processes

MDPI

Article

LightGBM-Integrated PV Power Prediction Based on Multi-Resolution Similarity

Yan Peng [1,*], Shichen Wang [2], Wenjin Chen [3], Junchao Ma [1], Chenxu Wang [1] and Jingwei Chen [3]

[1] State Grid Zhejiang Electric Power Research Institute, State Grid Zhejiang Electric Power Supply Co., Ltd., Hangzhou 310014, China
[2] School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China
[3] State Grid Zhejiang Electric Power Supply Co., Ltd., Hangzhou 310014, China
* Correspondence: misile8@gmail.com

Abstract: Improving the accuracy of PV power prediction is conducive to PV participation in economic dispatch and power market transactions in the distribution network, as well as safe dispatch and operation of the grid. Considering that the selection of highly correlated historical data can improve the accuracy of PV power prediction, this study proposes an integrated PV power prediction method based on a multi-resolution similarity consideration that considers both trend similarity and detail similarity. Firstly, using irradiance as the similarity variable, similar-days were selected using grey correlation analysis to form a set of similar data to control the similarity, with the overall trend of the day to be predicted at a macro level. Using irradiance to calculate the similarity at each specific point in time via Euclidean distance, similar-times were identified to form another set of similar data to consider the degree of similarity in detail. The above approach enables the selection of similarity data for both resolutions. Then, a 1DCNN-LSTM prediction model that considers the feature correlation of different variables and the temporal dependence of a single variable was proposed. Three important features were selected by a random forest model as inputs to the prediction model, and two similar data training models with different resolutions were used to generate a photovoltaic power prediction model based on similar-days and similar-times. Ultimately, the learning of the two predictions integrated with LightGBM compensate for each other, generating highly accurate predictions that combine the advantages of multi-resolution similarity considerations. Actual operation data of a PV power station was used for verification. The simulation results show that the prediction effect of ensemble learning was better than that of the single 1DCNN-LSTM model. The proposed method was compared with other commonly used PV power prediction models. In the data case of this study, it was found that the proposed method reduced the prediction error rate by 1.48%, 11.4%, and 6.45%, compared to the LSTM, CNN, and BP, respectively. Experiments show that model prediction results considering the selection of similar data at multiple resolutions can provide more extensive information to an ensemble learner and reduce the deviation in model predictions. Therefore, the proposed method can provide a reference for PV integration into the grid and participation in market-based electricity trading, which is of great significance.

Keywords: photovoltaic power forecasting; similar-day; similar-time; 1DCNN-LSTM; LightGBM

Citation: Peng, Y.; Wang, S.; Chen, W.; Ma, J.; Wang, C.; Chen, J. LightGBM-Integrated PV Power Prediction Based on Multi-Resolution Similarity. *Processes* **2023**, *11*, 1141. https://doi.org/10.3390/pr11041141

Academic Editor: Jesús Polo

Received: 7 March 2023
Revised: 29 March 2023
Accepted: 4 April 2023
Published: 7 April 2023

1. Introduction

Solar energy is a clean and efficient renewable energy source [1], and photovoltaic power generation is increasingly being promoted by various countries and regions to cope with energy shortages [2]. After years of development, China's photovoltaic industry has also successfully reached a leading level in international competitiveness, and it can be considered an important driving force for energy change in China. Photovoltaic power generation is susceptible to a variety of factors, such as installation methods, panel materials, and climate, and is therefore highly volatile, intermittent, and uncertain [3,4]. As more and

more photovoltaic power is connected to the grid, this uncertainty can pose a great threat to grid connection [5], so it is important to anticipate photovoltaic power in advance to prepare for grid connection.

Many methods have been proposed by researchers, and many artificial intelligence algorithms have been applied to photovoltaic power prediction, aiming at improving prediction accuracy. Among them, convolutional neural networks (CNNs) [6], back propagation (BP) neural networks [7], and long- and short-term memory (LSTM) neural networks [8] are frequently used methods for PV power prediction. Highly relevant historical data can be obtained by selecting similar data before selecting a model for prediction. The literature [9] uses Euclidean distance to select similar-days of data for the target day to be predicted and then decomposes the original photovoltaic power series using the empirical decomposition method (EMD) and then combines it with a least squares support vector machine (LSSVM) to predict photovoltaic power. However, this method only considers the data of similar-days as training data. The literature [10] uses Pearson correlation coefficients to select numerical weather prediction (NWP) data that is similar to the moment being predicted to obtain similar moments, and then estimates the power at the moment to be predicted based on the actual power at the similar moment. This method only calculates the similar moments as similarity degrees of the data. The literature [11] has used the entropy weighting method to calculate the weight of the influence of each meteorological factor on photovoltaic power and selected similar-days by calculating the meteorological factors of the historical day and the day to be predicted by weighted Euclidean distance and weighted correlation, and then predicted photovoltaic power by an optimized support vector machine (SVM). The literature [12] has constructed a weighted extended daily feature matrix from meteorological information and Pearson correlation coefficients and extracted the feature matrix with the smallest Euclidean distance from the date to be predicted in the historical data as the input to the LSTM for prediction. The above two methods only regard similar-days as similar data. The literature [13] has also used the k-nearest neighbor algorithm to identify data with high similarity as input to the neural network for photovoltaic outflow estimation.

The PV power prediction methods mentioned above are usually only the optimal combination of different algorithms, or a single method of similar data selection is envisaged. In addition, the above methods do not more fully account for the calculation of similarity among the data. There are also many forecasting methods in which only a single forecasting model is used, which does not take into account the full range of factors. Combining previous research performed on similar data selection and the use of neural networks in PV power prediction, this study proposes a PV power prediction method that considers both similar-day and similar-time LightGBM-integrated 1DCNN-LSTM prediction results. The major innovations are as follows:

(1) The similarity of the data is calculated in such a way that the trend similarity at the macro level and the detail similarity at the micro level are considered at the same time. It paves the way for combining the prediction results of similar-day and similar-time training data with more comprehensive considerations by using LightGBM.
(2) In the input data and feature selection for the model, selecting similar data is beneficial to improving the correlation between the model training data and the data to be predicted and improve the quality of model training. Removing features with low correlation and reducing the number of features can reduce overfitting of the model and enhance the understanding between features and eigenvalues.
(3) In the selection of forecasting models, correlations between different variables and the correlation between single-variable data series are considered. CNNs can capture the correlation between different features, while LSTMs pay more attention to the correlation between data series of the same variable. Therefore, we combined the two to get a 1DCNN-LSTM model, which improves the training effect of the model.

2. The Basic Idea of Picking Similar-Days and Similar-Times

Photovoltaic power generation is strongly influenced by meteorological factors [14], none of which is more influential than irradiance, so irradiance was used as a similar variable to select similar-days and similar hours. When photovoltaic power output is generally available from photovoltaic plants is from 05:00 to 19:00, the data during three periods was used as input data for the model.

To select data with a high degree of relevance and complementarity as input data for the prediction model, this paper selected similar-days and similar hours to determine the input data by using the characteristic quantity of irradiance, its historical measured value, and its numerical weather forecast prediction (NWP) value. The selection of similar-days focuses on the similarity of the general trend for all times of the day to control the similarity at a macro level, while the selection of similar-times focuses on the degree of similarity at a given time to consider the degree of similarity at a detailed level.

2.1. Grey Relational Analysis: Select Similar-Days

Grey relational analysis (GRA) [15] determines a correlation by the magnitude of the grey correlation coefficient and measures the degree of correlation between factors based on the similarity or dissimilarity of their trends. GRA is a simple and reliable method of analysis because it does not require an excessive sample size and data to find a typical distribution pattern, and is therefore computationally small.

Calculate the grey correlation coefficient r by means of Equation (1):

$$r_i = \frac{1}{n}\sum_{k=1}^{n} \frac{\min\limits_{i}\min\limits_{k}\left|x_0(k)-x_i(k)\right|+\rho\cdot\max\limits_{i}\max\limits_{k}\left|x_0(k)-x_i(k)\right|}{\left|x_0(k)-x_i(k)\right|+\rho\cdot\max\limits_{i}\max\limits_{k}\left|x_0(k)-x_i(k)\right|} \tag{1}$$

where $x_0(k)$ is the reference column and ρ is the discrimination coefficient. ρ is generally taken as 0.5. The grey correlation is represented by r_i in the equation; when $r_i < 0.5$, it represents little correlation; when $0.5 \le r_i \le 0.7$, it represents some correlation; and when $r_i > 0.7$, it represents a strong correlation.

The steps for selecting similar-days using GRA are as follows:

1. Extract the vector of measured historical daily irradiance values $x_0(k)$ and the vector of predicted NWP irradiance values $x_i(k)$ for the target day.

$$x_0(k) = [x_0(1), x_0(2), \ldots, x_0(k)];\ x_i(k) = [x_i(1), x_i(2), \ldots, x_i(k)] \tag{2}$$

2. Calculate the grey correlation coefficient r_i between the vector $x_0(k)$ and the vector $x_i(k)$ using Equation (1).

The historical days are sorted in descending order by the magnitude of the grey correlation coefficient r_i, and the top j historical days with the largest correlation coefficients are chosen as the target day's similar-days. The value of j can be determined based on the actual situation; using j = 10 yielded better results in this study's data.

2.2. Euclidean Distance: Select Similar-Times

The Euclidean distance (ED) [16] calculates the true distance between two points in n-dimensional space. The calculation of ED is intuitive and easy to implement. The data in this paper are one-dimensional, and the ED can be used to good effect in the case of low-dimensional data.

