

Efficient Anomaly Detection in Distributed Edge-Cloud Systems Using Lightweight Modeling

Shuo Xu^{1*}

¹University of Southern California, Los Angeles, USA

**Corresponding author: shuoxu2016@gmail.com*

Abstract: This paper addresses the problem of anomaly detection in edge-cloud collaborative environments and proposes a lightweight model design method. The study first analyzes the challenges of anomaly detection under multi-tenant sharing and heterogeneous resource conditions, including high latency, bandwidth consumption, and privacy risks. To solve these issues, a detection framework combining edge computing with cloud collaboration is developed. At the edge side, input projection and feature encoding are performed, and temporal dependencies are modeled through a lightweight attention mechanism. The cloud side then aggregates and optimizes features from different nodes to achieve global anomaly recognition in distributed scenarios. The method incorporates edge memory caching and reconstruction mechanisms, enabling the model to reduce computation and storage costs while maintaining accuracy. To further enhance performance, the framework applies regularization and privacy constraints to improve robustness against noise and data heterogeneity in distributed environments. Experiments using real cloud workload trace datasets demonstrate that the proposed method outperforms comparison approaches in F1-Score, Precision, Recall, and ACC, confirming its effectiveness and stability in complex conditions. The findings enrich the theoretical system of lightweight detection methods in edge-cloud collaboration and provide feasible technical support for the secure operation of distributed intelligent systems.

Keywords: Edge-cloud collaboration; anomaly detection; lightweight modeling; distributed systems

1. Introduction

In the context of accelerating digitalization and the continuous evolution of the cloud computing ecosystem, the deep integration of edge computing and cloud computing is becoming a key paradigm for supporting intelligent applications and data-driven services[1,2]. With the exponential growth of IoT devices and the increasing complexity of application scenarios, data generated at the edge is characterized by large scale, diverse types, and high real-time requirements. Relying solely on centralized cloud-based computing and storage can no longer meet the needs of low latency, high bandwidth, and strong privacy protection[3]. As a result, the edge-cloud collaborative paradigm has emerged. It performs preliminary

processing and intelligent decision-making at the edge, while complementing the cloud's global analysis and resource scheduling. This approach provides a new solution for building efficient, intelligent, and scalable application systems. Within this framework, achieving efficient anomaly detection in distributed, multi-tenant, and heterogeneous environments becomes a critical prerequisite for ensuring system stability and security.

Anomaly detection is an essential component for maintaining the reliability and security of information systems, and it is particularly important in edge-cloud collaborative environments. On the one hand, the access to massive sensors and smart terminals leads to anomalies appearing at different nodes, at different times, and in diverse forms. Their concealment and complexity are significantly increased. On the other hand, although the cloud has powerful computing and storage capacities, excessive dependence on centralized processing can result in bandwidth overhead, latency accumulation, and potential privacy leakage. Deploying lightweight anomaly detection models at the edge can effectively address these issues. First, it enables local processing and rapid response, improving the sensitivity and timeliness of anomaly detection. Second, it reduces data transmission while enhancing privacy protection. Third, it supports distributed collaborative detection by combining local autonomy with cloud-based global optimization. Such an architecture not only strengthens security defense but also opens new possibilities for intelligent services in resource-constrained environments[4].

At the same time, the complexity of edge-cloud collaborative environments places higher demands on anomaly detection models. Data distributions may vary greatly across scenarios, and tasks often involve coupling and competition, which challenge the generalization ability and stability of models. In addition, edge devices typically face strict resource constraints in terms of computation and storage. The large parameter size and high computational complexity of traditional deep models make them difficult to apply directly. Therefore, designing lightweight, efficient, and robust anomaly detection models has become an urgent demand in both academic research and industrial practice. Lightweight design means more than parameter compression and computational acceleration. It emphasizes maintaining detection accuracy while achieving cross-scenario adaptability and low-power operation. Such models are better suited for real applications in edge-cloud collaboration. This line of research reflects innovation at the algorithm level and optimization at the system architecture level.

