



## (51) International Patent Classification:

G06Q 10/04 (2012.01) G06K 9/62 (2022.01)

G06Q 50/06 (2012.01) G06N 3/04 (2006.01)

## (21) International Application Number:

PCT/CN2022/087787

## (22) International Filing Date:

20 April 2022 (20.04.2022)

## (25) Filing Language:

English

## (26) Publication Language:

English

(71) Applicants: SOUTHEAST UNIVERSITY [CN/CN]; No 2. Sipailou, Xinjiekou Street, Xuanwu District, Nanjing, Jiangsu 210018 (CN). STATE GRID SUZHOU POWER SUPPLY COMPANY [CN/CN]; No.555, Laodong Road, Gusu District, Suzhou, Jiangsu 215004 (CN). STATE GRID ELECTRIC POWER RESEARCH INSTITUTE

[CN/CN]; No.19 Chengxin Road, Jiangning District, Nanjing, Jiangsu 211106 (CN). CHINA ELECTRIC POWER RESEARCH INSTITUTE [CN/CN]; No 15. Qinghexi-aoying East Road, Haidian District, Beijing 100192 (CN). STATE GRID JIANGSU ELECTRIC POWER CO., LTD [CN/CN]; No.215, Shanghai Road, Gulou District, Nanjing, Qinghai 210024 (CN).

(72) Inventors: ZHAO, Jiaqing; No.555, Laodong Road, Gusu District, Suzhou, Jiangsu 215004 (CN). CHEN, Zhong; No 2. Sipailou, Xinjiekou Street, Xuanwu District, Nanjing, Jiangsu 210018 (CN). ZHAO, Qi; No.555, Laodong Road, Gusu District, Suzhou, Jiangsu 215004 (CN). WANG, Biao; No 2. Sipailou, Xinjiekou Street, Xuanwu District, Nanjing, Jiangsu 210018 (CN). LV, Yang; No.555, Laodong Road, Gusu District, Suzhou, Jiangsu 215004 (CN). XU, Chunlei; No.215, Shanghai Road, Gulou District, Nanjing, Jiangsu 210024 (CN). ZHUANG, Weijin; No 15. Qinghexi-

(54) Title: COUNTY-WIDE PHOTOVOLTAIC PREDICTION METHOD BASED ON CLUSTER DIVISION AND DATA ENHANCEMENT

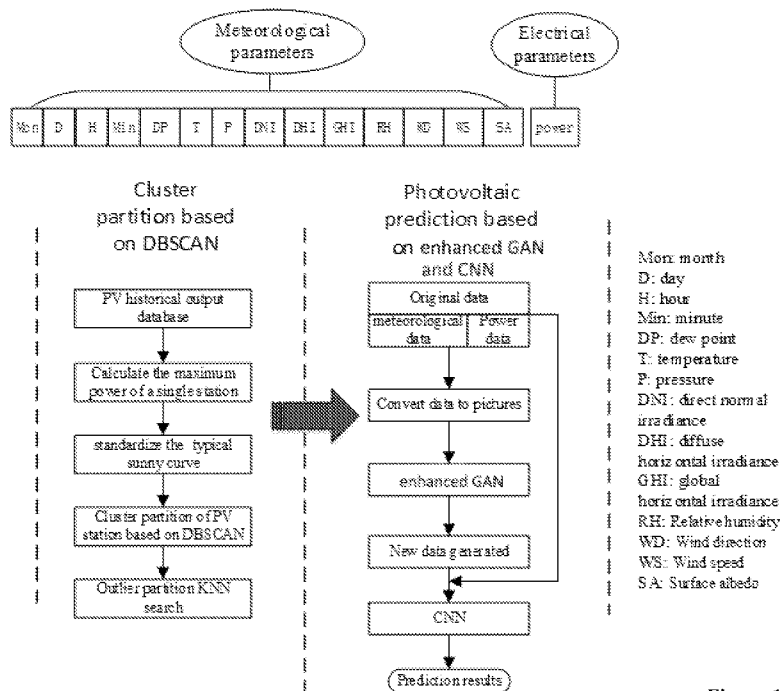


Figure 1

(57) Abstract: The present invention discloses a county-wide distributed photovoltaic prediction method based on cluster division and data enhancement. The prediction method is as follows: selecting a typical power curve in a sunny day from the photovoltaic (PV) output history database of the entire county, and standardizing the output with the maximum power of the station. Pearson correlation coefficient is used as the distance measure, and the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is used to cluster photovoltaic stations to form a cluster division. The outliers are divided into nearest clusters using k-nearest neighbor search method. An enhanced Generative Adversarial Networks (GAN) is used to augment the historical data; the original data and the generated data are jointly used to train the deep convolutional network prediction model. The present invention learns the original data distribution through the training process of the enhanced GAN and then generates correspondingly distributed data to enhance the entire county

iaoying East Road, Haidian District, Beijing 100192 (CN). **SHAN, Xin**; No.19 Chengxin Road, Jiangning District, Nanjing, Jiangsu 211106 (CN). **WANG, Yi**; No.19 Chengxin Road, Jiangning District, Nanjing, Jiangsu 211106 (CN). **WU, Haiwei**; No.215, Shanghai Road, Gulou District, Nanjing, Jiangsu 210024 (CN). **DING, Hongen**; No.555, Laodong Road, Gusu District, Suzhou, Jiangsu 215004 (CN). **TIAN, Jiang**; No.555, Laodong Road, Gusu District, Suzhou, Jiangsu 215004 (CN). **YU, Yu**; No.555, Laodong Road, Gusu District, Suzhou, Jiangsu 215004 (CN). **XU, Xiuzhi**; No.555, Laodong Road, Gusu District, Suzhou, Jiangsu 215004 (CN). **MENG, Yuting**; No.555, Laodong Road, Gusu District, Suzhou, Jiangsu 215004 (CN).

(74) **Agent: BEIJING NEWIPS INTELLECTUAL PROPERTY LTD.**; Room 427, 4th Floor, Building 6, No.1 Shangdi Tenth Street, Haidian District, Beijing 100085 (CN).

(81) **Designated States** (*unless otherwise indicated, for every kind of national protection available*): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) **Designated States** (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

**Published:**

— *with international search report (Art. 21(3))*

photovoltaic historical database. Finally the invention trains the deep convolutional neural network through the enhanced training set to improve the accuracy of prediction.

