energies

Article

Simplified Method for Predicting Hourly Global Solar Radiation Using Extraterrestrial Radiation and Limited Weather Forecast Parameters

Xinyu Yang [1,2], Ying Ji [1,2,*], Xiaoxia Wang [1,2], Menghan Niu [1,2], Shuijing Long [3], Jingchao Xie [1,2] and Yuying Sun [1,2]

[1] Beijing Key Laboratory of Green Building Environment and Energy Saving Technology, Beijing University of Technology, Beijing 100124, China
[2] Faculty of Architecture, Civil and Transportation Engineering, Beijing University of Technology, Beijing 100124, China
[3] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
* Correspondence: jyesse2015@163.com

Abstract: Solar radiation has important impacts on buildings such as for cooling/heating load forecasting, energy consumption forecasting, and multi-energy complementary optimization. Two types of solar radiation data are commonly used in buildings: radiation data in typical meteorological years and measured radiation data from meteorological stations, both of which are types of historical data. However, it is difficult to predict the hourly global solar radiation, which affects the application of relevant prediction models in practical engineering. Most existing methods for predicting hourly global solar radiation have issues such as difficulty in obtaining input parameters or complex data processing, which limits their practical engineering applications. This study proposed a simplified method to accurately predict the hourly horizontal solar radiation using extraterrestrial solar radiation, weather types, cloud cover, air temperature, relative humidity, and time as the input parameters. The back-propagation network, support vector machine, and light gradient boosting machine (LightGBM) models were used to establish the prediction model, and Shapley additive explanations were used to analyze the relationship between the input variables and the prediction results to simplify the structure of the prediction model. Taking Lanzhou New District in Gansu Province as an example, the results showed that the LightGBM model performed the best, with the root mean square error of 126.1 W/m^2. Shapley additive explanations analysis showed that weather type was not a significant factor in the LightGBM model. Therefore, the weather type was removed from the LightGBM model and the root mean square error was 135.2 W/m^2. The results showed that extra-terrestrial radiation and limited weather forecast parameters can be used to predict hourly global solar radiation with satisfactory prediction results.

Keywords: hourly global solar radiation; simplified prediction method; extraterrestrial solar radiation; LightGBM; SHAP analysis

Citation: Yang, X.; Ji, Y.; Wang, X.; Niu, M.; Long, S.; Xie, J.; Sun, Y. Simplified Method for Predicting Hourly Global Solar Radiation Using Extraterrestrial Radiation and Limited Weather Forecast Parameters. *Energies* **2023**, *16*, 3215. https://doi.org/10.3390/en16073215

Academic Editor: Jesús Polo

Received: 8 March 2023
Revised: 22 March 2023
Accepted: 28 March 2023
Published: 03 April 2023

1. Introduction

In recent years, global resources have become scarce and the environment is deteriorating. To improve the global environment, many countries have committed to energy conservation and an emissions reduction. The United States, Canada, the European Union, Japan, and other countries have promised to achieve carbon neutrality by 2050, and China has proposed achieving carbon neutrality by 2060. To achieve carbon neutrality, energy conservation in building operations is imperative. According to statistics from the International Energy Agency, the building and building construction sectors account for more than one-third of the total global final energy consumption, and the total amount of direct and indirect carbon emissions from electricity and commercial heat used in buildings has risen to 10 Gt, the highest level ever recorded. The building energy consumption increased

from 118 EJ in 2010 to 128 EJ in 2019 [1]. In 2019, carbon emissions from the operation of buildings in the European Union reached 980 million tons [2], and carbon emissions from residential and commercial sources in the United States were 1.856 billion tCO_2 [3], accounting for 36% of the total carbon emissions of the United States. International Energy Agency statistics indicate that residential energy consumption in various countries is at a high level. In 2019, the residential energy consumption was approximately 1.5 EJ in Canada, 1.8 EJ in Japan, and 1.6 EJ in the UK. In the same year, the total commodity energy consumption of China's building operation was 1.02 billion tCO_2, accounting for 21% of China's total energy consumption in that year [4]. Accurate prediction of a building's cooling and heating load, energy consumption, and solar energy availability for the next few days is an important way to achieve building energy savings. As the main impact factor of building load, meteorological conditions are an important factor in the load prediction model. Air temperature, relative humidity, wind speed, wind direction, cloud cover, and weather types for the next few days can be obtained through meteorological forecasts. However, the hourly global solar radiation cannot be predicted using weather forecasts. Hourly global solar radiation is the main factor affecting building load and is an important input for building load prediction. The accurate prediction of hourly global solar radiation can also improve solar energy utilization. Therefore, accurately and simply predicting hourly global solar radiation is a problem worth discussing.

At present, some scholars have conducted relevant research on the prediction of hourly global solar radiation, and the data processing methods, model infrastructure, input parameters, and original data scale vary. With the rapid development of machine learning, its application in the field of hourly global solar radiation prediction has gradually increased. Machine learning models can automatically learn the relationship between the input and target parameters, and many studies have used these models to establish hourly global solar radiation prediction models.

Wang et al. [5] decomposed the daily average solar radiation intensity into intrinsic mode functions as inputs to the daily solar radiation models. The hybrid empirical mode decomposition (EEMD) and regression model (RE) model had the best performance, with the root mean square error (RMSE) of 1.135 when the daily solar radiation was predicted. Bou-Rabee M A et al. [6] developed bidirectional long short-term memory (BiLSTM) to predict the daily solar radiation. This model took the historical time-series data as the input variable, and the RMSE during sunny and cloudy conditions were 4.24 W/m^2 and 20.95 W/m^2, respectively. These studies confirmed the feasibility of machine learning for solar radiation prediction. However, the above methods are suitable for the prediction of daily global radiation.

Jiménez-Pérez et al. [7] used a clustering algorithm to divide the types of days into four categories based on the solar global horizontal irradiance received in a period. According to the different types of days, support vector machines (SVM) and artificial neural networks (ANNs) were applied to establish hourly global solar radiation prediction models. The model parameters included air temperature, relative humidity, and atmospheric pressure. The results showed that the SVM model exhibited the best performance. When the input variables were the values of the meteorological parameters for the previous day, the RMSE was 147 W/m^2. When the input variables were the forecasts of the meteorological parameters for the same day, the RMSE was 119 W/m^2. Lan et al. [8] used discrete Fourier transform to extract the frequency features of historical solar radiation data from five locations: Dalian, Weihai, Qingdao, Dafeng, and Shanghai. Principal component analysis was applied to identify the crucial frequency features, which were input into an Elman-based neural network to predict Qingdao's solar radiation in the subsequent 24 h. The minimum RMSE value of the model appeared in autumn at 72.95 W/m^2, while the maximum appeared in spring at 191.33 W/m^2. Yong Zhou et al. [9] proposed an attention-based transformer model to predict the future 10 h global solar radiation, where the input parameters were the historical 70 h global solar radiation. The RMSE varied from 63.54 to 81.28 W/m^2. The above studies used historical solar radiation as the input parameter of

the model, and the flexibility of these methods was limited by the historical solar radiation acquisition.

