IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 April 2023, accepted 3 May 2023, date of publication 8 May 2023, date of current version 23 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3273596

RESEARCH ARTICLE

Short-Term Power Load Forecasting Based on VMD-Pyraformer-Adan

YIHAO TANG[1] AND HUAFENG CAI[1,2]

[1]School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan 430068, China
[2]Xiangyang Industrial Institute, Hubei University of Technology, Xiangyang 210023, China

Corresponding author: Huafeng Cai (whgkzj@hbut.edu.cn)

This work was supported in part by the Innovation Fund for Industry-University-Research in Chinese Universities under Grant 2021ITA05025.

ABSTRACT For the characteristics of fluctuation, periodicity and nonlinearity of power load data, this paper proposes a short-term power load forecasting model based on VMD-Pyraformer-Adan. Firstly, the variational modal decomposition (VMD) algorithm is used to modally decompose the electric load data, the over-zero rate and Pearson correlation coefficient are introduced to divide the modal components to obtain the low-frequency, mid-frequency and high-frequency parts, and the reconstructed data are formed with the original load data respectively. Secondly, the reconstructed data are input to the Pyraformer prediction network containing pyramidal attention module (PAM) and coarse-scale construction module (CSCM). Then a new momentum optimizer Adan is used to optimize the parameters of the prediction network. The final output prediction results. The experimental results show that the proposed model in the paper exhibits higher prediction accuracy compared with other models.

INDEX TERMS Short-term electric load forecasting, variational modal decomposition, pyramidal attention model, Adan optimizer.

I. INTRODUCTION

Electricity load forecasting plays an important role in the scheduling and maintenance of power grids and other aspects, and it is of great significance to deal with the challenges brought by power development [1], [2]. Therefore, power load forecasting has become an important research direction for power grid operation and maintenance.

Short-term power load data are characterized by high volatility and randomness, so the difficulty of power load forecasting is increasing, and the forecasting methods are constantly being improved and optimized. At present, the methods of electric load forecasting can be divided into three categories: mathematical statistics-based, traditional machine-learning-based and deep learning-based forecasting methods. Mathematical statistics-based methods include autoregressive model (AR) [3], autoregressive integrated moving average model (ARIMA) [4], Kalman filtering [5], and exponential smoothing [6]. Mathematical statistical methods are fast to fit, but require very high smoothness and accuracy of the data, and are less effective for fitting nonlinear series [7]. Traditional machine learning-based methods include random forest [8], support vector machine (SVM) [9], and decision trees [10]. These machine learning methods have the ability to learn nonlinear relationships in sequences, but they are not efficient and practical for analysis with large data sets, fail to take full advantage of the temporal information in load sequences, and have difficulty capturing the potential temporal dependence between outputs and inputs [11]. As an extension of machine learning, deep learning-based methods with powerful feature mining, nonlinear mapping, and adaptive capabilities include convolutional neural network (CNN) [12], deep belief network (DBN) [13], residual network (ResNet) [14], and recurrent neural network (RNN) [15]. In recurrent neural networks, long short-term memory (LSTM) [16] and gated recurrent unit (GRU) [17] are widely used. However, traditional RNN suffer from gradient disappearance and lack the ability to

The associate editor coordinating the review of this manuscript and approving it for publication was Nuno M. Garcia.

capture long-term dependencies [18]. For LSTM and GRU, problems such as dilution of historical information and loss of sequence information persist when the input sequence is too long. Based on this, the Google team proposed the Transformer model based on the attention mechanism [19], which has a stronger ability to capture long-range correlation information by learning the correlated temporal information in a sequence through its special attention mechanism. Based on the Transformer model, Liu et al. [20] proposed a low-complexity pyramidal attention model - Pyraformer for time series modeling and prediction. This method captures temporal dependencies at different scale ranges simultaneously in a multi-resolution manner through pyramidal attention, and is able to capture temporal correlations over longer distances with the same spatio-temporal complexity.

Since direct prediction of load data often fails to achieve optimal prediction, the current research trends among scholars are divided into the following three areas: (i) adding data processing methods for enhancing the characteristics of data before prediction networks; (ii) exploring more efficient and accurate prediction networks; (iii) pursuing more efficient parameter optimizers for parameter optimization. Fan et al. [21] proposed a DBN prediction model based on empirical mode decomposition (EMD). During the training process of DBN, multi-objective optimization models are constructed for accuracy and diversity, and the model parameters are optimized using the MOEA/D optimizer. But the number of EMD decompositions is unstable and will component with white noise, which increases the prediction difficulty [22]. Unlike EMD, variational mode decomposition (VMD) uses a variational model to determine the relevant frequency bands and extract the corresponding modal components with better noise immunity and theoretical basis [23]. Jia et al. [24] proposed a combined VMD-ISSA-GRU prediction model. The load data are first modally decomposed using the VMD algorithm, and then all subsequences and residuals of the VMD are predicted using the ISSA-optimized GRU network, and this method can effectively avoid the modal confounding phenomenon that occurs in EMD decomposition. Sun et al. [25] combined VMD with SG filter (Savitzky-Golay Filter) and proposed a combined VMD-SG-LSTM prediction model, where the data were noise reduced using SG filter after VMD decomposition, and then the reconstructed data were input to LSTM network for prediction to improve the model prediction accuracy.

Based on the advantages and shortcomings of the above prediction methods, a combined VMD-Pyraformer-Adan prediction model is proposed in this paper. First, the power load data are decomposed using the VMD algorithm, the over-zero rate and Pearson correlation coefficient are introduced to divide the modal components to obtain the low-frequency, mid-frequency and high-frequency parts, and the reconstructed data are formed with the original load data respectively. Secondly, the reconstructed data are fed into the Pyraformer network for training, and the parameters in the neural network are optimized using Adan optimizer to finally output the electric load prediction results. The contributions of this paper are as follows.

(1) The power load data are decomposed by VMD, and the low-frequency, medium-frequency and high-frequency parts contain different features, which are reconstructed with the original load data and input into the prediction network respectively, without causing feature loss, which can effectively improve the accuracy of the prediction model.

(2) The pyramidal attention-based Pyraformer is able to capture the long-range dependencies of the time series despite the low complexity, which further improves the prediction accuracy and convergence speed of the model in combination with the features of the Adan optimizer.

