

FedLGAN: a method for anomaly detection and repair of hydrological telemetry data based on federated learning

Zheliang Chen¹, Xianhan Ni¹, Huan Li², Xiangjie Kong^{Corresp. 2}

¹ Zhejiang Provincial Hydrological Management Center, Zhejiang Provincial Hydrological Management Center, Hangzhou, Zhejiang, China

² College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, Zhejiang, China

Corresponding Author: Xiangjie Kong

Email address: xjkong@ieee.org

The existing data repair methods primarily focus on addressing data missing issues by utilizing variational autoencoders to learn the underlying distribution and generate content that represents the missing parts, thus achieving data repair. However, this method is only applicable to data missing problems and cannot identify abnormal data. Additionally, as data privacy concerns continue to gain public attention, it poses a challenge to traditional methods. This paper proposes a generative adversarial network model based on the federated learning framework and long short-term memory network, namely the FedLGAN model, to achieve anomaly detection and repair of hydrological telemetry data. In this model, the discriminator in the generative adversarial network structure is employed for anomaly detection, while the generator is utilized for abnormal data repair. Furthermore, to capture the temporal features of the original data, a bidirectional long short-term memory network with an attention mechanism is embedded into the generative adversarial network. The federated learning framework avoids privacy leakage of hydrological telemetry data during the training process. Experimental results based on four real hydrological telemetry devices demonstrate that the FedLGAN model can achieve anomaly detection and repair while preserving privacy.

1 FedLGAN: A Method for Anomaly Detection 2 and Repair of Hydrological Telemetry Data 3 Based on Federated Learning

4 Zheliang Chen¹, Xianhan Ni¹, Huan Li², and Xiangjie Kong²

5 ¹Zhejiang Provincial Hydrological Management Center, Hangzhou, China

6 ²College of Computer Science and Technology Zhejiang University of Technology,
7 Hangzhou, China

8 Corresponding author:

9 Xiangjie Kong¹

10 Email address: czl@zjwater.gov.cn, szszy_nxh@zjwater.gov.cn,

11 zjutlihuan@outlook.com, xjkong@ieee.org

12 ABSTRACT

13 The existing data repair methods primarily focus on addressing data missing issues by utilizing variational
14 autoencoders to learn the underlying distribution and generate content that represents the missing parts,
15 thus achieving data repair. However, this method is only applicable to data missing problems and cannot
16 identify abnormal data. Additionally, as data privacy concerns continue to gain public attention, it poses a
17 challenge to traditional methods. This paper proposes a generative adversarial network model based on
18 the federated learning framework and long short-term memory network, namely the FedLGAN model,
19 to achieve anomaly detection and repair of hydrological telemetry data. In this model, the discriminator
20 in the generative adversarial network structure is employed for anomaly detection, while the generator
21 is utilized for abnormal data repair. Furthermore, to capture the temporal features of the original data,
22 a bidirectional long short-term memory network with an attention mechanism is embedded into the
23 generative adversarial network. The federated learning framework avoids privacy leakage of hydrological
24 telemetry data during the training process. Experimental results based on four real hydrological telemetry
25 devices demonstrate that the FedLGAN model can achieve anomaly detection and repair while preserving
26 privacy.

27 INTRODUCTION

28 In recent years, with the increasing uncertainty of global natural disasters, the construction of smart
29 hydrology has received more and more attention. Its purpose is to build an integrated hydrological
30 telemetry system that incorporates cloud computing, big data, and other technologies, in order to observe
31 and record hydrological phenomena occurring in nature in a more real-time and accurate manner, providing
32 a data foundation for hydrological research (Yan et al., 2019; Corbari et al., 2019; Karimi et al., 2019;
33 Kong et al., 2022). Obviously, as the primary source of hydrological data, hydrological telemetry
34 devices bear the responsibility of data collection and storage. The ability of telemetry devices to
35 provide accurate and reliable hydrological data directly affects fundamental decisions such as flood
36 control and drought resistance scheduling, ecological environmental protection, and comprehensive
37 development of water resources. However, in the actual operation process, telemetry devices often
38 encounter factors such as system failures, equipment aging, and weak signals in remote locations,
39 leading to abnormal situations such as numerical errors, partial data loss, and severe data gaps in the
40 collected hydrological data (Qin and Lou, 2019). This seriously affects the integrity, authenticity, and
41 accuracy of hydrological data, directly resulting in a significant reduction in the capabilities of various
42 hydrological model statistical analyses. Therefore, identifying anomalies in hydrological data, mining
43 the underlying data features, and simultaneously repairing abnormal data are of great significance in
44 improving hydrological forecasting performance and reducing losses caused by uncertainty in disasters.
45 For time series data such as hydrological telemetry data, existing abnormal detection methods mostly

utilize the advantages of Long Short-Term Memory (LSTM) networks in learning temporal features and construct coupled models in combination with other detection algorithms (Cook et al., 2019; Blázquez-García et al., 2021). (Malhotra et al., 2015) stacked LSTM networks to learn higher-level temporal features and made predictions on the data over multiple time steps. Considering the effectiveness and real-time requirements of abnormal detection algorithms, (Ding et al., 2019) proposed using LSTM models to evaluate the real-time anomalies of each univariate sensor time series, followed by a Gaussian mixture model for multidimensional joint detection of possible anomalies. (Xu et al., 2020) proposed a new fusion algorithm, LSTM-GAN-XGBOOST, to detect anomalies in deep features of massive time series data. (Niu et al., 2020) introduced an LSTM-based Variational Autoencoder-Generative Adversarial Network model (LSTM-based VAE-GAN) that jointly trains the encoder and generative adversarial network, leveraging the mapping ability of the encoder and the discriminative ability of the discriminator, significantly reducing the time required for anomaly detection. However, the aforementioned studies focus more on anomaly detection rather than repairing the detected abnormal data.

Considering that in practical scenarios, the identification and repair of abnormal data often need to be addressed synchronously, which involves detecting the abnormal data and then processing the abnormal portions. Even, compared to anomaly detection, data repair has more important significance. (Kong et al., 2023) proposed a dynamic graph convolutional recursive interpolation network (DGCRIN) to interpolate and repair traffic data, which employed a graph generator and dynamic graph convolutional gated recurrent unit (DGCGRU) to perform fine-grained modeling of the dynamic spatiotemporal dependencies of road network. In order to achieve both anomaly detection and repair for time-series data simultaneously, (Zhang et al., 2017) designed an iterative minimum-change-perception repair algorithm called IMR, which demonstrates high adaptability to existing anomaly detection techniques such as AR and ARX. (Park et al., 2021) proposed a robust sliding window-based LightGBM model, where anomalies are detected using a variational autoencoder (VAE), followed by the use of random forest to repair the anomalies.

