https://doi.org/10.3311/PPee.44105

Periodica Polytechnica Electrical Engineering and Computer Science

IoT Time-series Anomaly Detection Using a Hybrid Transformer-GRU Fusion Model

YiTao Liang[1*], Yuangang Li[2], Yingying Xu[3]

[1] Information and Network Center, Shanghai Business School, 458 Huancheng East Road, Fengxian District, 201400 Shanghai, China
[2] Faculty of Business Information, Shanghai Business School, 2271 Zhongshan West Road, Xuhui District, 200235 Shanghai, China
[3] Information and Network Center, Shanghai Business School, 458 Huancheng East Road, Fengxian District, 201400 Shanghai, China
[*] Corresponding author, e-mail: 21100025@sbs.edu.cn

Received: 25 March 2026, Accepted: 25 June 2026, Published online: 26 August 2026

Abstract

The rapid growth of Internet of Things (IoT) systems has generated massive, complex, and highly dynamic time-series data, making anomaly detection essential for system security and operational stability. However, traditional methods based on pointwise representations, statistical thresholds, or simple pairwise associations often struggle to capture complex temporal dependencies and feature interactions in IoT data. To address these challenges, we propose Residual GRU-Attention Anomaly Detector (RGAAD), an unsupervised framework for IoT time-series anomaly detection. RGAAD integrates residual GRU modeling, adaptive self-attention, and gated multi-scale feature fusion to jointly capture temporal dependencies and feature correlations. Extensive experiments on SMD, SWaT, and MSL demonstrate that RGAAD achieves highly competitive performance and consistently outperforms strong baseline methods. These results confirm that explicitly modeling pointwise anomalies and temporal relationships is effective for real-world IoT monitoring.

Keywords

Internet of Things, transformer, anomaly detection, unsupervised time series, self-attention mechanism

1 Introduction

With the rapid advancement of Internet of Things (IoT) technologies [1], smart devices and sensors have been widely deployed across diverse real-world scenarios, continuously generating massive time-series data. Accurate anomaly detection in these data is critical for ensuring system stability and security. However, the increasing scale, complexity, and temporal dynamics of IoT time series make reliable anomaly detection highly challenging. Traditional unsupervised anomaly detection methods, such as Local Outlier Factor (LOF) [2] and Support Vector Data Description (SVDD) [3], have achieved certain progress in identifying anomalous patterns. However, these methods often fail to fully exploit the temporal dynamics of time-series data and typically exhibit limited generalization when applied to unseen real-world scenarios.

Driven by recent advances in deep learning [4–6], self-supervised time-series anomaly detection methods have achieved promising results, especially through reconstruction-based and autoregressive learning objectives. These methods primarily learn pointwise temporal representations from normal patterns. Nevertheless, the sparsity of anomalies and the complexity of real-world time series make it difficult to reliably separate abnormal behaviors from normal variations. More importantly, their insufficient modeling of inter-variable dependencies limits their capability to capture relational and system-level anomalies.

Motivated by the limitations of existing methods, we propose Residual GRU-Attention Anomaly Detector (RGAAD), a Transformer-based anomaly detection framework for time-series data [7–10]. By exploiting self-attention, RGAAD captures global dependencies and produces time-dependent relevance distributions that reflect intrinsic temporal structures, including periodicity, trends, and dynamic fluctuations. We define these temporal relevance representations as feature dynamics. In addition, a gated feature fusion module is introduced to adaptively reweight multi-scale features over time, improving the model's ability to detect subtle anomalous variations.

Moreover, we observe that anomalous points often exhibit limited global correlations and primarily interact with neighboring timestamps. We define this property as temporal affinity, which characterizes the stronger local

Cite this article as: Liang, Y., Li, Y., Xu, Y. "IoT Time-series Anomaly Detection Using a Hybrid Transformer-GRU Fusion Model", Periodica Polytechnica Electrical Engineering and Computer Science, 2026. https://doi.org/10.3311/PPee.44105

temporal dependencies of anomalies. To capture this behavior, we introduce an adaptive temporal scaling mechanism that dynamically adjusts the influence of temporal distance according to data characteristics, thereby improving adaptability to both periodic and non-periodic series. In contrast, normal points usually maintain broader and more stable temporal dependencies. The resulting discrepancy between normal and anomalous points in feature dynamics and correlation patterns is defined as correlation deviation, which serves as an effective criterion for anomaly identification. Extensive experiments on real-world benchmark datasets, including SMD, SWaT, and MSL, show that our approach consistently outperforms existing methods. These results validate its effectiveness in handling complex IoT time-series data and highlight its potential as a robust solution for anomaly detection in dynamic environments.

2 Related work

Real-world time-series data are often collected from multiple sources, making inter-variable dependency modeling crucial for improving anomaly detection. To address this challenge, various approaches have been proposed. For example, Jones et al. [11] employed nonlinear modeling functions and statistical smoothed trajectory (SST) features to characterize relationships among correlated variables and detect anomalies. Similarly, MAD-GAN [12] leverages Long Short-Term Memory (LSTM) networks to capture sequential dependencies and integrates Generative Adversarial Networks (GANs) to learn complex interactions among multivariate time-series variables in an unsupervised manner.

