Research on Photovoltaic power prediction technology Based on Machine Learning

Hongxia Li 1, *, Jianlin Li 2, and Yang Mi 1

1 Shanghai University Of Electric Power, 200090 Shanghai, China
2 Inverter Technologies Engineering Research Center of Beijing (North China University of Technology), 100144 Beijing, China
* Corresponding author: 2324015280@qq.com

Abstract. In recent years, the photovoltaic power generation has obvious intermittent, randomness and volatility, and high permeability photovoltaic will have a huge impact on the safety and stability of the grid. The prediction of photovoltaic power generation is to improve the quality of photovoltaic grid, optimize grid scheduling, and ensure the basic technology of power grid safety and stability. In order to improve the prediction accuracy of photovoltaic power generation, this article will comprehensively carding and compare from 3 dimensions: photovoltaic power generation and meteorological factor correlation analysis, similar day selection, prediction method based on machine learning, and summarize the advantages and disadvantages of various methods. Further research has been put forward accordingly.

1. Introduction
The solar energy has environmentally friendly, renewable characteristics, in which the design requirements is low in its development and utilization, so the photovoltaic power generation is concerned about domestic and foreign scholars. As of the end of 2020, my country's solar energy-filled power installed capacity reached 25.343 million kilowatts, new and cumulative installed capacity ranked first in the world. Due to the randomness, volatility of photovoltaic power generation, large-scale photovoltaic access will cause sever impact, dynamically and steady state properties such as active frequency, reactive active voltage, angular stability, small disturbance stability, power quality and distribution system protection[1]. To solve this problem, it is necessary to accurately predict photovoltaic power generation to reasonably arrange grid scheduling, reduce impact, and improve the security and reliability of the grid.

Traditional photovoltaic power generation power methods mainly have physical model laws and statistics. The physical model method does not require the historical operation of the power station, the photovoltaic output power is predicted by building solar radiation model, photovoltaic conversion model, circuit model, and inverter model, but the irradiation of radiation, the change of cloud, rainwater and environment, battery The impact of temperature and other factors can lead to short-term predictions are not accurate; the statistical law includes ARMA, ARIMA, Malco chain, time sequence prediction method, etc[2]. Due to the complexity of the photovoltaic power factor, there is no practicality when the historical data collection fee is charged[3]. In this case, the output power prediction method based on artificial intelligence can consider and compensate for the impact of the above various factors. Artificial Intelligence, Machine Learning, ML) algorithm can learn, training
models, find optimal parameters through existing large amounts of data, get good forecast results or classification effect\cite{4}.

With the development of predictive technologies, the related methods of photovoltaic power generation based on machine learning are increasing\cite{5-7}. Most references are improved according to 3 dimensions: the correlation analysis of photovoltaic power generation and meteorological factors, similar day selection and prediction method based on machine learning. This paper will explain these methods from these three aspects, working principle, The parameters of the performance of algorithm and their advantages and disadvantages, and finally, the proposal of future research is given according to the machine learning method predicted by photovoltaic power generation, and there is a certain reference value for subsequent research work.

2. Correlation analysis of photovoltaic power generation and meteorological factor

The correlation analysis of meteorological factors affecting photovoltaic power generation can be found, which can identify key meteorological elements that affect photovoltaic power generation, avoid excessive input parameter data, slow computational speed, is one of the methods of improving the prediction accuracy of photovoltaic power generation. There are three main related analytical methods: Pearson-related coefficient method, spEARMAN correlation coefficient method and principal component analysis.

(1) Pearson related factor

The pearson correlation coefficient is a calculation method for calculating and measuring a linear correlation between a linear continuous variable, or a linear correlation coefficient. Gets the weight of each cluster characteristics through the Pearson correlation coefficient.

\[
\rho = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}}
\]

Where \(X\) represents environmental factors, \(Y\) represents photovoltaic power generation power

\[
\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \quad \overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}
\]

(kW),

The value range is [-1, 1]. If it is more than 0, the x and y are positive correlation, otherwise it is negatively related.

