Federated Spatio-Temporal Attention for Time Series Anomaly Detection

Weicheng Wang[1], Yue He[1], Xiaoliang Chen[1,2,5(✉)], Duoqian Miao[2], Hongyun Zhang[2], Xiaolin Qin[3], Shangyi Du[4], and Peng Lu[5]

[1] School of Computer and Software Engineering, Xihua University, Chengdu 610039, People's Republic of China
chenxl@mail.xhu.edu.cn, chexiaol@iro.umontreal.ca
[2] College of Electronic and Information Engineering, Tongji University, Shanghai 201804, People's Republic of China
[3] Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu 610041, People's Republic of China
[4] Department of Computer Science, McGill University, Montreal, QC H3A0G4, Canada
[5] Department of Computer Science and Operations Research, University of Montreal, Montreal, QC H3C3J7, Canada

Abstract. The accelerated expansion of Industrial Internet of Things (IIoT) systems has concurrently precipitated the generation of substantial multivariate time series data, where the implementation of precise anomaly detection mechanisms is indispensable for ensuring operational safety, and mitigating risks in complex industrial ecosystems.

Current methodologies predominantly leverage local spatial and temporal representations derived from adjacent nodes and recent time points. This approach, which focuses on local processing, frequently overlooks global topological relationships and temporal patterns—both crucial for precise anomaly detection. This limitation arises from insufficient modeling of global sensor-time dependencies and inadequate integration of spatial-temporal interdependencies, resulting in high false-positive rates.

To mitigate these issues, we introduce **FL-STAM**, a federated learning-enhanced framework with a spatio-temporal attention mechanism for unsupervised MTS anomaly detection.

First, a parallel graph attention architecture independently extracts global sensor dependencies and temporal patterns through serial-oriented and time-oriented modules. Second, a dual-branch transformer with Wasserstein distance quantification explicitly models spatial-temporal association discrepancies, amplifying discriminative features between normal and anomalous patterns. Third, this paper adopts a privacy-preserving federated learning paradigm, which can realize collaborative model training among distributed devices while effectively preventing the exposure of local data.

Extensive experiments on five IIoT datasets (SMAP, MSL, SMD, PSM, SWaT) demonstrate FL-STAM's superiority, achieving state-of-the-art (SOTA) F1 scores of 98.52% on PSM and 97.86% on SWaT. Abla-

M. Yoshikawa et al. (Eds.): ADMA 2025, LNAI 16197, pp. 129–141, 2026.
https://doi.org/10.1007/978-981-95-3453-1_9

tion studies verify component effectiveness, notably the parallel graph mechanism enhancing accuracy by 6.86% over baselines.

Keywords: Time series · Anomaly detection · Federated learning · Graph attention mechanism · Transformer

1 Introduction

A time series refers to a chronological arrangement of data points, commonly recorded at consistent intervals and organized in a sequential manner. Time series analysis involves applying various analytical methods to examine these data, with the goal of identifying significant statistical patterns and extracting meaningful insights.

The IIoT infrastructure widely collects multivariate time series (MTS) data, which possess complex topological relationships and significant temporal dynamic characteristics [13]. However, industrial networks and devices are vulnerable to potential threats and attacks, potentially leading to severe repercussions. Consequently, the development of efficient and precise anomaly detection systems for MTS is of paramount importance.

While temporal models like Recurrent Neural Networks (RNNs) [6], Autoencoders (AEs) [17], and Generative Adversarial Networks (GANs) [8] effectively capture temporal dependencies, they often neglect inter-series spatial correlations.

To mitigate these challenges, we introduce a distributed training framework, termed FL-STAM, grounded in the federated mean algorithm. This framework augments the STAM approach to enable the distributed training of anomaly detection models, improving their accuracy while safeguarding data privacy.