## COUNTY-WIDE PHOTOVOLTAIC PREDICTION METHOD BASED ON CLUSTER DIVISION AND DATA ENHANCEMENT

### FIELD OF THE INVENTION

The invention relates to the field of photovoltaic power prediction, in particular to a county-wide photovoltaic prediction method based on cluster division and data enhancement.

### BACKGROUND

Climate change is a global problem faced by mankind. With the rapid development of productivity, the amount of carbon dioxide emitted by countries has increased sharply, and the greenhouse effect has become more and more serious. Many countries have also proposed to reduce carbon emission. Photovoltaic power generation does not require the burning of fossil fuels for energy conversion, and it is a green and clean energy source. Centralized photovoltaics need to take up a lot of land resources, mostly concentrated in the sparsely populated region, power resources need to be transmitted to the load center through UHV transmission lines. The distributed photovoltaic on the roof does not occupy special land resources, and at the same time, it is close to the load center and has great potential for power generation. As of the end of May 2021, low-voltage distributed photovoltaic grid-connected 1,884,700 households and a capacity of 45,317,700 kilowatts have been put into operation in the operation area of State Grid Corporation of China. On September 14, 2021, the Chinese National Energy Administration formally issued the "Announcement on the List of the Pilot List of the entire County (City, District) Roof Distributed Photovoltaic Development" to actively promote the integration of photovoltaic users across the county. The penetration rate of photovoltaics continues to increase, and the randomness and volatility of its output put pressure on the safe and stable operation of the power grid. Therefore, the accuracy of the entire county's photovoltaic power generation forecast is becoming more and more important.

Distributed photovoltaics in the whole county currently have two main problems. On one hand, distributed photovoltaics started late compared with centralized photovoltaics, and historical weather data and power generation data are relatively scarce, making it difficult to

train high-precision data-driven models. On the other hand, with a large number of photovoltaics and a wide geographical distribution, it is difficult to predict one station at a time similar to a centralized site. Therefore, it is necessary to solve the problem of model training with poor data, and at the same time, it is necessary to deal with the problem of county-wide photovoltaic power prediction for massive sites.

## SUMMARY OF THE INVENTION

The purpose of the invention is to provide a county-wide photovoltaic prediction method based on cluster division and data enhancement. Taking Pearson correlation distance as the measure, the typical output in sunny day of each station is clustered by DBSCAN, and the stations with high similarity of output curve are divided into a cluster, so as to reduce the allocation of numerical weather forecast, reduce the calculation pressure and improve the cluster prediction accuracy. Through the confrontation training of the enhanced GAN, the characteristics of the original data are learned, so as to generate new data with similar distribution, supplement training samples. Then, the augmented data is used to train the deep convolution network. It can improve the prediction accuracy of the model.

The object of the invention can be realized by the following technical scheme:

A county-wide photovoltaic prediction method based on cluster division and data enhancement. The prediction method comprises the following steps:

S1: collect historical PV output and meteorological data of each PV station in sunny days for the set area and set time frame;

S2: calculate and standardize the maximum power of each PV station based on the historical PV output data in sunny days collected from S1, and then perform cluster division;

S3: construct a PV data augmentation neural network;

S4: augment the PV data using the PV data augmentation neural network constructed in S3, and input the augmented data into a CNN neural network to obtain the PV prediction neural network;

S5: input meteorological data into the PV prediction neural network obtained from S4 for PV prediction.

in the step 1, the collected PV output data is county-wide historical PV output data in sunny

days;

the meteorological data include month, day, hour, minute, direct normal irradiance, diffuse horizontal irradiance, global horizontal irradiance, temperature, pressure, relative humidity, wind direction, wind speed, surface albedo and power generation.

in the step 2, the typical historical sunny photovoltaic output curve is a standardized output curve obtained by first calculating the maximum power of each station based on the historical output data, and then calculating the unit value of the curve;

the formula for data standardization is as follows:

$$z = \frac{x}{x_{\max}}$$

where,  $x$  is the original sample data,  $x_{\max}$  is the maximum historical output value of PV station, and  $z$  is the standardized data.

in the S2, use DBSCAN to perform cluster division, and the distance measurement of DBSCAN adopts Pearson correlation distance; using KNN algorithm to search for the nearest  $k$  sites, counting the cluster numbers of  $k$  sites, and categorizing the abnormal sites into the cluster with the largest cluster number.

in the step 3, the chosen neural network for data augmentation is Generative Adversarial Networks.

the chosen neural network for data augmentation is an enhanced Generative Adversarial Networks with the following algorithm:

S3.1, initialize the learning rate  $\alpha$ , the clipping variable  $c$ , the batch training sample number  $m$ , iteration number threshold  $n_{\text{critic}}$ , the discriminator's network  $w_t$  and the generator's variable  $\theta_t$ ,

S3.2, check if the generator's variable  $\theta_t$  has converged, if converged, then stop the iteration; if not, then go in to S3.3;

S3.3, check if the current iteration number has reached the iteration number threshold  $n_{\text{critic}}$ , if so, go into S3.4; if not, then update the discriminator's network variable  $w_t$  and the current iteration number, and repeat S3.3;

S3.4, calculate the loss function of the generator  $g_{\theta_t}$  and update the generator's variable

$\theta_t$ , then go back to S3.2.

method to update the discriminator's network variable  $w_t$  is:

$$w_{t+1} = \text{clip}(w_t + \alpha \text{RMSProp}(w_t, g_{w_t}), -c, c)$$

where,  $\text{clip}(\cdot)$  is the clipping function in deep learning,  $\text{RMSProp}(\cdot)$  is the unpublished adaptive learning rate optimizer in deep learning,  $g_{w_t}$  is the gradient function of the loss function of the discriminator, that satisfies:

$$g_w = \nabla_{w_t} \left[ \frac{1}{m} \sum_{i=1}^m f_{w_t}(x^{(i)}) - \frac{1}{m} \sum_{i=1}^m f_{w_t}(g_{\theta_t}(z^{(i)})) \right]$$

where,  $x^{(i)}$  is the i-th batch of the input,  $z^{(i)}$  is the batch sample of the output distribution generated by the generator from the i-th training;  $f_{w_t}(\cdot)$  is a generator network.

the gradient of the generator's loss function is

$$g_{\theta_{t+1}} = -\nabla_{\theta_t} \frac{1}{m} \sum_{i=1}^m f_{w_t}(g_{\theta_t}(z^{(i)}))$$

where,  $f_{w_t}(\cdot)$  is a generator network,  $z^{(i)}$  is the batch sample of the output distribution generated by the generator from the i-th training.

the method to update the generator's variable  $\theta_t$  is:

$$\theta_{t+1} = \theta_t - \alpha \text{RMSProp}(\theta_t, g_{\theta_t})$$

where,  $\text{RMSProp}(\cdot)$  is the unpublished adaptive learning rate optimizer in deep learning.

in the S4, data augmentation means to convert the data collected from S1 into images by dividing the collected data into 48 points per day to form an image with the size of 48\*15 pixels; then use the enhanced GAN for adversarial learning to generate new data with the same distribution.

in the S4, the CNN neural network has a 5-layer structure; each of the first four layers includes a convolution layer, a batch normalization layer and a ReLU layer, and the last layer includes a full connection layer, a batch normalization layer and a ReLU layer.

