Computing and Informatics, Vol. 45, 2026, 82–104, doi: 10.31577/cai_2026_1_82

MULTI-SCALE MULTI-LOAD FEDERATED FORECASTING METHOD WITH MODE DECOMPOSITION

Chao Peng, Zikai Chen, Zhipeng Chen, Yong Zhang

School of Information and Control Engineering
China University of Mining and Technology
XuZhou, China
e-mail: yongzh401@cumt.edu.cn

Abstract. Accurate load forecasting is the premise of efficient and stable operation of integrated energy systems. For multiple integrated energy systems that have insufficient energy consumption data but similar energy consumption behavior, federated learning can establish a higher accuracy multi-load forecasting model for each system without disclosing data privacy. However, the existing federated learning methods cannot fully utilize common and individual characteristics in the energy consumption data of different nodes (that is, integrated energy systems), which obviously affects their prediction accuracy. In view of this, we propose a multi-time scale multi-load federated forecasting method based on mode decomposition (MD-MMFF). Firstly, a multivariate empirical mode decomposition method is introduced to decompose the energy consumption data of each node into two types, i.e., regular components and irregular components. Each node uses the local LSTM model to learn the regular components and predict their outputs. For the irregular components, a multi-load federated forecasting training mechanism based on knowledge distillation is proposed, and the corresponding multi-load forecasting model is jointly established for each node. Then, the predicted values of regular components and irregular components are integrated to obtain the final multivariate load forecasting results. Experimental results show that compared with the existing multi-load forecasting algorithms, the proposed MD-MMFF method can obtain higher accuracy multi-source load forecasting results.

Keywords: Integrated energy system, multi-load, federated learning, LSTM, multi-time scales, decomposition

1 INTRODUCTION

Social development and scientific and technological progress consume a large amount of fossil energy, resulting in an increasingly serious energy shortage. In order to solve these problems, integrated energy system (IES) technology is gradually showing its advantages. IES integrates various forms of energy supply, conversion and storage equipment, and can realize the coupling of different types of energy in the links of source, grid, load and storage [1]. It is an ideal way to reduce carbon emissions of energy systems and improve energy utilization. The premise of the reasonable operation of IES is accurate load forecasting [2]. Only when the accuracy of load forecasting reaches the standard can the output of each energy equipment be reasonably planned.

Machine learning methods such as long short-term memory (LSTM) [3] have been widely used in load forecasting due to their excellent learning ability. However, due to the coupling characteristics of heterogeneous energy flows in IES systems, multi-load forecasting is much more difficult than single-load forecasting [4]. Although scholars have proposed a variety of effective machine learning methods for multi-load forecasting [5, 6, 7, 8], most of them require sufficient load data. Obviously, for integrated energy systems with insufficient energy consumption data, a higher accuracy multi-load forecasting model can be established with the help of rich data from similar systems. However, in the face of a series of problems such as large data storage cost [9], data theft [10], tampering [11] and false data injection [12], many systems are unwilling to share their own data for the consideration of data security and other factors.

Faced with the problem of data islands, Google first proposed federated learning (FL) [13]. This is a way to build a learning model with the sharing mechanism while protecting the privacy of participants' data [14]. Federated learning can share training data without aggregating source data, which effectively protects user data privacy in the process of machine learning [15]. While FL has been widely used in medical care [16], communication [17], natural language processing [18] and other fields, it also provides an important way to deal with the problem of multi-load forecasting under privacy protection. Venkataramanan et al. [19] used the distributed characteristics of federated learning to co-train multiple energy prediction models and obtained obvious results. A model that combines federated learning and LSTM has been developed by Savi and Olivadese [20] to predict the electricity demand of houses. Chen et al. [21] combined the attention mechanism and the LSTM network. In order to improve the prediction accuracy of multi-load. Aiming at the problem of multi-load forecasting, Zhang et al. [22] combined the attention mechanism and LSTM network to design a multi-task hierarchical attention model, which was used to extract the global model of features and effectively improved the generalization ability of the model. These methods can effectively solve the problem of multi-load prediction, but the requirements for load data are more strict, such as the same scale. In many practical situations, due to different data collection devices, the load collection frequencies of

different participants or nodes are usually quite different, and the multi-load data obtained by them will have inconsistent time scales. It can be seen that how to use differential data to achieve accurate load prediction is a very challenging problem.

At the same time, due to the influence of exogenous attributes such as meteorological variations and holidays, the energy consumption data owned by each node often has certain common characteristics. Affected by the difference of energy consumption behavior of different users, the load data of each node also has its own personalized characteristics. This means that it is necessary to design a targeted federated learning mechanism to strengthen the learning of the common features between nodes while retaining the personalized characteristics of nodes. At present, there have been much work in personalized federated prediction. Zhang et al. [23] proposed a hierarchical federated learning method, which effectively improves the prediction efficiency of photovoltaic power generation by introducing personalized technology and semi-asynchronous aggregation strategy. Wang et al. [24] proposed a personalized federated learning prediction method, which effectively solves the problem of global model overfitting while improving the accuracy of residential load prediction.However, existing federated learning-based multi-load forecasting methods exhibit two key limitations. First, they fail to distinguish between common and personalized characteristics when analyzing the impact of energy consumption data on model performance. Second, these methods lack effective feature disentanglement mechanisms, resulting in inadequate separation and extraction of discriminative patterns from dataset.

In view of this, this paper studies a multi-time scale multi-load federated forecasting method based on mode decomposition, namely MD-MMFF. Firstly, each node locally uses a multivariate empirical mode decomposition method to decompose its own data into two categories: regular components and irregular components. To learn the regular components effectively, a local LSTM model is established on each node. For the irregular components, combining with the knowledge distillation mechanism of federated learning, a federated co-training mechanism for LSTM model is established. Subsequently, each node uses the ensemble strategy to obtain the final multi-load forecasting results.

The structure of this paper is as follows: Section 2 introduces the basic knowledge of LSTM, knowledge distillation, mode decomposition, etc. The proposed MD-MMFF model is described in detail in Section 3. In Section 4, the effectiveness of the proposed method is verified by comparative experiments. The conclusion of this paper is given in Section 5.

2 RELATED WORK

2.1 LSTM

As the scale of load forecasting data increases, its characteristics become more diverse, encompassing both linear and nonlinear features. Traditional methods excel

at handling linear data but struggle with large-scale nonlinear time-series data. To address these issues, scholars have proposed various methods including support vector machine [25], k-nearest neighbor [26], and gradient boosted decision tree [27]. Among them, the LSTM proposed by Hochreiter and Schmidhuber offers significant advantages. This recursive neural network architecture is widely used for time-series data prediction due to its exceptional performance in processing sequence data. Figure 1 shows the unit structure of LSTM, which incorporates forget gate, input gate, and output gate. In this diagram, x_t and h_{t-1} respectively represent the current input and the output from the previous time step, while c_{t-1} denotes the unit state from the previous time. They are used as inputs for the current unit. Similarly, h_t and c_t represent the output and state of the current unit, respectively.

