Sustainable Energy, Grids and Networks 34 (2023) 101019

Contents lists available at ScienceDirect

Sustainable Energy, Grids and Networks

journal homepage: www.elsevier.com/locate/segan

Short-term Forecasting Approach Based on bidirectional long short-term memory and convolutional neural network for Regional Photovoltaic Power Plants

Gang Li [a,*], Shunda Guo [a], Xiufeng Li [b], Chuntian Cheng [a]

[a] Institute of Hydropower and Hydroinformatics, Dalian University of Technology, Dalian 116024, China
[b] Power Dispatching Control Center, Yunnan Power Grid Co. LTD, Kunming 650011, China

ARTICLE INFO

Article history:
Received 1 September 2022
Received in revised form 10 February 2023
Accepted 11 February 2023
Available online 14 February 2023

Keywords:
Regional photovoltaic power
Short-term forecasting
Neural network
Deep learning
Up-scaling method

ABSTRACT

Accurate photovoltaic (PV) generation output prediction is one of the effective ways to ensure the safe operation of power grid, develop reasonable dispatching plan and improve the efficiency of clean energy. With the large-scale operation of PV power plants in recent years, forecasting regional PV output becomes more significant. We proposed a short-term forecasting approach based on bidirectional long short-term memory and convolutional neural network (BiLSTM-CNN) for regional PV power plants. First, the k-means algorithm is used to divide power plants with similar generation characteristics into the same output subregion. Second, a representative power plant in each subregion is selected based on three correlation coefficients. Then, we develop a regional prediction model based on BiLSTM-CNN method. This model takes historical operation and meteorological data of the representative power plant as input, and takes the total subregional power generation as output. Finally, this short-term forecasting approach is tested using real data from PV power plants in Chuxiong and Dali region, Yunnan province, China. The comparison of numerical results shows this proposed method can effectively improve the short-term prediction accuracy of regional PV generation output.

1. Introduction

As a clean and renewable energy source, solar energy can reduce environmental pollution and has strong development potential [1]. PV power generation has become the priority development direction in the energy field of all countries in the world. For instance, China's installed capacity of PV power plants reached 306 million kW by the end of 2021, with the nationwide annual newly installed PV capacity is increased to a highest of 54.88 million kW. In order to achieve the Dual Carbon Goal, which is presented by Chinese President at the 75th UN General Assembly in 2020 [2], the installed PV power capacity is expected to increase to 3.55 billion kW by 2060. PV power will become the dominant power source on the power grid, which has a significant impact on the power grid security as well as the planning and operation cost [3,4].

Unlike traditional power sources, PV power generation is affected by various meteorological factors such as solar radiation, temperature, cloud coverage, and duration of light. So it is intermittent, variable and uncertain. This uncertainty will affect the stability of the power grid and increase the difficulty of feeding into the grid from PV generation. According to statistics from the National Energy Administration (NEA) of China, PV power generation capacity reached 325.9 billion kWh, an increase of 25.1% year-on-year, and photovoltaic power generation utilization rate of 98.0%, but the amount of solar energy wasted about 6.52 billion kWh. So, we need to improve the flexibility of the power grid to respond to fluctuations in power generation, so as to relieve the difficulty of PV consumption and protect the stability of the grid. Among the many measures to ensure grid flexibility, accurate forecasting of renewable energy generation is considered cost-effective [5]. Therefore, it is necessary to improve the prediction accuracy of PV power generation.

Currently, PV forecasting models are divided into physical models [6–10], statistical methods [11–14], and artificial intelligence algorithms [15–18]. These prediction models and methods are mostly based on the characteristics of PV power plants and data collection, which can improve the prediction accuracy. However, with large-scale PV power sources connected to the grid, developing an accurate prediction model for each PV power plant is more challenging because of repetitive work that different historical data are required for different prediction models. On the other hand, the integration of PV power plants into the regional power grid becomes more common, and the plant number in the

* Corresponding author.
E-mail addresses: glee@dlut.edu.cn (G. Li), gsd@mail.dlut.edu.cn (S.D. Guo), lixiufeng198362@163.com (X.F. Li), ctcheng@dlut.edu.cn (C.T. Cheng).

https://doi.org/10.1016/j.segan.2023.101019
2352-4677/

region is increasing significantly. In addition, the grid dispatching department becomes more interested in regional PV power prediction, since the accurate prediction of relatively large and more stable regional PV generation can improve power dispatching scheme to ensure power grid stability and efficient energy consumption [19,20]. So it is necessary to conduct further research on regional PV power forecasts for better decision-making [21].

In recent literature, there are two main methods for regional PV power prediction, namely the superposition method and the upscaling method [22]. The superposition method predicts the output value of each PV power plant in the region, and then the predicted value of each power plant is superimposed to obtain the total output of the region [23]. For regions with a large number of power plants, the calculation time, the data dimension and the workload will increase considerably. Due to prediction error accumulation of the single power plant and data magnitude, if the prediction value of regional power is directly superimposed by the prediction value of each power plant, the error of the regional prediction value could be rather large. Zhang et al. [20] proposed a convolution neural network (CNN) prediction model to improve the superimposed method, which uses the power station and meteorological data in the whole region as model input. Yu et al. [21] proposed a new regional prediction model by combining the improved CNN with non-linear quantile regression. These methods reduce the heavy workload of the superposition method, but still require complete regional data as input. The upscaling method requires relatively less data and workload. It selects several representative power plants in the region to model, and the predicted output of representative power plants are combined or extrapolated to obtain the regional prediction results. Saint-Drenan et al. [24,25] analyzed the influencing factors of the error from upscaling method, and proposed an alternative method base on probability. Pierro et al. [26] reduced model input by clustering, and then proposed two improved models to predict the regional output. Shaker et al. [27] proposed a prediction model of wavelet neural network based on fuzzy algorithm, which only needs the representative power stations data in the region. Other scholars combine neural networks with intelligent algorithms. Zhou et al. [28] proposed hybrid model (SDA-GA-ELM). It finds a training set similar to the predicted day by customized similar day analysis (SDA), and searches the optimal weight value of extreme learning machine (ELM) by genetic algorithm (GA). Although the intelligent algorithm can optimize the weight of the network and make the result more accurate, it still has the problem of falling into the local optimal solution. Above analyses show that the upscaling method is more suitable for regional PV output prediction. But, the selection of representative power plants and the modeling of their predicted output to the regional PV output have a great influence on the prediction results. This requires a long time-series of historical data, and relatively high data integrity. However, most PV power plants are connected to the power grid in the past few years that can only provide limited historical operation data. Therefore, it is necessary to study on forecasting the total generation output for regional PV power plants with limited data.

We propose a short-term regional prediction method based on Bi-directional Long-Short Term Memory and Convolution Neural Network (BiLSTM-CNN) using the upscaling method. The total regional PV output at various scales can be predicted with few data and simple process. Firstly, we divide the photovoltaic region into several sub-regions by clustering algorithm. In order to make the representative power station more effective, we use three indicators to verify. Then, the prediction steps are divided into two steps, the representative power station prediction, and regional prediction. In the regional prediction, the BiLSTM-CNN model is used instead of the mathematical method to expand the predicted value of the representative power station to the predicted value of the whole region. This simplifies the complex extrapolation calculation process, and also enables better learning of the complex nonlinear power relationship between the representative power station and the whole region, resulting in better regional prediction results. Therefore, we only need to predict the representative power station. And by improving the prediction accuracy of representative power stations, the prediction accuracy of the region can be further improved.

The contributions are as follows:

(1) Using k-means algorithm, the entire region is divided into several output subregions, and the PV power plants with similar generation characteristics are grouped together.

(2) A representative power plant in each divided output subregion is selected based on three correlation coefficients, and its representativeness is verified through some indicators.

(3) The short-term forecasting approach based on BiLSTM-CNN is established. The BiLSTM is used to extract the time feature of input data, and the CNN is used to derive the relationship between the representative power plant and subregional power grid for generation output and meteorological data. This method could increase the accuracy of the prediction and reduce the complexity of the calculation.

2. Establish PV output subregion

Although the meteorological conditions in the same region are roughly the same, power generation characteristics of every PV power plant will differ slightly. To make more accurate predictions, the PV output subregions are established by using the clustering algorithm, where the characteristics of all PV power plants in each subregion are similar. The factors involved in clustering are generation output, minimum and maximum generation output, temperature, air pressure, rainfall, cloud cover, etc. We choose k-means method for clustering. Since the k-means method is heavily depend on the initial cluster center, we use the k-means++ method to initialize the centroid [29]. However, this method still needs to determine the number K of the output subregion. We use the minimum sum of squared errors as the evaluation function:

$$\phi = \sum_{i=1}^{K} \sum_{x \in \chi_i} \|x - c_i\|^2 \tag{1}$$

where K represents the number of categories required, χ_i is the sample set of class i, x is the sample in χ_i, c_i is the cluster center of the category i.

