



Research article

Multi-model fusion short-term power load forecasting based on improved WOA optimization

Xiaotong Ji^{1,*}, Dan Liu² and Ping Xiong²

¹ State Grid Hubei Electric Power Co., Ltd., Wuhan 430077, China

² State Grid Hubei Electric Power Research Institute, Wuhan 430077, China

* **Correspondence:** Email: tonyji77@hotmail.com.

Abstract: The high accuracy of short-term power load forecasting has a pivotal role in helping power companies to construct reasonable production scheduling plans and avoid resource waste. In this paper, a multi-model short-term power load prediction method based on Variational mode decomposition (VMD)-improved whale optimization algorithm (IWOA)-wavelet temporal convolutional network (WTCN)-bidirectional gated recurrent unit (BiGRU)-attention and CatBoost model fusion is proposed. First, VMD was employed to decompose the load data into different intrinsic mode functions. Second, a WTCN was utilized to extract the load data features, and multi-dimensional feature factors were integrated into the BiGRU network for model training. Moreover, an attention mechanism was added to enhance the influence degree of important information. The WTCN-BiGRU-attention model is improved by the WOA algorithm to optimize the hyperparameters of the network. Finally, the model was fused with the predicted data of the CatBoost network by the mean absolute percentage error-reciprocal weight (MAPE-RW) algorithm to construct the best fusion model. Compared with other forecasting models, the proposed multi-model fusion method has higher accuracy in short-term power load forecasting using the public data set for an Australian region.

Keywords: temporal convolutional networks; gated recurrent networks; improved whale optimization algorithm; catboost network; power load forecasting

1. Introduction

Power load forecasting can be divided into long-term forecasting, medium-term forecasting, short-term forecasting and ultra-short-term forecasting according to the forecasting time-scale. The forecasting period of short-term power load is typical, as it is a critical basis for maintaining the stable operation of the power system and improving economic benefits. The accuracy of the short-term power forecast can play an important role in addressing the issue of the power decision department controlling power

dispatch in the next step. Accurate short-term load forecasting can effectively reduce resource waste and improve economic benefits [1–3].

At present, load prediction methods primarily include a statistical prediction method composed of multiple linear regression [4], a Kalman filter [5, 6] an autoregressive moving average and a machine learning method composed of a support vector machine [7–9], an expert system and artificial neural networks [10]. Research has consistently shown that the calculation model of the statistical method is too ordinary, as it can only deal with linear data but cannot grasp the inherent characteristics of nonlinear data reasonably. Although the machine learning method can deal with nonlinear data well, it cannot extract time-series data features effectively. With the development of deep learning, it becomes the focus of load forecasting. A large number of deep neural networks are widely employed in load prediction, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs) [11] and short and long duration memory networks (LSTM) [12]. CNNs can effectively extract multidimensional data features, but it cannot deal with time-series features efficiently. RNNs can model long time-series data through a cyclic structure, but with the increase of load data, there are problems such as gradient disappearance or gradient explosion. As a special RNN, the LSTM network can better solve the deficiency of the RNN through the use of a gate structure. Nonetheless, with the increase in training data, it is difficult to select parameters for the LSTM network [13]. In order to effectively process multi-dimensional power load data, the CNN-LSTM hybrid neural network prediction method was proposed in the literature [14]. Feature extraction was carried out through a two-dimensional convolutional layer to reduce the training difficulty of the LSTM network model. Surveys such as that conducted by the authors of [15] showed that using the CNN to extract data features, using the gated recurrent neural (GRU) network to avoid the problem of multiple training parameters in the LSTM network and introducing an attention mechanism, can effectively improve the accuracy of power load forecasting. Reference [16] found that their CNN-BiGRU network improves data utilization in order to make data flow bidirectional in the network layer. According to the research, since CNN networks cannot predict time series data well, time series convolutional networks (TCNs) can be employed for sequence data prediction. And TCN can extract time series data features better than CNN and RNN [17]. Tian et al. [18] proposed a short-term wind speed prediction model employing empirical modal decomposition and an improved sparrow algorithm to optimize the LSTM neural network. The model decomposes the ultra-short-term wind speed by utilizing empirical modal decomposition, predicts it by employing the LSTM network and optimizes the LSTM network hyperparameters by improving the sparrow optimization algorithm. In [19] for short-term wind speed prediction, a prediction model based on local mean decomposition (LMD) with a combined kernel function least squares support vector machine (LSSVM) is proposed. Wind speed data are decomposed by the LMD algorithm and predicted by the LSSVM, and the firefly algorithm is employed to optimize the parameter selection. The authors of [20] proposed a time-series convolutional network with the multi-attentional mechanism. By introducing an initial structure into the TCN network, multidimensional information was extracted from convolutional kernels of disagreement scales, improving the accuracy of ultra-short-term load prediction effectively. The authors of [21] proposed a combined prediction model based on empirical modal decomposition to forecast traffic flow state information. The empirical modal decomposition is decomposed into components, the optimal prediction method is selected based on the results of adaptive analysis and the combined model weights are optimized by employing the fruit fly algorithm. The authors of [22] proposed a combined prediction model based on ensemble empirical modal decomposition and a reg-

ularized limit learning machine for wind speed prediction. The wind speed series of the ensemble empirical modal decomposition is predicted by employing the regularized limit learning machine, and the reliability of the prediction model is improved by cross-validation. Recent evidence suggests that complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) and variational model decomposition (VMD) can be employed to decompose power load data. Second, non-stationary components and stationary components are predicted by deep bidirectional long short-term memory (DBILSTM) and mixed logistic regression (MLR) networks respectively. Finally, the prediction accuracy is improved by combining the prediction structure with data reconstruction [23].

This paper presents a multi-model fusion method for short-term power load forecasting based on VMD, improved whale optimization (WOA), Wavelet temporal convolutional network (WTCN)-BiGRU and CatBoost methods. First, VMD was employed to decompose power load data into contrast intrinsic mode functions (IMFs), and weather characteristic factors were added for each intrinsic mode. Then, the TCN was utilized to extract multidimensional data features, and the extracted features were sent to the BiGRU network for model training. The influence degree of necessary information can be effectively retained by adding an attention mechanism. The model adopts the improved WOA (IWOA) algorithm to optimize the hyperparameter selection of the TCN-BiGRU-attention network, which designs the parameter selection of the network layer of the model more effectively and predicts the stationary component of the sequence in parallel with the CatBoost network. The parallel prediction results of two-layer networks were combined with the mean absolute percentage error-reciprocal weight (MAPE-RW) algorithm. The model accuracy was verified by utilizing commenced loading data from an Australian region, and the model performance was evaluated using the root mean square error (RMSE) and mean absolute percentage error (MAPE). The contributions of this paper can be summarized as follows:

- 1) A multi-model fusion load forecasting method is proposed. First, the IWOA-WTCN-BiGRU-attention prediction model of VMD was constructed as Model one. Second, the CatBoost prediction model based on a random search algorithm was adopted as Model two. Finally, the MAPE-RW algorithm has been utilized to fuse the load prediction results of the two models to achieve an accurate prediction of the power load value.

- 2) The Morlet wavelet function is used to improve the TCN. The Morlet wavelet basis function is introduced as residual block activation function of the TCN.

- 3) An improved WOA is proposed. The traditional WOA is improved by introducing a nonlinear convergence factor and adaptive weight, which improves the convergence speed and the convergence accuracy of the algorithm.

The paper has been organized in the following way. In Section 2, the network principles employed in the multi-model structure are introduced. In Section 3, the load prediction model structure of the multi-model fusion network is proposed. In Section 4, the experimental validation of the multi-model fusion network is performed and the experimental results are analyzed. Finally, the whole paper is summarized and future work to be carried out is presented.

