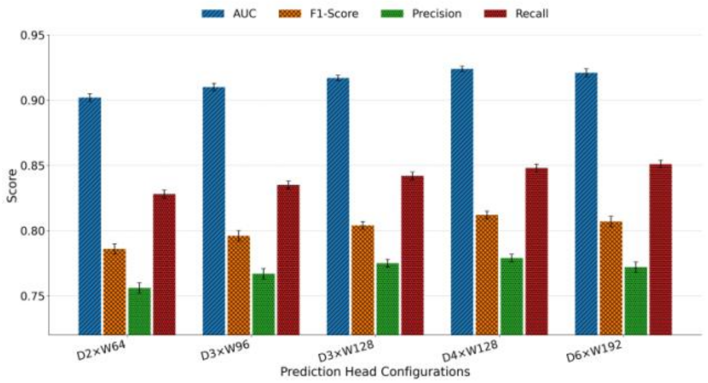

Figure 2. Hyperparameter sensitivity experiment of prediction head depth and width to boundary sample confidence interval

From the overall trend, as the prediction head increases from D2 × W64 to D4 × W128, AUC steadily improves from 0.902 to 0.924, and F1-Score rises from 0.786 to 0.812. This indicates that moderate depth and width significantly enhance the representation capacity for cross-service dependencies and temporal context. Combined with uncertainty estimation, variance constraints on boundary samples benefit both ROC space discrimination and threshold-level performance. The gains at this stage show that representation redundancy brought by increased capacity and uncertainty calibration complement each other. They improve global separability across all thresholds and stabilize prediction confidence at boundary points.

The precision curve reaches 0.779 at D4 × W128 and then slightly decreases to 0.772 at D6 × W192, while recall continuously increases from 0.828 to 0.851, showing a pattern of "stable precision and improved recall." This means that as model capacity continues to grow, the ability to capture weak signals and sparse anomalies further improves. However, the capture of correlation features also becomes more sensitive, making it easier to collect local non-causal noise, which slightly suppresses precision. At this stage, uncertainty gating plays a suppressive role, preventing a significant drop in precision, but it cannot fully offset the noise absorption effect caused by excessive depth or width.

F1-Score peaks at 0.812 with D4 × W128 and then slightly declines to 0.807, reflecting a balance point between capacity, calibration, and causality. Before D4 × W128, the synergy between representation capability and confidence interval modeling raises both precision and recall. Beyond this point, although recall continues to improve, the marginal gain of precision turns negative, leading to a slight decrease in the overall trade-off. For cloud service anomaly monitoring, this suggests that engineering deployments should prioritize head configurations near the peak to achieve a more robust Pareto point in the trade-off between false positives and false negatives.

AUC slightly drops from 0.924 to 0.921 at D6 × W192, consistent with the decline in precision. This indicates that excessive capacity may cause the decision boundary to widen and slightly dilute the global ranking. The causal inference branch can suppress part of the spurious correlation diffusion, ensuring that the continuous increase in recall does not come at the cost of a significant loss in precision. However, when representational freedom becomes too high, weak resonance from non-causal paths may still be amplified. Taken together, these results show that moderate capacity and the synergy between uncertainty and causality are key to improving robustness in complex cloud environments, while over-parameterization leads to minor regressions in both ranking and threshold performance.

This paper also analyzes the sensitivity of uncertainty calibration data under the condition of missing indicator channels and random noise injection. The experimental results are shown in Figure 3.

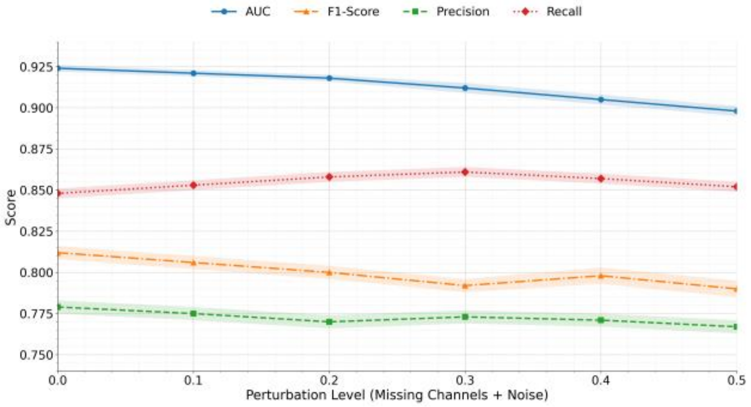

Figure 3. Uncertainty calibration data sensitivity experiment under missing indicator channels and random noise injection

As the proportion of missing channels and noise level increases from 0.00 to 0.50, the AUC decreases smoothly from 0.924 to 0.898, indicating that separability across the full threshold space weakens as information is lost. In multi-metric monitoring of cloud services, missing key channels reduce the discriminative power of cross-service dependencies, causing the margin between normal and abnormal states in the representation space to shrink. Even though uncertainty estimation provides confidence intervals for boundary samples, the overall ranking advantage is still diluted by the continuous information gap and amplified noise effects.