Beneficial effects of the invention:

1. The prediction method of the invention divides the highly correlated distributed stations in the whole county into clusters, and predicts the overall output of the photovoltaic cluster, so

as to reduce the allocation of numerical weather forecast stations, effectively reduce the calculation pressure and improve the overall prediction accuracy.

2. The prediction method of the invention learns the relationship and distribution characteristics of the original data through the confrontation training of the enhanced, generates new similar data, supplements the poor training data set and improves the prediction accuracy of the model.

## **BRIEF DESCRIPTION OF THE DRAWINGS**

Figure 1 is a flow chart of the invention;

Figure 2 is a histogram of DBSCAN clustering of the invention;

Figure 3 is a sequence diagram of the DBSCAN clustering results of the invention;

Figure 4 is the result of KNN search of the invention;

Figure 5 is the visualization of the original data;

Figure 6 is the generated data and training process of the enhanced GAN;

Figure 7 is a comparison diagram of CNN and the enhanced GAN-CNN prediction time series.

## **DETAILED DESCRIPTION OF EXEMPLARY EMBODIMENTS**

The technical scheme in the embodiment of the invention will be clearly and completely described below in combination with the accompanying drawings in the embodiment of the invention. Obviously, the described embodiments are only parts of the embodiments of the invention, not all of the embodiments of the invention. Based on the embodiments of the invention, all other embodiments obtained by those skilled in the art without creative work belong to the protection scope of the invention.

A county-wide photovoltaic prediction method based on cluster division and data enhancement is shown in Figure 1. The prediction method includes the following steps:

S1: collect historical PV output and meteorological data of each PV station in sunny days for the set area and set time frame;

In the present invention, the collected PV output data is county-wide historical PV output data in sunny days;

Meteorological data include month, day, hour, minute, direct normal irradiance, diffuse horizontal irradiance, global horizontal irradiance, temperature, pressure, relative humidity, wind direction, wind speed, surface albedo and power generation.

S2: calculate and standardize the maximum power of each PV station based on the historical PV output data in sunny days collected from S1, and then perform cluster division;

Further, the typical historical sunny photovoltaic output curve in S2 is a standardized output curve. First, calculate the maximum power of each station based on the historical output data, and then calculate the unit value of the curve;

The formula for data standardization is as follows:

$$z = \frac{x}{x_{\max}}$$

In the above formula,  $x$  is the original sample data,  $x_{\max}$  is the maximum historical output value of PV station, and  $z$  is the standardized data.

According to the PV station PV output curve of sunny days, DBSCAN is used to divide the PV stations of the whole county to form highly correlated PV clusters.

Those skilled in the art can choose any methods to perform cluster division depending on the situations. The proposed method is only a preferred example, and should not be used to limit the protection scope of the present invention.

Further, DBSCAN is used for cluster division, and the distance measurement of DBSCAN adopts Pearson correlation distance;

The formula of Pearson correlation distance is as follows.

$$D = 1 - \frac{(x_s - \bar{x}_s)(x_t - \bar{x}_t)'}{\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'} \sqrt{(x_t - \bar{x}_t)(x_t - \bar{x}_t)'}}$$

where

$$\bar{x}_s = \frac{1}{n} \sum_j x_{sj}, j = 1, 2, \dots, n, \quad \bar{x}_t = \frac{1}{m} \sum_i x_{ti}, i = 1, 2, \dots, m$$

$\mathcal{X}_s$ ,  $\mathcal{X}_t$  are the output data sequences of any two random PV stations,  $\mathcal{X}_{sj}$  is the j-th data point in  $\mathcal{X}_s$ ,  $\mathcal{X}_{ti}$  is the i-th data point in  $\mathcal{X}_t$ , n is the total number of data points in  $\mathcal{X}_s$ , m is the total number of data points in  $\mathcal{X}_t$ .

Further, for the abnormal sites generated by cluster analysis of DBSCAN, taking Pearson correlation distance as the measurement, using KNN algorithm to search for the nearest k sites, counting the cluster numbers of k sites, and categorizing the abnormal sites into the cluster with the largest cluster number.

S3: construct a PV data augmentation neural network;

Those skilled in the art may choose any neural networks to augment the data. The proposed method is only a preferred example, and should not be used to limit the protection scope of the present invention.

Preferably, the chosen neural network for data augmentation can be Generative Adversarial Networks (GAN);

In the present embodiment, the chosen neural network for data augmentation is an enhanced GAN with the following algorithm:

S3.1, initialize the learning rate  $\alpha$ , the clipping variable c, the batch training sample number m, iteration number threshold  $n_{critic}$ , the discriminator's network  $w_t$  and the generator's variable  $\theta_t$ ,

S3.2, check if the generator's variable  $\theta_t$  has converged, if converged, then stop the iteration; if not, then go in to S3.3;

S3.3, check if the current iteration number has reached the iteration number threshold  $n_{critic}$ , if so, go into S3.4; if not, then update the discriminator's network variable  $w_t$  and the current iteration number, and repeat S3.3;

Those skilled in the art can choose any methods to update the discriminator's network variable  $w_t$ . The proposed method is only a preferred example, and should not be used to limit the protection scope of the present invention.

Method to update the discriminator's network variable  $w_t$  is:

$$w_{t+1} = \text{clip}(w_t + \alpha \text{RMSProp}(w_t, g_{w_t}), -c, c)$$

where,  $\text{clip}(\cdot)$  is the clipping function in deep learning,  $\text{RMSProp}(\cdot)$  is the unpublished adaptive learning rate optimizer in deep learning,  $g_{w_t}$  is the gradient function of the loss function of the discriminator, that satisfies:

$$g_w = \nabla_{w_t} [\frac{1}{m} \sum_{i=1}^m f_{w_t}(x^{(i)}) - \frac{1}{m} \sum_{i=1}^m f_{w_t}(g_{\theta_t}(z^{(i)}))]$$

where,  $x^{(i)}$  is the i-th batch of the input,  $z^{(i)}$  is the batch sample of the output distribution generated by the generator from the i-th training;  $f_{w_t}(\cdot)$  is a generator network, those skilled in the art can choose it according to the situation.