2. Multi-model fusion analysis

2.1. VMD network

As mentioned in the literature [24], in 2014, Konstantin et al. proposed a VMD method for modal decomposition, which is an adaptive and completely non-recursive modal decomposition processing method. VMD has a more solid theoretical basis and can better suppress mode aliasing by controlling bandwidth. The decomposition method is suitable for non-stationary sequence data and can decompose the data set into multiple stationary sub-sequences with different frequency scales. The VMD solution procedure is as follows.

Structural constraint variational optimal problem is

$$\begin{cases} \min_{\{u_k\}, \{\omega_k\}} \left\{ \sum_{k=1}^K \left\| \partial_t \left[\left(d(t) + \frac{j}{\pi t} \right) g u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \\ s.t. \sum_{k=1}^K u_k(t) = S \end{cases} \quad (2.1)$$

where $\{u_k\}$, $\{\omega_k\}$ denote the corresponding modal set and center frequency after VMD decomposition respectively, and K is the number of IMFs.

The penalty factor α and Lagrange operator λ are introduced into the constrained variational problem to transform it into the following unconstrained variational problem:

$$\begin{aligned} L(u_k, \omega_k, \lambda) = & \alpha \sum_{k=1}^K \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) \cdot u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \\ & + \left\| f(t) - \sum_{k=1}^K u_k(t) \right\|_2^2 + \left\langle \lambda(t), f(t) - \sum_{k=1}^K u_k(t) \right\rangle \end{aligned} \quad (2.2)$$

The above unconstrained variational problems are solved by the alternating direction multiplier method, and the solving process is as follows:

$$\sum_{k=1}^K \left\| \hat{u}_k^{n+1} - \hat{u}_k^n \right\|_2^2 / \left\| \hat{u}_k^n \right\|_2^2 < \gamma \quad (2.3)$$

where γ is the allowable error, n is the number of iterations and the Fourier transform of \hat{u}_k^{n+1} , $f(t)$, $\lambda(t)$ is $\hat{u}_k^{n+1}(\omega)$, $\hat{f}(\omega)$, $\hat{\lambda}^n(\omega)$.

2.2. WTCN

The TCN was first proposed by Bai et al. [25]. in 2018 and is mainly employed for timing prediction, probability prediction, time prediction and traffic prediction. The TCN evolved from CNN results and can extract load data features effectively. In this paper, a multi-model fusion network is introduced, and a TCN is employed to extract data feature information from time series to remove invalid features to improve the accuracy of power load prediction. The TCN is composed of causal convolution, expansion convolution and a residual block [26].

Causal convolution adopts the one-dimensional full convolutional network framework. The zero-fill module is introduced into the network so that the input layer, hidden layer and output layer can keep the same length, to avoid the loss of effective information. The input y_t is related only to the input $(x_{t-1}, x_{t-2}, \dots, x_{t-n})$ before the current input x_t and t . The convolutional calculation is shown in Figure 1.

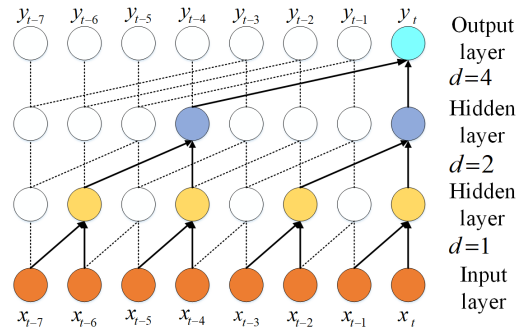


Figure 1. Convolutional calculation process.

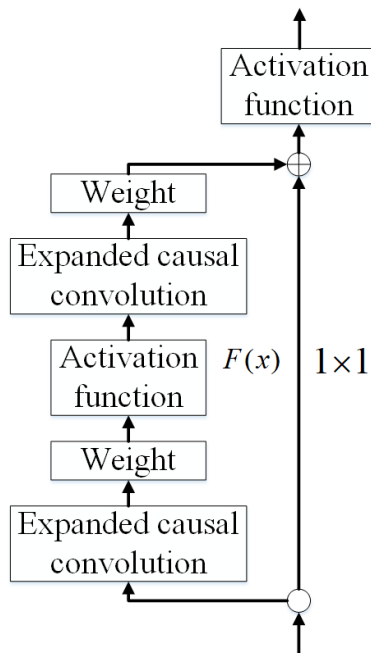


Figure 2. Structure of residual block.

Expansion convolution can increase the receptive field size of the output unit without increasing the number of parameters. The convolutional calculation method is as follows:

$$F(s) = (x \otimes f_d)(s) = \sum_{i=0}^{k-1} f(i) \cdot x_{t-d-i} \tag{2.4}$$

where f_d is the expansion rate, d corresponds to the filter and x_{t-d-i} is the input at the current time and the historical time.

The core idea of residual block is to introduce one or more layers of “hop connection” operation, and the network structure is shown in Figure 2. The left channel introduces weighted normalized accelerated gradient descent and a nonlinear activation function. The right channel is the convolution directly connected to the edge, which ensures that the input and output data dimensions are consistent. The residual block output is

$$h(x) = Activation(x + F(x)) \tag{2.5}$$

where x , $h(x)$ is the input and output of the residual block. The network output $h(x)$ is the result of linear transformation and activation function mapping. The WTCN is based on TCN topology, and the Morlet wavelet basis function is introduced into the residual block as its activation function. The Morlet wavelet basis function is expressed as

$$y = \cos(1.75x)e^{-x^2/2} \tag{2.6}$$

2.3. Bidirectional GRU network

The GRU network is simpler than the LSTM network with two gating units. The network inherits the advantages of the LSTM network and improves the training speed on the premise of ensuring training accuracy [27]. The GRU network structure is shown in Figure 3. By changing the GRU network into a bidirectional GRU network, information can be transmitted bidirectionally in the network layer, and the prediction accuracy of the network model is effectively improved [28]. The network structure is shown in Figure 4.

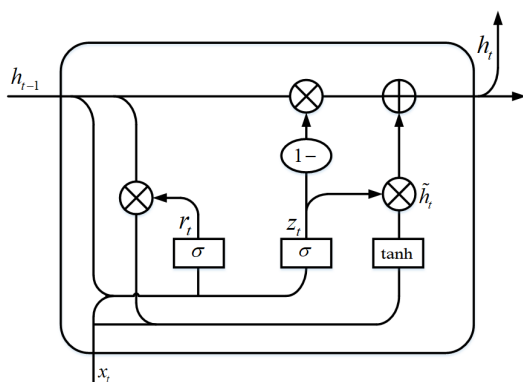


Figure 3. GRU network structure.

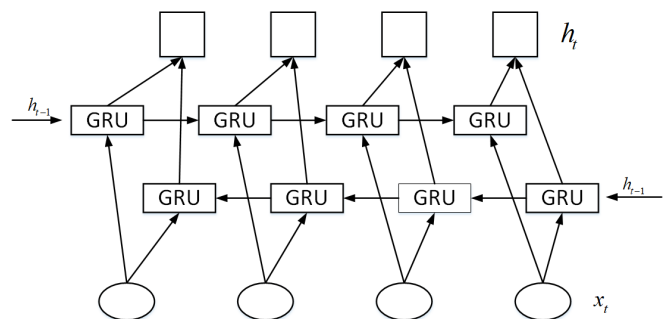


Figure 4. BiGRU network model.