Different K values can be obtained through the k-means method, and corresponding evaluation function values can be obtained. We can then draw a scatter diagram of the K values and evaluation function values, and set the K value to the one with an obvious inflection point. At the same time, we also calculate the silhouette coefficients and Davies–Bouldin index, together with the above method as a reference to determine the K value [30]. The silhouette coefficient is an index to evaluate the clustering effect. It can be understood as an index describing the contour clarity of each category after clustering. The larger the contour coefficient is, the better the clustering effect is. And the Davies–Bouldin index is known as the classification accuracy index. It comprehensively considers the similarity of intra-class samples and the difference of inter-class samples. The smaller the value is, the higher the clustering effectiveness is.

3. Select representative power plants

Correlation coefficients between two variables can measure how the change of one variable is correlated with the change

of the other variable, in the same or opposite direction [31]. Because the historical output of one power plant has a certain relationship with the historical output of the subregion, we use three correlation coefficients (Pearson, Spearman, and Kendall) to reflect the correlation between the power plant and the convergence subregion. The representativeness of the PV power plant is evaluated by the correlation coefficient to the subregion.

3.1. Correlation coefficients

The three correlation coefficients reflect the direction and degree of the changing trend between the two variables, and their value range is (−1,1). Zero means the two variables are not correlated, a positive value means positive correlation, and a negative value means negative correlation. A larger absolute value of the correlation coefficient means a stronger correlation.

3.1.1. Pearson correlation coefficient

Pearson correlation coefficient is calculated by a parametric test that requires a continuous normal distribution, and is most commonly used.

$$\rho_{X_{i,j},y_i} = \frac{\operatorname{cov}\left(X_{i,j}, Y_i\right)}{\sigma_{X_{i,j}}\sigma_{Y_i}} = \frac{E\left(X_{i,j}Y_i\right) - E\left(X_{i,j}\right)E\left(Y_i\right)}{\sqrt{E\left(X_{i,j}^2\right) - E^2\left(X_{i,j}\right)}\sqrt{E\left(Y_i^2\right) - E^2\left(Y_i\right)}} \tag{2}$$

where $X_{i,j}$ represents the average output process of PV station j in region i. Y_i represents the average output process for region i. E is the mathematical expectation. $cov(X_{i,j},Y_i)$ represents the covariance of $X_{i,j}$ and Y_i. σ is the standard deviation.

3.1.2. Spearman's rank correlation coefficient

Spearman's rank correlation coefficient has less strict requirements on data conditions than Pearson's. It can be used as long as the observed values of two variables are paired ranked data, or ranked data transformed from continuously observed data, regardless of the overall distribution and sample size.

3.1.3. Kendall correlation coefficient

Kendall correlation coefficient is an extension of Spearman, and it has the same data requirements as Spearman's. It can be used when the same rank is repeated many times in a small dataset [32]. Through the above three correlation coefficients, the correlation between the output of each power station in the subregion and the total output of the region is calculated. By comparing the average values of the three indicators, the most relevant one is selected as the representative power station.

3.2. Correlation analysis of representative power plant

To further verify the rationality of the representative power plant determined by the three correlation coefficients, the following three aspects are evaluated between the representative power plant and whole subregion.

(1) Seasonality

The historical generation output curves of typical days in different seasons are compared.

(2) Typical weather day

According to the existing meteorological information, the generation output curves of typical days (sunny days, cloudy days) in different seasons were found and compared respectively.

(3) Utilization hours of installation capacity

$$H = \frac{\sum_{t=1}^{T} P_t \times h_t}{P_{\mathrm{E}}} \tag{3}$$

where H represents the utilization hours. P_t and h_t are the generation output and time at the t period respectively. P_E is installation capacity. T is the number of periods.

The change process of daily installed utilization hours in representative power stations and subregion in a month is compared.

4. Short-term forecasting model for regional PV power plants

Since many factors affect the PV power plant generation, such as forecasting horizons, forecast model inputs, and performance estimation, making accurate output prediction output has become more complicated [33]. Prediction intervals are categorized into very short-term or ultra-short-term forecasting, short-term forecasting, medium-term forecasting, and long-term forecasting. Longer forecast horizon leads to greater chance of forecast error. CNN (convolutional neural network) or its hybrid form is the most promising method to improve the prediction accuracy. CNN can extract spatial features between representative power plant output and regional total output. LSTM (long short-term memory) neural network can deal with the temporal information that can extract the time characteristics of the output process. Jeff Donahue et al. [34] proposed the Long-term Recurrent Convolutional Network (LRCN) model, which combines CNN and LSTM. The results show that such models have advantages over a single model. The combined CNN and LSTM model can simultaneously extract the spatiotemporal characteristics between input and output [35,36]. By replacing the extrapolation process in the upscaling method with this model, we can more easily extract the complex spatiotemporal relationship between representative power plant and regional total output. Then given the output prediction of representative power plant, the output prediction of the whole region can be obtained using the above model. Only the data of representative power plant are needed, thus reducing the model input and calculation amount. The neural networks and hybrid networks are introduced below, respectively.

4.1. Convolutional neural network

CNN can use local operations to hierarchically abstract representations. It uses two essential techniques. First, CNN uses the 2D structure of the image, and pixels in adjacent areas are usually highly correlated. Therefore, CNN does not need to be fully connected but can use grouped partial connections. Second, the CNN architecture relies on feature sharing, each channel is generated by convolution using the same filter at all positions [37]. So CNN has good local feature extraction capabilities. The basic structure of CNN is shown in Fig. 1.

The input can be convolved through K convolution kernels, and K feature maps can be generated at the C1 layer through activation function. The output value of the unit j of the convolution layer l is calculated as formula (4). Where M_j^l is the set of selected input feature data, k_{ij}^l is the convolution kernel, and b_j^l is the bias.

$$a_j^l = f(b_j^l + \sum_{i \in M_j^l} a_i^{l-1} * k_{ij}^l) \tag{4}$$

After the K graphs of $C1$ layer are subjected to $M^l * M^l$ downsampling processing in the Pooling layer, a new feature map is obtained in the S1 layer. The formula of pooling layer is:

$$a_j^{l+1} = f(b_j^l + \beta_j^l \cdot down(a_j^l, M^l)) \tag{5}$$

where $down(\cdot)$ is a pooling function, commonly Mean-Pooling, Max-Pooling, Min-Pooling, and so on. β_j^l is multiplier residual. M^l is the size of the pooling box.

Generally, layer C is the feature extraction layer. Each neuron input is connected to the local receptive domain of the previous

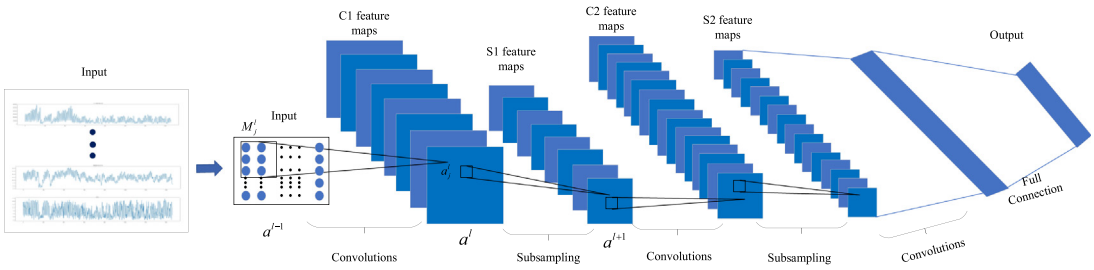

Fig. 1. CNN structure diagram.

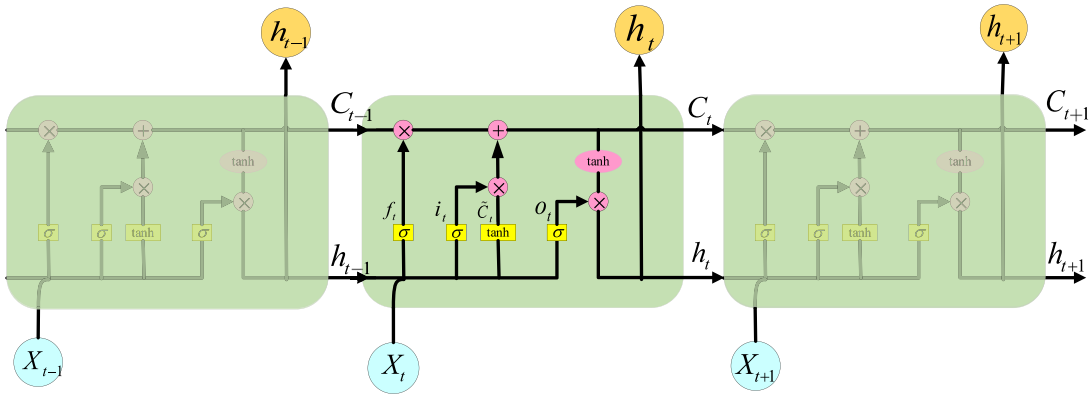

Fig. 2. LSTM structure diagram.

layer. Once the local feature is extracted, the location relationship between it and other features is also determined. The S-layer is a feature mapping layer that achieves some evaluation invariance after pooling, so the architecture is less affected by small changes in location. In addition, the number of network-free parameters and the complexity of network parameters is reduced because the neurons on a mapping surface share weights.

