Nonlinear and Trend-Aware Industrial Time Series Anomaly Detection with Federated Learning

Zhiqing Luo, Yan Qin

School of Automation, Chongqing University, Chongqing, 401331, China
zhiqing.luo@stu.cqu.edu.cn, yan.qin@cqu.edu.cn

ABSTRACT

Industrial anomaly detection aims to identify significant data deviations. However, it is hampered by the complex dynamics of time series, distributed data silos, and data heterogeneity. To overcome these challenges, we introduce a novel federated learning framework (FL) with two core modules: Multiple Definition Operators (MDO) to capture intricate temporal dynamics, and Temporal Trend Convolution (TTC) to extract interpretable trend patterns. FL enables multiple clients to collaboratively train a robust global model without centralizing raw data, thereby boosting generalization and preserving privacy. Critically, a tailored data-sharing strategy is implemented within the framework to mitigate the challenge of non-independent and identically distributed data. Experiments conducted on the Skoltech Anomaly Benchmark and other real-world datasets validate the efficacy of the MDO and TTC modules as well as confirm that the proposed framework significantly improves anomaly detection performance, demonstrating its practical potential for industrial applications.

1. INTRODUCTION

The proliferation of Internet of Things (IoT) and edge devices in industry has revolutionized data transmission. However, anomalies within these systems can lead to catastrophic outcomes, making their accurate and timely identification paramount. Owing to their powerful ability to learn complex features from temporal data, deep neural networks have become a cornerstone of time series anomaly detection. For instance, Zhang et al. (2023) addressed data challenges with self-supervised adaptive memory networks, while Jeong et al. (2023) employed data degradation schemes. In a similar vein, Deng & Hooi (2021) enhanced accuracy by combining graph neural networks with attention mechanisms.

Despite these advancements, the decentralized and privacy-sensitive nature of industrial data renders centralized model training impractical. Federated learning (FL) (McMahan et al. (2017)) offers a privacy-preserving alternative, enabling collaborative training across clients without exposing raw data. Consequently, FL has been explored for industrial anomaly detection using various architectures, including hybrid CNN-LSTMs (Liu et al. (2020)), stacked LSTMs (Sater & Hamza (2021)), and parameter-efficient models (Xu et al. (2024)). However, a major challenge in FL caused by non-independent and identically distributed (non-IID) data across industrial clients is client drift, which degrades global model performance. To address this challenge, this work makes a pragmatic trade-off by implementing a strategic data-sharing mechanism, which mitigates client drift and enhance model generalization.

However, designing a powerful local model for industrial time series is non-trivial. Many anomaly detection approaches frame the task as time series forecasting, where significant deviations from predictions indicate anomalies. While several advanced forecasting models like ITransformer (Liu et al. (2024)), TimesNet (Wu et al. (2023)) and TSLANet (Eldele et al. (2024)) have shown promise, they still have critical limitations. Specifically, approaches based on periodicity decomposition, such as TimesNet, can introduce redundancy and fail to capture comprehensive trend dynamics. More broadly, despite various strategies like inverted attention or interactive convolution and frequency denoising analysis, a fundamental gap remains: the inability to adaptively model nonlinear patterns while extracting interpretable temporal trends.

Based on the above motivations, this article proposes a new model named multiple definition operators with temporal trend convolution (MDOC), which is designed for nonlinear feature extraction and the acquisition of trend information. Moreover, we collaboratively develop this model to establish a federated framework named Fed-MDOC to address both data privacy and model generalization concerns. A data-sharing strategy is further being utilized to address the issue of client shift arising from the non-IID characteristics of distributed industrial data.