However, the aforementioned studies did not take into account the privacy issues present in the training data. Hence, in this study, we presents a generative adversarial network model based on the federated learning framework and long short-term memory network, which achieves both anomaly detection and repair for time-series data while ensuring data privacy protection. The model consists of three main components: the federated learning framework, the generative adversarial network model, and the attention-based long short-term memory network. The federated learning framework utilizes its unique mechanism of keeping data local to preserve privacy and employs the federated averaging algorithm to aggregate local training parameters for updating the global model. The generative adversarial network is the core part of the model, composed of a generator and a discriminator, which are optimized through adversarial training. We utilizes the property of the generator in the generative adversarial network to fit real data for data repair, while the discriminator's ability to distinguish between real and generated data enables anomaly detection. The attention-based bidirectional long short-term memory network is incorporated to better handle sequential data and further explore the temporal dependencies in hydrological data. Experimental results on four real datasets demonstrate that the generative adversarial network model based on federated learning outperforms other control group methods in multiple metrics, effectively achieving anomaly detection and repair for time-series data. The contributions of this paper are summarized as follows:

- (I) We propose a distributed model based on the federated learning framework, LSTM and GAN, called FedLGAN, which achieves efficient and accurate anomaly detection and data repair while ensuring data privacy. To the best of our knowledge, it is the first time to use the federated learning framework in the context of anomaly detection and data repair for hydrological telemetry data.
- (II) Integrating the attention-based bidirectional LSTM into the generative adversarial network enables effective capturing of the complex dynamics and temporal correlations in hydrological telemetry data. This enhancement strengthens the model's interpretability of anomalies and its capability for data repair.
- (III) By conducting extensive experiments on datasets from four real hydrological stations, we demonstrate the superiority and effectiveness of the proposed FedLGAN model.

The remainder of this paper is organized as follows: In the part of related work, we will review the cutting-edge methods for anomaly detection and data repair in hydrological telemetry data. The, we

100 describe the foundational knowledge of our framework in the preliminary. In methodology, we introduce
101 the proposed framework. The experiment part provides the performance evaluation, and concludes the
102 paper finally.

103 RELATED WORK

104 Anomaly Detection for Hydrological Telemetry Data

105 Hydrological data anomaly detection is an important research field that holds significant implications
106 for water resource management, flood forecasting, climate change studies, and more. Existing research
107 methods include statistical modeling approaches such as clustering and classification algorithms, as well
108 as deep learning methods such as convolutional neural networks and recurrent neural networks.

109 These methods leverage the learning of patterns and features within hydrological data to achieve more
110 accurate detection of anomalies. (Kulanuwat et al., 2021) developed a median-based statistical outlier
111 detection approach using a sliding window technique. (Shao et al., 2020) proposed a detection algorithm
112 called AR-iForest, which is a hydrological time series anomaly detection algorithm based on Isolation
113 Forest. It uses an autoregressive model to predict the current data and calculate the confidence interval.
114 Data that falls outside this interval is identified as an anomaly. To enhance the stability of anomaly
115 detection results, (Liu et al., 2020) proposed a parallel anomaly detection algorithm called Flink-iForest,
116 which combines the use of the iForest algorithm with the k-means algorithm to address the threshold
117 partitioning problem. In contrast to their approach, (Sun et al., 2017) proposed a density-based anomaly
118 pattern detection method specifically tailored for large-scale hydrological data with a significant amount of
119 noise. This method addresses the high time complexity issue of traditional anomaly detection algorithms.

120 Although the above methods demonstrate good performance in detecting extreme and specific value
121 anomalies, they are prone to missing small anomalies. Furthermore, these methods often struggle
122 to uncover the underlying spatiotemporal information in hydrological sequences and fail to provide
123 explanations for the types and causes of anomalies.

124 Data Repair for Hydrological Telemetry Data

125 Hydrological telemetry data has always been a scarce and valuable resource. However, these data are
126 susceptible to interference, leading to anomalies such as missing values and abrupt changes during the
127 collection and transmission processes. Therefore, the restoration of hydrological data anomalies has
128 always been a research problem of great significance (Gao et al., 2018).

129 The existing hydrological data repair methods primarily involve constructing time series models such
130 as autoregressive (AR) and moving average (MA) models to learn the distribution characteristics of the
131 data. These models are then used to predict, interpolate, or reconstruct the anomalous portions of the
132 data. Among these methods, deep learning-based time series models such as LSTM and RNN are widely
133 applied in practice. (He et al., 2023) proposed a deep learning model named Con-GRU for repairing water
134 level monitoring data with long-term anomalies, which captures both long-term and local time-dependent
135 features via one-dimensional convolution (Conv1D) and gated recurrent units (GRU). (Gill et al., 2007)
136 proposed a short-term prediction method for groundwater levels in well fields by combining artificial
137 neural networks (ANN) and support vector machines (SVM). They utilized interpolation techniques to
138 fill in missing data and tested their approach based on the observed data. (Heras and Matovelle, 2021)
139 used automatic learning machines of the Python Scikit Learn module, which integrates a wide range of
140 automated learning algorithms, such as Linear Regression and Random Forest.

141 Currently, there is limited research on the repair of abnormal parts in hydrological data, and previous
142 methods used by researchers have become somewhat outdated and may not be suitable for the current
143 characteristics of multimodal and complex hydrological data. Not only that, but there are also few
144 methods that can simultaneously achieve anomaly detection and data repair. Furthermore, the privacy
145 of hydrological data has been receiving increasing attention, leading to a decrease in available data.
146 Therefore, there is an urgent need for a new method that can ensure data privacy while achieving these
147 two important functionalities.

PRELIMINARY

Federated Averaging Algorithm

As the most widely used classical algorithm in federated learning framework, federated average (FedAvg) describes the process of server weighted aggregation of local model parameters, and then updates the global model. In this process, it is assumed that there are K clients in total, and the servers aggregate t times in total. First, the central server initializes the global model w_t , and then selects at least one up to k clients to participate in the training. Each selected client simultaneously receives the global model w_t delivered by the server, trains the respective local model w_{t+1}^k with their own data and sends it back to the server. The server will receive all local models and aggregate them in the way of weighted average to get the next round of global model weight w_{t+1} , which is calculated by Eqs. 1:

$$w_{t+1} \leftarrow \sum_{k=1}^K \frac{n_k}{n} w_{t+1}^k \quad (1)$$

In this paper, unsupervised time series anomaly detection based on federated learning framework. At the same time, FedAvg algorithm is used to coordinate the interaction between server and client. However, in view of the limitations of FedAvg in the NonIID scenario, we partially improve the FedAvg algorithm to obtain a relatively personalized federated anomaly detection model.

