Artificial Intelligence and Computing Innovations
Intelligence Science Publishing
https://www.ispublishing.com
ISSN: 3070-4944
Vo.5, No.7, 2025

Causal Aware Anomaly Detection for Modeling the Evolution of Backend Operational States in Ultra Large Scale Online Services

I-Hsien Lai[1*]
[1]University of Wisconsin-Madison, Madison, USA
**Corresponding author: j093760@gmail.com*

Abstract: To address the challenges of highly dynamic operational states, highly heterogeneous multi-source observations, and hidden anomaly causes that are susceptible to surface disturbances in ultra-large-scale online service backend systems, a causal perception anomaly detection method based on operational state evolution modeling is proposed. This method is grounded in multi-dimensional operational information such as logs, metrics, links, and topology. First, it constructs a state representation capable of characterizing the temporal correlations and structural dependencies of the backend system. Based on this, a state evolution coding mechanism characterizes the potential changes in the system's continuous operation, enhancing the perception of anomaly accumulation effects and implicit propagation behaviors. Subsequently, causal component extraction and intervention consistency constraints are introduced to distinguish between key driving information and background disturbance information in complex observations, reducing the interference of non-critical fluctuations on the anomaly identification process and thus improving the model's stability in identifying anomaly patterns in complex backend scenarios. Furthermore, evolutionary bias modeling is used to characterize the difference between the predicted state and the actual state, extending anomaly detection from static observation discrimination to structured identification oriented towards dynamic operational processes. The proposed method can collaboratively model state continuity, dependency propagation, and causal relationships within a unified framework, providing a solution for complex online service backend anomaly detection that combines expressive and discriminative capabilities. Related results demonstrate that this method exhibits good overall performance and application value in ultra-large-scale online service backend anomaly detection tasks.

Keywords: Backend anomaly detection; state evolution modeling; causal representation learning; intelligent operation and maintenance.

1 Introduction

With the continuous evolution of cloud computing, microservice architecture, container orchestration, and edge computing technologies, ultra-large-scale online services have become the core

infrastructure supporting the operation of the digital economy. Backend systems, facing massive user requests, complex business processes, and high-frequency resource scheduling, exhibit significant high concurrency, strong coupling, dynamic fluctuations, and cross-layer linkages during long-term operation[1,2]. The interactions between service instances, call chains, system logs, performance metrics, and resource states mean that the backend's operational state no longer exhibits static snapshot-like changes, but rather a complex dynamic process that evolves. Against this backdrop, accurately characterizing the evolutionary patterns of the backend's operational state and promptly identifying abnormal behavior in complex and turbulent environments has become a key scientific problem in ensuring the continuity, stability, and reliability of online services.

Traditional backend anomaly detection methods typically focus on statistical deviation analysis of single log sequences, single-dimensional performance metrics, or local behavioral patterns. While these methods can detect obvious faults or sudden anomalies to a certain extent, they often fail to reveal the deep-seated driving mechanisms behind system state changes when faced with the challenges of multi-source heterogeneous data coupling, enhanced temporal dependencies, and complex anomaly propagation paths in ultra-large-scale online services[3]. Especially under the combined effects of fluctuating business traffic, resource competition, link congestion, configuration drift, and changes in service dependencies, anomalies often exhibit characteristics such as delayed manifestation, cross-module diffusion, and mixed representations. Relying solely on surface-level correlation information can easily lead to false positives, false negatives, and biased identification of the root causes of anomalies. Therefore, from the perspective of operational state evolution, constructing anomaly detection theories and methods that can simultaneously understand temporal changes, structural correlations, and implicit driving factors has significant practical needs and theoretical value[4].

Causal perception modeling provides a new research approach to solving the above problems[5]. Compared to modeling paradigms that only focus on variable correlations, causal perception methods emphasize the interaction paths and influence mechanisms between the internal state of the system, event triggering factors, and abnormal results, helping to separate truly explanatory key factors from complex observed phenomena. In ultra-large-scale online service backend scenarios, there are often asymmetric, asynchronous, and hierarchical influence relationships between service load changes, resource scheduling strategies, dependency link disturbances, and local fault propagation. These relationships collectively shape the dynamic evolution process of the operational state. Introducing causal awareness mechanisms into anomaly detection tasks not only enhances the ability to identify complex anomaly patterns but also strengthens the model's adaptability to environmental changes, business disturbances, and system structural adjustments, thus laying the foundation for building more robust and interpretable intelligent backend operation and maintenance methods[6].

