TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 17/21 pp1181–1193
DOI: 10.26599/TST.2023.9010073
Volume 29, Number 4, August 2024

Univariate Time Series Anomaly Detection Based on Hierarchical Attention Network

Zexi Chen, Dongqiang Jia*, Yushu Sun, Lin Yang, Wenjie Jin, and Ruoxi Liu

Abstract: In order to support the perception and defense of the operation risk of the medium and low voltage distribution system, it is crucial to conduct data mining on the time series generated by the system to learn anomalous patterns, and carry out accurate and timely anomaly detection for timely discovery of anomalous conditions and early alerting. And edge computing has been widely used in the processing of Internet of Things (IoT) data. The key challenge of univariate time series anomaly detection is how to model complex nonlinear time dependence. However, most of the previous works only model the short-term time dependence, without considering the periodic long-term time dependence. Therefore, we propose a new Hierarchical Attention Network (HAN), which introduces seven day-level attention networks to capture fine-grained short-term time dependence, and uses a week-level attention network to model the periodic long-term time dependence. Then we combine the day-level feature learned by day-level attention network and week-level feature learned by week-level attention network to obtain the high-level time feature, according to which we can calculate the anomaly probability and further detect the anomaly. Extensive experiments on a public anomaly detection dataset, and deployment in a real-world medium and low voltage distribution system show the superiority of our proposed framework over state-of-the-arts.

Key words: anomaly detection; univariate time series; self-attention; edge computing

1 Introduction

To support the perception and defense of the operation risk of the medium and low voltage distribution system, it is crucial to conduct data mining on the time series generated by the system to learn anomalous patterns[1, 2] and carry out accurate detection of anomalies for timely discovery of anomalous conditions and early alerting[3, 4]. To adapt to the processing of large-scale Internet of Things (IoT) data, edge computing is also widely used[5–9]. In fact, the medium and low voltage distribution data are a time series with complex nonlinear time-dependent and periodic characteristics due to the influence of user behavior and schedule. When sudden abnormal events occur, this periodicity will be disrupted[10]. We can monitor the changes of medium and low voltage distribution time series in real time so as to carry out operational risk assessment and safety warning for key equipment, which provides the basis for the implementation of preventive maintenance strategy of distribution network and is crucial for the safe operation of highly reliable distribution system. In fact, accurate and efficient anomaly detection for the univariate time series has always been a huge challenge

• Zexi Chen, Dongqiang Jia, Lin Yang, Wenjie Jin, and Ruoxi Liu are with State Grid Beijing Electric Power Company, Beijing 100031, China. E-mail: chenzexijeff@126.com; jdq_epri@163.com; ylbert@163.com; 303293876@qq.com; liuruoxi0121123@163.com.
• Yushu Sun is with Institute of Electrical Engineering, Chinese Academy of Sciences, Beijing 100190, China. E-mail: yushusun@mail.iee.ac.cn.
* To whom correspondence should be addressed.
Manuscript received: 2023-06-01 ; revised: 2023-06-19; accepted: 2023-07-13

and has become the focus of researchers in recent years[11, 12].

At the beginning, some statistical methods such as Auto Regressive Integrated Moving Average (ARIMA)[13], Kalman filtering[14], and wavelet analysis were applied for anomaly detection of univariate time series. Later, some models based on traditional machine learning, such as One-Class Support Vector Machine (OCSVM)[15, 16] based on singular value detection, isolation forest[17] based on ensemble learning, and *k*-Nearest Neighbor (KNN)[18] based on clustering, were proposed. However, these methods cannot capture the complex nonlinear time dependence in time series. Some deep learning approaches, which can be broadly categorized into supervised model and unsupervised model, have been proposed and used to univariate time series anomaly detection in recent years in an effort to address this issue. The supervised model is suitable for the scene where there are many anomalous labels that are accurately marked so that the models can learn the anomalous features. The nonlinear time dependence of the time series is frequently extracted using popular deep learning networks, such as Feed-forward Neural Network (FNN)[19], Convolutional Neural Network (CNN), Long Short-Term Memory Network (LSTM)[20], and informer[21], and the anomaly probability is then determined for classification using a classifier. Unsupervised models are often used in scenarios where anomalous data are far less than normal data or even negligible. They train the model using normal samples to discover the typical normal pattern of the sequence and judge the data deviating from the normal pattern as anomalies. Common unsupervised models include forecasting methods based on CNN, LSTM, or transformer[22, 23], such as LSTM-AD[24] and DeepAnt[25], and some reconstruction models based on Auto-Encoders (AEs)[26], Variational Auto-Encoders (VAEs), or Generative Adversarial Networks (GANs), such as Donut[27], LSTM-VAE[28], and MAD-GAN[29, 30]. They use either the reconstruction error or prediction error as the anomalous score.

The previous work only considered the short-term time dependence of time series data. However, in the industrial context of medium and low voltage distribution, due to the fact that the variation of electricity related time series data may be affected by the workdays and seasons, there may be certain periodic patterns. If the long-term periodic dependence cannot be captured, it will lead to the inability to learn the correct potential patterns of time series. Figure 1 visualizes the time series data of a week in a real medium and low voltage distribution system, and it can be seen that the corresponding time of each day of the week has a strong periodicity, and this periodicity will be broken when an anomaly occurs. Therefore, our focus is how to simultaneously model short-term nonlinear time dependence and long-term periodic dependence patterns of time series within a unified framework. Our goal is to learn advanced time features that consider both long-term and short-term time dependencies, and further achieve more accurate anomaly detection results.

Along this line, we propose Hierarchical Attention Network (HAN), which is an end-to-end supervised univariate time series anomaly detection framework. The proposed hierarchical attention network is mainly composed of seven day-level multi-head self-attention networks and a week-level multi-head self-attention network. The day-level attention networks can capture the fine-grained short-term time dependence and obtain the day-level feature, then the day-level feature vectors of seven days are sent to the week-level attention network to model high-level periodic long-term time dependence and get the week-level feature. Finally, the day-level feature vector and the week-level feature vector are connected at the output layer, and then the anomaly probability of timestamp t is obtained through a Multiple Layer Perceptron (MLP) network with sigmoid function. Then, we use a binary cross entropy function as the goal function to reduce the discrepancy between the anomaly probability and the ground truth anomaly label to train our model and comprehend the pattern of the anomaly data. When detecting, if the

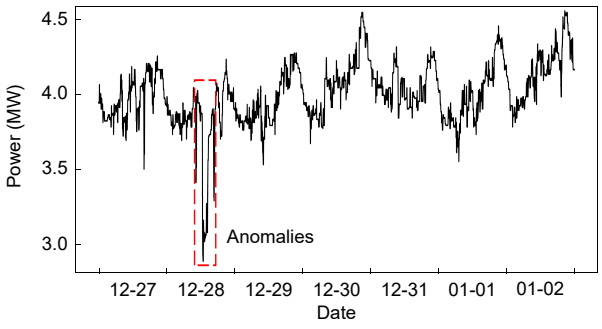


Fig. 1 An example of periodic time dependence on the ElecPower dataset. The daily data flow within a week is periodic, and the red dotted box indicates that anomalies have occurred.