Attention-based bidirectional LSTM

The bidirectional LSTM based on the attention mechanism was initially proposed by (Bahdanau et al., 2014) for sequence modeling and prediction. It combines the bidirectional LSTM and attention mechanism to better capture contextual information and important features in the input sequence. In the traditional Bidirectional LSTM, the input sequence is processed by two LSTM layers in both forward and backward directions. The forward LSTM computes in the order of the input sequence from the beginning to the end, while the backward LSTM computes in the reverse order. In this way, the forward and backward LSTMs capture the forward and backward context information of the input sequence, respectively, generating two sets of hidden state sequences.

To better utilize these hidden state sequences, the attention mechanism is introduced. The attention mechanism allows the model to dynamically assign weights to the inputs based on their importance. It calculates attention weights at each time step, focusing the attention on the most relevant parts of the input sequence for the current prediction. This enables the model to pay more attention to key information in the input sequence, thereby improving the performance of modeling and prediction.

GAN

The basic idea of GAN is derived from the "two-player zero-sum game" in game theory, and its main structure contains a generative model G and a discriminative model D . Among them, generator G is used to generate data, while discriminator D 's main task is to distinguish the real data from the fake data forged by G . G is committed to learning the distribution of real data to fool the discriminator, and the two are optimized in the process of confrontation. The loss function of GAN optimization is as follows:

$$\min_G \max_D V(D, G) = E_{x \sim P_{data}(x)} [\log D(x)] + E_{z \sim P_z(z)} [\log(1 - D(G(z)))] \quad (2)$$

When D can no longer distinguish the real data from the forged data, the ideal state of GAN training is reached. In our federated anomaly detection task, we use transformer as the generator G of GAN to reconstruct the original sequence, and strengthen the ability of transformer to learn data distribution by means of confrontation.

METHODOLOGY

In this section, we formally introduce our distributed framework, FedLGAN, which is designed for anomaly detection and data repair. More specifically, it includes the overall framework of FedLGAN, the basic idea behind the framework, and the key technologies.

Overall Framework

The overall framework of FedLGAN is depicted in Figure 1, which can be divided into three parts: the collaborative training part, the anomaly detecting part and the data repairing part. The basic idea of FedLGAN is to use the federated learning framework to cooperatively train the data of multiple edge devices, thereby improve the detection ability and generalization of the model. Among them, the federated learning framework is used to provide a secure distributed scene to protect the data privacy of edge devices, and the adversarial training mode of GAN is used to enhance the data repair capability of the generator and the anomaly discrimination ability of the discriminator. The LSTM is used to mine the degree of correlation and multi-scale sequence features of sequences. The LSTM is used to improve the GAN's ability to capture the temporal dependencies in time series. We will introduce it in more detail in the following section.

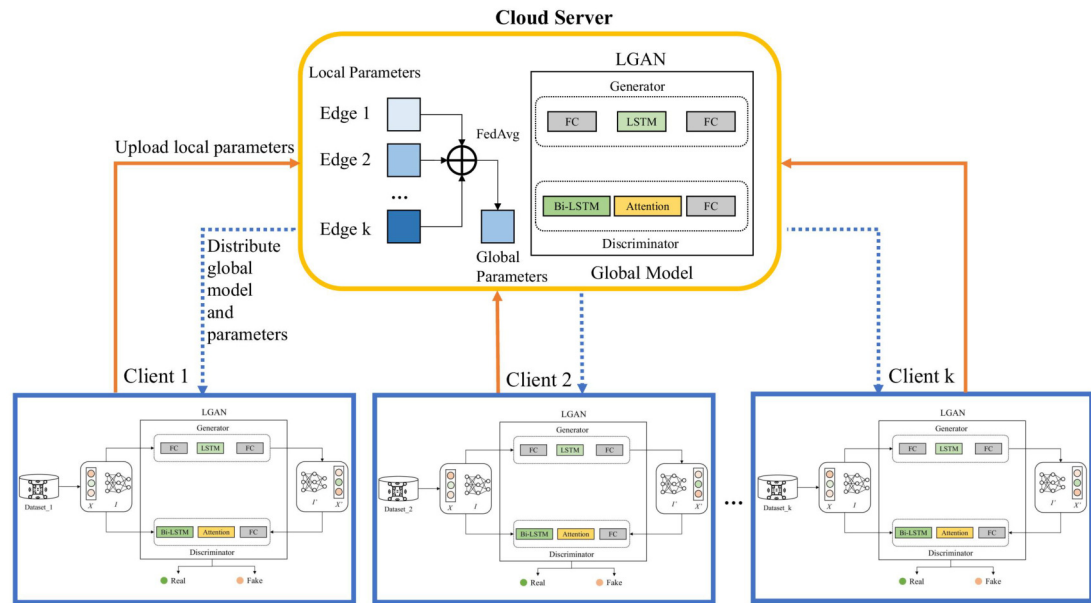


Figure 1. Overall Framework of FedLGAN

Collaborative Training

Figure 1 shows that the overall structure of the model training is based on the federated learning framework and GAN. In the process of local model collaborative training, we use the cloud server to initialize the global model and distribute it to each edge. After receiving it, the edge will input the local preprocessed normal data into the local model and start training. Figure 2 shows the structure of the local model, namely the LGAN.

We further explain the framework of adversarial training stage of FedLGAN in details. As shown in Figure 2, the generator G and the discriminator D have similar structures, both of which are composed of LSTM blocks. Firstly, we convert the input sequence X into the tensor form $I \in \mathbb{R}^{L \times f}$ with modality, where L represents the length of the sequence, and f is the dimension of potential representation. In the case of Vanilla GAN, neither the G nor the D has a specific structure to handle time series data. Therefore, the lack of consideration for the unique temporal characteristics of time series data during training is a major reason for the generator's weak ability to fit real data and the discriminator's low accuracy in detecting anomalies. To address these issues, we have introduced Long Short-Term Memory (LSTM) networks and bidirectional LSTM networks with attention mechanism in the G and D parts of the generative adversarial network model, respectively. At the same time, in the training process of GAN, the discrimination ability of the discriminator needs to be slightly greater than the camouflage ability of the generator, otherwise it will easily lead to the problem of mode collapse.