Research on causal awareness anomaly detection algorithms for modeling the evolution of backend operational states in ultra-large-scale online services has significant theoretical and engineering application value. Theoretically, this direction helps to shift backend anomaly detection from static pattern recognition to dynamic mechanism understanding, promoting the cross-integration of temporal evolution modeling, causal representation learning, and intelligent analysis of complex systems. In application, this research can provide a more reliable decision-making basis for fault early warning, risk

control, resource optimization, and stability assurance of online services, further enhancing the autonomous perception and adaptive control capabilities of digital infrastructure under high load, strong fluctuations, and complex coupling environments[7]. Against the backdrop of continuously increasing demands for intelligent operation and maintenance, conducting research in this direction is of great significance for ensuring the safe and stable operation of critical online service systems, supporting the supply of high-quality digital services, and promoting the development of next-generation intelligent operation and maintenance technologies.

2 BackGround

Large-scale online service backend systems typically consist of compute nodes, storage components, network links, scheduling platforms, middleware services, and business logic modules, exhibiting significant multi-layered collaborative characteristics in their operation. As service scale continues to expand, the backend system is no longer a simple combination of independently operating modules but has evolved into a high-dimensional dynamic system composed of numerous service instances, complex call relationships, and continuous resource exchange. In this system, the input pressure from user requests, changes in task scheduling strategies, fluctuations in resource allocation states, and performance degradation of local components all propagate their effects to other parts through dependencies, resulting in cross-level, cross-time-period, and cross-module state linkages. Due to the continuous existence of these linkages, the backend operating state often exhibits a coexistence of gradual changes and local mutations. Anomalies are therefore no longer limited to single-point failures, but are more likely to manifest as complex phenomena such as accumulated state shifts, intensified chain propagation, and overall service quality degradation. To facilitate the summarization of the core characteristics of the large-scale online service backend operating environment, Table 1 summarizes the relevant scenario attributes.

Table 1. Summary of Key Background Characteristics in Ultra-Large-Scale Online Service Backend Operation Scenarios

Background Dimension	Specific Manifestation	Impact on Operational State Analysis
System Scale	A large number of service instances, wide node distribution, and high business concurrency.	The state space expands rapidly, and local issues are more likely to affect overall system behavior.
Architectural Complexity	Dense microservice dependencies and collaborative operation among multiple components.	Anomaly propagation paths become more complex, increasing the difficulty of association analysis.
Data Heterogeneity	Logs, metrics, events, and trace data coexist.	Multi-source information fusion becomes more difficult, and state representation is less consistent.
Temporal Variability	Traffic fluctuates significantly, while configuration and deployment are continuously adjusted.	The system remains in a dynamic environment for long periods, making static analysis insufficient.
Diverse Sources of	Resource contention, version updates,	The causes of anomalies become more

Disturbance	network jitter, and sudden business surges can all trigger fluctuations.	entangled, and observable symptoms may not align with the true underlying triggers.
Complex Anomaly Manifestations	Single-point anomalies, chained anomalies, gradual degradation, and sudden failures are interwoven.	Anomaly boundaries become less clear, further increasing the difficulty of detection and discrimination.

In actual operation, backend status information typically originates from multiple dimensions, including log sequences, performance metrics, link events, resource utilization, and business behavior records. These data vary significantly in terms of time granularity, semantic expression, sampling frequency, and noise levels. Data from different sources reflects both the system's operational appearance and implicitly reveals internal state change trends. However, due to issues such as misalignment, incompleteness, and mutual interference among heterogeneous data, relying solely on local observations often fails to provide a stable portrayal of the overall operational status. Furthermore, online service scenarios are characterized by rapid changes in business rhythms, large differences in load peaks and troughs, and continuous updates to the deployment environment, causing backend systems to operate in a non-stable state for extended periods. This context dictates that anomaly detection tasks cannot remain merely at the level of isolated event identification; instead, a more systematic cognitive framework must be established to address the continuous evolution of the state, thereby providing a more solid modeling foundation for subsequent anomaly identification, risk warning, and operational assurance.

3 Methodology

Ultra large-scale online service backends continuously generate heterogeneous observations from metrics, logs, traces, and service dependency interactions, while the underlying operational state evolves under joint influences from workload fluctuation, resource contention, topology variation, and latent fault propagation. This paper presents the overall model architecture, as shown in Figure 1.