4.2. Long short-term memory neural network

Traditional feedforward neural networks (FNN) only accept information from input nodes. It operates on the input space and does not "remember" the time-specific inputs. In FNN, information only flows from the input layer to the hidden layer and then to the output layer [38]. On the contrary, Recurrent Neural Network (RNN) can carry out self-updating operations internally for time series information, but the problems of gradient disappearance and explosion may lead to learning and bridging long-lag information that fail the operation. LSTM is a unique RNN that truncates gradients and learns to bridge the minimum time lag of more than 1000 discrete time steps by forcing a constant error stream in a particular cell through constant error rotation [39].

Simply put, the X in the figure is the data processed by the CNN network at each moment, and h is the PV generate output result after a layer of LSTM.

The key to LSTM is the memory block, as shown in Fig. 2. It mainly contains three gates (forget gate, input gate, output gate) and a cell.

Forget gate: f_t calculated by h_{t-1} and X_t is used as forget gate to control how much of the memory cell C_{t-1}at the last moment is retained to the current moment C_t.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{6}$$

Input gate: i_t calculated by h_{t-1} and X_t is used as input gate to control how much $\tilde{C}_t$ is added to C_t. $\tilde{C}_t$ calculated from h_{t-1} and X_t by tan h function.

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{7}$$

$$\widetilde{C}_t = \tanh\left(W_c \cdot [h_{t-1}, x_t] + b_c\right) \tag{8}$$

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \widetilde{C}_t \tag{9}$$

Output gate: O_t calculated by h_{t-1} and X_t is used as output gate to control how much C_t outputs after tan h function to the current output value h_t.

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \tag{10}$$

$$h_t = o_t \otimes \tanh\left(C_t\right) \tag{11}$$

$$\sigma\left(x\right) = \frac{1}{1 - e^{-x}} \tag{12}$$

$$tanh\left(x\right) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{13}$$

Later, Alex Graves et al. [40] proposed a Bidirectional Long short-term Memory(BiLSTM). It consists of forward LSTM and backward LSTM. The original LSTM can only get front-to-back coding but not back-to-front information, while BiLSTM can capture the features of bi-directional time series well. The BiLSTM model structure is shown in Fig. 3.

4.3. Forecasting model for regional PV power plants

The prediction model is divided into two parts as a whole. One part is the output prediction of representative power stations in each region, and the other part is the total output prediction of the region. First, We propose a single station prediction model based on CNN-BiLSTM [33,34], which comprehensively considers meteorological factors and historical output data as input and predicts PV output the following day. The interrelated data of the representative power plant in each output area are formed into a two-dimensional matrix. The horizontal axis is full hour from 8:00 to 18:00 at 1-h interval because the PV power plants

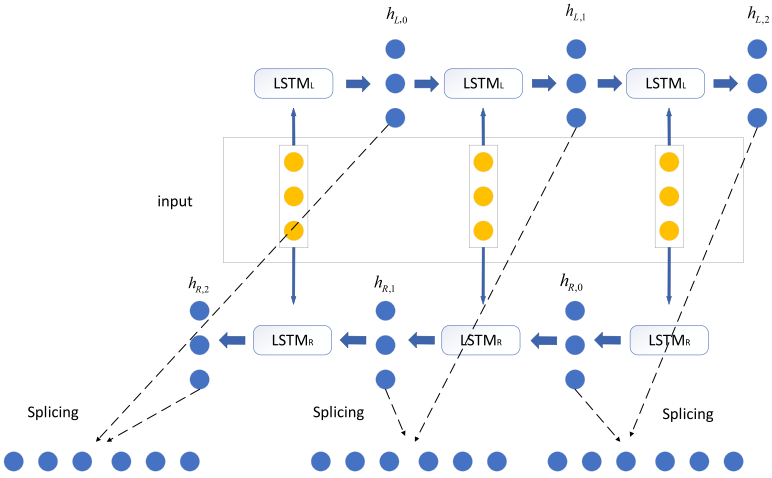

Fig. 3. BiLSTM structure diagram.

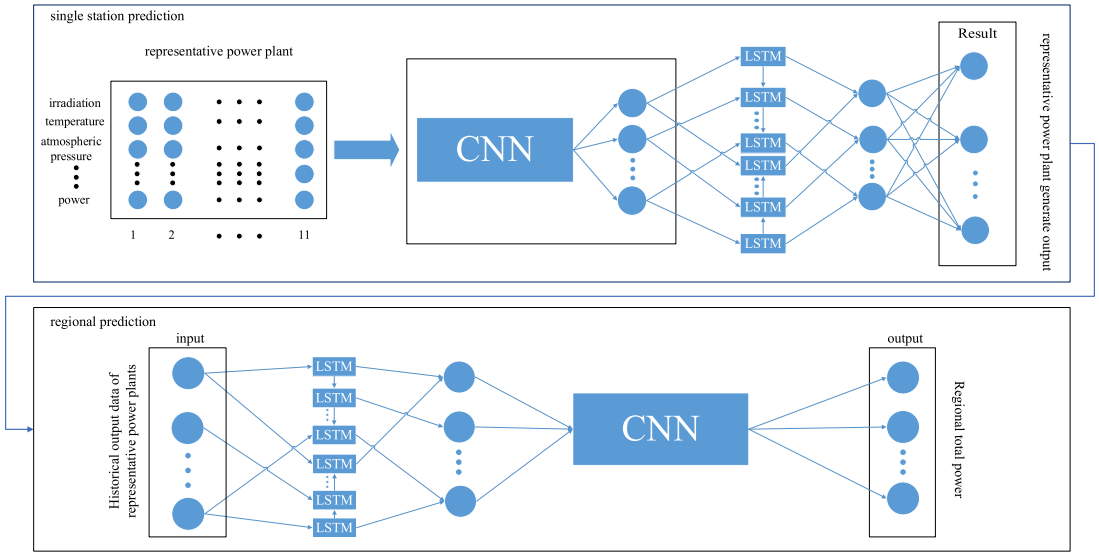

Fig. 4. The structure of forecasting model.

cannot generate electricity in the early morning or at night and the common available meteorological data is hourly. The vertical axis includes historical meteorological data, generation output data, and forecast meteorological data. Since there are many types of data input, it is necessary to extract the characteristics between the data. So the CNN network is put in front, and then the time characteristics are considered through LSTM. The two-dimensional matrix is successively transferred to CNN and BiLSTM, and the prediction results of representing power plants can be obtained through the full connection layer.

Then, we establish regional prediction model through BiLSTM-CNN. The historical output of regional representative power station is taken as input and the corresponding regional total output is taken as output for training. The total output of the region is predicted by the output of representative power stations. Since the input of this model is only the output of representative power stations, we extract the time characteristics of the output through BiLSTM, and then extract the local characteristics of the output curve through CNN. So we re-combine the CNN and BiLSTM to get the BiLSTM-CNN model [41]. The forecasting model is shown in Fig. 4.

4.4. Forecasting process

The overall flow chart of short-term forecasting approach based on BiLSTM-CNN for regional PV power plants is illustrated in Fig. 5. The specific steps are briefly described as follows:

(1) Divide the output subregions. PV power plants in the region are clustered into several output subregions through k-means method by using the historical output data, meteorological data, and other data.

(2) Select a representative power plant in each output subregion. Based on ascending scale method, three correlation coefficients of each power plant were calculated and compared. The power plant with the strongest correlation becomes the representative power plant. Then the correlation of representative power plant is verified according to indicators such as seasonality.

