Hybrid Network Model Based on Data Enhancement for Short-term Power Prediction of New PV Plants

Shangpeng Zhong, Xiaoming Wang, Bin Xu, Hongbin Wu, and Ming Ding

***Abstract*—This study proposes a hybrid network model based on data enhancement to address the problem of low accuracy in photovoltaic (PV) power prediction that arises due to insufficient data samples for new PV plants. First, a time-series generative adversarial network (TimeGAN) is used to learn the distribution law of the original PV data samples and the temporal correlations between their features, and these are then used to generate new samples to enhance the training set. Subsequently, a hybrid network model that fuses bi-directional long-short term memory (BiLSTM) network with attention mechanism (AM) in the framework of deep & cross network (DCN) is constructed to effectively extract deep information from the original features while enhancing the impact of important information on the prediction results. Finally, the hyperparameters in the hybrid network model are optimized using the whale optimization algorithm (WOA), which prevents the network model from falling into a local optimum and gives the best prediction results. The simulation results show that after data enhancement by TimeGAN, the hybrid prediction model proposed in this paper can effectively improve the accuracy of short-term PV power prediction and has wide applicability.**

***Index Terms*—New photovoltaic (PV) plant, short-term prediction, time-series generative adversarial network (TimeGAN), hybrid network, hyperparameter.**

I. Introduction

As one of the most productive energy sources available, photovoltaic (PV) power plays a crucial role in meeting the growing global demand for clean energy [1] - [4]. However, the intermittent and fluctuating nature of PV power generation due to variations in irradiance poses many challenges in terms of the safe and stable operation of the grid system when large-scale PV generators are connected to the power grid [5]-[7]. Improving the accuracy of PV power prediction offers an effective solution to overcome these challenges [8], [9].

The accuracy of PV power prediction not only depends on the model itself but also usually needs to be based on sufficient historical data. When the amount of training data is insufficient, it is extremely difficult to train the model to give a highly accurate prediction network with normal fitting. A lack of large amounts of historical data is a common problem for most new PV plants, and the inadequacy of the training data poses a great challenge in terms of PV power prediction. It is therefore crucial to find an effective method to reduce the impact of insufficient raw data on the output power prediction accuracy of new PV plants.

At present, when faced with insufficient raw PV data, the methods typically used to obtain accurate PV power prediction values can be divided into two categories, i.e., methods based on physical models and based on data synthesis.

1) Methods based on physical models mainly use numerical weather prediction to obtain the effective irradiance and then combine parameters such as the PV array conversion efficiency and the installation angle of the PV system to establish a physical model to obtain the output power of the plant [10], [11]. However, although this type of method does not require historical data (such as power and meteorology information) to train the prediction model, it relies on the geographical information of the PV plant and real-time meteorological data, and the physical model itself contains certain errors, making the model less resistant to disturbances [12], [13].

2) Methods based on data synthesis do not require complex physical modeling and can generate new samples with fewer sample data [14], [15]. This type of method uses neural networks to learn a sample distribution law for the PV data for different weather types, generates new samples that are similar to the original samples to expand the sample training set, and thus improves the prediction accuracy of the network model [16].

Unlike methods based on physical models, those based on data synthesis generate new samples to increase the sample capacity, which is more effective for solving the problem of insufficient raw data for PV plants. This type of method not only ensures that the generated samples are similar to the

Manuscript received: November 16, 2022; revised: March 18, 2023; accepted: May 25, 2023. Date of Crosscheck: May 25, 2023. Date of online publication: August 16, 2023.

This work was supported by the Regional Innovation and Development Joint Fund of National Natural Science Foundation of China (No. U19A20106), the Science and Technology Major Projects of Anhui Province (No. 202203f0702000 3), and the Science and Technology Project of State Grid Corporation of China (No. 52120522000F).

S. Zhong, X. Wang, H. Wu (corresponding author), and M. Ding are with the Anhui Province Key Laboratory of Renewable Energy Utilization and Energy Saving, Hefei University of Technology, Hefei 230009, China, (e-mail: zhongshangpeng@163. com; 316855097@qq. com; hfwuhongbin@hfut. edu. cn; mingding56@126.com).

B. Xu is with the State Grid Anhui Electric Power Research Institute, Hefei 230061, China (e-mail: xubin1980@sina.com).

DOI: 10.35833/MPCE.2022.000759

original samples, but also maintains a certain amount of diversity. In this paper, we therefore present a study of short-term power prediction of new PV plants based on data synthesis.

Current PV power prediction models based on data synthesis usually contain the following two links: data synthesis and model prediction, as described below.

1) Data synthesis. Generative adversarial networks (GANs) were proposed by Goodfellow and Pouget-Abadie to provide a proven solution for data synthesis, and have recently become the mainstream approach to data synthesis [17], [18]. A GAN generates new samples based on the distribution law of the original samples through a confrontation between a generative model and a discriminative model. Its powerful modeling capabilities not only enable the synthesis of new samples that are similar to the original sample data but can also maintain the diversity of the generated data. Reference [19] used Wasserstein GANs (WGANs) to generate wind power data based on limited historical information by augmenting the original dataset to improve the estimation of the worst-case probability distribution in distributed robust optimization. Reference [20] used a WGAN with penalized gradients to generate solar irradiance data and train a convolutional neural network based weather classification model with an augmented dataset consisting of the original and generated data to improve the classification accuracy of the model. In [21], a GAN based on financial time-series models was proposed to generate real data in a data-driven manner by learning the properties of the data. Although a large number of studies have shown that GANs are effective in terms of generating new samples, they have rarely considered the temporal correlations among variables when performing time-series data generation, which tends to reduce the diversity and accuracy of the generated samples.

2) Model prediction. Following the rapid development of machine learning, many researchers now use artificial neural networks (ANNs), long-short term memory (LSTM) networks, recurrent neural networks (RNNs), and support vector machines (SVMs) for the PV power prediction. In [22], a genetic-algorithm-based SVM (GA-SVM) model was proposed to classify historical weather data, in which a GA-optimized SVM was used to improve the short-term power prediction capability of residential PV power systems. Reference [23] proposed a deep learning based spatio-temporal correlation model that used a CNN to extract the spatial features from meteorological data and an LSTM to extract the temporal features from historical irradiance data to improve the prediction accuracy of solar irradiance. A variational modal decomposition method using a combination of ant colony optimization and neural networks was proposed [24] that combined ANNs with data processing, input variable selection, and external optimization techniques to predict the short-term output power of PV systems. Reference [25] used a short-term PV power prediction method based on an MFA-Elman neural network, which improved the prediction rate and accuracy by using MFA to solve problems such as the randomness of initial weights and thresholds of the Elman neural network and the slow training speed. A large number of studies have shown that although existing PV power prediction models could give good prediction results, they tend to favor the optimization of the network model, and few authors have mined deep information from the original data.

In summary, in order to improve the short-term power prediction accuracy of new PV plants, and to reduce the impact of insufficient original sample data on the accuracy of PV power prediction, we propose a hybrid network prediction method based on data augmentation, in which TimeGAN is used to generate new samples in the data synthesis phase. A deep & cross network with bi-directional long-short term memory and attention mechanism (DCN-BiLSTM-AM) hybrid network is then constructed, and finally, a hybrid network optimized using the whale optimization algorithm (WOA) is used to predict the PV output power on future days.

The main contributions of this paper are as follows.

1) TimeGAN is used to learn a sample distribution law from the PV data, and the temporal correlations between features, in order to generate new samples to enhance the original dataset. Principal component analysis (PCA) and t-distribution stochastic neighborhood embedding (t-SNE) are applied to analyze the similarity and diversity of the generated data and the original data, as a way of proving the validity of the generated data.

2) We extract higher-order cross-features using DCN in order to mine the deep information contained in the original features. The DCN-BiLSTM-AM hybrid network model is constructed by fusing BiLSTM and an AM in the framework of DCN to obtain the correlation between PV power output and historical timing information at the current moment, while the AM is used to assign different weights to the hidden states of BiLSTM.

3) The WOA is used to optimize the hyperparameters of the DCN-BiLSTM-AM hybrid network such as the number of neurons in the hidden layer, the number of model iterations, the learning rate, and the number of network layers in the deep and cross layers of the DCN. This gives the optimal hyperparameters for the network, thus optimizing the hybrid network model and giving the best prediction values while avoiding falling into local optima.

The rest of this paper is organized as follows. Section II introduces the overall prediction framework proposed in this paper. Section III investigates the data selection and synthesis. Section IV describes the components of the DCN-BiLSTM-AM hybrid network model and its overall prediction process. Section V explores the WOA optimization for the hybrid network model. The simulation verification results are presented in Section VI, and conclusions are summarized in Section VII.