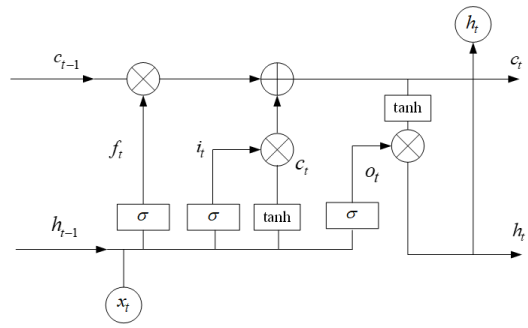

Figure 1. Unit structure of LSTM

The forget gate refers to the processing process of x_t and h_{t-1}, specifically using the sigmoid function to selectively process the information of the previous unit. The gate is represented by f_t and its formula is as follows.

$$f_t = \sigma\left(\omega_f \cdot [h_{t-1}, x_t] + b_f\right), \tag{1}$$

where ω_f is the weight matrix, b_f is the bias, σ is the sigmoid function.

The function of the input gate is to obtain a new cell state by fusing x_t and h_{t-1}. The process is mainly divided into three steps. Firstly, the update information is processed and stored for x_t and h_{t-1}, which is denoted as i_t, and its formula is shown as follows.

$$i_t = \sigma\left(\omega_i \cdot [h_{t-1}, x_t] + b_i\right). \tag{2}$$

Here, ω_i is the weight matrix, b_i is the bias.

Then, remember $\tilde{c_t}$ as the current input cell state, as shown in formula (3),

$$\tilde{c_t} = \tanh\left(\omega_c \cdot [h_{t-1}, x_t] + b_c\right), \tag{3}$$

where ω_c is the weight matrix, b_c is the bias.

Finally, by combining formula (2) and (3), the state c_t of this unit is obtained, and its formula is as follows.

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t. \tag{4}$$

In this way, the current information and the previous time information are combined to form a new unit state.

The output gate controls the effect of information from the previous moment on the current output. o_t is the output value of the status value of the control storage unit, and its formula is shown in (5). After integrating c_t and o_t, the output information of the current unit is obtained by formula (6).

$$o_t = \sigma\left(\omega_o \cdot [h_{t-1}, x_t] + b_o\right), \tag{5}$$

$$h_t = o_t \cdot \tanh\left(c_t\right). \tag{6}$$

2.2 Knowledge Distillation

Knowledge distillation [28] is a teacher-student-based training structure. In this method, the trained teacher model provides knowledge, and the student model is trained by distillation to acquire the teacher's knowledge to increase its model performance. The loss obtained by using the output of the teacher model to train the student model is called the soft loss, the loss between the predicted value and the expected value obtained by the trained model is called the hard loss, and the sum of the soft loss and the hard loss is called the total regression loss. The model is trained with the goal of minimizing the total regression loss of the student model, and the influence of the teacher model on the student model is adjusted by setting the proportion of soft loss and hard loss, in order to optimize the student model. The specific expression of the total regression loss of the student model is shown in Equations (7), (8), (9).

$$l_{soft} = \left\| y_{student} - y_{teacher} \right\|_2, \tag{7}$$

$$l_{hard} = \left\| y_{student} - y_{true} \right\|_2, \tag{8}$$

$$L_{distillation} = \alpha \cdot l_{soft} + (1 - \alpha) \cdot l_{hard}, \tag{9}$$

Where, $y_{student}$ represents the predicted value of the student network, $y_{teacher}$ represents the predicted value of the teacher model, and y_{true} is the true expected value. α is the hyperparameter that adjusts the proportion of soft and hard loss, and $L_{distillation}$ is the result of the summation of soft and hard loss multiplied by the corresponding proportion coefficient. By minimizing the loss value, the parameters of the student model are updated to help the student model training and obtain more accurate load prediction values. The specific process is shown in Figure 2.

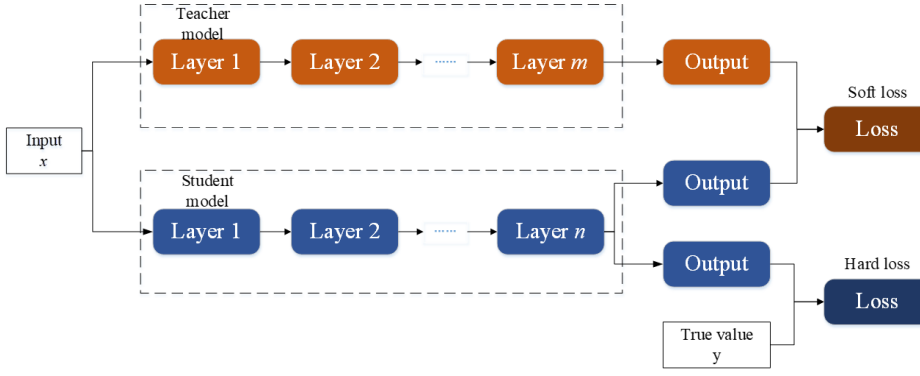

Figure 2. The framework of knowledge distillation

2.3 Multivariate Empirical Mode Decomposition

The traditional EMD decomposition method obtains the average value of the upper and lower envelopes to calculate the interpolation between the local maximum and minimum values, and then it obtains the local mean. However, in the decomposition problem of multi-load data, it is difficult to directly define local maxima and minima among multiple loads. In order to solve this problem, ur Rehman et al. [29] proposed the multivariate empirical mode decomposition method (MEMD). On the basis of EMD, MEMD extends the identification channel from one dimension to n dimensions, projects the n-dimensional signal along different directions, and generates the upper and lower envelopes in each projection direction, thus obtaining the n-dimensional envelope. The local mean is obtained by averaging the n-dimensional envelope, and the intrinsic mode functions (IMFs) in each channel are successively selected by subtracting the input signal from the local mean. With the increase of decomposition times, the frequency of components decomposed by IMF decreases in turn, and the last component with the lowest frequency is the Residual function (Res). Through MEMD, the multi-load historical data can be adaptively decomposed into a series of IMFs with the same number and matching frequency scales.

The basic steps of MEMD are as follows:

Step 1: Let an n-dimensional data sequence as $v(t) = [v_1(t), v_2(t), \ldots, v_n(t)]$, denoted it as the original signal, where one dimension represents a load type, $v(t)$ is the set of n kinds of load whose time series is t. Using the Hammersley sequence sampling method, Q appropriate sampling points are selected on a sphere as the direction of the projection vector, and Q direction vectors d^{θ_q}, $q = 1, 2, \ldots, Q$ are obtained, where q represents any point in the n-dimensional space and θ_q represents any projection direction.

Step 2: Project the n-dimensional original signal $v(t)$ along the direction vector d^{θ_q}, obtain the corresponding n-dimensional projection sequence P.

Step 3: Find the time set $\left\{t_J^{\theta_q}\right\}_{q=1}^{Q}$ when the n-dimensional projection sequence P reaches the maximum value, where $J \in [1, L]$ represents the time when P sequence reaches the maximum value, L is the length of historical load data.