The BiGRU network calculation formula is

$$\begin{cases} z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \\ r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \\ \tilde{h}_t = \tanh(W_{\tilde{h}} \cdot [r_t * h_{t-1}, x_t]) \\ h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \end{cases} \quad (2.7)$$

$$\begin{cases} h_t = GRU(x_t, h_{t-1}) \\ h_i = GRU(x_i, h_{i-1}) \\ h_t = w_i h_t + v_i h_i + b_t \end{cases} \quad (2.8)$$

where z_t is the update gate and r_t is the reset gate, both of which are jointly determined by the input x_t , hidden layer output x_t at the previous moment and activation function σ . h_t is the hidden layer output. W_z , W_r and $W_{\tilde{h}}$ are all trainable parameter matrices. h_t is the state of the forward hidden layer, h_i is the state of the backward hidden layer and b_t is the bias optimization parameter of the hidden layer at the current time.

2.4. Attention mechanism

The attention mechanism is an intuitive interpretation method that imitates human visual mechanisms. It is often exploited in deep learning tasks such as natural language processing, image analysis and load prediction. The human visual mechanism will pay attention to the critical information of the object deliberately and ignore the irrelevant information. Consequently, it has been found that the relevant time-series information can be effectively preserved by adding an attention mechanism and weight allocation principle in the network model [29]. The structure of attention is shown in Figure 5.

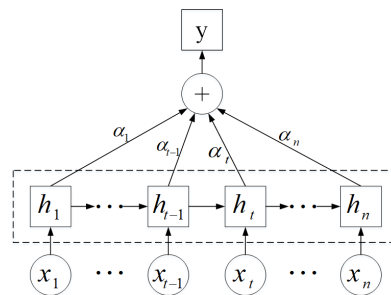


Figure 5. Attention mechanism.

2.5. IWOA

The WOA is a bionic meta-heuristic optimization algorithm proposed by Australian scholars Mirjalili and Lewis in recent years based on the predation behavior of model humpback whales. The algorithm highlights the local search behavior of the network model by imitating the whale hunting behavior and realizes the global search of the network through the random search strategy. The WOA has the advantages of faster speed and higher precision in solving model parameter optimization, so it

has wide application prospects. Nevertheless, the increase of power load data and influencing factors may cause the traditional WOA to have some limitations in the coordination of global search and local mining. Among them, the convergence factor a of the WOA cannot reflect the optimization process well with the linear decrease. Therefore, the nonlinear convergence factor a is proposed:

$$a = 2 - 2 \sin\left(u \frac{t}{\max_iter} \pi + \varphi\right) \quad (2.9)$$

where u and φ are the set parameters and $u = 2$ and $\varphi = 0$ represent. \max_iter is the maximum number of iterations. When the value is large at the initial stage of training, the searching range of optimal parameters can be effectively increased by the slowly decreasing convergence factor. With the increase in the number of iterations, the reduction speed of the convergence factor gradually increases and the convergence speed accelerates.

The introduction of the nonlinear factor a can improve the performance of the algorithm. However, in the traditional WOA, the whale motion position vector is not effectively utilized, so the population flexibility will be reduced and the optimization result will be affected. In this paper, the adaptive weight ω is introduced to enhance the global search capability of the algorithm and increase the total group diversity of the WOA. The formula for calculating ω is as follows:

$$\omega = 0.2 \cos\left(\frac{\pi}{2} \left(1 - \frac{t}{\max_iter}\right)\right) \quad (2.10)$$

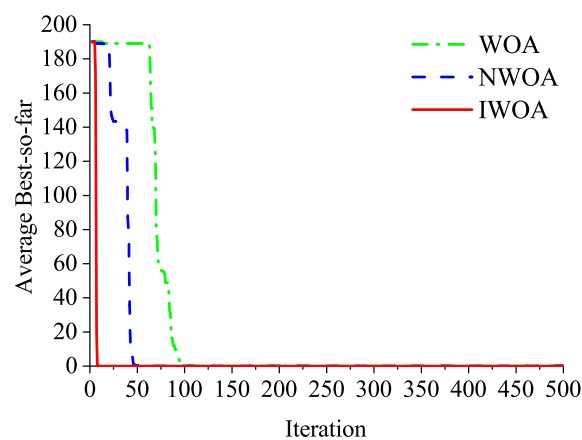


Figure 6. WOA convergence curve was tested by using reference functions.

In this paper, the performance of the IWOA is verified by introducing the benchmark test function $f(x) = \sum_{i=1}^n [x_i^2 - 10 \cos(2\pi x_i) + 10]$. The number of iterations of the algorithm was set to 500 and the dimension of the base test function is 30. To ensure the reliability of the optimization results, the average of 10 experimental results is employed to indicate its average level. The WOA, improved nonlinear convergence factor (NWOA) and IWOA with adaptive weights and a nonlinear convergence factor are compared for algorithm performance. The experimental results are shown in Figure 6. It is known that the WOA with improved adaptive weights and a nonlinear convergence factor (IWOA) not only improves the convergence speed of the algorithm, but also improves the convergence accuracy. The flow chart of the improved WOA is shown in Figure 7.

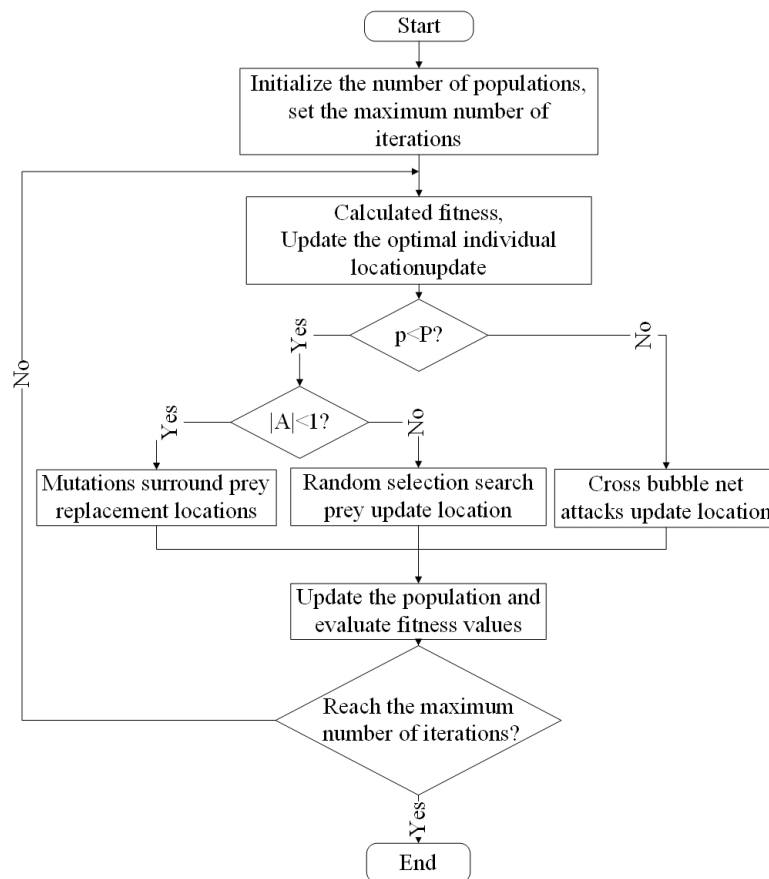


Figure 7. Flow chart of WOA.

2.6. CatBoost network

CatBoost is a machine learning library that the Russian search giant Yandex opened source in 2017, and it is an improvement on the gradient boosting decision tree (GBDT) algorithm [30]. The CatBoost algorithm has fewer parameters than the GBDT algorithm. The algorithm effectively solves the problems of gradient deviation and prediction deviation, reduces the risk of model overfitting and improves the generalization ability of the algorithm. CatBoost algorithms are often used in data mining and load forecasting.