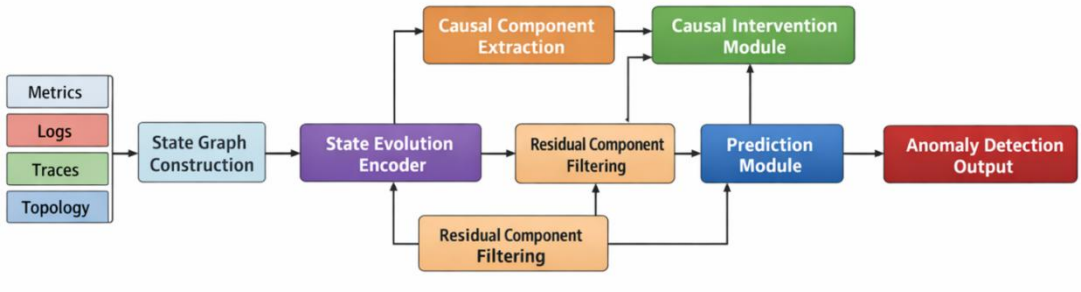


Figure 1. Overall model architecture

To capture this dynamic process in a causally meaningful manner, the proposed method first organizes the backend at time step t as a structured state graph, where $\mathcal{V}_t$ denotes service or resource entities, $\mathcal{E}_t$ represents inter-component dependencies, $\boldsymbol{X}_t$ contains node-level observations, and $\boldsymbol{A}_t$ encodes dependency strengths. Such a formulation is introduced to preserve both temporal continuity and cross-component influence paths, because anomaly emergence in large systems rarely originates

from isolated deviations and more often reflects the accumulation, transfer, and amplification of local disturbances across connected modules. In this setting, the observed backend state is written as:

Structure-aware root cause localization through joint graph representation learning and causal inference [8] provides a complementary foundation for treating service dependencies as explanatory pathways rather than as static adjacency information. Reliable monitoring of AI inference backends through behavior representation learning [9] likewise motivates encoding operational observations as evolving behavioral patterns. These principles guide the proposed state graph toward representations that retain both the topology of interactions and the behaviorally meaningful changes that precede anomaly escalation.

$$S_t = (\mathcal{V}_t, \mathcal{E}_t, \boldsymbol{X}_t, \boldsymbol{A}_t) \tag{1}$$

After this representation is established, a state evolution encoder is employed to map consecutive observations into latent trajectories so that transient fluctuations and persistent degradation trends can be distinguished within a unified space. Rather than treating each timestamp independently, the method constructs a recursive evolution process to retain contextual dependency between adjacent system states, which is essential for recognizing anomalies that only become identifiable through temporal accumulation. The hidden representation $\boldsymbol{H}_t$ is therefore updated as:

The temporal encoder further integrates multi-scale dynamics rather than relying only on adjacent-state similarity. Frequency-guided Transformer modeling with spectral enhancement [10] demonstrates the value of separating recurrent periodic structure from transient deviations, while attention-enhanced network traffic prediction [11] shows how load-sensitive temporal signals can improve the anticipation of state change. Accordingly, the evolution trajectory is designed to retain both short-lived operational shocks and longer cadence-dependent variations that may propagate across backend services.

$$\boldsymbol{H}_t = \phi\left(\boldsymbol{W}_h \boldsymbol{X}_t + \boldsymbol{W}_s \boldsymbol{A}_t \boldsymbol{X}_t + \boldsymbol{W}_r \boldsymbol{H}_{t-1}\right) \tag{2}$$

Here $\phi(\cdot)$ denotes a nonlinear transformation, while $\boldsymbol{W}_h$, $\boldsymbol{W}_s$, and $\boldsymbol{W}_r$ respectively control current observation encoding, structural aggregation, and temporal inheritance, thereby enabling the latent state to jointly reflect instantaneous behavior and historical evolution.