Existing time-series anomaly detection methods can be broadly categorized into representation learning, statistical thresholding, symbolic analysis, and residual-based modeling. Autoencoders have been used to learn low-dimensional representations for high-dimensional anomaly detection [13], while ODCA [14] exploits cross-correlation and outlier ranking to simplify high-dimensional observations. DSPOT focuses on adaptive threshold estimation under data drift, whereas symbolic discretization and grammar-based reduction identify rare subsequences as anomalies [15]. Residual-based methods, including ARX and ANN models [16], detect anomalies by extracting temporal characteristics and applying statistical tests to residual extrema.

To address the limitations of CNNs in modeling irregular and non-Euclidean data, recent studies have introduced Graph Neural Networks (GNNs) [17], which extend convolutional operations to graph-structured representations. By exploiting graph-based message passing, GNNs can effectively capture local structural patterns and inter-node dependencies. In IoT scenarios, where sensor relationships are often unknown, Graph Deviation Network (GDN) [18] learns adaptive pairwise correlations via cosine similarity to construct sensor graphs, and detects anomalies based on forecasting deviations with attention mechanisms. Similarly, MTAD-GAN [19] incorporates feature-wise and temporal graph attention to model multivariate dependencies, jointly optimizing forecasting and reconstruction objectives, followed by automatic thresholding for anomaly detection.

Attention mechanisms [7] have shown strong capability in modeling sequential data across natural language processing [20], audio analysis [21], computer vision [10], and time-series analysis [5], [8]. In time-series anomaly detection, GTA [1] combines graph-based sensor relationship modeling with temporal attention and reconstruction-based detection. Anomaly Transformer [22] further extends Transformer-based anomaly detection by introducing Anomaly-Attention, which leverages association discrepancy to distinguish abnormal temporal patterns and improve detection robustness.

3 Methods

As shown in Fig. 1, RGAAD adopts a multi-stage framework for multivariate time-series anomaly detection. After normalization and window segmentation, raw sensor data are processed by a bidirectional GRU encoder, a multi-head self-attention module, and a variational autoencoder to capture temporal dependencies, time-step importance, and robust normal representations. RGAAD then models feature dynamics and temporal affinity, which are adaptively integrated through a gated fusion module. Anomalies are identified by the correlation deviation between the reconstructed output and the original input under dynamic thresholding. By jointly modeling pointwise features, pairwise correlations, and temporal context, RGAAD enhances anomaly detection performance in IoT scenarios.

3.1 Data preprocessing and feature engineering

The foundation of our anomaly detection system begins with careful data preparation, and the processing flow is shown in Fig. 2, which utilizes standardization to process the data.

For multivariate time series data containing m different sensor measurements, we first apply standardization to normalize the characteristic scale:

$$X_{std} = \frac{X - \mu}{\sigma}, \tag{1}$$

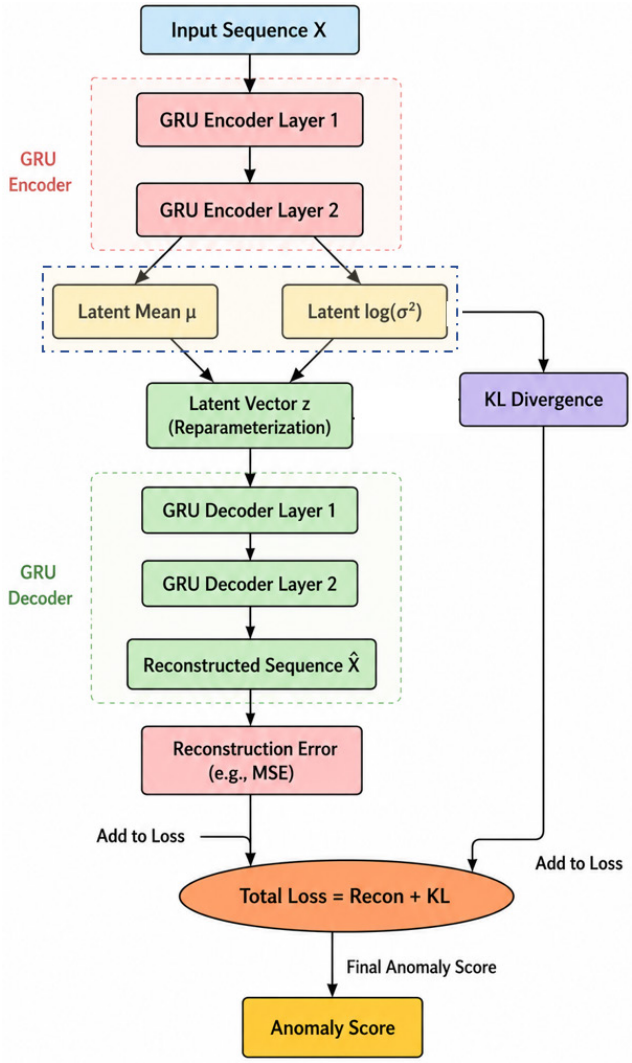


Fig. 1 Overview of residual GRU-attention anomaly detector

where μ and σ denote the feature-wise mean and standard deviation, respectively, computed exclusively from the training data. Therefore, the preprocessing procedure uses only training-set statistics and does not introduce information from the test set. This normalization step is essential for industrial sensor data, where different variables often exhibit substantially different numerical scales.