Step 4: Use the multivariate spline interpolation function to interpolate $\left[t_J^{\theta_q}, v\left(t_J^{\theta_q}\right)\right]$, obtain the multivariate envelope $\left\{e^{\theta_q}(t)\right\}_{q=1}^{Q}$, where $e^{\theta_q}(t)$ denotes the envelope after q point projection.

Step 5: Calculate the mean envelope $m(t)$ of Q direction vectors on the sphere, where, $m(t) = \frac{\sum_{q=1}^{Q} e^{\theta_q}}{Q}$;

Step 6: Define the intrinsic mode function $h(t) = v(t) - m(t)$, and verify whether $h(t)$ satisfies the stopping condition of the intrinsic mode function, that is, the number of passing through zero points in the function is the same as the number of passing through extreme points or the difference is 1, and the mean value of the upper and lower envelopes is zero at any time point. If so, let $v(t) = v(t) - h(t)$ and proceed to step 7. Otherwise, go back to step 1.

Step 7: Let the first $h(t)$ satisfying the stopping condition to be the first intrinsic mode function x_1, which corresponds to the high-frequency component. After subtracting x_1 from $v(t)$, obtain a new data sequence y_1 with the high frequency component removed. The second eigenmode component x_2 can be obtained by performing the same operation on the new data sequence y. This is repeated until the last data sequence x_n is not decomposable, where x_n represents the trend of data sequence $v(t)$. Finally, for a multivariate time series $v(t)$, n intrinsic mode components can be obtained.

3 THE PROPOSED MULTI-LOAD FEDERATED FORECASTING

3.1 Framework

The MD-MMFF model mainly includes three parts: data decomposition, local model building, and personalized federated learning based on knowledge distillation. In the first part, in each node the multi-load data is decomposed into multiple components by the MEMD decomposition method, and the zero-crossing rate is used as the judgment criterion. Based on this, the multi-load data is divided into three types of components: high frequency component, low frequency component and trend component. In the following, the low-frequency and trend components are collectively referred to as regular components, and the high-frequency components are referred to as irregular components. In the second part, for the regular component, the LSTM model is only used locally in each node for learning. Due to the high randomness of irregular components, it is challenging to build an accurate prediction

model using only local data. Even if a locally trained model performs well on historical data, it may struggle to maintain high accuracy on new datasets. Therefore, the federated learning is employed in the paper to conduct collaborative training among multiple nodes. This can enhance the model's generalization ability and adaptability to different datasets with random fluctuations.Therefore,for all irregular components, they are converted into one irregular component by stacking, from which the LSTM local model is built. In the third part, the parameter sharing federated learning is used between nodes to perform collaborative learning on the LSTM model established by the irregular components of each node. Based on this, each node can obtain a federated training model LSTM′. After that, taking the federated training model LSTM′ as the student model, and the LSTM local model built for the irregular components as the teacher model, the knowledge distillation method is used to fine tune LSTM′ on each node. Finally, the predicted values of the regular and irregular components are integrated to obtain the final prediction result.

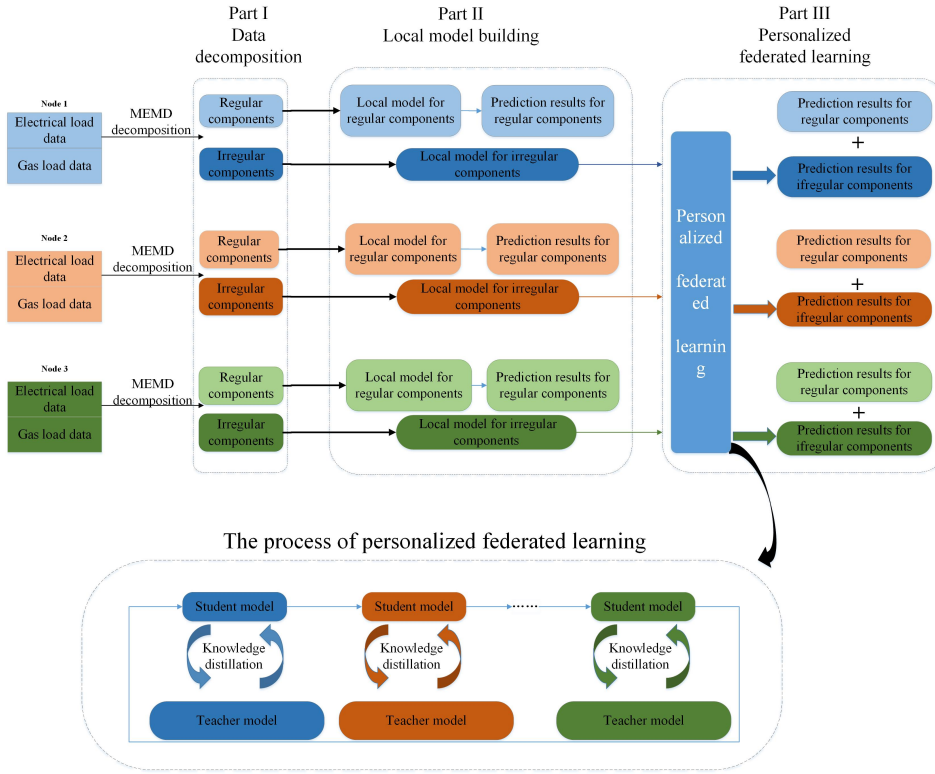

Figure 3. The basic framework of MD-MMFF

Taking the case of three participants/nodes as an example, Figure 3 shows the basic framework of the proposed MD-MMFF model. Furthermore, the main steps of the proposed MD-MMFF are as follows:

Step 1: Use MEMD to decompose the local multi-load data in each node.

Step 2: In each node, use the zero-crossing rate to classify those components of different energy data into high-frequency components, low-frequency components and trend components. For the decomposed data of each energy source, all high-frequency components are integrated into an irregular component.

Step 3: In each node, use the LSTM model to learn the regular component data of each energy source, and establish their LSTM local models. Similarly, establish the LSTM local models for the irregular component data of each energy source.

Step 4: Combine the improved FEDPS framework with knowledge distillation to co-train the local LSTM model obtained from training the irregular component.

Step 5: Repeat Step 4 until a predetermined number of iterations is reached, from which each node can obtain the final prediction model of the irregular component.

Step 6: Integrate the predicted values of regular and irregular components, output the final multi-load forecast value.

3.2 Multivariate Empirical Mode Decomposition for Multi-Load Data

Multi-load in integrated energy system is usually nonlinear, random and uncertain, and it is difficult to establish an accurate prediction model by simply using multi-load historical data. In order to better describe the complex characteristics of loads, it is necessary to extract different characteristics of loads. By using MEMD, in the paper different load data in a node is decomposed into the same number of IMF components.

Different frequency components obtained by decomposition represent the unique information in the original historical load data. For example, the high frequency component can well reflect the fluctuation degree of historical load data, and the low frequency component can reflect the change trend of load in a certain time. Therefore, the overall load forecasting accuracy can be improved by reasonably dividing the components and using different methods to establish forecasting models. By calculating the ratio of the number of times a component crosses the horizontal axis to the number of sample points of the component, the zero-crossing rate can well reflect the fluctuation of data. Therefore, this paper uses the zero-crossing rate to distinguish regular components from irregular components. The definition of the signal zero-crossing rate Z is shown in formula (10):

$$Z = \frac{n_{zero}}{U}, \tag{10}$$

where n_{zero} denotes the number of times which the component crosses the horizontal axis, and U denotes the number of component sample points. The specific process is shown in Figure 4.