II. Proposed Prediction Framework

The proposed prediction framework consists of three modules, as shown in Fig. 1.

1) Module 1: data analysis and synthesis. The main task of this part is to analyze the correlation between feature variables and historical power and use this as a basis for feature selection. Subsequently, TimeGAN is applied to learn the sample distribution law of the original PV data and to generate new PV data for different weather types to enhance the

original sample training set.

2) Module 2: DCN-BiLSTM-AM hybrid network model. The main task of this part is to fuse BiLSTM and AM in the DCN framework to obtain a DCN-BiLSTM-AM hybrid network. This network mines deep information from the original data to lay the foundation for model prediction while assigning different weights to important information to strengthen its influence on the final prediction results.

3) Module 3: WOA optimization. This part optimizes the parameters of the hybrid network model constructed in Module 2 to give the optimal hyperparameters of the network, thus allowing us to construct the optimal network model. Finally, the optimized hybrid network model is used to train and test the new sample training set to obtain the final PV power prediction results.

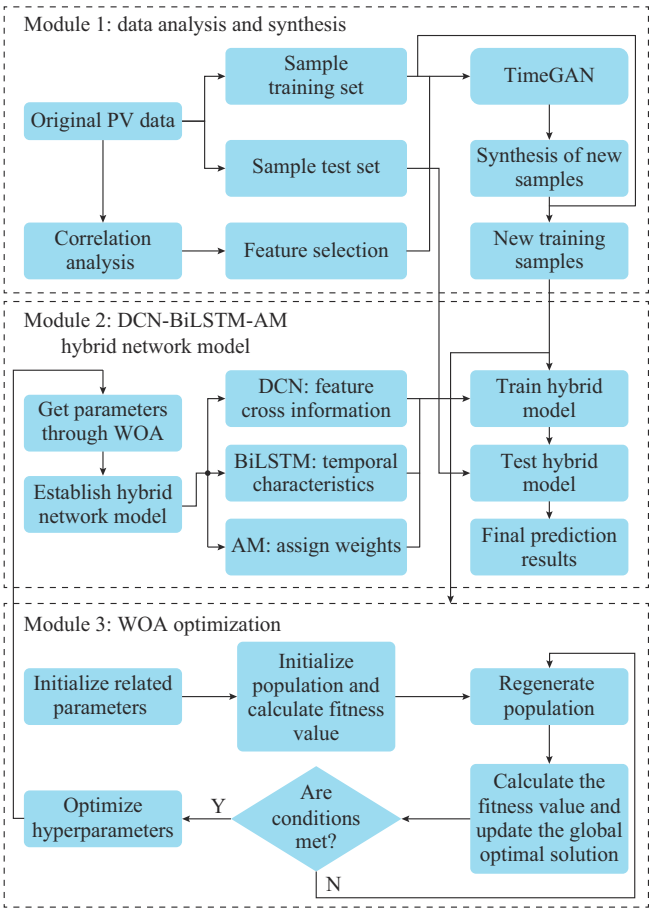

Fig. 1. Flow chart of proposed prediction framework.

We will introduce these three modules in detail in the subsequent sections, which describe in more detail the correlation among data analysis and synthesis, DCN-BiLSTM-AM hybrid network model, and WOA optimization.

III. Data Analysis and Synthesis

A. Correlation Analysis

When historical data containing eight-dimensional features such as temperature, humidity, and horizontal irradiance are used for PV power prediction and data synthesis, the correlation between the feature variables and historical power data and the correlation among the feature variables, are first calculated using Pearson correlation coefficient. The aim is to select the features with higher correlation with the PV power, to facilitate data synthesis and hence PV power prediction.

The Pearson correlation coefficient r is calculated as:

$$r=\frac{\sum_{i=1}^{n}(x_i-\hat{x})(p_i-\hat{p})}{\sqrt{\sum_{i=1}^{n}(x_i-\hat{x})^2\sum_{i=1}^{n}(p_i-\hat{p})^2}} \tag{1}$$

where n is the number of meteorological features; x_i represents the meteorological features; p_i is the PV output power; and $\hat{x}$ and $\hat{p}$ are the mean values of the corresponding variables. The closer the absolute value of Pearson correlation coefficient converges to one, the stronger the correlation between the two variables. The Pearson correlation coefficients between the features are shown in Fig. 2, and it can be observed that the PV output power maintains a strong correlation with six characteristic variables: radiation global tilted, radiation diffuse tilted, diffuse horizontal radiation, global horizontal radiation, temperature, and humidity. The correlation between daily rainfall and wind direction is small, so these variables will not be considered for PV power prediction and data synthesis in this paper.

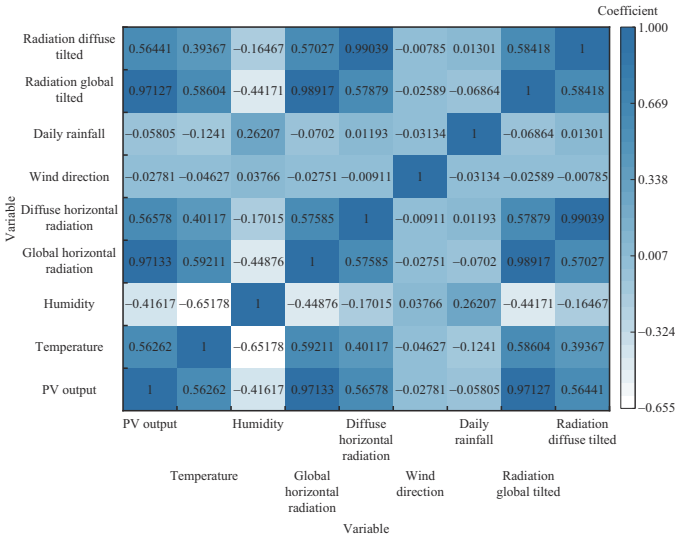

Fig. 2. Heat map of Pearson correlation coefficient.

B. Data Synthesis: TimeGAN

To address the problem of insufficient original sample data of PV plants, we apply sample expansion to the PV data using a GAN [26] for the purpose of increasing the amount of data available. However, since a traditional GAN does not fully consider the time correlation unique to time-series data, a TimeGAN is used for this purpose. TimeGAN enables the generated data to have temporal dynamic properties by considering the inter-row correlations of time series to satisfy the original relationships between the historical data and the feature variables, and the relationships among the feature variables [27].

As shown in Fig. 3, TimeGAN is a generative model based on RNNs and is used to deal with the inter-row correlations of time-series data by combining the GAN framework with autoregressive methods. The model consists of four networks: an embedding function, a recovery function, a sequence generator, and a sequence discriminator. The key to its model is the joint training of the autoencoder network (embedding function and recovery function) and the adversarial network (sequence generator and sequence discriminator), which makes TimeGAN simultaneously learn encoded features, generate representation morphology, and iterate over time. The autoencoder network is trained through supervised and reconstruction losses, where reconstruction losses ensure the accuracy of the learned potential representations and supervised losses help the sequence generator learn the temporal dynamic characteristics of the data.

The adversarial network operates in the potential space provided by the embedded network, making the potential dynamics of original and synthetic data synchronized by supervised losses. Ultimately, after the model training is completed, the time-series data that meet the sample requirements are obtained.

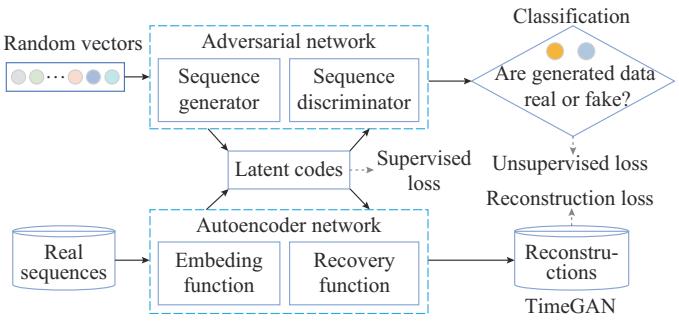

Fig. 3. Structure of TimeGAN.

C. Synthesis and Analysis of PV Data Under Different Meteorological Conditions

To verify the validity of the time-series PV data generated by TimeGAN, we learn the sample distribution laws for raw data of four weather types: sunny, cloudy, sunny to cloudy, and sunny to drizzly. This is done to generate new samples with temporal dynamic characteristics, as shown in Fig. 4. It can be observed that TimeGAN generates new samples with similar data on PV power and its meteorological characteristics for each weather type, which can enhance the dataset.