The F1-Score shows a "decrease – slight recovery – decrease" pattern (0.812→0.792→0.798→0.790), reflecting that threshold trade-offs are more sensitive to noise. Under moderate perturbations, confidence weighting can partially correct boundary instability caused by noise, temporarily easing the tension between Precision and Recall. However, when the missing ratio continues to increase, the accumulated uncertainty in the residual surpasses the buffering capacity of the gating mechanism, leading to more misclassifications near the threshold and causing the F1 score to decline again.

Precision exhibits a slight U-shape before trending downward (0.779→0.770→0.773→0.767), while Recall first increases and then decreases slightly (0.848→0.861→0.852). Their divergence reveals the distinct functional pathways of causal inference and uncertainty calibration in noisy environments. Under moderate perturbations, the causal branch filters out non-causal correlations and improves coverage of weak anomalies, which raises recall. However, as missing data and noise continue to grow, non-causal edges increasingly enter the candidate set, making it difficult for precision to improve correspondingly. In this case, uncertainty gating mainly acts as a "brake" to suppress false positives, but it cannot reverse the overall downward trend.

From a data sensitivity perspective, these results reveal the collaborative boundary of "observation sufficiency, uncertainty calibration, and causal constraint." Continuous loss of observation channels primarily degrades global ranking performance (AUC). In the moderate perturbation region, causal consistency and confidence gating can still maintain relatively stable threshold-level performance. Once the system enters a high-perturbation regime, the confidence intervals for boundary samples expand further, and incorrect evidence from non-causal noise increases, exerting more significant pressure on Precision and F1. For cloud service anomaly detection, this suggests that channel reconstruction or robust feature selection should be combined to enhance structural discriminability and calibration stability under channel loss and noise injection.

This paper also evaluates the sensitivity of the false alarm and false negative trade-off environment under multi-tenant interference intensity and cross-tenant coupling degree. The experimental results are shown in Figure 4.

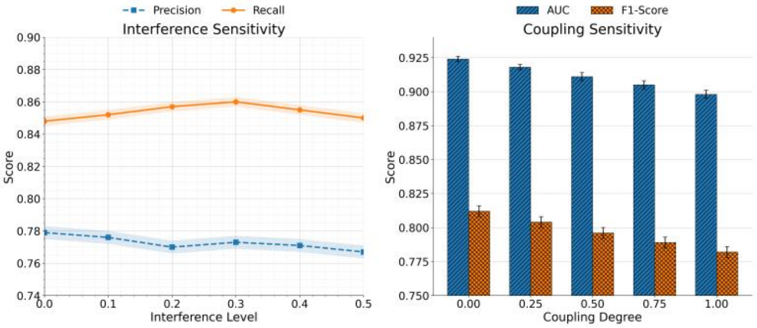

Figure 4. Environmental sensitivity experiment on the trade-off between false positives and false negatives under multi-tenant interference intensity and cross-tenant coupling degree

As multi-tenant interference gradually increases, Precision first drops slightly, then recovers, and finally declines again. This indicates that in the moderate interference region, uncertainty gating and evidence fusion can suppress some spurious correlations and temporarily improve the robustness of boundary decisions. However, as interference continues to intensify, non-causal noise introduced by cross-tenant interactions is more easily absorbed at the representation level, increasing the difficulty of controlling false positives and causing Precision to decline. This trajectory suggests that correlation modeling alone is insufficient to maintain stable alert quality. Confidence information must be treated as a primary component in the scoring function to avoid overreactions under amplified noise conditions.

Unlike the slight U-shape of Precision, Recall exhibits a unimodal "rise-then-fall" pattern (0.848→0.860→ 0.850). This shows that under moderate interference, the causal branch filters out non-causal paths and focuses on interpretable propagation chains, making it easier to capture weak signal anomalies and improving recall. When interference intensifies further, resonance between true causal chains and noise dilutes effective evidence, making it difficult to sustain recall gains. For cloud service operations, this suggests that in high-concurrency multi-tenant scenarios, the confidence threshold and the strength of causal regularization should be dynamically adjusted to maintain coverage without sacrificing too much precision.

The impact of cross-tenant coupling on global separability is reflected in a smooth decrease in AUC as coupling increases (0.924 → 0.898). As coupling deepens, synchronous fluctuations and topological resonance between services become more frequent, reducing the distance between normal and abnormal states in the representation space and shifting the ROC curve downward. Even with built-in uncertainty decomposition and causal constraints, if correlation-driven common-mode factors dominate, the ranking advantage is gradually eroded. Therefore, in highly coupled environments, structural decoupling methods such as graph sparsification and intervention-based representation learning, combined with channel-level robustness enhancement, are necessary to maintain separability across the full threshold space.