To effectively capture temporal patterns, we restructure the normalized time series into fixed-length subsequences using a sliding window approach:

$$S_t = \left[x_t, x_{t+1}, \ldots, x_{t+w-1}\right] \in \mathbb{R}^{w \times m} \tag{2}$$

The window size w determines the temporal context available to the model. Based on validation experimets, we set w = 100, which balances anomaly detectin reliability and computational efficiency. The sliding-window strategy converts continuous industrial sensor streams into fixed-length subsequences for model training and inference.

3.2 Neural network architecture design

As illustrated in Fig. 3, our model consists of three core components: a bidirectional GRU for temporal dependency modeling, a self-attention mechanism for capturing salient time-step relationships, and a probabilistic latent space module for robust representation learning. The detailed design is described below.

Bidirectional encoding with attention: The encoder component of our architecture employs a bidirectional GRU network to process input sequences in both temporal directions.

$$\begin{aligned} \vec{h}_t &= GRU\left(x_t, \vec{h}_{t-1}\right) \\ \overleftarrow{h}_t &= GRU\left(x_t, \overleftarrow{h}_{t-1}\right) \\ h_t &= \left[\vec{h}_t, \overleftarrow{h}_t\right] \end{aligned} \tag{3}$$

The bidirectional GRU models temporal dependencies in both directions, enabling the capture of cross-sensor lead–lag patterns. The concatenated hidden states provide a richer contextual representation by integrating forward and backward temporal information, which is beneficial for identifying early anomaly signals.

A multi-head self-attention mechanism processes the GRU outputs to identify the most diagnostically relevant time steps:

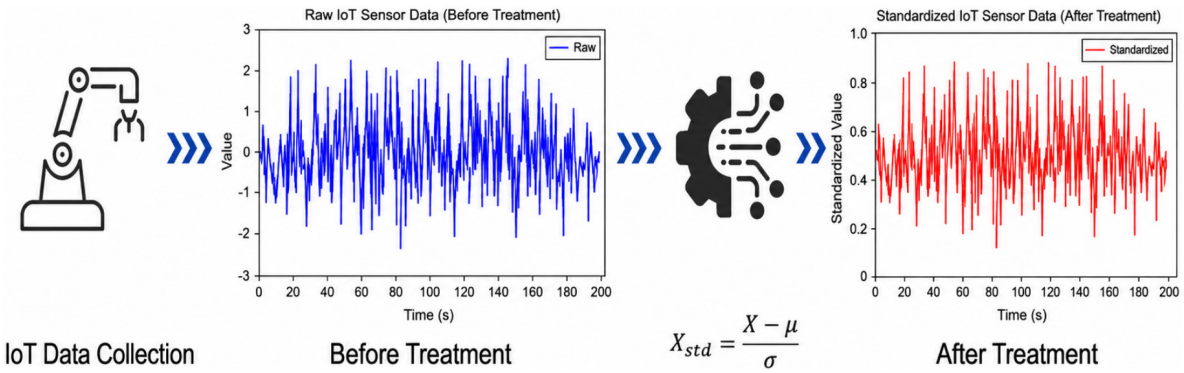


Fig. 2 Data processing process

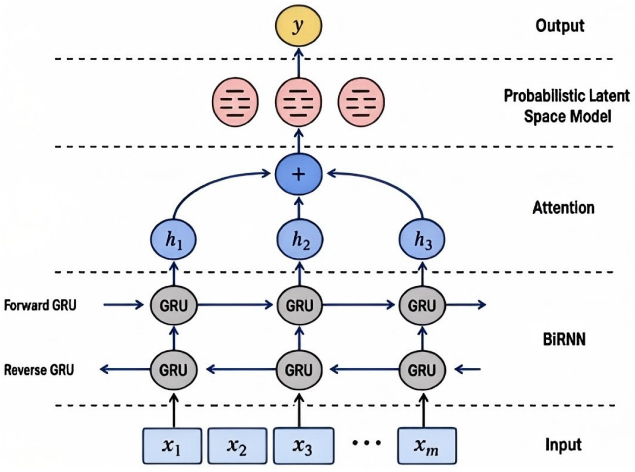


Fig. 3 RGAAD model architecture diagram

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V\ . \tag{4}$$

There are three main advantages to the attention mechanism: the mechanism permits the model to adaptively attend to the most informative elements of the input sequence, can be used to learn long-range dependencies, which may extend past the effective memory window of GRU, and can produce interpretable attention weights that can reveal which time steps are most important when making an anomaly detection decision.