(3) Forecast the subregion output value. As mentioned above, CNN-BiLSTM model can predict representative power plant output. If there is only one power plant in the output subregion, the predicted value is the total subregion output. Otherwise, establish a forecasting model BiLSTM-CNN in each subregion trained by the historical representative power plant output and the total subregion output, to obtain the prediction results of the subregional

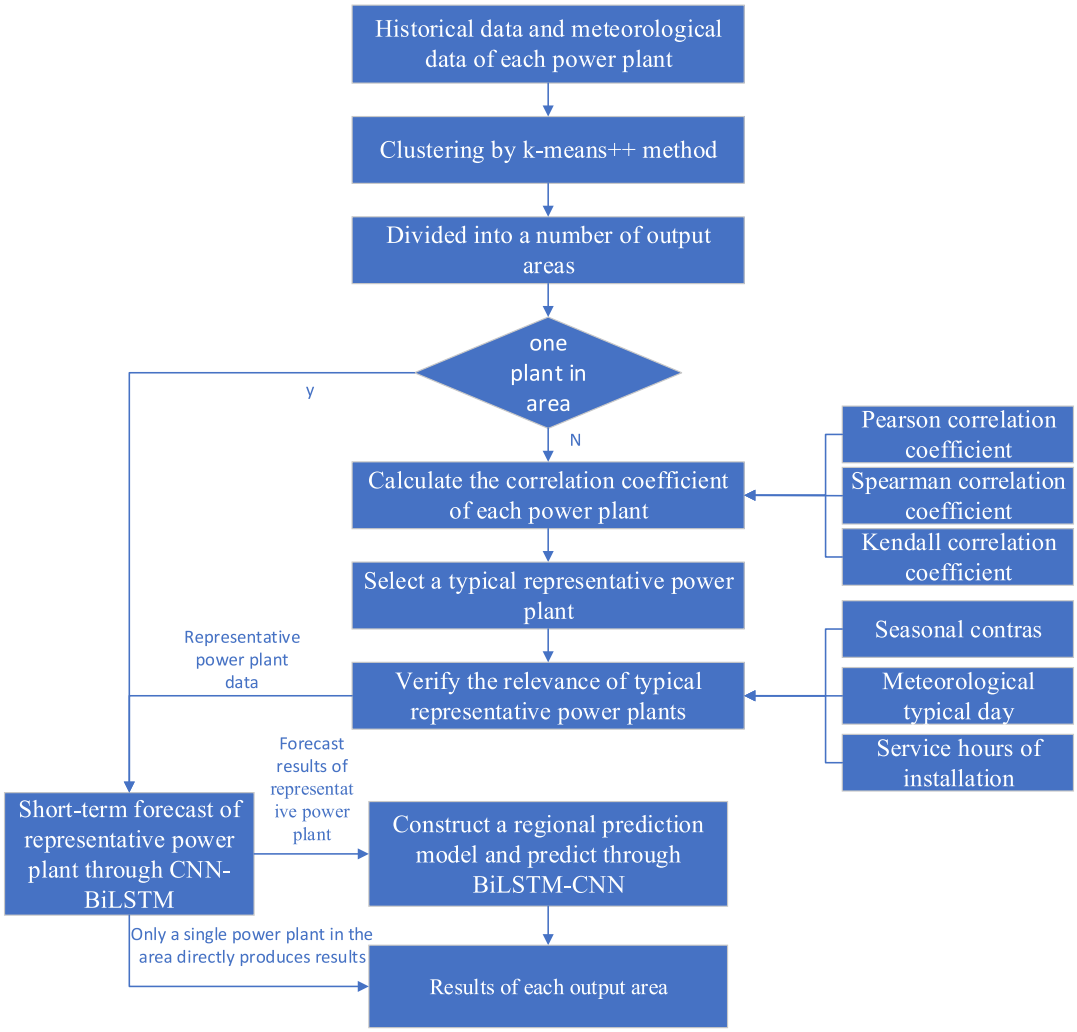

Fig. 5. Overall flow chart.

PV output by using the predicted output of representative power plants as input.

5. Case study

5.1. Engineering background and data description

Yunnan province is in a low latitude plateau with relatively high elevations. Its unique geographical location makes the rich solar radiation energy acceptable throughout the year. It is one of the provinces with the richest solar energy resources in China. The installed PV capacity in Yunnan is estimated to reach more than 18 million kilowatts, equivalent to the installed capacity of a Three Gorges power plant in China. Therefore, providing accurate forecast data of regional PV power plants for dispatching department will be more beneficial. We select two typical Dali and Chuxiong regions in Yunnan province as the research area. The experimental data is historical operation data, PV power generation data and meteorological data from January 1, 2017 to December 31, 2019. The meteorological data comes from ECMWF Reanalysis v5 (ERA5), mainly including temperature, dew point temperature, air pressure and wind speed.

5.2. Divide the PV output area

Through the k-means clustering algorithm, 9 PV power plants in Chuxiong region are iteratively clustered based on the output and meteorological data, and are divided into different output subregions. Fig. 6 shows relationship between the number of clusters and three indexes, including SSE, silhouette coefficients and Davies–Bouldin index.

It can be seen that the optimal K value of each index is different. The inflection point of SSE appears when $K = 3$. The silhouette coefficient reaches its maximum when $K = 2$, followed by $K = 3$. And the Davies–Bouldin reaches its optimum at $K = 6$, followed by $K = 8$ and 3. For $K = 2$, although the silhouette coefficient is the highest, the SSE is also the largest, and Davies–Bouldin index is also at a higher value, so $K = 2$ is excluded. For $K = 6$ and $K = 8$, their silhouette coefficients are low, and SSE has no obvious inflection point, so they are also excluded. After excluding them, the result of selected division is three output subregions. The classification results are as followed:

Output subregion 1: Xiaoxicun plant, Daguya plant, Dazhuang plant, Tianzishan plant, Changchong plant;

Output subregion 2: Xiutian plant, Hewai plant, Ganbala plant;

Output subregion 3: Banxing plant;

5.3. Select and verify representative power plants

The three correlation coefficients introduced in Section 3.1 are used to select representative power plant in each output

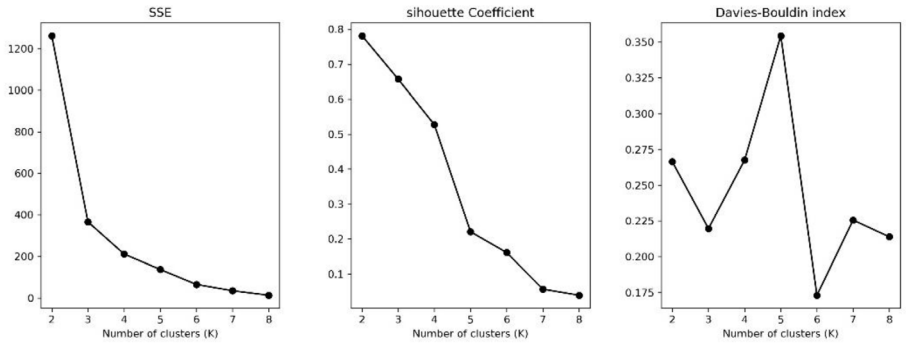

Fig. 6. Relationship between the number of clusters and three indexes.

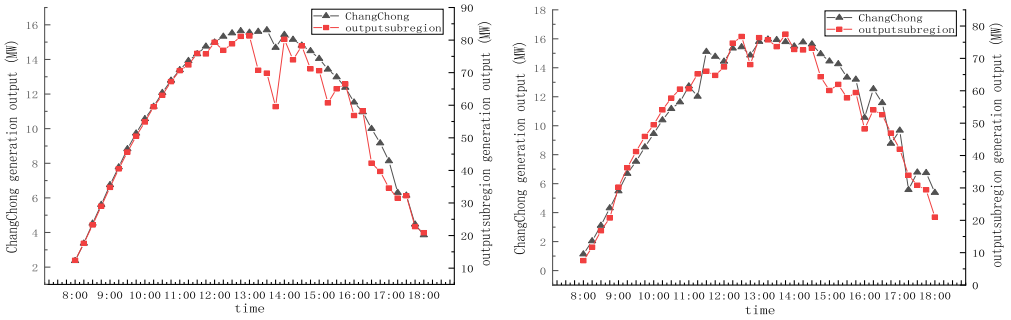

Fig. 7. Seasonal comparison between ChangChong power plant and output subregion 1.

Table 1
Correlation coefficient of each PV power plant in output subregion 1.

Power plant	Pearson	Spearman	Kendall	Weighted average
Xiaoxicun	0.9978	0.9893	0.9419	0.9763
Daguya	0.9933	0.9812	0.9117	0.9620
Dazhuang	0.9958	0.9891	0.9396	0.9748
Tianzishan	0.9957	0.9888	0.9372	0.9739
Changchong	0.9971	0.9922	0.9488	0.9794

Table 2
Correlation coefficients of three verification aspects.

