Energy Reports 9 (2023) 1490–1500

Contents lists available at ScienceDirect

Energy Reports

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

Research paper

Photovoltaic power prediction under insufficient historical data based on dendrite network and coupled information analysis

Tianhao Lu, Chunsheng Wang *, Yuan Cao, Hong Chen

School of Automation, Central South University, Changsha, 410083, Hunan, China

ARTICLE INFO

Article history:
Received 29 July 2022
Received in revised form 19 October 2022
Accepted 23 December 2022
Available online xxxx

Keywords:
Photovoltaic power prediction
Insufficient historical data
Dendrite network
Coupled information analysis

ABSTRACT

In recent years, the installed capacity of photovoltaic solar energy has been increasing year by year. The existing power prediction methods are difficult to achieve reliable power generation prediction for PV equipment that has just been put into service. This poses great difficulties for power system management and dispatching. Therefore, establishing a reliable PV power prediction model for this situation is important for the rapid integration of new-built PV installations into the power system for energy management and dispatching. In this paper, an ultra-short-term PV power prediction method based on coupled information analysis and a dendritic network is proposed. First, an improved Hampel filter is proposed to improve the accuracy of anomaly data processing by analyzing the information between strongly coupled variables. In addition, a dendritic network modeling method is introduced for PV generation prediction. Compared to other solutions, this prediction approach relies on very little historical data to achieve reliable predictions and also features a simple model structure and high generalization. Forty different sets of tests were conducted according to different weather conditions and data conditions. The experimental results show that the proposed method can achieve higher prediction performance and stability compared with the benchmark model.

© 2022 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

With the continuous consumption of fossil energy and the deterioration of the environment caused by carbon emissions, many countries worldwide are studying photovoltaic (PV) power generation as an essential alternative renewable energy resource. PV installed capacity has been growing rapidly in recent 20 years (IEA, 2021). However, PV power generation is greatly affected by weather conditions, and its output power is characterized by intermittent uncertainty and volatility. So PV generation significantly burdens the dispatching operation of the entire power system (Sobri et al., 2018).

In order to manage solar photovoltaic power stations efficiently and improve the utilization rate of solar resources (Barbieri et al., 2017), it is necessary to establish a high-precision and reliable ultra-short-term pV power prediction approach (Barbieri et al., 2017). The ultra-short-term prediction approach can provide valuable guidance for the intra-day scheduling of energy. Many prediction methods have been proposed that utilize different types of inputs such as satellite images, total sky cloud maps, numerical weather prediction (NWP).

Currently, ultra-short-term PV prediction approaches are mainly based on artificial intelligence (AI) methods (Pereira et al., 2022; Liu et al., 2015; Khan et al., 2017b; Xin, 2020; Khan et al., 2017a; Pu et al., 2021; Wan et al., 2017; Jang et al., 2016; Zeng and Qiao, 2013; Zazoum, 2022). Many studies have proved that these methods are superior to other prediction methods (Chang et al., 2017; Mellit et al., 2020). These AI-based methods are usually data-driven, and a large amount of data is required for model training to improve model accuracy. However, these methods cannot be applied to all scenarios, especially for these new-built PV plants where historical data is insufficient. The performance and accuracy of AI-based prediction approaches cannot be guaranteed due to the lack of sufficient data, which may lead to a wrong operation decision, and result in economic loss.

Under the condition of insufficient data, the physical method in other prediction methods does not require large amounts of historical data. Physical methods estimate the meteorological data provided by the meteorological station to establish the PV power prediction. Nevertheless, the modeling of this method is complex, and the prediction accuracy is greatly affected by weather forecasts (Dolara et al., 2015; Tuohy et al., 2015). In other areas of research, data generation is used to solve the problem of insufficient data, such as random oversampling (RO) (Mao et al., 2019), synthetic minority oversampling technique (SMOTE) (Sun et al., 2020) and autoencoder (Creswell et al., 2018), Etc. However, these methods have a common shortage: the newly generated data can only simulate the shape of the existing actual data and

* Corresponding author.
E-mail address: wangcsu@csu.edu.cn (C. Wang).

https://doi.org/10.1016/j.egyr.2022.12.076

2352-4847/© 2022 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Nomenclature

P_{actual}	Actual PV power
$P_{predict}$	Predicted PV power
R^2	Goodness of fit
ARIMA	Auto-regressive integrated moving average
CI-Hampel filter	Improved Hampel filter based on coupled information analysis
DD	Dendrite network
DHR	Diffuse horizontal radiation
GHR	Global horizontal radiation
GRA	Grey relation analysis
LSTM	Long short-term memory neural network
MAE	Mean absolute error
MIC	Maximal information coefficient
PCC	Pearson correlation coefficient
PV	Photovoltaic
RBFNN	Radial basis function neural network
RDT	Radiation diffuses tilted
RGT	Radiation global tilted
RMSE	Root-mean-square error
SVM	Support vector machine
WS	Wind speed

lack data diversity (Wang et al., 2019). Yin et al. (2021), Chen et al. (2018) uses generative adversarial networks (GANs) to solve the shortage. However, the modeling of GANs is complex, and to the authors' best knowledge, little research has proved that GANs are effective for PV power prediction under insufficient historical data. In recent studies, Luo et al. (2022) accomplished the power prediction problem for a new-built PV plant by a transfer learning approach, but the method requires a source domain with sufficient data and the same variables. This limits the application scenario of the method.

Therefore, how to accomplish a reliable prediction without other additional data is a research gap and a problem to be addressed in this study. In order to solve this problem it is necessary to fully exploit the information of existing data. In this paper, we will carry out research work on both abnormal data processing and prediction models.

Identifying and processing abnormal data is an effective way to reduce the adverse impact of abnormal data on the accuracy of the prediction model. The most common processing method is to remove outliers and missing data directly (Ernst and Gooday, 2019; Alkhayat and Mehmood, 2021; Lin et al., 2018; Ray et al., 2020) the case of sufficient training data, this method is efficient and straightforward, but under the condition of insufficient data, removing data means losing more information. Therefore, processing by replacing outliers is a more appropriate approach. The Hampel filter (Park et al., 2021) is exactly a method that meets the requirements. It judges abnormal data by calculating the distance between the center value and the median value in the data window and then replaces the abnormal data with the corresponding median value. Sharadga et al. (2020) verifies that this method can effectively improve the prediction effect. However, this method uses only information from the data itself in identifying outliers and cannot determine whether the outliers are normal fluctuations caused by drastic changes in weather. This situation will result in the misidentification of outliers, resulting in new abnormal data.

In order to solve this problem under insufficient data, the information between variables with strongly coupled linear relationships in PV is a key factor. The strongly coupled variables have the same variation trend; extracting this information will be able to correct the outlier judgment result of the Hampel filter and improve the accuracy of data processing. Based on this, an improved Hampel filter based on coupled information analysis (CI-Hampel filter) is proposed.

Similar to abnormal data processing, the prediction model is also the key factor affecting prediction accuracy. As mentioned above, prediction models based on AI methods have better prediction performance than other methods. Many scholars have researched PV power prediction based on these methods. The main AI algorithms used in PV power prediction models include back propagation neural network (BPNN) (Liu et al., 2015; Khan et al., 2017b), support vector machine (SVM) (Jang et al., 2016), Elman neural network (ElmanNN) (Khan et al., 2017a; Ma and Zhang, 2022), radial basis neural network (RBFNN) (Xin, 2020), extreme learning machine (ELM) (Pu et al., 2021; Wang et al., 2017), long short-term memory neural networks (LSTM) (Chen and Chang, 2021) and convolutional neural networks (CNN) (Ghimire et al., 2019), etc. Before making predictions, these algorithms require a large amount of data for training. The prediction models based on these algorithms cannot be trained effectively under insufficient historical data, which leads to poor prediction accuracy of the prediction models.