Probabilistic latent space representation: The model learns a variational latent representation that captures the essential characteristics of normal operating patterns:

$$\mu = W_\mu h_{attn}\ , \tag{5}$$

$$\log\sigma^2 = W_\sigma h_{attn}\ , \tag{6}$$

where h_{attn} Represents the attended sequence encoding. The latent variable z is sampled using the reparameterization trick:

$$z=\mu+\sigma\odot\in,\in\sim N(0,I)\cdot \tag{7}$$

By introducing probabilistic latent modeling, the model can regularize the latent representation, account for uncertainty in normal operating patterns, and better identify anomalous samples that deviate from the learned latent distribution.

3.3 Reconstruction and anomaly scoring

The decoder reconstructs the input sequence by progressively expanding the latent representation to the original feature dimension. To enhance reconstruction quality and training stability, we incorporate residual connections into the architecture:

$$\hat{S}=GRU(z)+S\cdot \tag{8}$$

The skip connections stabilize training by improving gradient propagation and mitigating vanishing gradients. They also enhance normal-pattern reconstruction while allowing the model to emphasize residual discrepancies associated with potential anomalies.

For each test sequence, we compute a comprehensive reconstruction error metric:

$$e_i=\frac{1}{w\times m}\sum_{t=1}^{w}\sum_{j=1}^{m}\left|S_{i,t,j}-\hat{S}_{i,t,j}\right|. \tag{9}$$

By considering reconstruction errors across both temporal and feature dimensions, this metric provides a robust anomaly score for each test sequence. The anomaly decision is then determined by a dynamic threshold:

$$\tau=\mu_e+k\sigma_e\ . \tag{10}$$

Here, μ_e and σ_e are the mean and standard deviation of reconstruction errors estimated from normal data. The coefficient k controls the detection sensitivity, and validation results indicate that k = 3 achieves a favorable balance between false alarms and missed detections, approximately matching the 99.7th percentile under the Gaussian assumption.

3.4 Training methodology and optimization

The training of the model uses a composite loss measuring both reconstruction accuracy and latent space regularization:

$$L=L_{recon}+\beta L_{KL}\ , \tag{11}$$

where β = 0.1 was determined through hyperparameter optimization to provide the best balance between these competing objectives. The reconstruction loss uses mean squared error:

$$L_{recon}=\frac{1}{n}\sum_{i=1}^{n}\left\|S_i-\hat{S}_i\right\|^2. \tag{12}$$

This strongly penalizes large deviations while being computationally efficient for continuous sensor data. The *KL* divergence term:

$$L_{KL}=-\frac{1}{2}\sum_{j=1}^{J}\left(1+\log\sigma_j^2-\mu_j^2-\sigma_j^2\right)\cdot \tag{13}$$

Ensures the learned latent distributions remain close to the standard normal distribution, preventing overfitting and enabling better generalization.

The model is implemented in PyTorch and optimized using Adam with a learning rate of 0.001 and a batch size of 64. To enhance training stability and generalization, we

adopt gradient clipping with a threshold of 1.0, Xavier uniform initialization, and dropout with a rate of 0.1 between layers. All experiments are conducted on a workstation equipped with an Intel Core i7-12700F CPU, 64 GB DDR4 RAM, and an NVIDIA RTX 5090 GPU with 32 GB VRAM under Ubuntu 20.04 LTS. The model typically converges within 10–20 epochs, with an average epoch time of approximately 30 s on SMD, 45 s on SWaT, and 35 s on MSL. Validation-based early stopping is used to prevent overfitting and avoid unnecessary computational overhead.

3.5 Evaluation metrics and performance assessment

We evaluate model performance using point-level Precision, Recall, and F1-score. By assessing individual time steps rather than entire windows, this evaluation protocol provides more fine-grained and actionable results for anomaly localization.

These metrics are particularly valuable for anomaly detection tasks where classes are inherently imbalanced (typically < 1% anomalies in industrial settings). The point-level evaluation ensures our results are directly comparable with other approaches in the literature while accurately reflecting real-world monitoring scenarios where precise anomaly localization is crucial.

To ensure accurate evaluation, predictions are temporally aligned with ground-truth labels while accounting for offsets caused by the sliding-window construction. We report results on the entire test set, enabling a comprehensive assessment of model performance under all available operating conditions and anomaly categories.

4. Experiments

RGAAD is evaluated on three benchmark datasets: SMD, SWaT, and MSL. Table 1 summarizes their statistical details.

4.1 Datasets

1. SMD (Server Machine Dataset) [23]: The SMD dataset was proposed by Su et al. [23] and contains five weeks of 38-dimensional monitoring data from server machines at a large Internet company. Since it reflects real-world server operation conditions, it serves as a representative benchmark for multivariate time-series anomaly detection.