F1-Score shows a monotonic decrease as coupling increases (0.812→0.782), indicating that threshold trade-offs become harder to manage in strongly dependent networks. Coupling amplifies the uncertainty region around boundary samples, leading to more frequent trade-offs between false positives and false negatives, and reducing overall performance at a unified threshold. Incorporating uncertainty estimation into residual weighting helps stabilize F1 under low to moderate coupling, but as dependency strength continues to rise, the marginal benefits of causal consistency constraints diminish, and F1 inevitably declines. This observation aligns with real-world cloud service environments. As cross-team and cross-tenant call relationships become increasingly tight, parameter tuning alone is insufficient to maintain threshold performance and must be complemented by topological governance and online calibration.

Next, this study conducted experiments on the data sensitivity of label scarcity and pseudo-label noise ratio to causal inference accuracy. The experimental results are shown in Figure 5.

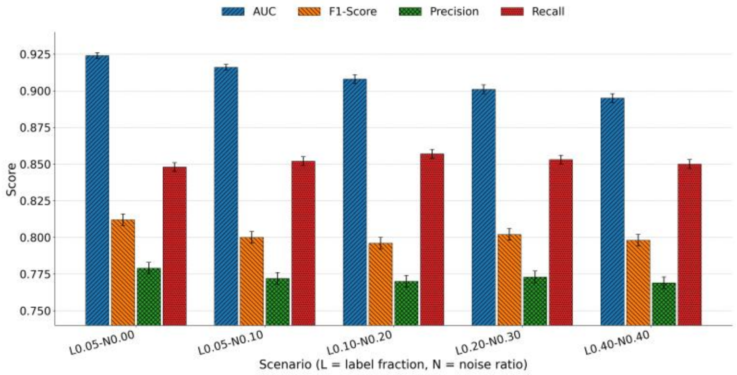

Figure 5. Data Sensitivity Experiment on Label Scarcity and Pseudo-label Noise Ratio to Causal Inference Accuracy

As label scarcity intensifies and the proportion of noisy pseudo-labels increases, the AUC decreases steadily from 0.924 to 0.895, indicating that the overall discriminative power of causal inference is significantly affected. Under conditions of sufficient labels and low noise, the model can maintain high-confidence anomaly detection through causal structure constraints and uncertainty quantification. However, as pseudo-label errors accumulate and supervisory signals become diluted, the identifiability of causal paths in the inference space decreases. This leads to a gradual reduction in the separability between normal and abnormal instances across the full threshold space, highlighting the importance of high-quality annotations for maintaining global discriminability.

The F1-Score follows a "decrease – slight recovery – decrease" pattern (0.812→0.796→0.802→0.798), indicating that under moderate noise, pseudo-labels, despite introducing errors, increase the diversity and coverage of the distribution, which temporarily improves threshold performance. However, when noise continues to rise, the spread of erroneous signals along non-causal paths causes boundary decisions to become inaccurate, leading to another drop in F1. This shows that there is a "tolerable noise region" in pseudo-label learning. Beyond this threshold, the advantage of causal modeling is gradually eroded by noise interference.

Precision declines initially, then recovers slightly before dropping again (0.779→0.770→0.773→0.769), while Recall first increases and then decreases (0.848→0.857→0.850). This divergence reflects the dual impact of pseudo-label quality on the model. At early noise levels, recall improves because pseudo-labels expand coverage of weak signals. However, as the proportion of incorrect labels grows, the model begins to treat irrelevant signals as part of causal paths, increasing false positives and reducing precision. Although uncertainty gating suppresses error propagation to some extent, it cannot fully offset the structural bias caused by degraded annotation quality.

Overall, the impact of label scarcity and pseudo-label noise on causal inference performance is reflected not only in the decline of global ranking performance (AUC) but also in the trade-offs between false positives and false negatives at the threshold level. A moderate amount of pseudo-labels can compensate for insufficient supervision and improve recall. However, excessive noise disrupts the semantic consistency of causal paths, leading to systematic degradation in accuracy and stability. This experiment suggests that in practical cloud service anomaly detection tasks, pseudo-label generation and filtering strategies should be designed in coordination with uncertainty estimation and causal regularization to maintain inference reliability and generalization under weak supervision conditions.