Noise and low-frequency data interference can be effectively reduced by adding prior distribution terms to the gradient decision tree. Its algorithm is as follows:

$$\hat{x}_k^i = \frac{\sum_j^{p-1} [x_{\sigma_j,k} = x_{\sigma_i,k}] \cdot Y_{\sigma_j} + a \cdot p}{\sum_j^{p-1} [x_{\sigma_j,k} = x_{\sigma_i,k}] + a} \quad (2.11)$$

where σ represents the weight coefficient of the prior term, and p represents the prior term. The CatBoost usually captures the mean of the data set as the first item when solving regression problems.

2.7. MAPE-RW algorithm principle

The MAPE-RW algorithm can fuse disagreement models according to the degree of prediction error and output the optimal prediction results. The proportion of the predicted value of each model was determined by finding the optimal weight. The final predicted value calculated by the algorithm is as follows:

$$\begin{cases} \omega_i = \frac{M_j}{M_i+M_j} \\ f_{final} = \omega_{VTWBA}f_{Catboost} + \omega_{Catboost}f_{VTWBA} \end{cases} \quad (2.12)$$

where ω_i is the corresponding model weight, and f_{final} is the final prediction output of multi-model fusion. $f_{Catboost}$ and f_{VTWBA} are the predicted outputs of the CatBoost model and VMD-decomposed WTCN-IWOA-BiGRU-attention model, respectively.

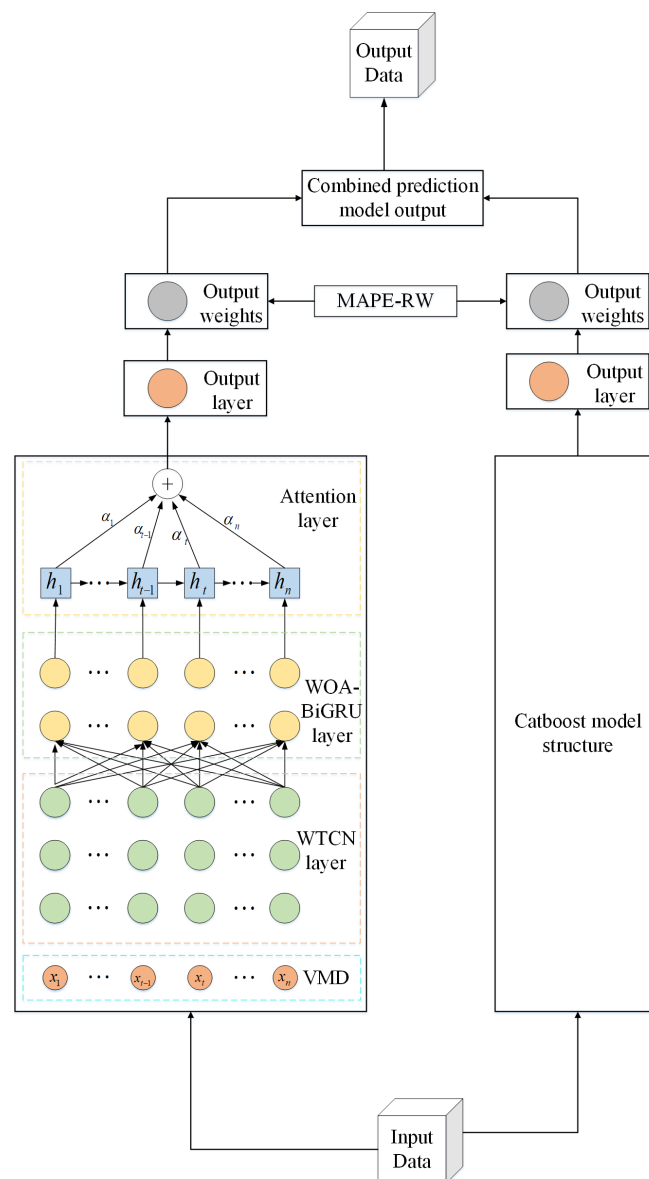


Figure 8. Multi-model fusion network load forecasting model.

3. Multi-model fusion network load forecasting model

There is a lot of power load influencing factors in time-series data, so the traditional prediction model cannot extract the data feature law effectively. In this paper, a multi-model fusion short-term power load forecasting model is proposed by combining a deep learning algorithm and a machine learning algorithm. Combined with the advantages of different algorithms, the characteristic information between data can be effectively mined to improve the accuracy of load prediction. The prediction model design of the multi-model fusion network is shown in Figure 8.

3.1. Structure of prediction model

1) Data processing. The model validation analysis is carried out by employing a public power load data set from 2006 to 2010 in an Australian region. This data set contains six-dimensional feature vectors, and the feature parameters are shown in Table 1 below. The data set captures 30 min as the sampling point, and load prediction is carried out by using a sliding window, with a sliding window size of 10 and sliding step size of 1. Therefore, 10 sets of historical data are employed to predict the electric load value at the next moment. The data set is divided into the training set, verification set and test set according to 3:1:1. The multi-fusion network model employs the verification set to conduct parameter tuning in the training process. In order to make the network model evaluation more accurate, the test set will not participate in the network model training.

Table 1. Characteristic parameters.

Characteristic parameters	Parameter types	Character descriptions
Data	Time	Samples were taken every 30 minutes
	Dew-point humidity	Equilibrium temperature
Weather factors	Dry-bulb humidity	Aerothermodynamic temperature
	Wet-bulb temperature	Thermodynamic saturation temperature
	Humidity	Degree of atmospheric dryness
Economic factors	Degree	Price per kWh

2) Input layer. Characteristic data and power load data are exploited as input for the prediction model. The input data with length n is filled with missing values and normalized into the prediction model.

3) VMD layer. The power load data are employed as the input of the prediction model. The long time-series data were input into the prediction Model one after missing-value filling and normalization. In the VMD, the values of k and α are determined by the central frequencies in the decomposition. The value of the central frequency is calculated by changing the values of k and α . By choosing a reasonable value of k , the phenomenon of model mixing can be avoided, and fewer network parameters for the WOA-based model one can also be generated. As can be seen from Table 2, at $k = 5$ and $\alpha = 1850$, the central frequency has been relatively stable with the least number of decomposition layers, which makes further training produce fewer parameters and improves the model training speed. The

penalty factor of decomposition of the variational model $\alpha = 1850$, the tolerance difference of collection $tol = 1e - 7$ and the number of decomposition modes $k = 5$ are set. The decomposition of each mode is shown in Figure 9.

Table 2. Center frequencies corresponding to different values of k .

k	Simulated signal center frequency					
2	0.0001	0.02243				
3	0.0001	0.02080	0.04274			
4	0.0001	0.02076	0.04167	0.06471		
5	0.0001	0.02076	0.04162	0.06469	0.34702	
6	0.0001	0.02076	0.04162	0.06238	0.08467	0.35872