(3) A VMD-Pyraformer-Adan forecasting model is proposed to combine the advantages of each module for power load forecasting, which shows better forecasting performance when compared with other models.

This paper is organized as follows: the section II conducts theoretical analysis of each module of the model and the overall structure of the power load forecasting model proposed in this paper is stated and model evaluation metrics are proposed; the section III conducts an arithmetic analysis to verify the forecasting performance of the model on a real data set; the section IV concludes the paper with conclusions and an outlook for future work.

II. FUNDAMENTALS OF THE MODEL

A. VARIATIONAL MODE DECOMPOSITION

The VMD is a data processing method for processing non-stationary time series signals that integrates Hilbert transform, alternating direction multiplier, and Wiener filter [23], [26]. The original signal $f(t)$ is decomposed into a predefined set of K eigenmodes with different frequency characteristics u_k with the constraint that the sum of all modes is equal to $f(t)$. The variational problem of the VMD algorithm under the constraint is formulated as follows.

$$\min_{\{u_k\}\{w_k\}} \left\{ \sum_k \left\| \partial_t \left[(\delta(t) + \frac{j}{\pi t}) * u_k(t) \right] e^{-jw_k t} \right\|_2^2 \right\} \tag{1}$$

$$s.t. \quad \sum_k u_k = f \tag{2}$$

where u_k is the K modal components obtained after decomposition; w_k is the center frequency of each mode; $\delta(t)$ is the unit pulse function; $*$ is the convolution operation.

By introducing the Lagrange multiplier operator λ and the quadratic penalty factor α into equation (1) and turning it into an unconstrained variational model, we obtain.

$$\begin{aligned} L(\{u_k\}, \{w_k\}, \lambda) = \alpha \sum_k \Bigg\| \partial_t \Bigg[\left(\delta(t) + \frac{j}{\pi t} \right) \\ * u_k(t) \Bigg] e^{-jw_k t} \Bigg\|_2^2 + \left\| f(t) - \sum_k u_k(t) \right\|_2^2 \\ + \langle \lambda(t), f(t) - \sum_k u_k(t) \rangle \end{aligned} \tag{3}$$

The variational model is optimized by the alternating direction multiplier method, and u_k^{n+1}, w_k^{n+1} and λ are updated iteratively to seek the "saddle point" of equation (3) in the iterative optimization sequence, and then to find the optimal solution of equation (1). The updated equation is as follows.

$$\hat{u}_k^{n+1}(w) = \frac{\hat{f}(w) - \sum_{i \neq k} \hat{u}(w) + \frac{\hat{\lambda}(w)}{2}}{1 + 2\alpha(w - w_k)^2} \tag{4}$$

$$w_k^{n+1} = \frac{\int_0^\infty w \left|\hat{u}_k(w)\right|^2 dw}{\int_0^\infty \left|\hat{u}_k(w)\right|^2 dw} \tag{5}$$

$$\hat{\lambda}^{n+1}(w) = \hat{\lambda}^n(w) + \tau(\hat{f}(w) - \sum\nolimits_k \hat{u}_k^{n+1}(w)) \tag{6}$$

where $\hat{u}_k^{n+1}(w)$ is the Wiener filter corresponding to each modal component; $\hat{f}(w)$, $\hat{u}_k(w)$, $\hat{\lambda}^n(w)$ are the Fourier transforms of $f(t)$, $u_k(t)$, $\lambda(t)$, respectively. w_k^{n+1} is the frequency center of each modal component, τ is the noise tolerance of the signal, and n is the number of iterations. Define the maximum number of iterations as N such that n satisfies $n \leq N$. There exists any $\varepsilon > 0$, and the convergence condition equation (7) is satisfied when the iteration is completed.

$$\sum\nolimits_k \frac{\left\|\hat{u}_k^{n+1} - \hat{u}_k^n\right\|}{\left\|\hat{u}_k^n\right\|} < \varepsilon \tag{7}$$

The overall flow of VMD decomposition is shown in Figure 1.

In the VMD decomposition, a method based on the improved signal energy (ISE) is used to select the value of K. That is, the parameter K can be determined when the ratio of the residual energy to the original energy (Erse) is sufficiently small and there is no significant downward trend [27]. The equation is as follows.

$$\text{Erse} = \frac{\sum_{n=1}^N \left|f(n) - \sum_{k=1}^K u_k(n)\right|^2}{\sum_{n=1}^N f(n)^2} \times 100\% \tag{8}$$

After the VMD decomposition is completed, K intrinsic mode functions (IMF) with different frequencies are obtained, and this paper uses the over-zero rate to measure the frequency of these K mode subsequences.

$$Z = n_{zero}/N \tag{9}$$

where Z is the over-zero rate, n_{zero} is the number of over-zeros, and N is the sample length.

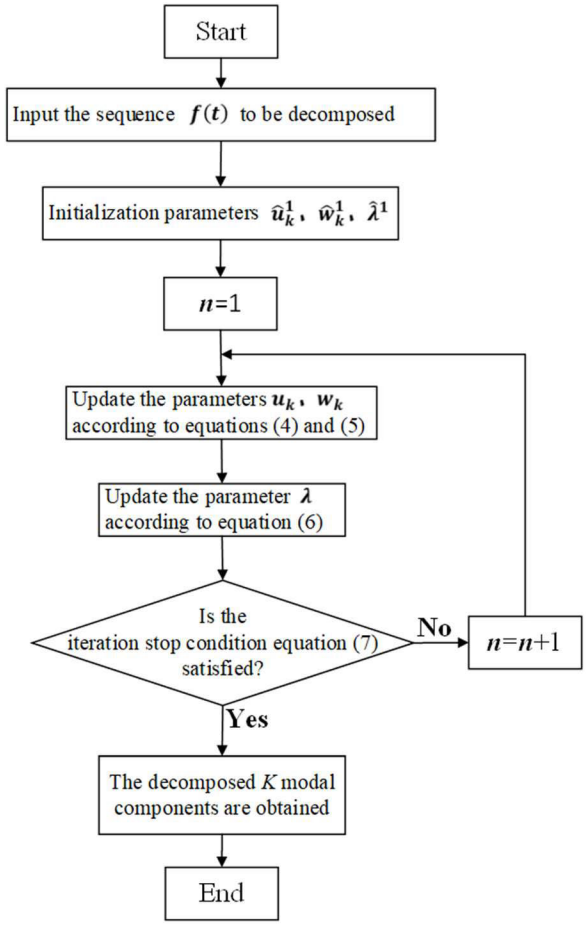

FIGURE 1. VMD decomposition flowchart.