Calculate the Euclidean distance d, as displayed in Equation (3):

$$d_i = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3}$$

where n represents the n-dimensional space and $x_i(i = 1, 2, \ldots, n)$ is a real number representing the i-th coordinate of x.

At a similar-time, the ED is calculated for point-to-point, 1-dimensional data, so n = 1 in Equation (3). The steps for selecting similar-times using ED are as follows:

1. Determine the measured historical irradiance value $X_i^{(0)}$ for the historical day and the predicted NWP irradiance value $X_i^{(1)}$ for the target day.

$$X_i^{(0)} = \left(X_1^{(0)}, X_2^{(0)}, \ldots, X_i^{(0)}\right);\ X_i^{(1)} = \left(X_1^{(1)}, X_2^{(1)}, \ldots, X_i^{(1)}\right) \tag{4}$$

2. Calculate the Euclidean distance d_i between the measured historical irradiance value $X_i^{(0)}$ and the predicted NWP irradiance value $X_i^{(1)}$ at each moment.

$$d_i = \sqrt{X_i^{(0)} - X_i^{(1)}} \tag{5}$$

The historical moments are sorted in ascending order by the magnitude of the ED d_i, and the top j historical moments with the smallest distance are chosen as the similar-times of the target moments. Take j = 10 to be consistent with the number of similar-days.

3. Input Feature Extraction with the 1DCNN-LSTM Model

The feature data for the corresponding moments of the selected similar-day and similar-time were used as input to the 1DCNN-LSTM prediction model and the power data were used as output. The input feature quantities were extracted in a streamlined manner by random forests.

3.1. Random Forest

Random forest is an algorithm based on decision trees [17], which has important applications in classification, prediction, and feature extraction. Random forests are mainly based on out-of-bag (OOB) error rates to perform feature extraction. For important features, there is a significant drop in the accuracy of the model (i.e., a significant increase in the OOB error rate) after introducing noise to them and performing RF training on data with only this feature variation; conversely, for unimportant features, there is little change in the accuracy of the model after retraining. The steps for performing feature extraction in a random forest are as follows:

1. For each decision tree, the corresponding OOB is selected, and its OOB error is calculated and recorded as $errOOB_1$.
2. Randomly add noise interference to feature X for all OOB samples and again calculate the OOB error, noted as $errOOB_2$.
3. Assuming that there are n trees in the forest, the degree of importance of feature X is given as follows:

$$importance = \frac{1}{n}\sum_{i=1}^{n}|errOOB_2 - errOOB_1| \tag{6}$$

4. The importance of each feature is calculated and sorted in descending order.

3.2. Principle of the 1DCNN

CNNs can be applied to computer vision and equally to time series [18,19]. The convolution kernel of a CNN applied to an image will convolve along the horizontal and vertical axes of the image and is called a two-dimensional convolutional neural network. The convolution kernel of a CNN applied to a time series will only convolve along the time step order and is called a one-dimensional convolutional neural network (1DCNN). Compared to models such as the RNN-improved LSTM, the 1DCNN has the advantage of being able to compute in parallel, has a fast training speed, and can obtain model

results that do not lose out to the LSTM in some application scenarios. A simple graphical explanation of the convolution process is shown in Figure 1.

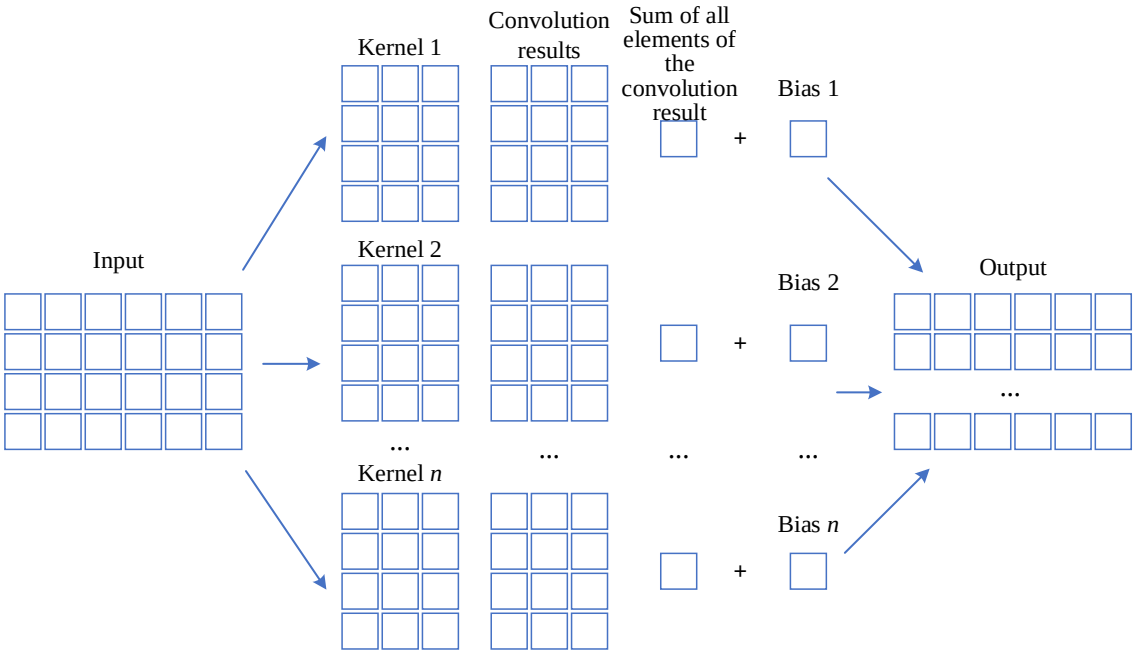

Figure 1. Schematic diagram of the 1DCNN process.

3.3. Structure of the LSTM

LSTMs are used to solve the long-term dependence problem prevalent in recurrent neural networks in general [20] and the gradient disappearance/explosion problem in RNNs [21]. The LSTM modifies the repetition unit based on the RNN; the structure of the LSTM is shown in Figure 2. Each repetition unit is controlled by several gates, from left to right: three gates (three σ), the first of which is the forget gate layer, identifying the information to be selectively forgotten. The second σ is the input gate layer, which determines which information needs to be updated, and the *tanh* layer determines how the information is to be updated. The third σ is the output gate layer; it determines what information is output.

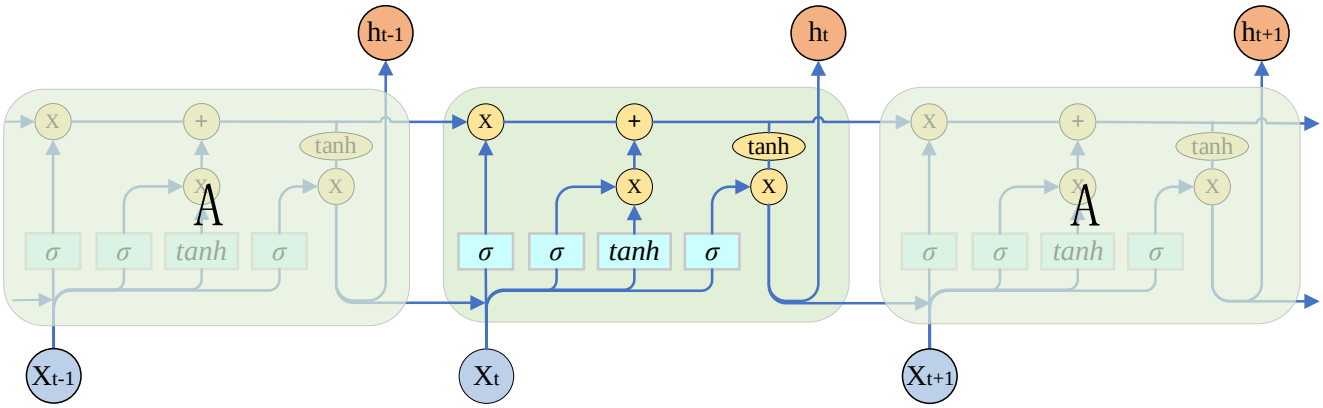

Figure 2. Structural diagram of the LSTM.

3.4. 1DCNN-LSTM

Neither 1DCNNs nor LSTMs alone can simultaneously extract correlations between feature sequences and long-term dependency patterns of features, as the intrinsic relationships between the various features such as irradiance, temperature, and humidity used to

predict photovoltaic power values are not easily described. The model uses a convolutional layer of 1DCNN for feature extraction and sample reconstruction to obtain the fluctuation pattern of the data. Time-domain modelling is then performed using an LSTM, where the LSTM layer performs forgetting learning on the features extracted by the CNN, retaining useful information that can compensate for the long-term dependency problem ignored by the CNN. The final prediction is made through a fully connected layer (Dense).

4. LightGBM Incorporates Similar-Day and Similar-Times

The prediction results of 1DCNN-LSTM with input data of similar-days and similar-times were integrated with the LightGBM algorithm to obtain better results.

4.1. LightGBM Algorithm

The LightGBM [22] algorithm is an improved algorithm in the framework of the GBDT algorithm based on decision trees, which has the advantages of being fast and efficient, having low memory consumption, and supporting data parallelism. "Light" is achieved through gradient-based one-side sampling (GOSS), exclusive feature bundling (EFB), and the histogram algorithm for fewer samples, fewer features, and less memory.