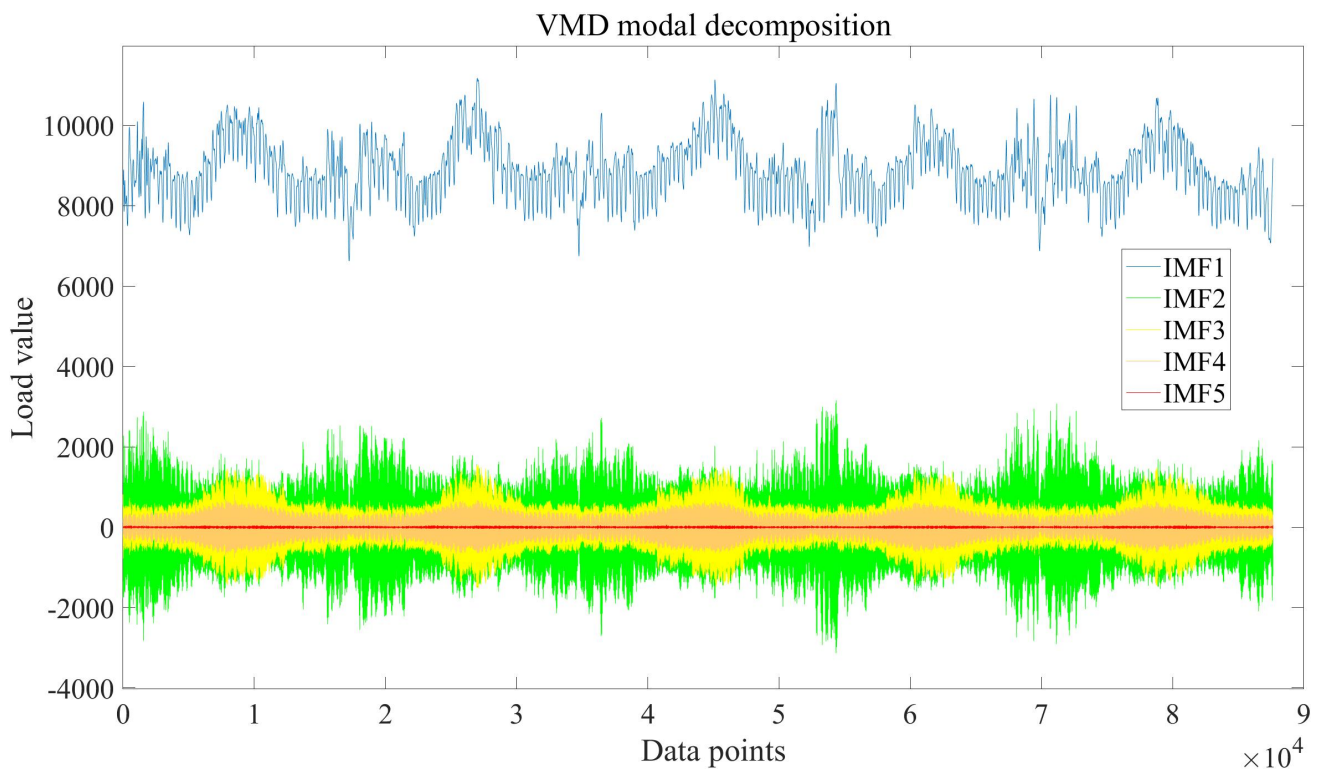


Figure 9. Results of VMD.

4) IWOA-based hyperparameter optimization. The optimal hyperparameters are obtained by employing WOA, and in order to perform the optimization search within the range of valid parameter selections, the range of network parameter values selected has been defined as shown in Table 3. The components and weather characteristics generated from the decomposition of load information by VMD are the input of the WTCN-BiGRU-attention network, respectively. And, the IWOA is employed to optimize the network hyperparameters. The optimal network hyperparameter search structure of each component is shown in Table 4.

Table 3. Parameters of IWOA.

Parameters	Parameter values	
Parameter settings	Population size	5
	Max iterations	5
	Constant b	2
	Learn rate	[0.001,0.01]
	Epoch	[10,100]
Search for upper and lower bounds	Batchsize	[16,128]
	BiGRU node number	[1,20]
	Number of nodes at the full	[1,100]

Table 4. Optimal parameter selection for power load forecasting.

Modal tags	IMF1	IMF2	IMF3	IMF4	IMF5
Learn rate	0.0047	0.0027	0.0064	0.0054	0.0024
Epoch	38	26	61	70	89
Batchsize	72	19	72	37	22
BiGRU node number	8	18	8	17	2
BiGRU node number	16	11	7	17	19
Number of nodes at the full	49	4	46	14	97

Table 5. Power load forecasting for CatBoost network hyperparameters.

Verbose	Learning rate	Iterations	Depth
50	0.03	900	12

5) WTCN layer. The influential factors of the load characteristics were added to the modes decomposed by the VMD layer, respectively. The Morlet wavelet function is used as a residual block activation function. The network extracts load characteristics and influencing factors through the WTCN layer. It normalizes the weight of the convolutional kernel. The dropout coefficient can be set to 0.2 to prevent over-fitting of the model. We set the expansion coefficient as (1, 2, 4, 8, 16, 32). We set the number of filters to 128.

6) BiGRU layer. The model builds two BiGRU layers to learn the features extracted from the WTCN, design full utilization of the data features and capture its internal change rules.

7) Attention layer. The input of the attention mechanism is the output data activated by the two-layer BiGRU network. The corresponding proportions of disagreement feature vectors are calculated according to the weight allocation principle, and the optimal weight parameter matrix is searched by using continuous updates and iteration.

8) CatBoost prediction model. A random search algorithm is employed to select the CatBoost network hyperparameters. The optimal network hyperparameters are shown in Table 5. The input power load and weather characteristic factors are modeled.

9) Output layer. The IWOA-WTCN-BiGRU-attention network is set as Model one, and the Cat-Boost network is set as Model two. The MAPE-RW algorithm was exploited to calculate the weight of the output results of Model one and Model two. Finally, the load prediction output of the multi-model fusion network is obtained by effective fusion of the model prediction results.

3.2. Model one loss function

Adam's optimization algorithm was selected as the parameter optimization method of network Model one. Adam is a first-order optimization algorithm that can effectively replace the traditional gradient descent process. The algorithm can update and iterate the weight of the network according to the data so that the loss function can be optimized. The loss function of the model is calculated by employing the mean square error, and its formula is

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (3.1)$$

where N is the number of samples; y_i and \hat{y}_i are the actual load value and predicted load value of model i , respectively.

4. Example analysis

4.1. Data normalization

The minimum-maximum normalization method is exploited to normalize the original data and increase the training speed of the model. The inverse normalization of the predicted data designs the comparison between the predicted value and the real value more intuitive. Its calculation formula is

$$x_n = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (4.1)$$

where x is the original load data. x_{\max} and x_{\min} are, respectively, the maximum value and minimum value of the sample data. x_n is the normalized data.

4.2. Model evaluation index

The RMSE, MAPE, mean absolute error (MAE) and R-square were utilized as evaluation indexes. The calculation formulas are as follows:

$$\left\{ \begin{array}{l} RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\tilde{x}_i - x_i)^2} \\ MAPE = \frac{100}{N} \sum_{i=1}^N \left| \frac{\tilde{x}_i - x_i}{P_0} \right| \\ MAE = \frac{1}{n} \sum_{i=1}^n |\tilde{x}_i - x_i| \\ R - square = 1 - \frac{\sum_{i=1}^n (\tilde{x}_i - x_i)^2}{\sum_{i=1}^n (\tilde{x}_i - \bar{\tilde{x}}_i)^2} \end{array} \right. \quad (4.2)$$

where N is the number of samples. \tilde{x}_i is the true value of the sample point i . x_i is the predicted value of the i_{th} sample point.

4.3. Analysis of model prediction results

4.3.1. Electricity load data forecast

The prediction results of the proposed model were compared with those of the traditional single model or mixed deep learning models such as GRU, LSTM, TCN, WTCN, WTCN-GRU, WTCN-LSTM, TCN-BiGRU, WTCN-BiGRU and TCN-BiGRU-attention models. The results of the load forecast data for December 20, 2010, are plotted to show more visually the accuracy advantages of the power load forecast model proposed in this paper. The load forecasting curves are shown in Figure 10. Different curves represent the prediction results and trends of disagreement prediction models. As can be seen from the prediction trends of the disagreement models shown in Figure 10, the prediction results of the prediction model proposed in this paper are more accurate, stable and closer to the real load. Table 6 shows the test lumped prediction and evaluation indexes of each model.