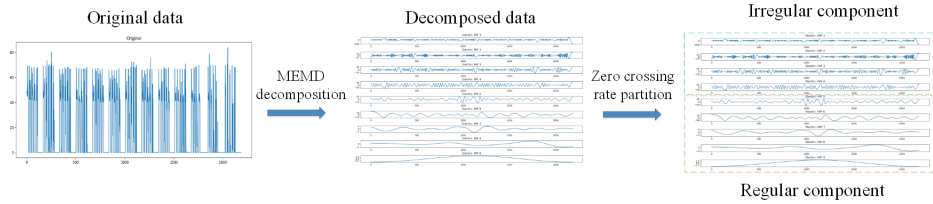

Figure 4. Data decomposition process

Due to the different historical load data of each node, the number of components obtained by each node after MEMD decomposition will be different. In order to facilitate collaborative training using federated learning, this paper overlays all irregular components to obtain a unified irregular component.

3.3 Knowledge Distillation Based Multi-Load Federated Training Mechanism

Aiming at the problem of heterogeneous prediction models between nodes caused by different time scales under federated learning, a parameter sharing federated learning prediction method for data decomposition is proposed in the paper. Data decomposition is used to extract the personality characteristics of node time series data, and federated learning is used to solve the problem of difficult collaborative training caused by differences in prediction models without leaking data privacy of each node.

To address the issue of differing time scales among loads within a node or across different nodes, spline interpolation is used to resample the load data with larger time scales, ensuring that all energy loads are standardized to the smallest time scale. Each node utilizes the multivariate load data at the smallest time scale to generate a local model. Without compromising data privacy, federated learning is employed to collaboratively train the models from different nodes. After training, fine-tuning is performed within each node to optimize its model.

In each node, multi-load data will be divided into regular components and irregular components by using the method in Section 3.2. On the basis of these decomposition results, the second part of the method is executed, and each node establishes the LSTM local model for the local regular components and irregular components. For the regular components, since they have certain regularity, LSTM with good prediction performance is used locally to learn and predict them. For the irregular components, due to its randomness, it is difficult to establish an accurate prediction model by using LSTM only locally. At the same time, since the load data of different nodes have different time scales, the LSTM prediction

models established by each node will have different values in terms of hyperparameters such as the number of LSTM layers and Dropout rate, and it is difficult to use the traditional federated learning framework to establish a unified prediction model. To this end, this paper proposes a personalized federated learning method, namely the improved parameter-sharing federated learning method with knowledge distillation (KD-FEDPS). Figure 5 shows the basic framework of the method.

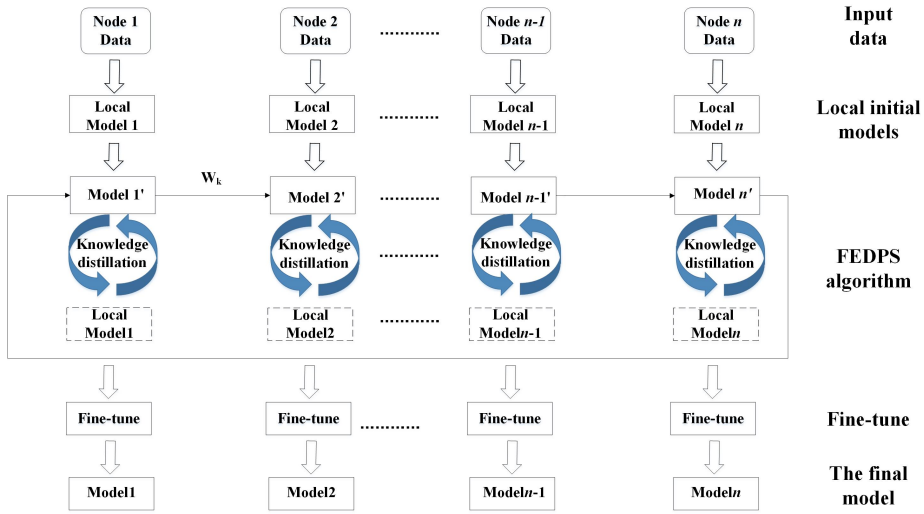

Figure 5. The framework of KD-FEDPS

The detailed execution of KD-FEDPS is describesd below. Based on the local models established on each node, we firstly establish a preliminary federated co-training model. Taking the first node (Node 1) as an example, firstly, the node passes the parameters of the local initial model Model1 to the second node. After the second node accepted these parameters, it updates the model Model1 with its own data, denoted the updated model as Model1′, and passes the parameters of Model1′ to the third node. And so on, after the last node completed local training, the parameters of Model1′ are fed back to Node 1. Because the initial model Model1 only uses local historical load data on Node 1 for training and has good personality characteristics, it is used as the teacher model. Using Model1′ as the student model, with the goal of minimizing the loss function, the knowledge distillation method is used to optimize the parameters of Model1′ by using a small batch stochastic gradient descent method. This can enable Model1 to learn the personality features of Node 1 well. We use the updated Model1′ as the prediction model for Node 1. A similar process is used for other nodes to operate. Then, each participant uses the data they own to adjust the local model with the Fine-tune technology. Here, the parameters of each layer before the n^{th} layer of each model are frozen, and only the

local data of nodes are used to retain the parameters of the last layer of the initial model. In this way, the final multivariate load federated collaborative forecasting model can be obtained.

Finally, the predicted values of the regular components predicted by the local LSTM model and the predicted values of the irregular components predicted by the LSTM prediction model under the KD-FEDPS framework are summed and integrated to obtain the final prediction results.

4 EXPERIMENTS

4.1 Experimental Data

This paper takes the data of three schools in Carbon Culture platform as research subjects. These three schools are named school A, B, and C. It is assumed from a security point of view that load data are not be exchanged among them. Here, the data contains the amount of electricity and gas consumed by the above school during a certain period of time. The load data of school A is collected every hour, while the data of school B and C can be collected every half an hour, which indicates the difference in data scale between them. Here, the electricity and gas data of these three schools from January 14 to April 14 are selected.

We used the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) as evaluation indexes in this experiment to illustrate the quality of the method. Their formulas are as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2}, \tag{11}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - x_i^{'}\right|, \tag{12}$$

where n is the number of predicted time points, $x_i^{'}$ is the predicted value at time i, and x_i is the true value at time i. For RMSE and MAE, the experimental results are the average values of 20 independent runs.

The same computer was used for all experiments. The experimental environment is as follows: AMD Ryzen 5 2400G with Radeon Vega Graphics 3.60 GHz processor, and the software uses Windows 10 operating system and PyCharm compiler with Python version 3.8.