S3.4, calculate the loss function of the generator  $g_{\theta_t}$  and update the generator's variable  $\theta_t$ , then go back to S3.2;

Those skilled in the art can choose any methods to set the loss function of the generator  $g_{\theta_t}$  and update the generator's variable  $\theta_t$ . The proposed method is only a preferred example, and should not be used to limit the protection scope of the present invention.

The gradient of the generator's loss function is

$$g_{\theta_{t+1}} = -\nabla_{\theta_t} \frac{1}{m} \sum_{i=1}^m f_{w_t}(g_{\theta_t}(z^{(i)}))$$

The method to update the generator's variable  $\theta_t$  is:

$$\theta_{t+1} = \theta_t - \alpha \text{RMSProp}(\theta_t, g_{\theta_t})$$

S4: Augment the PV data using the PV data augmentation neural network constructed in S3, and input the augmented data into a CNN neural network to obtain the PV prediction neural network;

Further, convert the data collected in S1 into images by dividing the collected data into 48 points per day to form an image with the size of 48\*15 pixels. Then, the enhanced GAN is used for adversarial learning to generate new data with the same distribution.

Further, the CNN neural network has a 5-layer structure. Each of the first four layers includes a convolution layer, a batch normalization layer and a ReLU layer, and the last layer includes a full connection layer, a batch normalization layer and a ReLU layer. The collected

and augmented data are used to train the deep convolution network to obtain the final prediction model.

S5: input meteorological data into the PV prediction neural network obtained from S4 for PV prediction. Preferably, the PV predication time frame is 1-3 days.

An example of the photovoltaic power prediction method of the present invention is as follows:

Based on the 2014 household photovoltaic data released by the Australian power grid, the historical photovoltaic data of 105 households are selected. The maximum photovoltaic output of each household in a year is counted, the data is standardized according to the formula of claim 2, and the typical sunny output sequence is selected. The Pearson correlation distance is used as the distance measure for clustering. Through parameter adjustment, the value of some key parameters of DBSCAN are finally determined. The domain radius is 0.04 and the minimum number of points is 4. The clustering results are shown in Figure 2. As can be seen from the histogram, 105 distributed sites are divided into two clusters, of which 9 sites are determined as outliers. The time sequence diagram of clustering is shown in Figure 3. From the time sequence diagram, the output curve can be roughly divided into two bands with high overlap, and the clustering results are basically consistent with the senses. KNN is used to search the nearest neighbor sites for 9 abnormal sites, k is 5, and the abnormal sites are classified into the cluster with the largest number. The final cluster division result is shown in Figure 4.

The output curve and the corresponding meteorological data are divided by day to form a data matrix with 15 characteristics and 48 time points, which is further imaged, as shown in Figure 5. New data is generated through the enhanced GAN's adversarial learning. As shown in Figure 6, it can be seen that the newly generated data has similar data characteristics to the original data. It can be seen from the training process that the enhanced GAN has stabilized.

The original data and the generated data are used to form the training set, which is used to train the deep convolutional network. It is compared with the deep convolutional network that uses the same structure and parameters and only trained by the original data. The predictive indicators on the test set are shown in Table 1. The prediction result sequence diagram is shown in Figure 7, which shows that the method can effectively improve the prediction accuracy.

Tab.1 The comparison of the two methods

| Error index | The present<br>method | CNN    |
|-------------|-----------------------|--------|
| RMSE        | 0.1368                | 0.1639 |

In the description of this specification, the description referring to the terms "one embodiment", "example", "specific example" and the like means that the specific features, structures, materials or features described in connection with the embodiment or example are included in at least one embodiment or example of the present invention. In this specification, the schematic expression of the above terms does not necessarily refer to the same embodiment or example. Further, the specific features, structures, materials or features described may be combined in a suitable manner in any one or more embodiments or examples.

The basic principles, main features and advantages of the invention are shown and described above. Those skilled in the industry should understand that the invention is not limited by the above embodiments. What is described in the above embodiments and specifications is only to explain the principle of the invention. On the premise of not departing from the spirit and scope of the invention, the invention will have various changes and improvements, which fall within the scope of the claimed invention.

## Claims

1. A county-wide photovoltaic prediction method based on cluster division and data enhancement is characterized in that the prediction method comprises the following steps:

S1: collect historical PV output and meteorological data of each PV station in sunny days for the set area and set time frame;

S2: calculate and standardize the maximum power of each PV station based on the historical PV output data in sunny days collected from S1, and then perform cluster division;

S3: construct a PV data augmentation neural network;

S4: augment the PV data using the PV data augmentation neural network constructed in S3, and input the augmented data into a CNN neural network to obtain the PV prediction neural network;

S5: input meteorological data into the PV prediction neural network obtained from S4 for PV prediction.

2. A county-wide photovoltaic prediction method based on cluster division and data enhancement according to claim 1, wherein,

in the step 1, the collected PV output data is county-wide historical PV output data in sunny days;

the meteorological data include month, day, hour, minute, direct normal irradiance, diffuse horizontal irradiance, global horizontal irradiance, temperature, pressure, relative humidity, wind direction, wind speed, surface albedo and power generation.

3. A county-wide photovoltaic prediction method based on cluster division and data enhancement according to claim 1 or 2, wherein,

in the step 2, the typical historical sunny photovoltaic output curve is a standardized output curve obtained by first calculating the maximum power of each station based on the historical output data, and then calculating the unit value of the curve;

the formula for data standardization is as follows:

$$z = \frac{x}{x_{\max}}$$

where,  $x$  is the original sample data,  $x_{\max}$  is the maximum historical output value of PV station, and  $z$  is the standardized data.

4. A county-wide photovoltaic prediction method based on cluster division and data enhancement according to claim 1 or 3, wherein,

in the S2, use DBSCAN to perform cluster division, and the distance measurement of DBSCAN adopts Pearson correlation distance; using KNN algorithm to search for the nearest  $k$  sites, counting the cluster numbers of  $k$  sites, and categorizing the abnormal sites into the cluster with the largest cluster number.

5. A county-wide photovoltaic prediction method based on cluster division and data enhancement according to claim 1, wherein,

in the step 3, the chosen neural network for data augmentation is Generative Adversarial Networks.