Beyond temporal dependence, backend anomalies are often driven by hidden triggers whose observable effects may be delayed, dispersed, or confounded by routine operational variation. For this reason, a causal awareness module is introduced to separate stable driving factors from superficial correlations, thereby reducing the risk that frequent but noncritical perturbations dominate anomaly judgment. Instead of directly scoring the raw latent state, the method first decomposes $\boldsymbol{H}_t$ into a causal component $\boldsymbol{C}_t$ and a residual component $\boldsymbol{B}_t$, where the former is designed to preserve structurally influential information, and the latter absorbs context-dependent background fluctuations. This decomposition is formulated as:

$$\boldsymbol{C}_t = \sigma\left(\boldsymbol{W}_c \boldsymbol{H}_t\right) \odot \boldsymbol{H}_t \tag{3}$$

$$\boldsymbol{B}_t = \boldsymbol{H}_t - \boldsymbol{C}_t \tag{4}$$

Such a separation is meaningful because complex backend systems frequently exhibit bursts of benign variability caused by workload seasonality, asynchronous scheduling, and elastic resource adjustment, whereas truly harmful anomalies are more closely associated with persistent causal disruptions that reshape the evolution path of the system state. Based on this consideration, a causal intervention consistency mechanism is further imposed so that the representation remains focused on invariant abnormal drivers under perturbed observational contexts. The intervened representation $\widetilde{\boldsymbol{H}}_t$ is constructed by suppressing residual disturbance and preserving causal semantics through:

The causal separation is coupled with robustness to limited and noisy observations. Unified semi-supervised and robust optimization for low-resource data mining [12] motivates preserving dependable structural signals when labels are scarce and multi-source measurements contain disturbance. This supports the residual pathway as a controlled buffer for contextual variation while concentrating the causal component on drivers that remain stable across perturbed operating conditions.

$$\widetilde{\boldsymbol{H}}_t = \boldsymbol{C}_t + \lambda \boldsymbol{B}_t \tag{5}$$

where $\lambda \in [0,1]$ controls the retained proportion of residual context, allowing the model to emphasize causal signals while avoiding excessive information loss in dynamic service environments.

Once the causally guided state has been obtained, anomaly identification is performed through evolution discrepancy modeling rather than through direct thresholding on raw observations. The core idea is that normal backend operation should follow a relatively coherent latent transition law, even under moderate load variation, while abnormal evolution introduces structural inconsistency between predicted and realized system states. In light of this principle, the method predicts the next latent state from the intervened representation and compares it with the actual evolved representation at the subsequent timestamp. The transition prediction is defined as:

$$\widehat{\boldsymbol{H}}_{t+1} = \psi\left(\widetilde{\boldsymbol{H}}_t, \boldsymbol{A}_t\right) \tag{6}$$

where $\psi(\cdot)$ denotes the evolution predictor conditioned on both the current causally filtered state and the dependency structure. Since anomalous behavior often manifests as deviation from expected evolution rather than as extreme absolute magnitude, the anomaly score is then measured by:

$$\alpha_t = \left\| \boldsymbol{H}_{t+1} - \widehat{\boldsymbol{H}}_{t+1} \right\|_2^2 \tag{7}$$

In this manner, the score α_t reflects whether the future system trajectory remains compatible with historically grounded causal evolution patterns, thereby improving sensitivity to progressive degradation, cross-service propagation, and structurally induced instability.

To make the learned representation simultaneously evolution aware and causally discriminative, the overall optimization objective combines state reconstruction, intervention consistency, and anomaly separability within a unified training framework. A joint objective is adopted because relying on a single learning signal may either overemphasize short-term fitting accuracy or underconstrain the preservation of causal structure. Specifically, reconstruction regularization maintains fidelity to observed backend behavior, consistency regularization aligns original and intervened trajectories, and anomaly-oriented

discrimination enlarges the gap between coherent evolution and disrupted evolution in latent space. The final loss function is given by:

$$\mathcal{L} = \mathcal{L}_{rec} + \beta\mathcal{L}_{con} + \gamma\mathcal{L}_{ano} \tag{8}$$

with $\mathcal{L}_{rec}$ preserving state expressiveness, $\mathcal{L}_{con}$ enforcing causal robustness under context perturbation, and $\mathcal{L}_{ano}$ enhancing deviation sensitivity for abnormal trajectories. Through this design, the method does not merely detect whether a current observation appears unusual, but instead evaluates whether the backend is evolving in a manner that violates the expected causal state transition law, which is particularly important for ultra-large-scale online services where critical failures are often preceded by subtle but cumulatively meaningful shifts in operational dynamics.

The joint objective also supports downstream operational control by distinguishing deviations that warrant attention from benign load-related variation. Risk-constrained reinforcement learning for SLO-driven cloud-native elastic scaling [13] emphasizes the need to balance service objectives with bounded intervention risk. In the present framework, this perspective is operationalized through intervention consistency and anomaly separability, so that detected evolution violations can provide a reliable basis for later alerting, scaling, or remediation decisions.