To solve this problem, prediction models are needed to mine more information from the existing data for learning. Ziane et al. (2021), Zhang et al. (2021) shows that the influence of meteorological factors on PV power generation is not independent but a complex interaction. By extracting this interactive information, the prediction model can obtain abundant information under insufficient historical data, thus improving the accuracy of the prediction model. However, current prediction models, especially neural network models, cannot achieve this process. This is because the input variables of these models are always input to the neurons individually, without establishing a connection between the input variables. Cheng et al. (2021) extracted the relationship between various meteorological factors by means of graph modeling. However, the method requires a sufficient amount of data for training. Therefore, to achieve the extraction of interaction information under insufficient historical data requires the introduction of new intelligent algorithms. In this paper, a prediction model based on a dendrite network (DD) (Liu and Wang, 2021) is proposed. DD can extract information about the logical relationships of input variables under insufficient historical data by a new network structure, which leads to more reliable predictions.

In this paper, an ultra-short-term PV generation prediction model based on coupled information analysis and a dendritic network is proposed. This will provide a data reference for the intra-day dispatch of power systems containing new-built PV generation. The main contributions of this paper can be summarized as follows:

- An ultra-short-term PV generation prediction method is proposed to provide reliable output forecasts for new-built PV installations under insufficient historical data.
- An improved Hampel filter (CI-Hampel filter) based on coupled information analysis is proposed to solve the problem of outlier misidentification and to provide more effective training data for prediction models.
- Introducing DD as the prediction algorithm, the prediction approach combining the CI-Hampel filter and DD can achieve effective training and prediction by the information mining method without additional data.

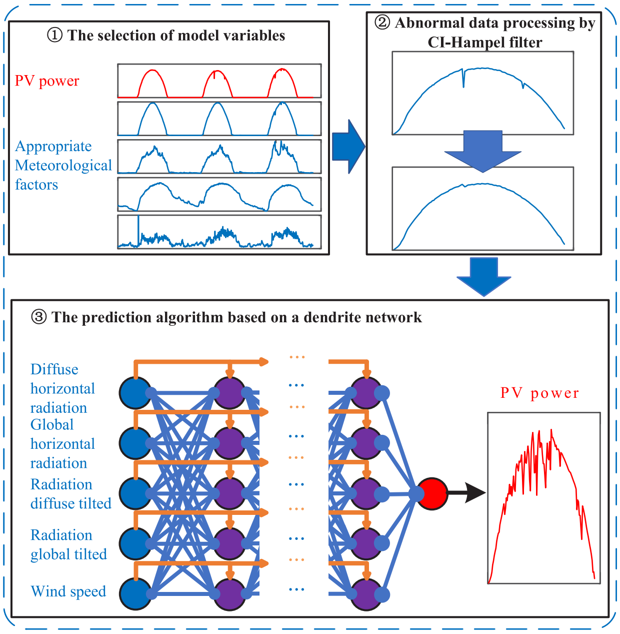
① The selection of model variables
PV power
Appropriate Meteorological factors
② Abnormal data processing by CI-Hampel filter
③ The prediction algorithm based on a dendrite network
Diffuse horizontal radiation
Global horizontal radiation
Radiation diffuse tilted
Radiation global tilted
Wind speed
PV power

Fig. 1. General architecture of the forecasting model.

Table 1
Meteorological factors and units.

Meteorological factors	Unit
Diffuse horizontal radiation	W/m^2
Global horizontal radiation	W/m^2
Radiation diffuse tilted	W/m^2
Radiation global tilted	W/m^2
Daily Rainfall	mm
Relative Humidity	%
Temperature Celsius	°C
Wind Direction	°
Wind Speed	m/s

- Performing experimental evaluation and verification of the proposed prediction model under different weather and season scenarios.

The remainder of this work is structured as follows: Section 2 introduces the basic architecture and method description of the prediction approach. The experimental evaluation metrics and experimental verification are given in Section 3. The discussion is given in Section 4 and finally the conclusion is given in Section 5.

2. Methodology

The overall architecture of the proposed ultra-short-term PV power prediction model is shown in Fig. 1. It consists of three modules: the selection of model variables, abnormal data processing by the CI-Hampel filter and the prediction algorithm based on a dendrite network. The details of the prediction approach are shown below.

2.1. The selection of model variables

The data studied in this paper are from the open-source dataset of Desert Knowledge Australia Solar Center (DKASC) (Dataset, 2021). The dataset consists of PV power generation data and meteorological factors data measured on-site. The information on meteorological factors is shown in Table 1.

Different selections of input variables will affect the prediction performance of the prediction model. It is a common method to select the key variables affecting the PV power by using the correlation coefficient. However, different correlation analysis methods may cause different conclusions. It is an appropriate solution to use different methods to conduct correlation analysis and choose the optimal combination of variables. This paper adopts the following three methods for analysis. (1) Maximal information coefficient (MIC); (2) Pearson correlation coefficient (PCC); (3) Grey Relation Analysis (GRA).

Table 2
The value range and characteristics of correlation analysis methods.

Method	Range	Characteristic
MIC	[0, 1]	Able to detect linear and nonlinear data relationships
PCC	[−1, 1]	Able to detect linear data relationships and indicate positive or negative correlations by a sign
GRA	[0, 1]	Able to detect linear data relationships and low requirements for data volume

Table 3
Correlation analysis results.

Meteorological factors	MIC	PCC	GRA
Diffuse horizontal radiation	0.4102	0.0609	0.5806
Global horizontal radiation	0.9503	0.9878	0.8575
Radiation diffuse tilted	0.4056	0.0847	0.5838
Radiation global tilted	0.9351	0.9833	0.8553
Daily Rainfall	0.0820	−0.1101	0.5294
Relative Humidity	0.1963	−0.4142	0.5688
Temperature Celsius	0.1764	0.3556	0.6854
Wind Direction	0.1855	0.0675	0.6456
Wind Speed	0.2185	0.4552	0.6820

The value range and characteristics of the three methods are shown in Table 2.

Due to the PV cells will not output at night, only the data correlation during the daytime is considered. The calculation results are shown in Table 3.

It can be seen from Table 3 that the three methods give different degrees of correlation for the same variable. Therefore, it is necessary to select appropriate input variables through experiments according to the results of the three methods. Through experiments in Section 3.3, the input variables of the prediction model are diffuse horizontal radiation (DHR), global horizontal radiation (GHR), radiation diffuse tilted (RDT), radiation global tilted (RGT) and wind speed (WS).

2.2. Abnormal data processing by CI-hampel filter

In order to reduce the negative impact of abnormal data and to have more information for prediction model training under insufficient historical data, abnormal data processing through data replacement is required. As mentioned earlier, Hampel filter is a suitable method. This method is able to identify and replace abnormal data using only a few data within the data window, but it also brings the problem of anomalous data identification errors. In this subsection, Hampel filter and the improvement process are described separately.