anomaly probability is greater than or equal to 0.5, we believe that an anomaly has occurred at this time. The following are main contributions of this paper:

- We are the first to model the fine-grained short-term time dependence and periodic long-term time dependence simultaneously in a unified framework, which has important practical significance for extracting the characteristics of time series.
- We propose an end-to-end hierarchical attention network framework consisting of seven day-level attention networks and a week-level attention network to model the fine-grained day-level short-term time dependence and periodic week-level long-term time dependence simultaneously, which can extract more abundant time features and thus obtain more accurate classification.
- We perform extensive experiments on a real-world medium and low voltage distribution system dataset and a public anomaly detection dataset to illustrate the superiority of our method over state-of-the-arts comparing with various univariate time series anomaly detection baselines.

The overall organization of the paper is as follows. Section 2 conducts research on related work. Section 3 introduces the definition of the problem studied in this article. Section 4 discusses the specific implementation of the proposed method, and Section 5 conducts experiments and result analysis to verify the superiority of our model. Section 6 concludes this paper.

2 Related Work

Univariate time series anomaly detection has attracted wide attention recently. There are a variety of conventional statistics based methods and machine learning algorithms for detecting anomalies. The latter usually includes supervised and unsupervised methods.

Statistical models. The statistical model includes some probabilistic methods based on distribution hypothesis, such as Markov inequality, Chebyshev inequality, 3-sigma, Z-score, boxplot, Chi-square test, Grubbs' test, and Extreme Studied Deviate test (ESD)[26]. These models presuppose that the data follow a particular probability distribution, and any data that deviate from this distribution are considered anomalies. Additionally, the autoregressive techniques ARIMA[13] and Kalman filtering[14] can suit the features of historical series to forecast the following timestamp. And the frequency domain properties are used by the wavelet analysis and Fourier transform[31] to find anomalies.

Supervised methods. Supervised anomaly detection methods rely significantly on accurate tagging. A common problem in supervised anomaly detection is the imbalance between anomalous classes and normal classes. To address this issue, some works have proposed the cost-sensitive learning and adaptive re-sampling methods based on traditional classification algorithms such as k-nearest neighbor, linear regression, Naive Bayes, random forests[32], decision trees, support vector machines[33], etc. With the popularity of some advanced deep learning models, some common deep learning networks such as FNN, CNN, and LSTM have been utilized for classification models. The Opprentice[34] trains the random forest classifier using features and labels after extracting anomalous characteristics from the original data using a variety of current deep learning detectors. The FNN[19] utilizes the raw data and labels to get a trained deep neural network. After that, the output layer applies a softmax function to get the probability of an anomaly. Informer[21] is an effective prediction scheme that has been put out to address the issue of long sequence time-series modeling, using ProbSparse self-attention and having the capacity to extract long-range dependence from the input time series. We utilize informer to extract time series features for the purpose of comparison with our model, and then we use the features to classify anomalies.

Unsupervised methods. The unsupervised anomaly detection methods concentrate on the normal mode and do not need to label the anomalies. The KNN[18] is a distance-based method, which calculates the average distance between each sample point and adjacent K samples in turn. This method can only find global anomalies by comparing distance and threshold and cannot detect local anomalies. The Principal Component Analysis (PCA)[35] converts the input into a low-dimensional space, reconstructs the original data using the low-dimensional feature, and then utilizes the reconstruction deviation as the anomalous score. The one-class SVM trains a hyperplane by mapping data to high-dimensional space using kernel function. During the test, if the data are within the hyperplane, it is considered as normal point, otherwise it is considered as anomaly. Isolation forest[17] is an unsupervised machine learning algorithm, which defines anomalies as points that are widely spaced out from dense populations and have a low density of occurrence. In

addition, the idea of other methods based on deep learning[36–41] is that if the normal samples are much larger than the anomalous one in the training set, the normal mode of time series will be learned by the reconstruction model or forcasting-based model during the training process. In the detection phase, the data deviating from the learned normal pattern, that is, the reconstruction error or forcasting error greater than the threshold, are defined as anomalies. Some typical reconstructed-based methods are based on AE, VAE, and GAN. For example, the AE[26] uses the neural network encoder to map the input sequence to a low-dimensional latent space and then uses a decoder to reconstruct the sequence. The Donut[27] takes fully connected network as the feed-forward network of VAE to reconstruct time series. To better model time dependence, the LSTM-VAE[28] uses the LSTM as encoder and decoder to learn complex nonlinear time dependence and reconstruct the input data. The MAD-GAN[29] is an unsupervised approach based on GAN network, which uses the Long Short-Term Memory-Recurrent Neural Networks (LSTM-RNN) as the discriminator and generator to extract the inner time dependency of the time series. LSTM-AD[24] utilizes LSTM encoding and decoding networks to learn time dependencies patterns, while the DeepAnt[25] utilizes CNN to achieve one-step prediction. Then the anomaly detector module takes the prediction error as anomaly score. The GCP_LSTM[42] uses LSTM to capture the nonlinear relationship of historical climate data and employs a 5-min long time window to focus on capturing short-term climate changes pattern to predict future climate change trends. The Interaction-enhanced and Time-aware Graph Convolution Network (ITGCN)[43] proposes a Graph Convolution Network (GCN) with improved time awareness and interaction capabilities and designs an aggregator to optionally embed the Point-of-Interest (POI) proximity relationship learned by GCN into each node feature, which can learn the relationship between all POIs and the dynamic time dependency within each POI at the same time. The Preference-based POI Category recommendation Model (PPCM)[44] uses local sensitive hashing to partition users while protecting privacy and then models time patterns using LSTM and attention to capture the preferences of each type of user and make POI recommendations.

In summary, previous work either used statistical methods to fit the distribution of data or used deep learning networks such as LSTM to learn short-term time dependencies within time series. However, there was a lack of analysis of long-term periodic dependencies, which may result in the loss of important periodic information and the inability to learn potential patterns of time series correctly, leading to the inability to detect anomalies in a timely and accurate manner. The solution to this problem has also created an advantage for our method compared to previous work. We consider the time dependencies within the sequence from the day and week levels, which can not only model short-term dependencies but also explore periodic patterns, thus obtaining richer time features. Specifically, in order to model the complex nonlinear time dependence and periodicity of univariate time series, we propose the hierarchical attention network, which adopts the day-level attention network and week-level attention network to capture the fine-grained short-term time dependence and periodic long-term time dependence simultaneously.