The contributions of this study are delineated as follows:

1. This research constructs a sequence-time parallel graph attention framework, designed to simultaneously handle global sequence dependencies and temporal dependencies in MTS data. The framework integrates two graph attention mechanisms, which precisely allocate weights to the influence of global nodes across both sequence and time dimensions, thereby significantly enhancing the optimization of time series feature representation.
2. Transformers are employed to extract spatial and temporal correlations. Furthermore, based on the extracted spatial and temporal correlations, the discrepancies among these correlations are computed through a spatio-temporal multi-head attention mechanism.
3. To ascertain the efficacy of FL-STAM, this study undertook a thorough experimental evaluation. Across five publicly available datasets, FL-STAM was benchmarked against eight prominent anomaly detection methods. The results illustrate that FL-STAM substantially surpasses other SOTA methods in performance.

2 Related Work

ADMTS has garnered substantial interest owing to its pivotal applications across diverse domains, including industrial monitoring, finance, and healthcare. The traditional approaches used in ADMTS are neatly sorted into three different types: clustering-based approaches [7], distance-based techniques [2], and isolation-based methods [9]. These methods have laid the foundational framework for identifying anomalies by leveraging statistical and machine learning techniques. However, with the advent of deep learning, more sophisticated and powerful models have emerged, significantly enhancing the feature representation and generalization capabilities for anomaly detection [15]. Deep learning methodologies predominantly encompass two intrinsic paradigms: reconstruction-centric and prediction-oriented models [3].

Reconstruction-based frameworks endeavor to encapsulate the intrinsic distribution of MTS data through latent variables, subsequently facilitating the reconstitution of original data samples from the acquired distribution [5].

AnomalyTrans [14] augments anomaly detection capabilities by leveraging a transformer architecture fortified with an anomaly attention mechanism. GLAD [16] extends AnomalyTrans by integrating a Gumbel-Softmax-based graph structure learning technique. These reconstruction-based methods have shown remarkable capabilities in automatically learning high-level data features, thereby enhancing the detection of complex anomaly patterns. Nevertheless, a significant drawback of these methods is their tendency to overlook sudden fluctuations in time series data, which may represent normal variations, leading to an increased false positive rate.

Prediction-based frameworks [11] center on anticipating forthcoming temporal instances or intervals based on historical MTS data, with anomalies detected through the evaluation of prediction errors. SCNN [4] is an adaptive, interpretable, and scalable prognostic architecture designed to independently model each constituent of spatio-temporal dynamics. By operating based on a predefined MTS generative process, SCNN captures the latent structure of spatio-temporal patterns, offering improved traceability and predictability compared to traditional methods. THOC [10] introduces a classification framework that captures temporal dynamics across various scales through the use of a dilated RNN with skip connections, and integrates multiscale time features by means of hierarchical clustering. This methodology efficiently addresses the complex temporal dependencies inherent in MTS data. However, a key limitation is that they primarily capture correlations between different timestamps in MTS, often neglecting the interdependencies between data from different sensors. This limitation highlights the need for more comprehensive approaches capable of efficaciously modeling both spatial and temporal dependencies within MTS data.

3 Methodology

In this chapter, we formally delineate the conceptual domain of unsupervised ADMTS problems and provide a detailed exposition of the Federated Learning

framework for Spatio-Temporal Anomaly Detection (FL-STAM). The objective is to identify temporal patterns that markedly diverge from normal behavior, independent of labeled anomaly data during the training phase.

3.1 Overview of Proposed Model

This study proposes a privacy-preserving anomaly detection architecture for MTS data within IIoT systems. The FL-STAM architecture identifies anomalies by analyzing discrepancies between spatial correlations and temporal interactions while integrating federated learning to enable distributed computation and preserve data privacy. FL-STAM comprises two principal components:

- **FL-STAM Model for ADMTS.** As depicted in Fig. 1, the process commences with data preprocessing to address discrepancies in values across different dimensions of the MTS. Subsequently, a parallel graph structure learning approach is deployed, encompassing a serial-oriented graph attention layer to capture sensor dependencies and a time-oriented graph attention layer to model temporal interdependencies within the time series. A transformer architecture incorporating an anomaly attention mechanism is then employed, enabling simultaneous modeling of spatial and temporal associations. Beyond reconstruction loss, the anomaly transformer is refined through a minimax optimization strategy, aimed at sharpening the distinction of association discrepancies and thereby improving the detection of time series anomalies.
- **FL-STAM Distributed Training Framework.** Sensor systems within distinct manufacturing areas stream temporal operational data to local edge devices. These edge devices train local FL-STAM models using resident data, enabling timely edge-based anomaly detection. This distributes computational load, preventing cloud server overload.