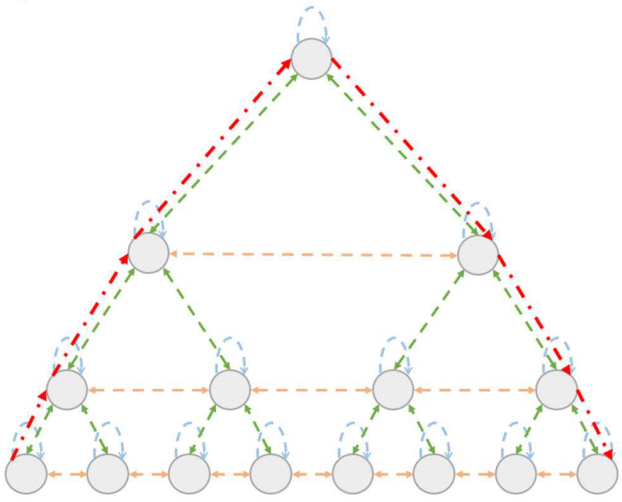

FIGURE 2. Pyramidal attention module, PAM.

B. PYRAFORMER NETWORK

Pyraformer is a neural network based on pyramidal attention. Each module of the Pyraformer network is described in detail next.

1) PYRAMIDAL ATTENTION MODULE

To capture different ranges of temporal dependencies, the pyramidal attention module (PAM) is introduced, as shown in Figure 2.

The module uses a tree structure to perform self-attention, extracting features at different resolutions through inter-scale connectivity and intra-scale connectivity to model the dependencies at different scales. In the pyramid graph structure,

the nodes at the bottom level represent the observed values at each moment, the nodes at the upper level extract features from the nodes at the lower level, and by connecting the nodes at each level, the relationship between each node can be found. Since the nodes in the upper layer contain information extracted from the nodes in the lower layer, and the nodes in the upper layer have already extracted and modeled features for long time sections of information, only the relationships between neighboring nodes need to be considered in each layer, reducing the complexity. This pyramidal graph structure is utilized to characterize the temporal dependencies in the sequence in a multi-resolution manner.

As shown in Figure 2, from inside to outside, the blue dashed line indicates the self-attention of each node, the orange dashed line indicates the information exchange between nodes within the same scale, the green dashed line indicates the information exchange between nodes of different scales, and the red dotted line indicates the maximum information propagation path required for information exchange between any two nodes.

2) COARSE-SCALE CONSTRUCTION MODULE

Coarse-scale construction module (CSCM) summarizes the embedded sequences at different scales and builds a multi-resolution tree structure, which in turn uses PAM to efficiently exchange information between nodes. The CSCM module introduces coarse-scale nodes scale by scale from the bottom to the top by performing a convolution on the corresponding sub-nodes C^S. As shown in Figure 3, several convolutional layers with kernel size C and step size C are applied sequentially to the embedded sequences in the time dimension, resulting in sequences of length L/C^S. These fine-to-coarse sequences are concatenated and fed to the PAM. To reduce the number of parameters and computation, each node is downscaled through a fully connected layer before feeding the sequence to the cascaded convolutional layers and recovered after all convolutions. Such a structure significantly reduces the number of parameters in the module and prevents overfitting.

Where B is the batch size, L is the sequence length, D represents the feature dimension of each node, and D_K denotes the feature dimension of the key vector.

The flowchart of pyraformer prediction network is shown in Figure 4.

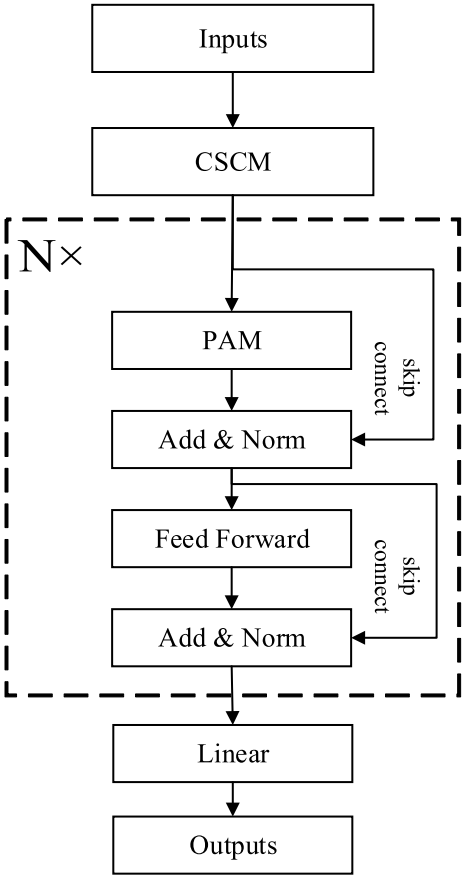

FIGURE 4. Pyraformer network flowchart.

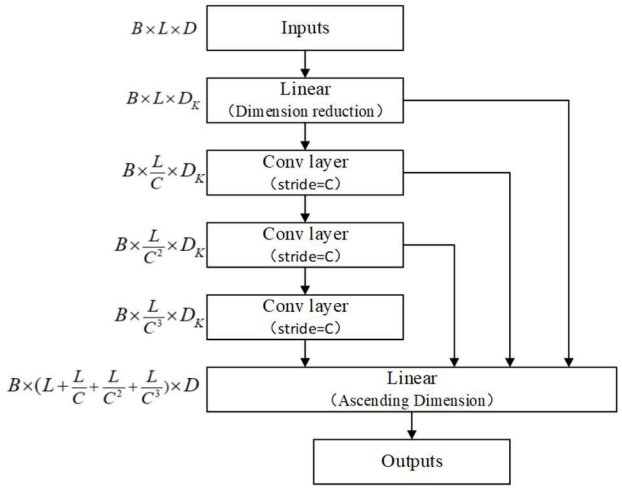

FIGURE 3. Coarse-scale construction module, CSCM.