3 Problem Statement

In this work, we concentrate on the univariate time series anomaly detection task. Since anomalous data account for a non-negligible proportion of our dataset, we use the supervised method to model the time series. Specifically, our hierarchical attention network models time dependence, gets the day-level feature vector y_t and the week-level feature vector y'_t from the historical time series $x_{t-7w+1}, x_{t-7w+2}, \ldots, x_t$, and then uses a classifier to obtain the anomaly probability O_t of timestamp t. Our goal is to get the anomaly label a_t. If $O_t \geqslant 0.5$, we assume that $a_t = 1$, which means that an anomaly occurs at timestamp t. The definition of symbols in the text is shown in Table 1.

4 Model Architecture

4.1 Model overview

The medium and low voltage distribution data are a time series with complex nonlinear time-dependent and periodic characteristics. To model the fine-grained short-term time dependence and periodic long-term time dependence in the medium and low voltage distribution time series simultaneously, we propose a hierarchical attention network, which consists of seven day-level self-attention networks and one week-level self-attention network. As shown in Fig. 2, the overall framework consists of three modules. The day-level

Table 1 Notations.

Index	Definition
w	Number of timestamps in a fixed window.
x_t	Input time series value at timestamp t.
g_t	Vector after embedding at timestamp t.
y_t	Day-level attention network output at timestamp t.
$y_t^{'}$	Week-level attention network output at timestamp t.
l	Dimension of vector after embedding.
l_k	Dimension of the query Q and key K in self-attention network.
l_v	Dimension of the value V in self-attention network.
H	Number of heads in multi-head attention networks.
O_t	Anomaly probability at timestamp t.
G_t	Ground truth label at timestamp t.
a_t	Anomaly label at timestamp t.

self-attention networks use seven multi-head self-attention networks to model the time series data of seven days before the current timestamp t, respectively. Specifically, we take the data of the last day as input of the last self-attention network to get the day-level feature y_t of the timestamp t and employ the data of the first six days as input of the other six self-attention networks to get the day-level features of each day. Finally, the day-level features of these seven days will be fed into the week-level self-attention network to further extract the week-level feature y_t'. In the output layer, the day-level feature vector of last day learned by the day-level attention network and the week-level feature vector learned by week-level attention networkwill be connected and then sent to an MLP network to get the anomaly probability of timestamp t. In the training stage, the difference between the anomaly probability and anomaly label is measured using a binary cross entropy loss function, and back-propagation is used to train the model to recognize the characteristics of the anomalous data. When performing anomaly detection, if the anomaly probability is greater than or equal to 0.5 at timestamp t, we assume that an anomaly has occurred at this moment.

4.2 Day-level attention network

In order to model the fine-grained short-term time dependence in the medium and low voltage distribution univariate time series, we utilize seven multi-head self-attention networks to model the time series data of seven days before the current time t, respectively.

First, we use the embedding module to convert the input value x_t at every timestamp t to the input embeddings vectors of dimension d. In addition, we also include "positional encodings" to the input embeddings to utilize the order of input data. In our work, we use sine and cosine functions of various frequencies as the positional encodings which can encode the relative positions information. The positional encodings are a vector having the same dimension with the input embeddings, and we add the input embeddings and positional encodings to get the input of day-level multi-head self-attention module:

$$g_t = \mathrm{emb}(x_t) + \mathrm{pos\_enc}(x_t) \tag{1}$$

where $g_t \in \mathbb{R}^l$, the emb is a linear mapping, and the pos_enc represents the sine cosine encoding process.

Through the above embedding layer, we can embed the sequence $x_{t-nw+1}, x_{t-nw+2}, \ldots, x_{t-(n-1)w}$ of the n-th

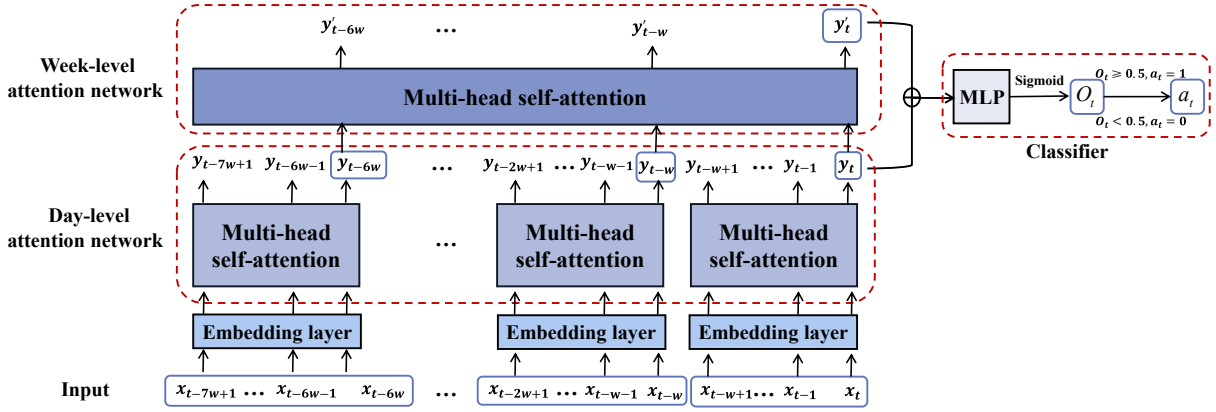


Fig. 2 Framework of the proposed HAN which mainly includes seven day-level self-attention networks, one week-level self-attention network, and a classifier.

day into the input token $g_{t-nw+1}, g_{t-nw+2}, \ldots, g_{t-(n-1)w}$ noted as $g_n \in \mathbb{R}^{w\times l}$. The input tokens are firstly mapped into queries Q, keys K, and values V:

$$Q = g_n \Phi_i^Q, K = g_n \Phi_i^K, V = g_n \Phi_i^V \tag{2}$$

where $Q \in \mathbb{R}^{w\times l}$, $K \in \mathbb{R}^{w\times l}$, and $V \in \mathbb{R}^{w\times l}$. The $\Phi_i^Q \in \mathbb{R}^{l\times l}$, $\Phi_i^K \in \mathbb{R}^{l\times l}$, and $\Phi_i^V \in \mathbb{R}^{l\times l}$ are linear projection matrixes. For the n-th day-level multi-head attention network with H heads, we linearly map the Q, K, and V for H times with various linear projections to l_k, l_k, and l_v dimensions. Next, we parallelize the attention function and apply the scaled dot product to determine each head's output, represented as head_h:

$$\text{head}_h = \text{Attention}(Q\Phi_h^Q, K\Phi_h^K, V\Phi_h^V) = \text{Softmax}\left(\frac{Q\Phi_h^Q K\Phi_h^{K\mathrm{T}}}{\sqrt{l_k}}\right)V\Phi_h^V \tag{3}$$

where $\Phi_h^Q \in \mathbb{R}^{l\times l_k}$, $\Phi_h^K \in \mathbb{R}^{l\times l_k}$, and $\Phi_h^V \in \mathbb{R}^{l\times l_v}$ are linear projection matrixes, and $l_k = l_v = l/H$.