3.2 FL-STAM Model for ADMTS

Serial-Time Parallel Graph Attention Mechanism. In the context of MTS, interdependencies among nodes are captured through correlation analysis in both serial and time dimensions. To address this, we employ two parallel graph attention layers, each tasked with the distinct extraction of features from the serial and temporal components of the multivariate time series data. Specifically, we use GATv2 [1], a modified version of the standard Graph Attention Network, to perform computations in both serial and time domains.

(1) Serial-Oriented Graph Attention Layer
The objective is to identify correlations among MTS without prior knowledge. To achieve this, the MTS is represented as a complete graph, wherein each node corresponds to a unique time series, exemplified by sensor data. The edges of the graph capture the relationships between pairs of time series. Specifically, the time series i is represented as a vector $\boldsymbol{x}^i = \{\nu_1^i, \nu_2^i, \ldots, \nu_s^i\}$, where s denotes the number of timestamps within a sliding window. The set of adjacent nodes

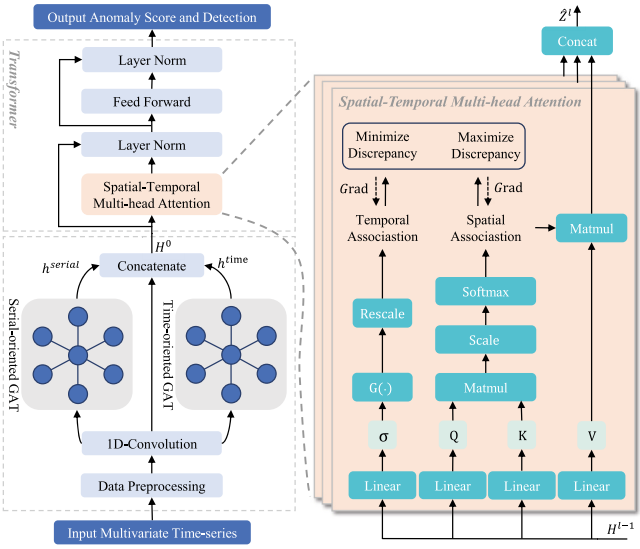

Fig. 1. The framework of FL-STAM.

$\mathcal{N}_i \in \mathbb{R}^n$ includes all other time series within the MTS. The serial correlation between time series i and j is denoted by c_{ij}^{serial}, with the corresponding formula provided below:

$$c_{ij}^{serial} = \boldsymbol{a}^\top \text{LeakyReLU}(W \cdot [x^i \parallel x^j]) \tag{1}$$

where $a \in \mathbb{R}^d$ and $W \in \mathbb{R}^{d \times 2s}$ are learned parameters, with d representing a configurable intermediate dimension, which is set to $2s$ in this context. The operator $||$ indicates vector concatenation, and LeakyReLU refers to a nonlinear activation function.

The serial attention scores α_{ij}^{serial} are normalized across all neighboring time series $j \in \mathcal{N}_i$ using the softmax function. The corresponding attention mechanism is then defined as follows:

$$\alpha_{ij}^{serial} = \frac{\exp\left(c_{ij}^{serial}\right)}{\sum_{j \in \mathcal{N}_i} \exp\left(c_{ij}^{serial}\right)} \tag{2}$$

Finally, the serial graph attention layer determines the output representation $\boldsymbol{h}^i$ for each node using the following process, ensuring that the dimensionality of $\boldsymbol{h}^i$ matches that of the input $\boldsymbol{x}^j$.

$$\boldsymbol{h}^i = \sigma\Big(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^{serial} \boldsymbol{x}^j\Big) \tag{3}$$

where σ represents the sigmoid activation function.