6. A county-wide photovoltaic prediction method based on cluster division and data enhancement according to claim 1 or 5, wherein,

The chosen neural network for data augmentation is an enhanced Generative Adversarial Networks with the following algorithm:

S3.1, initialize the learning rate  $\alpha$ , the clipping variable  $c$ , the batch training sample number  $m$ , iteration number threshold  $n_{\text{critic}}$ , the discriminator's network  $w_t$  and the generator's variable  $\theta_t$ ,

S3.2, check if the generator's variable  $\theta_t$  has converged, if converged, then stop the iteration; if not, then go in to S3.3;

S3.3, check if the current iteration number has reached the iteration number threshold  $n_{\text{critic}}$ , if so, go into S3.4; if not, then update the discriminator's network variable  $w_t$  and the current iteration number, and repeat S3.3;

S3.4, calculate the loss function of the generator  $g_{\theta_t}$  and update the generator's variable  $\theta_t$ , then go back to S3.2.

7. A county-wide photovoltaic prediction method based on cluster division and data enhancement according to claim 6, wherein,

method to update the discriminator's network variable  $w_t$  is:

$$w_{t+1} = \text{clip}(w_t + \alpha \text{RMSProp}(w_t, g_{w_t}), -c, c)$$

where,  $\text{clip}(\cdot)$  is the clipping function in deep learning,  $\text{RMSProp}(\cdot)$  is the unpublished adaptive learning rate optimizer in deep learning,  $g_{w_t}$  is the gradient function of the loss function of the discriminator, that satisfies:

$$g_w = \nabla_{w_t} [\frac{1}{m} \sum_{i=1}^m f_{w_t}(x^{(i)}) - \frac{1}{m} \sum_{i=1}^m f_{w_t}(g_{\theta_t}(z^{(i)}))]$$

where,  $x^{(i)}$  is the i-th batch of the input,  $z^{(i)}$  is the batch sample of the output distribution generated by the generator from the i-th training;  $f_{w_t}(\cdot)$  is a generator network.

8. A county-wide photovoltaic prediction method based on cluster division and data enhancement according to claim 6 or 7, wherein,

The gradient of the generator's loss function is

$$g_{\theta_{t+1}} = -\nabla_{\theta_t} \frac{1}{m} \sum_{i=1}^m f_{w_t}(g_{\theta_t}(z^{(i)}))$$

where,  $f_{w_t}(\cdot)$  is a generator network,  $z^{(i)}$  is the batch sample of the output distribution generated by the generator from the i-th training.

9. A county-wide photovoltaic prediction method based on cluster division and data enhancement according to claim 8, wherein,

the method to update the generator's variable  $\theta_t$  is:

$$\theta_{t+1} = \theta_t - \alpha \text{RMSProp}(\theta_t, g_{\theta_t})$$

where,  $\text{RMSProp}(\cdot)$  is the unpublished adaptive learning rate optimizer in deep learning.

10. A county-wide photovoltaic prediction method based on cluster division and data enhancement according to claim 1, wherein,

in the S4, data augmentation means to convert the data collected from S1 into images by dividing the collected data into 48 points per day to form an image with the size of 48\*15 pixels; then use the enhanced GAN for adversarial learning to generate new data with the same distribution.

11. A county-wide photovoltaic prediction method based on cluster division and data enhancement according to any one of claims 1-10, wherein,

in the S4, the CNN neural network has a 5-layer structure; each of the first four layers includes a convolution layer, a batch normalization layer and a ReLU layer, and the last layer includes a full connection layer, a batch normalization layer and a ReLU layer.