Finally, the final outputs of day-level attention network are obtained by connecting the outputs of all heads and then projecting them once more:

$$\text{MultiHead}(Q,K,V) = \text{Cat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_H)\Phi^O \tag{4}$$

where Cat is the abbreviation for concatenate, H is the total number of heads, and $\Phi^O \in \mathbb{R}^{Hl_v\times l}$. And we note the outputs of the n-th day-level attention network as the day-level features $y_{t-nw+1}, y_{t-nw+2}, \ldots, y_{t-(n-1)w}$, abbreviated as y_n:

$$y_n = \text{MultiHead}(Q,K,V) \tag{5}$$

We take the last day-level feature vector $y_{t-6w}, y_{t-5w}, \ldots, y_t$ of each day-level window to form a matrix $y \in \mathbb{R}^{7\times l}$ as the input of the week-level attention network.

4.3 Week-level attention network

Considering the periodic long-term time dependence in the medium and low voltage distribution time series, we utilize a week-level attention network whose inputs are seven day-level features learned by the day-level attention network to further extract the week-level feature.

The inputs of the week-level attention network are the day-level feature vector matrix $y \in \mathbb{R}^{7\times l}$. We firstly project the input matrix y into week-level queries Q', keys K', and values V':

$$Q' = y\Theta_i^Q, K' = y\Theta_i^K, V' = y\Theta_i^V \tag{6}$$

where $Q' \in \mathbb{R}^{7\times l}$, $K' \in \mathbb{R}^{7\times l}$, and $V' \in \mathbb{R}^{7\times l}$. The $\Theta_i^Q \in \mathbb{R}^{l\times l}$, $\Theta_i^K \in \mathbb{R}^{l\times l}$, and $\Theta_i^V \in \mathbb{R}^{l\times l}$ are learnable projection matrixs. Then, using various linear projection matrices, we linearly project the Q', K', and V' to H distinct subspaces. For the h-th head, we apply the scaled dot-product to obtain the output head'_h:

$$\text{head}'_h = \text{Attention}(Q'\Theta_h^Q, K'\Theta_h^K, V'\Theta_h^V) = \text{Softmax}\left(\frac{Q'\Theta_h^Q K'\Theta_h^{K\mathrm{T}}}{\sqrt{l_k}}\right)V'\Theta_h^V \tag{7}$$

where $\Theta_h^Q \in \mathbb{R}^{l\times l_k}$, $\Theta_h^K \in \mathbb{R}^{l\times l_k}$, and $\Theta_h^V \in \mathbb{R}^{l\times l_v}$ are linear projection matixs, and $l_k = l_v = l/H$. Connect the output of all heads, we can obtain the final outputs of the week-level attention network:

$$\text{MultiHead}(Q',K',V') = \text{Cat}\left(\text{head}'_1, \text{head}'_2, \ldots, \text{head}'_H\right)\Theta^O \tag{8}$$

where H represents the number of heads, and $\Theta^O \in \mathbb{R}^{Hl_v\times l}$. And we note the outputs of week-level attention network as week-level features y'_{t-6w}, $y'_{t-5w}, \ldots, y'_t$, abbreviated as y':

$$y' = \text{MultiHead}(Q',K',V') \tag{9}$$

The specific implementation and training process of the above HAN is shown in Algorithm 1.

4.4 Output layer

The day-level feature vectors and the week-level feature vector learned by the day-level attention networks and week-level attention network, respectively, are connected and then sent to an MLP network with sigmoid function to get the anomaly probability O_t at timestamp t:

$$O_t = \text{MLP}\left(\left[y_t, y'_t\right]\right) \tag{10}$$

4.5 Objective function and training

Since the anomaly samples in our datasets account up a substantial percentage of the total data, the univariate time series anomaly detection issue is seen as a classification problem, and the supervised learning approach is used.

When training, we use the binary cross entropy loss as the objective function to decrease the discrepancy between the anomaly probability O_t and ground truth label G_t:

Algorithm 1 Training of hierarchical attention networks

Input: A historical sequence $x_{t-7w+1}, x_{t-7w+2}, \ldots, x_t$ of one week before the current moment, classification labels G_t with anomaly 1 and normal 0, and N epochs.

Output: Trained HAN.

for epoch = 1 ∶ N **do**
  for each day n **do**
    for each moment t in day n **do**
      $g_t \leftarrow \text{emb}(x_t) + \text{pos\_enc}(x_t)$,
    end for
    $g_n \leftarrow g_{t-nw+1}, g_{t-nw+2}, \ldots, g_{t-(n-1)w}$
    $Q \leftarrow g_n\ \Phi_i^Q$, $K \leftarrow g_n\ \Phi_i^k$, and V $\leftarrow g_n\ \Phi_i^V$
    for each head h in day-level attention network **do**
      $\text{head}_h \leftarrow \text{Attention}(Q\Phi_h^Q, K\Phi_h^K, V\Phi_h^V)$
    end for
    Day-level features:
    $y_n \leftarrow \text{Cat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_H)\,\Phi^O$
  end for
  Draw the last day-level feature vector $y_{t-6w}, y_{t-5w}, \ldots, y_t$ of each day to form a matrix $y \in \mathbb{R}^{7\times l}$ as the input of the week-level attention network
  $Q' \leftarrow y\Theta_i^Q$, $K' \leftarrow y\Theta_i^K$, and $V' \leftarrow y\Theta_i^V$
  for each head h in week-level attention network **do**
    $\text{head}'_h \leftarrow \text{Attention}(Q'\Theta_h^Q, K'\Theta_h^K, V'\Theta_h^V)$
  end for
  Week-level features:
  $y' \leftarrow \text{Cat}\left(\text{head}'_1, \text{head}'_2, \ldots, \text{head}'_H\right)\Theta^O$
  Get the anomaly probability O_t at timestamp t: $O_t \leftarrow \text{MLP}\left(\left[y_t, y'_t\right]\right)$
  $L_{\text{BCE}} \leftarrow -G_t\log(O_t) - (1-G_t)\log(1-O_t)$
  Update HAN network parameters with L_{BCE}
end for

$$L_{\text{BCE}} = -G_t\log(O_t) - (1-G_t)\log(1-O_t) \tag{11}$$

And the network parameters are updated by gradient descent algorithm.

4.6 Anomaly detection

During anomaly detection, we extract the time-dependent and periodic characteristics of the input time series through the hierarchical attention network and obtain the anomaly probability through the output layer. If the anomaly probability O_t is greater than or equal to 0.5, we set the anomaly label a_t to 1, which means an anomaly occurs at the timestamp t. Otherwise, we assume the timestamp t is normal. Specifically,

$$a_t = \begin{cases} 1, & O_t \geqslant 0.5; \\ 0, & O_t < 0.5 \end{cases} \tag{12}$$

5 Experiment

5.1 Datasets description

We evaluate our model over two anomaly detection datasets: the private ElecPower and the public Dodgers.