The structure of this paper unfolds as follows: following this introduction, the subsequent methodology is detailed in sec-

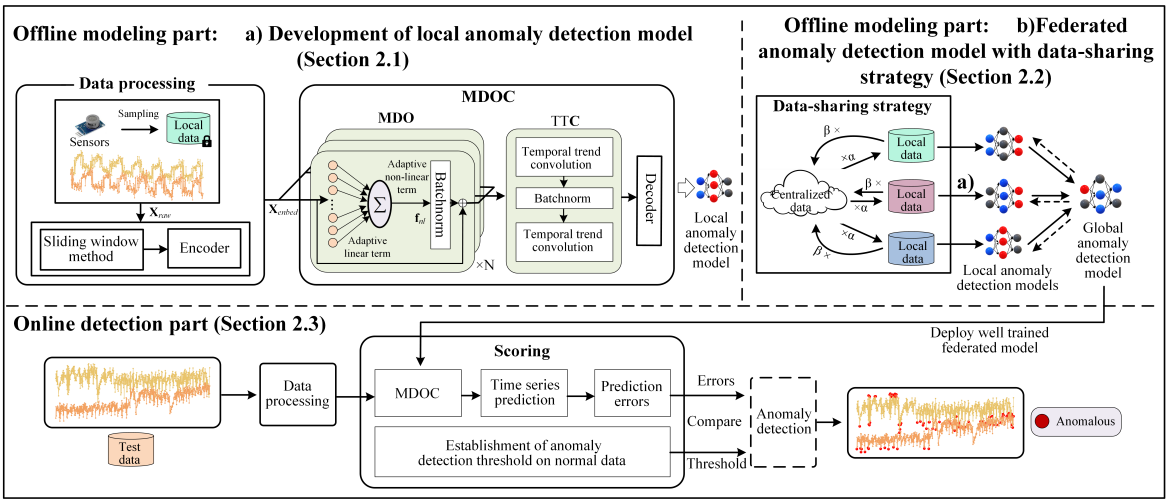

Figure 1. Overall anomaly detection process of our proposed Fed-MDOC.

tion 2. Section 3 presents the experiments and results, along with analysis. Finally, a conclusion is drawn in section 4.

2. METHODOLOGY

In this section, a federated anomaly detection method is designed as shown in Fig. 1, which includes offline modeling and online detection parts. The first part consists of local anomaly detection model development and FL with data-sharing strategy to aggregate a generalized global anomaly detection model. And the second part deploys the well trained global model for anomaly detection, which is achieved by comparing the errors between predicted and true values against a threshold learned from normal data.

2.1. Development of Local Anomaly Detection Model

The MDOC is proposed as the local anomaly detection model which leverages two novel modules, i.e., the MDO and the TTC.

2.1.1. Non-linear feature extraction with MDO

For each client k, the raw data matrix $\mathbf{X}_{raw} \in \mathbb{R}^{N_k \times D}$ from D sensors is transformed into a three-dimensional training tensor $\mathbf{X}_{train} \in \mathbb{R}^{(N_k - T) \times T \times D}$ using a sliding window of size T. To allow each variable to focus solely on its feature extraction, the raw batch sequences $\mathbf{X}_{batch} \in \mathbb{R}^{B \times D \times T}$, where B denotes the batch size, are encoded to a embedding input matrix $\mathbf{X}_{embed} \in \mathbb{R}^{B \times D \times E}$ independently through applying a linear layer, in which E is the embedding dimension of the encoder. Next, a linear coefficient $\mathbf{W} \in \mathbb{R}^{1 \times D \times E}$ is randomly initialized and the linear component output $\mathbf{f}_l \in \mathbb{R}^{B \times D \times 1}$ is below:

$$\mathbf{f}_l = \sum_{dim=-1} \mathbf{X}_{embed} \cdot Repeat(\mathbf{W}, (B, 1, 1)), \tag{1}$$

Eq. 1 reflects the interaction between the adaptive linear coefficient and the batch embedded data. Moreover, the adaptive nonlinear term $\widetilde{\mathbf{W}} \in \mathbb{R}^{1 \times D \times K}$ is designed to further derive nonlinear attributes as follows:

$$\begin{aligned} \mathbf{f} &= Repeat(\mathbf{f}_l, (1, 1, K)), \\ \widetilde{\mathbf{W}} &= Repeat(\widetilde{\mathbf{W}}, (B, 1, 1)), \\ \mathbf{f}_{nl} &= \left[\mathbf{f}_{bdk}^{\widetilde{\mathbf{W}}_{bdk}} \right]_{B,D,K}, \end{aligned} \tag{2}$$

where K means the length of predicted sequences, and $\mathbf{f}_{nl}$ is the output which contains abundant nonlinear information of each definition operator.