The abnormal data is mainly divided into the missing data and the outlier. The missing data means that the variable data is recorded as null, and outlier data means that the variable data exceeds the normal range or the variation trend range; the missing data can be regarded as special outliers. The Hampel filter identifies outliers by calculating the distance between the middle value of the data window and the median, then replaces the outliers with the median. The identification calculation formula is shown as follows.

$$|x_s - median\,(x_s)| > n \times \sigma \tag{1}$$

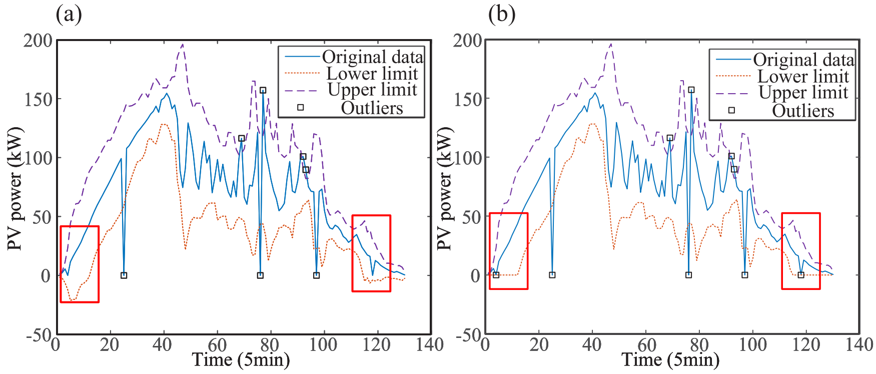

Fig. 2. (a) Hampel filter could not identify some outliers. (b) Improved Hampel filter based on physical characteristics.

where σ and $median(x_s)$ are the standard deviation and median of data in the window, respectively. The recognition range of outliers can be controlled by adjusting the window size and decision ratio n. For static data, the Hampel filter can effectively identify outliers. However, when the data in the window tends to be monotonically increasing or decreasing, the calculated standard deviation will be too large, resulting in a large range of outlier determination and failure to identify outliers, as shown in the red window in Fig. 2(a).

In the red window in Fig. 2(a), although these outliers conform to the filter's judgment of regular data, according to the physical characteristics of the data, the data should not be zero in the daytime range. Based on this feature, we make the following improvements to the filter's recognition formula.

$$\begin{cases} x_s \leq \max\{median\,(x_s) - n\cdot\sigma\,,0\} \\ x_s > median\,(x_s) + n\cdot\sigma \end{cases} \tag{2}$$

In this formula, we modified the lower limit of outlier determination. When the lower limit calculated according to formula (1) is less than zero, the lower limit will be changed to zero according to formula (2).

The outlier recognition effect based on formula (2) is shown in Fig. 2(b). It can be seen that outliers that could not be identified before can be identified after the improvement.

Although the Hampel filter is universal, it does not consider whether the outlier violently fluctuated due to weather changes. This defect will lead to misidentifying "normal outlier" data, especially for non-sunny day data. In this paper, "normal outliers" are defined as normal data fluctuations caused by drastic changes in weather rather than by abnormalities in sensors or measurement equipment.

In order to solve this problem, a coupled information analysis method is proposed to correct the outlier judgment results of the Hampel filter, which is named the CI-Hampel filter.

This method is mainly for PV power data and radiation data that are greatly affected by weather conditions. From the previous PCC analysis in Table 3, the PV power has a strong coupled linear relationship with global horizontal radiation (0.9878) and radiation global tilted (0.9833). This strong coupled linear relationship means that when one of these three variables changes, the remaining two variables will change with the same trend. Therefore, when the data of two or three variables are identified as outliers simultaneously, such outliers should be considered as "normal outliers" of normal changes rather than outliers caused by other faults. This process is shown in Fig. 3. The four outliers marked in the figure are considered "normal outliers".

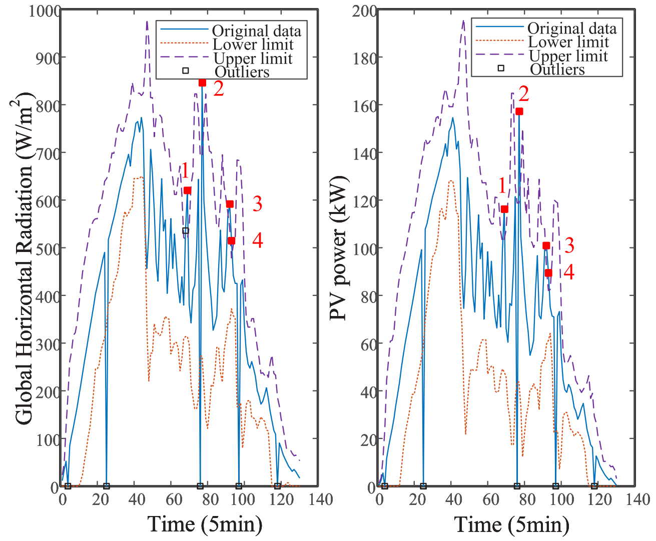

Fig. 3. Identifying "Normal Outliers" based on coupling Information Analysis.

Table 4
Linear relationship of selected variables.

	WS	GHR	DHR	RGT	RDT
WS	1	0.7337	0.5408	0.7332	0.5663
GHR	0.7337	1	0.6333	**0.9979**	0.6488
DHR	0.5408	0.6333	1	0.6384	**0.9911**
RGT	0.7332	**0.9979**	0.6384	1	0.6541
RDT	0.5663	0.6488	**0.9911**	0.6541	1

It is important to note that values identified as "normal outliers" should first be greater than zero to satisfy the physical characteristics of the data. In addition, by calculating the correlation between the selected variables, as shown in Table 4, a strongly coupled linear relationship (0.9911) also existed between diffuse horizontal radiation and diffuse tilted radiation.

Therefore, the same "normal outliers" identification method also applies to these two variables. In the outlier processing phase, retain the original value of the "normal outlier" and replace other outliers and missing values with the median value.

In order to achieve the best recognition effect of outliers (Pearson et al., 2016), different values of window length k and decision ratio n of the CI-Hampel filter should be set according to weather type or data type. The specific settings are shown in Appendix Table 1. The specific process of the CI-Hampel filter can be divided into the following steps.

Step 1: Input historical data.
Step 2: Determine the data type or weather type.

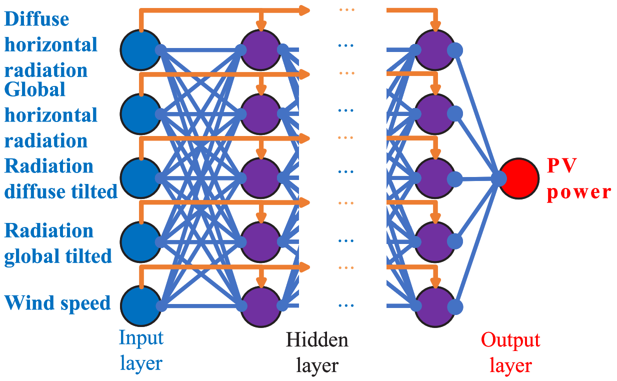

Fig. 4. The structure of DD.

Step 3: Determine the parameters k and n according to Table A.1.
Step 4: Calculate the upper and lower limit and median values.
Step 5: Set the lower limit to zero when the lower limit is less than zero.
Step 6: Determine the outliers based on the upper and lower limits.
Step 7: Identify "normal outliers" based on coupling information analysis; skip this step for variables without coupling variables.
Step 8: Outliers other than "normal outliers" are replaced by median values.
Step 9: Repeat the above steps until the last data.

2.3. The prediction algorithm based on a dendrite network

Dendrite Net (DD) is a basic machine learning algorithm with a new structure (Liu and Wang, 2021). Unlike the traditional neural network with multilayer perceptron (MLP) architecture, which simulates the function of the cell body part of the biological nervous system, DD simulates the information processing of dendrites in the biological nervous system by realizing multiple logical operations. The structure of DD is shown in Fig. 4.