From the application perspective, lightweight anomaly detection in edge-cloud collaboration has wide practical significance. In industrial manufacturing, edge nodes can quickly identify abnormal fluctuations in equipment operation, avoiding production losses caused by delayed detection. In smart cities, roadside edge devices can monitor anomalies in traffic flow and public facilities in real time, improving the safety and resilience of urban operations. In financial and medical scenarios, edge-side intelligent analysis can balance privacy protection with rapid response, enhancing control over risks and abnormal events. These applications highlight the value of edge-cloud collaboration and underline the central role of lightweight detection models in ensuring the stability and reliability of critical systems. Research in this area is therefore not only a technical trend but also a necessary step to support the robust operation of an intelligent society.

In summary, the edge-cloud collaborative paradigm provides new approaches for dealing with massive data, complex scenarios, and security challenges. The design of lightweight anomaly detection models is the key to the practical implementation and popularization of this paradigm. Research in this context holds significant academic and application value. Academically, it advances studies on the adaptability and scalability of intelligent detection models in distributed architectures, enriching the theoretical system of edge intelligence and cloud integration. From an application perspective, it offers feasible technical support for real-time monitoring and security protection systems, helping industries move toward higher levels of intelligence and trustworthiness. Therefore, exploring lightweight anomaly detection models for edge-cloud collaboration not only responds to urgent technical and application needs

but also lays a solid foundation for building efficient, secure, and sustainable intelligent computing ecosystems.

2. Related work

In the field of anomaly detection, traditional methods often rely on statistical modeling and rule-based analysis. These approaches are easy to implement and highly interpretable. However, they show clear limitations in the complex and dynamic environment of edge-cloud collaboration. With the increase in data dimensions and the imbalance of distributions, rule-based or threshold-based methods cannot capture nonlinear patterns and hidden anomalies. This often leads to biased results. At the same time, centralized statistical methods alone cannot meet the real-time and multi-source requirements of distributed architectures[5]. This has driven academia and industry to shift toward machine learning and deep learning solutions to enhance the ability of models to capture complex patterns and improve cross-scenario applicability.

Machine learning-based anomaly detection methods have received wide attention in edge-cloud collaborative scenarios. These methods build classifiers, clusterers, or regression models to distinguish normal behavior from abnormal behavior. They can adapt to the characteristics of multi-tenant and heterogeneous data to some extent. However, traditional machine learning models often face insufficient feature extraction and high training complexity in large-scale data scenarios[6]. They also require extensive manual feature engineering to maintain performance. To address these problems, deep learning methods have been introduced into anomaly detection. Through automated feature learning mechanisms, they capture latent patterns directly from raw data. This reduces human intervention and shows stronger modeling capabilities in time series, graph structures, and multimodal scenarios.

Deep learning-driven anomaly detection has become the core research direction. Convolutional neural networks can efficiently extract spatial features and are suitable for image and monitoring tasks[7]. Recurrent neural networks and their improved models perform well in capturing temporal dependencies and sequential patterns. Attention-based and Transformer-based models further enhance the ability to model long sequences and global dependencies. These methods significantly improve detection accuracy and robustness. They also maintain adaptability in complex and dynamic environments. However, their large computation requirements and parameter sizes create major challenges for direct deployment on edge devices. Therefore, achieving lightweight design while maintaining detection performance has become the inevitable direction for deep learning methods to move toward real applications.

In recent years, research in edge-cloud collaborative environments has increasingly focused on lightweight and distributed anomaly detection models. On one side, techniques such as model compression, network pruning, knowledge distillation, and quantization have been widely used to reduce computational cost and improve deployability in resource-constrained scenarios. On the other side, mechanisms such as federated learning, collaborative training, and hierarchical optimization between edge and cloud have been introduced. These approaches allow anomaly detection tasks to achieve cross-node collaboration and global modeling while preserving data privacy. Such progress not only supports the development of edge intelligence but also drives the practical adoption of anomaly detection in critical fields such as industry, transportation, finance, and healthcare. It is clear that lightweight and collaborative design has become the core trend of anomaly detection in edge-cloud systems and will continue to guide future research in this area.