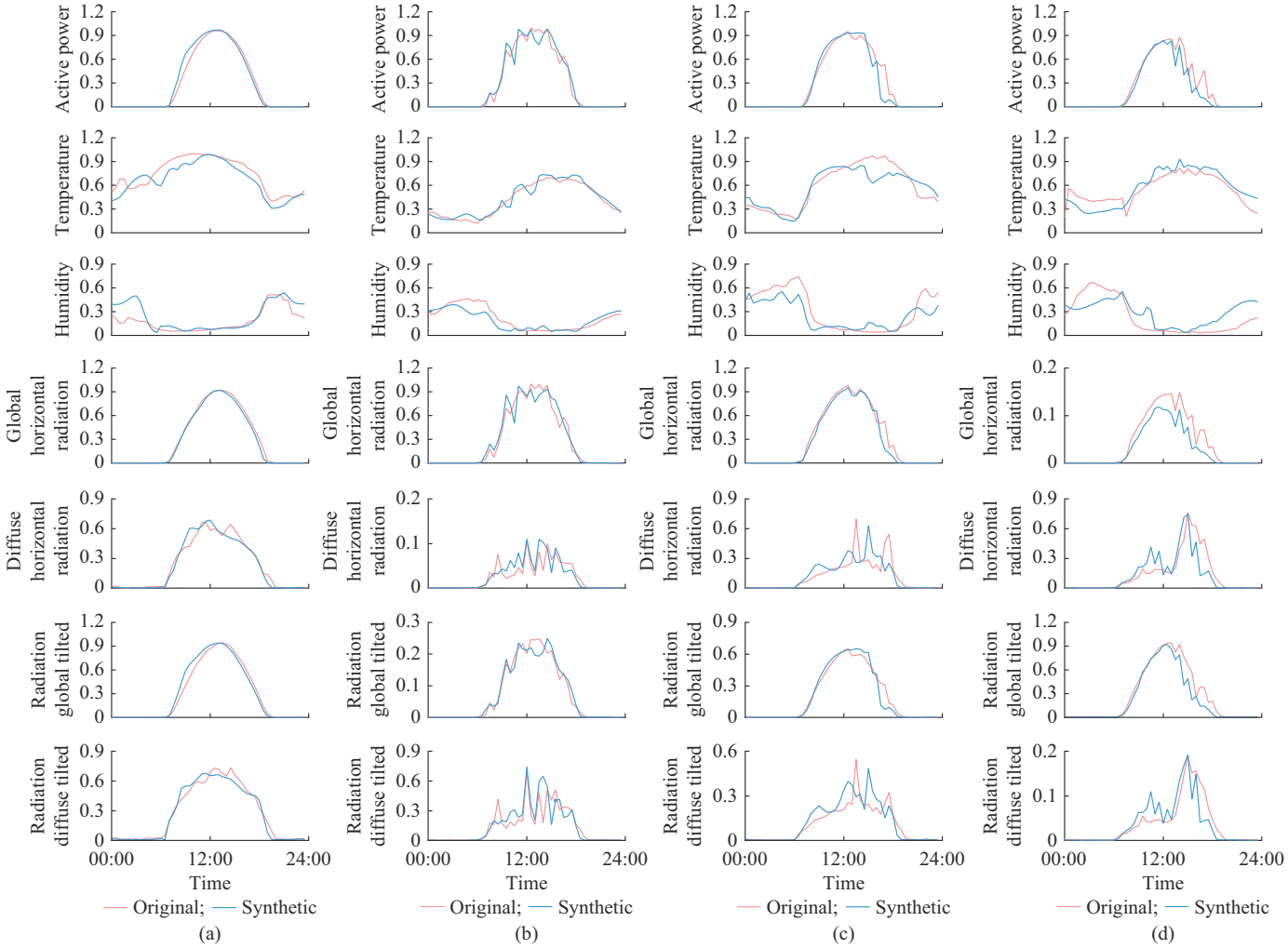

Fig. 4. New synthetic samples generated by TimeGAN under different meteorological conditions. (a) Sunny. (b) Cloudy. (c) Sunny to cloudy. (d) Sunny to drizzly.

This paper uses gated recurrent units (GRUs) and LSTM networks to build TimeGAN for learning the data distribution of a daily 48×7 tensor. This means there are 48 time points of PV data per day, and each time point contains PV power and 6 meteorological information data. The input of the sequence generator is a 72×1 dimensional noise conforming to the Gaussian normal distribution, and the output is a 48×24 tensor, for which the output of the sequence generator is transformed into a 48×7 tensor of synthetic data by the recovery function. The input of the sequence discriminator is the PV data of 48×7 tensor (at this point, PV data contains both synthetic and original data), which represents the input true and false samples, and the output is the discriminated result of one dimension. The structural parameters of TimeGAN are shown in Table I.

TABLE I
STRUCTURAL PARAMETERS OF TIMEGAN

Name of structure	Network	Parameter	Numerical value
Sequence generator	GRU	Number of hidden layer neurons	Two layers: 24, 24
		Number of output layer neurons	24
		Activation function	ReLU
Sequence discriminator	GRU	Number of hidden layer neurons	Two layers: 24, 24
		Number of output layer neurons	1
		Activation function	Sigmoid
Embedding function	LSTM	Number of hidden layer neurons	Two layers: 24, 24
		Number of output layer neurons	24
		Activation function	ReLU
Recovery function	LSTM	Number of hidden layer neurons	Two layers: 24, 24
		Number of output layer neurons	7
		Activation function	ReLU

In addition, to compare the similarity and diversity of the newly generated samples and the original samples, the new samples are analyzed using PCA and t-SNE, as shown in Figs. 5 and 6, respectively. From the PCA plot, it can be observed that the new samples generated by TimeGAN overlap well with the original sample data under different meteorological conditions. Some of the sample points are evenly distributed around the original sample points, which indicates that the newly generated synthetic data are similar to the original data and have sufficient diversity. From the t-SNE plot, we see that the original samples are uniformly distributed among the new synthetic samples, which indicates that TimeGAN generates high-quality samples that match the data distribution of the original sample training set.

IV. DCN-BILSTM-AM HYBRID NETWORK MODEL

When used for PV power prediction, historical PV data contain a large amount of potential information that can play a crucial role in predicting future PV power generation.

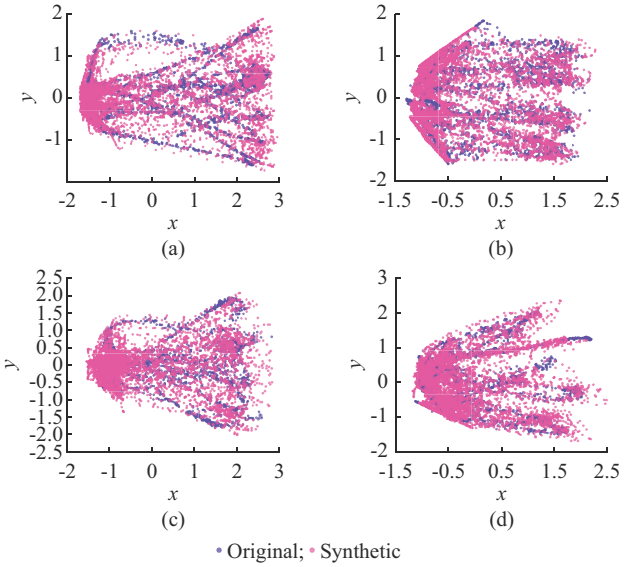

Fig. 5. Diversity and similarity analysis using PCA under different meteorological conditions. (a) Sunny. (b) Cloudy. (c) Sunny to cloudy. (d) Sunny to drizzly.

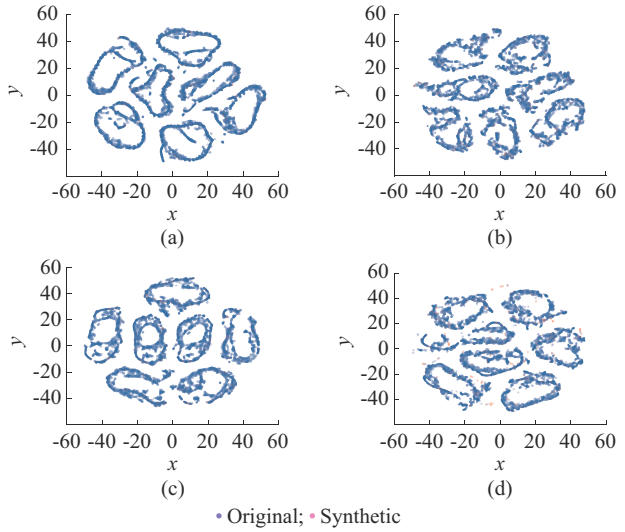

Fig. 6. Diversity and similarity analysis using t-SNE under different meteorological conditions. (a) Sunny. (b) Cloudy. (c) Sunny to cloudy. (d) Sunny to drizzly.