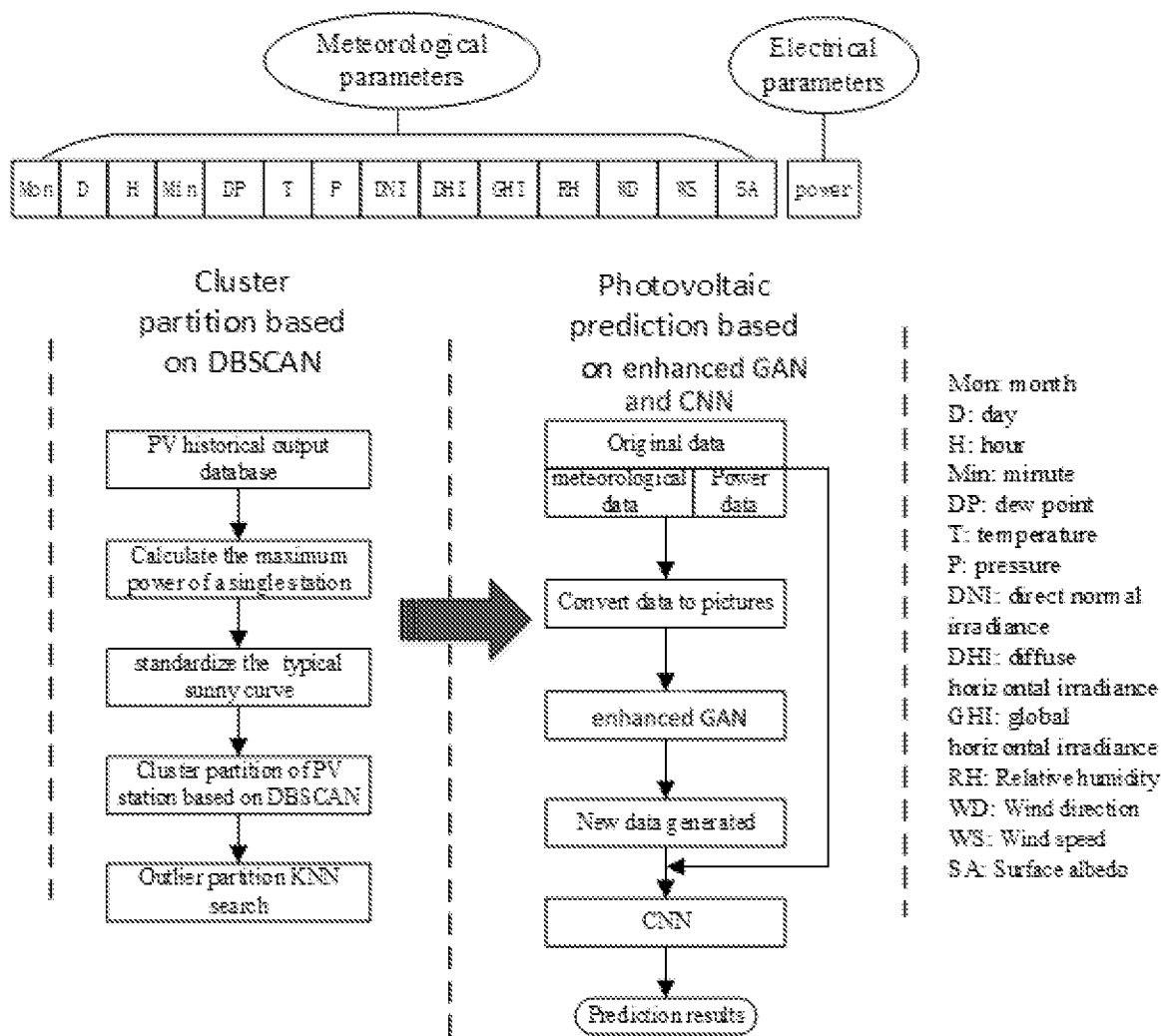


Figure 1

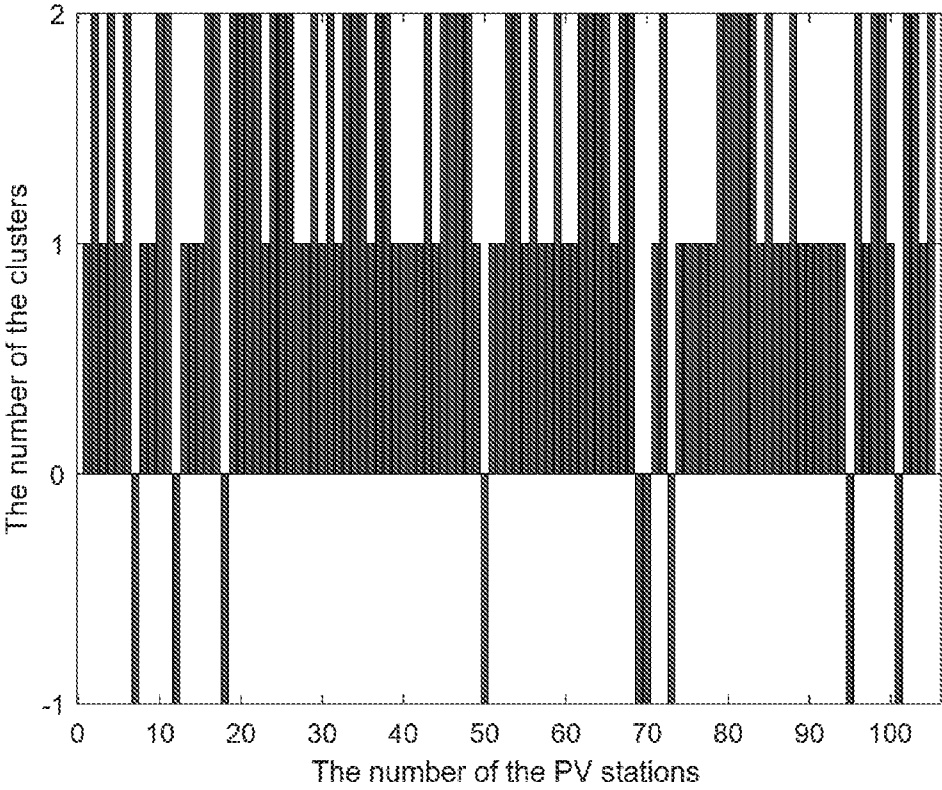


Figure 2

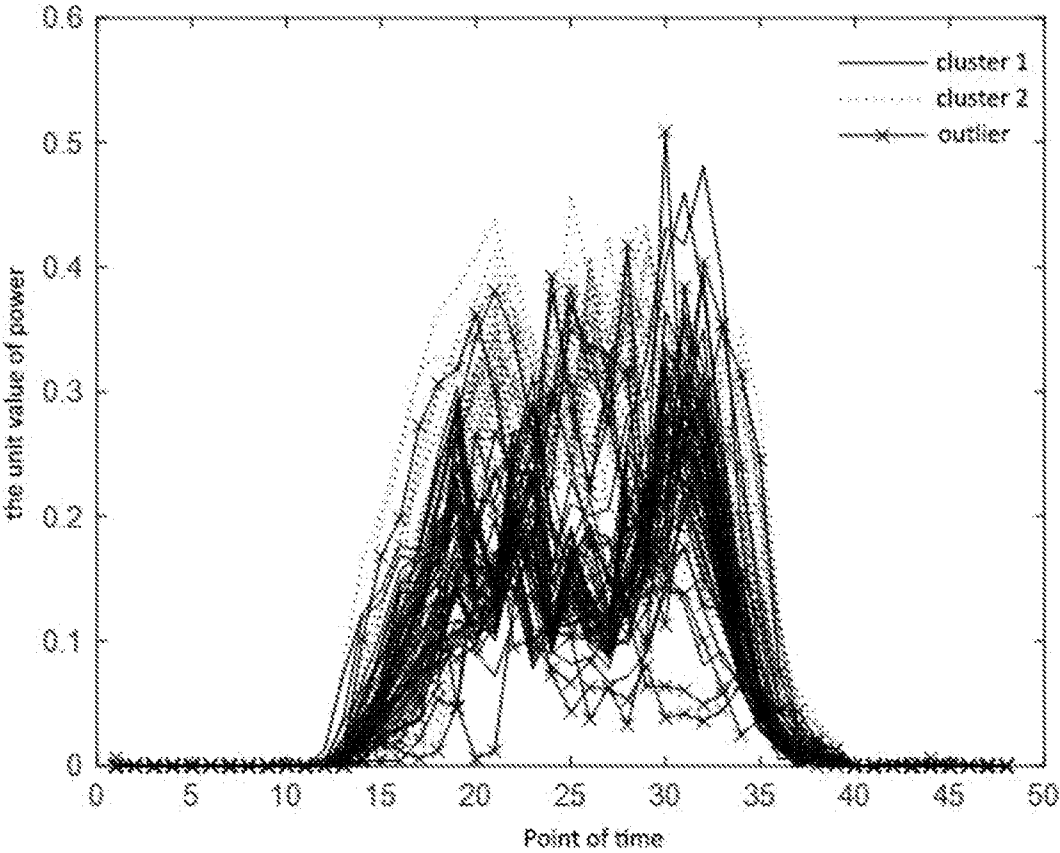


Figure 3