3) ADAN OPTIMIZER

The Adan optimizer was proposed by Xie et al. [28], which combines the advantages of adaptive optimizers, Nesterov impulses, and decoupled weight decay strategies to provide faster convergence than previous adaptive gradient algorithms, and can withstand larger learning rates and batch sizes, as well as enable dynamic L2 regularization of the model parameters. Adaptive moment estimation (Adam) optimizer is a common and effective gradient descent algorithm that uses the reball method impulse paradigm, but there are still some drawbacks: (i) the adaptive algorithm is similar to the effect of overlearning, and the generated model is overfitted when facing the overall distribution; (ii) the learning rate of Adam may change drastically during the optimization process, which may cause Adam not to converge, or miss the global optimal solution. While the Adan optimizer uses the Nesterov impulse algorithm, the Nesterov algorithm uses

the impulse to find an extrapolation point when calculating the gradient, completes the gradient calculation at that point and performs the impulse accumulation. The extrapolation point helps the Nesterov algorithm to sense the geometric information around the current point in advance, and does not simply rely on past impulses to bypass sharp local minima, but adjusts the direction of parameter updates by observing the surrounding gradients in advance. The operation of the Adan Optimizer is shown in Figure 5. This feature makes this optimizer more suitable for complex training paradigms and model structures.

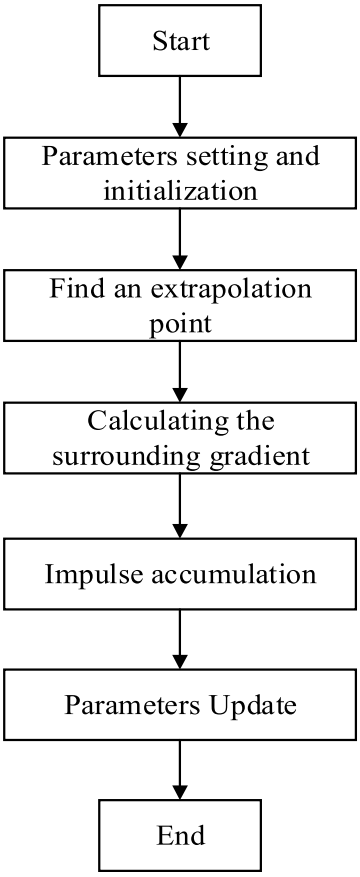

FIGURE 5. Adan optimizer flowchart.

C. PREDICTIVE MODEL STRUCTURE

In this paper, the VMD algorithm, Pyraformer network and Adan optimizer are selected for data processing, network structure and optimizer selection for power load forecasting, i.e. VMD-Pyraformer-Adan forecasting model, and the model structure is shown in Figure 6.

In the VMD decomposition module, the parameter K is selected according to equation (8). The power load data are decomposed by VMD to obtain IMF components with different frequencies, and the multiple components are divided into low-frequency, mid-frequency and high-frequency parts using the over-zero rate and PCC analysis. The divided components are separately formed into reconstructed data with

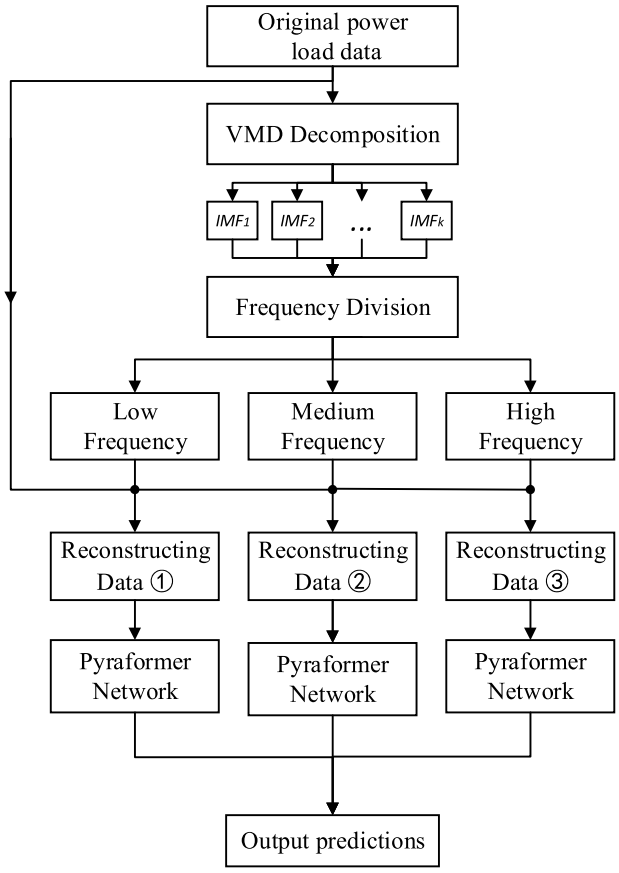

FIGURE 6. VMD-pyraformer-adan forecasting model.

the original data. Since the modal components of different frequency bands contain different features, the parameters of neural network learning are not uniform. Therefore, we choose to input the reconstructed data consisting of high, medium and low frequency components with the original data respectively to the respective prediction networks for learning. VMD decomposition of the load data before input to the prediction network can effectively decompose the load signal into several sets of signals with finite bandwidth and center frequency, which has a better signal processing effect of anti-noise and anti-volatility, and reduces the occurrence of modal blending phenomenon and endpoint effect, which can effectively extract the features of the electric load data, and the reconstructed data input to the network for feature extraction and prediction The prediction accuracy can be effectively improved.

In the Pyraformer prediction network, the reconstructed dataset is fed into the network model. The coarse-grained nodes on the pyramid structure are first initialized through the CSCM module, which summarizes the embedded sequences at different scales and builds a multi-resolution tree structure. Next, the PAM module is used to efficiently extract and exchange information between nodes to further capture different ranges of temporal dependencies. The Pyraformer model based on pyramidal attention is able to capture the