Finally, the batch normalized outputs are then summed and combined with a residual connection to prevent gradient vanishing.

2.1.2. Local Anomaly Detection Model with TTC

To address this limitation in trend discernment, the innovative TTC method is developed to augment the model's sensitivity of underlying trend dynamics.

As shown in Fig. 2, the TTC module first transforms the 1D temporal data into a 2D representation using a time-lag operation. A subsequent padding step aligns the temporal dimension of this 2D structure with the original sequence, making it compatible with convolutional operations. These processes ensure that each convolution kernel's receptive field encompasses three distinct temporal trend patterns: past-aware pattern, past and future-aware pattern, future-aware pattern. By iteratively applying above procedures across the entire origi-

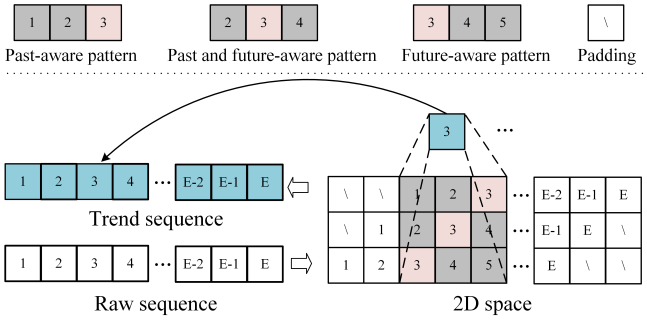

Figure 2. The proposed temporal trend convolution (TTC).

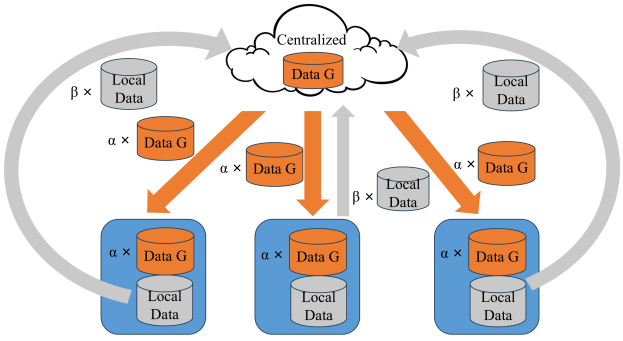

Figure 3. The data-sharing strategy.

nal series, a trend sequence that effectively encapsulates valuable underlying trend dynamics is costructed. Finally, the predicted values $\hat{\mathbf{Y}}$ are produced by a linear layer which decodes the generated trend-aware series to the prediction space. And a local anomaly detection model $\hat{\Theta}$ is derived through the gradient-based learning method:

$$\hat{\Theta} \leftarrow \hat{\Theta} - \eta \nabla \mathcal{L}(\hat{\mathbf{Y}}, \mathbf{Y}), \tag{3}$$

where η is the learning rate for weight learning, and $\mathcal{L}$ denotes the loss function evaluated by the mean squared error.

2.2. Federated Anomaly Detection Model with Data-sharing Strategy

The non-IID nature of data distributions across different clients leads to a biased global model during aggregation at the central server. Therefore, a data-sharing strategy within the federated framework is employed to mitigate these problems by reducing the impact of data distribution discrepancy among clients.

Initially, we employ a random client selection method, where a subset of clients is randomly chosen from the entire pool for each training round. Furthermore, as illustrated in Fig. 3, the core principle of our proposed strategy is to aggregate a small proportion, denoted by β, of each client's training data into a centralized dataset on the server, referred to as the concentrated data and represented by set $\mathbf{G}$. At the beginning of each round, a proportion α of this concentrated data (set $\mathbf{G}$) is sampled and combined with each participating client's local training data. This mixed dataset then serves as the new training data for each client. And then, for the aggregation of parameters from all trained clients into a global model, which subsequently initializes the local models for the next round, we utilize an element-wise averaging approach:

$$\mathbf{W}_{t+1} \leftarrow \sum_{c \in \mathbf{C}_s} \frac{1}{N} \mathbf{W}_t^c, \tag{4}$$

where $\mathbf{C}_s$ represents the set of trained clients in the ground t, N is the length of $\mathbf{C}_s$, and $\mathbf{W}_t^c$ is the parameters of client c in the ground t.