5. Conclusion

This study addresses the problem of anomaly monitoring in cloud service environments and proposes a modeling method that integrates uncertainty estimation with causal inference. It systematically tackles the challenges posed by multi-tenant interference, cross-tenant coupling, and pseudo-label noise to detection accuracy and stability. By introducing confidence intervals at the representation level and applying causal constraints at the structural level, the proposed method demonstrates significant advantages in boundary sample recognition, global ranking stability, and the balance between false positives and false negatives. The significance of this work lies not only in improving model robustness under dynamic and complex conditions but also in advancing anomaly detection from simple correlation modeling to a combination of causality-driven reasoning and uncertainty calibration. This provides a solid technical foundation for the secure and reliable operation of cloud service systems.

From a methodological perspective, this study emphasizes the importance of multidimensional modeling in scenarios with data sensitivity and environmental sensitivity. The experiments on label scarcity and pseudo-label noise reveal the damaging effect of spurious evidence on causal edge recognition while highlighting the value of uncertainty mechanisms in mitigating this risk. The interference and coupling experiments show the model's performance differences under complex multi-tenant interactions and further demonstrate the complementary roles of causal inference and uncertainty estimation in cross-level feature modeling. This exploration not only addresses practical challenges in cloud service environments but also provides new perspectives for anomaly detection and resource scheduling in broader distributed systems.

At the application level, the findings of this study have strong generalizability and transferability. The framework can be seamlessly integrated into existing cloud monitoring systems to provide more accurate technical support for service quality assurance, resource scheduling, and potential attack detection. More importantly, the method enables high-quality anomaly identification without relying on large amounts of clean labeled data, thereby reducing the maintenance and operational costs of large-scale cloud platforms. This low-dependence and high-robustness design can accelerate the deployment of automated monitoring solutions across industries, generating tangible benefits in service continuity and user experience.

Future research directions are worth further exploration. One important direction is to embed uncertainty estimation and causal inference mechanisms more deeply into cross-modal data fusion and real-time stream processing to enhance adaptability in complex environments. Another direction is to examine the method's scalability and adaptability in real-world large-scale cloud platforms, especially in cross-regional and multi-cloud collaborative scenarios. Furthermore, as cloud services increasingly integrate with edge computing and the Internet of Things, anomaly patterns will become more diverse and dynamic. The ideas of this study can provide theoretical and practical references for these emerging applications. Continued exploration of the integration between uncertainty estimation and causal inference is expected to drive intelligent operation systems toward greater autonomy, reliability, and efficiency.

References

[1] Soldani, J. and Brogi, A., "Anomaly detection and failure root cause analysis in (micro) service-based cloud applications: A survey", ACM Computing Surveys (CSUR), vol. 55, no. 3, pp. 1-39, 2022.

[2] Li, Z., Zhao, N., Zhang, S., Sun, Y., Chen, P., Wen, X. et al., "Constructing large-scale real-world benchmark datasets for AIOps", arXiv preprint arXiv:2208.03938, 2022.

[3] Zhang, C., Peng, X., Sha, C., Zhang, K., Fu, Z., Wu, X. et al., "DeeptraLog: Trace-log combined microservice anomaly detection through graph-based deep learning", Proceedings of the 44th International Conference on Software Engineering, pp. 623-634, May 2022.

[4] Gawlikowski J, Tassi C R N, Ali M, et al. A survey of uncertainty in deep neural networks[J]. Artificial Intelligence Review, 2023, 56(Suppl 1): 1513-1589.

[5] Mena J, Pujol O, Vitrià J. A survey on uncertainty estimation in deep learning classification systems from a bayesian perspective[J]. ACM Computing Surveys (CSUR), 2021, 54(9): 1-35.

[6] Jiao L, Wang Y, Liu X, et al. Causal inference meets deep learning: A comprehensive survey[J]. Research, 2024, 7: 0467.

[7] Frank K A, Lin Q, Xu R, et al. Quantifying the robustness of causal inferences: Sensitivity analysis for pragmatic social science[J]. Social Science Research, 2023, 110: 102815.

[8] Wang Z, Shu K, Culotta A. Enhancing model robustness and fairness with causality: A regularization approach[J]. arXiv preprint arXiv:2110.00911, 2021.

[9] Pearl, J., Causality, Cambridge University Press, 2009.

[10] Zhao H, Wang Y, Duan J, et al. Multivariate time-series anomaly detection via graph attention network[C]//2020 IEEE international conference on data mining (ICDM). IEEE, 2020: 841-850.

[11] Xu, J., Wu, H., Wang, J., & Long, M. (2021). Anomaly transformer: Time series anomaly detection with association discrepancy. arXiv preprint arXiv:2110.02642.

[12] Huang X, Chen N, Deng Z, et al. Multivariate time series anomaly detection via dynamic graph attention network and Informer[J]. Applied Intelligence, 2024, 54(17-18): 7636-765.

[13] Wang Y, Sun H, Wang C, et al. Interdependency matters: graph alignment for multivariate time series anomaly detection[C]//2024 IEEE International Conference on Data Mining (ICDM). IEEE, 2024: 869-874.