In this paper, DCN is used to obtain higher-order feature cross information, which exploits the concept of residuals to retain the low-dimensional feature information and to solve the problem of network performance degradation due to the increase in the number of network layers. Subsequently, BiLSTM is used to mine the intrinsic connection between the PV power output at the current moment and the historical timing information, to extract the temporal correlations between the data.

Finally, an AM is used to enhance the contribution of the important information to the final prediction results, so that the model can more easily capture the information on the long-range interrelated features in the sequence. To this end, a hybrid prediction model incorporating BiLSTM and AM, based on the framework of the DCN, is proposed to improve the accuracy of short-term PV power prediction through the

effective combination of multiple structures.

A. Principle of DCN Model

DCN [28] consists of two parts, a cross network and a deep network, which are used to obtain higher-order explicit and implicit feature combinations, respectively.

The cross network consists of several cross layers that can perform explicit feature cross-learning in an efficient manner on the original features of the input. The deep network is a feedforward neural network composed of multiple fully connected layers with strong nonlinear expression capability and is used here to perform implicit feature cross-learning on the original features of the input.

B. Principles of BiLSTM Model

As a variant of the RNN, LSTM gives better performance when dealing with long-time sequences, and can effectively solve the gradient explosion problem faced in the backpropagation step of RNNs when the time sequence is too long.

However, the information obtained by the LSTM is extracted before the moment of output, and does not take advantage of the reverse information. In a prediction made by a model, the output information for the current moment is not only related to past information, but may also be related to future information. Hence, we propose the use of a BiLSTM network for PV power prediction, as this can better capture the bi-directional sequence information by considering the changing patterns of previous and subsequent data simultaneously.

C. Principle of AM

AM [29] was developed to simulate the human brain's attention to different things at any given moment by highlighting key information and ignoring invalid information through the assignment of probabilities. The fundamental aim is to enhance the impact of important information on the final output of the model to allow for the effective mining of serial long-range data features with temporal correlations. An AM is introduced here to enhance the influence of important information by assigning different weights to the hidden states in BiLSTM, and hence to improve the training accuracy of the model.

D. Construction of DCN-BiLSTM-AM Hybrid Network

The proposed DCN-BiLSTM-AM hybrid network, as shown in Fig. 7, mainly consists of an input layer, a DCN layer, a BiLSTM layer, an AM layer, and an output layer. The proposed network takes PV data as input, extracts explicit and implicit higher-order feature information through the cross and deep networks in DCN, respectively, and then stitches the extracted feature information through the stacking layer to form the input to BiLSTM. The BiLSTM and AM layers then learn the intrinsic connections between PV power data from the extracted features and assign corresponding weights to different information to achieve PV power prediction. Finally, the final prediction results are generated by the output layer. The internal parameters of the DCN-BiLSTM-AM hybrid network model are shown in Table II.

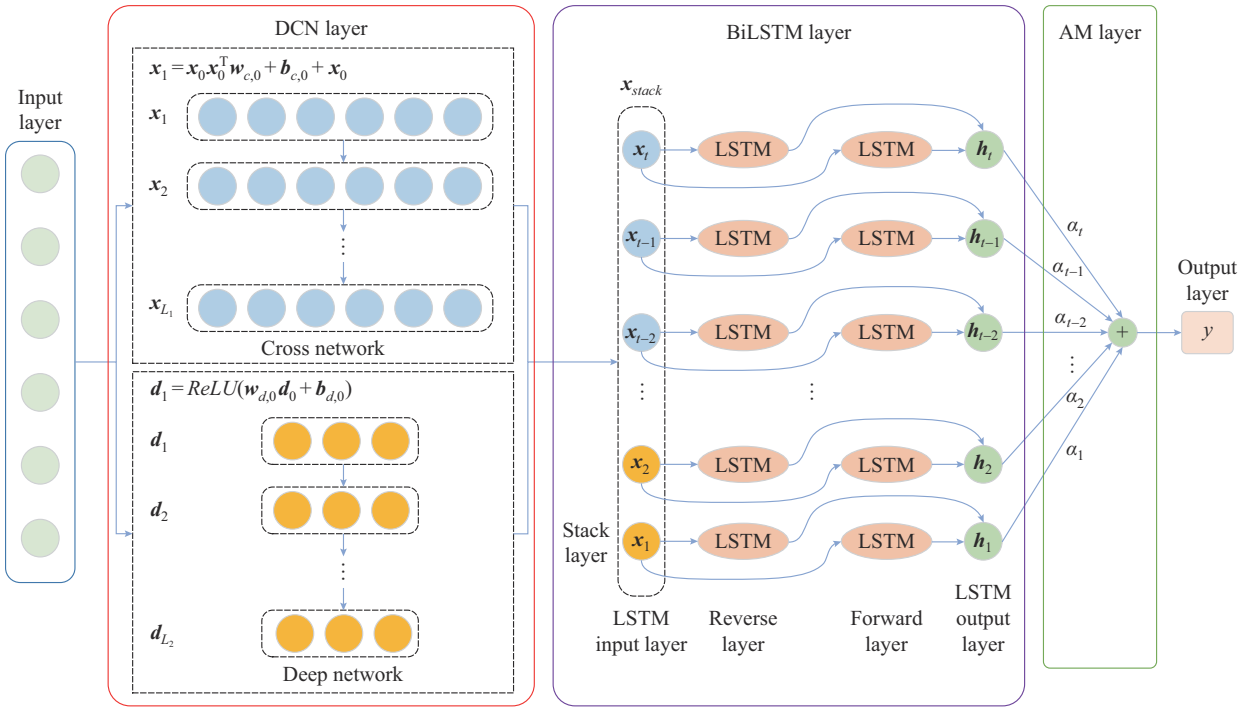

Fig. 7. Proposed DCN-BiLSTM-AM hybrid network framework.

The operation of the DCN-BiLSTM-AM hybrid network can be summarized as follows.

1) Input layer. The input layer accepts historical PV data as the input to the hybrid network model.

2) DCN layer. The DCN layer mainly extracts the features from the input historical PV data. After cross-combining the input features by DCN, the extracted feature information is stitched by stack layers to obtain the output $\boldsymbol{x}_{stack}$ of DCN, which can be expressed as:

TABLE II
INTERNAL PARAMETERS OF DCN-BILSTM-AM HYBRID NETWORK MODEL

Name of structure	Parameter	Numerical value
DCN (deep network)	Number of neurons	Three layers: 36, 32, 28
	Activation function	ReLU
DCN (cross network)	Number of cross network layers	3
Concatenate	Axis	−1
BiLSTM	Number of hidden layer neurons	Two layers: 18, 12
AM	Activation function	Softmax
Fully-connected layer	Number of neurons	10
	Activation function	ReLU
Output layer	Number of neurons	1
	Activation function	ReLU

$$\boldsymbol{x}_{L_1}=\boldsymbol{x}_0\boldsymbol{x}_{L_1-1}^{\mathrm{T}}\boldsymbol{w}_{c,L_1-1}+\boldsymbol{b}_{c,L_1-1}+\boldsymbol{x}_{L_1-1} \tag{2}$$

$$\boldsymbol{d}_{L_2}=ReLU(\boldsymbol{w}_{d,L_2-1}\boldsymbol{d}_{L_2-1}+\boldsymbol{b}_{d,L_2-1}) \tag{3}$$

$$\boldsymbol{x}_{stack}=concat(\boldsymbol{x}_{L_1},\boldsymbol{d}_{L_2}) \tag{4}$$

where $\boldsymbol{x}_{L_1}$ and $\boldsymbol{x}_{L_1-1}$ denote the outputs of layers L_1 and L_1-1 of the cross network, respectively; $\boldsymbol{d}_{L_2}$ and $\boldsymbol{d}_{L_2-1}$ denote the outputs of layers L_2 and L_2-1 of the deep network, respectively; $\boldsymbol{w}_{c,L_1-1}$, $\boldsymbol{b}_{c,L_1-1}$ and $\boldsymbol{w}_{d,L_2-1}$, $\boldsymbol{b}_{d,L_2-1}$ are the weights and biases of layers L_1-1 and L_2-1, and the subscripts c and d represent cross and deep network network, respectively; and *concat* denotes the stitching operation to be performed on the output sequence.