Afterwards, the parameters of local models are averaged to produce the next communication round global model $\mathbf{W}_{t+1}$. The training finally yields the final global model $\mathbf{W}_\zeta$ after ζ communication rounds.

2.3. Online Anomaly Detection

For the task of anomaly detection, squared prediction error (SPE) needs to be introduced to measure the control limits. In terms of normal samples:

$$SPE_{norm} = \sum_{d=1}^{D} (\hat{y}_d - y_d)^2, \tag{5}$$

D denotes variable dimension of each normal sample.

And the control limits δ_p can be derived based on the SPE_{norm}:

$$\delta_p = g\chi^2_{h,p}, \tag{6}$$

$$g = \frac{v_k}{2m_k}, \quad h = \frac{2m_k^2}{v_k} \tag{7}$$

$$m_k = Mean[SPE_{norm}], \quad v_k = Var[SPE_{norm}], \tag{8}$$

where $\chi^2_{h,p}$ represents the critical value of the chi-square distribution, with the confidence level of p and the degree of freedom of h. $Mean[\cdot]$ and $Var[\cdot]$ are the operations for taking the mean and taking the variance, respectively.

Following the training of the global anomaly detection model described in Section 2.2, it is deployed to all local clients to facilitate online anomaly detection, upon arrival of the test samples $\mathbf{X}_{test} = (\mathbf{X}_1, ..., \mathbf{X}_N)$, where $\mathbf{X}_i \in \mathbb{R}^{D \times T}$. They are fed into the deployed offline model $f_{\mathbf{W}_\zeta}$ to yield predictions $\hat{\mathbf{Y}}_{test} = f_{\mathbf{W}_\zeta}(\mathbf{X}_{test})$. The errors between the predictions $\hat{\mathbf{Y}}_{test}$ and the true values are then compared with a threshold learned from normal data, and errors exceeding this threshold indicate anomalies.

Table 1. MDO and TTC test on Anomaly_free

Models	LSTM		MDO+LSTM		MDO		TTC		MDO+TTC	
Metrics	$RMSE(\%)$	R^2	$RMSE(\%)$	R^2	$RMSE(\%)$	R^2	$RMSE(\%)$	R^2	$RMSE(\%)$	R^2
Accelerometer1RMS	8.91±0.09	0.7832±0.0037	8.98±0.20	0.7797±0.0686	8.00±0.18	0.8252±0.0067	7.57±0.46	0.8429±0.0203	6.96±0.66	0.8666±0.0480
Accelerometer2RMS	8.49±0.15	0.8179±0.0057	8.31±0.17	0.8253±0.0046	8.46±0.19	0.8190±0.0060	7.65±0.38	0.8516±0.0089	7.34±0.78	0.8621±0.0252
Current	21.86±0.12	-0.1750±0.0085	19.80±0.12	0.0362±0.0105	20.01±0.07	0.0159±0.0079	16.72±1.50	0.3071±0.0971	11.63±1.37	0.6627±0.0657
Pressure	11.44±0.04	-0.0647±0.0055	10.99±0.05	0.0177±0.0148	11.05±0.04	0.0063±0.0052	8.58±0.76	0.3967±0.0631	7.37±0.66	0.5541±0.0795
Temperature	6.18±0.03	0.5359±0.0050	5.85±0.06	0.5833±0.0077	5.89±0.03	0.5782±0.0062	5.09±0.17	0.6839±0.0211	4.83±0.26	0.7150±0.0088
Thermcouple	0.56±0.02	0.9653±0.0020	0.59±0.03	0.9608±0.0041	0.28±0.01	0.9909±0.0005	0.45±0.04	0.9769±0.0046	0.57±0.08	0.9627±0.0088
Voltage	26.02±0.32	-0.5096±0.0456	20.88±0.11	0.0276±0.0117	20.92±0.06	0.0243±0.0080	18.29±1.77	0.2465±0.1034	12.05±1.82	0.6686±0.0656
Volume Flow RateRMS	4.68±0.05	0.4870±0.0110	4.09±0.02	0.6094±0.0036	4.24±0.01	0.5793±0.0053	3.67±0.09	0.6839±0.0176	3.66±0.30	0.6844±0.0324
Avg	11.02±0.10	0.3500±0.0108	9.94±0.09	0.4800±0.0157	9.85±0.07	0.4799±0.0057	8.50±0.65	0.6236±0.0420	6.80±0.74	0.7470±0.0417