Kuk et al. [10] used the K-means clustering algorithm to collect meteorological data and divided the weather data into three classifications: sunny days, partially cloudy days, and cloudy or rainy days. Then, they established a prediction scheme for the hourly prediction of solar irradiance based on weather classification and the SVM model. The input parameters used in the model were sunshine duration, cloud cover, cloud type, sunshine, relative humidity, precipitation, air temperature, and wind speed. The RMSE of the model under the three weather types was 49.26, 62.57, and 57.87 W/m^2, respectively. Li et al. [11] conducted a sensitivity analysis to evaluate the contribution of each input parameter and the most significant five climatic variables were selected as inputs of various multivariate adaptive regression spline (MARS) models. Hourly global solar radiation prediction models were established based on horizontal extraterrestrial solar radiation, sunshine duration, visibility, amount of cloud cover, and wind speed. The lowest RMSE of the models was 76.1 W/m^2. Wang et al. [12] used correlation analysis to screen six parameters that were significantly related to the actual radiation level: atmospheric pressure, air temperature, relative humidity, precipitation, actual sunshine duration, and solar altitude angle. Using data from 2009 to 2019 from Haikou, the Elman neural network model was trained and established to predict the hourly solar radiation. The lowest RMSE of the model was 44.44 W/m^2. All of these studies used the actual sunshine durations for which predictions were unavailable. The actual sunshine duration can only be measured; therefore, the predicted value cannot be obtained. Therefore, the method proposed in this study has some limitations in practical applications for predicting the hourly global solar radiation.

Meenu et al. [13] established a convolutional long short-term memory fusion network (CNN-LSTM) to predict solar radiation for 15–150 s in advance. LSTM was applied to extract the time-series features of 10 past time steps of the solar radiation values, and the CNN was used to extract features from the cloud cover satellite images. The highest accuracy rate of the model was 99.23%. Francisco et al. [14] used satellite images, cloud data, direct solar radiation, and diffuse solar radiation data to predict the solar radiation level under different cloud types 90 min in advance. The model based on satellite data had the highest accuracy in predicting the radiation value under cumulus clouds, with an RMSE of approximately 100 W/m^2. These methods require the collection and analysis of satellite images, and the processing and analysis of satellite images is relatively complicated. Therefore, it is difficult to meet the requirements of general engineering practice and this type of model is unsuitable for practical engineering applications.

In summary, the limitations of most existing machine learning models for hourly solar radiation prediction are as follows. (1) Some methods that use historical solar radiation as the input parameter cannot eliminate the dependence on historical data. It is necessary to obtain the historical solar radiation to ensure the operation of the model. (2) Some hourly radiation prediction models require the actual sunshine duration as an input parameter, which refers to the time in a day when the sun is directly on the ground and can only be recorded by meteorological equipment. Because the parameter cannot be predicted, it affects the prediction function of solar radiation prediction models. (3) Some prediction models must acquire and analyze satellite images, which are difficult to obtain and analyze. Complex image processing limits the efficiency of radiation-prediction models.

To solve these problems, this study proposed a simplified method to predict the hourly global solar radiation, which has the input parameters of extraterrestrial solar radiation, weather types, cloud cover, air temperature, relative humidity, and time. The remainder of this paper is organized as follows. Section 2 introduces the method of the prediction model including the input parameters of the model, the algorithm of the model, the pre-processing of the original data, and the evaluation indicators of the method. Section 3 presents a case study and the performance of each model. Section 4 analyzes the importance of the model input parameters for the prediction results. Finally, Section 5 summarizes the study.

2. Methods

Extraterrestrial solar radiation varies with geographic location and the level of extraterrestrial radiation at the same location does not change significantly every year [15]. Extraterrestrial solar radiation reaches the surface of the Earth after being attenuated by the atmosphere; therefore, clouds in the atmosphere will cause the attenuation of extraterrestrial solar radiation through direct reflection or shortwave radiation [16]. Some studies have indicated that the presence of clouds and water vapor affects the level of global solar radiation [17–19]. The weather type and relative humidity can characterize the cloud amount and water vapor condition, respectively, to a certain extent. Weather type and relative humidity are the parameters for which the forecast values are easily obtained. The air temperature changes after the ground receives solar radiation. Relevant studies have also shown a relationship between air temperature and solar radiation [18,20]. According to the law of relative motion between the Sun and the Earth, there are inter-day and inter-annual variations in global solar radiation, so time is also one of the factors affecting global solar radiation. In summary, combined with the research summarized in Section 2.2 and the input parameters of the solar radiation prediction model summarized in the related literature review [21], it can be preliminarily determined that the input parameters selected in this study are extra-terrestrial solar radiation, weather types, cloud cover, air temperature, relative humidity, date, and hours.

Three typical algorithms were selected to establish hourly global solar radiation prediction models to demonstrate the feasibility of the simplified method. MAE and RMSE were applied as model evaluation indicators, and the algorithm structure with the highest prediction accuracy was selected. Then, the Shapley additive explanations (SHAP) model was used to analyze the influence of input parameters on the output value to try to further simplify the model. Figure 1 shows the flowchart for constructing the hourly global solar radiation prediction model.

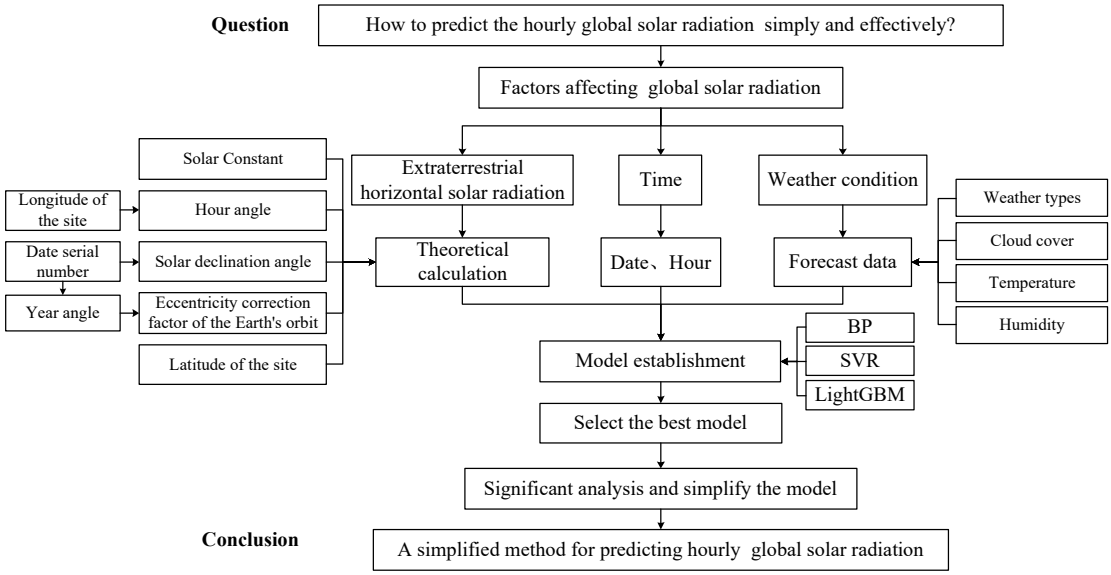

Figure 1. Flowchart of the establishment of the hourly global solar radiation prediction model.