Therefore, the discriminator is often trained multiple times before the generator is trained once. First initialize the generator G and fix it, start training the discriminator D , take the real data I and forged data I

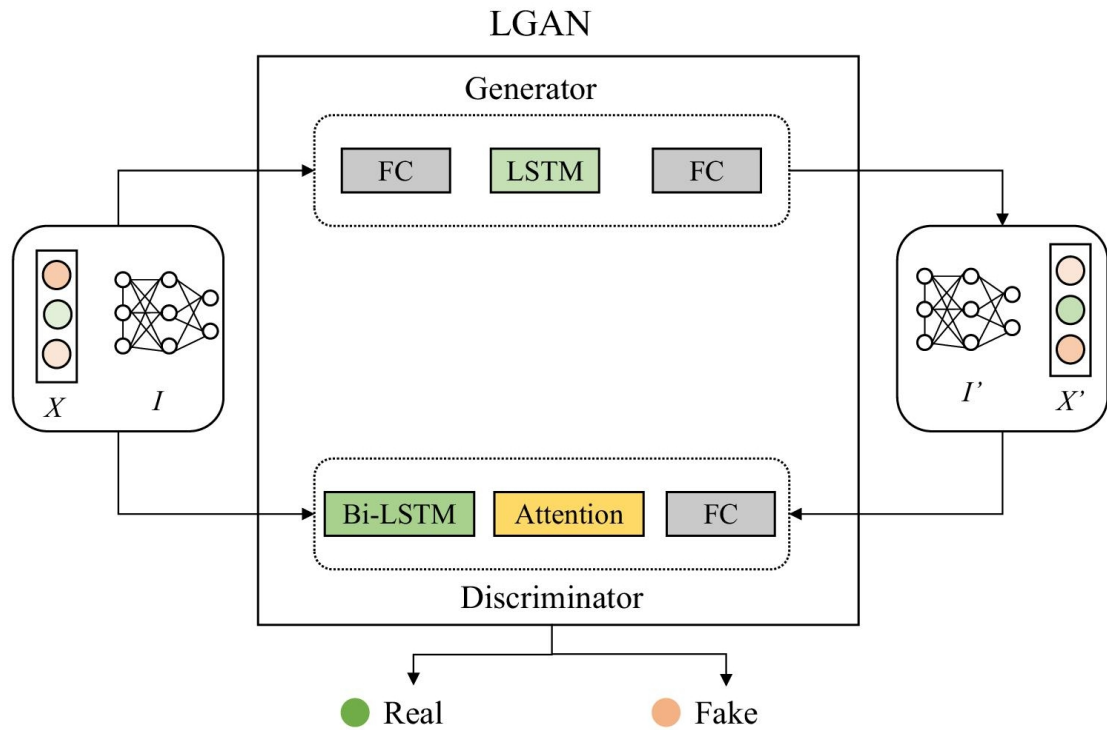


Figure 2. The structure of LGAN

as the input of D , and pass through the bidirectional LSTM layer, attention layer and The fully connected layer finally outputs the identification result. We feed the processed tensor I into the discriminator D . As shown in Figure 3 and Figure 4, the tensor I first enters the LSTM layer. The core of LSTM is the memory unit, which is cut or added information through a structure called gate to control the circulation and loss of features. This structure determines the degree to which the LSTM unit maintains the previous state and remembers the extracted features of the current data input. It has 3 gates of the control unit state, which are the input gates, forget gates and output gates are calculated by Eqs. 3, 4 and 5:

$$i_t = \sigma_g(W_i I_t + U_i h_{t-1} + b_i) \quad (3)$$

$$f_t = \sigma_g(W_f I_t + U_f h_{t-1} + b_f) \quad (4)$$

$$o_t = \sigma_g(W_o I_t + U_o h_{t-1} + b_o) \quad (5)$$

According to the above gate function and formula, the cell state at time can be calculated, and the calculation formula is as following equation:

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (6)$$

where i_t , f_t and o_t are the output values of the input gate, forget gate and output gate at time t respectively, I_t is the t th input sequence; h_{t-1} refers to the t th time The hidden layer state of, W , U and b are the weight vector, parameters and offset of the gating unit respectively. \tilde{C}_t represents the unit state update value, σ is the activation function, and the Sigmoid function is generally used. LSTM is composed of a series of memory unit chains, and controls the transmission state through gating settings, remembers information that needs to be memorized for a long time, and forgets unimportant content, so as to mine the time series variation rules of relatively long intervals and delays in the time series.

The bidirectional LSTM based on the attention mechanism is improved on the traditional LSTM, as shown in Figure 4, by adding a reverse LSTM layer to the original forward LSTM network layer, the purpose is to consider the context of the two directions To increase the available information of the network (Schuster and Paliwal, 1997). Therefore, different from the traditional LSTM, the network structure

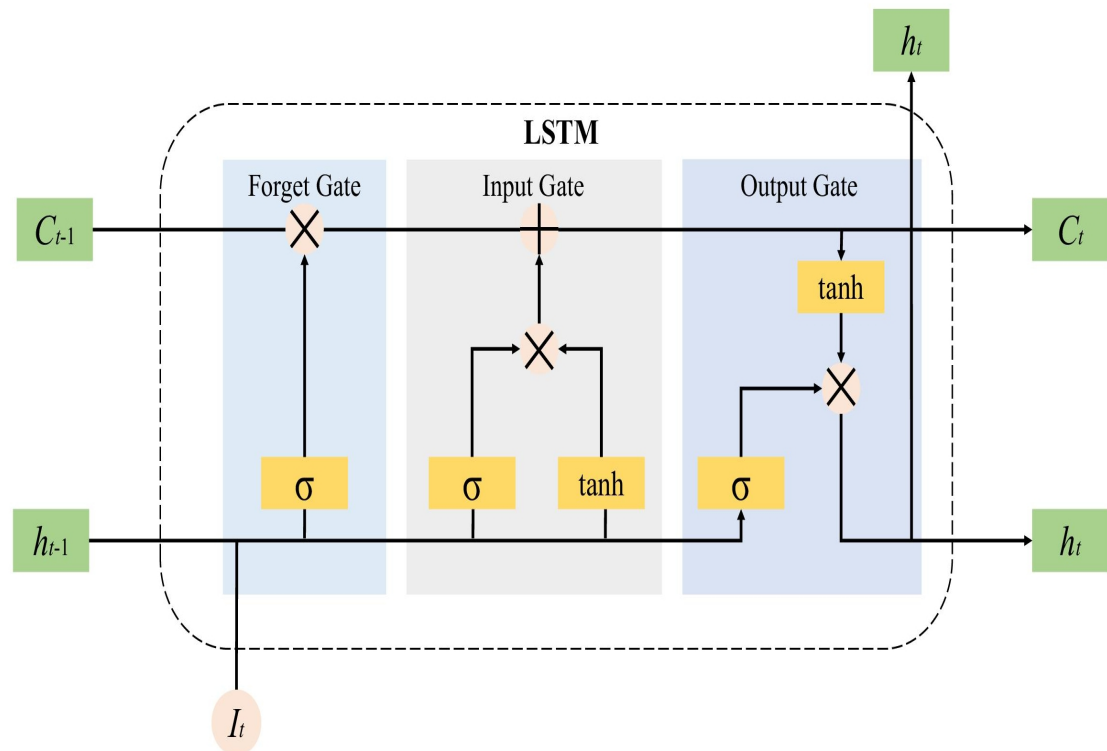


Figure 3. Internal unit of LSTM

contains two forward-passed \vec{h}_t and backward-passed \overleftarrow{h}_t respectively, then the calculation formula of the hidden layer state h_t at this time is as follows:

$$h_t = [\vec{h}_t \oplus \overleftarrow{h}_t] \quad (7)$$

To improve the learning ability of the discriminator, an attention mechanism is also introduced. The matrix for extracting the weights of this layer is

$$M = \tanh(H) \quad (8)$$

$$\alpha = \text{softmax}(w^T M) \quad (9)$$

And pass the product of and the weight matrix as the output of the attention layer:

$$r = H\alpha^T \quad (10)$$

213 where H is the output of the LSTM layer, w^L is the transposition of a parameter vector obtained by training
 214 and learning, α is the weight matrix, and r is the output of this layer. By adding the improvement of the
 215 above structure to the generative confrontation network, the ability of the D to detect anomalies and the
 216 ability of the G to fit the data can be enhanced at the same time, thereby improving the performance of the
 217 model as a whole. Finally, we input the result into the fully connected layer network and use the activation
 218 function sigmoid to fix the value in the $[0,1]$ interval, and then we can get the probability value P that each
 219 time stamp t of the sequence I is normal. If the input of the discriminator is I , that is, real and normal
 220 data, the judgment value of the output result at all time points is as close to 1 as possible, otherwise the
 221 output tends to 0. Obviously, for the discriminator D , whether it is generated data or abnormal data, it is
 222 hoped that the output result will be as close to 0 as possible. For the generator G , its training process is
 223 similar to that of the discriminator D . We input the tensor I into G , and pass through the LSTM layer and
 224 the fully connected layer in turn. After the output can be obtained from formulas 3, 4, and 5, we can get
 225 the reconstructed data I' by adding it to the fully connected layer. In this way, we can reconstruct and
 226 replace abnormal data, so as to realize data repair.

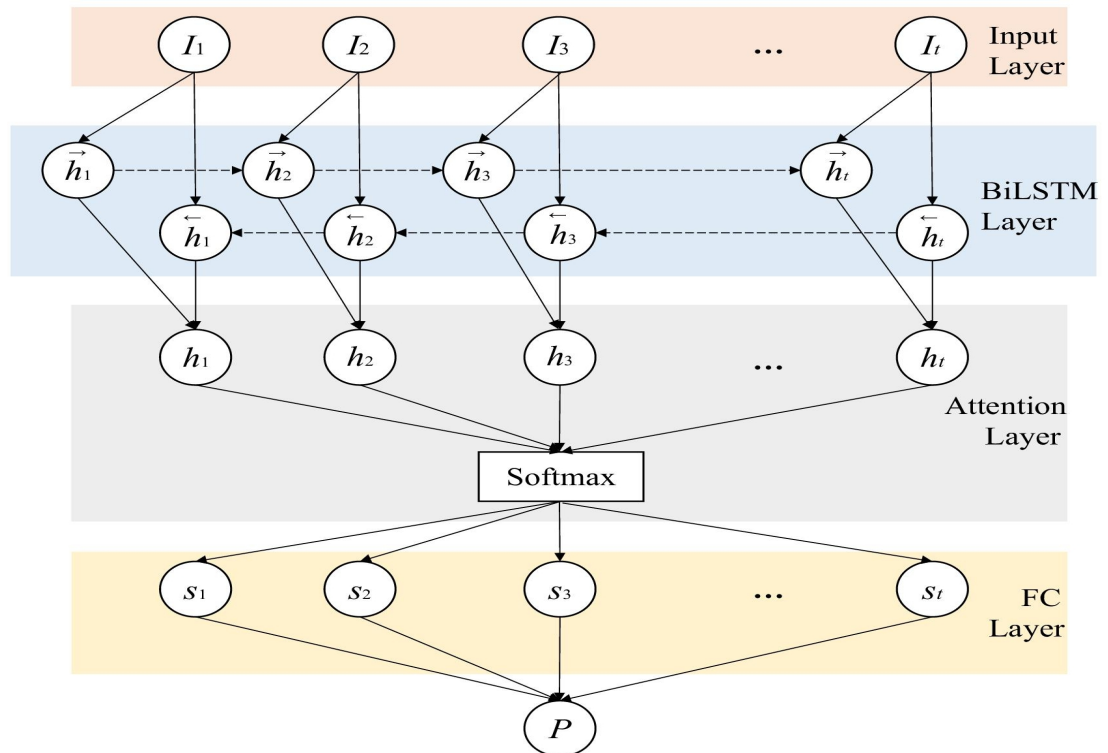


Figure 4. The details of BiLSTM

Our local model uses two LSTM blocks, mainly to form an adversarial structure. This way of confrontation forces the G to fully learn the characteristic information of normal data, thereby cheat the discriminator D in the training process. At the same time, the distinguish transformer D is also trying to distinguish between real data and reconstructed data, which are constantly optimized during process of confrontation. In the framework of federated learning, k edge devices use their local data for training. After a certain number of iterations, each client uploads its own training parameters to the cloud server for aggregation. Among them, we use the most classic federated average algorithm for aggregation. After that, the cloud server redistributes the aggregated parameters and models to each client to let them start training again, and so on until convergence. The specific process of model collaborative training is shown in Algorithm 1.