3. Experiments And Results

3.1. Data Preparation

To evaluate the prediction ability of the MDOC model, we used the time series real-world datasets, e.g., ETT (Zhou et al. (2021)), ECL[1], Exchange, and Traffic. Moreover, we conducted anomaly detection experiments on Skoltech Anomaly Benchmark (SKAB)[2] dataset.

- **ETT:** This dataset contains transformer temperature readings from electricity systems, divided into hourly-sampled categories (ETTh1, ETTh2) and minute-level sampled categories (ETTm1, ETTm2).
- **ECL:** This dataset contains electricity consumption records for 321 customers, recorded hourly from 2012 to 2014.
- **Exchange:** This dataset provides daily exchange rate data for eight countries spanning 1990 to 2016.
- **Traffic:** Hourly traffic flow data from 862 sensors on San Francisco freeways are included in this dataset.
- **SKAB:** This project offers a baseline platform for evaluating anomaly detection, containing one anomaly-free dataset and 34 datasets existing abnormal conditions. And each dataset contains eight variables.

3.2. Ablation Study

For the statistical robustness, all reported metrics are the average of ten independent experimental runs. Table 1 presents the results for the eight variables on the SKAB Anomaly_free dataset, reported as the mean ± standard deviation calculated from the ten trials. The results demonstrate a consistent improvement in the predictive performance of the LSTM model across most dimensions when MDO is applied. For the TTC model, significant gains were observed in the prediction of $Current$, $Pressure$, $Temperature$, and $Voltage$, while minimal differences were noted in the performance of other dimensions. Moreover, the MDO+TTC model exhibits the lowest mean $RMSE$ (6.80% ± 0.74%) and the highest R^2 (0.7470 ± 0.0417). Therefore, all these results clearly demonstrate the effectiveness of MDO.

Though had digged the nonlinear characteristics through MDO, the trend information is not extracted sufficiently. Consequently, Table 1 also records the enhancement of TTC to MDO to verify its influence. With the exception of the $Thermocouple$ dimension, the predictive performance for

Table 2. comparing MDOC with Others. And Avg means the average results from Outputs lengths $T \in \{1, 24, 48, 96\}$