(2) Time-Oriented Graph Attention Layer
Graph Attention Networks are utilized to model time dependencies within time series data. Specifically, a sliding window approach is utilized, representing each set of timestamps as a complete graph. To formalize the problem, let the vector at time i be denoted as $\boldsymbol{x}_i = \{\nu_i^1, \nu_i^2, \cdots, \nu_i^n\}$, where n represents the number of features in the MTS. The set of neighboring nodes $\mathcal{N}_i \in \mathbb{R}^s$ consists of all other timestamps within the current sliding window. The time correlation between timestamps i and j is denoted by c_{ij}^{time}, and the corresponding formulation is provided below:

$$c_{ij}^{time} = \boldsymbol{a}^\top \text{LeakyReLU}(\boldsymbol{W} \cdot [\boldsymbol{x}_i \parallel \boldsymbol{x}_j]) \tag{4}$$

where $a \in \mathbb{R}^d$ and $W \in \mathbb{R}^{d \times 2n}$ are learned parameters, with d representing a configurable intermediate dimension, which is set to $2n$ in this context. The operator $||$ indicates vector concatenation, and LeakyReLU refers to a nonlinear activation function.

The time attention scores α_{ij}^{time} are normalized across all neighboring timestamps $j \in \mathcal{N}_i$ at time i using the softmax function, and the attention mechanism is defined as follows:

$$\alpha_{ij}^{time} = \frac{\exp\left(c_{ij}^{time}\right)}{\sum_{j \in \mathcal{N}_i} \exp\left(c_{ij}^{time}\right)} \tag{5}$$

Finally, the time graph attention layer generates the output representation $\boldsymbol{h}_i$ for each node through the following process, ensuring that the dimensionality of $\boldsymbol{h}_i$ is consistent with that of the input $\boldsymbol{x}_i$.

$$\boldsymbol{h}_i = \sigma\Big(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^{time} \boldsymbol{x}_i\Big) \tag{6}$$

where σ represents the sigmoid activation function.

The output of the serial graph attention layer, h^{serial}, is represented as a matrix of dimensions $n \times s$, while the output of the time graph attention layer, denoted as h^{time}, is represented as a matrix of dimensions $s \times n$. To integrate information from multiple sources, the outputs from both the serial and time graph attention layers, along with the output from the one-dimensional convolutional neural network, are concatenated. This combined process is referred to as Embedding(X).

Transformer with Spatial and Temporal Association Discrepancies. The Transformer model introduced in this study, which accounts for spatial and temporal association discrepancies, is distinguished by its interleaved arrangement of spatial-temporal multi-head attention blocks and feedforward neural network layers. Let the model consist of L layers, with the input time series denoted as $H \in \mathbb{R}^{s \times d}$, where s indicates the series length and d its dimensionality. The general formulation for the l-th layer can be formally expressed as:

$$Z^l = \text{LayerNorm}(\text{ST-Attention}(H^{l-1}) + H^{l-1}) \tag{7}$$

$$H^l = \text{LayerNorm}(\text{FeedForward}(Z^l) + Z^l) \tag{8}$$

Here, $H^l \in \mathbb{R}^{s\times d}$, for $l \in \{1,\ldots,L\}$, signifies the output of the l-th layer with d channels. The initial input, $H^0 = \text{Embedding}(X)$, denotes the embedded form of the original sequence. $Z^l \in \mathbb{R}^{s\times d}$ is the hidden representation at the l-th layer. The function ST-Attention$(\cdot)$ is employed to calculate the spatial and temporal association discrepancies.

Spatial-Temporal Multi-head Attention. The single-branch self-attention mechanism [12] is limited in concurrently modeling spatial and temporal associations. To mitigate this limitation, we introduce a ST-Attention mechanism with a dual-branch structure [14].

Spatial-association is designed to examine the relationships within the original sequence, adaptively identifying the most relevant associations across the entire spatiotemporal domain, irrespective of temporal distance.

$$\mathrm{Q},\mathrm{K},\mathrm{V},\sigma = H^{l-1}W_{\mathrm{Q}}^l, H^{l-1}W_{\mathrm{K}}^l, H^{l-1}W_{\mathrm{V}}^l, H^{l-1}W_{\sigma}^l \tag{9}$$

$$SA^l = \text{Softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{d_{\text{model}}}}\right) \tag{10}$$

$$\widehat{Z}^l = SA^l\mathrm{V} \tag{11}$$

In the self-attention mechanism, $\mathrm{Q},\mathrm{K},\mathrm{V} \in \mathbb{R}^{s\times d}$ and $\sigma \in \mathbb{R}^{s\times 1}$ correspond to the Query, Key, Value, and the learned scale parameter, respectively. $W_{\mathrm{Q}}^l, W_K^l, W_V^l \in \mathbb{R}^{d\times d}$ and $W_\sigma^l \in \mathbb{R}^{d\times 1}$ denote the weights associated with $\mathrm{Q},\mathrm{K},\mathrm{V}$, and σ at the l-th layer, respectively.