Table 6. Total evaluation index results of the electricity load forecasting models.

Prediction model	RMSE	MAPE	MAE	R-square
IWOA-WTCN-BiGRU-attention	110.964	0.914	82.189	0.993
CatBoost	89.089	0.727	63.407	0.996
Fusion model of this paper	77.495	0.632	56.103	0.997
WTCN-BiGRU-attention	103.452	0.86	76.179	0.994
WTCN-BiGRU	97.573	0.803	69.903	0.995
WTCN-GRU	96.584	0.814	70.868	0.995
WTCN-LSTM	100.309	0.843	73.649	0.995
WTCN	103.558	0.869	75.975	0.994
TCN	103.969	0.89	78.024	0.994
LSTM	147.665	1.261	108.342	0.988
GRU	109.991	0.932	82.114	0.994
CNN	337.913	3.241	275.52	0.939

Therefore, the forecasting effects of different models were evaluated by analyzing the effects of the forecasting data of different models every month. Table 7 shows the error values of the monthly forecasting results of the different models. The analysis of the model evaluation metrics shows that the smaller the error values of MAPE, MAE and RMSE, the better the forecasting performance of the models. The larger the R-square value, the closer the predicted value is to the real value. After analyzing the data in Table 6 and Figure 10, the following conclusions can be drawn:

1) Compared with VMD-IWOA-WTCN-BiGRU-attention and CatBoost alone, the prediction results are more accurate by combining the proposed multiple models. The RMSE decreased by 33.469 and 11.594, MAPE decreased by 0.282 and 0.095% and MAE decreased by 26.086 and 7.304, respectively. By analyzing the reasons, it can be seen that VMD-IWOA-WTCN-BiGRU-attention has a large

prediction deviation in load fluctuation hours because VMD caused a loss of part of the data. The CatBoost model is more accurate in the prediction of stationary components, but the prediction deviation is larger when the data fluctuation is larger. Therefore, the MAPE-RW algorithm is used to integrate the advantages of the two models to create a prediction effect that is more accurate.

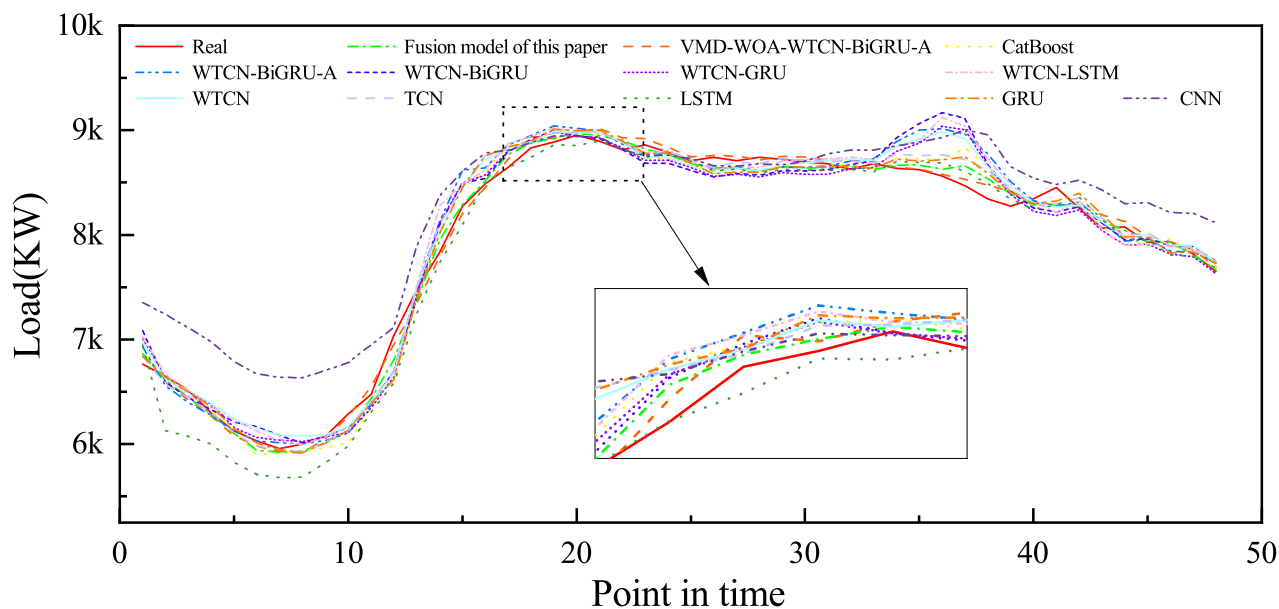


Figure 10. Results of the load forecast data for December 20, 2010.

Table 7. Each model predicts the evaluation index.

RMSE	Model one	Model two	Fusion model	WT-BG-a	WT-BG	WT-G	WT-L	WT
Jan.	108.372	106.039	97.048	126.77	116.382	96.321	107.057	121.456
Feb.	102.257	84.791	76.02	108.625	89.625	90.072	88.652	103.947
Mar.	85.485	72.585	61.48	84.297	76.175	80.666	85.047	86.109
Apr.	96.362	106.033	83.271	105.568	110.002	108.207	109.236	113.008
May	121.924	95.097	80	101.731	101.456	101.066	100.091	108.432
Jun.	141.496	86.97	82.599	100.258	91.622	93.661	98.118	104.279
Jul.	131.409	86.787	77.983	101.809	87.982	92.821	95.962	100.363
Aug.	170.315	84.826	102.867	105.915	103.802	105.615	112.917	107.68
Sept.	127.524	83.933	81.581	104.565	99.724	99.005	101.172	100.071
Oct.	85.224	89.977	65.933	95.965	93.855	99.785	101.756	97.116
Nov.	79.742	84.105	62.097	99.21	95.347	101.174	106.564	100.065
Dec.	79.927	82.657	63.727	103.142	101.035	92.024	101.432	101.644

WT-BG-a: WTCN-BiGRU-attention; WT-BG: WTCN-BiGRU; WT-G: WTCN-GRU; WT-L: WTCN-LSTM; WT: WTCN.

2) Compared with other independent prediction models, the prediction results of the model proposed in this paper are closer to the real value. Compared with the WTCN-BiGRU prediction model, the RMSE decreased by 20.078, MAPE decreased by 0.171% and MAE decreased by 13.4. It can be seen that the algorithm based on the bottom combination model also achieves good training results, but it has the disadvantage of low prediction accuracy.

4.3.2. Photovoltaic power generation data forecasts

To verify the feasibility and accuracy of the model in different forecasting areas. We employed the 2018 annual measured power generation of the domestic Ningxia Wuzhong Sun Mountain photovoltaic (PV) power plant for PV power generation prediction, as well as five environmental data types, i.e., total solar irradiation, PV panel module temperature, ambient temperature, atmospheric pressure and relative humidity, measured by the environmental detector corresponding to this PV array. The data sets were collected at 15-minute intervals. Since the PV array only emits energy during the daytime, the valid data of the daily 7:30–16:30 PV-emitted power were selected as the model validation data. The data are divided according to the ratio of 10:1:1, and the first 10 months are taken as the training data, November has been applied as the validation data during training and December data as the test set. The prediction model parameters are set in the same way as the power load prediction model parameters. The WTCN-BiGRU-attention network hyperparameters were selected by the WOA algorithm for the PV power generation prediction model, as shown in Table 8, and the best network hyperparameters were selected by the random search algorithm for the CatBoost network, as shown in Table 9.

Table 8. Selection of hyperparameters for the PV power prediction network.