Power plant	Seasonality	Typical weather day	Utilization hours
Daguya	0.7852	0.6809	0.8653
Dazhuang	0.8720	0.5294	0.7533
Tianzishan	0.8785	0.7488	0.7668
Xiaoxicun	0.8292	0.6509	0.8692
Changchong	0.9114	0.7906	0.8768

subregion, and then three aspects are analyzed to verify the representativeness of the selected power plants. According to [42], the PV power plant with the largest weighted average of the three correlation coefficients is the representative power plant in the output subregion. Table 1 shows the three correlation coefficients and weighted average coefficient of each PV power plant in the output subregion 1, where each correlation coefficient has the same weight.

From Table 1, Changchong power plant has slightly lower Pearson correlation coefficient than Xiaoxicun power plant, but it has larger Spearman and Kendall correlation coefficients, and the weighted average coefficient is also the largest. So Changchong power plant is considered having the strongest representation, and is selected as the representative power plant in output subregion 1. The representativeness of Changchong power plant is further verified from three aspects: seasonality, typical weather day and utilization hours of installed capacity. The comparison between the representative PV power plant and output subregion 1 are shown in Figs. 7–10. From any aspect, the generation output process of the Changchong power plant is similar to the whole subregion.

In addition, the correlation coefficients of these three verification aspects of each power plant are calculated as shown in Table 2. The indicators of Changchong power plant are superior to other power plants, so it is rational to select it as the representative power plant in output subregion 1.

Table 3
Model training parameters.

Iterations	Learning rate	Regularization coefficient	Attenuation strategy	Step size	Attenuation rate
150	0.01	1.00E−06	Decay gradually	15	0.1

We perform the similar selection and verification process for representative power plant in output subregion 2, where Xiutian power plant is chosen as the representative power plant. Banxing plant is the single and representative power plant in output subregion 3.

5.4. Forecasting results

We select the generation output data from 8:00 to 18:00 at one hour interval. The output process of 11 sequence points is used to train the model. Model training parameters are shown in Table 3.

The regularization coefficient is to reduce the over-fitting phenomenon in the training process. Learning rate represents the magnitude of parameter update in each training iteration. If the learning rate is too large, it will lead to non-convergence. Otherwise, it will lead to slow convergence of parameters with optimization. So we set an attenuation strategy for learning rate. Step

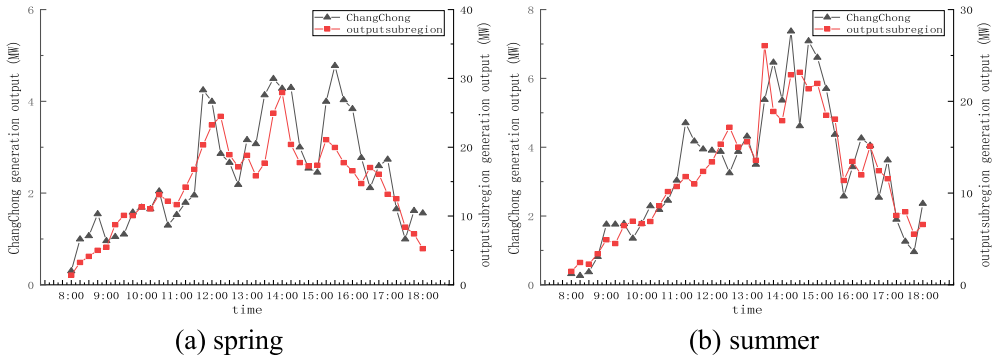

Fig. 8. Rainy day comparison between ChangChong power plant and output subregion 1.

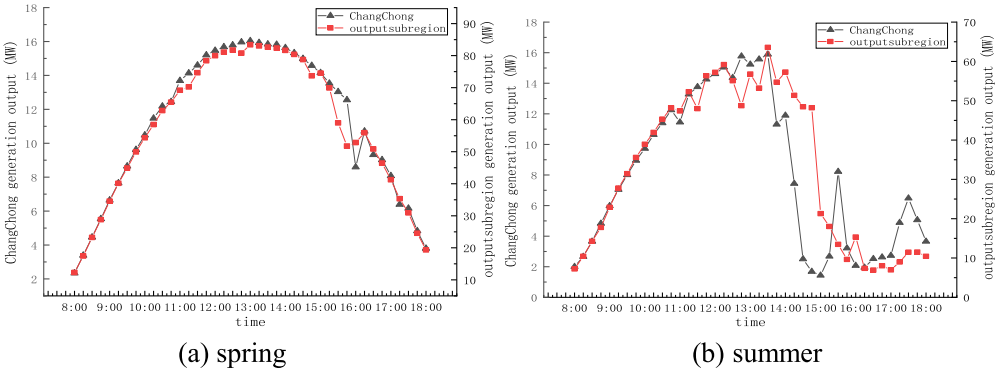

Fig. 9. Sunny day comparison between ChangChong power plant and output subregion 1.

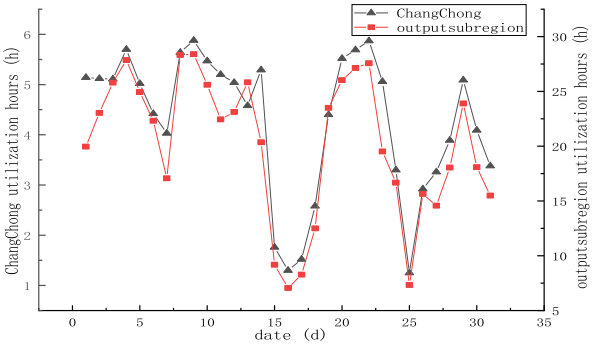

Fig. 10. Comparison of installed utilization hours between ChangChong power plant and output subregion 1.

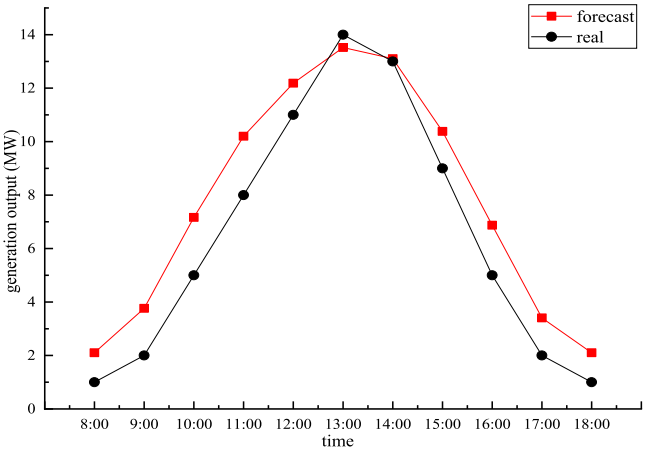

Fig. 11. Forecast and real output of the Changchong plant on January 6.

size represents the decay of learning rate per 15 iterations. The attenuation rate represents 0.1 per decay to the learning rate.

For the representative power plant in output subregion 1, we train the CNN-BiLSTM model by using the output and weather data from January 1, 2017 to December 31, 2018, and then predict the plant output on January 6, 2019. The result is shown in Fig. 11.

We then train the relationship between the generation output of Changchong plant and the output subregion 1 through the BiLSTM-CNN model. With the predicted out of Changchong plant on January 6, 2019 as input, the trained model predicts the generation output for output subregion 1. The results are in Fig. 12.

It can be seen from Table 4, Figs. 11 and 12 that improving the regional prediction network can further reduce the subregional output prediction error. The regional prediction network learns the relationship between the representative power plant and the total subregional output. It predicts the total output of the subregion close to the predicted output of the representative power plant based on the learnt relationship. Since the forecast result of the Changchong power plant did not predict the highest peak at 13:00, and the forecast output in the first half was greater than the actual value, the subregional forecast result did not reach the highest actual subregional output and the forecast output at the first half also exceeded the actual value. Improving the

Table 4
Prediction accuracy of representative power plant and output subregion.

	MSE	RMSE	MAE
Changchong	0.1453	0.1149	0.3194
Output subregion 1	0.0184	0.0409	0.1019

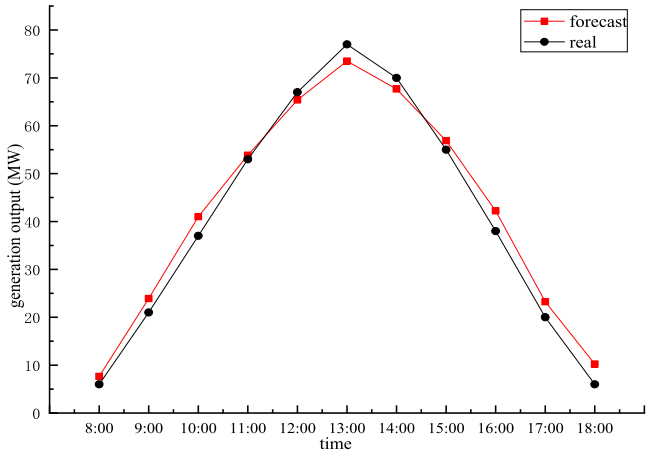

Fig. 12. Forecast and real generation output of output subregion 1.

accuracy of single plant prediction can improve the accuracy of subregional prediction.

