Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 12, No. 1, 2025

Research on Distributed Photovoltaic Data Prediction Based on Deep Learning Method

Nuolin Yu *

School of Science, Shandong Jianzhu University, Jinan, Shandong, China
*** Corresponding author Email:** 15689096889@163.com

Abstract: Aiming at the problem of distributed photovoltaic output fluctuation caused by weather state change and the technical bottleneck of the existing prediction model lacking refined meteorological information support, this paper constructs a long short-term memory (LSTM) network prediction model that integrates the characteristics of meteorological historical data. Firstly, based on the $k-means$ method meteorological information fusion technology, the historical meteorological data set is divided into four basic categories: sunny, cloudy, cloudy, rain and snow according to typical weather characteristics, and special attention is paid to the weather transition process between various categories to form a fusion data set covering the complete weather evolution scene. Then, based on the data set training, a distributed photovoltaic short-term power prediction model is constructed. Finally, by inputting the weather type parameters of the day to be predicted, the LSTM network power prediction results are output and the validity of the model is verified. The experimental data show that compared with the traditional BP neural network model, the average prediction error of the fusion model proposed in this paper is significantly reduced, which can provide more accurate decision-making basis for distributed energy access planning and optimal scheduling and resource allocation of power system.

Keywords: Distributed Photovoltaic; Long Short-Term Memory; Data Prediction.

1. Introduction

With the transformation of the global energy structure to low-carbon, distributed photovoltaic power generation, as an important part of clean and renewable energy, continues to rise in the penetration rate of power systems [1], [2], [3]. However, the distributed photovoltaic output is significantly affected by meteorological conditions such as light intensity, temperature, and cloud changes, showing strong intermittence and volatility, which poses a huge challenge to the stability and economic operation of the power system. Accurate distributed photovoltaic power prediction can not only effectively improve the new energy consumption capacity and optimize the grid scheduling strategy, but also provide a key basis for distributed energy investment decisions [4], [5].

At present, distributed photovoltaic power prediction still faces many technical bottlenecks. On the one hand, traditional forecasting methods rely on a single data source or a simple mathematical model, which makes it difficult to capture the complex nonlinear relationship between meteorological conditions and PV output [6]; on the other hand, the phenomenon of scattered distributed photovoltaic stations, missing or incomplete historical meteorological data is widespread, resulting in insufficient generalization ability of the model, and the prediction accuracy is difficult to meet the actual needs [7]. Many studies have also been carried out on this issue in the existing literature. Yan et al. [8] proposed an ultra-short-term photovoltaic power prediction model based on optimal frequency domain decomposition and deep learning. [9] proposed a short-term wind power forecasting method based on support vector machine optimized by hybrid algorithm. The comprehensive prediction of distributed photovoltaic short-term power generation by using weather classification technology was studied in [10]. The prediction accuracy of photovoltaic power is not only related to the selected prediction method, but also closely related to different weather conditions. Although some studies have tried to introduce deep learning algorithms to solve the above problems, there is still room for optimization in the fine processing and feature fusion of meteorological data.

Based on the above analysis, this paper focuses on the key technical problems of distributed photovoltaic data prediction, and proposes a prediction method that combines meteorological historical data and long-term and short-term memory network (LSTM). It aims to improve the accuracy and reliability of distributed photovoltaic power prediction through refined meteorological data classification and deep feature mining, and provide theoretical support and technical reference for the efficient operation of power system and the scientific planning of distributed energy.

2. Meteorological Information Fusion Clustering Process based on $k-means$ Method

Limited by the high construction and maintenance costs of meteorological monitoring stations, there may be practical difficulties in accurate meteorological data. At present, it mainly relies on non-precise weather forecast to realize weather condition discrimination. According to the 33 climate types defined by the GB/T22164-2008 standard of the National Meteorological Administration, this study classifies the weather types into four basic categories: sunny, cloudy, cloudy and rain and snow through meteorological fusion technology.