GOSS calculates the information gained by excluding most samples with small gradients and retaining those with larger gradients. EFB improves computational efficiency by binding features (usually mutually exclusive features) and reducing the dimension of features. GOSS and EFB's optimization of sample dimension and feature sampling makes it easier for the histogram algorithm to benefit. The histogram algorithm sorts the features in the sample before training. The features are first divided into histograms, and the histograms are used as "features" in subsequent training to build the decision tree, greatly reducing the complexity and memory consumption of the algorithm. The above is shown in Figure 3.

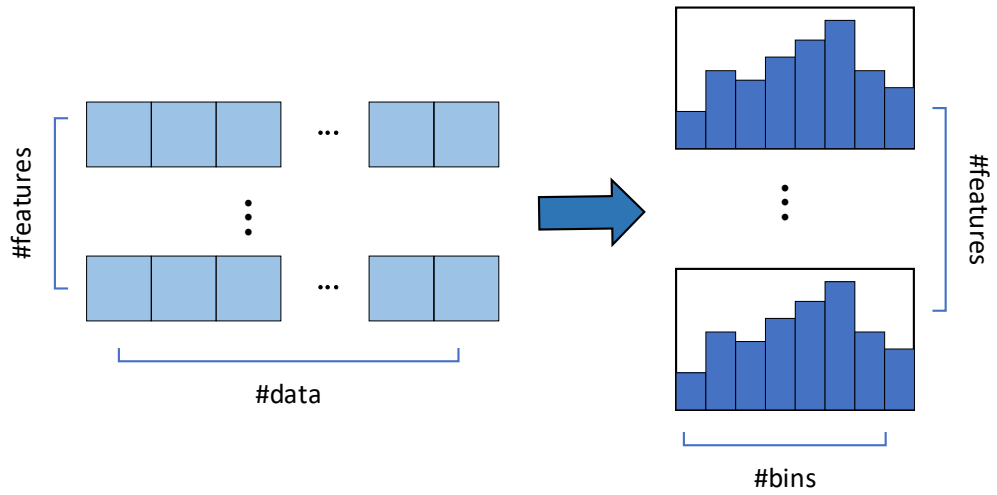

Figure 3. Histogram division of LightGBM.

4.2. LightGBM Integrates Similar-Day and Similar-Time Prediction Processes

When the training data of the prediction model are the similar-day data selected on the target day, the prediction result of the 1DCNN-LSTM prediction model is recorded as p_{sd}. When the training data of the prediction model are the similar-time data selected at the target date, the prediction result of the 1DCNN-LSTM prediction model is p_{st}. Using p_{sd} and p_{st} as inputs to LightGBM and their corresponding true values as outputs of LightGBM, the final predicted values are obtained after integrated learning. The prediction process for integrating 1DCNN-LSTM with similar day and similar time prediction results via LightGBM is shown in Figure 4.

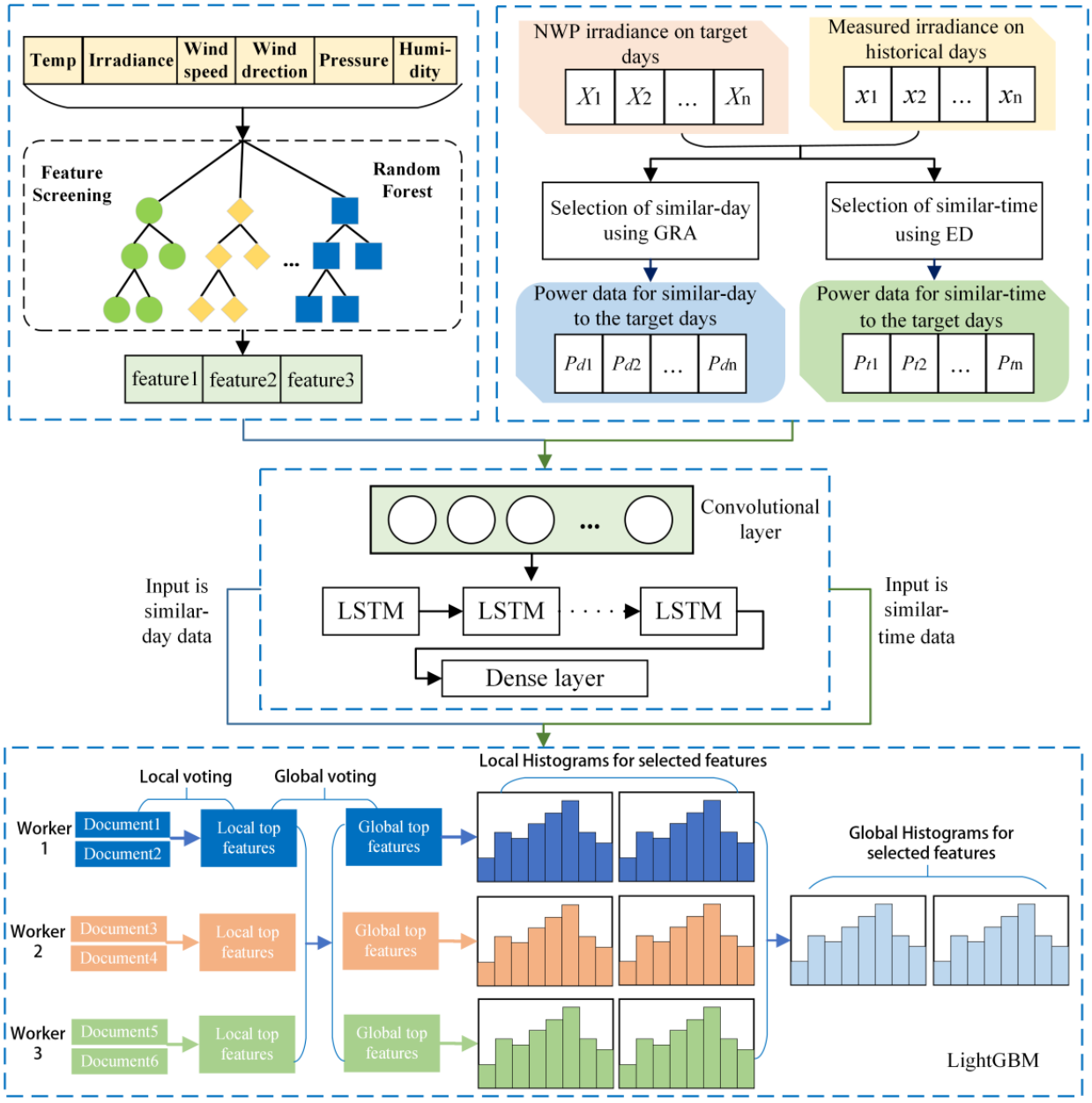

Figure 4. Flowchart of forecasting structure.

5. Discussion

The data set of power station 01, given in reference [23], was used to verify the experiment. From July 2018 to April 2019 were used as historical days, from which similar-day and similar-time were selected for the target date. Use a day in May 2019 as the target day and divide the sunny and non-sunny target days. In this paper, forecasts were made for all 31 days in May 2019, and the results show that the average forecast results for the 31 days after LightGBM integration improved the accuracy of the average forecast results by more than 1% over that of the 1DCNN-LSTM. Here, only the three days of 1 May 2019, 9 May 2019, and 29 May 2019 are shown for sunny target days; only the three days of 4 May 2019, 17 May 2019, and 22 May 2019 are shown for non-sunny target days.

5.1. Precision Evaluation Indicators

The model proposed in this paper was used to predict the photovoltaic output over the next 24 h at 15 min intervals, for a total of 96 moments in 24 h. The root mean squared error (*RMSE*) [24] was chosen as the evaluation metric for the prediction model:

$$RMSE = \frac{1}{P_{ic}} \sqrt{\frac{\sum_{i=1}^{n} \left(P_{pv}(i) - P_{av}(i)\right)^2}{n}} \times 100\%, n = 96 \quad (7)$$

where P_{ic} is the installed capacity of the photovoltaic plant, P_{pv} is the predicted photovoltaic power, and P_{av} is the actual photovoltaic power.

5.2. Selection of Day and Time Similarity

Similar-day and similar-time data were selected for sunny target days 1 May 2019, 9 May 2019, and 29 May 2019 and non-sunny target days 4 May 2019, 17 May 2019, and 22 May 2019, using the selection method for similar-days and similar-times shown in 1.1 and 1.2. A grey correlation analysis was used to calculate the overall correlation from 5:00 to 19:00 so as to select similar dates. We chose similar-days, regardless of sampling intervals, and only obtained one correlation value for one day. We compared the grey correlation between multiple historical dates and the date to be predicted, and selected the date with the highest correlation as the similar-day. Ten similar-days were selected for each of the three sunny target days. The curves of the selected similar-days and the grey correlation between the similar-days and the corresponding target days are shown in Figure 5. Similarly, 10 similar-days were chosen for each of the three non-sunny target days, and Figure 6 depicts their similar-day curves and grey correlations with the corresponding target days.