2.1. Calculation Method of Extraterrestrial Solar Radiation

Extraterrestrial solar radiation is only affected by the relative position between the Sun and Earth. Therefore, the extra-terrestrial solar radiation value of a certain place in one year can be calculated using known parameters such as longitude, latitude, date, and hour. The calculation formula is as follows [22]:

$$I_0 = I_{SC}E_0 \times (\sin\varphi \sin\delta + \cos\varphi \cos\delta \cos\omega) \quad (1)$$

where I_{SC} is the solar constant, taken as 1367 W/m^2; φ is the geographical latitude of the site; E_0 is the eccentricity correction factor of Earth's orbit [23]; δ is the solar declination [24]; ω is the hour angle [25].

2.2. Machine Learning Algorithms

A review by Cyril et al. [26] showed that artificial neural network models such as ANN have been the most popular algorithms used in solar radiation prediction in recent years. SVM and K-means algorithms have gradually been applied in this field, whereas methods such as boosting and regression trees are rarely used in the solar radiation field. Therefore, back propagation (BP) network, SVM, and light gradient boosting machine (LightGBM) were used to establish models for estimating the hourly global solar radiation to test whether the simplified prediction method proposed in this study is feasible. These three models are briefly described below.

2.2.1. BP Network

The main structure of the BP network includes an input layer, a hidden layer, and an output layer. The output of the neurons was determined by the input value, action function, and threshold [27]. The learning process of the BP network consists of the forward propagation of the signal and backward propagation of the error [28]. The BP network has strong self-adaptation and good fault tolerance. However, its training speed is slow and easily falls into the local minimum.

Theoretically, the three-layer BP network (with one hidden layer) shown in Figure 2 can realize any nonlinear mapping with a sufficient number of neurons [29]. The number of neurons in the hidden layer (m) can be determined according to Formula (2) [30,31]. The parameters of the BP network in this study are shown in Table 1.

$$m = \sqrt{n + l} + k \tag{2}$$

where n is the number of neurons in the input layer; l is the number of neurons in the output layer; k is a constant between 1 and 10.

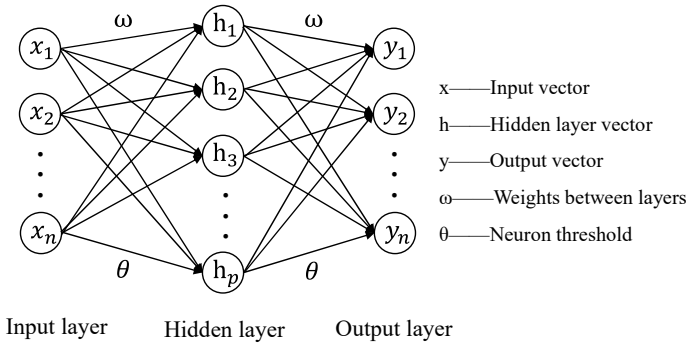

Figure 2. The BP network structure diagram.

Table 1. The parameters of the BP network.

Parameters	Value
Activation function	Relu
Optimizer	Adam
Epochs	10
Batch size	32
Loss function	Mean square error (MSE)
Hidden units	10

2.2.2. SVM

SVM is a prediction algorithm based on statistical principles. The basic principles of the SVM algorithm are structural risk minimization and the Vapnik–Chervonenkis dimension, which can process linear and nonlinear data simultaneously. The mechanism of SVM is to find an optimal classification hyperplane ($L : \omega \cdot x + b = 0$) that can not only guarantee classification accuracy, but also minimize the distance between the closest vector and the hyperplane [32]. In Figure 3, the sample points above parallel lines L1 and L2 are the support vectors. The SVM model can avoid overlearning and has a strong generalization ability and fast classification speed. It has significant advantages in solving nonlinear and high-dimensional pattern recognition problems; however, it is more sensitive to missing values.

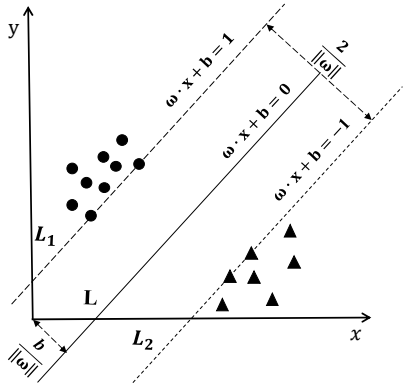

Figure 3. The optimal classification hyperplane(The circles and triangles represent the divided sample data, respectively).

2.2.3. LightGBM

LightGBM is based on the structure of the decision tree and boosting algorithm. A decision tree is a single classification algorithm, whose principle is to approximate discrete function values. The boosting algorithm is a commonly used ensemble learning algorithm that converts weak classifiers into ones through iterations [33]. LightGBM supports efficient parallel training, mainly through histogram optimization and the depth-limited leaf-wise tree growth strategy, to improve the calculation speed. As shown in Figure 4, the histogram algorithm divides the floating-point value into K ranges and constructs a histogram with width K. When traversing data, only discrete data are indexed. When searching for the optimal split point, the number of calculations can be reduced, and the calculation speed can be improved. As shown in Figure 5, a depth-limited leaf-wise tree growth strategy can find a leaf that has the greatest splitting gain and then split. LightGBM supports parallel training with high accuracy and can handle large amounts of data. Nevertheless, it is sensitive to noise and does not consider all characteristics of the data based on the optimal segmentation variable when searching for the optimal solution.

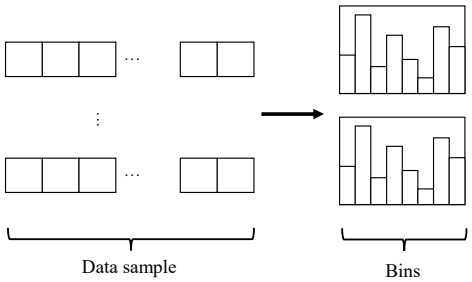

Figure 4. Histogram algorithm.

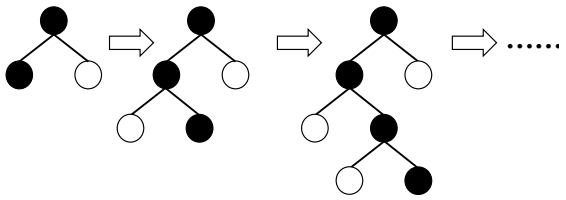

Figure 5. Leaf-wise growth strategy (Black circles represent leaf nodes with branches, and white circles represent leaf nodes without branches).