In view of the complex weather conditions including the weather transition process, such as sunny to cloudy, cloudy to rain, rain to snow, it is necessary to cluster the weather types in advance. Specifically, the $k-means$ clustering analysis method is adopted. Firstly, the clustering centers of the four basic weather types are extracted, and the Euclidean

distance between the weather sequence data with the transition process and each clustering center is calculated by the data approximation algorithm, and then it is merged into the nearest typical weather state. On this basis, combined with the real-time weather forecast information of the forecast day, the deep learning model is used to construct the distributed photovoltaic output power prediction model, and the accurate prediction is realized through the training of historical data set.

Taking the historical data of distributed photovoltaic as an example, the specific process of $k-means$ clustering algorithm is as follows:

(1) Set the distributed photovoltaic object set $H=\{H_1,H_2,\ldots,H_N\}$, N is the total number of data scenarios. Then the photovoltaic sequence on the d day is $H_d=(h_{d1},h_{d2},\ldots,h_{dM})$, the photovoltaic sequence on the f day is $H_f=(h_{f1},h_{f2},\ldots,h_{fM})$, and M is the number of photovoltaic data points counted in a day. Then the cosine similarity between H_d and H_f is:

$$S_{\cos}(H_d,H_f)=\frac{H_dH_f}{|H_d||H_f|}=\frac{\sum_{n=1}^{M}h_{dn}h_{fn}}{\sqrt{\sum_{n=1}^{M}h_{dn}^2}\sqrt{\sum_{n=1}^{M}h_{fn}^2}} \quad (1)$$

where the value range of $S_{\cos}(H_d,H_f)$ is $[0,1]$. The closer $S_{\cos}(H_d,H_f)$ is to 1, the more similar the photovoltaic curve shape represented by H_d and H_f is.

(2) In order to quickly find the optimal number of clusters and obtain better clustering results, the contour coefficient method is introduced to determine the appropriate number of clusters based on the $k-means$ -clustering algorithm. $s_c(d)$ is the contour coefficient value, which is defined as follows:

$$s_c(d)=\frac{b(d)-a(d)}{\max[a(d),b(d)]} \quad (2)$$

where

$$s_c(d)=\begin{cases}1-\dfrac{a(d)}{b(d)} & a(d)<b(d)\\ 0 & a(d)=b(d)\\ \dfrac{b(d)}{a(d)}-1 & a(d)>b(d)\end{cases} \quad (3)$$

where $a(d)$ is the average cosine similarity from sample H_d to other samples H_f in the same class. The smaller the $a(d)$, the closer the distance between the sample H_d and other samples in the same class, that is, the more similar. $b(d)$ is the average cosine similarity of all samples H_d from sample H_f to other categories. The larger the $b(d)$ is, the farther the distance between sample H_d and other categories is, that is, the less similar.

(3) The comprehensive contour coefficient of the clustering result $S_c(d)$ can be obtained by averaging the contour coefficients of all samples. The specific definition is as follows:

$$S_C=\frac{1}{N}\sum_{d=1}^{N}s_c(d) \quad (4)$$

where the S_C value range is $[-1,1]$. The closer the S_C value is to 1, the better the clustering effect is. The corresponding value k when the index value S_C reaches the maximum is the optimal number of clusters; the optimal clustering result is output, and the K typical time series scene corresponding to the optimal clustering number is obtained, denoted as $1,2,\cdots,k,\cdots,K$. The number of original scenes contained in the k-th typical scene is N_k, and N is the number of all scenes. Then the probability of the occurrence of the k -th typical scenario is $a_k=N_k/N$. According to the clustering results, the data of changing weather types are clustered into four basic categories: sunny, cloudy, cloudy and rain and snow. The process of changing weather is shown in Table.1.

Table 1. Changing weather processes

Weather categories	Changing weather types
Sunny categories	Sunny, sunny to cloudy, sunny to rain and snow
Cloudy categories	Cloudy, cloudy to sunny, cloudy to overcast sky, cloudy to rain and snow
Overcast sky categories	Overcast sky, overcast sky to sunny, overcast sky to cloudy, overcast sky to rain and snow
Rain and snow categories	Rain and snow, rain and snow to sunny, rain and snow to cloudy, rain and snow to overcast sky

The fusion of meteorological information processing on the original historical data can provide more obvious data for the prediction model to improve the overall accuracy of the prediction. After the deep learning model is built, the large-scale data set is extracted with reference to the fuzzy weather forecast. The characteristic quantity of the input data set and the output object are trained, and the neural network is trained to fully mine the relationship between the historical data to improve the accuracy and obtain the prediction results.