Table 1 Statistical details of SMD, SWaT and MSL datasets

Dataset	Window	Training	Validation	Test	Dimension
SMD	100	566,724	141,681	708,420	38
SWaT	100	396,000	99,000	449,919	51
MSL	100	46653	11664	73729	55

2. SWaT (Secure Water Treatment) [24]: The SWaT dataset was introduced by Mathur and Tippenhauer [24] and contains multivariate sensor data collected from a continuously operating water treatment testbed. With 51 sensor and actuator measurements, it provides a realistic benchmark for evaluating anomaly detection in critical infrastructure and industrial control systems.
3. MSL (Mars Science Laboratory) [25]: The MSL dataset, released by NASA and curated by Hundman et al., [25] contains 55-dimensional telemetry data from the Curiosity rover. Its anomaly labels are derived from official ISA reports. Since the training set contains only normal data and the test set includes real-world anomalies, MSL is widely used to evaluate generalization in aerospace anomaly detection.

4.2 Implementation details

The implementation combines a VAE with GRU-based sequence modeling, self-attention, and residual connections for time-series anomaly detection. After feature standardization and sliding-window segmentation, a GRU encoder captures temporal dependencies, while self-attention highlights informative time steps and long-range relationships. The latent representation is parameterized by mean and log-variance vectors and sampled using the reparameterization trick. A GRU decoder reconstructs the input sequence, with residual connections introduced to improve reconstruction fidelity and stabilize training. The model is optimized using a composite loss consisting of mean squared reconstruction error and KL divergence, weighted by $\beta = 0.1$. During inference, anomalies are detected by comparing reconstruction errors with a dynamic threshold defined as the mean plus three standard deviations of normal errors. The model is implemented in PyTorch with GPU acceleration, gradient clipping, early stopping, and mini-batch training, and is evaluated using Precision, Recall, and F1-score.

4.3 Baseline

We compared our model with nine existing baselines in different categories of anomaly detection methods in detail. In particular, we compared with traditional approaches that include OC-SVM [3], IsolationForest [26], and LOF [2], and also with time series models, which include LSTM [27]. We also compared our method to more recent and more advanced approaches, such as clustering-based ITAD [28], and reconstruction-based OmniAnomaly [23]. We also considered state-of-the-art deep learning models, such as InterFusion [29] and Anomaly Transformer [22], that are

the current state of the art in this area. We have chosen these baselines because our model is to be thoroughly tested in terms of how it performs over a broad range of anomaly detection methods.

4.4 Ablation study

To assess the contribution of key RGAAD components, we perform ablation studies by removing or replacing critical modules while keeping all hyperparameters and training settings fixed. The results on SMD and SWaT are reported using F1-score (%) as the main metric, as shown in Table 2.

4.5 Results analysis

As shown in Table 3, RGAAD achieves competitive performance across the SMD, SWaT, and MSL datasets. In particular, it obtains the best F1-score on MSL, indicating strong generalization capability on high-dimensional aerospace telemetry data. Although Anomaly Transformer achieves higher F1-scores on SMD and SWaT, RGAAD remains competitive while integrating GRU-based temporal modeling, self-attention, residual reconstruction, and probabilistic latent representation. These results suggest that the proposed hybrid architecture is effective for modeling complex temporal dependencies and feature correlations in multivariate IoT time-series data.

Table 2 Statistical details of SMD, SWaT and MSL datasets

Model variant	SMD	SWaT	MSL
Full RGAAD	87.00	85.37	94.10
w/o Attention	77.81	77.89	84.72
w/o GRU	83.22	81.45	89.47
w/o Residual	84.91	83.26	91.33
w/o VAE	82.13	79.04	87.51

Unlike traditional methods such as OC-SVM and Isolation Forest, which rely on shallow statistical features and have limited ability to model complex temporal dependencies, RGAAD learns intrinsic normal behavior through deep representation learning. Compared with advanced reconstruction-based models, including OmniAnomaly and InterFusion, RGAAD achieves further performance gains by integrating residual connections, gated feature fusion, and VAE-based probabilistic modeling, thereby improving the robustness of anomaly scoring.

Notably, even compared with the representative transformer-based method Anomaly Transformer (93.59% on MSL), RGAAD still achieves a further improvement to 94.10% on the MSL dataset, demonstrating its stronger generalization ability in handling high-dimensional aerospace telemetry data. On SMD and MSL, RGAAD also surpasses Anomaly Transformer by a clear margin, verifying the effectiveness of the proposed hybrid structure over pure transformer-based frameworks.

The results demonstrate that deep learning models consistently outperform traditional techniques, and RNN-based models such as LSTM (81.78% on SMD) have provided considerable improvement. However, the innovation of RGAAD lies in its joint modeling of feature dynamics and temporal affinity,which enables more accurate anomaly localization and better separation between normal and abnormal patterns.

The high precision (88.14% on SMD, 92.30% on MSL) and recall (85.89% on SMD, 95.80% on MSL) of RGAAD indicate its stable and reliable detection performance across diverse anomaly types. These results confirm that RGAAD is well-suited for real-world IoT monitoring scenarios where high detection accuracy and strong generalization are critical.