4.2 Comparison Algorithms

Two existing typical multi-load forecasting algorithms are used to verify the superiority of the proposed algorithm KD-FEDPS in the section. They are as follows:

- FedAvg. It is a basic federated learning algorithm. The algorithm uses the method of gradient descent to update the model parameters iteratively.
- Multi-attention mode [30]. It is a load forecasting method integrating attention mechanism. This method uses the attention mechanism strategy to learn the coupling relationship between the data, so as to form a multivariate load forecasting model.
- Transformer. It is a deep learning-based load forecasting method that leverages the Transformer architecture. This method utilizes the self-attention mechanism to capture long-range dependencies in load data, improving the model's ability to learn complex temporal patterns.

For fairness, all algorithms employ the same LSTM base algorithm. The model structure and parameters of each node in Multi attention model and the algorithm proposed in this paper are the same, and the model structure and parameters between each node in FedAvg are the same. In both the above method and the proposed method, the number of training iterations is set to 75. Additionally, all methods follow the same training iteration process, without any additional training on specific nodes. To reduce the impact of random model initialization, all reported prediction results are the average values of multiple independent runs. Specifically, in the KD-FEDPS algorithm, Table 1 shows the hyperparameter settings of the LSTM model used by the three nodes participating in federated learning.

	Number of LSTM layers	Dropout Coefficient	batch_size	Optimizer	Number of distillation
School A	3	0.3	21	adam	20
School B	5	0.4	42	adam	20
School C	4	0.4	42	adam	20

Table 1. Structure and parameters of the model used

4.3 Ablation Experiment

This section experimentally verifies the effectiveness of the proposed multi-load federated forecasting training mechanism and the mode decomposition method. To this end, two sets of ablation experiments are set up. In the first group, after decomposing the energy consumption data, the federated learning framework is used for co-training on each component, but the personalized characteristics of each node are learned without using knowledge distillation. The improved method without knowledge distillation is denoted as MD-MMFF/KD. Through this set of experiments, the effectiveness of the personalized characteristics of the learning nodes can be verified, and the rationality of the knowledge distillation mechanism introduced in this paper can also be proved. In the second group, no federated learning is used,

and each node is trained individually locally. The improved method without federated learning is denoted as MD-MMFF/FL. The ablation experiment adopts the same hyperparameter configurations as the proposed method. And all methods in the ablation experiment follow the same training strategy, including the number of training iterations, data distribution, and model update process, ensuring consistent experimental conditions and eliminating the potential impact of additional training stages on performance. This set of experiments can illustrate the necessity of learning common features between nodes and prove the effectiveness of the federated learning strategy used in this paper.

School		Error Index	MD-MMFF	MD-MMFF/KD	MD-MMFF/FL
School A	Electricity	RMSE	**0.340**	0.360	0.537
		MAE	**0.164**	0.215	0.420
	Gas	RMSE	**1.130**	1.410	3.441
		MAE	**0.503**	0.723	1.227
School B	Electricity	RMSE	0.623	**0.350**	0.866
		MAE	0.403	**0.244**	0.546
	Gas	RMSE	**1.103**	1.350	5.461
		MAE	**0.656**	0.751	4.444
School C	Electricity	RMSE	**0.120**	0.210	0.198
		MAE	**0.079**	0.098	0.162
	Gas	RMSE	**0.562**	0.772	0.949
		MAE	**0.316**	0.540	0.626

Table 2. Results of ablation experiment

Table 2 shows the RMSE and MAE results of ablation experiments. Figure 6 illustrates the electrical and gas load prediction curves obtained by MD-MMFF, MD-MMFF/KD and MD-MMFF/FL. In the figure, the blue line represents the prediction results obtained by MD-MMFF, the red line represents the prediction results obtained by MD-MMFF/FL, the black line represents the prediction results obtained by MD-MMFF/FL, and the yellow line represents the true value. The horizontal axis represents the time scale, that is, the moment of prediction, and the vertical axis represents the load value.

It can be seen from Table 2 that the proposed algorithm achieves good electricity and gas load prediction values in most cases. In particular, compared with the local model without federated learning mechanism (MD-MMFF/FL), the proposed algorithm achieves better prediction results with the help of federated learning mechanism.

Furthermore, we can see from Figure 6 that the proposed algorithm obtains better electrical load prediction values in most cases, compared to MD-MMFF/KD and MD-MMFF/FL. In particular, since the federated learning strategy was not used, the effect of each node trained separately is relatively the worst, and its prediction results are significantly worse than the results of the multi-load fore-