To further verify the feasibility of proposed model, we forecast the power generation of all output subregions of Chuxiong and Dali in each of the four seasons of 2019. The same data was used to compare with the superposition method and the CNN method. The superposition method adopts the same CNN-BiLSTM as forecasting model of the representative power plant we proposed. We also compare it with the SDA-GA-ELM model. Some of the predicted results for Chuxiong are shown in Figs. 13–16. And the calculation results of MSE, RMSE, and MAE of the three methods are shown in Tables 5 and 6.

It can be seen from Table 5 that the SDA-GA-ELM method in the output subregion1 has the worst performance among the three methods. The prediction accuracy of the proposed method is the best. The superposition method predicts the five power stations respectively, and then superimposes the results, resulting in the accumulation of errors. Since the superposition model is used the same CNN-BiLSTM model for single-plant prediction as the we proposed method, the overall error is slightly larger than the proposed method, but smaller than the CNN method. Due to a large amount of input data of five power stations, CNN does not learn the characteristics of the total subregional output well, which is most unable to reach the peak. In output subregion 2, the prediction effect of superposition method is improved compared with output subregion 1. The main reason is that there are few power plants in this output area, and the error accumulation is not large. Meanwhile, the error between it and our proposed method is further reduced. There is only one power plant in the output subregion 3, so the superposition method behaves the same as the method mentioned. CNN still does not predict the peak value, and the prediction effect is poor. Yunnan's climate basically belongs to the subtropical plateau monsoon type, with high temperatures and rainy summer, change is difficult to predict. By comparing the results for different seasons in each region, it can be seen that summer and autumn results are worse than other seasons. Mainly because Yunnan's climate basically belongs to the subtropical plateau monsoon type, with high temperatures and rainy summer, change is difficult to predict. Meanwhile, we compared the model in Dali, and the results are shown in Table 6. The results appear similar to Table 5, and the proposed method has the best results, further validating the validity of the model.

After aggregating the results for the two regions, the prediction effect of the proposed method is the best, followed by the superposition method and the CNN, the SDA-GA-ELM effect is the poorest. Due to the combination of genetic algorithm, SDG-GA-ELM not only improves the accuracy but also has the disadvantages of GA. The local optimal solution may appear, resulting in unstable results and long calculation time. The superposition method predicts each power plant in the subregion with CNN-BiLSTM model and adds up the results. This leads to the accumulation of errors, resulting in poor results. CNN method needs input data at large dimension and quantity and needs more data to train the network, and data preparation is time and energy-consuming. Besides, most CNN results do not predict the peak well. In contrast, the proposed method only predicts representative power plants using the same CNN-BiLSTM model and then predicts subregional output by learning the relationship between representative power plants and the total output subregion. The proposed method only requires the selected representative power plant have high-quality historical data, and the output of the whole subregion. Therefore, input data does not

Table 5
Comparison of forecasting methods in Chuxiong.

Subregion/region	Season	Method	MSE	RMSE	MAE
Output subregion 1	Spring	CNN	0.076	0.276	0.223
		Superposition	0.044	0.210	0.162
		SDA-GA-ELM	0.134	0.367	0.270
		Proposed method	0.037	0.191	0.144
	Summer	CNN	0.123	0.351	0.276
		Superposition	0.050	0.224	0.174
		SDA-GA-ELM	0.139	0.373	0.270
		Proposed method	0.044	0.209	0.156
	Autumn	CNN	0.074	0.272	0.207
		Superposition	0.070	0.264	0.199
		SDA-GA-ELM	0.144	0.379	0.282
		Proposed method	0.057	0.238	0.186
	Winter	CNN	0.090	0.300	0.224
		Superposition	0.062	0.249	0.179
		SDA-GA-ELM	0.092	0.304	0.219
		Proposed method	0.065	0.255	0.165
Output subregion 2	Spring	CNN	0.166	0.408	0.327
		Superposition	0.069	0.263	0.211
		SDA-GA-ELM	0.369	0.608	0.254
		Proposed method	0.057	0.240	0.182
	Summer	CNN	0.195	0.441	0.356
		Superposition	0.161	0.401	0.330
		SDA-GA-ELM	0.292	0.540	0.279
		Proposed method	0.113	0.336	0.259
	Autumn	CNN	0.481	0.694	0.549
		Superposition	0.337	0.581	0.424
		SDA-GA-ELM	0.524	0.724	0.371
		Proposed method	0.299	0.547	0.405
	Winter	CNN	0.208	0.456	0.304
		Superposition	0.158	0.398	0.274
		SDA-GA-ELM	0.377	0.614	0.426
		Proposed method	0.143	0.378	0.252
Output subregion 3	Spring	CNN	0.330	0.574	0.448
		Superposition	0.209	0.457	0.291
		SDA-GA-ELM	0.452	0.672	0.503
		Proposed method	0.209	0.457	0.291
	Summer	CNN	0.397	0.630	0.484
		Superposition	0.192	0.439	0.317
		SDA-GA-ELM	0.585	0.765	0.568
		Proposed method	0.209	0.457	0.291
	Autumn	CNN	0.578	0.761	0.632
		Superposition	0.340	0.583	0.423
		SDA-GA-ELM	0.599	0.774	0.740
		Proposed method	0.340	0.583	0.423
	Winter	CNN	0.329	0.574	0.435
		Superposition	0.184	0.429	0.308
		SDA-GA-ELM	0.344	0.586	0.494
		Proposed method	0.184	0.429	0.308

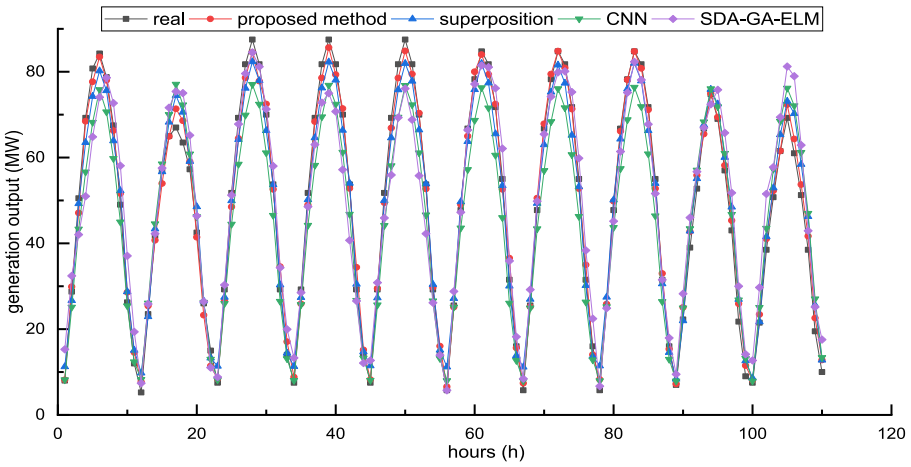

Fig. 13. The comparison curve of predicted and actual generation output in spring.

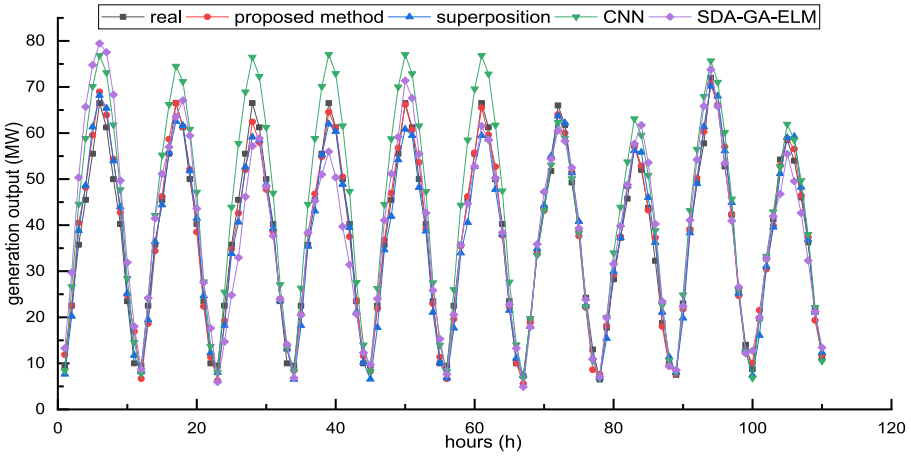

Fig. 14. The comparison curve of predicted and actual generation output in summer.