The spatial-association matrix $SA^l \in \mathbb{R}^{s\times s}$ is normalized along its last dimension by the Softmax function, ensuring that each row of SA^l forms a discrete probability distribution. $\widehat{Z}^l \in \mathbb{R}^{s\times d}$ represents the hidden state obtained after the l-th layer of the ST-Attention mechanism.

Output Anomaly Score and Detection. The discrepancy between spatial and temporal associations is assessed through the symmetric Wasserstein Distance, a metric that quantifies the divergence between two probability distributions, wherein a reduced Wasserstein Distance signifies a lesser disparity between the distributions. The association discrepancy across various layers is aggregated to consolidate multi-level feature associations into a holistic measure, as delineated below:

$$\text{Dis}(TA, SA; X) = \left(\frac{1}{L}\sum_{l=1}^{L}\left[\mathcal{W}\left(TA_{i,:}^l, SA_{i,:}^l\right) + \mathcal{W}\left(SA_{i,:}^l, TA_{i,:}^l\right)\right]\right)_{i=1}^{s}. \tag{12}$$

where Wasserstein$(\cdot)$ denote the Wasserstein Distance between two discrete probability distributions. The vector $\text{Dis}(TA, SA; X) \in \mathbb{R}^{s\times 1}$ quantifies the pointwise association discrepancy of the time series X with respect to the temporal

associations TA and spatial associations SA across multiple layers. The i-th element of Dis corresponds to the association discrepancy at the i-th time point of X. Empirical findings suggest that anomalous time points tend to exhibit lower $\text{Dis}(TA, SA; X)$ values than normal time points, thereby rendering Dis a distinctive and effective measure for anomaly detection.

As an unsupervised task, the model is optimized using reconstruction loss during the training phase. This loss function directs the spatial association component to discern the most relevant correlations. To effectively distinguish between normal and anomalous time points, an auxiliary loss term is introduced to accentuate the association discrepancy. Given the unimodal nature of the temporal association, the discrepancy loss encourages the spatial association to emphasize non-adjacent regions, thereby intensifying the challenge of reconstructing anomalies and thereby enhancing their detectability. The loss function for the input series $X \in \mathbb{R}^{s \times d}$ is formulated as follows:

$$\mathcal{L}_{\text{Total}} = \|X - \widehat{X}\|_{\text{F}}^2 - \lambda \times \|\text{Dis}(TA, SA; X)\|_1 \tag{13}$$

where $\widehat{X} \in \mathbb{R}^{s \times n}$ represents the reconstructed matrix of X. $\|\cdot\|_{\text{F}}, \|\cdot\|_1$ indicate the Frobenius norm and 1-norm. The parameter λ serves to balance the various terms within the loss function. For $\lambda > 0$, the optimization process aims to amplify the association discrepancy.

3.3 FL-STAM Distributed Training Framework

This section implements a FL framework for collaborative anomaly detection model training across distributed edge devices, enhancing detection accuracy while ensuring data privacy preservation. Each edge device operates as a client node running the FL-STAM anomaly detection model (Sect. 3.2), with parameter aggregation performed at the central server. The framework maintains data privacy by restricting transmission to model parameters while retaining raw data locally. The proposed framework's workflow comprises seven sequential stages:

(1) Sensor data acquired by the edge device constitutes the local dataset for model training.
(2) The edge device performs local model training following the methodology described in Sect. 3.2, utilizing the collected local dataset.
(3) Upon completion of local training, the edge device transmits the updated model parameters to the global server.
(4) The global server executes a model aggregation process, integrating parameters from all participating clients to produce an updated global model.
(5) The aggregated global model parameters are subsequently disseminated to all edge devices.
(6) Steps (2) through (5) are repeated until the global model converges.
(7) Th edge device performs anomaly detection using the converged global model.