The traditional multilayer perceptron structural (MLP) model can be written as $f(\Sigma wx)$ or $f(\Sigma wx+b)$, where f is the nonlinear mapping activation function. It can be seen that the input variable x is always input into the activation function independently, while DD adopts matrix multiplication and Hadamard product as the neuron operation function to build functional expressions containing information about logical relationships between inputs, as shown in Eq. (3).

$$A^{l}=W^{l,l-1}A^{l-1}\circ X \tag{3}$$

where A^{l-1} and A^{l} represents the input and output of the module at layer l, X is the initial input of the network, and $W^{l,l-1}$ represents the weight matrix between layer $l-1$ and layer l, respectively. The symbol $\circ$ represents the Hadamard product. The functional expressions contain information about logical relationships between inputs, which can be illustrated by a three-input-one-output-two-layer DD calculation process, as shown in Eq. (4).

$$\begin{aligned}
f\left(X\right) &= W^{21}\left(\left(W^{10}X\right)\circ X\right)\\
&= \begin{bmatrix} W_{11}^{21} & W_{12}^{21} & W_{13}^{21}\end{bmatrix}\\
&\times \left(\begin{bmatrix} W_{11}^{10} & W_{12}^{10} & W_{13}^{10}\\ W_{21}^{10} & W_{22}^{10} & W_{23}^{10}\\ W_{31}^{10} & W_{32}^{10} & W_{33}^{10}\end{bmatrix}\begin{bmatrix} x_0\\ x_1\\ x_2\end{bmatrix}\circ\begin{bmatrix} x_0\\ x_1\\ x_2\end{bmatrix}\right)\\
&= W_{11}^{21}W_{11}^{10}x_0^2+W_{12}^{21}W_{22}^{10}x_1^2+W_{13}^{21}W_{33}^{10}x_2^2\\
&+\left(W_{11}^{21}W_{12}^{10}+W_{12}^{21}W_{21}^{10}\right)x_0x_1+\left(W_{11}^{21}W_{13}^{10}+W_{13}^{21}W_{31}^{10}\right)x_0x_2\\
&+\left(W_{12}^{21}W_{23}^{10}+W_{13}^{21}W_{32}^{10}\right)x_1x_2
\end{aligned} \tag{4}$$

Where x_0 can be set as 1. The logical relation information can be expressed as: "And": multiplication (e.g., x_1x_2). "Or": addition (e.g., $x_1 + x_2$); "Not": minus (e.g., $-x_1$ or $-x_2$).

In this paper, DD uses a simple gradient descent rule for learning. The forward propagation of DD hidden layers and the output layer are shown in Eq. (5).

$$\begin{cases} A^{l}=W^{l,l-1}A^{l-1}\circ X\\ A^{L}=W^{L,L-1}A^{L-1}\end{cases} \tag{5}$$

The error-back propagation of DD hidden layers and the output layer are shown in Eqs. (6), (7) and (8).

$$dA^{L}=\hat{Y}-Y \tag{6}$$

$$\begin{cases} dZ^{L}=dA^{L}\\ dZ^{l}=dA^{l}\circ X\end{cases} \tag{7}$$

$$dA^{l-1}=\left(W^{l,l-1}\right)^{T}dZ^{l} \tag{8}$$

The weight adjustment of DD is made according to the following equations:

$$dW^{l,l-1}=\frac{1}{m}dZ^{l}\left(A^{l-1}\right)^{T} \tag{9}$$

$$W_{new}^{l,l-1}=W_{old}^{l,l-1}-\alpha dW^{l,l-1} \tag{10}$$

where $\hat{Y}$ and Y are DD's outputs and real values, respectively. m denotes the number of training samples in one batch. α is the learning rate.

The error backpropagation calculation of DD is similar to that of the BP network. However, without a nonlinear mapping function, the calculation speed of DD is faster under the same number of hidden layers (Liu et al., 2015). The computational complexity of overall DD is $O(2n+1)$.

In this study, the ability of DD to extract logical relationship information between selected variables is used to obtain more abundant information from insufficient historical data and achieve higher performance prediction than existing prediction methods.

3. Experiment and analysis

In this section, the effectiveness and advancement of proposed approaches are verified by experiments, respectively.

3.1. Evaluation metrics

In order to evaluate the performance of the proposed prediction model, mean absolute error (MAE), root-mean-square error (RMSE) and goodness of fit (R^2) are used as evaluation metrics. These metrics are defined by the following equations:

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|P_{actual}\left(i\right)-P_{predict}\left(i\right)\right| \tag{11}$$

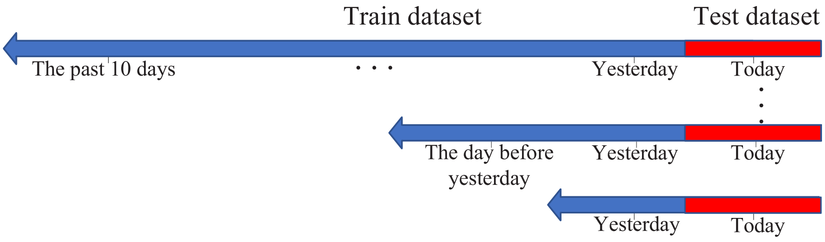

Fig. 5. Schematic diagram of experimental data.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(P_{actual}\left(i\right) - P_{predict}\left(i\right)\right)^2} \tag{12}$$

$$R^2 = 1 - \frac{Var\left(P_{actual} - P_{predict}\right)}{Var\left(P_{predict}\right)} \tag{13}$$

where P_{actual} and $P_{predict}$ represent the actual and predicted PV power, respectively, n represents the sampling points of the PV power generation period.

3.2. Data of the experiments

The experimental data for the study in this paper includes the total power generation data of 38 solar power stations in Alice Springs, Australia, in 2021, with a total installed capacity of 263.0 kW. The data is measured on-site, and the data interval is 5 min. Hence, our study focuses on the 5-minute ultra-short-term PV power prediction.

The Four cases of test data were set according to the season and weather type: Mar. 07, 2021 (Overcast); Jun. 02, 2021 (Cloudy); Sept. 18, 2021 (Rainy) and Dec. 29, 2021 (Sunny). Since the PV equipment has no power output at night, only the daytime data is considered. Historical data ranging from $t = 1, \ldots, 10$ days before the test data date were used as the train dataset to simulate varying degrees of insufficient data, as shown in Fig. 5.

All prediction experiments were performed by first training the prediction model using the training data set and then inputting the input variables in the test set to predict the corresponding PV power.

3.3. Experiments for selecting model variables

The subjective selection of variables lacks reasonable explanation, especially when the results obtained by the three methods are different, so the selection is made through experiments. The variables were sorted according to the correlation results of three correlation analysis methods. The top i = 1, 2, …, 9 variables were used as the input of the prediction model according to the ranking, respectively, to test the prediction performance of the model. The training data of each test set is the historical data of 10 days. DD-based model is the prediction model, and the hyperparameters are shown in Table A.2. The average MAE of the four test sets was used as the evaluation index of the prediction effect. The experimental results are shown in Fig. 6.

The results show that the prediction model using the top five variables, ranked according to MIC correlation as the input variables, has the lowest prediction error (4.6792 kW). Further increase of input variables will increase the prediction error, so these five variables (diffuse horizontal radiation, global horizontal radiation, radiation diffuse tilted, radiation global tilted and wind speed) are the most appropriate input variables for the prediction model.

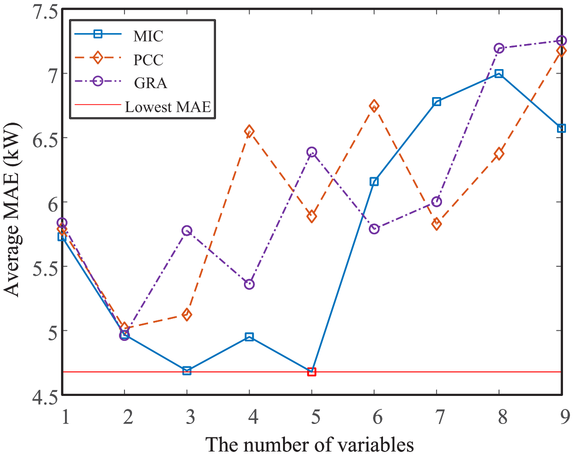

Fig. 6. Average MAE of prediction for different input variables.