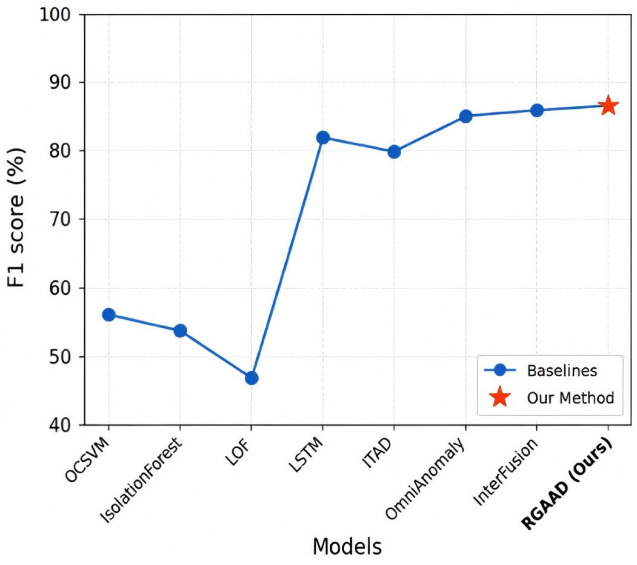
Table 3 Quantitative results for RGAAD and other methods on SMD, SWaT and MSL datasets. The P, R, and F1 represent Precision, Recall, and F1-score, respectively.

Dataset	SMD			SWaT			MSL		
Metric	P	R	F1	P	R	F1	P	R	F1
OCSVM	44.34	76.72	56.19	45.39	49.22	47.23	59.78	86.87	70.82
IsolationForest	42.31	73.29	53.64	49.29	44.95	47.02	53.94	86.54	66.45
LOF	56.34	39.86	46.68	72.15	65.43	68.62	47.72	85.25	61.18
LSTM	78.55	85.28	81.78	86.15	83.27	84.69	85.45	82.50	83.95
ITAD	86.22	73.71	79.48	63.13	52.08	57.08	69.44	84.09	76.07
OmniAnomaly	83.68	86.82	85.22	81.42	84.30	82.83	89.02	86.37	87.67
InterFusion	87.02	85.43	86.22	80.59	85.58	83.01	81.28	92.70	86.62
Anomaly-Transformer	87.64	91.37	88.23	91.55	96.37	94.07	92.09	95.15	93.59
Ours	88.14	85.89	87.00	83.45	87.73	85.37	92.30	95.80	94.10

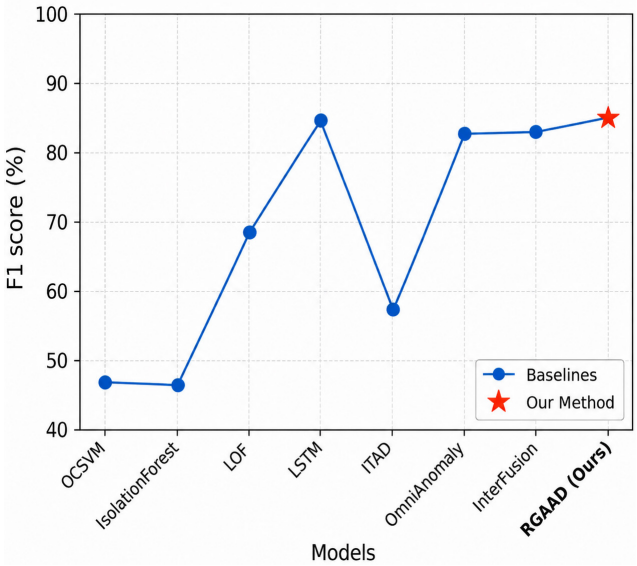
As illustrated in Fig. 4, the F1-score comparison across models further validates the superiority of RGAAD over existing methods on both SMD and SWaT datasets. The continuous green lines connecting baseline methods reveal a gradual performance improvement from classical algorithms (e.g., OC-SVM at 56.19%) to deep learning approaches (e.g., InterFusion at 86.22%), while our method (highlighted in red with bold labels) demonstrates a decisive leap in anomaly detection capability.

Key observations include:

- Consistent Dominance: RGAAD achieves the highest F1-values at both ends of the curves (87.00% for SMD, 85.37% for SWaT), with its star-shaped marker (*) in the SWaT plot distinctly separating from the cluster of baseline results.
- Architectural Impact: The steep rise from InterFusion (86.22%) to RGAAD (87.00%) on SMD reflects the effectiveness of integrating Transformer-based temporal modeling with pairwise association learning, overcoming the limitations of pure reconstruction-based methods.
- Generalization Advantage: Despite SWaT's complex industrial scenarios, RGAAD maintains a balanced performance (Precision = 83.45%, Recall = 87.73%), as evidenced by its stable position above the 85% threshold, whereas methods like ITAD suffer significant performance drops (57.08% F1).



(a)



(b)

Fig. 4 F1 Value Chart: (a) F1 scores on the SMD dataset; (b) F1 scores on the SWaT dataset

The visual gaps between the red markers (our method) and green trajectory (baselines) quantitatively emphasize RGAAD's dual strength: localized anomaly sensitivity from local temporal dependency modeling and global pattern awareness through adaptive temporal scaling. This aligns with the model's design philosophy of bridging feature dynamics and contextual dependencies.