2.3. Model Evaluation Indicators

The RMSE and relative error (RE) were used to evaluate the performance of the models. RMSE is the square root of the ratio of the square of the deviation between the predicted and actual values to the total amount of data, which can measure the error between the real and predicted values. The RE is the ratio of the absolute error to real values, which reflects the reliability of the prediction. In this study, the RMSE were applied to evaluate the overall prediction results of the model, and the RE was used to evaluate the coincidence between the predicted and actual values. The smaller the RMSE and RE values, the higher the model accuracy. The calculation formulae are as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_p - y_d)^2}{n}} \tag{3}$$

$$RE = \frac{|y_p - y_d|}{y_d} \times 100\% \tag{4}$$

where y_p is the output value of the prediction model, W/m^2; y_d is the actual solar radiation value, W/m^2; n is the total value.

3. Case Study

This study took Lanzhou New District as an example, selected the actual meteorological data of the area in 2020, and used the above algorithms to establish the models to predict the hourly global solar radiation. The original data were from the National Meteorological Information Center of China (http://data.cma.cn/ (accessed on 20 January 2021)). Simultaneously, verification and comparative analyses were performed. After cleaning and pre-processing the original meteorological data according to the method described in Section 3.2, 4453 groups of valid data were obtained. After data cleaning, there were six weather types. The following section shows and analyzes the operational effects of each model. After data cleaning, hourly global solar radiation prediction models were established, and the prediction results of each model are shown.

3.1. Calculation of Extraterrestrial Solar Radiation

The geographical coordinates of Lanzhou New District are 36°44′ N and 103°15′ E. Taking 10:00 on 1 February 2020, as an example, the calculation formula introduced in Section 2.1 was used to calculate the extraterrestrial solar radiation value. According to the above information, we can calculate the eccentricity correction factor of the Earth's orbit at this time as $E_0 = 1.0301$, the solar declination angle as $\delta = 17.75^{\circ}$, and the hour angle as $\omega = -197.47^{\circ}$. Taking the solar constant, Lanzhou New District latitude, eccentricity correction factor of the Earth's orbit, solar declination angle, and hour angle into Equation (1), the extraterrestrial solar radiation at that time can be calculated as follows:

$$I_0 = 438.87\ \mathrm{W/m^2}$$

The hourly extraterrestrial solar radiation value of Lanzhou New District in 2020 was calculated, and its variation tendency is shown in Figure 6. Extraterrestrial radiation

showed a trend of first increasing and then decreasing over time, which satisfies the natural law of solar radiation changes.

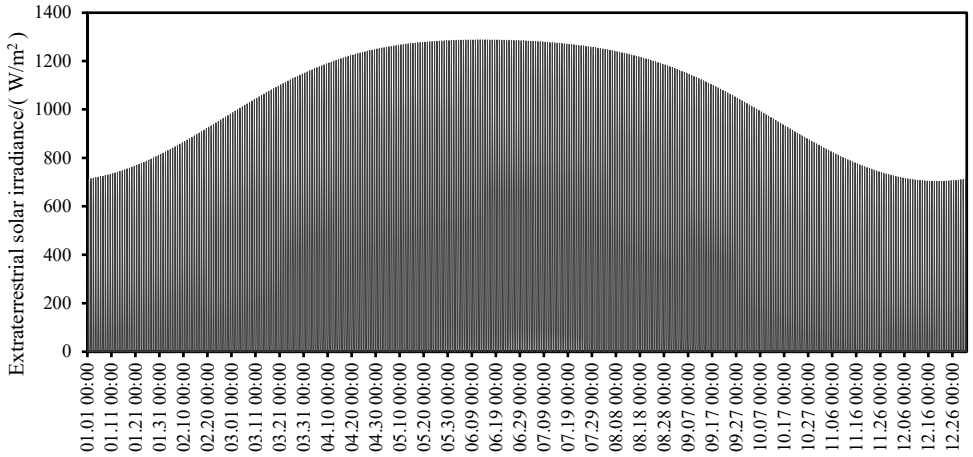

Figure 6. Hourly extraterrestrial horizontal solar radiation changes in 2020 in Lanzhou New District.

3.2. Establishment of the Database

The input parameters, range of parameters, and data preprocessing methods of the model are listed in Table 2. According to the International Commission on Illumination (CIE), radiation data with an altitude angle of less than 4° or extraterrestrial solar irradiance of less than 20 W/m^2 should be excluded [34]. This standard eliminates data that do not meet the requirements of the original database. Because there is no solar radiation at night, datasets with a historical solar radiation of 0 and extraterrestrial solar radiation of 0 were excluded.

Table 2. Statistical table of the model input parameters.

Parameter	Unit	Range of Change	Pre-Processing Method
Extraterrestrial solar radiation	W/m^2	0~1287	Normalization
Weather types	–	0~17	One-hot encoding
Cloud cover	%	0~100	Normalization
Air temperature	°C	−22~31	Normalization
Relative humidity	%	0~100	Normalization
Date	–	0~366	Normalization
Hour	h	0~23	Normalization

The normalization method used in this study and its calculating formula is:

$$x = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{5}$$

where x is the normalized data; X is the original data of the sample; $X_{\max}$ is the maximum value of the sample data; $X_{\min}$ is the minimum value of the sample data.

The weather types in the original meteorological data were recorded in text form; therefore, the weather types needed to be coded. In this study, weather types were divided into 18 types and numbered. The number for each weather type is listed in Table 3. After data preprocessing, the weather types existing in the database are numbered 0–5, and one-hot encoding is required before the model uses the weather types.

Table 3. Statistical table of the number of weather types.

Weather Type	No.
Sunny	0
Cloudy	1
Overcast	2
Light rain (rain, sleet)	3
Moderate rain	4
Light snow(snow)	5
Heavy rain	6
Rainstorm	7
Moderate snow	8
Heavy snow	9
Blizzard	10
Mist	11
Gale	12
Sand dust	13
Floating dust	14
Severe haze	15
Moderate haze	16
Mild haze	17

3.3. Establishment Testing and Evaluation of Prediction Model

A model training database after the original data pre-processing was established. We randomly selected 80% of the data for the model training set and verification; the remaining 20% of the data were used as the model test set. Hourly global solar radiation prediction models were established based on the BP network, SVM, and LightGBM. The RMSE of the prediction results for the training and testing sets are listed in Table 4.

Table 4. The RMSE of the three models.

Model	Training Set	Testing Set
BP	127.0	138.7
SVM	134.4	135.5
LightGBM	88.5	126.1

Zhang et al. [35] compared the accuracy of various hourly solar radiation prediction models and found that the RMSE of most models ranged from 88.33 to 142.22 W/m^2. The statistical results in Table 4 show that the RMSE values of the three models were within this range. The LightGBM model had the lowest RMSE and highest prediction accuracy. For the LightGBM model, the RMSE was 88.5 W/m^2 in the training set and 126.1 W/m^2 in the testing set.