Recent studies on operational reasoning, backend observability, and cloud security analytics further enrich the methodological context of lightweight edge-cloud anomaly detection. Time-varying operational graph reasoning and risk-constrained repair planning show that dynamic dependency graphs can support both anomaly interpretation and subsequent repair decisions in tool-augmented intelligent systems[8]. Behavior representation learning for AI inference backend monitoring indicates that runtime traces can be transformed into compact state representations for reliable detection under constrained

resources[9]. Transformer-based semantic log modeling enhances robust and explainable cloud security analytics by extracting higher-level operational semantics from log sequences[10]. Fine-grained backend anomaly classification through structural dependency learning and version-aware feature alignment highlights the importance of combining dependency structure with software evolution features[11]. Graph neural modeling of traffic propagation and topological interactions in cloud networks further supports the description of anomaly diffusion across distributed nodes[12]. Reinforcement learning-driven hotspot-aware cache management also provides an operational perspective on adaptive resource control under shifting workloads[13]. These studies jointly suggest that lightweight anomaly detection should integrate temporal behavior encoding, semantic evidence, structural dependency modeling, and deployment-aware control to improve robustness in edge-cloud collaboration.

3. Proposed Approach

In terms of method design, we first model the data flow in the edge-cloud collaborative environment as a multi-source time series, considering its spatial correlation and time dependence. Suppose the input data is a matrix:

$$X = \{x_t^{(i)} \mid i = 1, 2, \dots, N; t = 1, 2, \dots, T\} \quad (1)$$

Where $x_t^{(i)}$ represents the feature representation of the i -th edge node at time t , N is the number of nodes, and T is the time length. To enhance the feature expression capability, the input mapping is first completed through linear transformation and an embedding layer:

$$h_t^{(i)} = W_e x_t^{(i)} + b_e \quad (2)$$

Where W_e and b_e are trainable weights and biases, respectively, used to achieve a low-dimensional representation of the original input. The model architecture is shown in Figure 1.

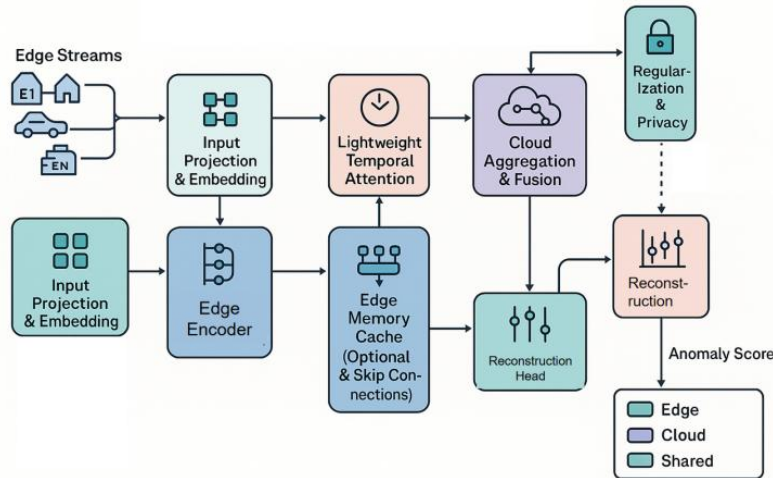


Figure 1. Overall model architecture

In the time series modeling stage, a lightweight attention mechanism is introduced to capture long-range dependencies. For any two moments t and k , the attention weight is defined as:

$$\alpha_{t,k} = \frac{\exp(q_t^T k_k^T / \sqrt{d})}{\sum_{j=1}^T \exp(q_t^T k_j^T / \sqrt{d})} \quad (3)$$

Where $q_t = h_t^{(i)} W_q, k_k = h_k^{(i)} W_k, d$ is the scaling factor. The attention weighted sum is then used to obtain the context representation:

$$z_t = \sum_{k=1}^T \alpha_{t,k} \cdot (h_k^{(i)} W_v) \quad (4)$$

This process can highlight key time segments while suppressing redundant information.