4 Experimental Results and Analysis

4.1 Dataset introduction

This study uses the publicly available Train Ticket microservice anomaly dataset as the foundation for our experimental data. Released on Zenodo and made available for download and use through open access, this dataset contains 10 sets of anomaly scenario data related to microservice version configurations, and simultaneously provides multi-source observation information, including logs, Jaeger distributed call chain tracing data, and Prometheus KPI monitoring metrics. Because its data organization method can simultaneously reflect service dependencies, request propagation processes, and performance state changes, this dataset can comprehensively characterize the evolution of the backend system's operational state under anomaly disturbances, which is highly consistent with the research goals of modeling the operational state of ultra-large-scale online service backends and detecting causal anomalies.

Furthermore, the Train Ticket itself is a typical benchmark system built for microservice scenarios. The data not only retains multimodal backend observable information but also provides brief descriptions of identified anomalies, providing excellent data support for subsequent modeling around state evolution, cross-service impact, and anomaly propagation patterns. Compared to data resources containing only a single log or a single metric, this dataset can simultaneously cover log semantics, link topology behavior, and metric fluctuation characteristics. It is more suitable for studying anomaly detection problems in complex online service backends driven by local perturbations, dependency propagation, and structural changes. Therefore, it can serve as a representative publicly available data source for the task scenario presented in this paper.

4.2 Experimental setup

To ensure the stability of the model training process and the consistency of result evaluation, all experiments were conducted in a unified hardware and software environment, and the number of training epochs, batch size, optimizer type, learning rate, and regularization parameters were standardized. Considering that the backend runtime state evolution modeling task involves multi-source heterogeneous inputs, temporal dependency modeling, and structural association learning, the experimental platform uses a graphics processing unit with strong parallel computing capabilities to meet the memory and computing resource requirements of complex models during the training phase. Meanwhile, the software environment adopts mainstream deep learning frameworks and standardized scientific computing libraries to improve the reproducibility and deployment compatibility of the model implementation. Specific hardware and software configurations and core hyperparameter settings are shown in Table 2.

Table 2 Experimental hardware and software environment and main hyperparameter settings

Category	Configuration Item	Specific Setting
Hardware Environment	Operating System	Ubuntu 22.04
Hardware Environment	Processor	Intel Xeon Silver 4314 CPU
Hardware Environment	Graphics Processing Unit	NVIDIA RTX 4090 24GB
Hardware Environment	Memory	128 GB
Software Environment	Python Version	Python 3.10
Software Environment	Deep Learning Framework	PyTorch 2.1
Software Environment	CUDA Version	CUDA 12.1
Software Environment	Data Processing Libraries	NumPy, Pandas, Scikit-learn
Training Parameters	Number of Epochs	100
Training Parameters	Batch Size	32
Training Parameters	Optimizer	AdamW
Training Parameters	Initial Learning Rate	0.0001
Training Parameters	Weight Decay	0.00001
Training Parameters	Learning Rate Scheduling Strategy	Cosine Annealing
Training Parameters	Random Seed	42
Model Parameters	Hidden Dimension	128
Model Parameters	Time Window Length	20
Model Parameters	Dropout	0.1
Model Parameters	Causal Intervention Coefficient	0.3
Model Parameters	Loss Weight Coefficient β	1.0
Model Parameters	Loss Weight Coefficient γ	0.5

4.3 Experimental Results and Analysis

To further illustrate the comprehensive performance of the proposed method in the anomaly detection task of ultra-large-scale online service backends, a comparative analysis was conducted with representative works that are closely related to the research direction of this paper. These methods mainly focus on log anomaly identification, link anomaly detection, indicator anomaly discrimination, and multi-source operational status analysis in microservice scenarios, reflecting the main research paths

in the current field of intelligent backend operation and maintenance from different perspectives. The comparison results are summarized in Table 3.