Models		MDOC		TSLANet		ITransformer		TimesNet	
Metric		RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
$ETTh1$	1	**0.0331**	**0.9687**	0.0537	0.9178	0.0540	0.9167	0.0549	0.9138
	24	0.0926	0.7549	**0.0884**	**0.7763**	0.0972	0.7300	0.1023	0.7006
	48	0.0994	0.7174	**0.0955**	**0.7387**	0.1030	0.6963	0.1056	0.6805
	96	0.1070	0.6716	**0.1031**	**0.6950**	0.1111	0.6460	0.1199	0.5874
	Avg	**0.0830**	0.7782	0.0852	**0.7820**	0.0913	0.7473	0.0957	0.7206
$ETTh2$	1	**0.0183**	**0.9936**	0.0272	0.9858	0.0265	0.9866	0.0269	0.9862
	24	**0.0456**	**0.9601**	0.0470	0.9577	0.0500	0.9521	0.0512	0.9498
	48	0.0537	0.9447	**0.0533**	**0.9456**	0.0565	0.9388	0.0587	0.9339
	96	0.0630	0.9238	**0.0620**	**0.9264**	0.0653	0.9184	0.0705	0.9047
	Avg	**0.0452**	**0.9556**	0.0474	0.9539	0.0496	0.9490	0.0518	0.9437
$ETTm1$	1	**0.0194**	**0.9888**	0.0357	0.9620	0.0357	0.9621	0.0345	0.9646
	24	**0.0816**	**0.8017**	0.1063	0.6640	0.1008	0.6974	0.0959	0.7263
	48	**0.1038**	**0.6794**	0.1345	0.4614	0.1267	0.5220	0.1174	0.5900
	96	**0.1024**	**0.6876**	0.1217	0.5589	0.1164	0.5967	0.1203	0.5690
	Avg	**0.0768**	**0.7893**	0.0995	0.6616	0.0949	0.6946	0.0920	0.7125
$ETTm2$	1	**0.0114**	**0.9975**	0.0174	0.9941	0.0174	0.9941	0.0169	0.9944
	24	**0.0317**	**0.9804**	0.0385	0.9711	0.0400	0.9687	0.0406	0.9678
	48	**0.0418**	**0.9659**	0.0485	0.9541	0.0490	0.9531	0.0538	0.9436
	96	**0.0468**	**0.9572**	0.0502	0.9507	0.0518	0.9476	0.0554	0.9401
	Avg	**0.0329**	**0.9753**	0.0387	0.9675	0.0395	0.9659	0.0417	0.9615
ECL	1	0.0419	0.9607	0.0438	0.9515	**0.0349**	**0.9719**	0.0423	0.9587
	24	0.0730	0.8809	0.0712	0.8866	**0.0681**	**0.8964**	0.0715	0.8857
	48	0.0797	0.8580	0.0789	0.8608	**0.0771**	**0.8671**	0.0793	0.8594
	96	0.0842	0.8418	0.0840	0.8425	**0.0826**	**0.8476**	0.0861	0.8343
	Avg	0.0697	0.8854	0.0695	0.8854	**0.0657**	**0.8958**	0.0698	0.8845
$Exchange$	1	0.0122	0.9969	0.0116	0.9972	**0.0114**	**0.9973**	0.0124	0.9968
	24	**0.0257**	**0.9861**	0.0283	0.9830	0.0349	0.9742	0.0349	0.9743
	48	0.0382	0.9691	**0.0381**	**0.9693**	0.0472	0.9529	0.0453	0.9567
	96	0.0572	0.9285	**0.0533**	**0.9379**	0.0637	0.9115	0.0673	0.9013
	Avg	0.0333	0.9702	**0.0328**	**0.9719**	0.0393	0.9590	0.0400	0.9573
$Traffic$	1	**0.0445**	**0.9226**	0.0501	0.8682	0.0435	0.8824	0.0495	0.8479
	24	0.0653	0.8328	0.0693	0.8117	**0.0628**	**0.8456**	0.0672	0.8230
	48	0.0694	0.8113	0.0756	0.7764	0.0682	0.8182	**0.0659**	**0.8301**
	96	0.0772	0.7657	0.0808	0.7434	**0.0751**	**0.7786**	0.0769	0.7680
	Avg	0.0641	**0.8331**	0.0690	0.7999	**0.0624**	0.8312	0.0649	0.8173
1st count		17	17	8	9	9	8	1	1

[1]The specific description of ECL dataset can refer to https://archive.ics.uci.edu/dataset/321/electricityloaddiagrams20112014

[2]SKAB is a specially developed by Skoltech used for evaluation of anomaly detection core open source project. (https://github.com/waico/SkAB)