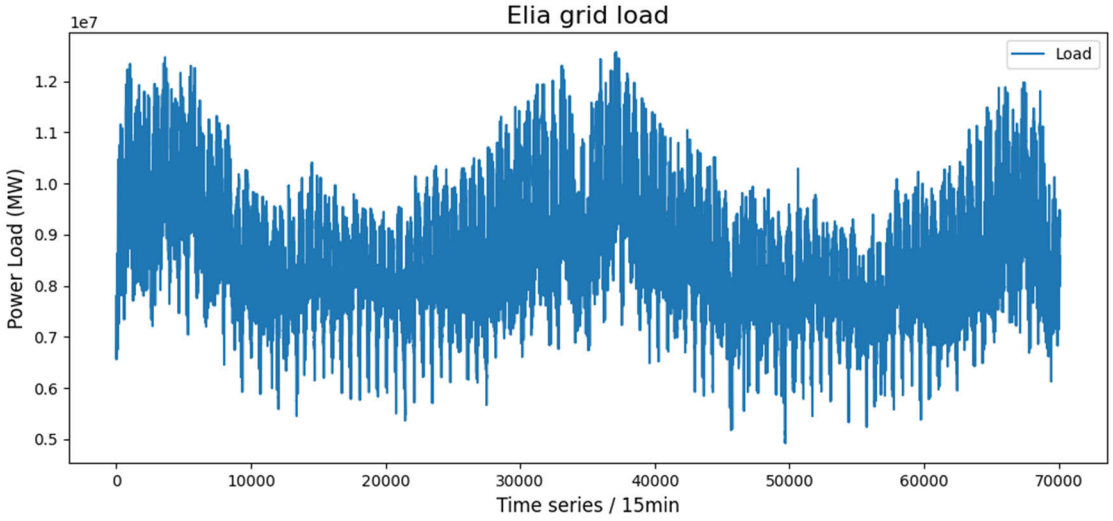

FIGURE 7. Power load data.

long-range dependencies of time series and has a faster convergence rate with the addition of Adan optimizer. Therefore, the VMD-Pyraformer-Adan power load forecasting model proposed in this paper makes improvements in data processing, network structure and optimizer selection, and the low-frequency components containing the main features of the original data are obtained by VMD decomposition and data reconstruction with the original data. The feature extraction capability of Pyraformer for time series and the parameter optimization capability of Adan optimizer are used to obtain the final power load prediction values to achieve accurate prediction of power load data and improve the prediction performance of the model.

D. MODEL EVALUATION METRICS

In order to evaluate the prediction performance of the model, root mean square error (RMSE) and mean absolute percentage error (MAPE) are chosen as the evaluation indicators in this paper. RMSE is often used to measure the deviation between the predicted and true values, and MAPE considers not only the deviation between the predicted and true values, but also the ratio between the deviation and the true values. y_i denotes the true value, $\bar{y}_i$ denotes the predicted value, and n denotes the total number of samples. Smaller values of RMSE and MAPE represent more accurate load prediction. The calculation formula is as follows.

$$RMSE = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n} y_i - \bar{y}_i\right)^2} \tag{10}$$

$$MAPE = \sum_{i=1}^{n} \frac{|y_i - \bar{y}_i|}{y_i} \times 100\% \tag{11}$$

III. SIMULATION ANALYSIS OF ALGORITHMS

A. DATASET

The dataset used in this paper is derived from the public dataset of the Belgian grid company Elia's electricity load [29]. It includes net generation measured from local power stations injecting electricity into the Elia grid, net inflows from distribution to the Elia grid and net imports at the border. The dataset is selected from the electricity load at 15-minute intervals from January 1, 2018 to December 31, 2019, with a total length of 70080. The data set is divided into a training set and a test set, where the training set is used for model learning training and parameter tuning, and the test set experimental results are used to evaluate the prediction effectiveness of the proposed model. The electric load data situation is shown in Figure 7.

B. VMD DECOMPOSITION

The power load data are input to the VMD module for decomposition. In the VMD algorithm, the quadratic penalty factor $\alpha = 2000$, the noise tolerance $\tau = 0$, the K value is selected by the improved signal energy (ISE) method, and the remaining parameters are default values.

To determine the K value, keep the other parameters constant, decompose the power load data into the corresponding IMF components and residual signals, and find the ratio of residual energy to the original energy at different K values according to equation (8), as shown in Figure 8.

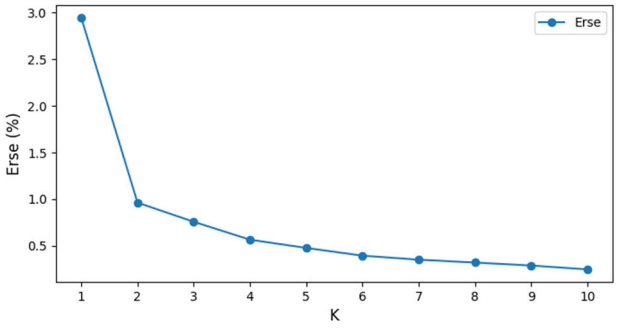

FIGURE 8. Erse at different K values.