Table 3. Experimental results compared with other models

Method	ACC	PRE	REC	F1
Zhang et al.[14]	91.84	90.76	92.11	91.43
Xie et al.[15]	92.37	91.20	92.85	92.02
Zhang et al.[16]	93.15	92.44	93.56	92.99
Liu et al.[17]	90.68	89.94	91.02	90.48
Chen et al.[18]	89.97	89.21	90.33	89.77
Duan et al.[19]	91.42	90.88	91.67	91.27
Yu et al.[20]	92.76	91.95	93.08	92.51
Ours	94.63	93.88	95.14	94.50

The proposed algorithm demonstrates superior comprehensive recognition capabilities in ultra-large-scale online service backend anomaly detection tasks. This indicates that the method can more effectively capture abnormal features in complex operational states, while maintaining accuracy in discrimination and balancing positive class recognition ability with overall stability. This demonstrates that constructing a causal perception modeling mechanism around the evolution of backend operational states can more fully explore the intrinsic correlations between multi-source operational information, reduce the interference of surface disturbances on anomaly judgment, and thus enable the model to maintain strong robustness and adaptability in scenarios involving complex service dependencies, dynamic load fluctuations, and multi-dimensional anomalies. In conclusion, the proposed algorithm effectively demonstrates its effectiveness and application value for anomaly detection tasks in complex backend systems.

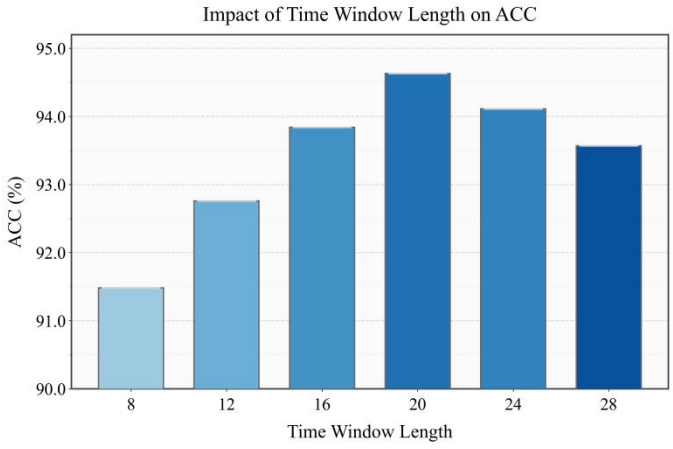


Figure 2. The impact of time window length on ACC

As shown in Figure 2, the model achieves its best ACC when the time window length is set to 20, while shorter or longer windows both lead to a decline in detection performance. This suggests that an appropriate temporal range is necessary for backend anomaly detection, since overly short windows may fail to capture the accumulation and propagation of abnormal signals, whereas overly long windows may

introduce redundant fluctuations and weaken the discriminability of anomaly patterns. The result indicates that the current setting provides a more suitable balance between temporal context preservation and effective anomaly identification in complex backend operational environments.

To further verify the practical role of the key designs in this paper for anomaly detection in ultra-large-scale online service backends, ablation analysis was conducted focusing on state evolution coding, causal component extraction, and causal intervention consistency mechanisms. The comprehensive performance of different module combinations was evaluated under a unified setting. The relevant results are shown in Table 4.

Table 4. Ablation test results

Ablation Setting	ACC	PRE	REC	F1
w/o Causal Component Extraction	92.84	91.96	93.12	92.53
w/o Causal Intervention Consistency	93.27	92.41	93.65	93.03
w/o Evolution Discrepancy Prediction	92.68	91.87	92.94	92.40
Ours	94.63	93.88	95.14	94.50

The results in Table 4 show that the complete model achieves the best performance across all evaluation metrics, while removing any core component leads to a noticeable decline in anomaly detection effectiveness. In particular, the decreases observed after excluding causal component extraction, causal intervention consistency, or evolution discrepancy prediction suggest that these modules contribute complementary roles in the overall framework. This indicates that backend anomaly detection benefits from jointly modeling temporal state evolution and disturbance-resistant causal representations, rather than relying only on a single discrimination mechanism.

The learning rate is a key factor affecting the magnitude of model parameter updates and convergence stability, and its value directly impacts the learning quality of the backend runtime state evolution representation. For causal perception anomaly detection tasks, an appropriate learning rate setting helps maintain the stability of the representation extraction process in complex dynamic scenarios and enhances the model's ability to perceive key anomaly information. Based on this consideration, it is necessary to examine the performance changes of the algorithm under different learning rates, thereby providing a reasonable basis for model parameter configuration. The experimental results are shown in Figure 3.

The results show that the proposed algorithm maintains good anomaly detection capabilities under the current learning rate settings, indicating that the constructed causal-aware state evolution modeling mechanism has strong parameter adaptability. This method can stably extract key anomaly representations under the joint constraints of multi-source backend operational information and effectively maintain the discriminative ability and state representation quality during model training, thus demonstrating good robustness and usability.