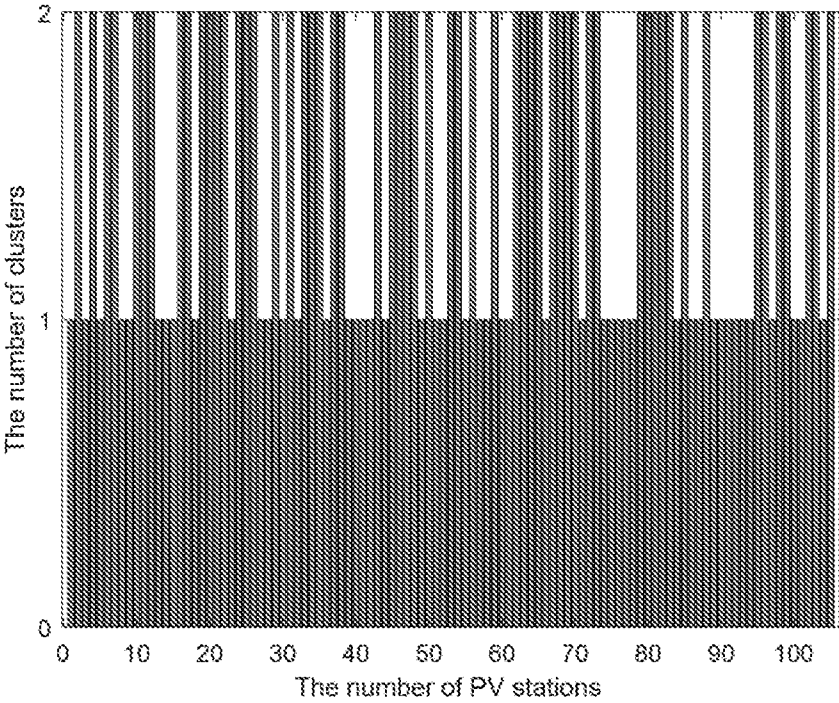


Figure 4

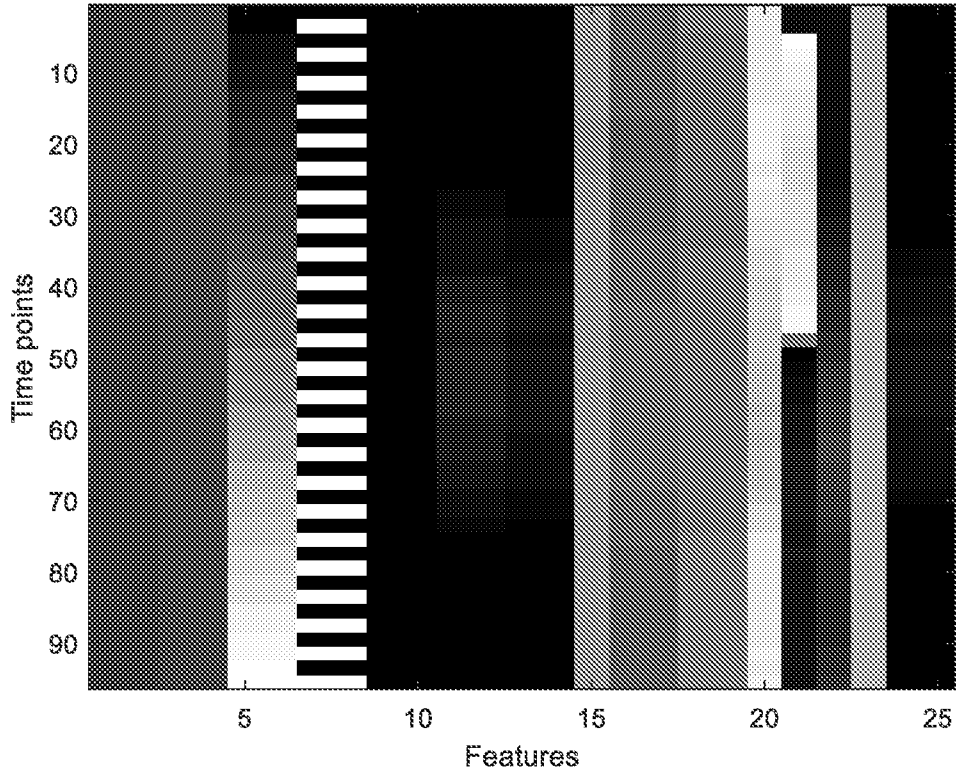


Figure 5

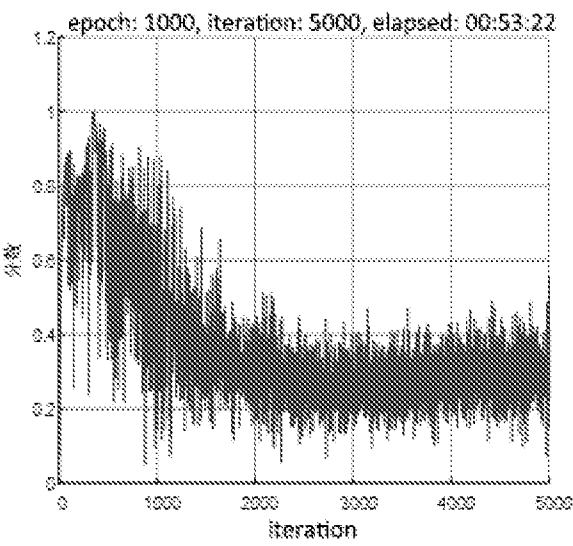
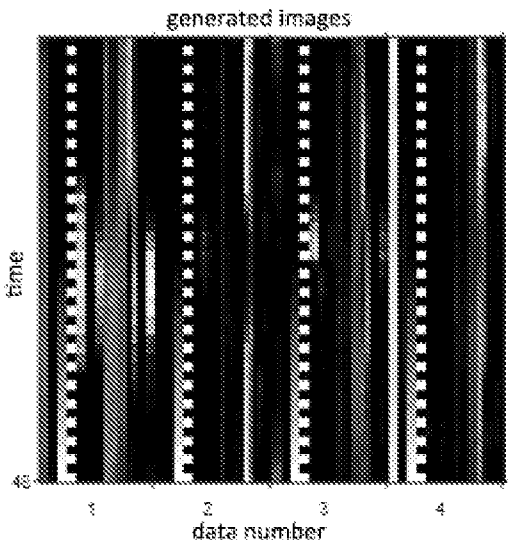