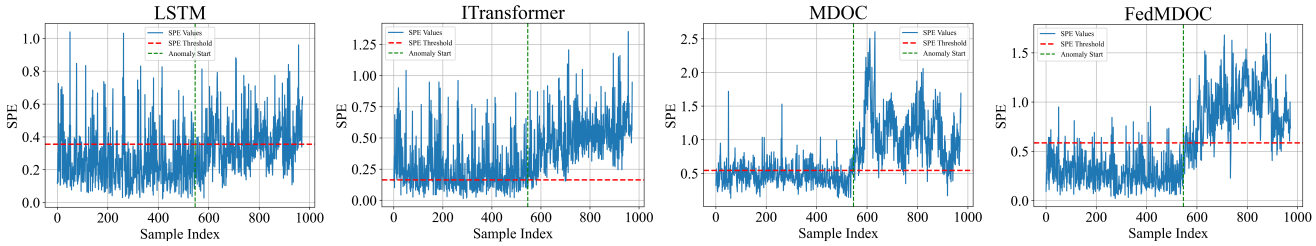

Figure 4. The SPE values and thresholds of different models.

all other dimensions of MDO+TTC (MDOC) exhibited significant improvements, and the averaged $RMSE$ across all dimensions was reduced by 30.96% compared to the MDO.

3.3. Time Series Forecasting Using the Proposed MDOC

To further validate the superior capabilities of the MDOC model, we compared it against several advanced models (i.e., TSLANet, ITransformer and TimesNet) in the time series forecasting domain with the output lengths $T \in \{1, 24, 48, 96\}$ while the size of lookback window is 24. Moreover, the complete forecasting results are presented in Table 2, with the best results highlighted in red. The term Avg refers to the arithmetic mean across all considered output lengths. As evidenced by the results in Table 2, the MDOC model exhibited a notable level of long-term prediction performance across the evaluated models, achieving 17 best results out of 35 instances for both $RMSE$ and R^2. All the aforementioned evidence further supports the efficacy of our proposed approach, also validating the MDOC model as the cornerstone model for the subsequent federated anomaly detection task.

3.4. Federated Anomaly Detection

To maintain the distinction between normal and abnormal data during forward propagation, the anomaly detection experiments did not employ a series-stationarization process. For the Fed-MDOC model, datasets were divided into 13 training, 5 validation, and 16 testing working conditions, respectively. During the testing phase, a SPE value exceeding the threshold $\delta_{0.95}$, which was predetermined using the whole normal data from the $anomaly\_free$ dataset in the confidence level of 0.95, was classified as anomaly.

Moreover, the SPE values of $Other11$ are presented in contrasted Fig. 4, data points above the horizontal dotted red decision line are classified as outliers, while those below are considered normal points. Values on the right side of the green line are labeled as anomalies. The MDOC-based models demonstrate superior anomaly detection, effectively separating normal and anomalous data both visually and quantitatively (Table 3). And the federated approach (FedMDOC) further enhances this by learning from diverse operational conditions. Notably, the federated model makes a strategic trade-off, accepting a minor loss in accuracy for a significantly higher F1-score, indicating a better balance between false positives and negatives.

Furthermore, the FedMDOC with data-sharing strategy (FedDS) is primarily characterized by the two aforementioned parameters, α and β. The effectiveness of this strategy is illustrated by comparing its performance with that of FedAvg on the normal data of all test working conditions. As presented in Table 4, underlined values represent results that outperform FedAvg, and the best performing value is highlighted in red among these. It demonstrates that greater values of α and β generally result in improved performance, as indicated by lower MSE values. This can be attributed to the fact that larger α and β values introduce a greater proportion of diverse data from other clients, facilitating the learning of global data characteristics by the local model. Moreover, the anomaly detection performance of FedDS is also recorded in Table 4, the FedDS can also maintain the detection ability while improving forecasting precision.

4. Conclusion

We proposed a MDOC model to extract the nonlinearity and abundant trend information of time series in this paper. And we have testified the promotion of MDO and TTC modules to models. Moreover, experiments show the superior ability of anomaly detection of Fed-MDOC because it owns a vision of vary fault conditions what can have a better generalization. Finally, the impact of introducing a data-sharing strategy along with its key parameters α and β on the performance of the federated model is discussed. In real-world anomaly detection scenarios, the operational states of equipments are influenced by numerous factors, leading to significant heterogeneity in data collected from the same client. Consequently, future research will focus on enhancing the domain generalization capabilities of federated learning in such contexts.