3) BiLSTM layer. The feature information $\boldsymbol{x}_{stack}$ extracted by DCN is used as the input to BiLSTM. The extracted feature information is fully learned by using BiLSTM to capture the effect of bidirectional information on the final prediction results and consider the changing patterns of the data before and after the time series. The output of BiLSTM is represented as:

$$\boldsymbol{h}_t=concat(\boldsymbol{h}_{ta},\boldsymbol{h}_{tb}) \tag{5}$$

where $\boldsymbol{h}_t$ denotes the output of the BiLSTM hidden layer at moment t; and $\boldsymbol{h}_{ta}$ and $\boldsymbol{h}_{tb}$ denote the outputs of the forward and inverse LSTM hidden layers, respectively.

4) AM layer. The output of $\boldsymbol{h}_t$ is used as the input to the AM layer, and the weights of different input features are calculated using AM to highlight important information. The output $\boldsymbol{s}_t$ of AM at moment t can be expressed as:

$$\boldsymbol{s}_t=\sum_{t=1}^{i}\alpha_t\boldsymbol{h}_t \tag{6}$$

$$\alpha_t=\frac{\exp(\boldsymbol{e}_t)}{\sum_{j=1}^{t}\boldsymbol{e}_j} \tag{7}$$

$$\boldsymbol{e}_t=\boldsymbol{u}_a\tanh(\boldsymbol{w}_a\boldsymbol{h}_t+\boldsymbol{b}_a) \tag{8}$$

where $\boldsymbol{e}_t$ is the value of the probability distribution of the AM over the BiLSTM output at moment t; $\boldsymbol{u}_a$ and $\boldsymbol{w}_a$ are the weighting factors; and $\boldsymbol{b}_a$ is the bias.

5) Output layer. The output layer takes the output of the AM layer as input and computes the final prediction y_i via the fully-connected layer. In this paper, we adopt the ReLU function as the activation function of the fully-connected layer. Therefore, y_i can be expressed as:

$$y_i=ReLU(w_o s_i+b_o) \tag{9}$$

where w_o and b_o denote the weights and biases of the output layer, respectively.

V. WOA OPTIMIZATION

A. Optimizing Model Hyperparameters

The prediction accuracy of DCN-BiLSTM-AM hybrid network is affected by hyperparameters such as the number of iterations, the number of hidden layer neurons of BiLSTM, the learning rate, and the number of cross and deep layers of the DCN model. To prevent the hybrid network from falling into local optimal solutions in the prediction process, which will affect the final prediction results of the hybrid network model, we carry out hyperparameter optimization of the model using WOA.

WOA, a meta-revelation optimization algorithm, was proposed to simulate the bubble hunting of humpback whales [30]. The bubble hunting strategy has the local exploitation ability of whale shrinkage surrounding prey and the spiral hunt method as well as the global optimal search features of a random search. It has a simple mode of operation, strong search ability, fast convergence speed, and can jump out of a local optimal solution. In this paper, we therefore optimize the parameters of the hybrid network model using WOA, in which each whale represents an optimization strategy, and the number of dimensions of individual whale positions corresponds to the number of parameters to be optimized in the hybrid network model. The optimal hyperparameters of the hybrid network are obtained as the individual whales continuously update their positions.

B. Steps Used for Parameter Optimization

WOA was used to optimize the parameters of the hybrid network, as shown in Fig. 8, with the following steps. In Fig. 8, P is the probability of the predation mechanism, a random number with value domain [0,1]; and A is the vector coefficient for selecting the path to search for prey.

Step 1: initialize the whale population size M, the maximum number of iterations $T_{\max}$, and the upper and lower bounds of the search range.

Step 2: initialize the number of iterations in the DCN-BiLSTM-AM hybrid network, the number of the hidden layer neurons of the BiLSTM, the learning rate, and the number of network layers in the cross and deep layers of the DCN.

Step 3: compute the fitness value of individual whale according to the fitness function. Based on the magnitude of the fitness value, M individual whales are selected as the initial population.

Step 4: compute the fitness value of each of the M individual whales. The position of the individual whale with the smallest fitness value is taken as the optimal position.

Step 5: update the position of the next generation of individual whales.

Step 6: continuously optimize the hyperparameters of the

hybrid network until the maximum number of iterations is reached, to obtain the optimal hyperparameters of the DCN-BiLSTM-AM hybrid network.

Step 7: substitute the obtained optimal hyperparameters into the DCN-BiLSTM-AM hybrid network for training and testing. In this paper, the hyperparameter optimization time in different cases is kept around 40 min, and the number of iterations of WOA optimization is 20.

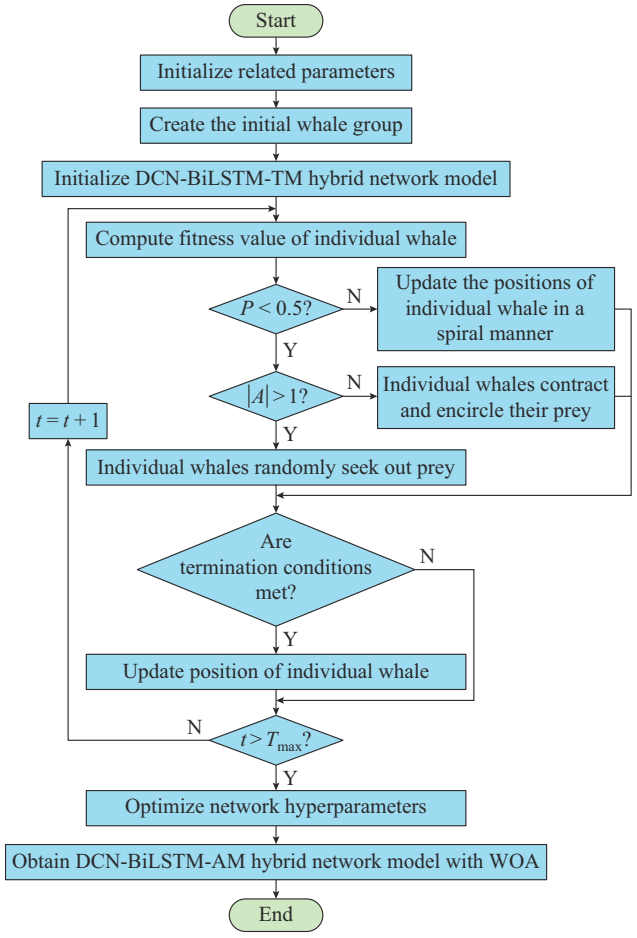

Fig. 8. Flow chart of hybrid network based on WOA optimization.

VI. Simulation Validation

In this paper, the mean absolute error (MAE) E_{MAE} and root mean square error (RMSE) E_{RMSE} are used to evaluate the model prediction accuracy after obtaining the PV power prediction values, as shown in (10) and (11), respectively.

$$E_{MAE}=\frac{1}{m}\sum_{t=1}^{m}\left|y_t-y_{st}\right| \tag{10}$$

$$E_{RMSE}=\sqrt{\frac{1}{m}\sum_{t=1}^{m}(y_t-y_{st})^2} \tag{11}$$

where m is the number of prediction points; y_t is the predicted value of the PV output power at moment t; and y_{st} is the measured value of PV output power at moment t.

A. Analysis of Prediction Results

To analyze the performance of the hybrid network model proposed in this paper, the summer data from an Australian PV power station drawn from December 2014 to February 2015 are used as the original sample data. During this period, sample data of 16 days are selected for the four meteorological conditions, of which 15 are used as the original sample training set and one as the test set, to build the network model. In this paper, we use historical data for single-step prediction and predict the PV output power 1 hour in advance, then gradually output the prediction results for 24 hours with a prediction interval of 30 min. The normalized PV power prediction curves are obtained by predicting the PV output power under different meteorological conditions, as shown in Fig. 9. The prediction errors of the proposed hybrid network model under different meteorological conditions are shown in Table III.

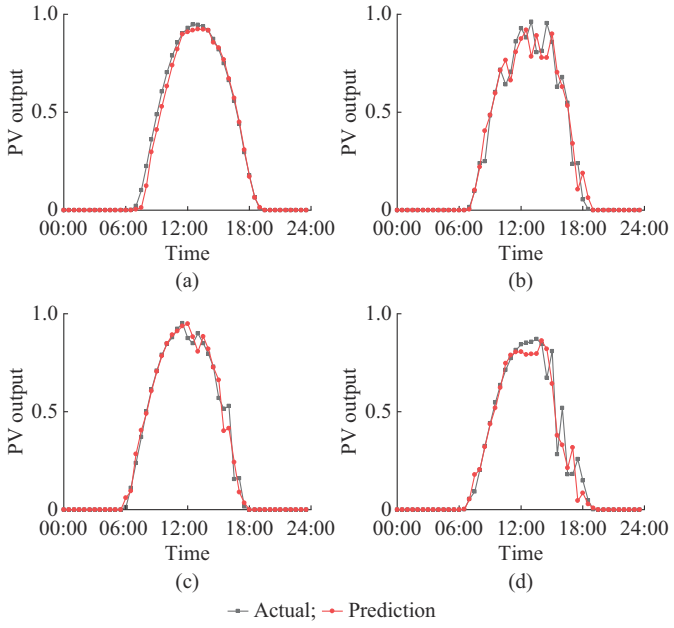

Fig. 9. PV power prediction results under different meteorological conditions. (a) Sunny. (b) Cloudy. (c) Sunny to cloudy. (d) Sunny to drizzly.