3. Short-term Distributed Photovoltaic Prediction Model based on LSTM

In this paper, after data fusion, data prediction methods are mainly divided into two kinds. The first method is the real meteorological data included in the historical data of photovoltaic power, which determines the meteorological categories corresponding to the historical data. The second is the forecast of the weather type on the day of the forecast day, which is generally provided by the meteorological department. Based on this information, it is possible to decide which type of weather type data to use for prediction. Compared with the traditional shallow learning method, the LSTM deep learning method can better mine the relationship in historical data.

3.1. Analysis of Experimental Results

In order to improve the long-term dependence of recurrent neural networks, a gate mechanism is introduced to control the speed of information storage, including selectively adding

new information and selectively forgetting previously stored information. In this paper, a class of recurrent neural networks based on gating is proposed. The long-term and short-term memory network changes according to the characteristics of the recurrent neural network, and can properly deal with the problem of simple recurrent neural network-gradient explosion and disappearance. Its operation principle is shown in Figure 1.

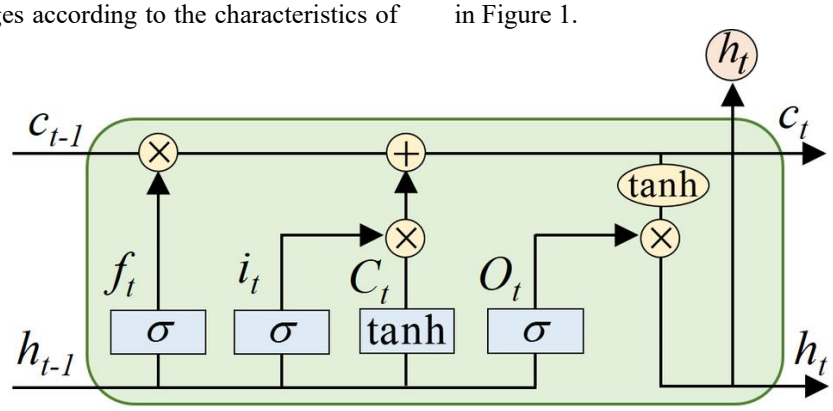

Figure 1. Loop structure diagram of LSTM unit network

LSTM mainly controls the transmission of data through three gates, namely forgetting gate, input gate and output gate.

$$\begin{aligned} i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ \tilde{s}_t &= \tanh(W_s x_t + U_s h_{t-1} + b_s) \end{aligned} \tag{5}$$

where x_t is the input of the current time and h_{t-1} is the external state of the previous time. W_*, U_*, b_* are the network structure parameters that need to be updated in the LSTM network. x represents the long-term memory information of the time series. The activation functions σ and $\tanh$ map the output to the $[0,1]$ interval and $[-1,1]$ interval respectively. The larger the value is, the more information is transmitted forward at the previous moment. The smaller the value is, the less information is transmitted forward at the previous moment.

Each gate consists of a sigmoid layer (which outputs a vector from 0 to 1, representing 'the degree to which it is allowed to pass') and a dot product operation. The three gates control the retention or forgetting of information:

(1) The Forgetting Gate: Decide which old information in the cell state is discarded.

(2) Input gate: Decide what new information is added to the cell state.

(3) Output gate: determine which part of the cell state is output as the current hidden state.

The characteristic of the LSTM network structure is that with the increase of stratification, errors do not decay rapidly, thus deepening the learning depth. Compared with traditional learning algorithms such as shallow neural networks, the learning effect has not changed much after more than two levels, and its advantages are obvious. Its core idea is to make the neural network have the ability of 'selective memory' through adaptive information screening, which is suitable for complex scenes that need to model long-term dependence.