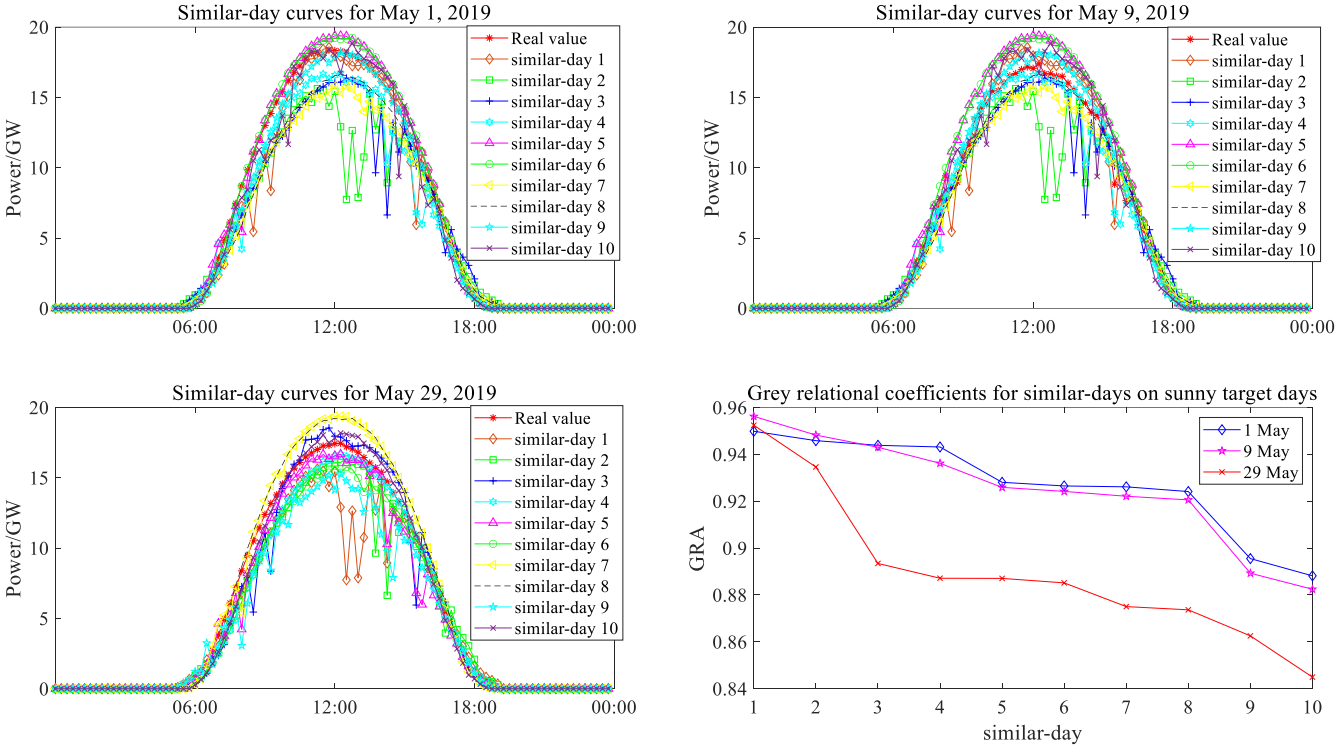

Figure 5. Similar -day curves for sunny target days and their grey relational coefficients.

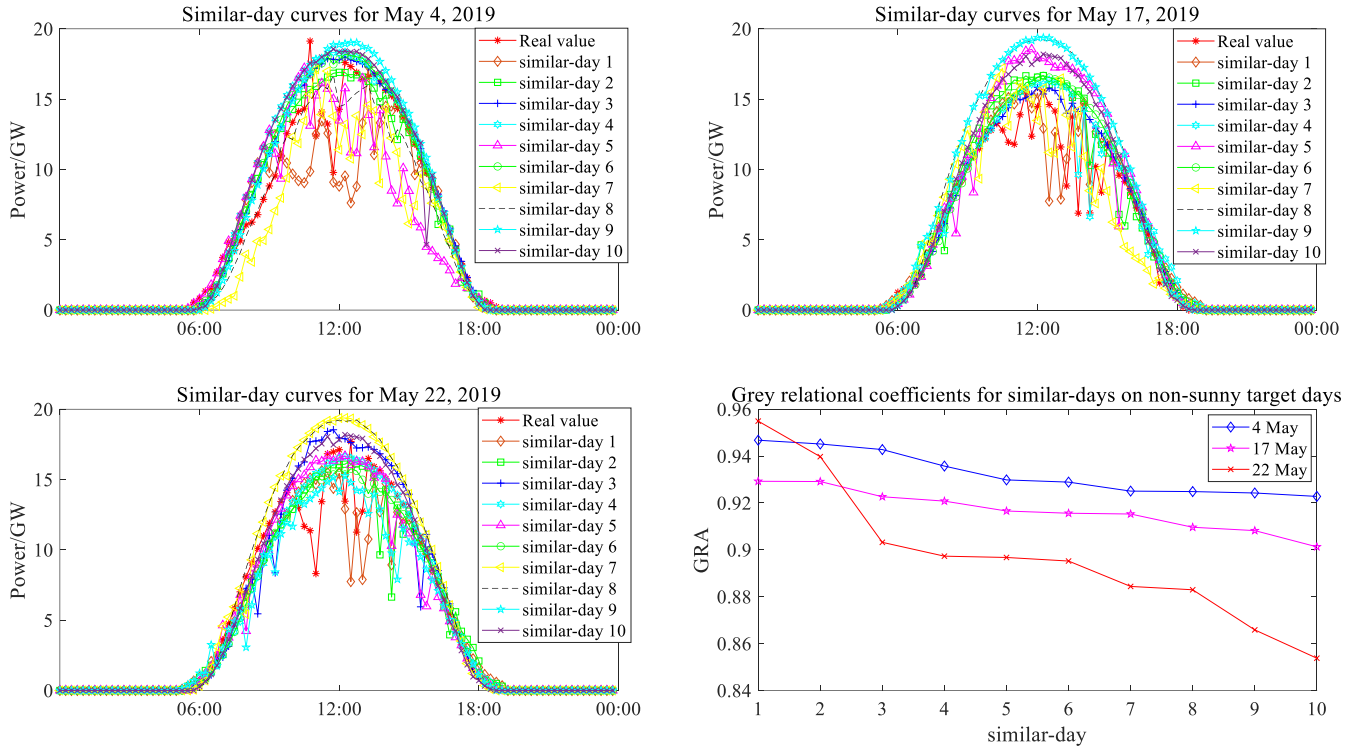

Figure 6. Similar-day curves for non-sunny target days and their grey relational coefficients.

By calculating the Euclidean distance between each sampling point (15 min) from 5:00 to 19:00 and the same time in history, similar-times can be selected. Similar-times were calculated and 57 values were obtained for one day. For each moment, the corresponding moment in history with the smallest Euclidean distance was selected as the similarity moment. The photovoltaic plant has power output from 05:00 to 19:00, with one data point every 15 min, for a total of 57 moments. For each moment of the target day, 10 corresponding similar-time data points were selected, and the similar-time points with the same Euclidean distance ranking among these 57 moments were reconstituted into one day of data. The full-day curve consisting of similar-times selected for sunny target days and the Euclidean distances for all similar moments are shown in Figure 7. The complete one-day curve consisting of similar-times for non-clear target days and the Euclidean distances for all similar-times of the day are shown in Figure 8.

5.3. Feature Extraction for Predictive Model Inputs

There are many factors that affect PV power generation, and too many variables can hinder the method from finding the model. Removing features with low correlation and reducing the number of features can reduce overfitting of the model and enhance the understanding between features and eigenvalues. The six features associated with the PV power factor were therefore feature-extracted and downscaled to three features, which were used as input to the subsequent prediction model. A total of six feature quantities—temperature, irradiance, wind speed, wind direction, air pressure, and humidity (NWP data are used; NWP has an inherent forecast error, which can be offset to some extent by replacing the actual measurement of historical weather data with NWP data)—were -extracted using the random forest algorithm according to the steps shown in Section 3.1. The three features with the highest importance for power were extracted and used as input to the 1DCNN-LSTM prediction model in the next step. Six features were used as inputs to the random forest model, power was used as the labelled output, and the degree of importance of the individual features obtained (retaining the last four decimal places) was ranked in descending order, as shown in Table 1. The importance of the features calculated by the random forest model is shown in Figure 9.