4 Experiments

This segment delineates the experimental framework, including the specification of the data set, evaluation metrics, and baseline comparisons. Comprehensive experiments substantiate the efficacy of the proposed model across multiple performance dimensions (Table 1).

Table 1. Comprehensive statistical summary of the experimental datasets.

Dataset	Applications	Dimensions	Train	Test	Anomaly rate(%)
MSL	Space	55	58,317	73,729	10.72
SMAP	Space	25	135,183	427,617	13.13
PSM	Server	25	132,481	87,841	27.75
SMD	Server	38	708,405	708,420	4.16
SWaT	Water	51	496,800	449,919	11.98

4.1 Experimental Setup

Computationally, all experiments were conducted on Ubuntu 20.04 with an RTX 4090 GPU (24 GB VRAM), utilizing PyTorch 2.0 and CUDA 11.8. Model configuration included a non-overlapping sliding window of size 100, dataset-specific anomaly thresholds (SWaT: 0.1%, SMD: 0.5%, others: 1%), and a 3-layer transformer (8 heads, 512 hidden units). Optimization used the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$) with a learning rate of 10^{-4} and a loss balance coefficient $\lambda = 3$ (Eq. 13).

4.2 Experimental Results

This segment presents the experimental outcomes that validate the effectiveness of the FL-STAM model in anomaly detection across five diverse datasets. The assessment is grounded in the comparison of FL-STAM with multiple SOTA anomaly detection techniques, employing precision, recall, and F1 score as key performance metrics. Each dataset was evaluated through ten test iterations to ensure dependability and robustness of the reported performance metrics.

Baseline Comparisons. Table 2 provides a detailed comparison of the FL-STAM model with eight SOTA anomaly detection methods across five distinct datasets. The best results are highlighted in bold, with rankings provided in parentheses. Both average F1 scores and their corresponding rankings are reported in the table. The technique recognized as the best performer is the one that captures the highest average F1 scores and achieves top rankings. If a

method secures high average F1 scores but falls short in rankings, it might perform especially well on certain datasets while showing inconsistent results. On the other hand, a method with excellent rankings but lower average F1 scores could suggest stable performance across all datasets, even if it doesn't excel particularly on any single dataset.

Table 2. Experimental results for all methods on five public datasets, measured by F1 score.

Method	MSL			SMAP			PSM		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
IF (2008)	53.94	86.54	66.45(9)	52.39	59.07	55.53(9)	75.91	57.36	65.34(9)
OmniAnomaly (2019)	89.02	86.37	87.67(6)	92.49	81.99	86.92(7)	88.39	74.46	80.83(8)
MTAD-GAT (2020)	87.54	94.40	90.84(4)	89.06	91.23	90.13(4)	95.28	75.65	84.34(5)
InterFusion (2021)	81.28	92.70	86.62(7)	89.77	88.52	89.14(5)	83.61	83.45	83.52(6)
GDN (2021)	91.35	86.12	88.66(5)	89.32	88.72	89.02(6)	80.65	83.57	82.09(7)
AnomalyTrans (2022)	92.09	95.15	93.59(3)	94.13	99.40	96.69(3)	96.91	98.90	97.89(3)
Dcdetector (2023)	93.69	99.69	**96.60(1)**	95.63	98.92	97.02(2)	97.14	98.74	97.94(2)
RWKV-TS (2024)	78.11	77.74	77.92(8)	89.04	55.94	68.71(8)	98.32	95.92	97.10(4)
FL-STAM	92.79	98.13	95.39(2)	96.16	99.45	**97.78(1)**	98.67	98.37	**98.52(1)**