Table 5
Comparison results of CI-Hampel filter and Hampel filter.

Test dataset	CI-Hampel filter			Hampel filter		
	MAE	RMSE	R^2	MAE	RMSE	R^2
Overcast	3.4607	4.989	0.9877	4.6595	6.402	0.9797
Cloudy	2.4761	3.7941	0.9933	3.4738	4.5484	0.9903
Rainy	2.7294	3.5551	0.9907	3.1423	4.0373	0.9881
Sunny	2.9022	3.3334	0.9965	3.5642	4.1077	0.9946

3.4. Comparison between Hampel filter and CI-Hampel filter

In order to test the performance of the CI-Hampel filter, the data processed by the two methods are respectively used as DD training data. Compare which abnormal data processing method performs better by predicting performance under the same parameter settings; the results are shown in Fig. 7, and the numerical results are shown in Table 5.

It can be seen from the results that the prediction model can achieve better prediction performance by training the data processed by the CI-Hampel filter. Compared with the prediction model, which was trained with Hampel filter processing data, MAE and RMSE of the prediction model were reduced by 13%–28% and 12%–22%, respectively, and R^2 increased by 0.001-0.003.

3.5. Comparison of our prediction model with benchmark models

3.5.1. Benchmark models

Since existing studies have demonstrated that AI-based models have more reliable prediction performance, RBFNN, SVM (Chang and Lin, 2011) and LSTM, which are the more prominent performers among such methods, are used as benchmark methods. In addition, as a classical method for time series prediction, ARIMA is also used as the benchmark method for comparison in order to explore the prediction effectiveness of other non-ML models. All benchmark methods and the proposed DD-based

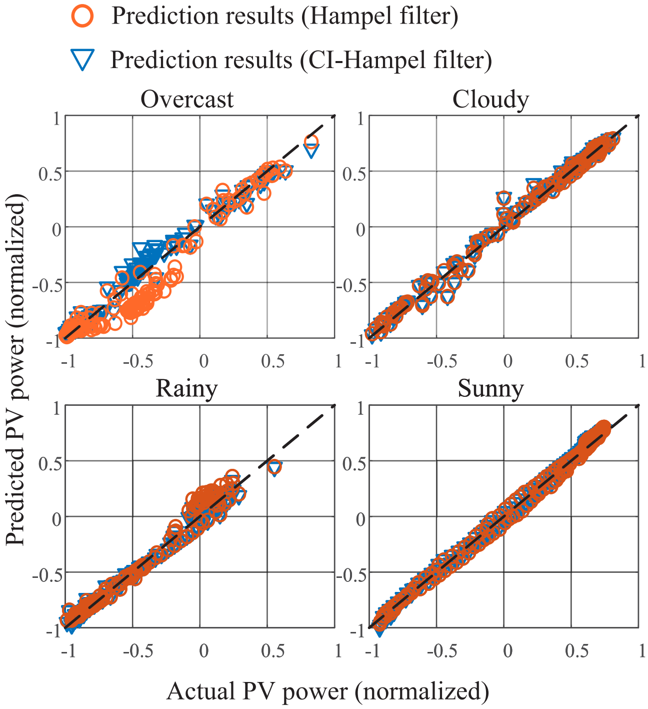

Fig. 7. Prediction results using two data processing methods.

model are trained with historical data processed by CI-Hampel filter, and the prediction performance of each model is tested under the same input conditions. Since the ARIMA model could not be run with less than three days of training data, no results were evaluated when the training data was one day of data and two days of data. The hyperparameters of DD and each benchmark model were determined by the grid search method as shown in Table A.2.

3.5.2. Prediction performance with varying degrees of insufficient historical data

In order to test the performance of prediction models with insufficient training data of varying degrees, the historical data from $t = 1, \ldots, 10$ days before the test data were used as the training data, respectively. The following is the analysis of PV prediction performance under four different cases of weather conditions.

Case 1: Overcast

When the weather on the forecast day is overcast, the prediction performance of each prediction model is shown in Fig. 8. The difference in performance of the different models can be clearly seen in the figure. From the results, our DD-based model has the best prediction performance, and the prediction error is always smaller than other models under the same training conditions. The next best performer is the LSTM model. The worst prediction performance is the ARIMA model. From the result of R^2, its predictive performance is even lower than the prediction using the average of the data.

Case 2: Cloudy

The prediction performance of each prediction model is shown in Fig. 9. It can be seen that the benchmark models produce a large prediction error with only one day of training data. And with the increase in training data, the prediction performance has improved significantly. Our model performs the best among these methods. When there are more than two days of training data, the prediction performance of SVM approaches that of our model.

Case 3: Rainy

As shown in Fig. 10, The performance of our DD-based prediction model for such rainy days is similar to that for cloudy days, as reflected in the similarity of MAE, RMSE and R^2 results. For the benchmark models, the prediction performance improves as the training data increases, but the MAE and RMSE are still higher than our model.

Case 4: Sunny

The prediction performance is shown in Fig. 11. From the results, it can be seen that all prediction models perform better when the forecast day is sunny compared to other forecast weather cases. The MAE and RMSE of LSTM are slightly smaller than our model when the training data is from 1 day to 3 days of data. However, when the amount of training data is larger than three days of data,our model is superior to all benchmark models.

From the above four cases, a total of forty sets of comparative experiments, it can be seen that our predictive models can achieve reliable prediction under varying degrees of insufficient historical data. Especially for overcast, cloudy and rainy days, our method has a substantial advantage compared with benchmark models. Compared with LSTM, which has the best prediction performance among the benchmark models, MAE and RMSE of our DD-based model are reduced by 12.63%–67.21% and 9.51%–68.16%, respectively. Moreover, R^2 is improved by 0.0017–0.2517.

As the amount of training data increases, the prediction errors of all models show an overall decreasing trend. This trend illustrates that an increase in the amount of training data can lead to more effective training of the model. However, not all increases in the amount of training data lead to more effective training of the prediction model. As shown in case 4, the prediction error of the LSTM increases when the amount of training data is more than three days of historical data. This is because the new training data and the test data are not the same weather type, so the new training data has a negative impact on the prediction model resulting in poor prediction performance. Compared to benchmark models, our model is more robust and has less change in prediction performance.

Summarize all evaluation results and draw a box-plot. The distribution of MAE, RMSE and R^2 for each prediction model for the forty experiments is shown in Fig. 12. Our prediction model can always maintain little MAE, RMSE, and large R^2 under various conditions. it is obvious that our model outperforms the benchmark models with superior stability (tighter bound of the box).

4. Discussion

The purpose of this study is to design a method that can achieve reliable ultra-short-term PV power prediction under insufficient historical data. The method consists of three modules: the selection of model variables, abnormal data processing by the CI-Hampel filter and the prediction algorithm based on a dendrite network.