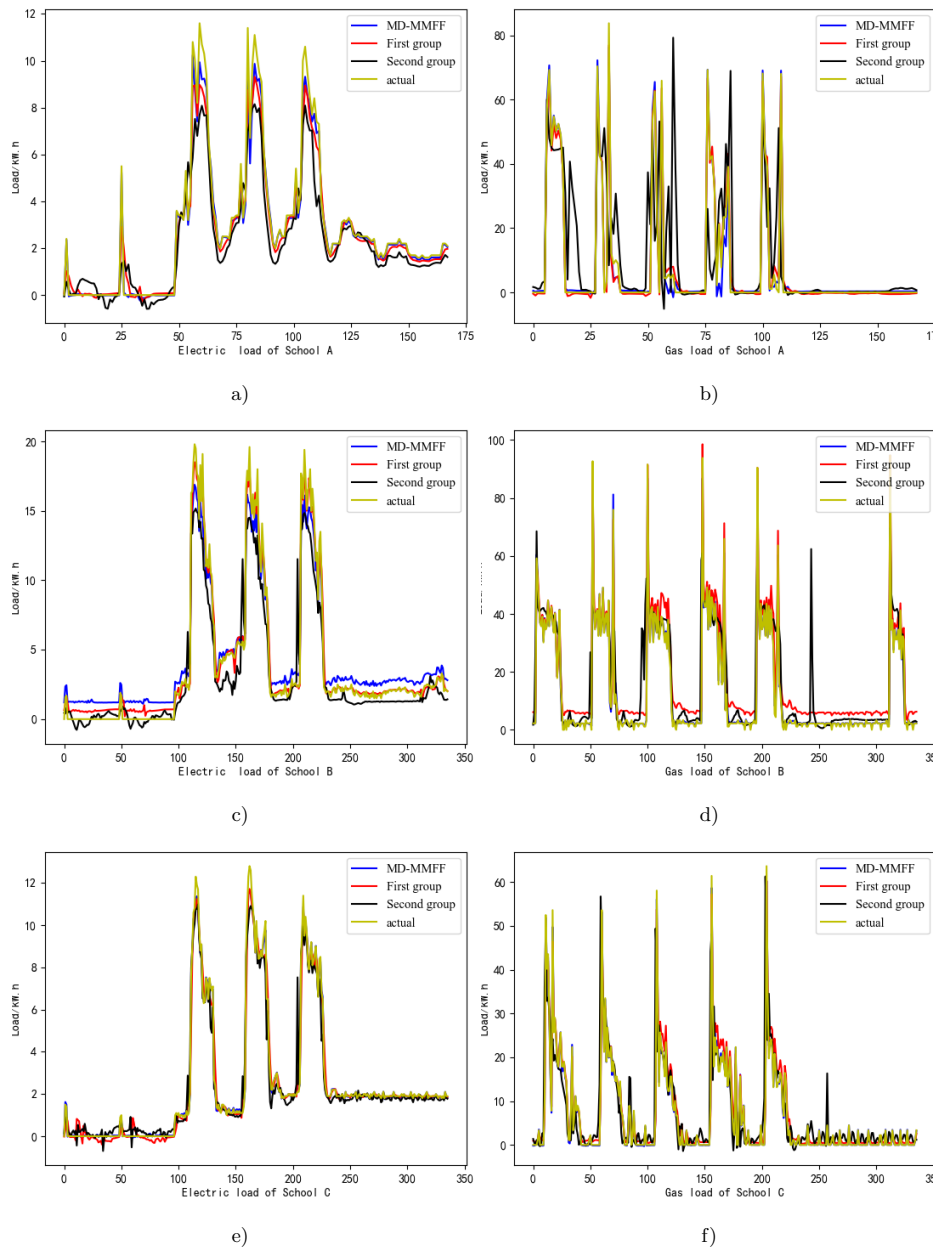

Figure 6. Load prediction curves obtained by MD-MMFF, MD-MMFF/KD, and MD-MMFF/FL

School		Error Index	MD-MMFF	FedAvg	Multi-attention	Transformer Model
School A	Electricity	RMSE	**0.340**	0.598	1.340	1.370
		MAE	**0.164**	0.275	0.682	0.707
	Gas	RMSE	**1.130**	1.523	1.421	1.700
		MAE	**0.503**	1.144	0.782	0.879
School B	Electricity	RMSE	0.623	**0.305**	1.081	1.594
		MAE	0.403	**0.238**	0.685	0.818
	Gas	RMSE	**1.103**	1.182	6.641	1.404
		MAE	**0.656**	0.655	4.205	0.748
School C	Electricity	RMSE	**0.120**	0.290	0.418	0.868
		MAE	**0.079**	0.186	0.286	0.477
	Gas	RMSE	**0.562**	1.057	0.858	0.826
		MAE	**0.316**	0.628	0.515	0.537

Table 3. RMSE and MAE values obtained by the four algorithms

casting method under the federated learning framework. Due to the lack of effective learning of personalized temporal characteristics on each node, the prediction results of MD-MMFF/KD without knowledge distillation are inferior to that of the proposed algorithm in the time periods with large and frequent load fluctuations.

Due to the small fluctuation amplitude and certain regularity of electric load, MD-MMFF, MD-MMFF/KD and MD-MMFF/FL have achieved relatively good electric load prediction results. However, at time 0 to 50 in school A, time 0 to 100 in School B and time 0 to 100 in school C, time 125 to 168 in School A and time 276 to 336 in School B and C, the prediction errors of the three methods become worse due to the high frequency and small fluctuation amplitude of electric load fluctuation. However, the prediction results of MD-MMFF are better than those of MD-MMFF/KD and MD-MMFF/FL at these moments, because in MD-MMFF the common features between nodes are effectively learned and the individual characteristics of nodes are preserved.

Compared with the electric load, the gas load fluctuates greatly. Since the federated learning strategy was not used, the prediction results of each node trained separately are significantly weaker than the results of MD-MMFF with federated learning. In particular, at the moments with obvious regularity such as 0–100 in School A and 0–200 in School B and C, the proposed method achieves higher prediction accuracy due to the effective learning of Personalized temporal characteristics of each node.

4.4 Comparative Experiments

This section experimentally verifies the effectiveness of the proposed MD-MMFF by comparing to FedAvg, Multi-attention and Transformer model. Table 3 shows

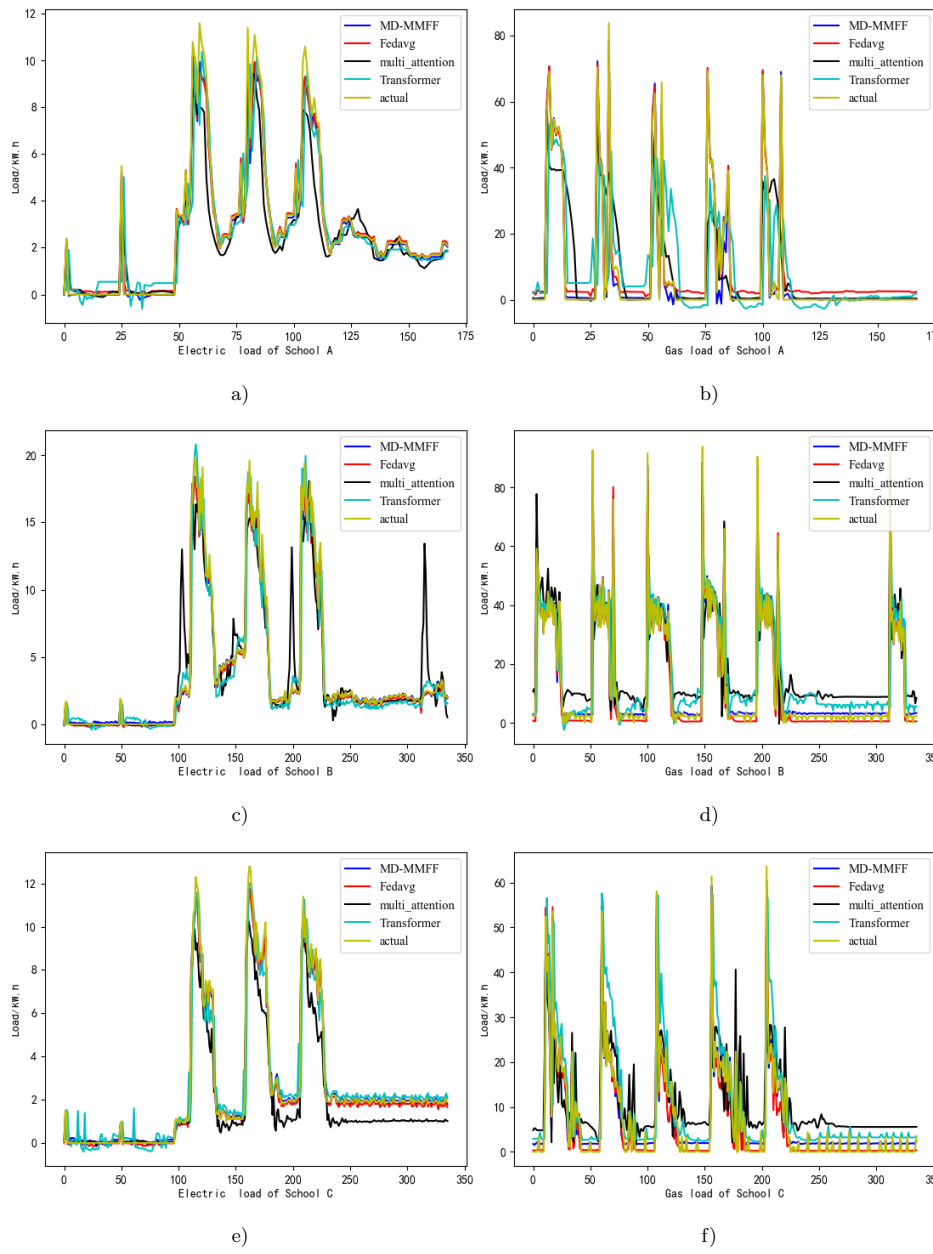

Figure 7. Load prediction curves obtained by MD-MMFF, FedAvg, Multi-attention and Transformer

their RMSE and MAE results. It can be seen from Table 3 that the proposed algorithm achieves good prediction results in most cases. Compared with MD-MMFF and FedAvg, the prediction results of Multi-attention and Transformer model are poor.