3.2. Distributed Photovoltaic Short-Term Output Prediction Process

The value of LSTM model parameters has a great influence on its fitting performance. Experiments show that the selection of super parameters often depends on the nature of historical data itself, and too large or too small can not necessarily get the ideal results. The prediction model proposed in this paper includes two parts: meteorological fusion and deep learning network. The problem of fuzzy weather information is solved by meteorological fusion, and the problem is solved by deep learning network. The overall prediction process is shown in Figure2, as follows:

(1) Firstly, the data of the output power period from 8: 00 to 18: 00 is extracted, and then the input sequence of the prediction model is sorted out and the training set and test set are divided. Since the inputs are the same dimensional data, there is no need for normalization.

(2) The input of the prediction model is extracted by correlation analysis. Based on the $k-means$ -clustering algorithm, four clustering centers are obtained, and the distance between various types of weather data sets and clustering centers is calculated to realize the meteorological fusion of historical data sets.

(3) Based on the LSTM network, the four types of data sets after meteorological fusion are used for training to obtain the trained LSTM model.

(4) Enter the weather type of the day to be predicted to obtain the predicted daily power prediction value.

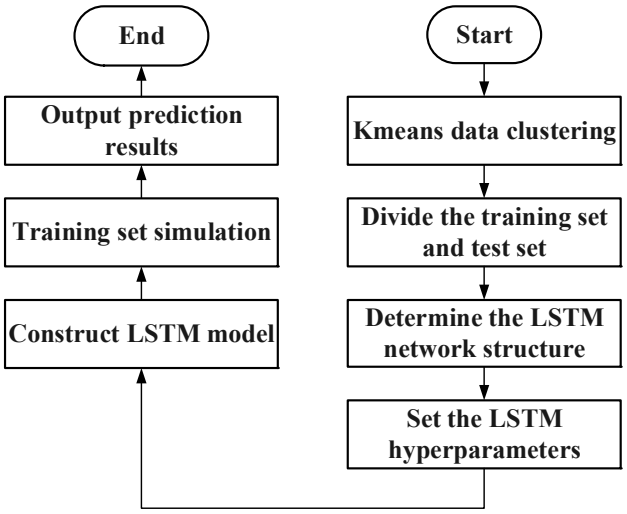

Figure 2. Photovoltaic short-term output prediction process

4. Example Analysis

This paper takes the historical photovoltaic data of a certain

place for example analysis. The data occurred from January 1 to December 31, 2020. By analyzing historical data, considering that the annual photovoltaic output period is 8: 00-18: 00, this period is used as the predicted data set. The time interval of a single period in this paper is 15 min.

Through the $k-means$ -clustering algorithm, the historical photovoltaic power sequence is classified, and the following results are obtained:

From the specific situation of the various types of weather coverage in the Table.2, sunny and cloudy categories contain more days.

In order to highlight the advantages of meteorological fusion and LSTM algorithm, the prediction results of unclassified data are compared with those after classification, and the prediction results of LSTM algorithm combined with weather characteristics are compared with those of BP shallow neural network.

Table 2. Change weather clustering results

Weather categories	Sunny categories	Cloudy categories	Overcast sky categories	Rain and snow categories
Number of categories	124	101	67	73

In the experiment, the super parameter setting of the LSTM model is as follows: the initial learning rate is 0.005, the number of hidden layer neurons is 230, the maximum number of iterations is 100, and the minimum package size is 242.