The study in the model variable selection section found that MIC was the most suitable input variable screening method for DD. This is because MIC is able to detect linear and nonlinear data relationships in the data, which fits well with the way DD is computed. From the experimental results, it appears that using DHR, GHR, RDT, RGT and WS as the input variables of the prediction model can obtain more accurate prediction results. However, if there are not so many multiple meteorological factor variables, similar accuracy prediction results can be obtained with

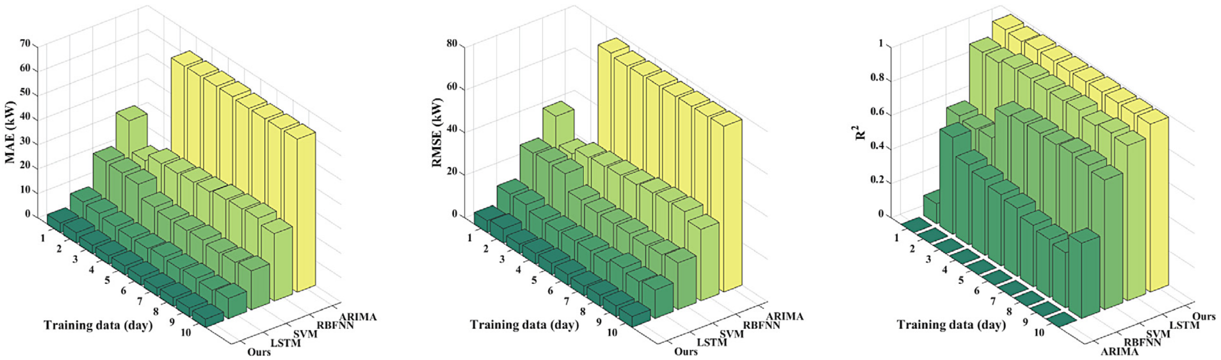

Fig. 8. Prediction performance, test set on Mar. 07 with the weather condition of overcast.

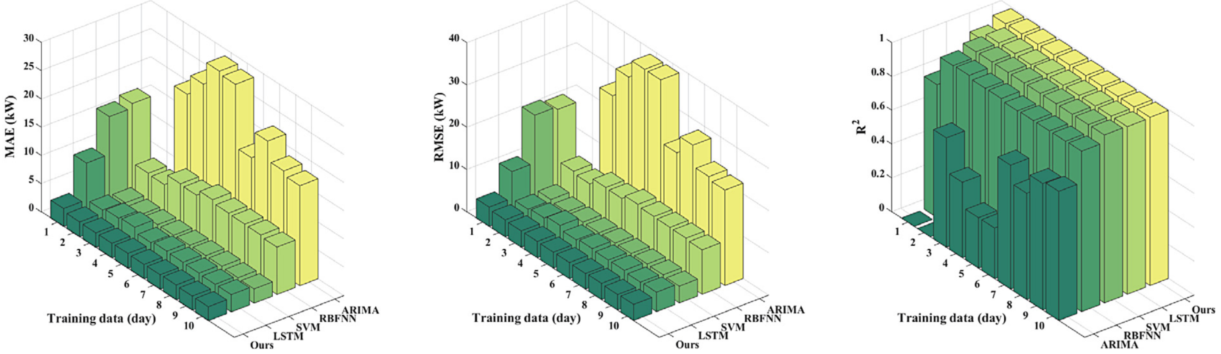

Fig. 9. Prediction performance, test set on June 02 with the weather condition of cloudy.

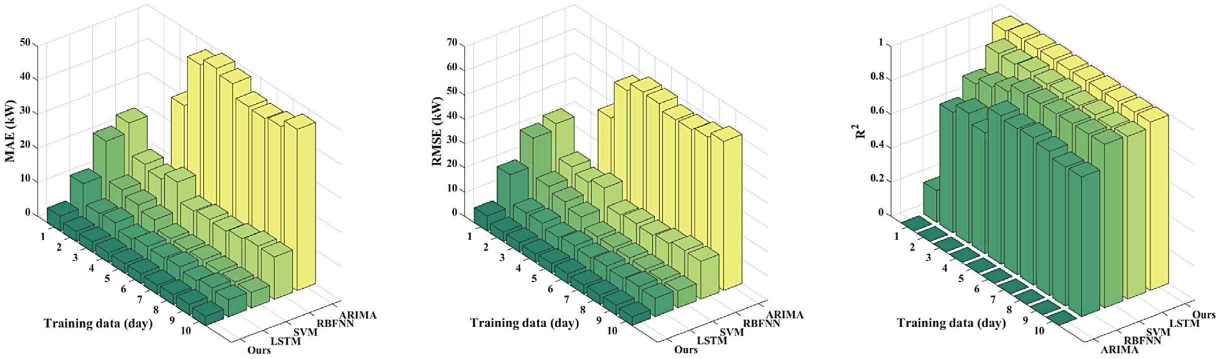

Fig. 10. Prediction performance, test set on Sept. 18 with the weather condition of rainy.

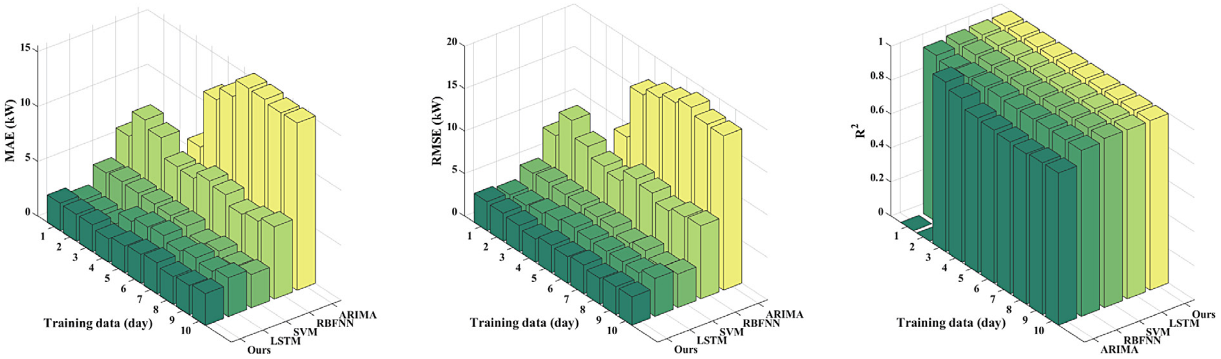

Fig. 11. Prediction performance, test set on Dec. 29 with the weather condition of sunny.

only GHR, RGT and DHR. It should be noted that the experiments in this study are based on open-source data from DKASC, so the experimental conclusions only apply to PV installations with the same type of data. Due to differences in climate and time,

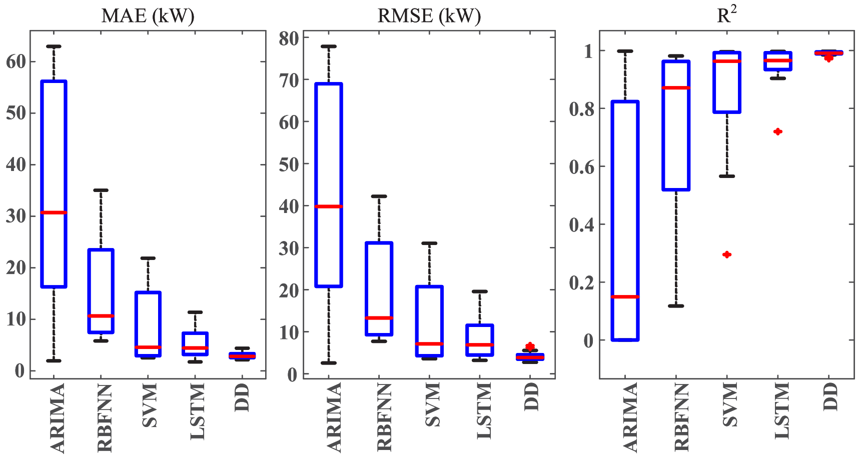

Fig. 12. Box-plot summarizing all evaluation results.