Method	SMD			SWaT			Avg F1	Avg F1 Ranking	
	Pre	Rec	F1	Pre	Rec	F1		Rank	(value)
IF (2008)	42.31	73.29	53.64(9)	49.42	44.95	47.02(9)	57.60(9)	9(9)	
OmniAnomaly (2019)	83.68	86.82	85.22(6)	81.42	84.30	82.83(7)	84.69(7)	6.8(7)	
MTAD-GAT (2020)	88.28	84.92	86.57(4)	92.46	75.12	82.89(6)	86.95(4)	4.6(4)	
InterFusion (2021)	84.02	88.41	86.16(5)	80.59	85.58	83.01(5)	85.69(5)	5.6(5)	
GDN (2021)	71.70	99.74	83.42(8)	99.35	68.12	80.82(8)	84.80(6)	6.8(7)	
AnomalyTrans (2022)	89.40	95.45	92.33(2)	91.55	96.73	94.07(3)	94.91(3)	2.8(3)	
Dcdetector (2023)	83.59	91.10	87.18(3)	93.11	99.77	96.33(2)	95.01(2)	2(2)	
RWKV-TS (2024)	87.45	81.43	84.33(7)	88.20	94.85	91.40(4)	83.89(8)	6.2(6)	
FL-STAM	93.79	95.29	**94.53(1)**	97.33	98.41	**97.86(1)**	**96.82(1)**	**1.2(1)**	

The FL-STAM model achieves an average F1 score of 96.82 across all datasets, surpassing all competing methods and establishing itself as the top performer. With an average F1 ranking of 1.2, FL-STAM occupies the leading position. In contrast, the InterFusion, AnomalyTrans, and Dcdetector models represent the SOTA methods for 2021, 2022, and 2023, respectively, serving as the primary benchmarks in this comparison. On the SWaT dataset, FL-STAM exhibits a 17.89% improvement over InterFusion, a 4.03% advantage over AnomalyTrans, and a 1.59% edge over Dcdetector. In the PSM dataset, FL-STAM achieves an F1 score 17.96% higher than InterFusion, 0.64% superior to AnomalyTrans, and 0.59% better than Dcdetector. Across all five datasets, FL-STAM

demonstrates a 12.99% increase in F1 score over InterFusion, a 2.01% improvement over AnomalyTrans, and a 1.91% gain over Dcdetector. In all other datasets assessed, FL-STAM consistently outperforms the baseline models.

The analysis of traditional models such as Isolation Forest (IF) reveals their limitations; they consistently underperform relative to deep learning models across all datasets, primarily due to their inability to effectively capture complex patterns and spatiotemporal correlations in MTS data. For example, while OmniAnomaly demonstrates improvements in temporal modeling, it fails to capture the spatial dependencies critical for accurate anomaly detection. The GDN methodology, encompassing node embeddings and graph attention mechanisms, underscores the inter-nodal correlations while disregarding temporal dependencies. In contrast, both InterFusion and MTAD-GAT effectively address both time and spatial dependencies, leading to improved detection precision. Nonetheless, these models still rely on basic judgment criteria. AnomalyTrans pioneers anomaly detection criterion that integrates association discrepancy, facilitating the differentiation between anomalous and normal data. However, it overlooks the spatial dependencies within MTS, which limits the enhancement of anomaly detection performance. In contrast, DCdetector accounts for both spatial and time dependencies, leading to moderate improvements.

The proposed FL-STAM distinguishes itself by minimizing reconstruction errors while simultaneously considering both spatial and temporal correlations in the data. This methodological enhancement is pivotal in enabling the model to achieve more precise anomaly detection. For example, FL-STAM not only integrates association discrepancy to differentiate anomalous from normal data but also strengthens spatial and temporal dependency modeling, resulting in its superior performance.

5 Conclusion

This study proposes FL-STAM, a novel transformer-based framework for MTS anomaly detection, enhanced by a parallel graph structure learning mechanism. Our key contributions are: (1) a parallel graph attention mechanism capturing global serial and temporal dependencies independently, overcoming the common limitation of isolated spatial or temporal modeling; (2) a spatial-temporal multihead attention mechanism within the transformer, simultaneously modeling spatiotemporal associations and utilizing their Wasserstein Distance discrepancy for nuanced anomaly understanding, significantly boosting accuracy; (3) the integration of federated learning, enabling privacy-preserving distributed training across devices, enhancing scalability and addressing IIoT data security. Evaluations on five benchmark datasets confirm FL-STAM's consistent superiority over SOTA methods in F1-score and computational efficiency, with ablation studies underscoring the critical role of the parallel graph structure. While demonstrating strong performance, ongoing challenges in dynamic IIoT environments warrant future research into adaptive continuous learning mechanisms, domain knowledge integration, and framework extension to diverse time-series data.