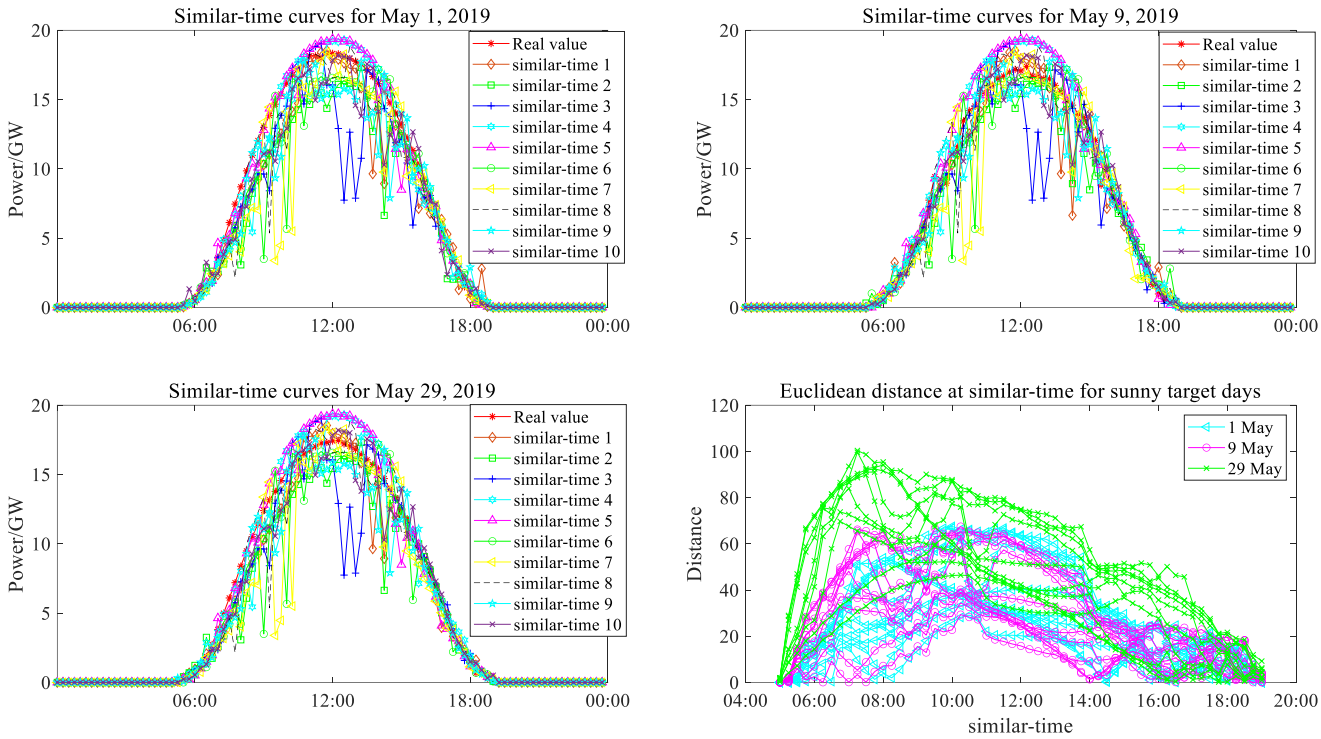

Figure 7. Similar-time curves and their Euclidean distances for sunny target days.

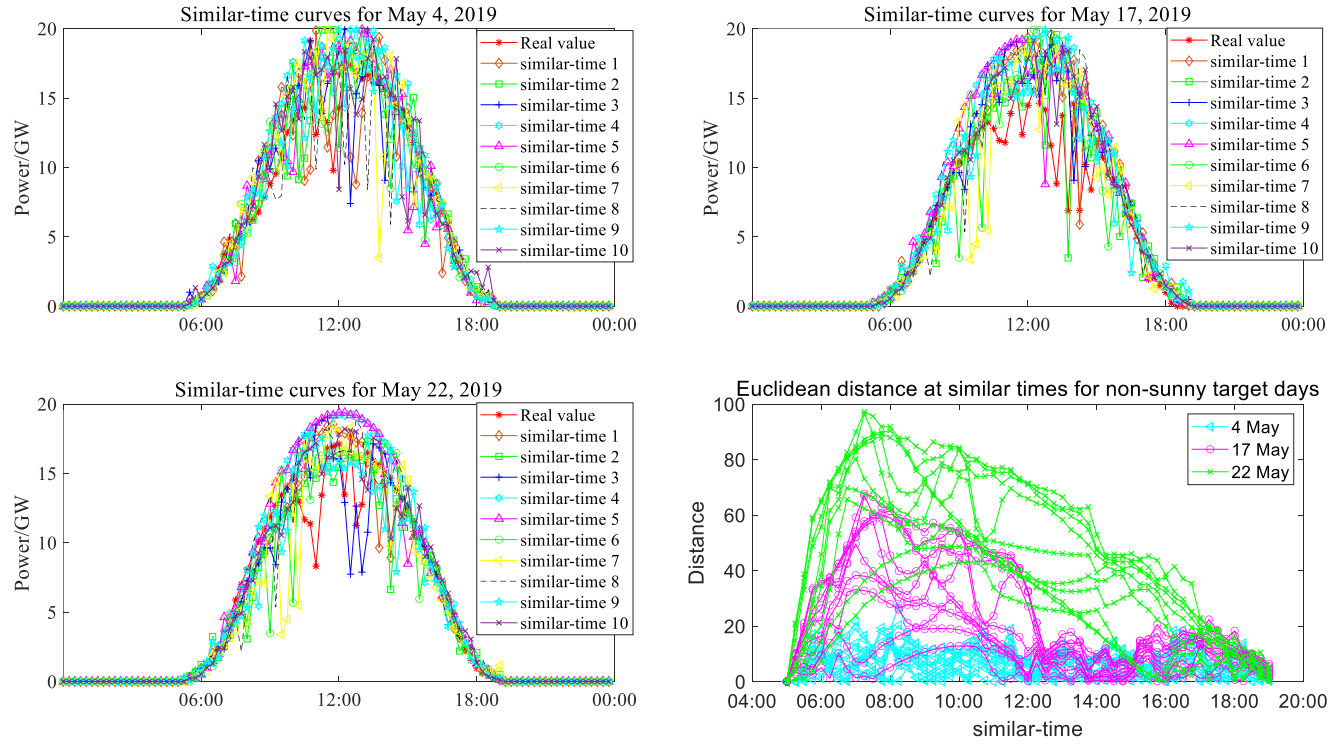

Figure 8. Similar-time curves and their Euclidean distances for non-sunny target days.

Table 1. Ranking of feature importance.

Feature	Irradiance	Humidity	Temp	Wind Speed	Wind Direction	Pressure
Importance	0.6549	0.1151	0.0767	0.0563	0.0496	0.0474

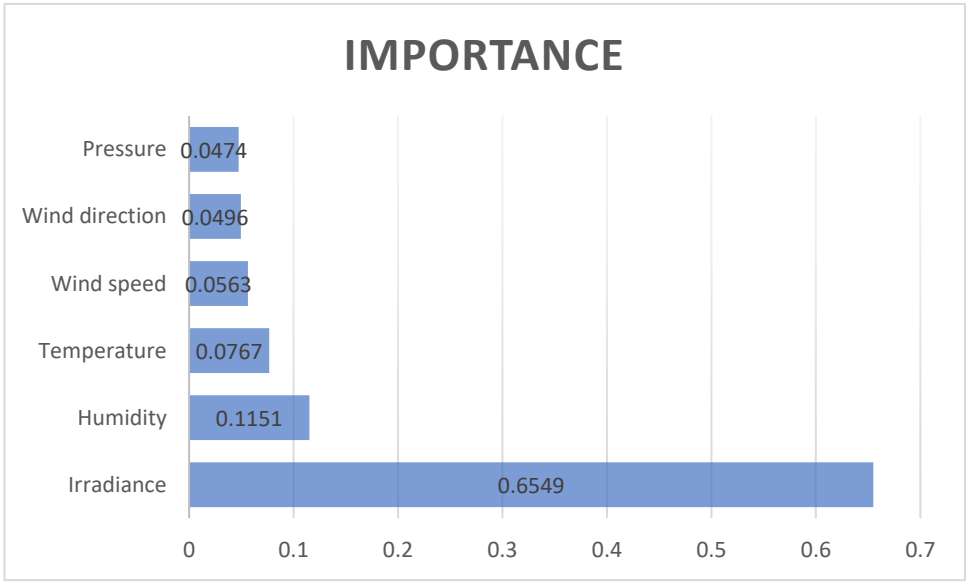

Figure 9. Feature importance.

As shown in Table 1 and Figure 9, the most important characteristic of PV power is irradiance, and its percentage is as high as 0.6549, which is in line with common sense. The degree of importance of humidity and temperature on PV output power is also high. Wind speed, wind direction, and air pressure are less important, indicating that these three characteristics have little influence on photovoltaic power output values. From the six features, three features with high importance to the PV power output value were selected. Irradiance, humidity, and temperature were used as inputs to the 1DCNN-LSTM prediction model, and the predicted values of photovoltaic power were obtained by model training.

5.4. 1DCNN-LSTM Prediction

In this paper, Python 3.6 was used for programming, and the deep learning framework was TensorFlow 1.13.2. The network structure parameters of 1DCNN-LSTM are shown in Table 2. The model input has three features and is trained one row at a time, so the output shape of the input layer is one row and three columns. Since the input layer is the layer of input data, there are no neuron parameters. The input of the *Conv*1*D* layer is the output of the previous layer, and its output has the same shape as the input, all in one row and three columns. The neuronal parameters of the *Conv*1*D* layer are calculated as follows:

$$parameters\_Conv1D = j \times [n \times m \times i + 1] = 3 \times [1 \times 1 \times 3 + 1] = 12$$

where the convolutional kernel volume is $n \times m$, the number of convolutional kernels is j, and the input dimension is i.

Table 2. Structural parameters of 1DCNN-LSTM.

Layer (Type)	Output Shape	Parameter #
Input layer	(None, 1, 3)	0
Conv1D	(None, 1, 3)	12
LSTM	(None, 5)	180
Dense	(None, 1)	6

The output shape dimensions of the *LSTM* layer are user-defined and can be different from the input feature dimensions. This is because there is a nonlinear transformation in

the *LSTM* that converts input dimension to output size. The output size of the *LSTM* here is five. The *LSTM* neuron parameters are calculated as follows:

$$parameters\_LSTM = n \times [d_h \times (d_h + d_x) + d_h] = 4 \times [5 \times (5 + 3) + 5] = 180$$

where n is the number of nonlinear transformations, three σ plus one tanh, with a value of four. d_h is the output dimension, with a value of five. d_x is the input dimensions, with a value of three. The output of the Dense layer is 1-dimensional data with power values. The number of neuron parameters is the output dimension five of the *LSTM* layer plus the output dimension one of the Dense layer, giving a value of six.