Modal tags	IMF1	IMF2	IMF3	IMF4	IMF5
Learn rate	0.0038	0.0012	0.0059	0.0067	0.0030
Epoch	72	54	50	73	96
Batchsize	32	57	118	117	70
BiGRU node number	5	1	16	10	3
BiGRU node number	6	7	15	13	12
Number of nodes at the full	44	12	54	84	12

Table 9. PV power forecasting CatBoost network hyperparameters.

Verbose	Learning rate	Iterations	Depth
50	0.05	500	9

The prediction results of the model are compared with the hybrid neural network models WTCN-BiGRU-attention, TCN-BiGRU-attention and CNN-BiGRU-attention. The daily power generation forecasting results of PV panels for two days are plotted to show more intuitively the advantages of the multi-model fusion forecasting network proposed in this paper. The accuracy of the load curve prediction results is shown in Figures 11 and 12, where different curves represent the prediction results and trends of different models. From the figures, it can be seen that the proposed multi-modal fusion

forecasting network has higher accuracy. Table 10 shows the total prediction evaluation index of each model test set.

Table 10. PV power generation forecast model evaluation index.

Prediction model	RMSE	MAPE	MAE	R-square
Model one	1.996	24.380	1.190	0.964
Model two	1.434	13.119	1.023	0.982
Fusion model of this paper	1.285	12.285	0.924	0.985
WTCN-BiGRU-attention	3.737	26.8	2.315	0.875
TCN-BiGRU-attention	3.739	23.808	2.466	0.875
CNN-BiGRU-attention	3.738	26.243	2.365	0.875

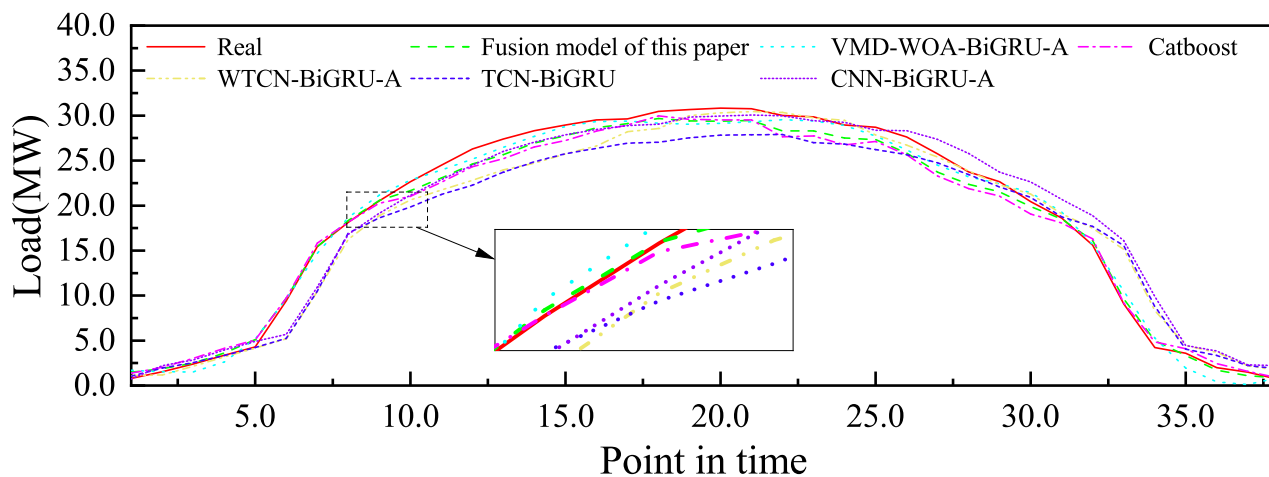


Figure 11. December 10, 2018 PV power forecast.

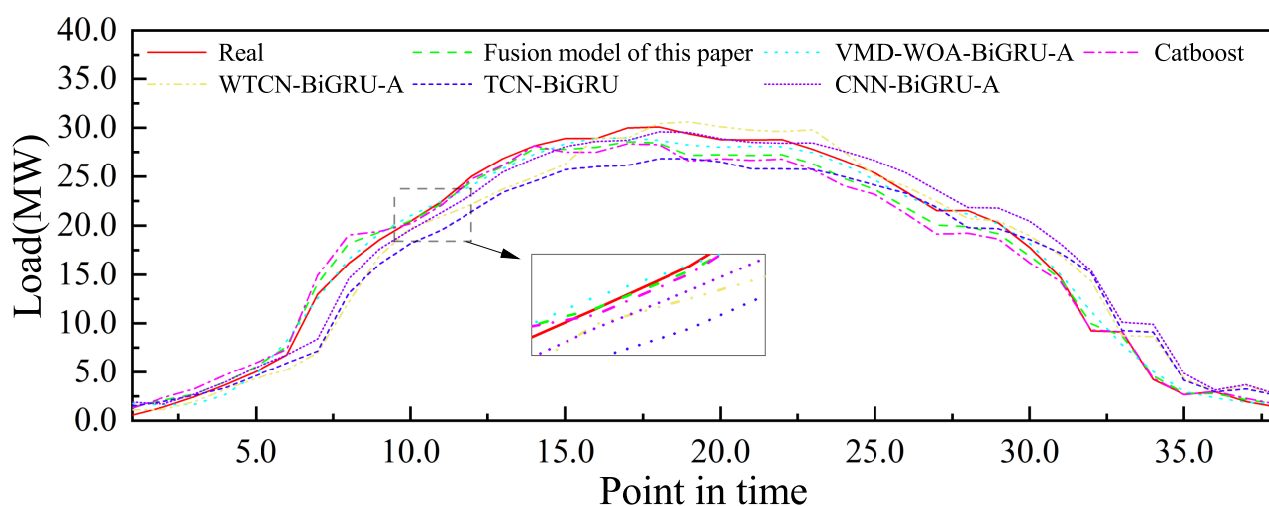


Figure 12. December 18, 2018 PV power forecast.

To sum up, the multiple models proposed in this paper combined with the short-term power load prediction model have more outstanding prediction performance, and the prediction results are relatively more stable, meaning that the model can be better used to predict the power load data with multidimensional feature inputs.

5. Conclusions

The major objective of this study was to build a forecasting model by integrating multiple models to improve the accuracy of power load forecasting. In Model one, decompose the data into multiple components by VMD decomposition. Then, an IWOA is exploited to optimize the super parameters of the WTCN-BiGRU-attention network model. At the same time, Model two is designed for the parallel prediction of multi-dimensional load data by the CatBoost algorithm. Finally, the MAPE-RW algorithm is employed to fuse the prediction results of the two models to achieve accurate and personal measurements of short-term power load data. Taking multidimensional load data of an area in Australia as a model example, the feasibility verification analysis was carried out, and the main conclusions are as follows:

1) Based on the power load data of a certain region in Australia, we constructed a multi-dimensional power load feature set to better predict the non-stationary components with strong fluctuation of power load data, i.e., Model one.

2) The stationary components of multi-dimensional power load data are predicted by Model two, and the model prediction results are fused by the MAPE-RW algorithm, which improve the power load prediction accuracy of multi-model fusion.

To sum up, the hybrid neural network combined model for multidimensional characteristic power load data prediction has been proposed in this paper. This research sheds new light and not only provides reference and choices for short-term power load forecasting methods, but also has good reference significance for other power fields, such as wind power generation forecasting and energy storage unit service-life forecasting. However, the structure of the multi-model fusion network is overly complex, which increases the model prediction time and wastes computer resources while improving the accuracy of power load prediction. Therefore, in the future, the authors will work on designing a more concise interval prediction model for power loading with suitable accuracy. While overcoming the training time, the interval prediction makes the prediction results more meaningful for the power sector to conduct power dispatching and effectively avoid the waste of power resources.