Furthermore, this also shows that the proposed method can effectively coordinate the relationship between state modeling, causal component extraction, and anomaly detection in complex backend scenarios of ultra-large-scale online services, enabling the model to maintain good recognition performance in complex dependency relationships and dynamic disturbance environments. Overall, the proposed algorithm demonstrates strong effectiveness and application value in backend anomaly detection tasks, providing reliable support for subsequent research on intelligent operation and maintenance in complex operational scenarios.

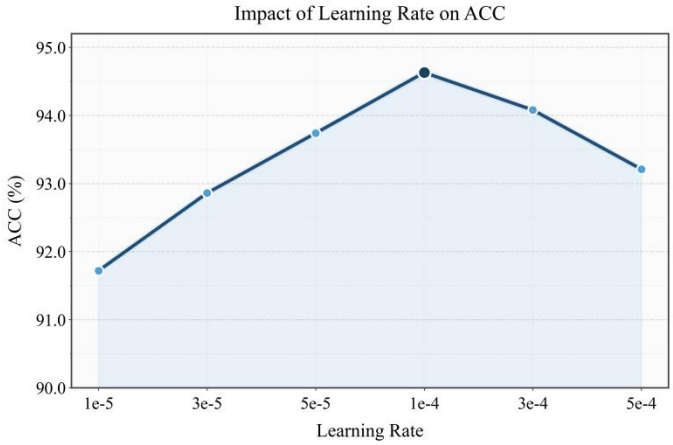


Figure 3. The impact of learning rate on Acc

5 Conclusion

This paper addresses the challenges of complex operational states, dynamic evolution processes, heterogeneous multi-source observations, and hidden anomaly drivers in the anomaly detection of ultra-large-scale online service backends. It proposes a causal-aware anomaly detection method based on backend operational state evolution modeling. Starting from the practical needs of long-term continuous operation of backend systems, this research organically combines state evolution modeling, causal component extraction, and intervention consistency constraints to enhance the model's ability to identify complex anomaly patterns and suppress surface disturbances. Compared to traditional approaches that rely solely on local observation deviations or static correlation analysis, this paper focuses more on the continuity of internal state changes within the backend system, the structural nature of dependency propagation, and the key driving factors in the anomaly formation process. This allows anomaly detection to move beyond instantaneous pattern discrimination and provide more interpretative identification and judgment of system state changes in a more realistic operational environment. Results demonstrate that the proposed method exhibits good overall performance in complex backend anomaly detection tasks, indicating the feasibility and application potential of introducing a causal-aware mechanism into backend operational state evolution analysis.

From an application perspective, this research has significant practical implications for intelligent operation and maintenance, online service assurance, cloud platform stability maintenance, and risk early warning for complex digital infrastructure. With the continuous development of internet platforms, industrial internet, fintech systems, smart city service networks, and large-scale edge collaboration platforms, backend systems are handling increasingly frequent, complex, and sensitive business requests.

If anomaly identification is not timely or accurate, it often leads to service degradation, link congestion, resource waste, and even critical business interruptions. The method proposed in this paper can extract more discriminative state information from multi-source operational data and maintain strong anomaly perception capabilities under complex dependencies and dynamic disturbances. This provides new technical support for building more reliable automated monitoring mechanisms, improving fault early warning efficiency, and enhancing system resilience. At the same time, this research also provides a scalable approach for the application of causal modeling methods in backend operation and maintenance scenarios, which is of positive significance for promoting the transformation of anomaly detection technology from experience-driven to mechanism-aware driven.

Future work can be further expanded and deepened in several directions. On the one hand, the method's adaptability to larger-scale service topologies, sparser anomaly samples, and more complex cross-domain operating environments can be further enhanced, enabling it to better serve unified anomaly perception tasks in multi-cluster, multi-tenant, and cross-platform collaborative scenarios. On the other hand, based on existing state evolution modeling, we can further enhance the capabilities of anomaly root cause localization, anomaly propagation path tracing, and interpretable analysis for operational decision-making. This will enable the model to not only identify anomalies but also provide more direct support for subsequent handling strategies. Furthermore, with the development of large-scale model-based operational agents, adaptive resource regulation, and closed-loop fault recovery technologies, causal awareness anomaly detection is expected to be further integrated with automatic diagnosis, dynamic scheduling, and intelligent repair mechanisms, forming a full-process intelligent operation and maintenance framework for complex online service backends. Therefore, this research not only has clear theoretical value but also provides a practically significant reference path for the continuous evolution and high-reliability operation of related application fields.