(2) Spearman correlation coefficients

The existing method uses the pearson correlation coefficient to analyze the input variable, but the variable data when using the pearson correlation coefficient is required to meet the normal distribution. Since the spearman correlation coefficient is widely applicable, the spEARMAN correlation coefficient can be used as long as the two variables are paired, so this article uses the Spearman correlation coefficient for correlation analysis\cite{8}. 2 N-dimensional vector \(x\), \(y\)'s Spearman correlation coefficient calculation formula is as follows:

\[
\rho_s = \frac{\sum_{i=1}^{N} \left( R_i - \bar{R} \right) \left( S_i - \bar{S} \right)}{\sqrt{\sum_{i=1}^{N} \left( R_i - \bar{R} \right)^2} \sqrt{\sum_{i=1}^{N} \left( S_i - \bar{S} \right)^2}} = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N(N^2 - 1)}
\]

In the formula: \(R_i\) and \(S_i\) are the rank of the sequencing after the vector \(X\), \(Y\), respectively; the average rank of the vector \(x\), \(y\), respectively; \(n\) is the total number of observations.
The Spearman correlation is the real number in [-1, 1]. When $> 0$, the two variables are positively correlated, and that is negative. The larger, the higher the variable $x$ and $y$.

(3) Principal Component Analysis

Through the main component analysis method, a plurality of meteorological factors affecting the photovoltaic force are reduced, extracting less variables containing the original information, and simplifies the dimension of the model input vector under the premise of ensuring accuracy\[9-10\].

The original $n$ weather variables are represented by $X_1$-$X_n$, and $Y_1$-$Y_n$ is $n$ new variables extracted after the dimension, and the expression of the main components can be written through the feature vector:

$$
\begin{align*}
Y_1 &= a_1X_1 + a_2X_2 + \ldots + a_nX_n \\
Y_2 &= b_1X_1 + b_2X_2 + \ldots + b_nX_n \\
&\vdots \\
Y_n &= c_1X_1 + c_2X_2 + \ldots + c_nX_n
\end{align*}
$$

(3)

3. Similar day selection

Similar Day selection usually has 3 categories: direct method, K-Means clustering algorithm and gray correlation method.

(1) Direct method

The similar day of the direct method has 2: one is the date with the date of the day to be predicted, and the other is the date of the same period over the years. The reference base point solar radiation illuminance and its corresponding gradient value are selected as the similar variable, and the power generation power of each base value of the obtained similarity is finally obtained by screening\[11\].

(2) Kmeans

The use of the Kmeans algorithm to clinically class a historical training sample data affected by factors such as solar radiation intensity and ambient temperature, thereby achieving the similarities to select the photovoltaic power prediction day. The basic principle is as follows:

Step 1: Select K samples from the sample set as the initial centroid;

Step 2: Calculate each sample to the European downtown of each initial cluster center, select the nearest cluster center to form a K cluster, and update the K cluster according to the distance formula;

Direct until the maximum number of iterations or the centroid update amplitude is less than the threshold. Several all sample points are shown in fig1.
Grey Relational Analysis is a method of associating the relationship between the relationship between the gray system theory of gray system proposed in 1982, which judges the degree of correlation between the various factors based on the degree of similarity of the curve geometry of each factor. With gray correlation degree \( R \) indicates the overall correlation between the historical power generation day and the pre-sustain meteorological feature vector intermediate value, the closer the \( R \) value is, the more similar.

\[
\rho_i(k) = \frac{\min \left( \min_{i} x_i(k) - x_j(k) \right) + \rho \max \left( \min_{i} x_i(k) - x_j(k) \right)}{\min \left( \max_{i} x_i(k) - x_j(k) \right) + \rho \max \left( \max_{i} x_i(k) - x_j(k) \right)}
\]

Where: indicates the \( k \)th meteorological characteristic component of the day to predict the day; indicating the \( k \)th meteorological feature component of the \( i-I \) Historical Day;

4. Machine-based photovoltaic power prediction method

At present, there are usually three categories: support vector machines, neural network methods and multi-model fusion methods.