Figures 7 and 8 show the performance of the three models on the training and testing sets, respectively. Figure 7 shows a comparison between the predicted and actual solar radiation values of the three models. It can be seen that the deviation between the predicted and actual values of the SVM model was the largest and its prediction accuracy was the lowest. The LightGBM model was the most accurate, and the predicted value of the model was close to the actual value. Figure 8 shows the cumulative probability distribution curves of the RE of the three models and the RE distribution of each data point in the three models. For the SVM, BP and LightGBM model, the proportions of samples with RE less than 10% in the training set were 27.6%, 14.7%, and 55.9%, and in the testing set, the proportions of samples with RE less than 10% were 32.4%, 13.5%, and 33.6%, respectively. Figure 8 also indicates that LightGBM was the most accurate and effective algorithm among the three models.

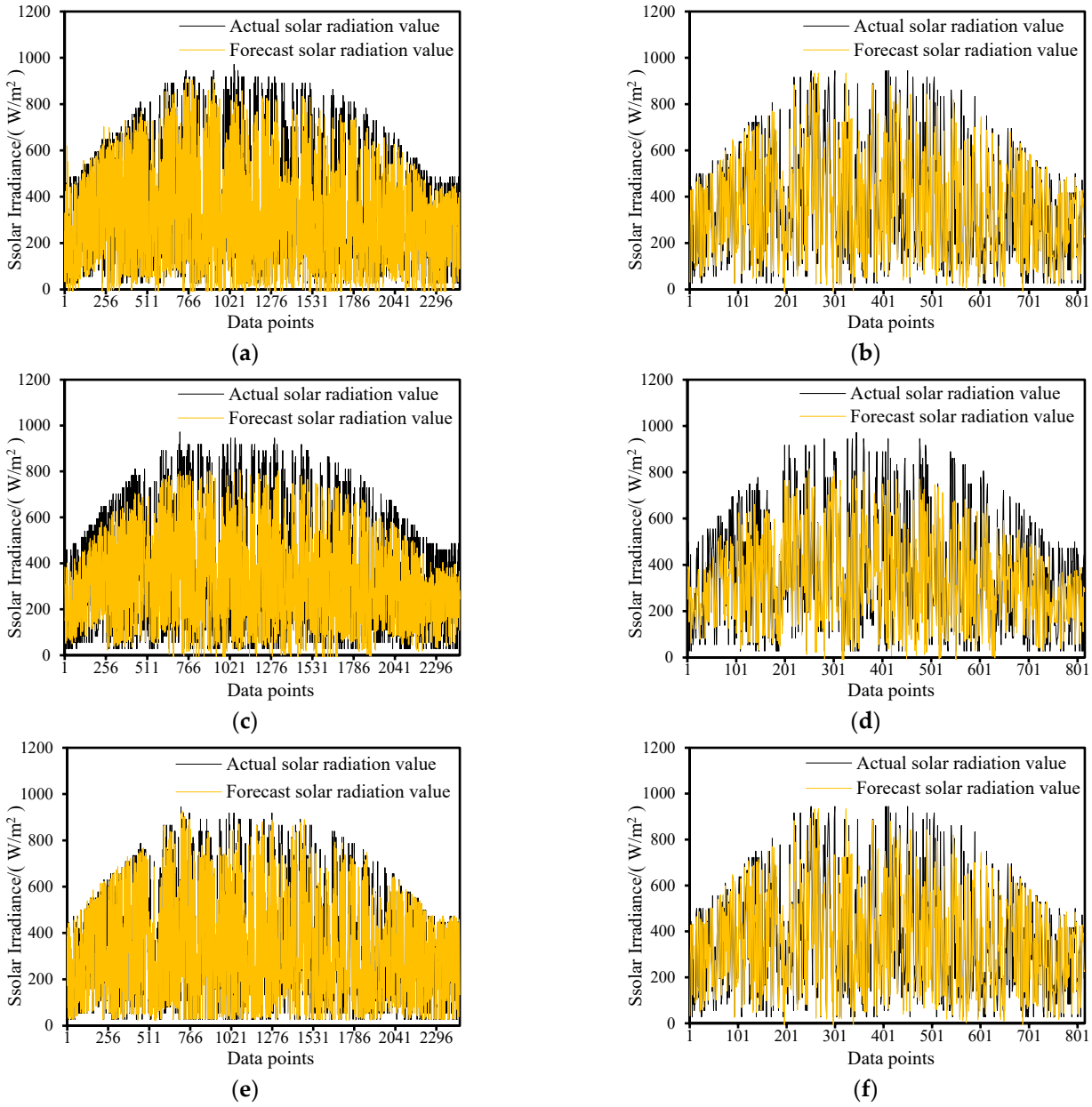

Figure 7. Prediction results of three models. (**a**) Training set of BP. (**b**) Testing set of BP. (**c**) Training set of SVR. (**d**) Testing set of SVR. (**e**) Training set of LightGBM. (**f**) Testing set of LightGBM.

To compare the performance of the hourly global solar radiation prediction method proposed in this study, Table 5 summarizes the algorithm structures, inputs, output, and prediction accuracy used in relevant research. The comparison indicates that the prediction performance of the method proposed in this study was similar to other models. Moreover, the parameters marked in red in Table 5 could not be obtained from the forecast values, while the input parameters of our study can be easily obtained and the data pre-processing is simple.

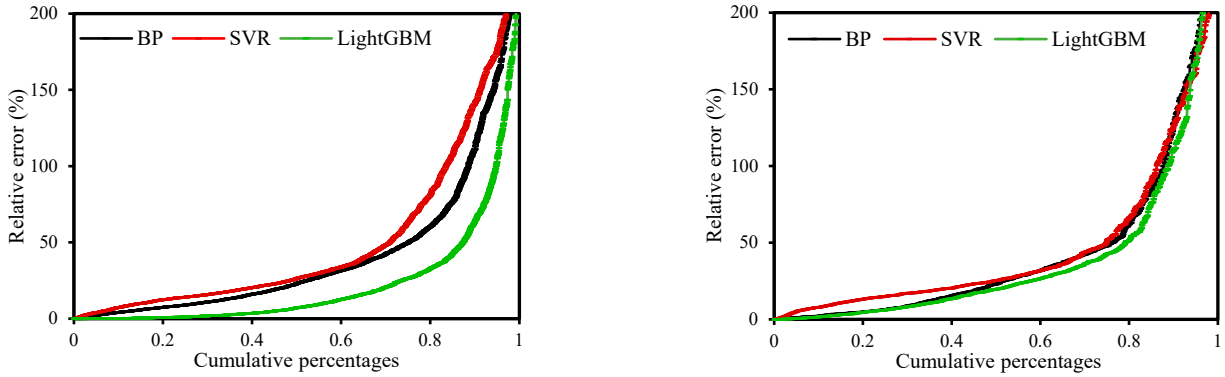

Figure 8. Probability cumulative curve of samples in three models. (**a**) Training set. (**b**) Testing set.