References

[1] M. Panahandeh, A. Hamou-Lhadj, M. Hamdaqa, et al., "ServiceAnomaly: An anomaly detection approach in microservices using distributed traces and profiling metrics," Journal of Systems and Software, vol. 209, p. 111917, 2024.

[2] J. Chen, F. Liu, J. Jiang, G. Zhong, D. Xu, Z. Tan and S. Shi, "TraceGra: A trace-based anomaly detection for microservice using graph deep learning", Computer Communications, vol. 204, pp. 109-117, 2023.

[3] Z. Zeng, Y. Zhang, Y. Xu, et al., "Traceark: Towards actionable performance anomaly alerting for online service systems," in 2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Practice, 2023, pp. 258-269.

[4] C. Lee, T. Yang, Z. Chen, et al., "Eadro: An end-to-end troubleshooting framework for microservices on multi-source data," in 2023 IEEE/ACM 45th International Conference on Software Engineering, 2023, pp. 1750-1762.

[5] S. Zhang, P. Jin, Z. Lin, et al., "Robust failure diagnosis of microservice system through multimodal data," IEEE Transactions on Services Computing, vol. 16, no. 6, pp. 3851-3864, 2023.

[6] Y. Zhu, J. Wang, B. Li, et al., "Microirc: Instance-level root cause localization for microservice systems," Journal of Systems and Software, vol. 216, p. 112145, 2024.

[7] W. Tian, H. Zhang, N. Yang, et al., "Graph-based root cause localization in microservice systems with protection mechanisms," International Journal of Software Engineering and Knowledge Engineering, vol. 33, no. 08, pp. 1211-1238, 2023.

[8] L. Wang and C. Y. Chang, "Structure-Aware Root Cause Localization for Microservice Backends via Joint Graph Representation Learning and Causal Inference," 2024.

[9] T. Ding and Y. Yuan, "Behavior Representation Learning for Reliable Monitoring of AI Inference Backends," 2024.

[10] Y. Lyu, J. Jiang, and J. Hu, "Frequency-Guided Transformer Modeling with Spectral Enhancement for Accurate Time Series Forecasting," 2024.

[11] Z. Qing and F. Huang, "Attention-Enhanced Network Traffic Prediction for Intelligent Load Balancing Optimization," 2025.

[12] Y. Mei, "Unified Semi-Supervised and Robust Optimization Framework for Low-Resource Noisy Data Mining," 2024.

[13] P. Hu, G. Uziel, and X. Ma, "Risk-Constrained Reinforcement Learning for SLO-Driven Cloud-Native Elastic Scaling," 2024.

[14] C. Zhang, X. Peng, C. Sha, et al., "Deeptralog: Trace-log combined microservice anomaly detection through graph-based deep learning," in Proceedings of the 44th International Conference on Software Engineering, 2022, pp. 623-634.

[15] Z. Xie, H. Xu, W. Chen, et al., "Unsupervised anomaly detection on microservice traces through graph vae," in Proceedings of the ACM Web Conference 2023, 2023, pp. 2874-2884.

[16] S. Zhang, Z. Pan, H. Liu, et al., "Efficient and robust trace anomaly detection for large-scale microservice systems," in 2023 IEEE 34th International Symposium on Software Reliability Engineering, 2023, pp. 69-79.

[17] J. Liu, T. Yang, Z. Chen, et al., "Practical anomaly detection over multivariate monitoring metrics for online services," in 2023 IEEE 34th International Symposium on Software Reliability Engineering, 2023, pp. 36-45.

[18] Z. Chen, J. Liu, Y. Su, et al., "Adaptive performance anomaly detection for online service systems via pattern sketching," in Proceedings of the 44th International Conference on Software Engineering, 2022, pp. 61-72.

[19] C. Duan, T. Jia, H. Cai, et al., "Afalog: A general augmentation framework for log-based anomaly detection with active learning," in 2023 IEEE 34th International Symposium on Software Reliability Engineering, 2023, pp. 46-56.

[20] Z. Yu, C. Pei, S. Zhang, et al., "Autokad: Empowering kpi anomaly detection with label-free deployment," in 2023 IEEE 34th International Symposium on Software Reliability Engineering, 2023, pp. 13-23.