![Fig. 2. Typical improvement machine learning method](image)

4.1. BP

Artificial neural network has a non-linear, non-limit, very qualitative, non-convexity, etc., because it has its own ability to learn, Lenovo storage, and high speed, often used in terms of identification, prediction, etc. The short-term prediction method based on improved NARX neural network (Garbm-NARX) is established on the artificial neural network. Compared with traditional BP neural networks, NARX neural network adds time series learning capabilities, and uses genetic algorithms (Genetic Algorithm) UA) Parameter optimization of depth RBM, solving problems such as slow-in-depth RBM trap-local optimal convergence speed; reference[15]The photovoltaic power prediction model of the genetic algorithm (Ga) -furangular (RBF) neural network, combines the advantages of the fuzzy theory and the RBF neural network, which essentially imparts fuzzy input signals and blurred rights to the RBF neural network Value, make it all nodes and parameters have a specific meaning, namely corresponding to the affiliate function and the reasoning process of the fuzzy system, and improve the accuracy of the prediction; reference[16]Considering the influence of abnormal data on the prediction accuracy in the training sample, on the basis of a long-short memory network (LSTM) neural network, the forest (Iforest) algorithm is cleaned, filtering out the disabled data, Extreme data, error data; reference[17]Using digital twin technology, on the initial prediction
result of the GA-BP neural network, refer to the historical meteorological database, the initial forecast results are corrected, and the multi-dimensional assessment of the photovoltaic battery status and the surrounding environment, real-time, high-precision prediction power generation power; reference[18]A new type of photovoltaic power prediction system utilizing a deep volume neural network (CNN) structure and the input signal decomposition algorithm is proposed, and the signal decomposition algorithm for empirical mode decomposition (EMD) is used to decompose historical power signals into sub-assemblies. To extract depth functions, all input parameters will be converted to 2D feature mapping and feed to CNN input.

However, these methods require adequate historical data support to supply artificial neural network training. Typically, at least a continuous and complete data of the photovoltaic system output power is required for statistical regression. The lack of historical data will invalidate the relevant artificial intelligence prediction method. reference[19]Aiming at the lack of historical data, a digital twin model based on LSTM-oriented PV power prediction, the proposed digital twin model, can use the knowledge of the high-priced photovoltaic system in the historical data to assist the limited historical data limited photovoltaic system. Establishing power generation prediction digital twin models, having high prediction accuracy and fast training time.

4.2 SVM
SVM as a new learning machine, which is based on structural risk minimization guidelines, and has a good problem with the actual risk minimum, and has a good problem in realities such as small samples, nonlinear, local minority and high dimensional numbers. Solution, it has strong generalization. reference[20]Establish SVM prediction model, predicting data of different weather types, respectively; reference[21]The power prediction model based on SVM-based photovoltaic system is established. This model replaces the principle of experience risk in traditional machine learning methods with the principle of structural risks, and has excellent performance in small sample machine learning.;reference[22]Cluster analysis of training samples through the K mean algorithm, training support vector machines on all kinds of data obtained by clustering; reference[23]The support vector machine has a minimum multiplier support vector machine (LSSVM) and the segmentation support vector machine (PSVM), which proves that the prediction effects of the segmentation support vector machine is smaller. The multiplier support vector machine model is improved; reference[24]Establish a photovoltaic power prediction model of the vector machine, and use the remaining method for the optimization of nuclear parameters and penalty parameters for the support vector machine model constructed.

