Photovoltaic Power Prediction and SHAP Interpretability Analysis Based on Multi-Model Comparative Learning

Xiaoxi Dai #, Yicheng Yang #, *

College of Information Science and Technology, Hebei Agricultural University, Baoding, China, 071001

* Corresponding Author Email: yangyichenghb@163.com

#These authors contributed equally.

Abstract. In recent years, photovoltaic (PV) power generation has been developing rapidly, but its power instability poses a challenge to the stable operation of power systems. In this study, three machine learning models, CatBoost, Random Forest, and Support Vector Regression (SVR), were used for PV power prediction, and it was found that the SVR model performed the best, with the highest fit between predicted and true values ($R^2 \approx 0.95$) and the smallest fluctuation in residual error (-50% to 50%). Through SHAP analysis, the study reveals that illumination intensity is the most critical variable affecting PV power generation with a contribution of more than 70%. Additional machine learning models and optimization algorithms can be further explored in the future to improve prediction performance. In summary, this study provides an effective combinatorial machine learning method for high-precision PV power prediction, revealing the dominant role of key factors such as illumination intensity on the prediction results. Future research directions include optimizing the model structure and integrating multi-source data to further improve the accuracy and reliability of PV power prediction.

Keywords: machine learning, photovoltaic power, model prediction, SHAP.

1. Introduction

Driven by the "dual-carbon" strategy, PV power generation, as an important force in the energy transition, is developing rapidly around the world[1]. In order to reduce dependence on non-renewable energy sources, optimization of the energy mix has become a top priority[2]. The importance of PV as a core component of renewable energy is self-evident[3]. However, the stochastic and fluctuating nature of PV power generation poses a huge challenge to grid stability[4]. China's PV industry continues to expand[5], however, its power generation efficiency is highly dependent on weather conditions, which poses a significant challenge to the stable operation of the power grid[6]. Therefore, accurate prediction of PV power generation not only helps to ensure the stability of the power grid, but also reduces energy waste, promotes the further development of the PV industry, and enhances the power system's ability to absorb new energy sources[7].

At present, PV power prediction methods are mainly divided into two categories: indirect prediction models and direct prediction models [8]. Indirect prediction models extrapolate power generation by predicting solar radiation intensity and combining it with the physical parameters of PV panels, which are more accurate but complicated modeling process [9]. Direct prediction models, on the other hand, utilize historical data directly and make predictions through machine learning algorithms, which has the advantage of being easy to operate [10]. However, existing studies are still deficient in feature selection, model optimization, and interpretability of prediction results [11]. For example, the limitations of a single machine learning model in mining PV feature data have not yet been adequately addressed [12], and lack of in-depth analysis of the mechanisms that influence key features [13]. The existence of these problems limits the further improvement of prediction accuracy and affects the reliability of prediction models in practical applications [14].

The research innovation of this paper is to construct a more accurate input feature set by deeply analyzing the correlation between PV data features and meteorological factors [15]. In the data preprocessing stage, we strictly dealt with outliers and missing values to ensure data quality. To

address the limitations of a single machine learning model in feature mining, this paper uses a variety of machine learning models (including CatBoost, Random Forest, and Support Vector Regression models) to conduct comparative experiments, and ultimately finds that the Support Vector Regression (SVR) model performs particularly well in predicting the effect of [16]. In addition, this paper introduces the SHAP (SHapley Additive exPlanations) analysis method, which provides an interpretable analysis of feature importance and reveals that illumination intensity is the feature that has the greatest impact on the prediction results, with a contribution of more than 70 percent [17]. These findings provide new ideas for PV power prediction, especially in regions with high illumination intensity, where the SVR model is able to effectively capture complex nonlinear relationships and maintain good prediction performance when dealing with high-dimensional data containing multiple meteorological factors and historical data [18]. The research in this paper not only provides a new technical means for PV power prediction, but also provides a strong support for the stable operation of the power grid and the efficient utilization of energy.

2. Analysis of data sets

2.1. Description of data sources

The dataset on which this study relies is a collection of publicly available information dedicated to the prediction of PV power, derived from experimental tests, and contains a number of variables closely related to PV power generation, which adequately reflect the parameters of power generation formation, and detailed information about the dataset and links to access it can be found at https://www. heywhale. com/mw/dataset/637c65af7b4e5f9b3d5c2fee/content.