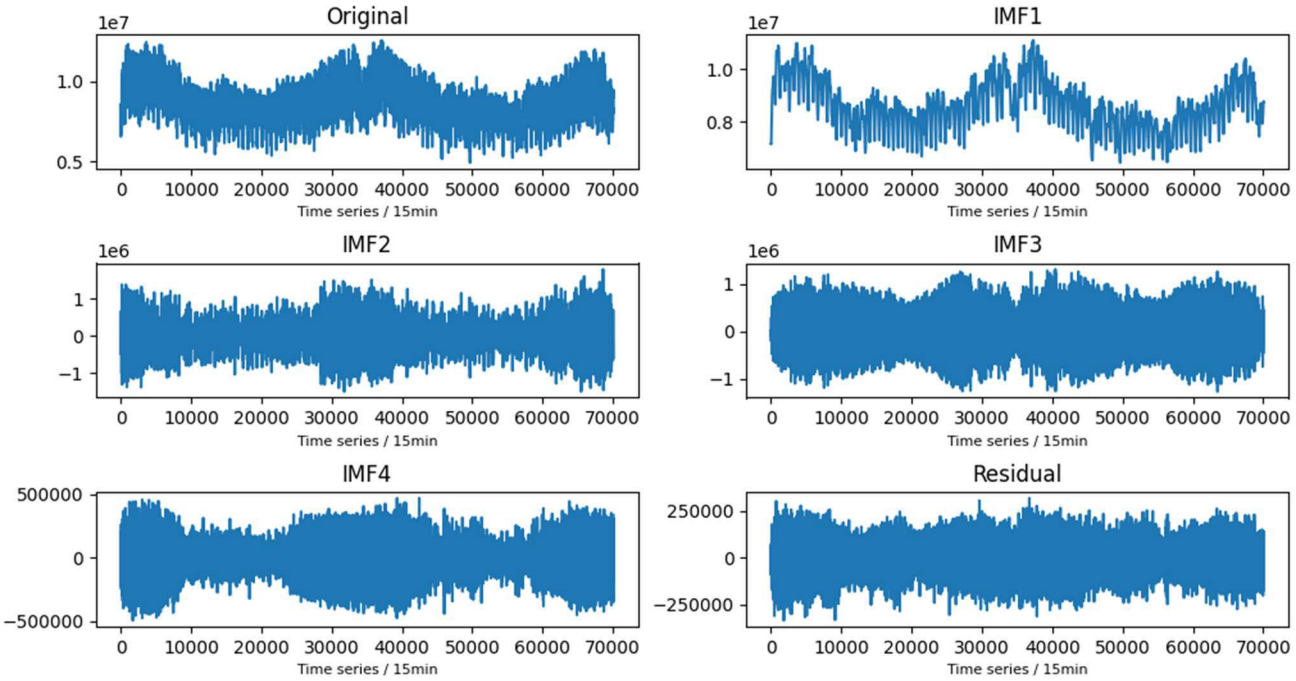

FIGURE 9. Original signal and decomposed IMF components.

As can be seen from Figure 8, when $K = 3$, Erse $\leq 1\%$, but at this time the Erse curve still has a significant downward trend; when $K = 4$, Erse $= 0.567\%$, and the downward trend tends to level off. Therefore, it can be determined that $K = 4$ is the appropriate modal number. The results of the power load data after VMD decomposition are shown in Figure 9.

The power load data are decomposed into multiple IMF components by VMD decomposition, and the over-zero rate of each component is calculated according to equation (9), as shown in Table 1.

TABLE 1. Over-zero rate of each intrinsic mode function.

Intrinsic Mode Function	Over-zero rate
IMF_1	0
IMF_2	0.020
IMF_3	0.042
IMF_4	0.081
Residual	0.355

As can be seen from Table 1, the modal components IMF_1, IMF_2, IMF_3, IMF_4 and Residual increase in frequency in that order.

C. PCC ANALYSIS

Pearson correlation coefficient (PCC) analysis refers to the analysis of two or more elements of variables with correlation, with the aim of analyzing the degree of correlation between two variables, so as to measure the closeness of the correlation between two variable factors, calculated as in equation (12).

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{\sum_{i=1}^{n}(X_i - E(X))(Y_i - E(Y))}{\sigma_X \sigma_Y}}{n} \tag{12}$$

where $\boldsymbol{\rho}_{XY}$ is the correlation coefficient between variable X and variable Y; $Cov(X,Y)$denotes the covariance between vectors; σ_X and $\sigma_{\mathbf{Y}}$ are the standard deviations of variable X and variable Y, and E denotes the mathematical expectation.

PCC analysis was performed on each modal component and the results are shown in Table 2.

TABLE 2. Correlation of each modal component.

Intrinsic Mode Function	Relevance
IMF_1	0.792
IMF_2	0.478
IMF_3	0.431
IMF_4	0.173
Residual	0.153

According to the results in Table 1 and Table 2, we divide IMF_1 as the low-frequency part, IMF_2 and IMF_3 as the medium-frequency part, IMF_4 and Residual as the high-frequency part. Next, the data from each of the three parts and the original load data are composed into reconstructed data and fed into the prediction network.

D. ANALYSIS OF SIMULATION RESULTS

1) PYRAFORMER NETWORK PARAMETERS SELECTION

The selection of appropriate model parameters plays a crucial role in the predictive performance of the model. Here, the appropriate model parameters are determined by univariate comparison experiments. First, the number of attention layers, attention heads and Batch size are set as default values, and the parameters to be adjusted are changed according to the univariate principle. During the experiment, epoch = 10, the number of neighboring nodes for information transfer of child nodes in PAM is 3 (including itself), and the number of child nodes for information transfer of parent nodes is 4.

In the process of selecting the model parameters, the MAPE of the optimal generation and the average training time per generation were selected as evaluation indicators. According to the single variable principle, keeping other parameters constant and changing the number of attention layers, the results were obtained as in Table 3.

TABLE 3. Batch size comparison.

Batch size	MAPE(%)	Training time / epoch
2	0.686	17.15min
4	0.674	9.97min
8	0.691	7.25min
16	0.692	5.45min
32	0.754	4.89min
64	0.929	4.59min

In the training process, if the Batch size is set too small, it is easy to fall into the local optimum and cannot get the global optimum solution; if it is set too large, the gradient update direction is not accurate and the error loss is not easy to converge to the minimum value.

According to the experimental results in Table 4, the optimal Batch size is selected as 4. According to the same method, the optimal number of attention layers is 6 and the optimal number of attention heads is 4.

2) OPTIMIZER COMPARISON

In the network parameter optimization section, the Adan optimizer selected in this paper is compared with the current mainstream Adam optimizer. The MSE loss curves under the two optimizers are shown in Figure 10.

It can be seen from Figure 10 that the Adan optimizer converges faster than the Adam optimizer, verifying the fast convergence of the Adan optimizer.

3) OVERALL MODEL COMPARISON

In order to verify the superiority of the proposed model, experiments are conducted to compare the model in this paper with other models under the same conditions to predict the future electric load data. Experiments 1, 2, 3, and 4 are used to verify the significant effect after adding each module algorithm. Experiments 6, 7, and 8 experiments are compared with the prediction models proposed by others [30], [31], [32]

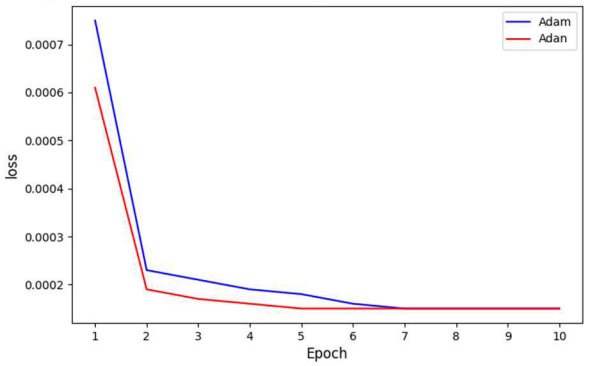

FIGURE 10. Loss curves under two optimizers.