Table 5. Comparison of the performance of the models in this study.

Reference	Model	Inputs	Output	RMSE (W/m^2)
[7]	SVM	Historical hourly solar irradiation (for clustering), air temperature, relative humidity, and atmospheric pressure (previous day or the same day)	Hourly global solar radiation	119–163
	ANNs			145–180
[8]	DFT-PCA-Elman	Historical hourly solar irradiation data from nearby sites	24 h-ahead radiation	191.33 (Spring) 142.26 (Summer) 72.95 (Autumn) 102.61 (Winter)
[9]	seq2seq-LSTM seq2seq-at-LSTM transformer model	Historical 70 h global solar radiation	The future 10 h global solar radiation	109.24–159.41 91.96–115.41 63.54–81.28
[10]	SVM	Sunshine duration, cloud cover, cloud type, sunshine, relative humidity, precipitation, air temperature, wind speed	Hourly global solar radiation	49.26 (Sunny) 62.57 (Partially cloudy) 57.87 (Cloudy or rainy)
[11]	MARS	Horizontal extraterrestrial solar radiation, sunshine duration, visibility, cloud cover amount, wind speed	Hourly global solar radiation	76.1
[12]	Elman with similar day	Atmospheric pressure, air temperature, relative humidity, precipitation, actual sunshine duration, solar altitude angle	Hourly global solar radiation	66.67
	Elman without similar day			130.56
This study	BP	Extraterrestrial solar radiation, weather types, cloud cover, air temperature, relative humidity, and time	Hourly global solar radiation	138.7
	SVM			135.5
	LightGBM			126.1

In conclusion, the simplified method proposed in this study to predict the hourly global solar radiation is feasible. This method takes easily accessible parameters of extraterrestrial solar radiation, cloud cover, weather types, air temperature, relative humidity, and time as model inputs, and the LightGBM model had the best prediction performance among the models established in this study.

4. Analysis of Input Variables

Through the case study, it can be observed that the LightGBM model had the best prediction performance among the three models. Therefore, based on this algorithm, the feature importance of each input parameter in the prediction model was analyzed. Since

the LightGBM model is a black-box model, it is impossible to directly know the internal calculation process of the model and to intuitively display the influence of each input parameter on the model operation and prediction results. Therefore, a suitable method is required to explain it, and the SHAP model proposed by Lundberg and Lee can meet the needs of explaining the black-box model. The SHAP model is an additive explanatory model based on cooperative game theory. Its core is to calculate the SHAP values of each feature of the model and summarize the contribution of the feature to the predictive ability of the model. Traditional feature importance ranking cannot judge the relationship between the feature and the output result, whereas the SHAP model can intuitively reflect the impact of each feature on the predicted value and the positive or negative impact. Therefore, the SHAP model was applied in this study to explain the hourly global solar radiation prediction model based on the LightGBM algorithm.

4.1. Significance Analysis

Figure 9 summarizes the feature importance with a density scatter plot. The ordinate in the figure is the name of each input parameter, arranged in descending order of the SHAP average absolute value. In the figure, the weather types are encoded as w_i (i is 0~5), which are shown in Table 2. The abscissa is the SHAP value, and each point in the figure represents a sample data. The redder the color, the larger the value and the bluer the color, the smaller the value. Figure 10 shows the results of ranking the importance and the average absolute value of each feature. The ranking of the feature importance in this figure was the same as that in Figure 9.

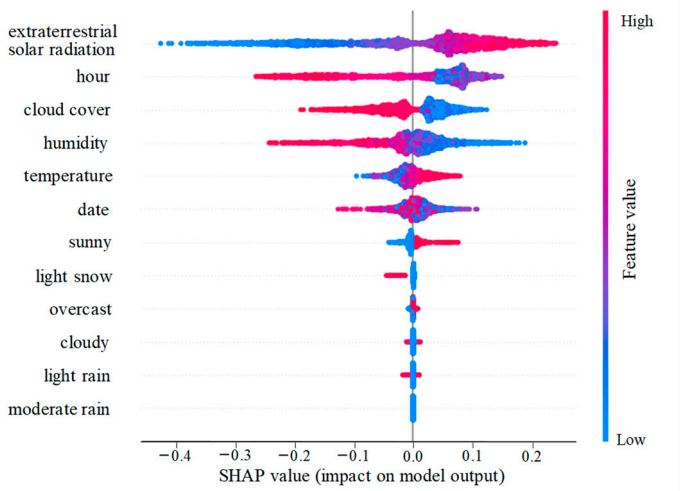

Figure 9. Density scatter plot of the SHAP analysis results.

Figures 9 and 10 show that the importance of the input parameters in the LightGBM model from high to low are as follows: extraterrestrial solar radiation, hour, relative humidity, cloud cover, date, air temperature, and weather types.

The average SHAP value of extra-terrestrial solar radiation was the largest, indicating that this parameter had the greatest impact on the output value of the model. The red sample points were distributed on the side with positive SHAP values, suggesting that this parameter positively affects the output value of the mode. Hour is the second most significant factor affecting the predicted value of the solar radiation. When the data point was red, most of the SHAP values were less than 0, and when the data point was blue, the SHAP value was greater than 0. This indicates that the solar radiation reached its maximum after noon and then gradually weakened. This phenomenon shows that the solar radiation first increased and then decreased with time.

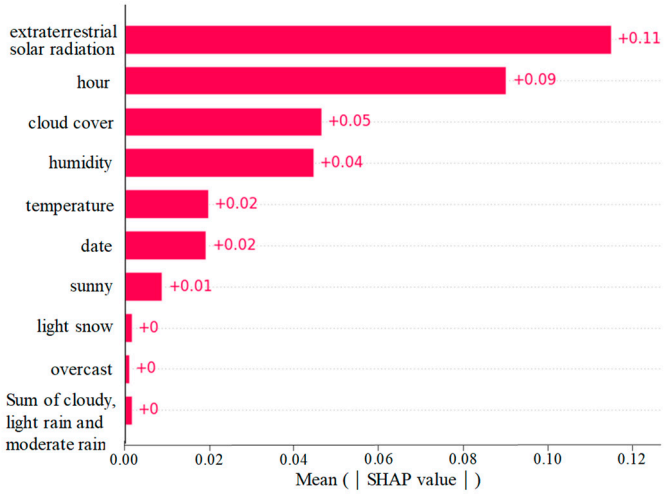

Figure 10. Sort bar graph of the significance of SHAP features.

Relative humidity and cloud cover were negatively correlated with the SHAP value, so the decrease in relative humidity and cloud cover indicates that the hourly global solar radiation is increasing. The red and blue sample points were uniformly distributed and were concentrated in the area where the SHAP value was zero, showing no obvious trend. The reason for this phenomenon may be that the actual solar radiation is highest in the middle of the year and decreases in the early and late parts of the year. The air temperature positively affected the predicted value of solar radiation, and the solar radiation increased with an increase in temperature.