4.3 Multi-model fusion
In addition, there is a multi-model fusion method, a separate prediction model is often limited, and the resulting prediction results are not satisfactory in terms of prediction accuracy, and the weighting average method can calculate the weight of each model according to the error of the single predictive model. To assign a large weight to a small error, a smaller weight is allocated to the error of the error, amplify the advantages of the fused prediction model and weaken the influence of its disadvantages on the prediction, and retain the advantages of each model.
Table 1. Typical improvement machine learning method

| Method                  | References | Main Improvements                                                                                                                                                                                                 | Advantages and Disadvantages |
|-------------------------|------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------|
| XGBoost-Light GBMNBoost | 25         | Three integrated learning models perform dynamic weight fusion to retain feature correlation, and introduce correction factors to reduce the influence of error accumulation on prediction accuracy. | Higher accuracy, cumbersome operation |
| CEEMD-PE-LSTM           | 26         | Decomposing the original sequence of the data into a plurality of inherent modal components with CEEMD, reconstructed with PE; LSTM is modeled by the reconstructed subsequences. | Higher accuracy, cumbersome operation |
| Chaos-EEMD--PFBD-GA-BP  | 27         | EEMD and PFBD are used to optimize the chaotic characteristics, build a GA-BP neural network prediction model, perform single step and three-step prediction of photovoltaic power generation. | Infusing data with deep excavation data, extracting well, predictable polymerization components, predictive accuracy |
| LSTM-FC                 | 28         | Dual branch input allows it not only to consider the impact of meteorological data on power generation, but also considers time continuity and periodic dependence. | Higher accuracy, cumbersome operation |
| VMD-SE-LSSVM            | 29         | VMD technology decomposes historical photovoltaic power generation into a series of limited bandwidth, with the least squares support vector machine to predict photovoltaic power generation power and error, respectively, power value after the error compensation is the final prediction result. | Avoid the impact of modal aliasing and noise impact, taking into account the impact of weather factors and time dimensions on predicted values |

Reference[25] Extreme Gradient Boosting XGBoost, Gradient Single Sampling, Goss Light GBM and Mutual Exclusive Bind, EXCLUSIVE FEATURE BUND-LING, EFB) NGBOST. The integrated learning model performs dynamic weight fusion to retain feature correlation, and introduce the correction factor to reduce the impact of error accumulation on prediction accuracy;

Reference[26] Combining the overall empirical modal decomposition (CEEMD), arrangement (PE), LSTM neural network, using CEEMD to decompose the original sequence of power generation and influencing factor data into multiple inherent modal components, and use the algorithm algorithm to modality The component is reconstructed; finally the reconstructed subsequences are predicted by LSTM modeling, and then superimpose the child sequence prediction result to obtain the predicted value, compared to a single model, the power prediction accuracy is significantly improved;

Reference[27] A photovoltaic power generation power combination prediction method based on Chaos Empirical Mode DecomPosition (EEMD) and the PFBD) and GA-BP Neural Network is proposed. First, on the basis of the photovoltaic power generation power sequence phase spatial reconstruction, EEMD and PFBD are used to optimize the chaotic characteristics, and the GA-BP nerve is constructed by GA-optimized BP neural network (bpnn). Network prediction model, photovoltaic power generation power combination step and three-step prediction showing good prediction performance;

Reference[28] The LSTM-FC depth learning algorithm composed of LSTM and full connection (FC) layer is proposed. The two branch input of the model makes it not only considering the impact of meteorological data on power generation, but also considers time continuity and periodic dependence, thereby increasing a certain degree of prediction accuracy. And SVM, GBDT, GBDT, Generalized Regression Neural Network (GRNN), feedforward neural network (FFNN) and LSTM and other models, the accuracy is higher than the model in other input forms;

Reference[29] A prediction method based on VMD-SE-LSSVM and iterative error correction is proposed. This method first uses...
variational modal decomposition (VMD) technology to decompose historical photovoltaic power generation into a series of limited bandwidth, avoiding the impact of modal aliasing and noise impact; then predicting photovoltaic power and error using LSSVM method. The power value after the error compensation is the final prediction result. In addition, this paper also uses the sample entropy (SE) principle to quantify the weather type as a feature vector input SVM participation in prediction, and taking into account the impact of weather factors and time dimensions on predicted values.