The NWP irradiance, humidity, and temperature data were used as inputs to the 1DCNN-LSTM, and the power data as outputs. Training set 1 is the data of the three characteristic quantities and corresponding power of the selected similar-days.The second training data set was three characteristic quantities with power data for the corresponding moments at similar-times. The three features of the target day with the power data at the corresponding moments were used as the test data set. The 1DCNN-LSTM model learns the association between irradiance, humidity, and temperature and power data at each sampling moment (15 min) to make predictions of PV power values. The prediction error rates of 1DCNN-LSTM for different training datasets are shown in Table 3.

Table 3. Prediction error rate using 1DCNN-LSTM on target day.

Weather	Sunny Target Days			Non-Sunny Target Days		
Date	2019.5.1	2019.5.9	2019.5.29	2019.5.4	2019.5.17	2019.5.22
Similar-day	3.18%	5.41%	5.43%	5.90%	7.00%	11.48%
Similar-time	3.09%	5.54%	5.00%	6.17%	6.74%	11.08%

As can be seen from Table 3, sometimes better predictions were obtained with similar-days as training data and sometimes better predictions were obtained with similar-times as training data. The reason why there is such a result may be because the calculation of trend similarity and detail similarity has its own advantages. Therefore, when it is impossible to determine which of these two data will give a better prediction as the training data for 1DCNN-LSTM, we can consider these two similar data selection methods, for example, the comprehensive learning of two predicted results. That is why we suggest considering both trend similarity and detail similarity to select similar data.

5.5. LightGBM Integration Forecast and Comparison

From the analysis in Section 5.4, it can be concluded that when it is uncertain which method will produce better results, we can consider the prediction results obtained from the two similar datasets from the 1DCNN-LSTM model together. Integrating learning is a good option. The photovoltaic power from 1 March 2019 to 30 April 2019 was predicted using the same prediction model and prediction method as shown in 4.4, and the predictions were used as the training set for LightGBM. The prediction results of the 1DCNN-LSTM with different training data on the target day were the test set for LightGBM. The LightGBM model learns the relationship between the two predictions obtained by the 1DCNN-LSTM and the real value at each time period (15 min) of the prediction, thus producing better prediction.

In the following, the prediction results of 1DCNN-LSTM with similar-days as training data are referred to as SD-C-L; the prediction results of 1DCNN-LSTM with similar-times as training data are referred to as ST-C-L; and the results of integrating the two predictions are still referred to as LightGBM.

As can be seen from Table 4, in most cases, the last prediction method keeps the error rate at a low level. Moreover, as far as the average error over all target days is concerned, the forecast error rate after the integration of LightGBM is even lower. The results of the LightGBM-integrated learning reduced the error rate by 1.23% compared to the similar-day C-L approach and by 1.1% compared to the similar-time C-L approach. The reason for this

result is that LightGBM learned both the prediction results obtained from trend similarity and the prediction results obtained from detail similarity, so it is more likely to reduce the prediction deviation. The PV power and its average error curve for sunny target days and non-sunny target days obtained by these three methods from Table 4 are shown in Figures 10 and 11, respectively.

Table 4. Comparison of prediction error rates for different methods.

Weather	Sunny Target Days			Non-Sunny Target Days			Mean Error
Date	1 May 2019	9 May 2019	29 May 2019	4 May 2019	17 May 2019	22 May 2019	Rate
SD-C-L	3.18%	5.41%	5.43%	5.90%	7.00%	11.48%	6.4%
ST-C-L	3.09%	5.54%	5.00%	6.17%	6.74%	11.08%	6.27%
LightGBM	3.37%	3.06%	3.04%	7.98%	5.41%	8.16%	5.17%

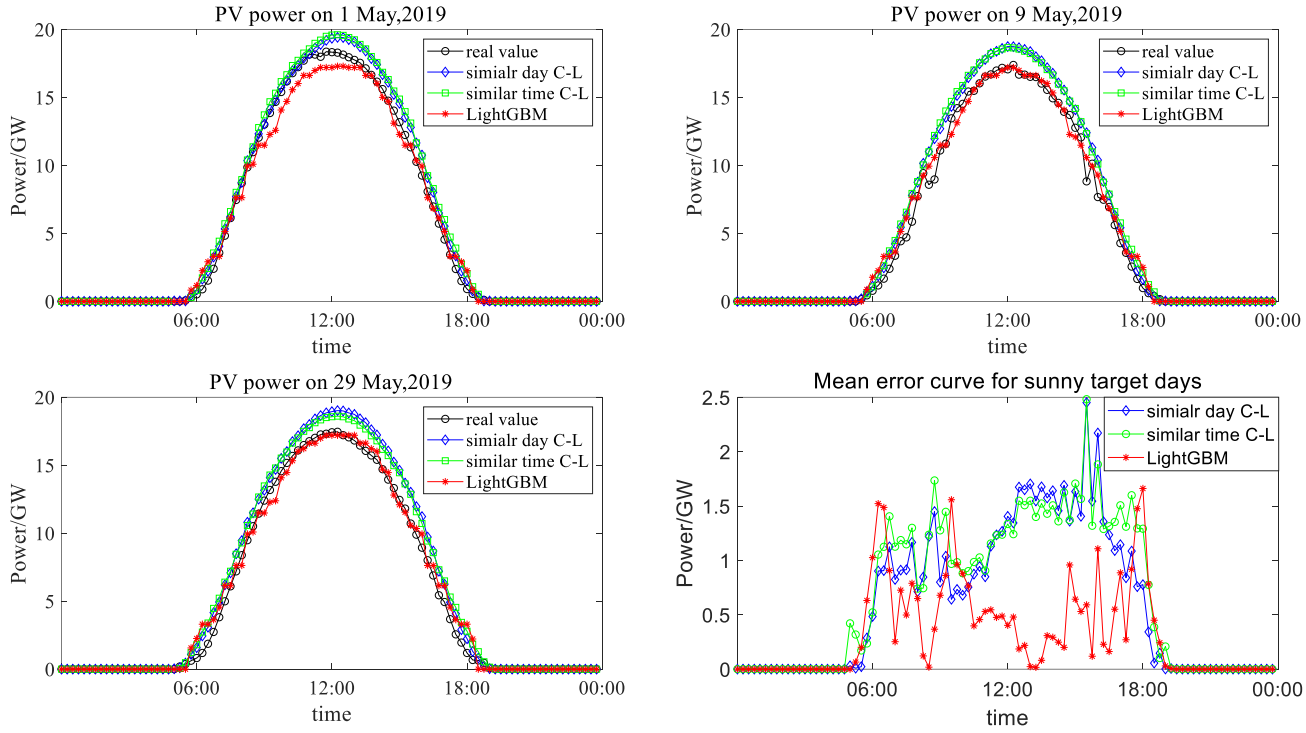

Figure 10. Predicted power and mean error curves for sunny target days.

Similar to the nomenclature in Table 4, the prediction results obtained by using similar-day data as training data for the 1DCNN-LSTM are referred to as "similar-day C-L". The prediction results obtained by using similar-time data as training data for the 1DCNN-LSTM are referred to as "similar-time C-L" and the end results achieved by integrating the two prediction results using LightGBM are called "LightGBM". As can be seen from Figures 10 and 11, the average error curve is closer to a zero value both for sunny and non-sunny target days after the integration of LightGBM. In most cases, the LightGBM curve fits better with the true value curve. It can be seen that the integrated learning of trend similarity and detail similarity provides a better forecast.

In order to verify the effectiveness of this method, this method was compared to other forecasting models. The PV power prediction method based on the multi-resolution 1DCNN-LSTM prediction and integration by LightGBM is called "SD-CL-LGBM" for short. The LSTM, CNN, and BP commonly used in PV power prediction were compared with the SD-CL-LGBM method in the same data situation and their respective prediction error rates are shown in Table 5.