Weather type had no remarkable effect on the model. Sunny and light snow were relatively significant weather types that affected the level of global solar radiation. When the weather was sunny, the SHAP value was positive, indicating that sunny weather had a positive effect on global solar radiation. When the weather was light snow, the SHAP value was negative. This indicates that the global solar radiation level decreased when snow was present. However, the SHAP values for overcast, cloudy, light rain, and moderate rain were mainly concentrated at 0, which did not significantly affect the prediction results of the LightGBM model.

4.2. Simplification of the Model

To simplify the model, we attempted to eliminate the characteristic parameters of weather types, establish and test the LightGBM model again, and compare the test results of the model before and after the simplification. Table 6 shows the running results of the LightGBM model after excluding the weather types.

Table 6. The RMSE of the LightGBM model test set with or without weather types.

LightGBM Model	RMSE (W/m^2)
Weather types included	126.1
Weather types excluded	135.2

From the above analysis, among the input parameters, weather types had the least importance to the model. Therefore, the LightGBM model was established and tested again after eliminating the weather types. Table 4 shows the running results of the LightGBM model after excluding the weather types. The RMSE of the LightGBM model was 135.2 W/m^2 after eliminating the weather types, which was 9.1 W/m^2 higher than those with weather types included. The RMSE of the LightGBM model remained unchanged after excluding weather types. Figure 11 shows the cumulative probability distribution

curves of the RE before and after excluding the weather types. It can be observed that the RE distributions of the two models were not significantly different.

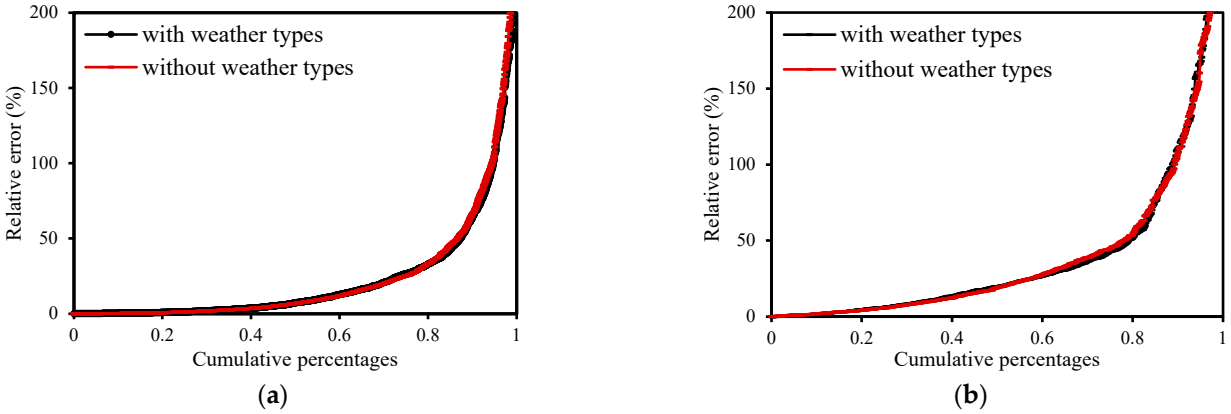

Figure 11. Probability cumulative curve of the LightGBM model with and without weather types. (**a**) Training set. (**b**) Testing set.

The above results confirm the analysis results of the SHAP model; that is, the weather types were the least significant to the LightGBM model. There was no significant change in the performance of the LightGBM with and without weather types. Therefore, when using the LightGBM model to predict the hourly global solar radiation, the input parameters can be simplified into six groups: extra-terrestrial solar radiation, cloud cover, air temperature, relative humidity date, and hour.

5. Conclusions

In summary, the methods for predicting hourly global solar radiation proposed in this study can achieve aa satisfactory performance. The RMSE of the BP network, SVR, and LightGBM were 138.7 W/m^2, 135.5 W/m^2, and 126.1 W/m^2, respectively, where it can be seen that the LightGBM model exhibited the best performance.

Based on the SHAP analysis results of the LightGBM model, weather types were not the main factors that affected the prediction result of the model. The accuracy of the model did not change significantly after excluding the weather types. Therefore, the input parameters of the LightGBM model were simplified to extraterrestrial solar radiation, cloud cover, air temperature, relative humidity date, and hour.

In conclusion, this method is applicable to engineering applications that need to predict the hourly ground solar radiation and provides a convenient and effective method for research and the engineering of building load calculation, energy consumption prediction, solar energy utilization, etc. Unfortunately, due to the limitation of data sources, the model was only validated in Lanzhou and still needs to be popularized and verified in other regions.

Author Contributions: Conceptualization, Y.J.; Methodology, J.X. and Y.S.; Software, X.Y. and S.L.; Validation, X.Y. and X.W.; Formal analysis, M.N.; Investigation, X.Y. and M.N.; Resources, Y.J.; Data curation, Y.J.; Writing—original draft, X.Y.; Writing—review & editing, X.W.; Visualization, X.Y. and S.L.; Supervision, J.X. and Y.S.; Funding acquisition, Y.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Youth Program of the National Natural Science Foundation of China (No. 51908006).

Data Availability Statement: Not Applicable.