4.6 Hyperparameter sensitivity analysis

To evaluate the robustness of RGAAD, we conduct a sensitivity analysis on three key hyperparameters, β, k, and window size w, across the SMD, SWaT, and MSL datasets, as shown in Table 4. The results show a consistent unimodal trend, with the best F1-scores achieved at $\beta = 0.1$, $k = 3$, and $w = 100$ across all datasets. The model remains stable under moderate deviations from these optimal settings, with F1-scores staying within 3–4 percentage points of the peak values. Performance degradation is observed when β is too small or too large, when the threshold is overly aggressive or conservative, and when the window size is insufficient to capture temporal context. These findings demonstrate that

Table 4 Hyperparameter sensitivity results

Hyperparameter	Value	SMD (F1)	SWaT (F1)	MSL (F1)
	0.01	85.12	83.21	91.45
β	0.1	87.00	85.37	94.10
	1.0	83.54	81.63	89.82
	1	84.71	82.52	90.84
k	3	87.00	85.37	94.10
	5	85.42	83.15	91.53
	10	85.53	83.42	91.71
w	50	86.81	85.12	93.24
	100	87.00	85.37	94.10

RGAAD is robust to hyperparameter variations and reliable for practical anomaly detection scenarios.

5 Conclusion

In this study, we proposed RGAAD, a Transformer-based framework for unsupervised anomaly detection in IoT time-series data. The model integrates temporal affinity and feature dynamics to jointly capture local anomalous patterns and global temporal dependencies. In addition, adaptive temporal scaling and gated feature fusion improve anomaly discrimination by modeling correlation deviations between normal and abnormal patterns. Experiments on SMD, SWaT, and MSL demonstrate that RGAAD consistently achieves superior performance compared with existing methods. These results highlight the effectiveness and robustness of RGAAD for practical anomaly detection in dynamic IoT monitoring environments.

Acknowledgement

The datasets used in this study are publicly available from official repositories. SMD and MSL can be accessed via the OmniAnomaly repository, while SWaT is available from the iTrust Labs repository.

References

[1] Chen, Z., Chen, D., Zhang, X., Yuan, Z., Cheng, X. "Learning Graph Structures with Transformer for Multivariate Time-Series Anomaly Detection in IoT", IEEE Internet of Things Journal, 9(12), pp. 9179–9189, 2021. https://doi.org/10.1109/JIOT.2021.3100509

[2] Breunig, M. M., Kriegel, H. P., Ng, R. T., Sander, J. "LOF: identifying density-based local outliers", In: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, Dallas, Texas, USA, 2000, pp. 93–104. ISBN 978-1-58113-217-5 https://doi.org/10.1145/342009.335388

[3] Tax, D. M. J., Duin, R. P. W. "Support Vector Data Description", Machine learning, 54(1), pp. 45–66, 2004. https://doi.org/10.1023/B:MACH.0000008084.60811.49

[4] Lai, K. H., Zha, D., Xu, J., Zhao, Y., Wang, G., Hu, X. "Revisiting Time Series Outlier Detection: Definitions and Benchmarks", In: 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks, online conference, 2021. ISBN 9781713871095

[5] Tariq, S., Lee, S., Shin, Y., Lee, M. S., Jung, O., Chung, D., Woo, S. S. "Detecting Anomalies in Space using Multivariate Convolutional LSTM with Mixtures of Probabilistic PCA", In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York, NY, United States, 2019, pp. 2123–2133. ISBN 978-1-4503-6201-6 https://doi.org/10.1145/3292500.3330776

[6] Tang, J., Chen, Z., Fu, A. W. C., Cheung, D. W. "Enhancing Effectiveness of Outlier Detections for Low Density Patterns", In: Advances in Knowledge Discovery and Data Mining, Springer, 2002, pp. 535–548. ISBN 978-3-540-43704-8 https://doi.org/10.1007/3-540-47887-6_53

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. "Attention Is All You Need", In: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017. ISBN 978-1-5108-6096-4 https://doi.org/10.48550/arXiv.1706.03762

[8] Wu, H., Xu, J., Wang, J., Long, M. "Autoformer: decomposition transformers with auto-correlation for long-term series forecasting", In: NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2021, pp. 22419–22430. ISBN 978-1-7138-4539-3 https://doi.org/10.5555/3540261.3541978

[9] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., …, Amodei, D. "Language Models are Few-Shot Learners", In: Advances in Neural Information Processing Systems 33 (NeurIPS 2020), Curran Associates Inc., 2020, pp. 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

[10] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows", In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 9992–10002. ISBN 978-1-6654-2813-2 https://doi.org/10.1109/ICCV48922.2021.00986

[11] Jones, M., Nikovski, D., Imamura, M., Hirata, T. "Exemplar learning for extremely efficient anomaly detection in real-valued time series", Data Mining and Knowledge Discovery, 30(6), pp. 1427–1454, 2016. https://doi.org/10.1007/s10618-015-0449-3