- **ElecPower.** The ElecPower is a dataset composed of medium and low voltage distribution univariate time series collected from the State Grid Beijing Electric Power Company. We continuously collect the medium and low voltage distribution data at 5-min intervals between December 27, 2020 and February 6, 2021. The whole dataset contains 12 098 data, including 4578 marked anomaly data, accounting for 37.84% of the whole dataset. In details, normal data are marked as 0, and anomalous data are marked as 1. And the attack time varies from 5 to 200 min.
- **Dodgers[45].** The Dogers dataset collects the vehicle count of the 101 freeway on-ramp near the Dodge Stadium, where the Los Angeles Dodgers baseball team's home game is located. Under normal circumstances, the traffic flow at the intersection of expressway presents a normal peak in the morning and evening and when there are competitions. When a traffic accident occurs, the traffic flow changes abnormally. Each data point in the dataset represents the traffic flow within 5 min of the adjacent history, and the entire dataset contains 133 attacks in total.

We split the training set, verification set, and test set depending on the ratio of 7 ∶ 1 ∶ 2. Additionally, we have standardised the datasets. Especially, for the training data:

$$\tilde{x} = \frac{x-\mu}{\sigma} \tag{13}$$

where x is the original training set data. $\tilde{x}$ is the normalized training set data. μ and σ are the mean and variance of the training set, respectively.

5.2 Experimental setup

Baselines. We compare the performance of the proposed method with ten popular time series anomaly detection methods, including:

- **ARIMA[13]:** The ARIMA is an autoregressive integrated moving average model, which predicts the future based on the historical lag value and takes the

prediction error as anomaly score.

● **PCA[35]:** The PCA can determine the hyperplane of the least squares error, that is, a low-dimensional subspace having the lowest reconstruction error after projection. The points with a large reconstruction error are regarded as anomalies.

● **KNN[18]:** The KNN is a distance-based unsupervised method. It calculates the average distance of the nearest K samples for each sample point, whose distance exceeding the threshold is considered as an anomaly.

● **Isolation forest[17]:** The isolation forest integrates a group of decision trees and introduces a random attribute selection method for segmentation. The depth of branches containing anomalies is small, because the anomalies are located in sparse areas. And the anomaly score is calculated using the leaf-to-root distance.

● **FNN[19]:** The FNN is a supervised model which can get hidden features of time series from raw datasets and then obtain the anomaly probability through a softmax function.

● **LSTM-AD[24]:** The LSTM networks' recurrent hidden layers are able to learn the high level temporal features. And the prediction error is regarded as anomaly score.

● **Informer[21]:** The informer is designed for long time-series forecasting, which proposes the ProbSparse self-attention and uses a distilling operation to solve the problems of quadratic time and space complexity. We use the infomer to model time series and predict anomaly probability.

● **AE[26]:** The AE maps the input data into low-dimensional latent variables and uses them to reconstruct the input. The points deviating from the normal mode, that is, the points with high reconstruction deviation are taken as anomalies.

● **Donut[27]:** The Donut takes fully connected neural network as the feedforward network of VAE to reconstruct time series and proposes the modied Evidence Lower Bound (ELBO). In addition, Donut applies the reconstruction probability to measure anomaly scores.

● **LSTM-VAE[28]:** The LSTM-VAE takes LSTM as the feedforward network of VAE to learn the features of input and treats data difficult to recover as anomalies.

Metrics. We take the Precision (P), Recall (R), and $F1$ score as the evaluation metrics.

$$P = \frac{N_{\text{tp}}}{N_{\text{tp}} + N_{\text{fp}}} \tag{14}$$

$$R = \frac{N_{\text{tp}}}{N_{\text{tp}} + N_{\text{fn}}} \tag{15}$$

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{16}$$

where N_{tp} stands for the number of accurately recognized anomalies, N_{fp} stands for the amount of normal samples incorrectly identified as anomalies, and N_{fn} stands for the quantity of anomalies filtered incorrectly.

Training settings. We perform all experiments on an NVIDIA GeForce RTX 3090 GPU using Pytorch version 1.10.0 with CUDA 11.1. The number of heads of all multi-head self-attention networks is 8. The hidden layer dimension l of the attention network is 128 on both datasets. On the ElecPower dataset, the size of the input window w of the day-level attention network is 288, which is the amount of data in a day, with 5 min intervals. The model is trained using the Adam optimizer for up to 50 iterations with a learning rate of 0.0001 and an early stopping policy with patience of 5.

5.3 Performance and analysis

In Table 2, we evaluate the anomaly detection performance of our HAN on three metrics: precision, recall, and $F1$ score, and compare the HAN with other ten popular time series anomaly detection methods on the ElecPower and Dodgers datasets. It is obvious that our HAN performs the best compared with other ten baselines. Moreover, we further observe the following phenomenons from Table 2.

● Our HAN has achieved the best result in recall and $F1$. On the ElecPower dataset, compared with the suboptimal method, we have achieved 4.00% and 8.12% improvement on recall and $F1$, respectively. On the Dodgers dataset, our HAN's precision, recall, and $F1$ are 5.21%, 1.46%, and 10.32% higher than the second best methods, respectively.

● The performance of deep learning methods including RNN or attention mechanism is much better than that of statistical methods such as ARIMA and traditional machine learning methods such as PCA, KNN, and isolation forest, which shows the importance of modeling time dependence. Especially, the $F1$ of our HAN is 105.10% and 122.58% higher than that of ARIMA, and 64.11% and 328.52% higher than that of

Table 2 Experimental results of different methods on ElecPower and Dodgers using precision, recall, and *F*1 as metrics. The best performances are in bold and the second are underlined.

(%)

Method	ElecPower			Dodgers		
	P	*R*	*F*1	*P*	*R*	*F*1
ARIMA	24.82	45.00	31.99	22.49	65.50	33.48
PCA	41.58	38.50	39.98	11.13	39.74	17.39
KNN	28.09	2.22	4.12	23.81	0.79	1.54
Isolation forest	50.04	13.17	20.85	45.19	48.16	46.63
FNN	**70.57**	49.05	57.87	68.15	<u>66.96</u>	<u>66.75</u>
LSTM-AD	53.29	48.95	51.03	63.17	51.24	56.58
Informer	46.43	<u>87.54</u>	<u>60.68</u>	<u>78.43</u>	57.84	66.58
AE	51.31	40.91	45.52	52.30	38.91	44.62
Donut	<u>60.18</u>	18.18	27.92	34.64	59.27	43.72
LSTM-VAE	59.66	10.22	17.46	53.43	29.92	38.36
HAN (ours)	51.28	**91.04**	**65.61**	**82.52**	**67.94**	**74.52**

PCA on the two datasets significantly.