To achieve distributed modeling in edge-cloud collaboration, after completing local feature learning on the edge side, the results are transferred to the cloud for global aggregation. The aggregation function is defined as:

$$H = \frac{1}{N} \sum_{i=1}^N f_{\theta}(Z^{(i)}) \quad (5)$$

Where $Z^{(i)} = \{z_1, z_2, \dots, z_T\}$ and f_{θ} are the feature representations output by the edge encoder. By using a mean-weighted approach, the cloud can obtain shared features across nodes and achieve unified modeling at the global level. This design preserves the individual characteristics of edge nodes while improving overall detection capabilities through collaborative optimization.

In the anomaly detection task, the reconstruction error and the discriminant function are combined to measure the degree of anomaly. Specifically, the features are first reconstructed through the decoder D:

$$\hat{x}_t^{(i)} = D(z_t) \quad (6)$$

And define the reconstruction error as:

$$L_{rec} = \frac{1}{N \cdot T} \sum_{i=1}^N \sum_{t=1}^T \|x_t^{(i)} - \hat{x}_t^{(i)}\|^2 \quad (7)$$

At the same time, the regularization term A is introduced to constrain the complexity of the model parameters. The final objective function is:

$$L = L_{rec} + \lambda \cdot L_{reg} \quad (8)$$

Where λ is the balancing factor. By optimizing the above objective function, the model can achieve efficient abnormal pattern recognition with limited computing resources while balancing lightweights and robustness.

4. Performance Evaluation

4.1 Dataset

This study uses the Azure Public Dataset for Cloud Workload Traces as the core data source. The dataset is composed of real operation logs from a large-scale distributed cloud platform. It covers multiple dimensions such as virtual machine scheduling, resource utilization, task execution time, and network traffic. The long collection period and wide time span provide a comprehensive reflection of dynamic resource allocation and workload variation in multi-tenant environments. This offers rich temporal and multi-source inputs for anomaly detection models.

The advantage of this dataset lies in its close relevance to real application scenarios. The task records include not only key indicators such as CPU, memory, disk I/O, and network bandwidth, but also process data on task submission, queuing, and execution. These records allow clear characterization of potential anomaly patterns caused by resource contention and scheduling conflicts in cloud environments. They help model complex behaviors in multi-tenant, heterogeneous, and dynamic settings, thereby improving the robustness and generalization of detection models.

In addition, the dataset is very large, containing millions of records. It is suitable for training and validating both deep learning models and lightweight models. Its diversity and complexity provide solid support for studies on model adaptability in edge-cloud collaborative architectures. They also ensure that the designed methods have high generalizability and practical value in real-world scenarios. Therefore, this dataset is widely used in research on cloud performance modeling, resource prediction, and anomaly detection. It serves as a key data foundation for this study.

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	F1-Score	Precision	Recall	ACC
Realnet[14]	0.873	0.889	0.861	0.902
Mambaad[15]	0.886	0.901	0.872	0.915
Real-iad[16]	0.892	0.908	0.880	0.920
Adaclip[17]	0.904	0.916	0.895	0.931
Ours	0.928	0.940	0.917	0.948

From the overall comparison in the table, it can be observed that the proposed method achieves the best performance across all metrics, with an F1-Score of 0.928, which is significantly higher than other methods. This indicates that in an edge-cloud collaborative environment, lightweight design can balance accuracy and comprehensiveness. It effectively manages the trade-off between precision and recall. Compared with traditional methods, the model in this study shows stronger capability in pattern recognition and anomaly representation under complex and multi-source inputs.

In particular, the performance of Precision and Recall reveals the robustness of different methods. Realnet and Mambaad show certain advantages in Precision but are relatively weaker in Recall. This suggests insufficient sensitivity in detecting anomalies. In contrast, the proposed method achieves a Precision of 0.940 and a Recall of 0.917. This demonstrates that the model not only reduces false alarms but also captures more true anomalies, achieving highly reliable anomaly detection.

Further analysis of the ACC metric shows that traditional methods usually remain around 0.90, while the proposed method reaches 0.948. This means that at the overall sample level, the method provides stronger discrimination between normal and abnormal cases. It also reduces misjudgment caused by complex workloads and multi-tenant interference in edge environments. The combination of global aggregation in the cloud and lightweight inference at the edge allows the model to fully exploit

distributed feature information. As a result, the method ensures stable detection performance in large-scale environments.