5. Conclusion
In summary, whether it is traditional physical model or statistical law or existing machine-learning-based photovoltaic power prediction technology, there will be some problems that are still resolved, and how to design an effective way in multiple scenes to predict photovoltaic Power generation power, reducing the uncertainty of prediction, is still a challenge. With wide demand for photovoltaic power prediction, there will give several following recommendations to provide reference for related researchers.

(1) If the historical data is sufficient, it can be predicted according to the basic machine learning method. If historical data has an incomplete, abnormal and data and complex issues, the multi-model fusion method is required to apply different advantages to various models to Improvement in different scenarios, improve prediction accuracy.

(2) In order to improve the accuracy of the prediction, it should be more concerned and actively explore more universal parameter optimization algorithm applications in photovoltaic power generation.

(3) With the increasing computer technology, the artificial intelligence algorithm based on machine learning is widely used in the field of prediction, based on deep learning, strengthening learning and edge computing technology, which should be fully utilized, and play the role of intelligence algorithm.

Acknowledgments
Supported by: Beijing Natural Science Foundation Project (NO: 21JC0026)

References
[1] Ding Ming, Wang Weisheng, Wang Xioli, Song Yunting, Chen Chi, Sun Ming. Related on the Impact of Power System in Power System [J]. Chinese Journal of Motor Engineering, 2014, 34 (01): 1-14.
[2] Qian Zhen, Cai Shibo, Gu Yuxi, Tong Jianjun, Bao Guanjun. Summary of Study on PV Generation Power Prediction Method [J]. Mechanical Engineering, 2015, 32 (05): 651-659.
[3] Yang Li Feng. Summary of application research of artificial intelligence algorithm in photovoltaic power generation [J]. Solar, 2020 (08): 30-35.
[4] Jiao extraordinary. Research on short-term photovoltaic power generation power prediction method based on machine learning [D]. Zhongyuan Institute of Technology, 2021.
[5] Lai Changwei, Li Jinghua, Chen Bo, Huang Yujin, Wei Yanyang. Summary of Research on Prediction Technology of Photovoltaic Power Generation [J]. Journal of Electrician Technology, 2019, 34 (06): 1201-1217.
[6] Wang Hongkun, Ge Lei, Li Hongwei, Chi Fujian. Summary of Characteristics Analysis and Forecasting Method of Distributed Photovoltaic Power [J]. Power Construction, 2017, 38 (07): 1-9.
[7] Li An Shou, Chen Qi, Prince Cel, Li Tiecai. Summary of Power Forecasting Method of Photovoltaic Power Generation System [J]. Electrical Transmission, 2016, 46 (06): 93-96.
[8] Tian Jian Gang, Zhang Pei, Peng Chunhua, Time, Wang Tieqiang, Yin Rui, Wang Yifeng. Photovoltaic power generation of photovoltaic power generation Based on the short-term memory neural network [J]. Modern Power, 2020, 37 (06): 629-638.
[9] Wang Wei, Huang Ke, Zheng Yizhui, Li Li, Shao Fengpeng, Jia Likai, Xu Qingshan. Based on Firefly Algorithm? Based on PV Generation Power Combination Prediction of Generalized Regression Neural Network [J]. Technical Technology, 2017,41 (02): 455-461.

[10] Wang Wei, Huang Ke, Zheng Yizhui, Li Li Xue, Lang Yongbo, Wu Hao. Conturated PVD Prediction Method Based on PN / PCA / SS-SVR [J]. Power System Automation, 2016,40 (17): 156-162.

[11] Wang Junhui, Li Min, Chang Peng Bo, Liu Ning. Prediction based on similar days of photovoltaic power error correction [J] grid and clean energy, 2020,36 (11): 134-138.

[12] Liu Jialin. Research on short-term photovoltaic power generation power prediction based on Similar Day Theory and PCA-IPSO-Elman Model [D]. Lanzhou Jiaotong University, 2019.

[13] He Yajing, Wang Deng. Summary of the application of artificial neural network in photovoltaic [J]. Solar, 2019 (01): 17-20.

[14] Zhu Xi, Ding Yunlong, Guo Li, Shi Haoqi. Based on the short-term prediction of photovoltaic power generation based on improved NARX neural network algorithm [J]. Journal of Wuhan University (Science Edition), 2020, 66 (05): 505-511.