PV installations in other regions may have different data types and data sampling intervals, so the variable selection for the prediction model should be re-run using MIC.

The study in the abnormal data processing section found that the problem of the Hampel filter misidentifying "normal outliers" could be solved by analyzing the information between strongly coupled variables. The experimental results verify that the training data processed based on the CI-Hampel filter enables the prediction model to obtain more accurate prediction results. The method can be applied to other areas of data processing provided that there is a strong coupling linear relationship between at least two variables.

The prediction model part of the study introduced DD as the prediction algorithm of the model. Forty different sets of prediction experiments were conducted to verify that DD can achieve reliable prediction by extracting information about logical relationships between input variables. It is important to note that the hyperparameters of both DD and benchmark models are determined by means of the grid search method. The hyperparameters settings of many works of literature were referred to in this process to ensure the reasonableness of the experimental comparison results.

Most importantly, our method has a simple structure. This means that there is much room for improvement, such as using an optimization algorithm for weight optimization or a similar-day selection method for selecting training data would further improve the prediction performance. In the future, the proposed approach will be tested in more cases with different feature variables, such as sky satellite images, total-sky cloud images, and NWP. In addition, the method proposed in this work has the potential to be applied to solve other engineering problems where historical data are insufficient, such as load demand forecasting and wind power generation.

5. Conclusion

In this paper, a prediction approach of ultra-short-term PV power generation based on DD and coupled information analysis is proposed. This approach consists of a CI-Hampel filter and a DD-based prediction model. It can realize reliable ultra-short-term PV power generation prediction under the condition of insufficient historical data. Unlike solutions in other studies, this approach does not require data generation models or transfer learning.

Firstly, the CI-Hampel filter is proposed to achieve more effective abnormal data processing by analyzing the information between strongly coupled variables and providing higher-quality training data for the prediction model. Secondly, a DD-based prediction model for PV power generation is proposed to achieve reliable prediction with a small amount of training data by extracting the information of logical relationships between input variables. Finally, the prediction approach combining the CI-Hampel filter and DD is tested on an open source dataset with forty groups. Experiments show fewer prediction errors can be obtained by applying the method under various weather conditions, and its prediction performance and stability are better than benchmark models. Compared with LSTM, which has the best prediction performance among the benchmark models, MAE and RMSE of our DD-based model are reduced by 12.63%–67.21% and 9.51%–68.16%, respectively. Moreover, R^2 is improved by 0.0017–0.2517.

CRediT authorship contribution statement

Tianhao Lu: Conceptualization, Methodology, Software, Writing – original draft. **Chunsheng Wang:** Funding acquisition, Project administration, Writing – review & editing. **Yuan Cao:** Investigation, Resources, Supervision. **Hong Chen:** Visualization.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

This research work was supported by the National Natural Science Foundation of China (61973322 and 62103443), and by Hunan Natural Science Foundation under Grant 2022JJ40630.

Appendix. Supplementary material

The parameters setting of the Hampel filter and the hyperparameters setting of algorithms utilized in experiments are shown in Tables A.1 and A.2. In Table A.1, the parameters of CI-Hampel filter for PV power and irradiance data are set according to weather conditions, and separate parameters are set for wind speed data.

Table A.1
Linear relationship of selected variables.

Weather type or variable	Window size (k)	n
Sunny	3	1
Cloudy	5	2
Overcast	10	3
Rainy	10	3
Wind speed	5	3

Table A.2
Hyperparameters setting of each prediction model in experiments.

Algorithm	Hyperparameters
DD	DD module number: 4; Batch size: 1; Max epoch: 50; Learning rate: 0.05;
ARIMA	Stationarity test: ADF, KPSS; Difference order d:0; P Q Determination method: AIC+BIC;
SVM	SVM type: $\epsilon - SVR$ (regression); kernel type: radial basis function $exp(-\gamma * \|u-v\|^2)$; cost: 2.2; γ : 2.8; ϵ : 0.01;
RBFNN	Mean squared error goal: 0; Spread of radial basis functions: 0.7; Maximum number of neurons: 5; Number of neurons to add between displays: 25;
LSTM	Hidden units: 10; Max epoch: 100; Gradient Threshold: 1; Initial learn rate: 0.005; Learn rate schedule: piecewise; Learn rate drop period: 50; Learn rate drop factor: 0.2; Verbose: 0;

References

Alkhayat, G., Mehmood, R., 2021. A review and taxonomy of wind and solar energy forecasting methods based on deep learning. Energy AI 4, 100060. http://dx.doi.org/10.1016/j.egyai.2021.100060.

Barbieri, F., Rajakaruna, S., Ghosh, A., 2017. Very short-term photovoltaic power forecasting with cloud modeling: A review. Renew. Sustain. Energy Rev. 75, 242–263. http://dx.doi.org/10.1016/j.rser.2016.10.068.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. URL: https://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Chang, G., Lu, H., Chang, Y., Lee, Y., 2017. An improved neural network-based approach for short-term wind speed and power forecast. Renew. Energy 105, 301–311. http://dx.doi.org/10.1016/j.renene.2016.12.071.

Chen, H., Chang, X., 2021. Photovoltaic power prediction of LSTM model based on pearson feature selection. Energy Rep. 7, 1047–1054. http://dx.doi.org/10.1016/j.egyr.2021.09.167, URL: https://www.sciencedirect.com/science/article/pii/S2352484721009768, 2021 International Conference on Energy Engineering and Power Systems.

Chen, Y., Wang, Y., Kirschen, D., Zhang, B., 2018. Model-free renewable scenario generation using generative adversarial networks. IEEE Trans. Power Syst. 33 (3), 3265–3275. http://dx.doi.org/10.1109/TPWRS.2018.2794541.

Cheng, L., Zang, H., Ding, T., nong Wei, Z., qiang Sun, G., 2021. Multi-meteorological-factor-based graph modeling for photovoltaic power forecasting. IEEE Trans. Sustain. Energy 12, 1593–1603.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A., 2018. Generative adversarial networks: An overview. IEEE Signal Process. Mag. 35 (1), 53–65. http://dx.doi.org/10.1109/MSP.2017.2765202.

Dataset, 2021. Desert knowledge Australia solar center. URL: http://dkasolarcentre.com.au/.

Dolara, A., Leva, S., Manzolini, G., 2015. Comparison of different physical models for PV power output prediction. Sol. Energy 119, 83–99. http://dx.doi.org/10.1016/j.solener.2015.06.017.

Ernst, M., Gooday, J., 2019. Methodology for generating high time resolution typical meteorological year data for accurate photovoltaic energy yield modelling. Sol. Energy 189, 299–306. http://dx.doi.org/10.1016/j.solener.2019.07.069.

Ghimire, S., Deo, R.C., Raj, N., Mi, J., 2019. Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms. Appl. Energy 253, 113541. http://dx.doi.org/10.1016/j.apenergy.2019.113541.

IEA, 2021. World Energy Outlook 2021. Technical Report, IEA, Paris, URL: https://www.iea.org/reports/world-energy-outlook-2021.

Jang, H.S., Bae, K.Y., Park, H.-S., Sung, D.K., 2016. Solar power prediction based on satellite images and support vector machine. IEEE Trans. Sustain. Energy 7 (3), 1255–1263. http://dx.doi.org/10.1109/TSTE.2016.2535466.

Khan, I., Zhu, H., Yao, J., Khan, D., 2017a. Photovoltaic power forecasting based on Elman Neural Network software engineering method. In: 2017 8th IEEE International Conference on Software Engineering and Service Science. ICSESS, IEEE, Beijing, China, pp. 747–750. http://dx.doi.org/10.1109/ICSESS.2017.8343021.