References

Deng, A., & Hooi, B. (2021). Graph neural network-based anomaly detection in multivariate time series. , *35*(5), 4027–4035.

Table 3. The detection performance of models

Models	**FedDS**		**FedMDOC**		**MDOC**		**ITransformer**		**LSTM**		**Transformer**	
Metrics	*Accuracy*	*F*1	*Accuracy*	*F*1	*Accuracy*	*F*1	*Accuracy*	*F*1	*Accuracy*	*F*1	*Accuracy*	*F*1
*Other*9	1.0	0.9654	1.0	0.9703	1.0	0.8231	1.0	0.6684	1.0	0.8696	1.0	0.6860
*Other*10	0.9430	0.9523	0.9608	0.9591	0.9946	0.8823	0.9822	0.7355	0.9661	0.8800	0.9875	0.7355
*Other*11	0.6112	0.7290	0.8805	0.9060	0.9461	0.8236	0.9789	0.7066	0.4660	0.5489	0.9718	0.7100
*Other*12	0.8561	0.8776	0.8526	0.8741	0.8877	0.7574	0.9578	0.5889	0.8842	0.7753	0.9684	0.5974
*Other*14	1.0	0.9553	1.0	0.9488	1.0	0.7886	1.0	0.5840	1.0	0.7819	1.0	0.5889
*Valve*1_9	0.8439	0.8799	0.8571	0.8888	0.9179	0.7494	0.9629	0.6346	0.8835	0.7582	0.9550	0.6417
*Valve*1_12	0.9786	0.9557	0.9893	0.9636	1.0	0.8650	1.0	0.6619	0.9866	0.8361	1.0	0.6538
*Valve*1_15	0.8210	0.8690	0.8263	0.8758	0.8815	0.7528	0.9500	0.6383	0.8526	0.7448	0.9552	0.6385

Table 4. The MSE of parameters α and β

α \ β	**0.05**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
0.05	0.0259	0.0252	0.0236	0.0230	0.0252	0.0231
0.1	0.0281	0.0244	0.0240	0.0219	0.0224	0.0226
0.2	0.0241	0.0249	0.0229	0.0249	0.0242	0.0215
0.3	0.0247	0.0253	0.0216	0.0225	0.0214	0.0201
0.4	0.0256	0.0242	0.0215	0.0191	0.0201	0.0187
0.5	0.0244	0.0220	0.0223	0.0206	0.0186	**0.0166**
FedAvg				0.0227		

Eldele, E., Ragab, M., Chen, Z., Wu, M., & Li, X. (2024). TSLANet: Rethinking Transformers for Time Series Representation Learning. In *41 international conference on machine learning.*

Jeong, Y., Yang, E., Ryu, J. H., Park, I., & Kang, M. (2023). AnomalyBERT: Self-Supervised Transformer for Time Series Anomaly Detection using Data Degradation Scheme. In *International conference on learning representations.*

Liu, Y., Garg, S., Nie, J., Zhang, Y., Xiong, Z., Kang, J., & Hossain, M. S. (2020). Deep anomaly detection for time-series data in industrial iot: A communication-efficient on-device federated learning approach. *IEEE Internet of Things Journal*, *8*(8), 6348–6358.

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., & Long, M. (2024). itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations.*

McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273–1282).

Sater, R. A., & Hamza, A. B. (2021). A federated learning approach to anomaly detection in smart buildings. *ACM Transactions on Internet of Things*, *2*(4).

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., & Long, M. (2023). Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International conference on learning representations.*

Xu, R., Miao, H., Wang, S., Yu, P. S., & Wang, J. (2024). Pefad: a parameter-efficient federated framework for time series anomaly detection. In *Proceedings of the 30th acm sigkdd conference on knowledge discovery and data mining* (pp. 3621–3632).

Zhang, Y., Wang, J., Chen, Y., Yu, H., & Qin, T. (2023). Adaptive memory networks with self-supervised learning for unsupervised anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, *35*(12), 12068-12080.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 35, pp. 11106–11115).