● The models based on attention mechanism, such as informer and our HAN, can perform more effectively than models based on RNN, such as LSTM-AD, which indicates that attention mechanism has better long-term time-dependent modeling capability than RNN. This is because the attention network can pay equal attention to any other position with a short network path, while LSTM pays less attention to the information at a distance. In particular, the HAN's recall and *F*1 are 86.00% and 28.57% higher than LSTM-AD on the ElecPower dataset and 32.59% and 31.71% higher on the Dodgers dataset, respectively.

● Our hierarchical attention network considering high-level weekly time dependence can achieve better performance than informer that only considers day-level time dependence, which proves the necessity of modeling periodic long-term time dependence. Especially, the precision, recall, and *F*1 of our HAN are 10.45%, 4.00%, and 8.12% higher than the informer on the ElecPower dataset and 5.21%, 17.46%, and 11.93% higher on the Dodgers dataset, respectively.

● The models based on reconstruction, such as AE, Donut, and LSTM-VAE, do not achieve good performance. This is because the number of normal data in training set is not far greater than that of anomalies, so the models can not learn the normal mode of the data and thus cannot accurately detect anomalies according to the reconstruction error during the anomaly detection phase. Especially, the *F*1 of our HAN is 134.99% and 70.45% higher than Donut and 275.77% and 94.26% higher than LSTM-VAE on the two datasets significantly.

5.4 Ablation study

To study the efficiency of each part of our HAN, we gradually drop or replace each component to observe the change of model's performance, and the results are shown in Table 3.

● $\text{HAN}_{w/o\ \text{Attention Network}}$. To study the long-term time-dependent modeling effect of attention network, we use the LSTM network to replace the attention network. As shown in Table 3, the performance of this variant declines compared with that of HAN. The precision, recall, and *F*1 of $\text{HAN}_{w/o\ \text{Attention Network}}$

Table 3 Experimental results of ablation studies on ElecPower and Dodgers using precision, recall, and *F*1 as metrics. The HAN without removal of any components achieves the best performance.

(%)

Method	ElecPower			Dodgers		
	P	*R*	*F*1	*P*	*R*	*F*1
HAN (ours)	**51.28**	**91.04**	**65.61**	**82.52**	**67.94**	**74.52**
w/o Attention Network	44.82	59.66	51.19	68.67	56.68	62.10
w/o Hierarchical Mechanism	49.48	82.68	61.91	64.17	65.21	64.69

decrease by 12.60%, 34.47%, and 21.98% on the ElecPower dataset and 16.78%, 16.57%, and 16.67% on the Dodgers dataset, respectively, which proves that the attention network has superior time dependence modeling ability and also proves that it is important to consider the time dependence for univariate time series anomaly detection.

• $\text{HAN}_{w/o\ \text{Hierarchical Mechanism}}$. To study the impact of hierarchical scheme, we do not consider the time dependence at the week-level but only the time dependence at the day-level. Specifically, we use the one-day long history sequence and a day-level multi-head self-attention network to get the anomaly label at current time t. From Table 3, we can observe that the precision, recall, and $F1$ of $\text{HAN}_{w/o\ \text{Hierarchical Mechanism}}$ decrease by 3.51%, 9.18%, and 5.64% on the ElecPower dataset and 22.24%, 4.02%, and 13.19% on the Dodgers dataset, respectively, which demonstrates the importance of modeling periodic long-term time dependence and the effectiveness of our proposed hierarchical mechanism.

5.5 Parameter analysis

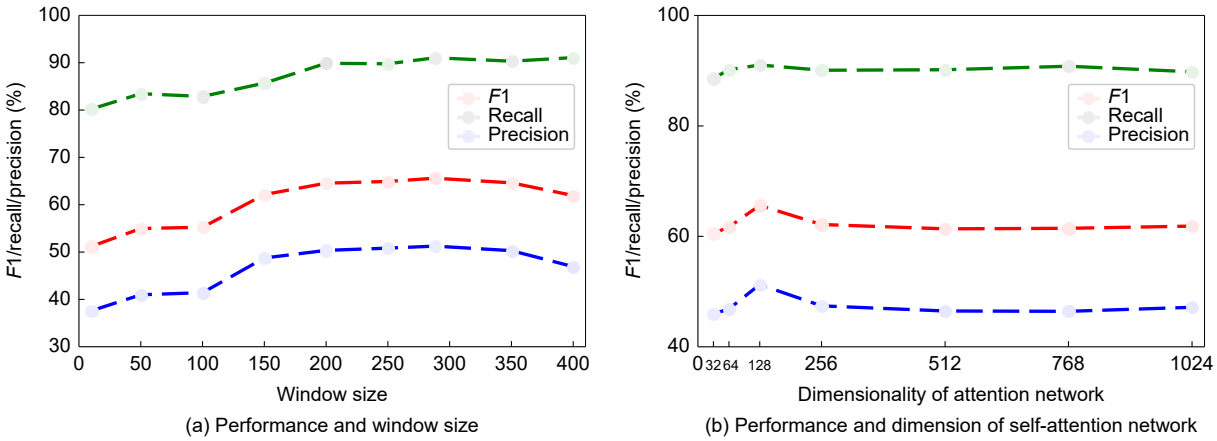
To study the effect of various hyper parameters on the anomaly detection performance of our model, we carry out experiments under different parameter values and analyze the results.

• **Window size.** The long-term time dependent modeling ability of hierarchical attention network is related to the length of the historical window of the day-level attention networks. If the window length is too small, the attention network cannot learn the periodic rule of time series. If the window length is too large, the network will be too complex and reduce efficiency. From Fig. 3a, we can see that with the increase of window length, the precision, recall, and $F1$ continue to increase and reach the best when the window length is 288, then the $F1$ decreases. The HAN is optimal when the window length is 288 because 288 is the data volume of one day. Using seven day-level attention networks and one week-level attention network, the model can learn the periodicity of week-level long-term time dependence.

• **Dimensionality l of self-attention networks.** The dimensions l_k and l_v of Q, K, and V in the hierarchical attention network will affect the long-term time dependent modeling ability of the attention network. Since $l_k = l_v = l/H$, we focus on the l. If the dimension is too small, it tends to lead to insufficient representation ability of the network, so that it can not learn the complex nonlinear time dependence. If the dimension is too big, it will lead to too complex and inefficient network. Figure 3b shows the anomaly detection result of our HAN corresponding to the different dimensions of self-attention networks. As seen in Fig. 3b, the precision, recall, and $F1$ increase with the increase of dimension and reach peak at 128. After that, the three metrics decrease as the dimension increases and then tend to be stable.