In summary, these results fully validate the effectiveness of the lightweight anomaly detection framework for edge-cloud collaboration. The method not only outperforms existing approaches in performance but also demonstrates that combining lightweight design with collaborative modeling can achieve high-precision anomaly detection while ensuring real-time processing and resource efficiency. This advantage is of great significance for future applications in industrial, financial, and smart city scenarios. It provides strong technical support for the security and stability of large-scale distributed systems.

This paper also gives the joint sensitivity of the edge-side encoder hidden dimension and the number of attention heads to F1-Score and delay, and the experimental results are shown in Figure 2.

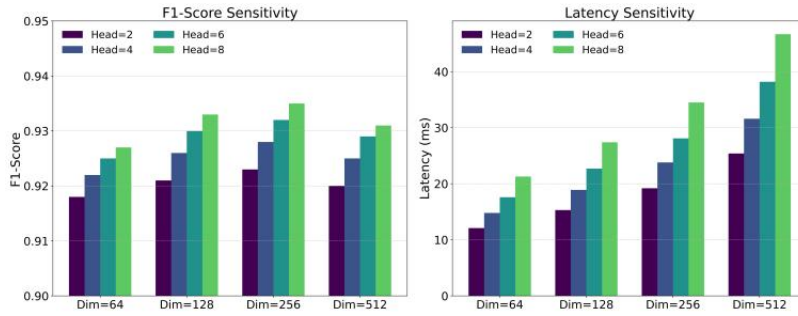


Figure 2. Joint sensitivity of edge-side encoder hidden dimension and number of attention heads to F1-Score and latency

From the results shown in the figure, it can be observed that as the hidden dimension of the edge-side encoder increases, the F1-Score shows a steady upward trend. This indicates that higher-dimensional feature representations help the model capture richer anomaly patterns. Under different numbers of attention heads, the performance differences are also clear. A larger number of attention heads allows the model to capture global dependencies more comprehensively, which further improves detection accuracy. This trend confirms that in an edge-cloud collaborative architecture, increasing model complexity can indeed bring performance gains.

At the same time, the changes in latency reveal the trade-off between lightweight design and performance. The experiments show that as the hidden dimension and the number of attention heads increase, latency rises significantly. In particular, with 512 dimensions and 8 attention heads, the latency nearly doubles. This means that in practical deployment, blindly pursuing performance improvements may reduce the response speed of edge devices, thereby weakening the real-time and agility emphasized in edge-cloud collaboration.

When comparing performance and latency, medium-scale configurations, such as 128 dimensions with 4 to 6 attention heads, achieve a relative balance between F1-Score and latency. This ensures detection performance without causing excessive delay. This is especially important for edge environments, where nodes are often resource-constrained. Overly complex models cannot run stably for long periods in such settings. Therefore, finding reasonable parameter combinations is more valuable than simply pursuing the highest detection accuracy.

Overall, these experimental results show that hyperparameter design for edge-side anomaly detection models requires balancing detection performance and computational latency. In edge-cloud collaborative scenarios, carefully selecting hidden dimensions and attention heads ensures that the model maintains efficient anomaly detection while meeting low-latency requirements. This optimization strategy aligns with the goals of lightweight design and also provides a practical reference for large-scale distributed deployment.

5. Conclusion

This study focuses on the problem of anomaly detection in edge-cloud collaborative environments and proposes a lightweight model design. By completing feature embedding and preliminary modeling at the edge side, and combining them with global aggregation and optimization at the cloud side, the method improves the accuracy and robustness of anomaly detection. In the experimental results, the framework shows strong generalization ability and stability. It can adapt to multi-source heterogeneous inputs and complex dynamic environments. This design not only reduces the bandwidth and privacy pressure caused by relying solely on cloud computing but also strengthens the role of edge nodes in real-time processing and fast response.