[15] Ye Lin, Chen Zheng, Zhao Yongjiong, Zhu Qianwen. Photovoltaic power prediction model based on genetic algorithm - fuzzy radial nerve network [J]. Power System Automation, 2015,39 (16): 16-22.

[16] Zhang Qian, Ma Wi, Li Guo Li, Ma Jinhui, Dingjijin. Lopement of Frequency Domain Decomposition and Depth Learning Algorithm in Short Term Load and Photovoltaic Power Prediction [J]. China Motor Engineering Journal, 2019,39 (08): 2221-2230 +5.

[17] Sun Rongfu, Wang Longyang, Wang Yulin, Ding Rong, Xu Hai Xiang, Wang Jingran, Li Qiang. Themboic Power Based on Digital Twins [J]. Print Technology, 2021,45 (04): 1258-1264.

[18] Korkmaz Deniz, Acikgoz Hakan, Yildiz Ceyhun. A Novel Short-Term Photovoltaic Power Forecasting Approach based on Deep Convolutional Neural Network [J]. International Journal of Green Energy, 2021,18 (5):

[19] Shi Kai, Zhang Dongxia, Han Shaoting, Jiezhihun. Photovoltaic Power Prediction Digital Twin Model Based on LSTM and Migration Learning [J / OL]. Power Grid Technology: 1-11 [2021-08-10]. https://doi.org/10.13335 /J.1000-3673.pst.2021.0738.

[20] Zhao Xinyu. Research and implementation of power forecasting of photovoltaic power generation system [D]. North China Electric Power University, 2012.

[21] Jiang Enyu, Qi Liang, Xia Yuan Han, Mi Yang, Deng Wei. Photovoltaic Power Prediction Based on Support Vector Machine [J]. Journal of Shanghai Electric Power Institute, 2015,31 (06): 511-513 + 524.

[22] Zhang Yujin, Yang Lan, Ge Shuang Ye, Zhou Hangxia. Prediction of Short-term Photovoltaic Power Generation Based on KMeans-SVM [J]. Power System Protection and Control, 2018,46 (21): 118-124.

[23] Shi Jie, Liu Xiaofei. Research on the optimization of new energy power prediction algorithm [J]. Power Generation Technology, 2019,40 (01): 78-82.

[24] Yu Qiu Ling, Xu Changqing, Li Shan, Liu Hong, Song Yi, Liu Xiaou. Short-term photovoltaic power prediction based on fuzzy clustering and support vector [J]. Power System and Automation, 2016, 28 (12): 115-118 + 129.

[25] Xin Junjie, Xu Liang, Li Yongjie, Li Wenlei. Dynamic correction prediction of multi-model fusion photovoltaic power generation based on integrated learning [J]. Sensors and Microsystems, 2021,40 (04): 117-121.

[26] Li Bing Chen, Yu Hui, Liu Jingyu. Prediction of Short-term Photovoltaic Power Generation Based on KMEANS and CEEMD-PE-LSTM [J]. Hydropower Energy Science, 2021,39 (04): 204-208.
[27] Wang Yufei, Fu Yuchao, Xue Hui. Based on CHAOS-EEMD-PFBD Decomposition and Over Short-term Prediction Law of PV Generation of GA-BP Neural Network [J]. Journal of Solar Energy, 2020, 41 (12): 55-62.

[28] Li Yang, Ye Feng, Liu Zihao, Wang Zhijian, Mao Yupeng, Ma Xin. A Short-Term Photovoltaic Power Generation Forecast Method Based on LSTM [J]. Mathematical Problems in Engineering, 2021:

[29] Yu Jingyang, Zhao Yizhen, Yang Ningning, Yue Tongjun, Gao Chunyang. Prediction of Photovoltaic Power Power Based on VMD-SE-LSSVM and Iterative Error Correction [J]. Journal of Solar Energy, 2020, 41 (02): 310-318.