Optimization Method

Since our local model is based on two LSTM blocks, the optimization process of the model meets the training standard of GAN. That is, we update generator G and discriminator D alternately. In the m -th iteration, when the D is trained, we fixed G and D is updated according to the following equation:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{m=1}^M [\log D(I_k^{(m)}) + \log(1 - D(G(I_k^{(m)})))] \quad (11)$$

Similarly, we fixed D , and updated G according to the following equation:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{m=1}^M \log(1 - D(G(I_k^{(m)}))) \quad (12)$$

where M denotes the maximum iterations of local training. In order to make the model converge more easily, we also use gradient punishment to force Lipschitz constraint. Therefore, the loss function of

Algorithm 1 Model Collaborative Training Stage

Input: The Generator G and Discriminator D ; The total optimization round M ; The edge devices indexed by k and their training samples I_k ; Initialized global model parameters: W_{global}^0 ; The local model parameters: $W_{local}^{0,k}$; The ratio of D and G training times per round: N ;

Output: A well-trained G ; A well-trained D ;

```

1: for each round  $m = 1, 2, \dots, M$  do
2:   for each edge devices  $k$  do
3:      $W_{local}^{m,k} \leftarrow W_{global}^m$ 
4:     for each round  $n = 1, 2, \dots, N$  do
5:        $\nabla_{\theta_d} [\log D(I) + \log(1 - D(G(I_k)))]$ 
6:     end for
7:      $\nabla_{\theta_g} \log(1 - D(G(I_k)))$ 
8:      $W_{global}^{m+1} \leftarrow \text{FedAvg}[W_{local}^{m,0}, \dots, W_{local}^{m,k}]$ 
9:      $W_{local}^{m+1,k} \leftarrow W_{local}^{m+1}$ 
10:   end for
11: end for
```

model training stage in m -th iteration for D can be defined as follows:

$$L_D = \frac{1}{m} \sum_{m=1}^M [\log D(I_k^{(m)}) + \log(1 - D(G(I_k^{(m)})))] + \lambda (\|\nabla_{\tilde{I}_k} D(\tilde{I}_k^{(m)})\|_2 - 1)^2 \quad (13)$$

where $\tilde{I}_k^{(m)} = \epsilon I_k^{(m)} + (1 - \epsilon)G(I_k^{(m)})$ refers to the data that random interpolation sampling on the line of $I_k^{(m)}$ and $G(I_k^{(m)})$. In order to keep the reconstructed data close to the original data, we also take reconstruction loss as the optimization strategy of the generator. As a result, for the G , we get a new loss function of model training phase in m -th iteration:

$$L_G = \frac{1}{m} \sum_{m=1}^M \log(1 - D(\tilde{I}_k^{(m)})) + \|I_k^{(m)} - \tilde{I}_k^{(m)}\|^2 \quad (14)$$

Anomaly Detection and Data Repair

Figure 1 shows the steps of federated anomaly detection and data repair. First, the cloud server distributes the final trained model parameters to all edge clients, and each client updates the parameters in the local model after receiving them. Then, it enters the anomaly detection and data repair stage. In this stage, we input sequence I into generator G and discriminator D , and finally we can get the reconstruct sequence I' and the anomaly time points in the detection sequence.

EXPERIMENT

In this section, we introduce the details of the experiment, including datasets, model settings, evaluation indicators, etc. Then, compare the performance of our model FedLGAN with other methods. In addition, we also analyzed the hyperparameters of the model.

Datasets

We used the hydrological data collected by four hydrological telemetry devices in Hangzhou, Jinhua, Shaoxing, and Lishui in Zhejiang Province in the past three months to conduct experiments to ensure that the data sources for model training and testing are authentic and reliable. However, due to the differences in the equipment models and geographical locations of different telemetry sites, the data recording interval and the attributes of the collected data may be different. Therefore, in the experiment, the common attributes of the hydrological equipment of the four telemetry stations are extracted, and the collection records of data such as water level, rainfall, and voltage are counted at intervals of 5 minutes. In addition, according to the actual situation, the unreasonable abnormal data is divided into a separate test set for the

abnormal detection part of the experiment. Considering that the data collected by the device is abnormal only in a few cases, it is necessary to artificially dirty part of the normal data to provide a sufficient amount of abnormal data for testing. Since the data set of each hydrological telemetry equipment contains a total of 90 days of data from January 1 to March 31, 2022, the author screened all normal data within 90 days as the training set, and the last 15 days of normal data and The data after they are dirty is used as the test set.

Experimental Settings

All experiments are run on the same server, the host operating system is Ubuntu 18.04, the memory is 128 GB, the CPU is Intel(R) Xeon(R) Gold, 16-core dual-thread, and the graphics card is NVIDIA Quadro P6000. The Pytorch version is v1.6.0, the initial learning rate is set to 0.000 1, and the batch size is set to 64. In addition, the experiment uses mean absolute error (Mean Absolute Error, MAE), mean square error (Mean Square Error, MSE), root mean square error (Root Mean Square Error, RMSE), and mean absolute percentage error (Mean Absolute Percentage Error, MAPE) as the evaluation index. Among them, the data of various indicators of anomaly detection are calculated by comparing the detection results of the discriminator in the model with the real labels, while the index data of data repair is obtained by using the formula of normal data before dirtying and data repaired by the generator calculated.

Comparison Experiments

In order to reflect the superiority of the generative adversarial network based on federated learning, this experiment compares it with four control algorithms. The comparative algorithms cover parametric methods, non-parametric methods, and deep learning methods. Since this experiment includes both anomaly identification and data repair, it is considered to compare these two parts separately, where anomaly detection part includes LSTM(Hochreiter and Schmidhuber, 1997) and GRU(Cho et al., 2014), and data repair part includes VAE(Kingma and Welling, 2013) and GAN(Goodfellow et al., 2020).

- **LSTM(Hochreiter and Schmidhuber, 1997)**: a special RNN that performs better on longer sequences.
- **GRU(Cho et al., 2014)**: a variant of LSTM that removes the forget gate and consists only of an update gate and a reset gate.
- **VAE(Kingma and Welling, 2013)**: a structure composed of an encoder and a decoder, which is trained to minimize the reconstruction error between the encoded and decoded data and the initial data, and its essence is to estimate the density of a function with hidden variables
- **GAN(Goodfellow et al., 2020)**: a deep learning model of unsupervised learning, which consists of a generator and a discriminator, and uses the idea of confrontation to continuously optimize the model.

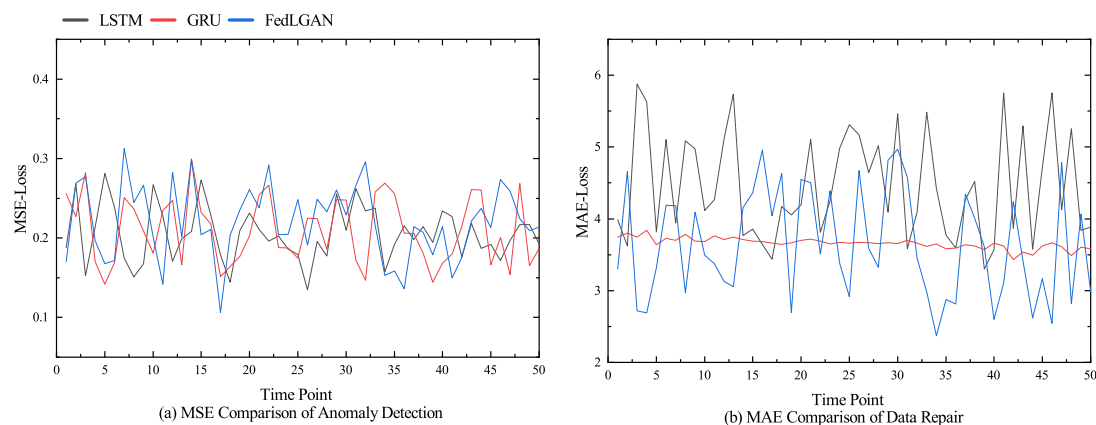


Figure 5. Performance Comparison of Different Models.