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

References

1. IEA. Tracking Buildings 2020. Paris. Available online: https://wwwieaorg/reports/tracking-buildings-2020 (accessed on 8 June 2021).
2. Europe Environment Agency. Greenhouse Gas Emissions from Energy Use in Buildings in Europe. 26 October 2020. Available online: https://www.eea.europa.eu/ims/greenhouse-gas-emissions-from-energy (accessed on 3 December 2021).
3. U.S. Energy Information Administration. U.S. Energy-Related Carbon Dioxide Emissions, 2019. 30 September 2020. Available online: https://www.eia.gov/environment/emissions/carbon/archive/2020/ (accessed on 8 December 2021).
4. Building Energy Research Center, Tsinghua University. *2021 Annual Report on China Building Energy Efficiency*; China Architecture Publishing & Media Co., Ltd.: Beijing, China, 2021.
5. Wang, S.-Y.; Qiu, J.; Li, F.-F. Hybrid Decomposition-Reconfiguration Models for Long-Term Solar Radiation Prediction Only Using Historical Radiation Records. *Energies* **2018**, *11*, 1376. [CrossRef]
6. Bou-Rabee, M.A.; Naz, M.Y.; Albalaa, I.E.D.; Sulaiman, S.A. BiLSTM Network-Based Approach for Solar Irradiance Forecasting in Continental Climate Zones. *Energies* **2022**, *15*, 2226. [CrossRef]
7. Jiménez-Pérez, P.F.; Mora-López, L. Modeling and forecasting hourly global solar radiation using clustering and classification techniques. *Sol. Energy* **2016**, *135*, 682–691. [CrossRef]
8. Lan, H.; Zhang, C.; Hong, Y.-Y.; He, Y.; Wen, S. Day-ahead spatiotemporal solar irradiation forecasting using frequency-based hybrid principal component analysis and neural network. *Appl. Energy* **2019**, *247*, 389–402. [CrossRef]
9. Zhou, Y.; Li, Y.; Wang, D.; Liu, Y. A multi-step ahead global solar radiation prediction method using an attention-based transformer model with an interpretable mechanism. *Int. J. Hydrogen Energy* **2023**, *in press*. [CrossRef]
10. Bae, K.Y.; Jang, H.S.; Sung, D.K. Hourly Solar Irradiance Prediction Based on Support Vector Machine and Its Error Analysis. *IEEE Trans. Power Syst.* **2017**, *32*, 935–945. [CrossRef]
11. Li, D.H.W.; Chen, W.; Li, S.; Lou, S. Estimation of hourly global solar radiation using Multivariate Adaptive Regression Spline (MARS)—A case study of Hong Kong. *Energy* **2019**, *186*, 115857. [CrossRef]
12. Wang, X.; Chen, Z.; Wang, X.; Xiaochen, S.; Liu, X. Hourly total solar radiation prediction based on similar day ang Elman neural network. *J. Hainan Univ. Nat. Sci. Ed.* **2020**, *38*, 347–355.
13. Ajith, M.; Martínez-Ramón, M. Deep learning based solar radiation micro forecast by fusion of infrared cloud images and radiation data. *Appl. Energy* **2021**, *294*, 117014. [CrossRef]
14. Rodríguez-Benítez, F.J.; López-Cuesta, M.; Arbizu-Barrena, C.; Fernández-León, M.M.; Pamos-Ureña, M.Á.; Tovar-Pescador, J.; Santos-Alamillos, F.J.; Pozo-Vázquez, D. Assessment of new solar radiation nowcasting methods based on sky-camera and satellite imagery. *Appl. Energy* **2021**, *292*, 116838. [CrossRef]
15. Che, H.Z. Analysis of 40 years of solar radiation data from China, 1961–2000. *Geophys. Res. Lett.* **2005**, *32*, L06803. [CrossRef]
16. Blal, M.; Khelifi, S.; Dabou, R.; Sahouane, N.; Slimani, A.; Rouabhia, A.; Ziane, A.; Neçaibia, A.; Bouraiou, A.; Tidjar, B.; et al. A prediction models for estimating global solar radiation and evaluation meteorological effect on solar radiation potential under several weather conditions at the surface of Adrar environment. *Measurement* **2020**, *152*, 107348. [CrossRef]
17. Alam, M.S.; Al-Ismail, F.S.; Hossain, M.S.; Rahman, S.M. Ensemble Machine-Learning Models for Accurate Prediction of Solar Irradiation in Bangladesh. *Processes* **2023**, *11*, 908. [CrossRef]
18. He, P.; Cui, Y.; Li, J.; Liu, S. Analysis of solar radiation and relative factors in southeast coastal cities of China. *Prog. Geogr.* **2019**, *38*, 1793–1801. [CrossRef]
19. Sansa, I.; Boussaada, Z.; Mazigh, M.; Bellaaj, N.M. Solar radiation prediction for a winter day using ARMA model. In Proceedings of the 2020 6th IEEE International Energy Conference (ENERGYCon), Gammarth, Tunisia, 13–16 April 2020; pp. 326–330.
20. Fraihat, H.; Almbaideen, A.A.; Al-Odienat, A.; Al-Naami, B.; De Fazio, R.; Visconti, P. Solar Radiation Forecasting by Pearson Correlation Using LSTM Neural Network and ANFIS Method: Application in the West-Central Jordan. *Future Internet* **2022**, *14*, 79. [CrossRef]
21. Zang, H.; Cheng, L.; Liu, L.; Wei, Z.; Sun, G. Research and Prospect for Data-driven Estimation and Prediction of Solar Radiation. *Autom. Electr. Power Syst.* **2021**, *45*, 170–183.
22. Khatib, T.; Elmenreich, W. A Model for Hourly Solar Radiation Data Generation from Daily Solar Radiation Data Using a Generalized Regression Artificial Neural Network. *Int. J. Photoenergy* **2015**, *2015*, 968024. [CrossRef]
23. Cao, Q.; Liu, Y.; Sun, X.; Yang, L. Country-level evaluation of solar radiation data sets using ground measurements in China. *Energy* **2022**, *241*, 122938. [CrossRef]
24. Xiao, M.; Yu, Z.; Cui, Y. Evaluation and estimation of daily global solar radiation from the estimated direct and diffuse solar radiation. *Theor. Appl. Climatol.* **2020**, *140*, 983–992. [CrossRef]
25. Yıldırım, H.B.; Çelik, Ö.; Teke, A.; Barutçu, B. Estimating daily Global solar radiation with graphical user interface in Eastern Mediterranean region of Turkey. *Renew. Sustain. Energy Rev.* **2018**, *82*, 1528–1537. [CrossRef]
26. Voyant, C.; Notton, G.; Kalogirou, S.; Nivet, M.-L.; Paoli, C.; Motte, F.; Fouilloy, A. Machine learning methods for solar radiation forecasting: A review. *Renew. Energy* **2017**, *105*, 569–582. [CrossRef]
27. Wen, L. Cloud Computing Intrusion Detection Technology Based on BP-NN. *Wirel. Pers. Commun.* **2021**, *126*, 1917–1934. [CrossRef]
28. Sun, J.; Chen, M.; Kong, L.; Hu, Z.; Veerasamy, V. Regional Load Frequency Control of BP-PI Wind Power Generation Based on Particle Swarm Optimization. *Energies* **2023**, *16*, 2015. [CrossRef]

29. Chen, Z.; Liu, Q.; Ou, Y. Cross-talk Resistant adaptive Neural Network algorithm for Speech Enhancement. *Chin. J. Sci. Instrum.* **2008**, *29*, 623–626.
30. Wang, Z.; Wang, F.; Su, S. Solar Irradiance Short-Term Prediction Model Based on BP Neural Network. *Energy Procedia* **2011**, *12*, 488–494. [CrossRef]
31. Feng, R.; Wencheng, L. LSSA-BP-based cost forecasting for onshore wind power. *Energy Rep.* **2023**, *9*, 362–370. [CrossRef]
32. Chauhan, V.K.; Dahiya, K.; Sharma, A. Problem formulations and solvers in linear SVM: A review. *Artif. Intell. Rev.* **2018**, *52*, 803–855. [CrossRef]
33. Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **2020**, *54*, 1937–1967. [CrossRef]
34. CIE. *Guide to Recommended Practice of Daylight Measurment*; Central Bureau of the CIE: Vienna, Austria, 1994.
35. Zhang, J.; Zhao, L.; Deng, S.; Xu, W.; Zhang, Y. A critical review of the models used to estimate solar radiation. *Renew. Sustain. Energy Rev.* **2017**, *70*, 314–329. [CrossRef]