6 Conclusion

To support the perception and defense of the operation risk of the medium and low voltage distribution system, we proposed an end-to-end hierarchical attention network anomaly detection framework, which can simultaneously model the fine-grained short-term time dependence and periodic long-term time dependence of univariate time series. We introduced seven day-level multi-head attention networks to model



(a) Performance and window size

(b) Performance and dimension of self-attention network

Fig. 3 Hyper-parameter analysis on ElecPower dataset.

the day-level time dependence of the first seven days in the historical sequence and get seven day-level time feature vectors, and then sent them into a week-level attention network to model the periodic week-level time dependence and get the week-level time feature vector. Finally, the day-level time feature vector and week-level time feature vector of the current moment went through a classifier to obtain the anomaly probability of this moment, according to which we can further judge whether the moment is anomalous or not. Extensive experiments on a real-world medium and low voltage distribution electric power dataset and a publicly traffic flow anomaly detection dataset demonstrated that our method outperforms other univariate time series anomaly detection approaches. In future work, we may further introduce month-level and season-level periodicity to model longer term time dependent patterns, and we hope to study some effective strategies to improve the efficiency of the proposed network.

Acknowledgment

This work was supported by the Science and Technology Project named "Research on Risk Perception and Defense System for Medium and Low Voltage Distribution System Operation Based on Data Mining" of State Grid Beijing Electric Power Company (No. 520202220002).

References

[1] R. A. Leon, V. Vittal, and G. Manimaran, Application of sensor network for secure electric energy infrastructure, *IEEE Trans. Power Deliv.*, vol. 22, no. 2, pp. 1021–1028, 2007.

[2] N. Laptev, S. Amizadeh, and I. Flint, Generic and scalable framework for automated time-series anomaly detection, in *Proc. 21th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Sydney, Australia, 2015, pp. 1939–1947.

[3] Z. Li, X. Xu, T. Hang, H. Xiang, Y. Cui, L. Qi, and X. Zhou, A knowledge-driven anomaly detection framework for social production system, *IEEE Trans. Comput. Soc. Syst*., doi: 10.1109/TCSS.2022.3217790.

[4] Y. Jia, B. Liu, W. Dou, X. Xu, X. Zhou, L. Qi, and Z. Yan, CroApp: A CNN-based resource optimization approach in edge computing environment, *IEEE Trans. Ind. Inform.*, vol. 18, no. 9, pp. 6300–6307, 2022.

[5] S. Chen, Y. Tao, D. Yu, F. Li, and B. Gong, Distributed learning dynamics of multi-armed bandits for edge intelligence, *J. Syst. Archit.*, vol. 114, p. 101919, 2021.

[6] S. Chen, Y. Tao, D. Yu, F. Li, B. Gong, and X. Cheng, Privacy-preserving collaborative learning for multiarmed bandits in IoT, *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3276–3286, 2021.

[7] H. Yu, Z. Chen, X. Zhang, X. Chen, F. Zhuang, H. Xiong, and X. Cheng, FedHAR: Semi-supervised online learning for personalized federated human activity recognition, *IEEE Trans. Mob. Comput.*, vol. 22, no. 6, pp. 3318–3332, 2023.

[8] V. Zeufack, D. Kim, D. Seo, and A. Lee, An unsupervised anomaly detection framework for detecting anomalies in real time through network system's log files analysis, *High Confid. Comput.*, vol. 1, no. 2, p. 100030, 2021.

[9] J. Chen, H. Huang, and H. Chen, Informer: Irregular traffic detection for containerized microservices RPC in the real world, *High Confid. Comput.*, vol. 2, no. 2, p. 100050, 2022.

[10] S. Wu, S. Shen, X. Xu, Y. Chen, X. Zhou, D. Liu, X. Xue, and L. Qi, Popularity-aware and diverse web APIs recommendation based on correlation graph, *IEEE Trans. Comput. Soc. Syst.*, vol. 10, no. 2, pp. 771–782, 2023.

[11] X. Zhou, X. Xu, W. Liang, Z. Zeng, and Z. Yan, Deep-learning-enhanced multitarget detection for end-edge-cloud surveillance in smart IoT, *IEEE Internet Things J.*, vol. 8, no. 16, pp. 12588–12596, 2021.

[12] L. Qi, Y. Yang, X. Zhou, W. Rafique, and J. Ma, Fast anomaly identification based on multiaspect data streams for intelligent intrusion detection toward secure industry 4.0, *IEEE Trans. Ind. Inform.*, vol. 18, no. 9, pp. 6503–6511, 2022.

[13] H. Z. Moayedi and M. A. Masnadi-Shirazi, ARIMA model for network traffic prediction and anomaly detection, in *Proc. 2008 Int. Symp. Information Technology*, Kuala Lumpur, Malaysia, 2008, pp. 1–6.

[14] F. Knorn and D. J. Leith, Adaptive Kalman filtering for anomaly detection in software appliances, in *Proc. IEEE INFOCOM Workshops*, Phoenix, AZ, USA, 2008, pp. 1–6.

[15] M. Amer, M. Goldstein, and S. Abdennadher, Enhancing one-class support vector machines for unsupervised anomaly detection, in *Proc. ACM SIGKDD Workshop on Outlier Detection and Description*, Chicago, IL, USA, 2013, pp. 8–15.

[16] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning, *Pattern Recognit.*, vol. 58, pp. 121–134, 2016.

[17] Z. Ding and M. Fei, An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window, *IFAC Proc. Vol.*, vol. 46, no. 20, pp. 12–17, 2013.

[18] F. Angiulli and C. Pizzuti, Fast outlier detection in high dimensional spaces, in *Proc. 6th European Conf. Principles of Data Mining and Knowledge Discovery*, 2002, Helsinki, Finland, 2002, pp. 15–27.

[19] R. Zhang, S. Dong, X. Nie, and S. Xiao, Feedforward neural network for time series anomaly detection, arXiv preprint arXiv: 1812.08389, 2018.

[20] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, Informer: Beyond efficient transformer for long sequence time-series forecasting, *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 12, pp. 11106–11115, 2021.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, in *Proc. 31st Int. Conf. on Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[23] Z. Chen, D. Chen, X. Zhang, Z. Yuan, and X. Cheng, Learning graph structures with transformer for multivariate time-series anomaly detection in IoT, *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9179–9189, 2022.

[24] P. Malhotra, L. Vig, G. M. Shroff, and P. Agarwal, Long short term memory networks for anomaly detection in time series, in *Proc. 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges, Belgium, 2015, pp. 89–94.

[25] M. Munir, S. A. Siddiqui, A. Dengel, and S. Ahmed, DeepAnT: A deep learning approach for unsupervised anomaly detection in time series, *IEEE Access*, vol. 7, pp. 1991–2005, 2018.

[26] C. C. Aggarwal, Time series and multidimensional streaming outlier detection, in *Outlier Analysis*, C. C. Aggarwal, ed. Cham, Switzerland: Springer, 2017, pp. 273–310.