Model	MAE	MSE	MAPE%
LSTM	0.371	0.212	13.541
GRU	0.393	0.230	17.411
FedLGAN	0.480	0.238	74.139

Table 1. Comparison of anomaly detection performance.

Model	MAE	MSE	MAPE%
VAE	3.447	2.151	72.607
GAN	4.420	1.843	85.940
FedLGAN	3.430	1.708	54.824

Table 2. Comparison of data repair performance

290 The performance comparison results of different models for anomaly detection and data repair are
 291 shown in Figure 5, and Figure 6 shows the performance of various indicators for model data restoration.
 292 The more details are shown in Table 1 and Table 2, which respectively list the average results of each
 293 group's final testing in anomaly detection and data repair. From Table 1, it can be seen that LSTM
 294 and GRU have significant advantages in time series prediction compared to the discriminator of GAN
 295 model, with various indicator data of 0.371, 0.212, and 13.541%, as well as 0.393, 0.230, and 17.411%,
 296 respectively. Although our proposed model embeds bidirectional LSTM into it, its discriminative ability is
 297 slightly inferior to these two traditional prediction models due to the constraints of single mode generated
 298 by the generator, which makes it impossible to consider all abnormal situations. From Table 2, it can be
 299 seen that GAN, as the backbone of image processing, also performs well in hydrological data restoration
 300 work, with various indicator data of 4.420, 1.843, and 85.940%, respectively. However, due to the
 301 algorithm not taking into account the temporal characteristics of the data and the potential for pattern
 302 collapse, there is still a significant gap between the repaired data and the original normal data. Unlike
 303 GAN, VAE explicitly models the distribution of potential variables in hydrological data using encoders
 304 and decoders, allowing for the specified distribution of generated data. Therefore, compared to GAN,
 305 VAE has a slight performance improvement in hydrological data restoration. Our proposed generative
 306 adversarial network model based on federated learning achieved the optimal experimental results, with
 307 three indicator data lower than the control group, which were 3.430%, 1.708%, and 54.824%, respectively.

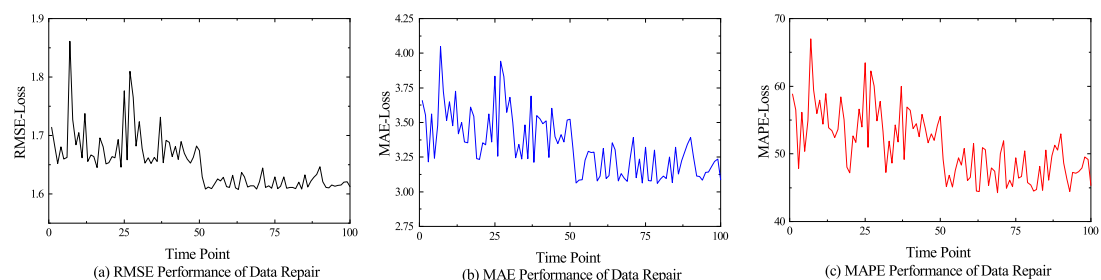


Figure 6. Performance of Data Repair.

308

309 Ablation Experiment

310 In order to demonstrate the effectiveness of the federated learning framework and the attention based
 311 long-term and short-term memory network, these two parts were removed and ablation experiments were
 312 conducted. The experimental results are shown in Table 3. Among them, FedLGAN* represents the
 313 experimental results of using only generative adversarial networks and long short-term memory networks
 314 after deleting the federated learning framework, while FedLGAN** represents the experimental results
 315 of combining the federated learning framework with generative adversarial networks after deleting the

Model	Anomaly Detection			Data Repair			Training Time/s
	MAE	MSE	MAPE%	MAE	RMSE	MAPE%	
GAN	1.0831	1.930	92.438	4.442	1.843	85.940	3868.069
FedLGAN*	0.740	0.729	75.037	4.831	2.013	82.639	3123.836
FedLGAN**	2.279	6.390	91.495	6.349	2.674	93.927	4967.078
FedLGAN	0.480	0.238	74.139	3.430	1.708	54.824	5334.658

Table 3. Performance Comparison of Ablation Experiment.

316 LSTM.

317 From Table 3, it can be seen that when only using the generative adversarial network model, good
 318 experimental results were not achieved in both anomaly detection and data repair. Considering the
 319 temporal characteristics of experimental data, FedLGAN* has a MAPE index of 75.037% and 82.639%
 320 in anomaly detection and data repair, respectively, which is significantly improved compared to the
 321 original generative adversarial network model. Unlike FedLGAN, which focuses more on data privacy
 322 and security, FedLGAN utilizes a federated learning framework to improve the original model. Its MSE
 323 index in anomaly detection is 6.390, while its RMSE index in data repair is 2.674. Although compared to
 324 the performance improvement brought by long-term and short-term memory networks, federated learning
 325 frameworks may even have a negative impact on certain indicators, slightly sacrificing the performance of
 326 the model in exchange for data privacy and security has significant practical significance and value. It
 327 is worth mentioning that due to the unique distributed training of federated learning architecture, it has
 328 high requirements for communication, so its model training time is often longer. Considering both data
 329 privacy security and its temporal characteristics, all indicators achieved optimal experimental results in
 330 the hydrological dataset. Therefore, the introduction of a federated learning framework and a bidirectional
 331 long short-term memory network based on attention mechanism in this study have both played a significant
 332 role in improving the performance of the model.