Figure 7 shows the forecasting result curves of electricity and gas load obtained by the proposed algorithm and the three comparison algorithms. In the figure, the blue solid line represents the prediction result obtained by the proposed MD-MMFF, the red solid line represents the prediction result obtained by FedAvg, the black solid line represents the prediction result obtained by the Multi-attention model, the cyan solid line represents the prediction result obtained by the Transformer model and the yellow solid line is the true value. We can see that:

1. For School A, the fluctuation of electric load from time 50 to 125 has certain regularity, but the fluctuation range is large. Compared with the three comparison algorithms, the proposed method in this paper achieves better prediction results. In terms of gas load, the FedAvg's prediction accuracy is relatively poor due to the lack of learning personalized characteristics. The Multi-attention and Transformer models perform poorly at the extreme points with sharp load variations, and the Transformer model struggles to make accurate predictions when the gas load is zero.

2. For School B, in terms of electricity load prediction, although the proposed algorithm in this paper captures the fluctuation of load well, its prediction accuracy is general. However, for the gas load with better regularity, the proposed algorithm obtains better prediction results compared to the three comparison algorithms.

3. For School C, in terms of electricity load prediction, the proposed algorithm and FedAvg,Transformer algorithm obtains good prediction results in most time periods. For the time of 225–336 with large fluctuation frequency and small fluctuation amplitude, the prediction results of the proposed algorithm are better than that of the FedAvg algorithm. For gas load, at the moment of small fluctuation amplitude and high fluctuation frequency, the gas load prediction results obtained by FedAvg are relatively poor because the client-specific temporal patterns of a time series are not well learned.The Transformer model performs poorly in the peak value section and is unable to capture the declining trend of the gas load peak.

In summary, for the vast majority of moments, the load forecasting results obtained by the proposed MD-MMFF are closer to the true value. In particular, when the load fluctuation amplitude is small and the frequency is high, and when the load fluctuation amplitude is large, the proposed algorithm has better prediction accuracy on the electricity and gas loads.

5 CONCLUSIONS

This paper proposed a multi-time scale multi-load federated forecasting method based on mode decomposition, called MD-MMFF, which effectively solves the problem that the existing federated learning methods do not mine the common and individual characteristics of energy consumption data. In the MD-MMFF, the MEMD-based multi-load data decomposition strategy can effectively decompose different load data held by each node into regular and irregular components. Combining with federated learning, LSTM and a knowledge distillation mechanism, the established load forecasting model can fully learn the irregular components on nodes. Finally, the proposed method was applied to multiple integrated energy systems, the experimental results show its effectiveness.

Although the proposed MD-MMFF shows better performance, two challenges remain for its practical deployment. First, the communication and computation costs between heterogeneous nodes are still relatively high. Future work should study lightweight federated frameworks with adaptive parameter pruning to reduce overhead. Second, extreme scenarios may degrade the robustness of prediction results. This may be a good solution method to integrate anomaly detection mechanisms with the current decomposition strategy for enhancing the model's adaptability. Additionally, combining blockchain-based secure aggregation and edge computing could further address privacy-efficiency trade-offs in multi-energy systems.

Acknowledgments

This work was jointly supported by The Key Research and Development Program of Xuzhou (Modern Agriculture) under Grant No. KC23129, The National Natural Science Foundation of China under Grant No. 62203446 and Grant No. 62273348, and Qing Lan Project of Jiangsu University.

REFERENCES

[1] Rifkin, J.: The Third Industrial Revolution: How Lateral Power Is Transforming Energy, the Economy, and the World. St. Martin's Press, 2011.

[2] Zhu, J.—Dong, H.—Li, S.—Chen, Z.—Luo, T.: Review of Data-Driven Load Forecasting for Integrated Energy System. Proceedings of the CSEE, Vol. 41, 2021, No. 23, pp. 7905–7924, doi: 10.13334/j.0258-8013.pcsee.202337 (in Chinese).

[3] Hochreiter, S.—Schmidhuber, J.: Long Short-Term Memory. Neural Computation, Vol. 9, 1997, No. 8, pp. 1735–1780, doi: 10.1162/neco.1997.9.8.1735.

[4] Wang, B.—Zhang, L.—Ma, H.—Wang, H.—Wan, S.: Parallel LSTM-Based Regional Integrated Energy System Multienergy Source-Load Information Interactive Energy Prediction. Complexity, Vol. 2019, 2019, No. 1, Art. No. 7414318, doi: 10.1155/2019/7414318.

[5] Zhu, R.—Guo, W.—Gong, X.: Short-Term Load Forecasting for CCHP Systems Considering the Correlation Between Heating, Gas and Electrical Loads Based on Deep Learning. Energies, Vol. 12, 2019, No. 17, Art. No. 3308, doi: 10.3390/en12173308.

[6] Tan, Z.—De, G.—Li, M.—Lin, H.—Yang, S.—Huang, L.—Tan, Q.: Combined Electricity-Heat-Cooling-Gas Load Forecasting Model for Integrated Energy System Based on Multi-Task Learning and Least Square Support Vector Machine. Journal of Cleaner Production, Vol. 248, 2020, Art. No. 119252, doi: 10.1016/j.jclepro.2019.119252.

[7] Zhou, B.—Meng, Y.—Huang, W.—Wang, H.—Deng, L.—Huang, S.—Wei, J.: Multi-Energy Net Load Forecasting for Integrated Local Energy Systems with Heterogeneous Prosumers. International Journal of Electrical Power & Energy Systems, Vol. 126, Part A, 2021, Art. No. 106542, doi: 10.1016/j.ijepes.2020.106542.

[8] Wang, X.—Wang, S.—Zhao, Q.—Wang, S.—Fu, L.: A Multi-Energy Load Prediction Model Based on Deep Multi-Task Learning and Ensemble Approach for Regional Integrated Energy Systems. International Journal of Electrical Power & Energy Systems, Vol. 126, Part A, 2021, Art. No. 106583, doi: 10.1016/j.ijepes.2020.106583.

[9] Xu, Y.—Yang, X.—Zhang, J.—Zhu, J.—Sun, M.—Chen, B.: Proof of Engagement: A Flexible Blockchain Consensus Mechanism. Wireless Communications & Mobile Computing, Vol. 2021, 2021, Art. No. 6185910, doi: 10.1155/2021/6185910.

[10] Jindal, A.—Dua, A.—Kaur, K.—Singh, M.—Kumar, N.—Mishra, S.: Decision Tree and SVM-Based Data Analytics for Theft Detection in Smart Grid. IEEE Transactions on Industrial Informatics, Vol. 12, 2016, No. 3, pp. 1005–1016, doi: 10.1109/TII.2016.2543145.