Khan, I., Zhu, H., Yao, J., Khan, D., Iqbal, T., 2017b. Hybrid power forecasting model for photovoltaic plants based on neural network with air quality index. Int. J. Photoenergy 2017, 1–9. http://dx.doi.org/10.1155/2017/6938713.

Lin, P., Peng, Z., Lai, Y., Cheng, S., Chen, Z., Wu, L., 2018. Short-term power prediction for photovoltaic power plants using a hybrid improved Kmeans-GRA-Elman model based on multivariate meteorological factors and historical power datasets. Energy Convers. Manage. 177, 704–717. http://dx.doi.org/10.1016/j.enconman.2018.10.015.

Liu, J., Fang, W., Zhang, X., Yang, C., 2015. An improved photovoltaic power forecasting model with the assistance of aerosol index data. IEEE Trans. Sustain. Energy 6 (2), 434–442. http://dx.doi.org/10.1109/TSTE.2014.2381224.

Liu, G., Wang, J., 2021. Dendrite Net: A white-box module for classification, regression, and system identification. IEEE Trans. Cybern. 1–14. http://dx.doi.org/10.1109/TCYB.2021.3124328.

Luo, X., Zhang, D., Zhu, X., 2022. Combining transfer learning and constrained long short-term memory for power generation forecasting of newly-constructed photovoltaic plants. Renew. Energy 185, 1062–1077. http://dx.doi.org/10.1016/j.renene.2021.12.104, URL: https://www.sciencedirect.com/science/article/pii/S0960148121018310.

Ma, X., Zhang, X., 2022. A short-term prediction model to forecast power of photovoltaic based on MFA-Elman. Energy Rep. 8, 495–507. http://dx.doi.org/10.1016/j.egyr.2022.01.213, URL: https://www.sciencedirect.com/science/article/pii/S235248472200213X, 2021 International Conference on New Energy and Power Engineering.

Mao, W., Liu, Y., Ding, L., Li, Y., 2019. Imbalanced fault diagnosis of rolling bearing based on generative adversarial network: A comparative study. IEEE Access 7, 9515–9530. http://dx.doi.org/10.1109/ACCESS.2018.2890693.

Mellit, A., Massi-Pavan, A., Ogliari, E., Leva, S., Lughi, V., 2020. Advanced methods for photovoltaic output power forecasting: A review. Appl. Sci. 10 (2), 487. http://dx.doi.org/10.3390/app10020487.

Park, J.-i., Lee, Y., Suh, E., Jeon, H., Yoon, K.-J., Kim, K.-s., 2021. Improvement of optical flow estimation by using the Hampel filter for low-end embedded systems. IEEE Robot. Autom. Lett. 6 (4), 7233–7239. http://dx.doi.org/10.1109/LRA.2021.3095927.

Pearson, R.K., Neuvo, Y., Astola, J., Gabbouj, M., 2016. Generalized Hampel filters. EURASIP J. Adv. Signal Process. 2016 (1), 87. http://dx.doi.org/10.1186/s13634-016-0383-6.

Pereira, S., Abreu, E.F., Iakunin, M., Cavaco, A., Salgado, R., Canhoto, P., 2022. Method for solar resource assessment using numerical weather prediction and artificial neural network models based on typical meteorological data: Application to the south of Portugal. Sol. Energy 236, 225–238. http://dx.doi.org/10.1016/j.solener.2022.03.003.

Pu, S., Li, Z., Wan, H., Chen, Y., 2021. A hybrid prediction model for photovoltaic power generation based on information entropy. IET Gener. Transm. Distrib. 15 (3), 436–455. http://dx.doi.org/10.1049/gtd2.12032.

Ray, B., Shah, R., Islam, M.R., Islam, S., 2020. A new data driven long-term solar yield analysis model of photovoltaic power plants. IEEE Access 8, 136223–136233. http://dx.doi.org/10.1109/ACCESS.2020.3011982.

Sharadga, H., Hajimirza, S., Balog, R.S., 2020. Time series forecasting of solar power generation for large-scale photovoltaic plants. Renew. Energy 150, 797–807. http://dx.doi.org/10.1016/j.renene.2019.12.131.

Sobri, S., Koohi-Kamali, S., Rahim, N.A., 2018. Solar photovoltaic generation forecasting methods: A review. Energy Convers. Manage. 156, 459–497. http://dx.doi.org/10.1016/j.enconman.2017.11.019.

Sun, J., Li, H., Fujita, H., Fu, B., Ai, W., 2020. Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. Inf. Fusion 54, 128–144. http://dx.doi.org/10.1016/j.inffus.2019.07.006.

Tuohy, A., Zack, J., Haupt, S.E., Sharp, J., Ahlstrom, M., Dise, S., Grimit, E., Mohrlen, C., Lange, M., Casado, M.G., Black, J., Marquis, M., Collier, C., 2015. Solar forecasting: Methods, challenges, and performance. IEEE Power Energy Mag. 13 (6), 50–59. http://dx.doi.org/10.1109/MPE.2015.2461351.

Wan, C., Lin, J., Song, Y., Xu, Z., Yang, G., 2017. Probabilistic forecasting of photovoltaic generation: An efficient statistical approach. IEEE Trans. Power Syst. 32 (3), 2471–2472. http://dx.doi.org/10.1109/TPWRS.2016.2608740.

Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., Wang, F.-Y., 2017. Generative adversarial networks: introduction and outlook. IEEE/CAA J. Autom. Sinica 4 (4), 588–598. http://dx.doi.org/10.1109/JAS.2017.7510583.

Wang, F., Zhang, Z., Liu, C., Yu, Y., Pang, S., Duić, N., Shafie-khah, M., Catalão, J.P., 2019. Generative adversarial networks and convolutional neural networks based weather classification model for day ahead short-term photovoltaic power forecasting. Energy Convers. Manage. 181, 443–462. http://dx.doi.org/10.1016/j.enconman.2018.11.074.

Xin, W., 2020. Photovoltaic power prediction based on RBF neural network optimized by Gray wolf algorithm. In: 2020 3rd International Conference on Control and Robots. ICCR, IEEE, Tokyo, Japan, pp. 226–230. http://dx.doi.org/10.1109/ICCR51572.2020.9344245.

Yin, H., Ou, Z., Zhu, Z., Xu, X., Fan, J., Meng, A., 2021. A novel asexual-reproduction evolutionary neural network for wind power prediction based on generative adversarial networks. Energy Convers. Manage. 247, 114714. http://dx.doi.org/10.1016/j.enconman.2021.114714.

Zazoum, B., 2022. Solar photovoltaic power prediction using different machine learning methods. Energy Rep. 8, 19–25. http://dx.doi.org/10.1016/j.egyr.2021.11.183, URL: https://www.sciencedirect.com/science/article/pii/S2352484721013287, 2021 The 8th International Conference on Power and Energy Systems Engineering.

Zeng, J., Qiao, W., 2013. Short-term solar power prediction using a support vector machine. Renew. Energy 52, 118–127. http://dx.doi.org/10.1016/j.renene.2012.10.009.

Zhang, M., Liu, W., Qi, W., 2021. Experimental study on the influence of temperature and radiation on photovoltaic power generation in summer. IOP Conf. Ser.: Earth Environ. Sci. 621 (1), 012030. http://dx.doi.org/10.1088/1755-1315/621/1/012030.

Ziane, A., Necaibia, A., Sahouane, N., Dabou, R., Mostefaoui, M., Bouraiou, A., Khelifi, S., Rouabhia, A., Blal, M., 2021. Photovoltaic output power performance assessment and forecasting: Impact of meteorological variables. Sol. Energy 220, 745–757. http://dx.doi.org/10.1016/j.solener.2021.04.004.