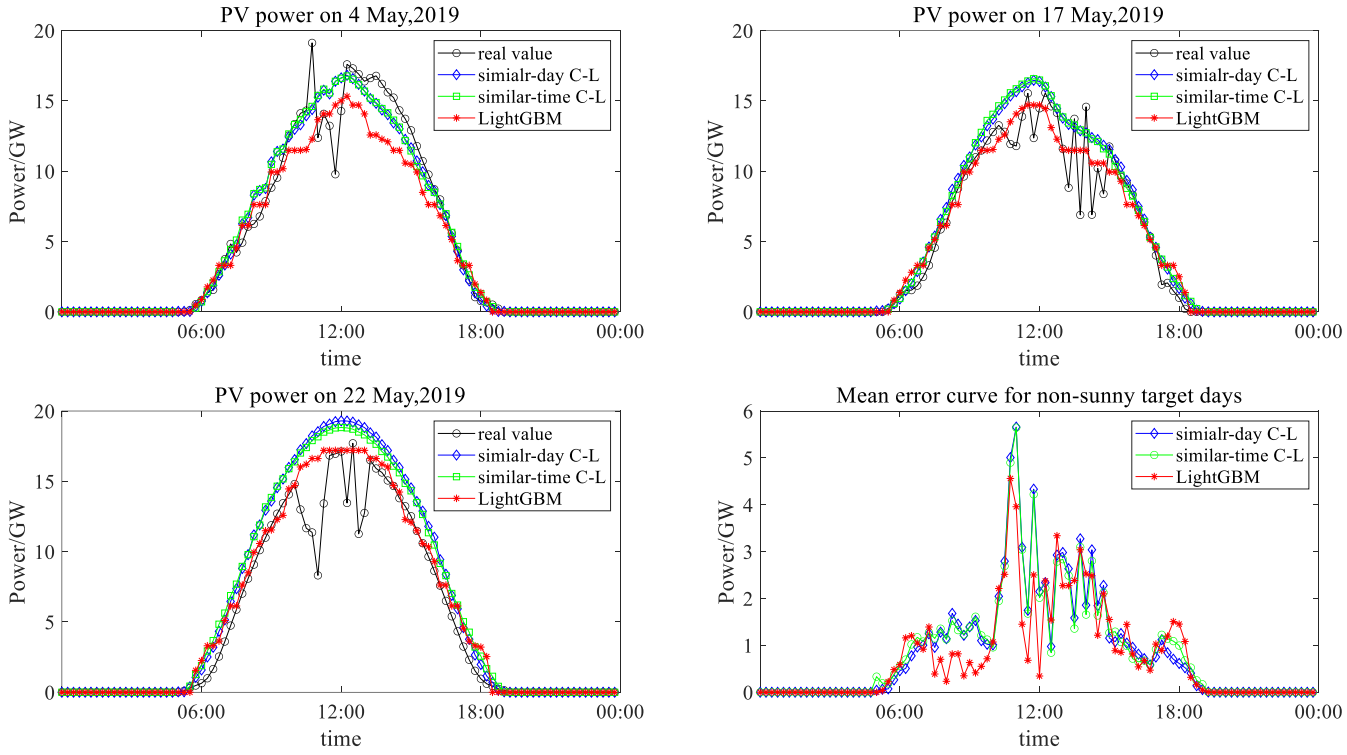

Figure 11. Predicted power and mean error curves for non-sunny target days.

Table 5. Comparison of multiple model prediction errors.

Weather	Sunny Target Days			Non-Sunny Target Days			Mean Error
Date	**2019.5.1**	**2019.5.9**	**2019.5.29**	**2019.5.4**	**2019.5.17**	**2019.5.22**	**Rate**
LSTM	5.74%	5.68%	5.27%	7.75%	6.89%	8.57%	6.65%
CNN	10.80%	15.72%	18.30%	12.02%	17.84%	24.74%	16.57%
BP	8.49%	10.88%	11.29%	9.80%	15.31%	13.94%	11.62%
SD-CL-LGBM	3.37%	3.06%	3.04%	7.98%	5.41%	8.16%	5.17%

The LSTM model is a powerful time-series forecasting model. In order to maintain the consistency of data and make use of the time series characteristics of the LSTM, the power data from July 2018 to April 2019 were used as historical training data for the LSTM to predict the data for May 2019. CNNs can be used for multi-feature regression prediction; therefore, features such as NWP irradiance, NWP humidity, and NWP temperature were used as the CNN input, and power data were used as outputs. Likewise, we used July 2018 to April 2019 as historical data and May 2019 as the data to be predicted. Because the training of the BP neural network used for multi-feature regression prediction was poor with the current data, only the NWP irradiance was used as an input for the BP model power prediction. The time range of the historical data and the data to be predicted were consistent with the other models. It can be seen in Table 5 that the SD-CL-LGBM forecast was better for almost all target days. In terms of the average error rate for all target days, the proposed method was lower than any other models. The average error rate of the SD-CL-LGBM model was 1.48% lower than that of the LSTM model, 11.4% lower than that of the CNN model, and 6.45% lower than that of the BP model. This shows that the proposed method has better prediction in the case of the data in this paper. Figures 12 and 13 show the comparison of the forecast results of the SD-CL-LGBM model with the LSTM model, the CNN model, and the BP model.

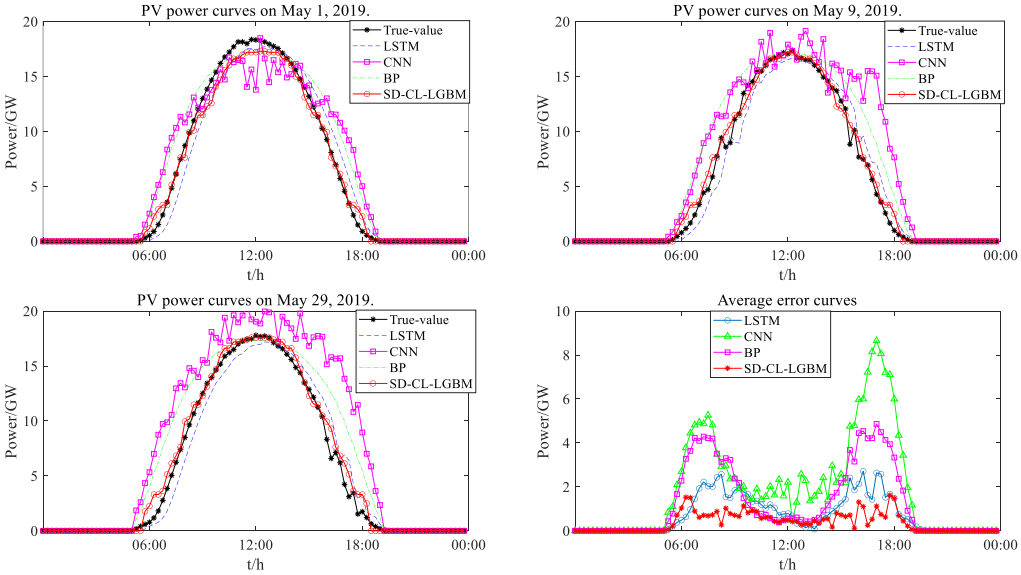

Figure 12. Comparison of prediction results of different models for sunny target days.

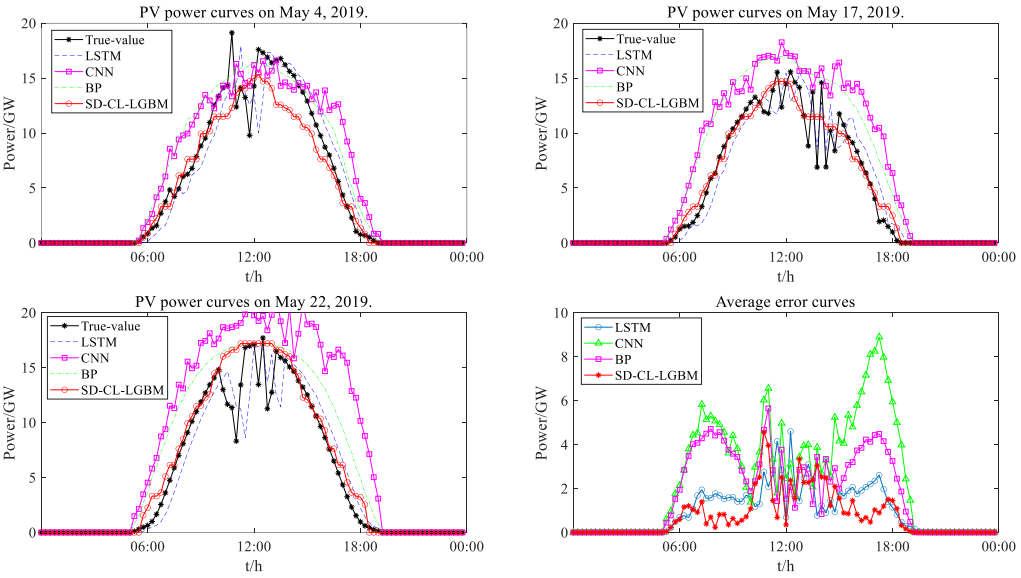

Figure 13. Comparison of prediction results of different models for non-sunny target days.

As can be seen from Figures 12 and 13, the predicted curve of the CNN model is obviously different from the true value curve, and has the maximum average error curve. The reason for this phenomenon may be that CNNs' strong ability to perceive relationships between multiple features has resulted in overfitting. The shape of the prediction curve of the LSTM model is closest to the real value curve, but there is a prediction lag. This may be related to the prediction mechanism of the LSTM, in which the prediction is inferred every n steps, which has a certain lag. The prediction curve of the BP model is not lagging behind, and its shape is consistent with the real value, but the prediction accuracy needs to be improved. Considering the training results with similar trend and details, an integrated learning method, SD-CL-LGBM, which combines CNN and LSTM, has the best prediction effect. The prediction curves of the proposed method are closest to the true value curves and the average prediction error curve is closest to the zero value. Therefore, this method is effective to some extent.

6. Conclusions

(1) Firstly, grey correlation analysis is applied to calculate the correlation between historical days and target days to select similar-days, and Euclidean distance is applied to calculate the distance between historical moments and target moments to select similar-times, so as to obtain highly correlated training data as the input of the prediction model. The training data were selected considering the degree of similarity at multiple resolutions, both from a macro perspective of the overall trend over the whole day and from a detailed perspective of the individual moments. It is helpful to extract similar data more comprehensively by combining trends and details in the selection of similar data.

(2) Secondly, the three characteristic quantities with the highest degree of importance were screened by the random forest algorithm, and the data of the three characteristic quantities at each time point between similar-days and similar moments were used as the input data of the prediction model. After screening the random forest model, it was decided that irradiance, humidity, and temperature data for similar-days and similar-times, together with NWP irradiance, NWP humidity, and NWP temperature data for the day to be predicted would be used as sample data to construct the prediction model. Reducing the number of features input to the model is helpful for the model to find the pattern between features and tags, and improves the training efficiency of the model.