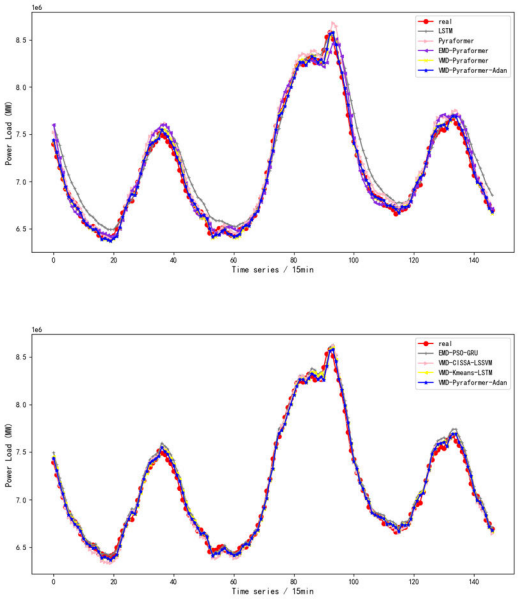

FIGURE 11. Comparison of model prediction results.

TABLE 4. Comparison of the models.

Models	MAPE(%)	RMSE(MW)
① LSTM	1.46	159364
② Pyraformer	0.92	101518
③ EMD-Pyraformer	0.86	94032
④ VMD-Paraformer	0.69	74516
⑤ VMD-Pyraformer-Adan	0.65	70732
⑥ EMD-PSO-GRU	0.75	82506
⑦ VMD-CISSA-LSSVM	0.72	79214
⑧ VMD-Kmeans-LSTM	0.68	73648

for verifying the superiority of the combined model in this paper. The experimental results are shown in Table 4.

Comparing the model in this paper with the models in Table 4, the model in this paper outperforms the other models in terms of MAPE and RMSE. Compared with the other models, MAPE is reduced by 0.03% - 0.81% and RMSE is reduced by 2916-88632 MW, finally reaching the optimal accuracy. The model prediction results are shown in Figure 11.

As can be seen from Figure 11, the model in this paper can learn the potential characteristics of the power load data when the power load data changes sequentially or abruptly, and the prediction curve fits the real data curve more closely than other models, achieving a better prediction effect.

IV. CONCLUSION

Due to the volatility, periodicity and nonlinearity of power loads, this paper proposes a VMD-Pyraformer-Adan model for short-term power load forecasting. The VMD algorithm is used to decompose the electric load data, and the frequency of the modal components is measured using the over-zero rate, and then combined with the Pearson correlation coefficient analysis to divide the high, medium and low frequency components. Next, the original data are combined with each of the three modal components to obtain three sets of reconstructed data, which are fed into the Pyraformer network for training. While the Pyraformer network is being trained, the parameters in the neural network are optimized using the advantage of fast convergence of the Adan optimizer, and the final output of the electric load prediction results. Then comparing the prediction model proposed in this paper with other prediction models, the model in this paper shows good performance and fitting effect, which can improve the accuracy of short-term power load prediction, with the following advantages.

(1) The Pyraformer model based on pyramidal attention is able to capture the long-range dependencies of time series and can accurately predict power load data, with faster convergence speed after adding Adan optimizer.

(2) After the nonlinear and volatile power load series are decomposed by VMD, the modal components of different frequency bands contain different features, which can improve the accuracy of the prediction model as one of the inputs of the prediction network.

(3) Compared with the existing models, the VMD-Pyraformer-Adan model proposed in this paper has a higher accuracy in short-term power load forecasting as the predicted values match the true values more closely.

Although the model in this paper shows good prediction performance, it still has certain shortcomings. The next step will be to further explore the data characteristics of electric load, improve the prediction network structure, and find the parameter automatic optimization algorithm to optimize the parameters in order to further improve the short-term electric load prediction accuracy and efficiency.

REFERENCES

[1] X. Yao, X. Fu, and C. Zong, "Short-term load forecasting method based on feature preference strategy and LightGBM-XGboost," *IEEE Access*, vol. 10, pp. 75257–75268, 2022, doi: 10.1109/ACCESS.2022.3192011.

[2] C. Feng, M. Sun, and J. Zhang, "Reinforced deterministic and probabilistic load forecasting via *Q*-learning dynamic model selection," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1377–1386, Jun. 2020, doi: 10.1109/TSG.2019.2937338.

[3] C. Fan, Y. Ding, and Y. Liao, "Analysis of hourly cooling load prediction accuracy with data-mining approaches on different training time scales," *Sustain. Cities Soc.*, vol. 51, Nov. 2019, Art. no. 101717, doi: 10.1016/j.scs.2019.101717.

[4] S. Yan and M. Hu, "A multi-stage planning method for distribution networks based on ARIMA with error gradient sampling for source–load prediction," *Sensors*, vol. 22, no. 21, p. 8403, Nov. 2022, doi: 10.3390/s22218403.

[5] Y. Huang, F. Zhu, G. Jia, and Y. Zhang, "A slide window variational adaptive Kalman filter," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 67, no. 12, pp. 3552–3556, Dec. 2020, doi: 10.1109/TCSII.2020.2995714.

[6] G. Dudek, P. Pelka, and S. Smyl, "A hybrid residual dilated LSTM and exponential smoothing model for midterm electric load forecasting," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 7, pp. 2879–2891, Jul. 2022, doi: 10.1109/TNNLS.2020.3046629.

[7] M. A. Mahmud, "Isolated area load forecasting using linear regression analysis: Practical approach," *Energy Power Eng.*, vol. 3, no. 4, pp. 547–550, 2011.

[8] J. Moon, Y. Kim, M. Son, and E. Hwang, "Hybrid short-term load forecasting scheme using random forest and multilayer perceptron," *Energies*, vol. 11, no. 12, p. 3283, Nov. 2018, doi: 10.3390/en11123283.

[9] Y. Fu, Z. Li, H. Zhang, and P. Xu, "Using support vector machine to predict next day electricity load of public buildings with sub-metering devices," *Proc. Eng.*, vol. 121, pp. 1016–1022, Jan. 2015.

[10] T. Liu, Z. Tan, C. Xu, H. Chen, and Z. Li, "Study on deep reinforcement learning techniques for building energy consumption forecasting," *Energy Buildings*, vol. 208, Feb. 2020, Art. no. 109675, doi: 10.1016/j.enbuild.2019.109675.