[12] Li, D., Chen, D., Jin, B., Shi, L., Goh, J., Ng, S. K. "MAD-GAN: Multivariate Anomaly Detection for Time Series Data with Generative Adversarial Networks", In: Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series. ICANN 2019, Springer, 2019, pp. 703–716. ISBN 978-3-030-30489-8 https://doi.org/10.1007/978-3-030-30490-4_56

[13] Sakurada, M., Yairi, T. "Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction", In: Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis, Gold Coast, Australia, 2014, pp. 4–11. ISBN 9781450331593 https://doi.org/10.1145/2689746.2689747

[14] Lu, H., Liu, Y., Fei, Z., Guan, C. "An Outlier Detection Algorithm Based on Cross-Correlation Analysis for Time Series Dataset", IEEE Access, 6, pp. 53593–53610, 2018. https://doi.org/10.1109/ACCESS.2018.2870151

[15] Senin, P., Lin, J., Wang, X., Oates, T., Gandhi, S., Boedihardjo, A. P., Chen, C., Frankenstein, S. "Time series anomaly discovery with grammar-based compression", In: Proceeding of the 18th International Conference on Extending Database Technology (EDBT), Brussels, Belgium, 2015, pp. 481–492. https://doi.org/10.5441/002/edbt.2015.42

[16] Akouemo, H. N., Povinelli, R. J. "Data Improving in Time Series Using ARX and ANN Models", IEEE Transactions on Power Systems, 32(5), pp. 3352–3359, 2017. https://doi.org/10.1109/TPWRS.2017.2656939

[17] Defferrard, M., Bresson, X., Vandergheynst, P. "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering", In: 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 2016. ISBN 978-1-5108-3881-9 https://doi.org/10.48550/arXiv.1606.09375

[18] Deng, A., Hooi, B. "Graph Neural Network-Based Anomaly Detection in Multivariate Time Series", Proceedings of the AAAI Conference on Artificial Intelligence, 35(5), pp. 4027–4035, 2021. https://doi.org/10.1609/aaai.v35i5.16523

[19] Zhao, H., Wang, Y., Duan, J., Huang, C., Cao, D., Tong, Y., Xu, B., Bai, J., Tong, J., Zhang, Q. "Multivariate Time-Series Anomaly Detection via Graph Attention Network", In: 2020 IEEE International Conference on Data Mining (ICDM), Sorrento, Italy, 2020, pp. 841–850. https://doi.org/10.1109/ICDM50108.2020.00093

[20] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", [preprint] arXiv preprint, 24 May 2019. https://doi.org/10.48550/arXiv.1810.04805

[21] Huang, C. Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A. M., Hoffman, M. D., Dinculescu, M., Eck, D. "Music transformer: Generating music with long-term structure", [preprint] arXiv preprint, 12 December 2018. https://doi.org/10.48550/arXiv.1809.04281

[22] Xu, J., Wu, H., Wang, J., Long, M. "Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy", In: International Conference on Learning Representations, virtual conference, 2022. https://doi.org/10.48550/arXiv.2110.02642

[23] Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., Pei, D. "Robust Anomaly Detection for Multivariate Time Series through Stochastic Recurrent Neural Network", In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 2019, pp. 2828–2837. ISBN 9781450362016 https://doi.org/10.1145/3292500.3330672

[24] Mathur A. P., Tippenhauer, N. O. "Swat: a water treatment testbed for research and training on ICS security" In: 2016 International Workshop on Cyber-physical Systems for Smart Water Networks (CySWater), Vienna, Austria, 2016, pp. 31–36. ISBN 978-1-5090-1161-2 https://doi.org/10.1109/CySWater.2016.7469060

[25] Hundman, K., Constantinou, V., Laporte, C., Colwell, I., Söderström, T. "Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding", In: KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, United Kingdom, 2018, pp. 387–395. ISBN 978-1-4503-5552-0 https://doi.org/10.1145/3219819.3219845

[26] Liu, F. T., Ting, K. M., Zhou, Z. H. "Isolation Forest", In: 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 2008, pp. 413–422. ISBN 978-0-7695-3502-9 https://doi.org/10.1109/ICDM.2008.17

[27] Graves, A. "Long Short-Term Memory", In: Supervised Sequence Labelling with Recurrent Neural Networks, Springer, 2012, pp. 37–45. ISBN 978-3-642-24796-5 https://doi.org/10.1007/978-3-642-24797-2_4

[28] Shin, Y. Lee, S., Tariq, S., Lee, M. S., Jung, O., Chung, D., Woo, S. S. "Itad: Integrative Tensor-based Anomaly Detection System for Reducing False Posivitives of Satellite Systems", In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 2733–2740. https://doi.org/10.1145/3340531.3412716

[29] Li, Z., Zhao, Y., Han, J.,Su, Y., Jiao, R., Wen, X., Pei, D. "Multivariate Time Series Anomaly Detection and Interpretation using Hierarchical Inter-Metric and Temporal Embedding", In: KDD '21: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Virtual Event Singapore, 2021, pp. 3220–3230. ISBN 978-1-4503-8332-5 https://doi.org/10.1145/3447548.3467075