(3) Then, the prediction model uses a 1DCNN-LSTM model. The 1DCNN model can focus on the correlation between different feature sequences, such as temperature, humidity, irradiance, etc.; the LSTM model can extract long-term dependency patterns between data sequences; and the combination of the two can compensate for each other's shortcomings. It is helpful to improve the prediction accuracy of the model to consider the correlation between different features and the dependence between single variable sequences.

(4) Finally, multiple predictions from the 1DCNN-LSTM considering multi-resolution similarity were integrated using the LightGBM-integrated learner to obtain the final prediction results. Experiments in this paper show that the proposed LightGBM-integrated 1DCNN-LSTM with a similar-day and similar-time prediction model has improved accuracy compared to the single prediction model. This idea of integrated learning for PV power prediction, considering the similarity degree of multiple resolutions on similar-days and at similar moments, as proposed above, has some practical implications for PV grid connection and grid scheduling. Providing more comprehensive and extensive information for integrated learners will help to improve their training effectiveness. Future work can try to increase the range of information provided for integrated learners to improve prediction accuracy.

(5) As this study uses NWP data, the NWP itself has certain errors and therefore will have some impact on the prediction results. Subsequent research will be enhanced to study how to reduce the impact of NWP errors on the prediction results in order to improve the prediction method in this paper.

Author Contributions: Y.P.: conceptualization, methodology, investigation, writing—original draft. S.W.: formal analysis, visualization, revision of manuscript. W.C.: conceptualization, resources, writing—review and editing. J.M.: conceptualization, validation, writing—reviewing and editing. C.W.: resources, formal analysis, writing—review and editing. J.C.: methodology, data curation, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Project, grant number No. 5211DS220009. The APC was funded by the State Grid Zhejiang Electric Power Co., Ltd.

Data Availability Statement: The data used are from an article published in *Solar Energy*. For more information see reference [20].

Acknowledgments: All authors would like to thank to the State Grid Zhejiang Electric Power Co., Ltd. for financial support and facilities.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

PV	photovoltaic
1DCNN	one-dimensional convolutional neural network
CNN	convolutional neural network
LSTM	long- and short-term memory neural network
LightGBM	light gradient boosting machine
BP	back propagation
NWP	numerical weather prediction
GRA	grey relational analysis
ED	Euclidean distance
OOB	out-of-bag
RNN	recurrent neural network
GOSS	gradient-based one-side sampling
EFB	exclusive feature bundling
GBDT	gradient boosting decision tree
RF	random forest

References

1. Xu, Y.; Zhang, B.; Huang, J.; Xie, X.; Wang, R.; Shen, D.; He, L.; Yang, K.L. Forecast of Photovoltaic Power Based on IWPA-LSSVM Considering Weather Type and Similar Day. 2022. Available online: https://kns-cnki-net.webvpn.ncepu.edu.cn/kcms/detail/11.3265.TM.20211109.1202.002.html (accessed on 22 November 2022).
2. Serrano Ardila, V.M.; Maciel, J.N.; Ledesma, J.J.G.; Ando Junior, O.H. Fuzzy time series methods applied to (In) direct short-term photovoltaic power forecasting. *Energies* **2022**, *15*, 845. [CrossRef]
3. Li, F.; Li, C.; Mi, Q.; Song, Q.; Cui, Y.; Zhao, J. The time-varying weight ensemble forecasting of short-term photovoltaic power based on GRA-BPNN. *Renew. Energy Resour.* **2018**, *36*, 1605–1611.
4. An, Y.; Sun, K. Photovoltaic Power Prediction Based on Similar Day and Echo State Networks. *Smart Power* **2020**, *48*, 38–43.
5. Wu, S. Review of Power Forecasting Methods of Photovoltaic Power Generation System. *J. Eng. Therm. Energy Power* **2021**, *36*, 1–7.
6. Si, Z.; Yang, M.; Yu, Y.; Ding, T.; Li, M. A Hybrid Photovoltaic Power Prediction Model Based on Multi-source Data Fusion and Deep Learning. In Proceedings of the 2020 IEEE 3rd Student Conference on Electrical Machines and Systems (SCEMS), Jinan, China, 4–6 December 2020; pp. 608–613. [CrossRef]
7. Zhang, H.; Liu, D.; Zhu, T.; Liu, G.; Hu, H.; Xiao, C. Short—Term PV Power Prediction Based on BP Neural Network Optimized by Similar Daily and Momentum Method. *Smart Power* **2021**, *49*, 46–52.
8. Sharadga, H.; Hajimirza, S.; Balog, R.S. Time series forecasting of solar power generation for large-scale photovoltaic plants. *Renew. Energy* **2020**, *150*, 797–807. [CrossRef]
9. Yang, S.; Luo, D.S.; He, H.; Yang, J.W.; Hu, S. Output Power Forecast of Photovoltaic Power System Based on EMD-LSSVM Model. *Acta Energ. Sol. Sin.* **2016**, *37*, 1387–1395.
10. Zhang, S.; Dong, L.; Ji, D.Y.; Hao, Y.; Zhang, X.F. Power forecasting of ultra-short-term photovoltaic station based on NWP similarity analysis. *Acta Energ. Sol. Sin.* **2022**, *43*, 142–147.
11. Ge, L.; Lu, W.; Yuan, X.; Zhou, Q. Power forecasting of photovoltaic plant based on improved similar day and ABC-SVM. *Acta Energ. Sol. Sin.* **2018**, *39*, 775–782.
12. Zheng, R.; Li, G.; Hai, B.; Wang, K.; Peng, D. Day-ahead power forecasting of distributed photovoltaic generation based on weighted expanded daily feature matrix. *Electr. Power Autom. Equip.* **2022**, *42*, 99–105.
13. Cheng, Z.; Liu, C.; Liu, L. A Method of Probabilistic Distribution Estimation of Photovoltaic Generation Based on Similar Time of Day. *Power Syst. Technol.* **2017**, *41*, 448–455.
14. Su, R.; Ding, X.; Sun, F.; Han, Y.; Liu, H.; Yan, J. Ultra-short-term photovoltaic power prediction based on Wide & Deep-XGB2LSTM model. *Electr. Power Autom. Equip.* **2021**, *41*, 31–37.
15. Tong, Z.; Zhong, J.; Li, Z.; Wu, J.; Li, J. Short-Term Load Forecasting Based on Grey Relational Analysis and CNN-LSTM. *Electrotech. Electr.* **2022**, *8*, 17–22.
16. Liao, W.; Zhang, R.; Yu, W.; Wang, G. Prediction of output power of photovoltaic based on similar samples and principal component analysis. *Acta Energ. Sol. Sin.* **2016**, *37*, 2377–2385.
17. Chen, T.; Wang, Y.; Ji, Z. Combination Forecasting Model of Photovoltaic Power Based on Empirical Wavelet Transform. *J. Syst. Simul.* **2021**, *33*, 2627–2635.

18. Mao, M.; Feng, X.; Chen, S.; Wang, L. A Novel Maximum Power Point Voltage Forecasting Method for Pavement Photovoltaic Array Based on Bayesian Optimization Convolutional Neural Network. 2022. Available online: https://kns-cnki-net-443.webvpn.ncepu.edu.cn/kcms/detail/detail.aspx?dbcode=CAPJ&dbname=CAPJLAST&filename=ZGDC20221117000&uniplatform=NZKPT&v=66MZTQRHAarHZy5shy2NCBoemqrmklu9nEvCwIbQqyY7QyPkbry7hcfzUV6RcW8O (accessed on 9 December 2022).
19. Liu, X.; Mu, Y.; Wu, Z.; Yan, K. Super-Short-Term Photovoltaic Power Forecasting Based on DWT-CNN-LSTM. *J. Zhengzhou Univ. (Nat. Sci. Ed.)* **2022**, *54*, 86–94.
20. Wang, K.; Du, H.; Jia, R.; Liu, H.; Liang, Y.; Wang, X. Short-term Interval Probability Prediction of Photovoltaic Power Based on Similar Daily Clustering and QR-CNN-BiLSTM Model. *High Volt. Eng.* **2022**, *48*, 4372–4388.
21. Yang, X.; Zhao, Z.; Yang, Y. Research on distributed photovoltaic power prediction method based on combination of spatiotemporal information. *Therm. Power Gener.* **2022**, *51*, 64–72.
22. Wang, J.; Bi, L.; Zhang, K.; Sun, P.; Ma, X. Short-Term Photovoltaic Generation Prediction Based on Multi-Feature Fusion and Xgboost-Lightgbm-Convlstm. 2022. Available online: https://kns-cnki-net-443.webvpn.ncepu.edu.cn/kcms/detail/detail.aspx?dbcode=CAPJ&dbname=CAPJLAST&filename=TYLX20220906000&uniplatform=NZKPT&v=4-88attwqNOWBLwQLAgNB9v_lrolXXL0vOOw-wFLesBWmaUp2Xc57f5uSEvrtxIe (accessed on 20 November 2022).
23. Yao, T.; Wang, J.; Wu, H.; Zhang, P. A photovoltaic power output dataset: Multi-source photovoltaic power output dataset with Python toolkit. *Sol. Energy* **2021**, *230*, 122–130. [CrossRef]
24. Yang, X.; Yang, Y.; Meng, L. Data Sharing and GRA Weight Optimization for Power Prediction of Distributed Photovoltaic Power Plant Considering Missing NWP Information. 2022. Available online: https://kns-cnki-net.webvpn.ncepu.edu.cn/kcms/detail/23.1202.TH.20220704.1020.004.html (accessed on 28 November 2022).