[11] X.-B. Jin, W.-Z. Zheng, J.-L. Kong, X.-Y. Wang, Y.-T. Bai, T.-L. Su, and S. Lin, "Deep-learning forecasting method for electric power load via attention-based encoder–decoder with Bayesian optimization," *Energies*, vol. 14, no. 6, p. 1596, Mar. 2021, doi: 10.3390/en14061596.

[12] H. H. Goh, B. He, H. Liu, D. Zhang, W. Dai, T. A. Kurniawan, and K. C. Goh, "Multi-convolution feature extraction and recurrent neural network dependent model for short-term load forecasting," *IEEE Access*, vol. 9, pp. 118528–118540, 2021, doi: 10.1109/ACCESS.2021.3107954.

[13] Z. Li, S. Bao, and Z. Gao, "Short term prediction of photovoltaic power based on FCM and CG-DBN combination," *J. Electr. Eng. Technol.*, vol. 15, no. 1, pp. 333–341, Jan. 2020, doi: 10.1007/s42835-019-00326-3.

[14] Z. Jia, L. Yang, Z. Zhang, H. Liu, and F. Kong, "Sequence to point learning based on bidirectional dilated residual network for non-intrusive load monitoring," *Int. J. Electr. Power Energy Syst.*, vol. 129, Jul. 2021, Art. no. 106837, doi: 10.1016/j.ijepes.2021.106837.

[15] N. Q. Dat, N. T. N. Anh, N. N. Anh, and V. K. Solanki, "Hybrid online model based multi seasonal decompose for short-term electricity load forecasting using ARIMA and online RNN," *J. Intell. Fuzzy Syst.*, vol. 41, no. 5, pp. 5639–5652, Nov. 2021, doi: 10.3233/JIFS-189884.

[16] B. Farsi, M. Amayri, N. Bouguila, and U. Eicker, "On short-term load forecasting using machine learning techniques and a novel parallel deep LSTM-CNN approach," *IEEE Access*, vol. 9, pp. 31191–31212, 2021, doi: 10.1109/ACCESS.2021.3060290.

[17] S. Jung, J. Moon, S. Park, and E. Hwang, "An attention-based multilayer GRU model for multistep-ahead short-term load forecasting," *Sensors*, vol. 21, no. 5, p. 1639, Feb. 2021, doi: 10.3390/s21051639.

[18] M. Morchid, "Parsimonious memory unit for recurrent neural networks with application to natural language processing," *Neurocomputing*, vol. 314, pp. 48–64, Nov. 2018, doi: 10.1016/j.neucom.2018.05.081.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.

[20] S. Liu, H. Yu, C. Liao, J. Li, W. Lin, A. X. Liu, and S. Dustdar, "Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–20.

[21] C. Fan, C. Ding, J. Zheng, L. Xiao, and Z. Ai, "Empirical mode decomposition based multi-objective deep belief network for short-term power load forecasting," *Neurocomputing*, vol. 388, pp. 110–123, May 2020, doi: 10.1016/j.neucom.2020.01.031.
[22] L. Lv, Z. Wu, J. Zhang, L. Zhang, Z. Tan, and Z. Tian, "A VMD and LSTM based hybrid model of load forecasting for power grid security," *IEEE Trans. Ind. Informat.*, vol. 18, no. 9, pp. 6474–6482, Sep. 2022, doi: 10.1109/TII.2021.3130237.
[23] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 531–544, Feb. 2014, doi: 10.1109/TSP.2013.2288675.
[24] P. Jia, H. Zhang, X. Liu, and X. Gong, "Short-term photovoltaic power forecasting based on VMD and ISSA-GRU," *IEEE Access*, vol. 9, pp. 105939–105950, 2021, doi: 10.1109/ACCESS.2021.3099169.
[25] Q. Sun and H. Cai, "Short-term power load prediction based on VMD-SG-LSTM," *IEEE Access*, vol. 10, pp. 102396–102405, 2022, doi: 10.1109/ACCESS.2022.3206486.
[26] J. Duan, P. Wang, W. Ma, S. Fang, and Z. Hou, "A novel hybrid model based on nonlinear weighted combination for short-term wind power forecasting," *Int. J. Electr. Power Energy Syst.*, vol. 134, Jan. 2022, Art. no. 107452, doi: 10.1016/j.ijepes.2021.107452.
[27] S. Shen and J. J. He, "SGCS: A signal reconstruction method based on Savitzky–Golaysgz filtering and compressed sensing for wavelength modulation spectroscopy," *Opt. Exp.*, vol. 29, no. 22, pp. 35848–35863, 2021, doi: 10.1364/OE.437649.
[28] X. Xie, P. Zhou, H. Li, Z. Lin, and S. Yan, "Adan: Adaptive Nesterov momentum algorithm for faster optimizing deep models," 2022, *arXiv:2208.06677*.
[29] *ELIA: Elia Grid Load Data*. Accessed: Nov. 14, 2022. [Online]. Available: https://www.elia.be/en/grid-data/data-download-page
[30] W. Zhang and T. Wang, "Short-term power load forecasting model design based on EMD-PSO-GRU," *Sci. Program.*, vol. 2022, Aug. 2022, Art. no. 4755519, doi: 10.1155/2022/4755519.
[31] G. Wang, X. Wang, Z. Wang, C. Ma, and Z. Song, "A VMD–CISSA–LSSVM based electricity load forecasting model," *Mathematics*, vol. 10, no. 1, p. 28, Dec. 2021, doi: 10.3390/math10010028.
[32] Z. Sun, S. Zhao, and J. Zhang, "Short-term wind power forecasting on multiple scales using VMD decomposition, K-means clustering and LSTM principal computing," *IEEE Access*, vol. 7, pp. 166917–166929, 2019, doi: 10.1109/ACCESS.2019.2942040.

YIHAO TANG was born in Henan, China, in 2002. He is currently pursuing the bachelor's degree with the School of Electrical and Electronic Engineering, Hubei University of Technology. His research interests include control theory and control engineering.

HUAFENG CAI was born in Hubei, China, in 1978. He received the master's degree in power electronics and power transmission from the Hubei University of Technology, in 2005. He is currently with the Hubei University of Technology as an Associate Professor in automation research direction for power electronics technology, fault prediction, and optimization.