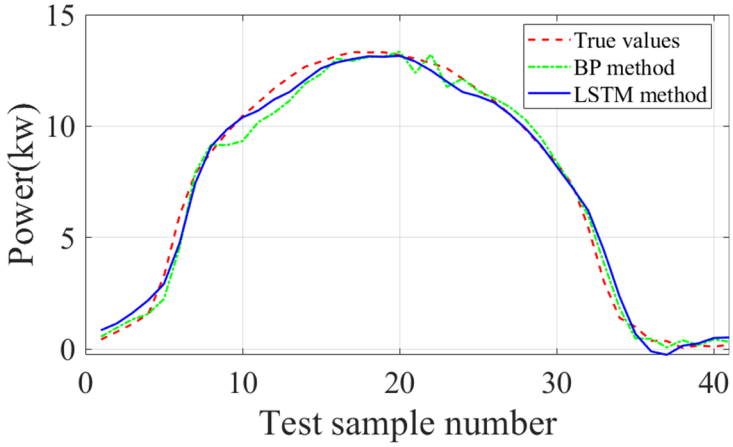

Figure 3. Comparison of typical sunny forecast results

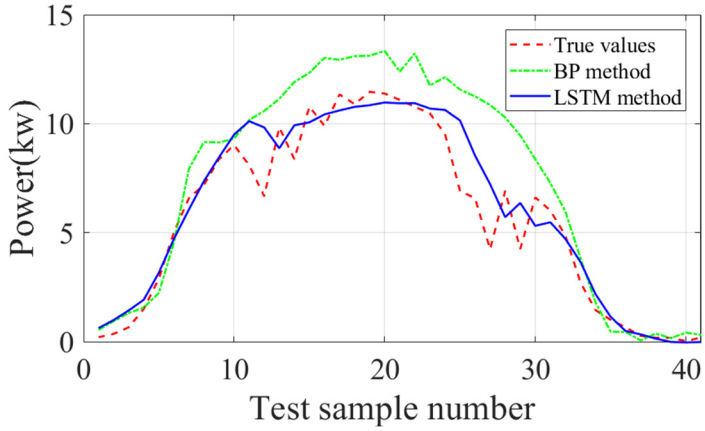

Figure 4. Comparison of typical cloudy forecast results

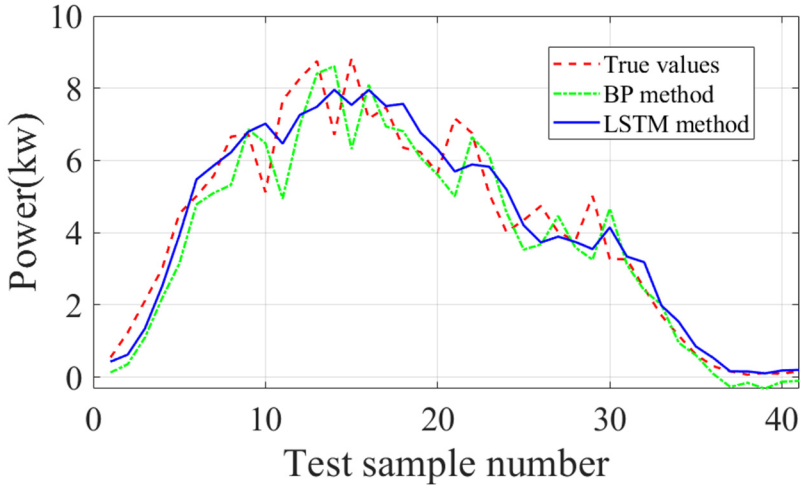

Figure 5. Comparison of typical overcast sky forecast results

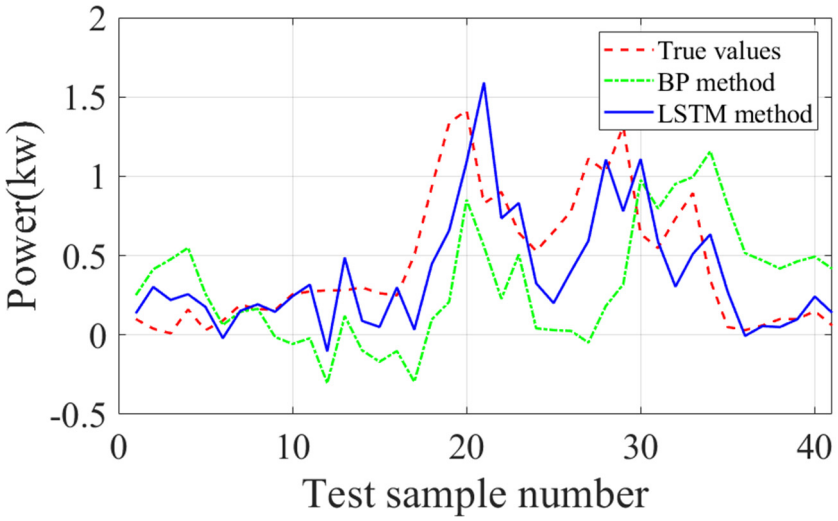

Figure 6. Comparison of typical rain and snow forecast results

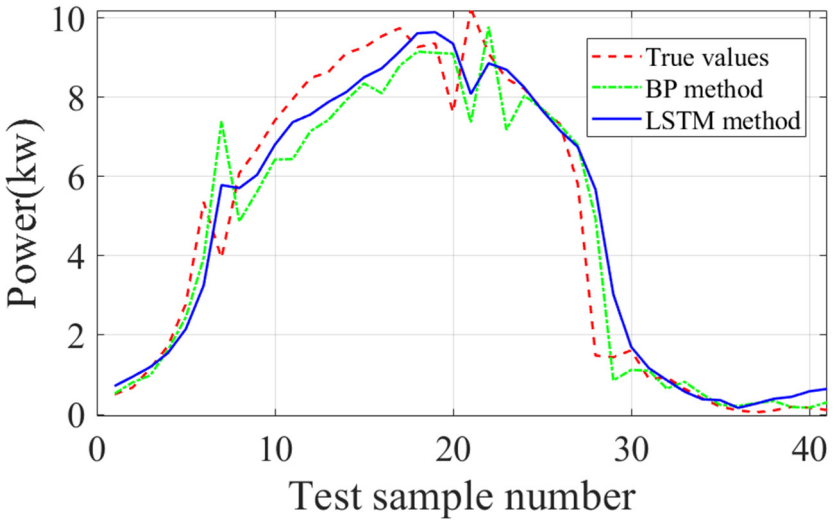

Figure 7. Comparison of prediction results of sunny to cloudy days

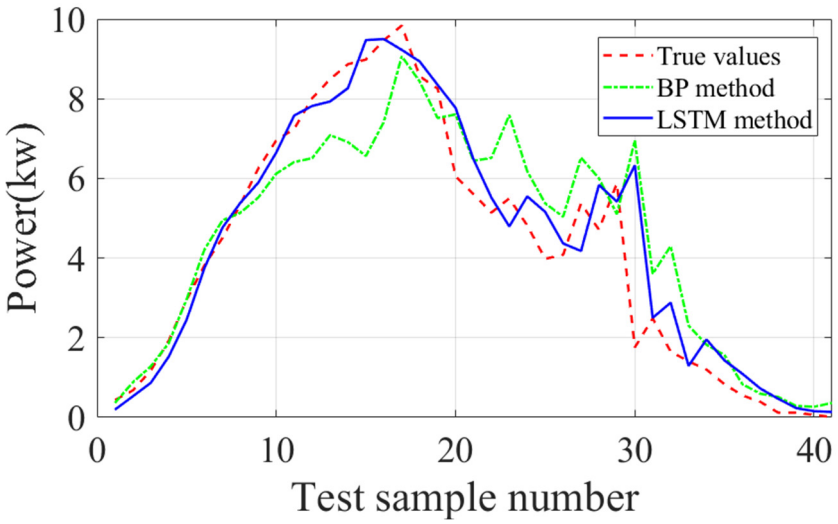

Figure 8. Comparison of prediction results of overcast sky to rain and snow

The above figures 3 to 8 show the comparison of the four algorithms of the photovoltaic output prediction curve under different weather conditions. It can be intuitively seen that the accuracy of the LSTM algorithm is higher than that of the BP neural network and is closer to the real value.

In order to comprehensively analyze the effectiveness and accuracy of the prediction model, the following three indicators are selected to evaluate the performance of various prediction models. The formula is as follows:

$$RMSE(X,h)=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x_i)-y_i)}$$
$$MAPE(X,h)=\frac{100\%}{m}\sum_{i=1}^{m}\left|\frac{h(x_i)-y_i}{y_N}\right| \quad (6)$$

where $h(x_i)$ is the predicted value of the model, y_i is the true value, and y_N is the rated installed capacity. $RMSE$ is used to judge the difference between the predicted value and the actual value, and $MAPE$ reflects the average percentage error on the rated installed capacity.