TABLE III
Prediction Errors of Hybrid Network Model Under Different Meteorological Conditions

Condition	E_{RMSE} (%)	E_{MAE} (%)
Sunny	0.0655	0.0351
Cloudy	0.1279	0.0749
Sunny to cloudy	0.0812	0.0483
Sunny to drizzly	0.1437	0.0761

In Fig. 9, the trend of the PV power prediction curve for sunny days largely coincides with the actual PV output. However, for cloudy weather, the prediction values do not fit well with the actual values due to the uncertainty in the cloud cover. There are more fluctuations in the prediction curve after 12:00 for sunny to cloudy days than for sunny days due to the changes in the weather. Abrupt changes in weather, lack of light intensity, and uncertainty due to changes in other meteorological factors under sunny to drizzly conditions mean that the actual output power curve is more volatile compared with that of sunny to cloudy weather, and we

can observe the deviation of a certain magnitude range between the prediction and actual values at some instants on the graph.

B. Verification of Model Validity and Effectiveness

To verify the validity and effectiveness of the proposed model, we compare the prediction effects of six different models under four meteorological conditions (sunny, cloudy, sunny to cloudy, and sunny to drizzly), as follows.

Model 1: the proposed DCN-BiLSTM-AM hybrid network model with WOA optimization in this paper.

Model 2: the BP model.

Model 3: the LSTM model.

Model 4: the BiLSTM model.

Model 5: the BiLSTM-AM model.

Model 6: the DCN-BiLSTM-AM hybrid network model.

The prediction curves and error results for the above models are shown in Figs. 10 and 11, respectively.

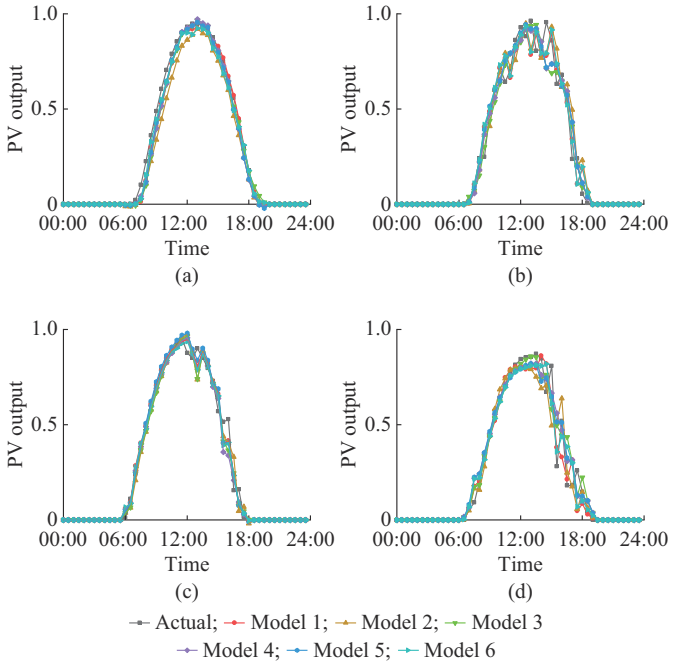

Fig. 10. Short-term PV power prediction results of different models under different meteorological conditions. (a) Sunny. (b) Cloudy. (c) Sunny to cloudy. (d) Sunny to drizzly.

It can be observed that model 6 has better prediction accuracy and the minimum error value for short-term PV power prediction under different meteorological conditions compared to models 2-4. Meanwhile, the prediction accuracy of model 6 is improved to some extent by introducing the WOA.

By combining the data in Fig. 11, it can be observed that the prediction accuracy of model 1 is higher than the other five prediction models. On sunny days, for example, the RMSE of model 1 is reduced by 52.22%, 19.53%, 14.15%, 6.96%, and 2.53%, respectively, compared with that of models 2-6, while the MAE of model 1 is reduced by 65.42%, 26.88%, 24.03%, 23.36%, and 14.18%, respectively.

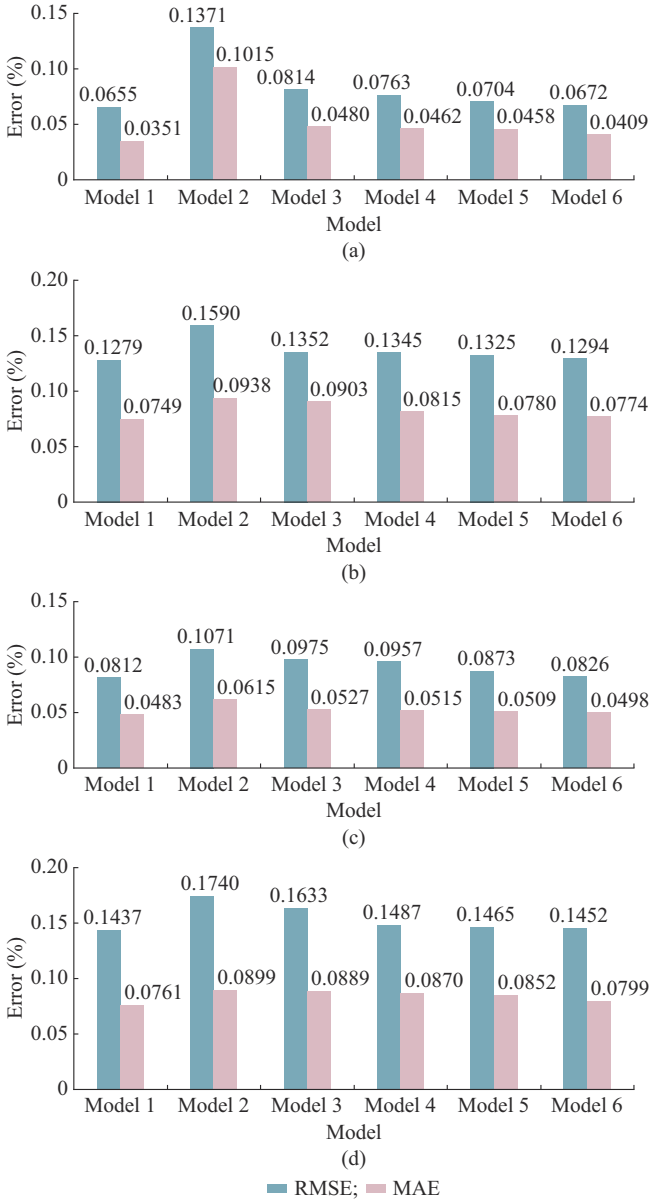

Fig. 11. PV power prediction errors of different models under different meteorological conditions. (a) Sunny. (b) Cloudy. (c) Sunny to cloudy. (d) Sunny to drizzly.

C. Validation of PV Data Generated by TimeGAN

To verify the effectiveness of the data generated by the TimeGAN for actual predictions, sample data are generated to enhance the dataset to improve the prediction accuracy of the TimeGAN. TimeGAN is used to expand the original sample training set by 20%, 40%, and 60% of the data volume for each weather condition of the PV plant in summer; in other words, we expand the sample data for three, six, and nine days, respectively. The original training set is merged with the new sample data to train a hybrid network model to predict the PV output power under different meteorological conditions. Meanwhile, nine days of sunny, cloudy, sunny to cloudy, and sunny to drizzly weather are selected from the summer data of this PV station from December 2015 to February 2016. Original samples are used to expand the original sample training set for PV power prediction as a comparison

validation experiment with the synthetic expanded training set. The different schemes used for expansion of training sets are shown in Table IV.

TABLE IV
SCHEMES USED FOR EXPANSION OF TRAINING SET

Case	Scheme	Data
Case A		Raw data
Case B	Expansion of 20%	Synthetic data
Case C		Original data
Case D	Expansion of 40%	Synthetic data
Case E		Original data
Case F	Expansion of 60%	Synthetic data
Case G		Original data

To better demonstrate the prediction results, the output power prediction curves under sunny, cloudy, sunny to cloudy, and sunny to drizzly weather conditions for different expansion schemes in summer are shown in Fig. 12. The PV power prediction errors under different meteorological conditions are shown in Fig. 13.