333 CONCLUSION

334 We propose a generative adversarial network model based on a federated learning framework, in which
 335 the federated learning framework acts on data privacy protection, and the discriminator and generator in
 336 the generative adversarial network are used for data anomaly detection and data restoration, respectively.
 337 In order to improve the ability of the model to extract temporal features, the two-way long-short-term
 338 memory network and the ordinary long-short-term memory network based on the attention mechanism are
 339 respectively embedded in the model's discriminator and generator. The model processes the hydrological
 340 data of the hydrological telemetry equipment into a time series matrix sequence as input, and extracts
 341 relevant time series information from the bidirectional long short-term memory network layer in the
 342 discriminator, and uses the result, namely the state of the hidden layer, as the input of the attention layer
 343 to obtain weights Matrix, finally, output the identification result through the fully connected layer to
 344 complete the abnormal identification of the data. In addition, the matrix sequence judged as abnormal
 345 data by the discriminator is also input to the generator, and its ability to fit the data distribution is used to
 346 complete data restoration. The experiment uses the real hydrological data sets of four telemetry devices in
 347 Hangzhou, Jinhua, Shaoxing and Lishui provided by the Zhejiang hydrological communication platform.
 348 The results fully prove the feasibility and superiority of the model. However, due to the limitations of data
 349 sources, it may lead to poor performance on other hydrological telemetry equipment, so the validity of the
 350 model will continue to be verified on the data collected by other hydrological telemetry equipment in the
 351 province. In addition, considering the distributed training method of the federated learning framework,
 352 compared with the centralized model, the operation efficiency is not high. The follow-up work will also
 353 focus on reducing the number of communications in federated learning and reducing the training time, so
 354 as to further improve the practicability of the network model.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 62072409, in part by the Zhejiang Provincial Natural Science Foundation under Grant LR21F020003, and in part by the R&D Program of of Zhejiang Provincial Department of Water Resources under Grant RB2216.

Grant Disclosures

The following grant information was disclosed by the authors:
Zhejiang Provincial Hydrological Management Center.
Zhejiang University of Technology.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Zheliang Chen conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and tables, authored or reviewed drafts of the paper, and approved the final draft.
- Xianhan Ni conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Huan Li conceived and designed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the paper, and approved the final draft.
- Xiangjie Kong conceived and designed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The code is available at GitHub: <https://github.com/2450848351/FedLGAN>.

Additional information:

The partially processed hydrological telemetry data of the four hydrological stations in Zhejiang Province can be obtained at: <https://github.com/2450848351/FedLGAN/tree/master/data>.

REFERENCES

- Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Blázquez-García, A., Conde, A., Mori, U., and Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)*, 54(3):1–33.
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Cook, A. A., Mısırlı, G., and Fan, Z. (2019). Anomaly detection for iot time-series data: A survey. *IEEE Internet of Things Journal*, 7(7):6481–6494.
- Corbari, C., Salerno, R., Ceppi, A., Telesca, V., and Mancini, M. (2019). Smart irrigation forecast using satellite landsat data and meteo-hydrological modeling. *Agricultural Water Management*, 212:283–294.
- Ding, N., Ma, H., Gao, H., Ma, Y., and Tan, G. (2019). Real-time anomaly detection based on long short-term memory and gaussian mixture model. *Computers & Electrical Engineering*, 79:106458.
- Gao, Y., Merz, C., Lischeid, G., and Schneider, M. (2018). A review on missing hydrological data processing. *Environmental Earth Sciences*, 77(2):47.
- Gill, M. K., Asefa, T., Kaheil, Y., and McKee, M. (2007). Effect of missing data on performance of learning algorithms for hydrologic predictions: Implications to an imputation technique. *Water Resources Research*, 43(7).

- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.
- He, L., Ji, S., Xin, K., Chen, Z., Chen, L., Nan, J., and Song, C. (2023). Application of deep learning in drainage systems monitoring data repair—a case study using con-gru model. *Water*, 15(8):1635.
- Heras, D. and Matovelle, C. (2021). Machine-learning methods for hydrological imputation data: analysis of the goodness of fit of the model in hydrographic systems of the pacific-ecuador. *Revista Ambiente & Água*, 16.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.
- Karimi, H. S., Natarajan, B., Ramsey, C. L., Henson, J., Tedder, J. L., and Kemper, E. (2019). Comparison of learning-based wastewater flow prediction methodologies for smart sewer management. *Journal of Hydrology*, 577:123977.
- Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Kong, X., Wu, Y., Wang, H., and Xia, F. (2022). Edge computing for internet of everything: A survey. *IEEE Internet of Things Journal*, 9(23):23472–23485.
- Kong, X., Zhou, W., Shen, G., Zhang, W., Liu, N., and Yang, Y. (2023). Dynamic graph convolutional recurrent imputation network for spatiotemporal traffic missing data. *Knowledge-Based Systems*, 261:110188.
- Kulanuwat, L., Chantrapornchai, C., Maleewong, M., Wongchaisuwat, P., Wimala, S., Sarinnapakorn, K., and Boonya-aroonnet, S. (2021). Anomaly detection using a sliding window technique and data imputation with machine learning for hydrological time series. *Water*, 13(13):1862.
- Liu, Y., Lou, Y., and Huang, S. (2020). Parallel algorithm of flow data anomaly detection based on isolated forest. In *2020 International Conference on Artificial Intelligence and Electromechanical Automation (AIEA)*, pages 132–135. IEEE.
- Malhotra, P., Vig, L., Shroff, G., Agarwal, P., et al. (2015). Long short term memory networks for anomaly detection in time series. In *ESANN*, volume 2015, page 89.
- Niu, Z., Yu, K., and Wu, X. (2020). Lstm-based vae-gan for time-series anomaly detection. *Sensors*, 20(13):3738.
- Park, S., Jung, S., Jung, S., Rho, S., and Hwang, E. (2021). Sliding window-based lightgbm model for electric load forecasting using anomaly repair. *The Journal of Supercomputing*, 77:12857–12878.
- Qin, Y. and Lou, Y. (2019). Hydrological time series anomaly pattern detection based on isolation forest. In *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, pages 1706–1710. IEEE.
- Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.
- Shao, P., Ye, F., Liu, Z., Wang, X., Lu, M., and Mao, Y. (2020). Improving iforest for hydrological time series anomaly detection. In *Algorithms and Architectures for Parallel Processing: 20th International Conference, ICA3PP 2020, New York City, NY, USA, October 2–4, 2020, Proceedings, Part III 20*, pages 170–183. Springer.
- Sun, J., Lou, Y., and Ye, F. (2017). Research on anomaly pattern detection in hydrological time series. In *2017 14th Web Information Systems and Applications Conference (WISA)*, pages 38–43. IEEE.
- Xu, X., Zhao, H., Liu, H., and Sun, H. (2020). Lstm-gan-xgboost based anomaly detection algorithm for time series data. In *2020 11th International Conference on Prognostics and System Health Management (PHM-2020 Jinan)*, pages 334–339. IEEE.
- Yan, L., Wan, D., Zhao, Q., and Yang, Y. (2019). Research on implementation methods of edge computing in intelligent hydrology. In *Proceedings of the 13th International Conference on Ubiquitous Information Management and Communication (IMCOM) 2019 13*, pages 211–224. Springer.
- Zhang, A., Song, S., Wang, J., and Yu, P. S. (2017). Time series data cleaning: From anomaly detection to anomaly repairing. *Proceedings of the VLDB Endowment*, 10(10):1046–1057.