[27] H. Xu, W. Chen, N. Zhao, Z. Li, J. Bu, Z. Li, Y. Liu, Y. Zhao, D. Pei, Y. Feng, et al., Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications, in *Proc. 2018 World Wide Web Conf.*, Lyon, France, 2018, pp. 187–196.

[28] D. Park, Y. Hoshi, and C. C. Kemp, A multimodal anomaly detector for robot-assisted feeding using an LSTM-based variational autoencoder, *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 1544–1551, 2018.

[29] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, and S. K. Ng, MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks, in *Proc. 28th Int. Conf. Artificial Neural Networks*, Munich, Germany, 2019, pp. 703–716.

[30] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, USAD: Unsupervised anomaly detection on multivariate time series, in *Proc. 26th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*, Virtual Event, CA, USA, 2020, pp. 3395–3404.

[31] Y. Dwivedi and S. S. Rao, A test for second-order stationarity of a time series based on the discrete Fourier transform, *J. Time Ser. Anal.*, vol. 32, no. 1, pp. 68–91, 2011.

[32] L. Breiman, Random forests, *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[33] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, Support vector machines, *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, 1998.

[34] D. Liu, Y. Zhao, H. Xu, Y. Sun, D. Pei, J. Luo, X. Jing, and M. Feng, Opprentice: Towards practical and automatic anomaly detection through machine learning, in *Proc. 2015 Internet Measurement Conference*, Tokyo, Japan, 2015, pp. 211–224.

[35] W. Qiu, Y. Wu, G. Wang, S. X. Yang, J. Bai, and J. Li, A novel unsupervised anomaly detection based on robust principal component classifier, in *Proc. Defense and Security Symposium*, Orlando (Kissimmee), FL, USA, 2006, pp. 260–268.

[36] S. Chen, Y. Wang, D. Yu, J. Ren, C. Xu, and Y. Zheng, Privacy-enhanced decentralized federated learning at dynamic edge, *IEEE Trans. Comput.*, vol. 72, no. 8, pp. 2165–2180, 2023.

[37] Y. Yuan, J. Yu, X. Cheng, Z. Zou, D. Yu, and Z. Cai, Decentralized parallel SGD based on weight-balancing for intelligent IoV, *IEEE Trans. Intell. Transp. Syst*., doi: 10.1109/TITS.2022.3216709.

[38] Z. Xie, Y. Huang, D. Yu, R. M. Parizi, Y. Zheng, and J. Pang, FedEE: A federated graph learning solution for extended enterprise collaboration, *IEEE Trans. Ind. Inform.*, vol. 19, no. 7, pp. 8061–8071, 2023.

[39] S. Chen, D. Yu, Y. Zou, J. Yu, and X. Cheng, Decentralized wireless federated learning with differential privacy, *IEEE Trans. Ind. Inform.*, vol. 18, no. 9, pp. 6273–6282, 2022.

[40] S. Chen, D. Yu, F. Li, Z. Zou, W. Liang, and X. Cheng, PPAR: A privacy-preserving adaptive ranking algorithm for multi-armed-bandit crowdsourcing, in *Proc. 2022 IEEE/ACM 30th Int. Symp. Quality of Service* (*IWQoS*), Oslo, Norway, 2022, pp. 1–10.

[41] D. Yu, Z. Zou, S. Chen, Y. Tao, B. Tian, W. Lv, and X. Cheng, Decentralized parallel SGD with privacy preservation in vehicular networks, *IEEE Trans. Veh. Technol.*, vol. 70, no. 6, pp. 5211–5220, 2021.

[42] Y. Liu, D. Li, S. Wan, F. Wang, W. Dou, X. Xu, S. Li, R. Ma, and L. Qi, A long short-term memory-based model for greenhouse climate prediction, *Int. J. Intell. Syst.*, vol. 37, no. 1, pp. 135–151, 2022.

[43] Y. Liu, H. Wu, K. Rezaee, M. R. Khosravi, O. I. Khalaf, A. A. Khan, D. Ramesh, and L. Qi, Interaction-enhanced and time-aware graph convolutional network for successive point-of-interest recommendation in traveling enterprises, *IEEE Trans. Ind. Inform.*, vol. 19, no. 1, pp. 635–643, 2023.

[44] L. Qi, Y. Liu, Y. Zhang, X. Xu, M. Bilal, and H. Song, Privacy-aware point-of-interest category recommendation in Internet of Things, *IEEE Internet Things J.*, vol. 9, no. 21, pp. 21398–21408, 2022.

[45] A. Ihler, J. Hutchins, and P. Smyth, Adaptive event detection with time-varying Poisson processes, in *Proc. 12th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Philadelphia, PA, USA, 2006, pp. 207–216.

Zexi Chen received the bachelor degree in electrical engineering and automation from Huazhong University of Science and Technology, China in 2013, the master degree in electrical engineering from University of Southern California, Los Angeles, CA, USA in 2015, and the PhD degree in renewable energy from North China Electric Power University, China in 2022. He is currently working as a senior engineer at State Grid Beijing Electric Power Company, China. His research areas include smart grid, integrated energy technologies, renewable energy, and energy storage.

Dongqiang Jia received the bachelor degree in electrical engineering from China Agricultural University, China in 2009 and the PhD degree from University of Chinese Academy of Sciences, China in 2015. He is currently working as a senior engineer at State Grid Beijing Electric Power Company, China. His research interests include power quality and integrated energy technologies.

Yushu Sun received the BS and MS degrees from North China Electric Power University, Beijing, China in 2010 and 2013, respectively, and the PhD degree from University of Chinese Academy of Sciences, Beijing, China in 2022. He is currently working as an associate professor at Institute of Electrical Engineering, Chinese Academy of Sciences, China. His research interests include renewable energy power generation, energy storage, and microgrid.

Lin Yang received the bachelor, master, and PhD degrees in electrical engineering from Harbin Institute of Technology, China in 2002, 2004, and 2008, respectively. He is currently working as an assistant chief engineer at State Grid Beijing Electric Power Company, China. His research areas include power system dispatching and operation technology management, intelligent electrical equipment, and business digital transformation. He is also a senior expert in the field of distribution network automation and communication.

Wenjie Jin received the bachelor degree in electric engineering from Huazhong University of Science and Technology, China in 2007 and the master degree in electric engineering from Xi'an Jiaotong University, China in 2010. He is currently working as a senior engineer at State Grid Beijing Electric Power Company, China. He is mainly engaged in distribution network operation, maintenance management, and power supply guarantee.

Ruoxi Liu received the bachelor, master, and PhD degrees in electrical power system and automation from North China Electric Power University, China in 2005, 2008, and 2012, respectively. He is currently working as a senior engineer at State Grid Beijing Electric Power Company, China. His research areas include power system risk assessment, distribution network fault analysis, and distribution network automation.