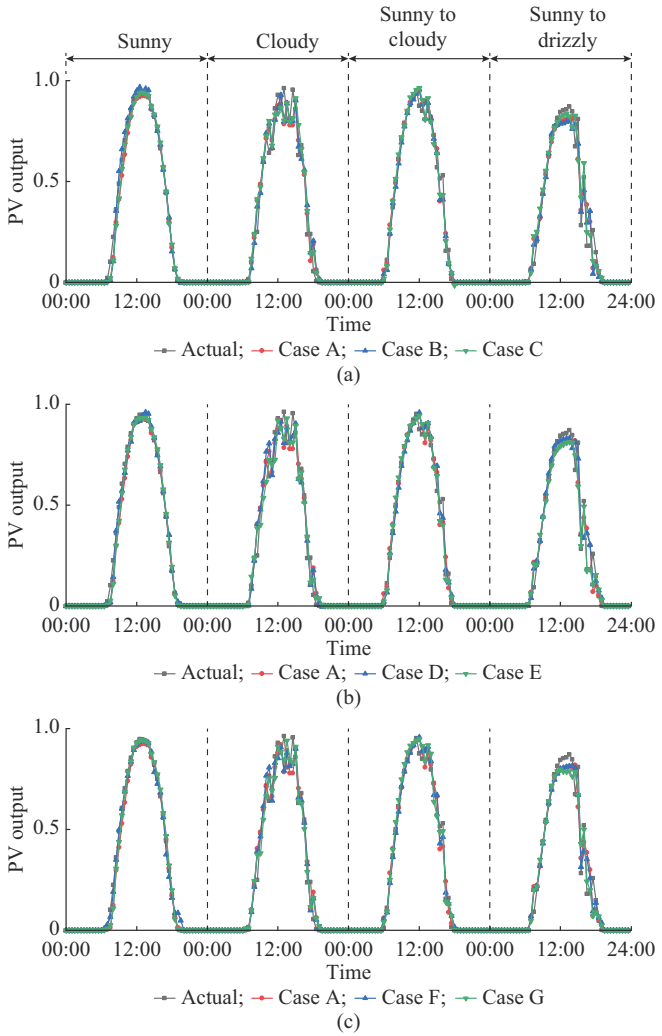

Fig. 12. PV power prediction curves under differnet weather conditions for different expansion schemes. (a) Expansion of 20%. (b) Expansion of 40%. (c) Expansion of 60%.

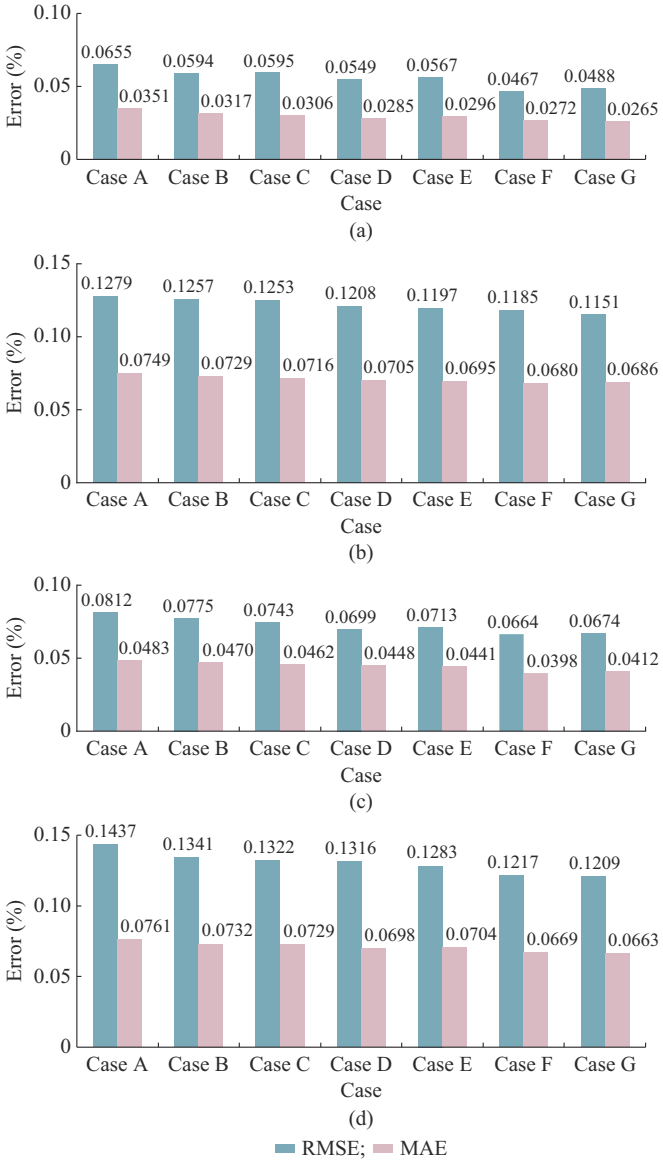

Fig. 13. PV power prediction errors of different expansion schemes under different meteorological conditions. (a) Sunny. (b) Cloudy. (c) Sunny to cloudy. (d) Sunny to drizzly.

In the proposed prediction model, we use TimeGAN to generate new samples to expand the original sample training set under different meteorological conditions and show that this could effectively improve the accuracy of prediction model for PV output power. We also find that the use of TimeGAN to expand the original training set under different meteorological conditions under the same circumstances gives prediction results very similar to those when the training set is expanded with original data.

Table V shows a comparison of the results for the prediction errors in Cases A-G under different meteorological conditions. The relative optimization rates (RORs) E_{RMSE} and E_{MAE} in the table denote the degree to which each method (Cases B-G) is optimized with regard to the original solution (Case A).

As shown in Table V, compared with the original data prediction, the prediction accuracy is significantly improved

when the original data are expanded by 20%, 40%, and 60% of the data volume using TimeGAN. As the amount of data expanded by TimeGAN from the original sample training set becomes larger, the model predicts the PV output power with higher accuracy. It can also be observed that the improvement in prediction accuracy is approximately the same for both training sets (i. e., training sets expanded with synthetic and original data) under meteorological conditions with the same expansion ratio. Taking the data volume expanded by 20% on sunny days as an example, RMSE decreases by 9.31% and MAE decreases by 9.69% during the synthetic data expansion.

TABLE V
COMPARISON OF RESULTS FOR PREDICTION ERRORS IN CASES A-G UNDER DIFFERENT METEOROLOGICAL CONDITIONS

Case	Sunny		Cloudy		Sunny to cloudy		Sunny to drizzly	
	E_{RMSE} (%)	E_{MAE} (%)	E_{RMSE} (%)	E_{MAE} (%)	E_{RMSE} (%)	E_{MAE} (%)	E_{RMSE} (%)	E_{MAE} (%)
Case A	0	0	0	0	0	0	0	0
Case B	–9.31	–9.69	–1.72	–2.67	–4.56	–2.69	–6.68	–3.81
Case C	–9.16	–12.82	–2.03	–4.41	–8.50	–4.35	–8.00	–4.20
Case D	–16.18	–18.80	–5.55	–5.87	–13.92	–7.25	–8.42	–8.28
Case E	–13.44	–15.66	–6.41	–7.21	–12.19	–8.70	–10.72	–7.49
Case F	–28.70	–22.51	–7.35	–9.21	–18.23	–17.60	–15.31	–12.09
Case G	–25.50	–24.50	–10.01	–8.41	–17.00	14.70	–15.87	–12.88

During real data expansion, RMSE decreases by 9.16% and MAE decreases by 12.82%. It can be observed that the accuracy improvement of PV power prediction is approximately the same for the same expansion ratio, regardless of whether synthetic data expansion or real data expansion is used, which proves the feasibility of using synthetic data instead of real data expansion for training. This indicates that expanding the samples with synthetic data can achieve the same prediction accuracy as using original data, which also demonstrates the effectiveness of the proposed method.

VII. CONCLUSION

A hybrid network prediction model based on data augmentation has been proposed for new PV plants with insufficient original sample data, and the correctness and validity of this model are verified. The conclusions are as follows.

1) When original sample data are scarce, TimeGAN can generate new samples of high quality that conform to the sample distribution pattern, thereby enhancing the sample dataset for different meteorological conditions and preserving the temporal correlation between sample features.

2) WOA is used to maximize the capability of the hybrid network model, and to effectively improve the accuracy and robustness of PV power prediction. Compared with other prediction models, it yields a higher prediction accuracy.