[11] Chuwa, M. G.—Wang, F.: A Review of Non-Technical Loss Attack Models and Detection Methods in the Smart Grid. Electric Power Systems Research, Vol. 199, 2021, Art. No. 107415, doi: 10.1016/j.epsr.2021.107415.

[12] Cui, L.—Qu, Y.—Gao, L.—Xie, G.—Yu, S.: Detecting False Data Attacks Using Machine Learning Techniques in Smart Grid: A Survey. Journal of Network and Computer Applications, Vol. 170, 2020, Art. No. 102808, doi: 10.1016/j.jnca.2020.102808.

[13] McMahan, B.—Moore, E.—Ramage, D.—Hampson, S.—Aguera y Arcas, B.: Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Singh, A., Zhu, J. (Eds.): Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research (PMLR), Vol. 54, 2017, pp. 1273–1282, https://proceedings.mlr.press/v54/mcmahan17a.html.

[14] Geiping, J.—Bauermeister, H.—Dröge, H.—Moeller, M.: Inverting Gradients – How Easy Is It to Break Privacy in Federated Learning? In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., Lin, H. (Eds.): Advances in Neural Information Processing Systems 33 (NeurIPS 2020). Curran Associates, Inc., 2020, pp. 16937–16947, doi: 10.48550/arXiv.2003.14053.

[15] Yang, Q.: AI and Data Privacy Protection: The Way to Federated Learning. Journal of Information Security Reserach, Vol. 5, 2019, No. 11, pp. 961–965 (in Chinese).

[16] Liu, Z.—Chen, Y.—Zhao, Y.—Yu, H.—Liu, Y.—Bao, R.—Jiang, J.—

Nie, Z.—Xu, Q.—Yang, Q.: Contribution-Aware Federated Learning for Smart Healthcare. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, No. 11, pp. 12396–12404, doi: 10.1609/aaai.v36i11.21505.

[17] Subramanya, T.—Riggio, R.: Centralized and Federated Learning for Predictive VNF Autoscaling in Multi-Domain 5G Networks and Beyond. IEEE Transactions on Network and Service Management, Vol. 18, 2021, No. 1, pp. 63–78, doi: 10.1109/TNSM.2021.3050955.

[18] Wu, X.—Liang, Z.—Wang, J.: FedMed: A Federated Learning Framework for Language Modeling. Sensors, Vol. 20, 2020, No. 14, Art. No. 4048, doi: 10.3390/s20144048.

[19] Venkataramanan, V.—Kaza, S.—Annaswamy, A. M.: DER Forecast Using Privacy-Preserving Federated Learning. IEEE Internet of Things Journal, Vol. 10, 2023, No. 3, pp. 2046–2055, doi: 10.1109/JIOT.2022.3157299.

[20] Savi, M.—Olivadese, F.: Short-Term Energy Consumption Forecasting at the Edge: A Federated Learning Approach. IEEE Access, Vol. 9, 2021, pp. 95949–95969, doi: 10.1109/ACCESS.2021.3094089.

[21] Chen, Y.—Ning, Y.—Chai, Z.—Rangwala, H.: Federated Multi-Task Learning with Hierarchical Attention for Sensor Data Analytics. 2020 International Joint Conference on Neural Networks (IJCNN), 2021, pp. 1–8, doi: 10.1109/IJCNN48605.2020.9207508.

[22] Zhang, G.—Zhu, S.—Bai, X.: Federated Learning-Based Multi-Energy Load Forecasting Method Using CNN-Attention-LSTM Model. Sustainability, Vol. 14, 2022, No. 19, Art. No. 12843, doi: 10.3390/su141912843.

[23] Zhang, W.—Chen, X.—He, K.—Chen, L.—Xu, L.—Wang, X.—Yang, S.: Semi-Asynchronous Personalized Federated Learning for Short-Term Photovoltaic Power Forecasting. Digital Communications and Networks, Vol. 9, 2023, No. 5, pp. 1221–1229, doi: 10.1016/j.dcan.2022.03.022.

[24] Wang, Y.—Gao, N.—Hug, G.: Personalized Federated Learning for Individual Consumer Load Forecasting. CSEE Journal of Power and Energy Systems, Vol. 9, 2023, No. 1, pp. 326–330, doi: 10.17775/CSEEJPES.2021.07350.

[25] Zhang, C.—Liu, Y.—Zhang, H.—Huang, H.: Research on the Daily Gas Load Forecasting Method Based on Support Vector Machine. 2011 Fourth International Conference on Intelligent Computation Technology and Automation, IEEE, Vol. 1, 2011, pp. 222–227, doi: 10.1109/ICICTA.2011.65.

[26] Lv, X.—Cheng, X.—Yan, S.—Tang, Y.: Short-Term Power Load Forecasting Based on Balanced KNN. IOP Conference Series: Materials Science and Engineering, Vol. 322, 2018, No. 7, Art. No. 072058, doi: 10.1088/1757-899X/322/7/072058.

[27] Liu, S.—Cui, Y.—Ma, Y.—Liu, P.: Short-Term Load Forecasting Based on GBDT Combinatorial Optimization. 2018 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2), 2018, pp. 1–5, doi: 10.1109/EI2.2018.8582108.

[28] Hinton, G.—Vinyals, O.—Dean, J.: Distilling the Knowledge in a Neural Network. CoRR, 2015, doi: 10.48550/arXiv.1503.02531.

[29] ur Rehman, N.—Xia, Y.—Mandic, D. P.: Application of Multivariate Empirical Mode Decomposition for Seizure Detection in EEG Signals. 2010 Annual International

Conference of the IEEE Engineering in Medicine and Biology, 2010, pp. 1650–1653, doi: 10.1109/IEMBS.2010.5626665.

[30] Zhang, J.—Wang, S.—Tan, Z.—Sun, A.: An Improved Hybrid Model for Short-Term Power Load Prediction. Energy, Vol. 268, 2023, Art. No. 126561, doi: 10.1016/j.energy.2022.126561.

Chao Peng received his Bachelor's degree in the School of Automation from Shandong University, Jinan, China in 2010. He is a doctoral student at the China University of Mining and Technology. His research interests include evolutionary algorithms, software testing, and machine learning.

Zikai Chen received his Bachelor's degree in electronic information engineering from the China University of Mining and Technology, Xuzhou, China, in 2023. He is currently pursuing a Master's degree in control science and engineering at the China University of Mining and Technology, Xuzhou, China. His current research interest is load forecasting in integrated energy systems.

Zhipeng Chen received his Master of Science degree in electronic information engineering from the China University of Mining and Technology, Xuzhou, China in 2024. His research focuses on load forecasting for integrated energy systems.

Yong Zhang is Professor and Doctoral Supervisor at the China University of Mining and Technology, Xuzhou, China. He received his Ph.D. from the China University of Mining and Technology. He is a member of the Committee on Natural Computing and Digital Intelligent Cities, Chinese Association for Artificial Intelligence. His main research interests include intelligent optimization and data mining.