Acknowledgment. This work is supported by the National Key R&D Plan "Key Special Project of Cyberspace Security Governance" (No. 2022YFB3104700), the National Natural Science Foundation (Grant nos. 62402395, 62376198, 62076182), the Science and Technology Program of Sichuan Province (Grant no. 2023YFS0424), the Science and Technology Service Network Initiative (No. KFJ-STS-QYZD-2021-21-001), and the Talents by Sichuan provincial Party Committee Organization Department, and Chengdu - Chinese Academy of Sciences Science and Technology Cooperation Fund Project (Major Scientific and Technological Innovation Projects).

References

1. Brody, S., Alon, U., Yahav, E.: How attentive are graph attention networks? arXiv preprint arXiv:2105.14491 (2021)
2. Chaovalitwongse, W.A., Fan, Y.J., Sachdeo, R.C.: On the time series k-nearest neighbor classification of abnormal brain activity. IEEE Trans. Syst. Man Cybernetics-Part A: Syst. Humans **37**(6), 1005–1016 (2007)
3. Darban, Z.Z., Webb, G.I., Pan, S., Aggarwal, C., Salehi, M.: Deep learning for time series anomaly detection: A survey. ACM Comput. Surv. **57**(1), 15:1–15:42 (2025). https://doi.org/10.1145/3691338
4. Deng, J., Chen, X., Jiang, R., Yin, D., Yang, Y., Song, X., Tsang, I.W.: Disentangling structured components: Towards adaptive, interpretable and scalable time series forecasting. IEEE Trans. Knowl. Data Eng. **36**(8), 3783–3800 (2024). https://doi.org/10.1109/TKDE.2024.3371931
5. Goodge, A., Hooi, B., Ng, S.K., Ng, W.S.: Robustness of autoencoders for anomaly detection under adversarial impact. In: Bessiere, C. (ed.) Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence, pp. 1244–1250. International Joint Conferences on Artificial Intelligence Organization (2021). https://doi.org/10.24963/IJCAI.2020/173
6. Hundman, K., Constantinou, V., Laporte, C., Colwell, I., Soderstrom, T.: Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In: Guo, Y., Farooq, F. (eds.) Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 387–395. ACM, New York (2018). https://doi.org/10.1145/3219819.3219845
7. Kiss, I., Genge, B., Haller, P., Sebestyén, G.: Data clustering-based anomaly detection in industrial control systems. In: 2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP), pp. 275–281. IEEE (2014)
8. 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: International Conference on Artificial Neural Networks, pp. 703–716. Springer (2019)
9. Liu, F.T., Ting, K.M., Zhou, Z.: Isolation forest. In: Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy, pp. 413–422. IEEE Computer Society (2008).https://doi.org/10.1109/ICDM.2008.17
10. Shen, L., Li, Z., Kwok, J.: Timeseries anomaly detection using temporal hierarchical one-class network. Adv. Neural. Inf. Process. Syst. **33**, 13016–13026 (2020)
11. Tealab, A.: Time series forecasting using artificial neural networks methodologies: a systematic review. Future Comput. Inform. J. **3**(2), 334–340 (2018). https://doi.org/10.1016/j.fcij.2018.10.003

12. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
13. Wang, H.z., Li, G.q., Wang, G.b., Peng, J.c., Jiang, H., Liu, Y.t.: Deep learning based ensemble approach for probabilistic wind power forecasting. Appl. Energy **188**, 56–70 (2017)
14. Xu, J., Wu, H., Wang, J., Long, M.: Anomaly transformer: Time series anomaly detection with association discrepancy. CoRR abs/2110.02642 (2021). https://arxiv.org/abs/2110.02642
15. Zhang, C., et al.: A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 1409–1416 (2019)
16. Zhou, X., Dai, C., Wang, W., Qiu, T.: Global-local association discrepancy for multivariate time series anomaly detection in iiot. IEEE Internet Things J. **11**(7), 11287–11297 (2024). https://doi.org/10.1109/JIOT.2023.3330696
17. Zong, B., et al.: Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: International Conference on Learning Representations (2018)