3) The proposed model has good applicability and application prospects for the power prediction of new PV plants with few samples.

Due to the complex meteorological changes of turning weather, this is a feasible direction for future research to propose a more refined way to classify weather types when expanding different types of weather data to further improve the effectiveness of synthetic data.

REFERENCES

[1] H. Sharadga, S. Hajimirza, and R. S. Balog, "Time series forecasting of solar power generation for large-scale photovoltaic plants," *Renewable Energy*, vol. 150, pp. 797-807, May 2020.

[2] M. N. Akhter, S. Mekhilef, H. Mokhlis *et al.*, "A hybrid deep learning method for an hour ahead power output forecasting of three different photovoltaic systems," *Applied Energy*, vol. 307, pp. 1-19, Jan. 2022.

[3] G. G. Platero, M. G. Casado, M. P. García *et al*., "CECRE: supervision and control of Spanish renewable energies in the last 15 years," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 2, pp. 269-276, Mar. 2022.

[4] K. Wang, X. Qi, and H. Liu., "A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network," *Applied Energy*, vol. 251, pp. 1-14, Oct. 2019.

[5] B. Gu, H. Shen, and X. Lei *et al.*, "Forecasting and uncertainty analysis of day-ahead photovoltaic power using a novel forecasting method," *Applied Energy*, vol. 299, pp. 1-14, Jun. 2021.

[6] V. Kushwaha and N. M. Pindoriya, "A SARIMA-RVFL hybrid model assisted by wavelet decomposition for very short-term solar PV power generation forecast," *Renewable Energy*, vol. 140, pp. 124-139, Sept. 2019.

[7] N. Q. Nguyen, L. D. Bui, B. V. Doan *et al.*, "A new method for forecasting energy output of a large-scale solar power plant based on long short-term memory networks a case study in Vietnam," *Electric Power Systems Research*, vol. 199, pp. 1-14, Jun. 2021.

[8] H. Zhang, J. Shi, and C. Zhang, "A hybrid ensembled double-input-fuzzy-modules based precise prediction of PV power generation," *Energy Reports*, vol. 8, pp. 1610-1621, Mar. 2022.

[9] D. Korkmaz, "SolarNet: a hybrid reliable model based on convolutional neural network and variational mode decomposition for hourly photovoltaic power forecasting," *Applied Energy*, vol. 300, pp. 1-20, Jul. 2021.

[10] Y. Ma, Q. Lv, R. Zhong *et al.*, "Short-term photovoltaic power forecasting method based on irradiance correction and error forecasting," *Energy Reports*, vol. 7, pp. 5495-5509. Nov. 2021.

[11] Y. Zhang and L. Kong, "Photovoltaic power prediction based on hybrid modeling of neural network and stochastic differential equation," *ISA Transactions*, vol. 128, pp. 181-206, Sept. 2022.

[12] S. Bhavsar, R. Pitchumani, and M. A. Ortega-Vazquez, "Machine learning enabled reduced-order scenario generation for stochastic analysis of solar power forecasts," *Applied Energy*, vol. 293, pp. 1-11, Apr. 2021.

[13] J. Zeng and W. Qiao, "Short-term solar power prediction using a support vector machine," *Renewable Energy*, vol. 52, pp. 118-127, Apr. 2013.

[14] N. M. M. Bendaoud, N. Farah, and S. B. Ahmed, "Comparing generative adversarial networks architectures for electricity demand forecasting," *Energy Build*, vol. 247, pp. 1-12, Jun 2021.

[15] W. Li, J. Chen, Z. Wang *et al.*, "IFL-GAN: improved federated learning generative adversarial network with maximum mean discrepancy model aggregation," *IEEE Transactions on Neural Networks and Learning Systems*, doi: 10.1109/TNNLS.2022.3167482

[16] S. Asre and A. Anwar, "Synthetic energy data generation using time

variant generative adversarial network," *Electronics*, vol. 11, no. 3, pp. 1-14, Feb. 2022.
[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp.139-144, Nov. 2020.
[18] B. Franci and S. Grammatico. "Training generative adversarial networks via stochastic nash games," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 3, pp. 1319-1328, Mar. 2023.
[19] B. Zhao, T. Qian, and W. Tang *et al.*, "A data-enhanced distributionally robust optimization method for economic dispatch of integrated electricity and natural gas systems with wind uncertainty," *Energy*, vol. 243, pp. 1-15, Mar. 2022.
[20] F. Wang, Z. Zhang, C. Liu *et al*., "Generative adversarial networks and convolutional neural networks based weather classification model for day ahead short-term photovoltaic power forecasting," *Energy Conversion and Management*, vol. 181, pp. 443-462, Feb. 2019.
[21] S. Takahashi, Y. Chen, and K. Tanaka-Ishii, "Modeling financial time-series with generative adversarial networks," *Physica A – Statistical Mechanics and its Applications*, vol. 527, pp. 1-12, Aug. 2019.
[22] W. V. Deventer, E. Jamei, and G. Thirunavukkarasu *et al.*, "Short-term PV power forecasting using hybrid GASVM technique," *Renewable Energy*, vol. 140, pp. 367-379, Sept. 2019.
[23] H. Zang, L. Liu, L. Cheng *et al.*, "Short-term global horizontal irradiance forecasting based on a hybrid CNN-LSTM model with spatiotemporal correlations," *Renewable Energy*, vol. 160, pp. 26-41, Nov. 2020.
[24] S. Netsanet, Z. Dehua, Z. Wei *et al.*, "Short-term PV power forecasting using variational mode decomposition integrated with ant colony optimization and neural network," *Energy Reports*, vol. 8, pp. 2022-2025, Nov. 2022.
[25] X. Ma and X. Zhang, "A short-term prediction model to forecast power of photovoltaic based on MFA-Elman," *Energy Reports*, vol. 8, pp. 495-507, Jul. 2022.
[26] M. Cheng, F. Fang, I. Navon *et al.*, "A real-time flow forecasting with deep convolutional generative adversarial network: application to flooding event in Denmark. Phys," *Physics of Fluids*, vol. 33, no. 5, pp. 1-14, May 2021.
[27] J. Yoon, D. Jarrett, and M. van der Schaar, "Time-series generative adversarial networks," in *Proceedings of the 33rd Conference on Advances in Neural Information Processing Systems*, Vancouver, Canada, Dec. 2019, pp. 1-11.
[28] R. Wang, B. Fu, G. Fu *et al.*, "Deep & cross network for Ad click predictions," in *Proceeding ACM-SIGKDD International 23rd Conference on Knowledge Discovery and Data Mining*, Halifax, Canada, Aug. 2017, pp. 1-7.
[29] Neeraj, J. Mathew, and R. K. Behera, "EMD-Att-LSTM: a data-driven strategy combined with deep learning for short-term load forecasting," *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 5, pp. 1229-1240, Sept. 2022.
[30] S. Mirjalili and A. Lewis, "The whale optimization algorithm," *Advances in Engineering Software*, vol. 95, pp. 51-67, May 2016.

Shangpeng Zhong received the B.S. degree in automation from Wanjiang College of Anhui Normal University, Wuhu, China, in 2018, and the M.S. degree in physics from Anhui Normal University, Wuhu, China, in 2021. He is currently pursuing the Ph.D. degree in electrical engineering at Hefei University of Technology, Hefei, China. His research interests include integrated energy systems and renewable energy technologies.

Xiaoming Wang received the B.S. degree in electrical engineering from Hefei University of Technology, Hefei, China, in 2011, and the M.S. degree in electrical engineering from North China Electric Power University, Beijing, China, in 2014. He is currently pursuing the Ph.D. degree in electrical engineering at Hefei University of Technology, Hefei, China. And he is a Senior Engineer at the State Grid Anhui Electric Power Research Institute, Hefei, China. His research interests include photovoltaic power generation, new energy consumption, and power quality.

Bin Xu received the M.S. degrees in electrical engineering and automation from Hefei University of Technology, Hefei, China, in 2012. He is now a Senior Engineer of State Grid Anhui Electric Power Research Institute, Hefei, China. His research interests include distribution network system planning and design techniques.

Hongbin Wu received the B.Sc., M.Sc. and Ph.D. degrees in electrical engineering from Hefei University of Technology, Hefei, China, in 1994, 1998 and 2005, respectively. He is now a Professor of Hefei University of Technology. His research interests include distributed generation and microgrids, active distribution networks, and integrated energy systems.

Ming Ding is now a Professor of Hefei University of Technology, Hefei, China. His research interests include power system reliability, planning, power market, and renewable energy resource techniques.