Acknowledgments

This work was supported by the State Grid Corporation of China Headquarters Science and Technology Project (5400-202122573A-0-5-SF). The authors thank the editors and the anonymous reviewers for their helpful comments and suggestions that have improved the presentation of this manuscript.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. X. Shan, X. Lu, M. Y. Zhai, Z. Gao, C. Xu, X. Teng, et al., Analysis of key technologies for artificial intelligence applied to power grid dispatch and control, *Autom. Electr. Power Syst.*, **43** (2019), 49–57.
2. S. Fan, L. Li, S. Wang, X. Liu, Y. Yu, B. Hao, Application analysis and exploration of artificial intelligence technology in power grid dispatch and control, *Power Syst. Technol.*, **44** (2020), 401–411.
3. C. Bian, S. Liu, H. Xing, Y. Jia, Research on fault-tolerant operation strategy of rectifier of square wave motor in wind power system, *CES Trans. Electr. Mach. Syst.*, **5** (2021), 62–69. <https://doi.org/10.30941/cestems.2021.00008>
4. F. L. Tan, J. Zhang, H. Z. Ma, Combined forecasting method of power load based on trend change division, *J. North China Electr. Power Univ.*, **47** (2020), 17–24.
5. P. Y. Chen, Y. J. Fang, Short-term load forecasting of power system for holiday point-by-point growth rate based on Kalman filtering, *Eng. J. Wuhan Univ.*, **53** (2020), 139–144.
6. B. Li, F. Qin, Y. Wu, J. Huang, Short-Term daily load curve forecasting based on fuzzy information granulation and multi-strategy sensitivity, *Trans. China Electrotech. Soc.*, **32** (2017), 149–159.
7. D. X. Niu, S. Y. Dai, A short-term load forecasting model with a modified particle swarm optimization algorithm and least squares support vector machine based on the denoising method of empirical mode decomposition and grey relational analysis, *Energies*, **10** (2017), 408–428. <https://doi.org/10.3390/en10030408>
8. Q. Liu, Z. Z. Huang, S. Li, Research on power load forecasting based on support vector machine, *J. Balkan Tribol. Assoc.*, **22** (2016), 151–159.
9. C. B. Li, S. K. Li, Y. Q. Liu, A least squares support vector machine model optimized by moth-flame optimization algorithm for annual power load forecasting, *Appl. Intell.*, **45** (2016), 1166–1178. <https://doi.org/10.1007/s10489-016-0810-2>
10. J. J. Sun, S. Z. Zhang, M. D. Zeng, Multi-objective optimal control for flexible load in active distribution network considering time-of-use tariff, *Trans. China Electrotech. Soc.*, **33** (2018), 401–412.
11. S. Zhang, J. X. Yang, J. C. Liu, J. Y. Liu, Power load recovery based on multi-scale time-series modeling and estimation, *Trans. China Electrotech. Soc.*, **35** (2020), 2736–2746.
12. M. Tan, S. P. Yuan, S. H. Li, Y. Su, H. Li, F. He, Ultra-short-term industrial power demand forecasting using LSTM based hybrid ensemble learning, *IEEE Trans. Power Syst.*, **35** (2020), 2937–2948. <https://doi.org/10.1109/TPWRS.2019.2963109>
13. J. F. Rendon-Sanchez, L. M. de Menezes, Structural combination of seasonal exponential smoothing forecasts applied to load forecasting, *Eur. J. Oper. Res.*, **27** (2019), 916–924. <https://doi.org/10.1016/j.ejor.2018.12.013>
14. J. Lu, Q. Zhang, Z. Yang, M. Tu, J. Lu, H. Peng, Short-term load forecasting method based on CNN-LSTM hybrid neural network model, *Autom. Electr. Power Syst.*, **43** (2019), 131–137.

15. B. Zhao, Z. Wang, W. Ji, X. Gao, X. Li, A short-term power load forecasting method based on attention mechanism of CNN-GRU, *Power Syst. Tech.*, **43** (2019), 4370–4376.
16. L. J. Zhu, Z. Xun, Y. X. Wang, Short-term power load forecasting based on CNN-BiLSTM, *Power Syst. Tech.*, **45** (2021), 4532–4539.
17. Y. Zhao, H. Wang, L. Kang, Z. Zhang, Temporal convolution network-based short-term electrical load forecasting, *Trans. China Electrotech. Soc.*, **5** (2022), 1243–1251.
18. Z. D. Tian, H. Chen, A novel decomposition-ensemble prediction model for ultra-short-term wind speed, *Energy Convers. Manage.*, **248** (2021), 1–18. <https://doi.org/10.1016/j.enconman.2021.114775>
19. Z. D. Tian, Short-term wind speed prediction based on LMD and improved FA optimized combined kernel function LSSVM, *Eng. Appl. Artif. Intell.*, **91** (2020), 1–24. <https://doi.org/10.1016/j.engappai.2020.103573>
20. C. Tong, L. Zhang, H. Li, Y. Ding, Temporal inception convolutional network based on multi-head attention for ultra-short-term load forecasting, *IET Gener. Transm. Distrib.*, **16** (2021), 1680–1696. <https://doi.org/10.1049/gtd2.12394>
21. Z. D. Tian, Approach for short-Term traffic flow prediction based on empirical mode decomposition and combination model fusion, *IEEE Trans. Intell. Transp. Syst.*, **22** (2021), 5566–5576. <https://doi.org/10.1109/TITS.2020.2987909>
22. Z. D. Tian, S. J. Li, Y. H. Wang, A prediction approach using ensemble empirical mode decomposition-permutation entropy and regularized extreme learning machine for short-term wind speed, *Wind Energy*, **23** (2020), 177–206. <https://doi.org/10.1002/we.2422>
23. S. E. Haupt, S. Dettling, J. K. Williams, J. Pearson, T. Jensen, T. Brummet, et al., Blending distributed photovoltaic and demand load forecasts and deep bidirectional long short-term memory and multiple linear regression, *Sol. Energy*, **157** (2017), 542–551.
24. K. Dragomiretskiy, D. Zosso, Variational mode decomposition, *IEEE Trans. Signal Process.*, **62** (2014), 531–544. <https://doi.org/10.1109/TSP.2013.2288675>
25. S. J. Bai, J. Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, preprint, arXiv:1803.01271.
26. S. H. Rafi, N. AI Masood, S. R. Deeba, S. R. Deeba, E. Hossain, A short-term load forecasting method using integrated CNN and LSTM network, *IEEE Access*, **9** (2021), 32436–32448. <https://doi.org/10.1109/ACCESS.2021.3060654>
27. X. B. Jin, W. Z. Zheng, J. L. Kong, X. Y. Wang, Y. T. Bai, T. L. Su, et al., Deep-learning forecasting method for electric power load via attention-based encoder-decoder with Bayesian optimization, *Energies*, **14** (2021), 1406–1596. <https://doi.org/10.3390/en14061596>
28. Y. Y. Wang, J. Chen, X. Q. Chen, X. Zeng, Y. Kong, S. Sun, et al., Short-term load forecasting for industrial customers based on TCN-LightGBM, *IEEE Trans. Power Syst.*, **36** (2021), 1984–1997. <https://doi.org/10.1109/TPWRS.2020.3028133>
29. M. Lysaker, A. Lundervold, X. C. Tai, Noise removal using fourth-order partial differential equation with applications to medical magnetic resonance images in space and time, *IEEE Trans. Image Process.*, **12** (2003), 1579–1590. <https://doi.org/10.1109/TIP.2003.819229>

-
30. Y. Zhang, Q. Ai, L. Lin, S. Yuan, Z. Y. Li, A very short-term load forecasting method based on deep LSTM RNN at zone level, *Power Syst. Tech.*, **43** (2019), 1884–1892.



AIMS Press

©2022 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)