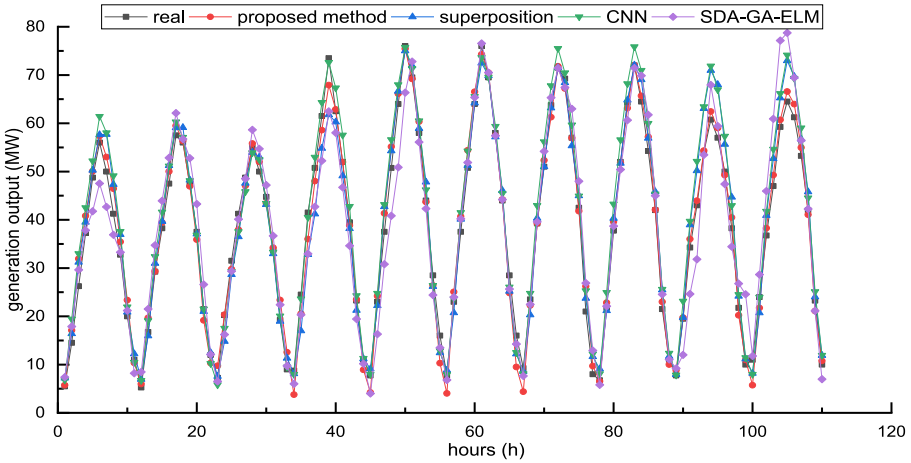

Fig. 15. The comparison curve of predicted and actual generation output in autumn.

necessarily have large dimensions or quantities. It is suitable for regions or subregions lacking some power plants' historical data or having new power plants. In addition, the proposed method can also improve the prediction accuracy of the entire output subregion by improving the prediction accuracy of the representative power plants.

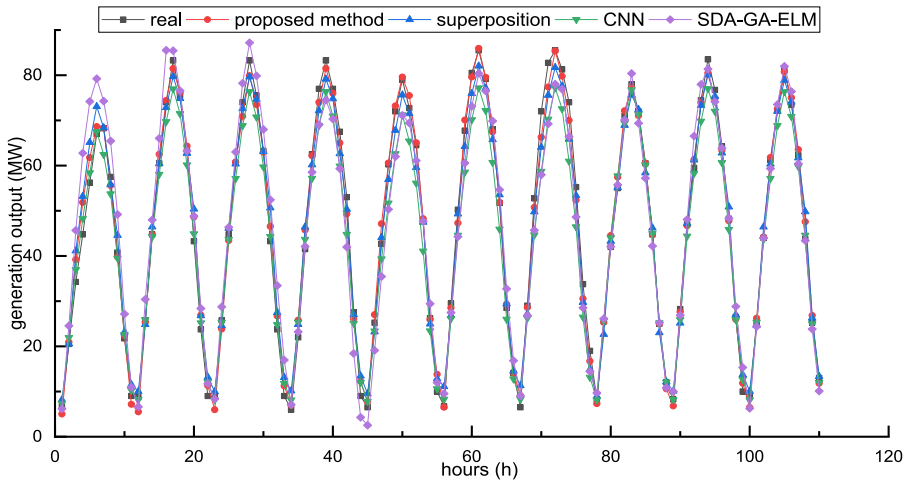

Fig. 16. The comparison curve of predicted and actual generation output in winter.

Table 6
Comparison of forecasting methods in Dali.

Subregion/region	Season	Method	MSE	RMSE	MAE
Output subregion 1	Spring	CNN	0.213	0.461	0.390
		Superposition	0.181	0.425	0.303
		SDA-GA-ELM	0.264	0.514	0.429
		Proposed method	0.061	0.246	0.194
	Summer	CNN	0.229	0.478	0.369
		Superposition	0.409	0.639	0.478
		SDA-GA-ELM	0.291	0.540	0.427
		Proposed method	0.155	0.394	0.286
	Autumn	CNN	0.367	0.606	0.468
		Superposition	0.420	0.648	0.406
		SDA-GA-ELM	0.409	0.639	0.419
		Proposed method	0.328	0.572	0.366
	Winter	CNN	0.458	0.677	0.461
		Superposition	0.408	0.639	0.355
		SDA-GA-ELM	0.440	0.663	0.424
		Proposed method	0.292	0.540	0.323
Output subregion 2	Spring	CNN	0.112	0.335	0.270
		Superposition	0.087	0.295	0.243
		SDA-GA-ELM	0.148	0.384	0.270
		Proposed method	0.057	0.239	0.186
	Summer	CNN	0.191	0.437	0.341
		Superposition	0.194	0.440	0.354
		SDA-GA-ELM	0.201	0.448	0.379
		Proposed method	0.151	0.389	0.311
	Autumn	CNN	0.366	0.605	0.509
		Superposition	0.282	0.531	0.434
		SDA-GA-ELM	0.259	0.509	0.403
		Proposed method	0.228	0.478	0.376
	Winter	CNN	0.249	0.499	0.260
		Superposition	0.203	0.451	0.286
		SDA-GA-ELM	0.272	0.522	0.297
		Proposed method	0.234	0.484	0.246

6. Conclusion

Accurate forecast of PV generation output can help dispatching departments make more reasonable plans and improve the utilization efficiency of renewable energy. With the large-scale grid connection of PV power plants, regional prediction has attracted more and more attention. we proposed a short-term forecasting approach based on BiLSTM-CNN for regional PV power plants, aiming to improve the regional prediction accuracy and solve the prediction difficulties caused by incomplete historical data of some power plants. PV output subregions are established by using k-means method, where PV power plants with similar power generation characteristics are grouped into the same output subregion. A representative power plant in each output subregion is selected based on three correlation coefficients, and the rationality of the selected power plant is verified from three aspects. On this basis, the BiLSTM-CNN method is used to study the correlation of generation output between the representative power plant and subregional PV power plants. Given the output of representative power plant predicted by CNN-BiLSTM model, we can predict the subregional output based on such correlation.

Finally, we take the PV power plants in Chuxiong and Dali region, Yunnan province as an example. From the numerical results, the proposed method can obtain higher prediction accuracy compared with the superposition method, the CNN method and the SDG-GA-ELM method. The method can find the most representative PV power plants in the region through the analysis of various indicators. And it requires less data, only the complete data of representative power stations in the region can be used to predict the output of the whole region. Therefore, this method can also be applied to areas where the historical data of some power plants are insufficient, or new power plants continue to join.

CRediT authorship contribution statement

Gang Li: Investigation, Conceptualization, Writing – review & editing, Supervision. **Shunda Guo:** Methodology, Software, Data curation, Writing – original draft. **Xiufeng Li:** Resources, Validation. **Chuntian Cheng:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

This study is supported by the National Natural Science Foundation of China (No. 51879030). The authors are very grateful to the anonymous reviewers and editors for their constructive comments.

References

[1] N. Dong, J.F. Chang, A.G. Wu, et al., A novel convolutional neural network framework based solar irradiance prediction method, Int. J. Electr. Power Energy Syst. 114 (2020) 105411, http://dx.doi.org/10.1016/j.ijepes.2019.105411.

[2] Z. Ya-Xin, L.U.O. Hui-Lin, W. Can, Progress and trends of global carbon neutrality pledges, Adv. Clim. Chang. Res. 17 (1) (2021) 88, http://dx.doi.org/10.12006/j.issn.1673-1719.2020.241.

[3] X. Bai, L. Liu, D. Wei, et al., Research on security threat and evaluation model of new energy plant and station, in: 2020 International Conference on Computer Communication and Network Security, CCNS, IEEE, Xi'an, China, 2020, pp. 75–80, [2022-08-15]. https://ieeexplore.ieee.org/document/9240740/.

[4] X. Zhao, X. Kong, W. Lu, Two-stage robust stochastic optimal dispatch of regional integrated energy system considering renewable energy and load uncertainty, in: 2021 24th International Conference on Electrical Machines and Systems, ICEMS, IEEE, Gyeongju, Korea, Republic of, 2021, pp. 2240–2245, [2022-08-15]. https://ieeexplore.ieee.org/document/9634644/.

[5] K. Kim, J. Hur, Weighting factor selection of the ensemble model for improving forecast accuracy of photovoltaic generating resources, Energies 12 (17) (2019) 3315, http://dx.doi.org/10.3390/en12173315.

[6] M.P. Almeida, M. Muñoz, I. De La Parra, et al., Comparative study of PV power forecast using parametric and nonparametric PV models, Sol. Energy 155 (2017) 854–866, http://dx.doi.org/10.1016/j.solener.2017.07.032.

[7] K. Bakker, K. Whan, W. Knap, et al., Comparison of statistical post-processing methods for probabilistic NWP forecasts of solar radiation, Sol. Energy 191 (2019) 138–150, http://dx.doi.org/10.1016/j.solener.2019.08.044.