Figure 6

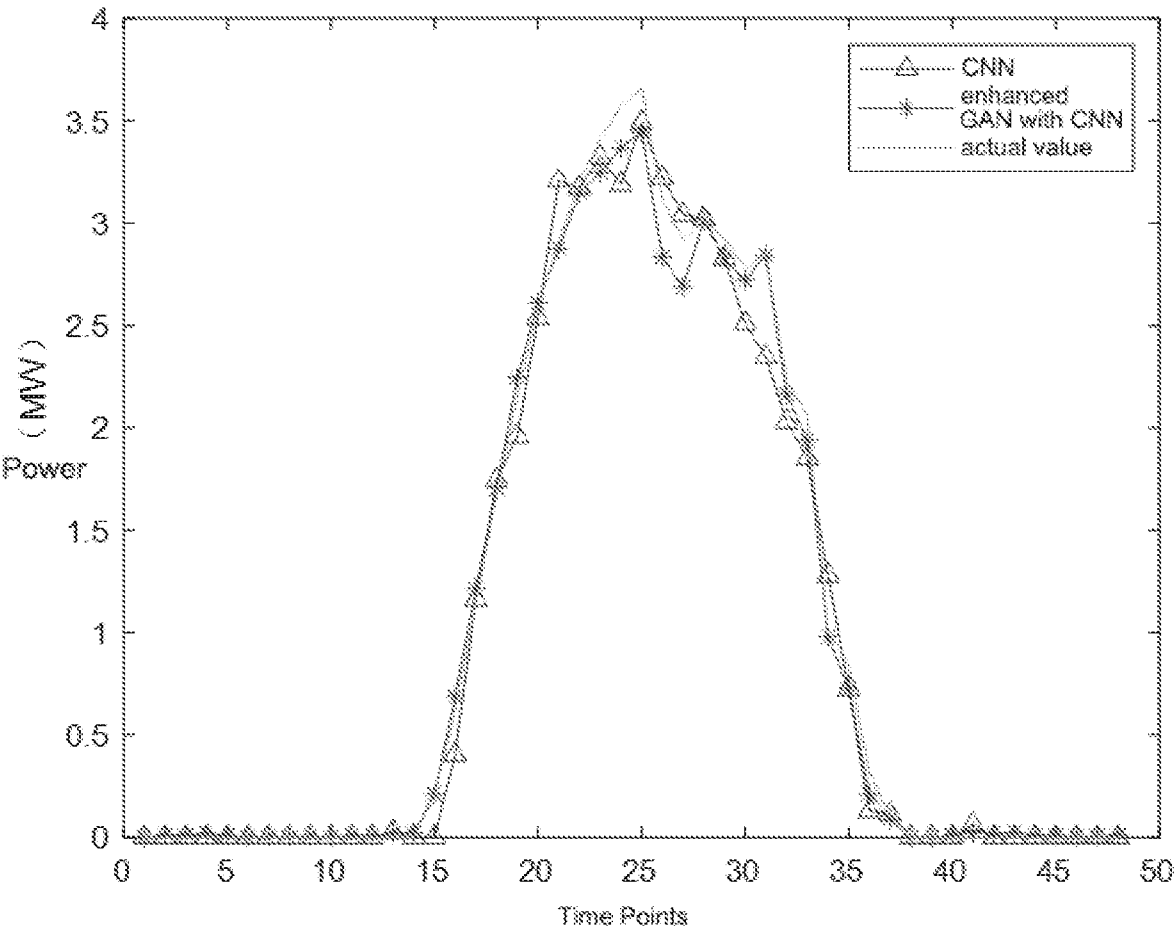


Figure 7

## INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/087787

**A. CLASSIFICATION OF SUBJECT MATTER**

G06Q 10/04(2012.01)i; G06Q 50/06(2012.01)i; G06K 9/62(2022.01)i; G06N 3/04(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

G06Q;G06K;G06N;H02J

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, CNKI, EPODOC, WPI, IEEE: PV, solar, photovoltaic, prediction, forecast, cluster, DBSCAN, picture, image, GAN, CNN

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                                                                                                                                                                                        | Relevant to claim No. |
|-----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| E         | CN 114492941 A (SOUTHEAST UNIVERSITY et. al.) 13 May 2022 (2022-05-13)<br>description, paragraphs [0006]-[0095] and figure 1                                                                                                                                                              | 1-11                  |
| A         | CN 110414748 A (HEFEI SUNGROW NEW ENERGY TECHNOLOGY CO., LTD.) 05 November 2019 (2019-11-05)<br>description, paragraphs [0073]-[0198] and figures 1-8                                                                                                                                     | 1-11                  |
| A         | CN 114169445 A (HOHAI UNIVERSITY et. al.) 11 March 2022 (2022-03-11)<br>the whole document                                                                                                                                                                                                | 1-11                  |
| A         | CN 110705760 A (GUANGDONG UNIVERSITY OF TECHNOLOGY) 17 January 2020 (2020-01-17)<br>the whole document                                                                                                                                                                                    | 1-11                  |
| A         | CN 110689183 A (SHANDONG UNIVERSITY) 14 January 2020 (2020-01-14)<br>the whole document                                                                                                                                                                                                   | 1-11                  |
| A         | WANG, Fei et. al. "Generative adversarial networks and convolutional neural networks based weather classification model for day ahead short-term photovoltaic power forecasting" <i>Energy Conversion and Management</i> , Vol. 181, 01 February 2019 (2019-02-01),<br>the whole document | 1-11                  |



Further documents are listed in the continuation of Box C.



See patent family annex.

\* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

08 December 2022

Date of mailing of the international search report

20 December 2022

Name and mailing address of the ISA/CN

National Intellectual Property Administration, PRC  
6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing  
100088, China

Facsimile No. (86-10)62019451

Authorized officer

HUANG,Shan

Telephone No. 86-(10)-53961235

**INTERNATIONAL SEARCH REPORT**  
**Information on patent family members**

International application No.

**PCT/CN2022/087787**

| Patent document<br>cited in search report |           |   | Publication date<br>(day/month/year) | Patent family member(s) | Publication date<br>(day/month/year) |
|-------------------------------------------|-----------|---|--------------------------------------|-------------------------|--------------------------------------|
| CN                                        | 114492941 | A | 13 May 2022                          | None                    |                                      |
| CN                                        | 110414748 | A | 05 November 2019                     | None                    |                                      |
| CN                                        | 114169445 | A | 11 March 2022                        | None                    |                                      |
| CN                                        | 110705760 | A | 17 January 2020                      | None                    |                                      |
| CN                                        | 110689183 | A | 14 January 2020                      | None                    |                                      |