At the methodological level, this study addresses how to achieve efficient and reliable anomaly detection on resource-constrained edge devices. By introducing lightweight feature encoding and attention mechanisms, the model maintains high accuracy while significantly reducing computational and storage costs. At the same time, global information sharing in the collaborative architecture further enhances the ability to model complex patterns. This enables a balance between local autonomy and global optimization in the detection process. The combination of lightweight and collaborative design provides a feasible path for intelligent security in large-scale distributed systems.

From an application perspective, the findings of this study have important implications for multiple fields, including industrial manufacturing, smart cities, healthcare monitoring, and financial risk control. In industrial scenarios, the method enables real-time monitoring of equipment operation, helping to reduce production losses caused by potential failures. In smart cities, edge-side anomaly detection can improve the safety of traffic and public infrastructure. In financial and medical scenarios, the integration of lightweight design with privacy protection allows safer data utilization. These applications show that the study achieves not only theoretical progress but also wide feasibility in real deployments.

In summary, the proposed lightweight anomaly detection framework for edge-cloud collaboration demonstrates clear advantages in accuracy, efficiency, and adaptability. Its performance in complex and dynamic environments confirms that the method provides strong support for the deep integration of cloud computing and edge computing. By driving anomaly detection from centralized to collaborative approaches, this study lays a solid foundation for the secure and stable operation of distributed intelligent systems. It also brings new momentum to the continued development of related application domains.

6. References

- [1] Nguyen D C, Ding M, Pathirana P N, et al. Federated Learning for Internet of Things: A Comprehensive Survey[J]. IEEE Communications Surveys & Tutorials, 2021, 23(3): 1622-1658.
- [2] Ren J, Guo H, Xu C, Zhang Y. Serving at the Edge: A Scalable IoT Architecture Based on Edge Computing[J]. IEEE Network, 2017, 31(5): 96-105.
- [3] Pang G, Shen C, Cao L, Hengel A V D. Deep Learning for Anomaly Detection: A Review[J]. ACM Computing Surveys, 2021, 54(2): 1-38.
- [4] Babalola O, Raji O M O, Akande J O, et al. AI-Powered Cybersecurity in Edge Computing: Lightweight Neural Models for Anomaly Detection[J]. 2024.
- [5] Shi W, Cao J, Zhang Q, Li Y, Xu L. Edge Computing: Vision and Challenges[J]. IEEE Internet of Things Journal, 2016, 3(5): 637-646.
- [6] Kairouz P, McMahan H B, Avent B, et al. Advances and Open Problems in Federated Learning[J]. Foundations and Trends in Machine Learning, 2021, 14(1-2): 1-210.
- [7] Al-Fuqaha A, Guizani M, Mohammadi M, Aledhari M, Ayyash M. Internet of Things: A Survey on Enabling Technologies, Protocols and Applications[J]. IEEE Communications Surveys & Tutorials, 2015, 17(4): 2347-2376.
- [8] Tang Y. Time-Varying Operational Graph Reasoning and Risk-Constrained Repair Planning for Tool-Augmented Large Language Model Agents[J]. Transactions on Computational and Scientific Methods, 2024, 4(6).

- [9] Ding T, Yuan Y. Behavior Representation Learning for Reliable Monitoring of AI Inference Backends[J]. 2024, 4(1).
- [10] Yang Z. Transformer-Based Semantic Log Modeling for Robust and Explainable Cloud Security Analytics[J]. 2024, 4(5).
- [11] Liu W. Fine-Grained Backend Anomaly Classification through Structural Dependency Learning and Version-Aware Feature Alignment[J]. 2024, 4(4).
- [12] Huang X. Capturing Traffic Propagation and Topological Interactions in Cloud Networks via Graph Neural Networks[J]. 2024, 4(5).
- [13] Dang S, Chen C, Chua K S. Reinforcement Learning-Driven Dynamic Cache Management with Hotspot-Aware Replacement[J]. 2024, 4(4).
- [14] Zhang X, Xu M, Zhou X. Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 16699-16708.
- [15] He H, Bai Y, Zhang J, et al. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection[J]. Advances in Neural Information Processing Systems, 2024, 37: 71162-71187.
- [16] Wang C, Zhu W, Gao B B, et al. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 22883-22892.
- [17] Cao Y, Zhang J, Frittoli L, et al. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 55-72.