[8] P. Aillaud, J. Lequeux, J. Mathe, et al., Day-ahead forecasting of regional photovoltaic production using deep learning, in: 2020 47th IEEE Photovoltaic Specialists Conference, PVSC, IEEE, Calgary, AB, Canada, 2020, pp. 2688–2691, [2022-04-16]. https://ieeexplore.ieee.org/document/9300538/.

[9] R.A. Verzijlbergh, P.W. Heijnen, S.R. De Roode, et al., Improved model output statistics of numerical weather prediction based irradiance forecasts for solar power applications, Sol. Energy 118 (2015) 634–645, http://dx.doi.org/10.1016/j.solener.2015.06.005.

[10] H. Böök, A.V. Lindfors, Site-specific adjustment of a NWP-based photovoltaic production forecast, Sol. Energy 211 (2020) 779–788, http://dx.doi.org/10.1016/j.solener.2020.10.024.

[11] F.O. Hocaoglu, F. Serttas, A novel hybrid (Mycielski-Markov) model for hourly solar radiation forecasting, Renew. Energy 108 (2017) 635–643, http://dx.doi.org/10.1016/j.renene.2016.08.058.

[12] D.S. Tripathy, B.R. Prusty, D. Jena, Short-term PV generation forecasting using quantile regression averaging, in: 2020 IEEE International Conference on Power Systems Technology, POWERCON, 2020, pp. 1–6, http://dx.doi.org/10.1109/POWERCON48463.2020.9230535.

[13] H. Panamtash, Q. Zhou, T. Hong, et al., A copula-based Bayesian method for probabilistic solar power forecasting, Sol. Energy 196 (2020) 336–345, http://dx.doi.org/10.1016/j.solener.2019.11.079.

[14] D.S. Tripathy, B.R. Prusty, K. Bingi, A k-nearest neighbor-based averaging model for probabilistic PV generation forecasting, Int. J. Numer. Modelling, Electron. Netw. Devices Fields 35 (2) (2022) e2983, http://dx.doi.org/10.1002/jnm.2983.

[15] J. Shi, W.J. Lee, Y. Liu, et al., Forecasting power output of photovoltaic systems based on weather classification and support vector machines, IEEE Trans. Ind. Appl. 48 (3) (2012) 1064–1069, http://dx.doi.org/10.1109/TIA.2012.2190816.

[16] M.N. Akhter, S. Mekhilef, H. Mokhlis, et al., Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques, IET Renew. Power Gener. 13 (7) (2019) 1009–1023, http://dx.doi.org/10.1049/iet-rpg.2018.5649.

[17] D. Lee, K. Kim, Recurrent neural network-based hourly prediction of photovoltaic power output using meteorological information, Energies 12 (2) (2019) 215, http://dx.doi.org/10.3390/en12020215.

[18] P. Gupta, R. Singh, Pv power forecasting based on data-driven models: a review, Int. J. Sustain. Eng. 14 (6) (2021) 1733–1755, http://dx.doi.org/10.1080/19397038.2021.1986590.

[19] Yang, Fu, Yao, et al., A regional photovoltaic output prediction method based on hierarchical clustering and the mRMR criterion, Energies 12 (20) (2019) 3817, http://dx.doi.org/10.3390/en12203817.

[20] X. Zhang, Y. Yang, H. Wang, et al., A convolutional neural network for regional photovoltaic generation point forecast, in: E3S Web of Conferences, Vol. 185, 2020, p. 01079, http://dx.doi.org/10.1051/e3sconf/202018501079.

[21] Y. Yu, M. Wang, F. Yan, et al., Improved convolutional neural network-based quantile regression for regional photovoltaic generation probabilistic forecast, IET Renew. Power Gener. 14 (14) (2020) 2712–2719, http://dx.doi.org/10.1049/iet-rpg.2019.0949.

[22] J.G. da S. Fonseca Jr., T. Oozeki, H. Ohtake, et al., Regional forecasts of photovoltaic power generation according to different data availability scenarios: a study of four methods: Regional forecasts of photovoltaic power generation, Prog. Photovolt., Res. Appl. 23 (10) (2015) 1203–1218, http://dx.doi.org/10.1002/pip.2528.

[23] B.D. Dimd, S. Völler, U. Cali, et al., A review of machine learning-based photovoltaic output power forecasting: Nordic context, IEEE Access 10 (2022) 26404–26425, http://dx.doi.org/10.1109/ACCESS.2022.3156942.

[24] Y.M. Saint-Drenan, G.H. Good, M. Braun, et al., Analysis of the uncertainty in the estimates of regional PV power generation evaluated with the upscaling method, Sol. Energy 135 (2016) 536–550, http://dx.doi.org/10.1016/j.solener.2016.05.052.

[25] Y.M. Saint-Drenan, G.H. Good, M. Braun, A probabilistic approach to the estimation of regional photovoltaic power production, Sol. Energy 147 (2017) 257–276, http://dx.doi.org/10.1016/j.solener.2017.03.007.

[26] M. Pierro, M. De Felice, E. Maggioni, et al., Data-driven upscaling methods for regional photovoltaic power estimation and forecast using satellite and numerical weather prediction data, Sol. Energy 158 (2017) 1026–1038, http://dx.doi.org/10.1016/j.solener.2017.09.068.

[27] H. Shaker, D. Manfre, H. Zareipour, Forecasting the aggregated output of a large fleet of small behind-the-meter solar photovoltaic sites, Renew. Energy 147 (2020) 1861–1869, http://dx.doi.org/10.1016/j.renene.2019.09.102.

[28] Y. Zhou, N. Zhou, L. Gong, et al., Prediction of photovoltaic power output based on similar day analysis, genetic algorithm and extreme learning machine, Energy 204 (2020) 117894, http://dx.doi.org/10.1016/j.energy.2020.117894.

[29] D. Arthur, S. Vassilvitskii, k-means++: The Advantages of Careful Seeding. 11.

[30] F. Martinez Alvarez, A. Troncoso, J.C. Riquelme, et al., Energy time series forecasting based on pattern sequence similarity, IEEE Trans. Knowl. Data Eng. 23 (8) (2011) 1230–1243, http://dx.doi.org/10.1109/TKDE.2010.227.

[31] P. Schober, C. Boer, L.A. Schwarte, Correlation coefficients: Appropriate use and interpretation, Anesth. Analg. 126 (5) (2018) 1763–1768, http://dx.doi.org/10.1213/ANE.0000000000002864.

[32] H. Akoglu, User's guide to correlation coefficients, Turk. J. Emerg. Med. 18 (3) (2018) 91–93, http://dx.doi.org/10.1016/j.tjem.2018.08.001.

[33] R. Ahmed, V. Sreeram, Y. Mishra, et al., A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization, Renew. Sustain. Energy Rev. 124 (2020) 109792, http://dx.doi.org/10.1016/j.rser.2020.109792.

[34] J. Donahue, L.A. Hendricks, M. Rohrbach, et al., Long-term recurrent convolutional networks for visual recognition and description, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 677–691, http://dx.doi.org/10.1109/TPAMI.2016.2599174.

[35] Y. Chen, S. Zhang, W. Zhang, et al., Multifactor spatio-temporal correlation model based on a combination of convolutional neural network and long short-term memory neural network for wind speed forecasting, Energy Convers. Manage. 185 (2019) 783–799, http://dx.doi.org/10.1016/j.enconman.2019.02.018.

[36] S. Zhang, Y. Chen, J. Xiao, et al., Hybrid wind speed forecasting model based on multivariate data secondary decomposition approach and deep learning algorithm with attention mechanism, Renew. Energy 174 (2021) 688–704, http://dx.doi.org/10.1016/j.renene.2021.04.091.

[37] I. Hadji, R.P. Wildes, What do we understand about convolutional networks? 2018, arXiv:1803.08834 [cs], [2022-04-16]. http://arxiv.org/abs/1803.08834.

[38] K. Wang, X. Qi, H. Liu, A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network, Appl. Energy 251 (2019) 113315, http://dx.doi.org/10.1016/j.apenergy.2019.113315.

[39] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[40] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Netw. 18 (5–6) (2005) 602–610, http://dx.doi.org/10.1016/j.neunet.2005.06.042.

[41] K. Wang, X. Qi, H. Liu, Photovoltaic power forecasting based LSTM-convolutional network, Energy 189 (2019) 116225, http://dx.doi.org/10.1016/j.energy.2019.116225.

[42] K. Zheng, Y. Niu, Research on renewable power basement output characteristics, Taiyangneng Xuebao/Acta Energiae Sol. Sin. 39 (2018) 2591–2598.