Table 3. Comparison of prediction accuracy of different algorithms

Weather types	MAPE		RMSE	
	BP	LSTM	BP	LSTM
Typical sunny	6.12%	3.45%	1.27	0.94
Typical cloudy	12.4%	8.23%	2.45	1.73
Typical overcast sky	6.37%	5.67%	1.65	1.24
Typical rain and snow	3.15%	2.12%	0.56	0.38
Sunny to cloudy	6.92%	5.56%	1.32	1.20
Overcast sky to rain and snow	12.55%	6.21%	2.87	1.30
Average error	7.92%	5.21%	1.69	1.13

Table 3 shows the specific error values of different algorithms under the two indicators, and the smaller the value, the better. The MAPE value of LSTM is 2.71% lower than that of BP, and the RMSE value of LSTM is also reduced, which is 0.56 lower than that of BP.

5. Conclusion

Aiming at the problem of insufficient power prediction accuracy of distributed photovoltaic power stations due to the lack of historical meteorological data, this paper proposes a distributed photovoltaic output prediction method that integrates meteorological information and LSTM algorithm. Through the test and simulation analysis under different weather scenarios, the following core conclusions are formed:

1) Aiming at the problem of photovoltaic power prediction caused by the lack of historical meteorological data, this study fully extracts the characteristics of historical data sets through algorithm-based meteorological data fusion technology, and accurately clusters various transitional weather types into different weather categories. Based on the meteorological fusion data set training model, the prediction value of the success rate is generated by inputting the weather characteristic parameters of the day to be predicted. This method significantly improves the prediction accuracy under complex weather conditions.

2) LSTM deep learning algorithm can effectively capture the long-term dependence characteristics in time series data. Compared with shallow BP neural network, it has significant advantages in historical data feature mining ability, to achieve higher precision photovoltaic output prediction.

References

[1] Gupta P, Singh R. PV power forecasting based on data-driven models: a review[J]. International Journal of Sustainable Engineering, 2021, 14(6): 1733-1755.

[2] Iheanetu K J. Solar photovoltaic power forecasting: A review[J]. Sustainability, 2022, 14(24): 17005.

[3] Liu Z, Du Y. Evolution towards dispatchable PV using forecasting, storage, and curtailment: A review[J]. Electric Power Systems Research, 2023, 223: 109554.

[4] Scott C, Ahsan M, Albarbar A. Machine learning for forecasting a photovoltaic (PV) generation system[J]. Energy, 2023, 278: 127807.

[5] Li Y, Song L, Zhang S, et al. A TCN-based hybrid forecasting framework for hours-ahead utility-scale PV forecasting[J]. IEEE Transactions on Smart Grid, 2023, 14(5): 4073-4085.

[6] Massidda L, Bettio F, Marrocu M. Probabilistic day-ahead prediction of PV generation. A comparative analysis of forecasting methodologies and of the factors influencing accuracy[J]. Solar Energy, 2024, 271: 112422.

[7] Son Y, Zhang X, Yoon Y, et al. LSTM–GAN based cloud movement prediction in satellite images for PV forecast[J]. Journal of Ambient Intelligence and Humanized Computing, 2023, 14(9): 12373-12386.

[8] Yan J, Hu L, Zhen Z, et al. Frequency-domain decomposition and deep learning based solar PV power ultra-short-term forecasting model[J]. IEEE Transactions on Industry Applications, 2021, 57(4): 3282-3295.

[9] Konstantinou M, Peratikou S, Charalambides A G. Solar photovoltaic forecasting of power output using lstm networks[J]. Atmosphere, 2021, 12(1): 124.

[10] Abdel-Basset M, Hawash H, Chakrabortty R K, et al. PV-Net: An innovative deep learning approach for efficient forecasting of short-term photovoltaic energy production[J]. Journal of Cleaner Production, 2021, 303: 127037.