2.2. Analysis of data

Figure 1 illustrates the frequency distributions of the variables, which all approximately follow a normal distribution, with more dispersed data distributions for board temperature, ambient temperature, illumination intensity, and voltage, and relatively concentrated distributions for the other eigenvalues. In a specific analysis of the temperatures, it was found that the board temperatures were overall higher than the ambient temperatures, which may indicate the superiority of the photovoltaic boards in terms of thermal conductivity or heat absorption properties. Comparison in terms of voltage shows that the voltage peaks in zone A are significantly higher than those in zones B and C, which may be related to higher settings of certain parameters (e. g. current intensity) in zone A, whereas the voltage distribution parameters in zones B and C exhibit smoother characteristics.

Regarding the PV conversion efficiency, although the most common conversion efficiency in all four zones of ABCD is 20%, the distribution of the conversion efficiency shows a clear left skewness, and the observations indicate that the overall conversion efficiency fluctuates more despite the fact that the most frequent conversion efficiency stabilizes at 20%.

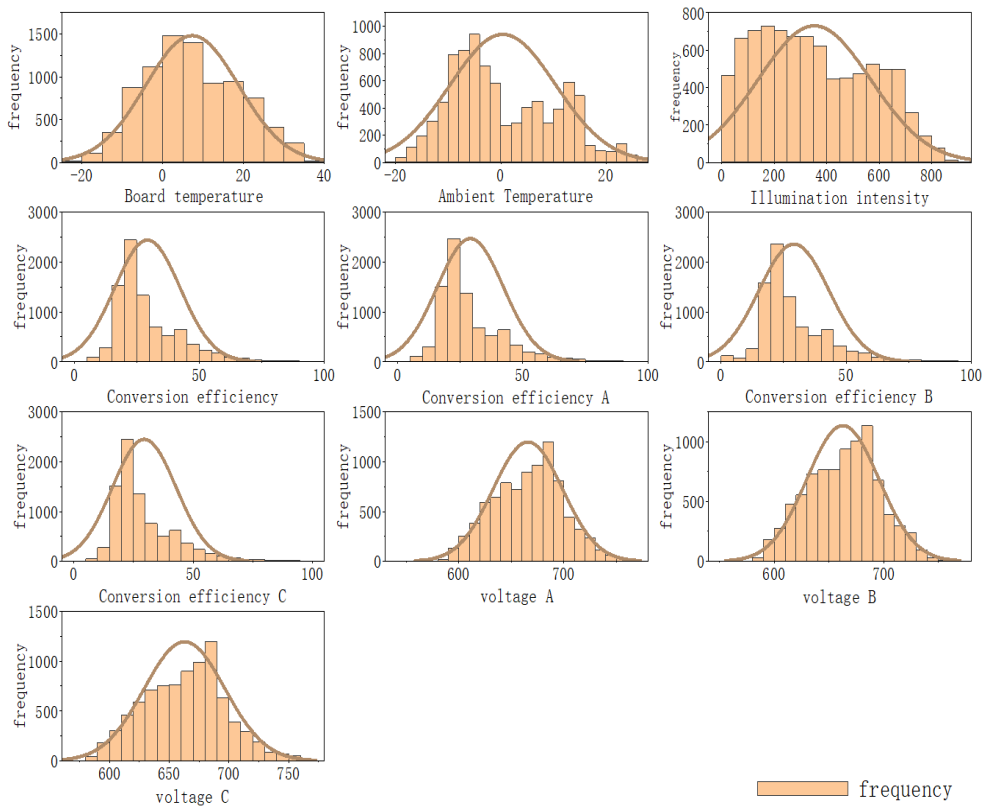

Figure 1. Histogram of frequency distribution

2.3. Data preprocessing

Considering the data parameters of the original dataset and, the features including board temperature, ambient temperature, and illumination intensity were selected among the columns of features related to the target power analysis, and the dataset was divided into a training set and a validation set. In order to reduce the risk of overfitting when validating the model performance at a later stage, the validation set was divided to 20% of the total data. The data was also normalized to increase the convergence speed of the machine learning algorithm and to improve the performance of the model. The formula is:

$$X' = \frac{X - \mu}{\sigma} \tag{1}$$

where μ is the sample mean and σ is the sample standard deviation.

First, it is proposed to fit and transform the features and target variables of the training data separately, and then transform the validation dataset using the same mean and standard deviation, which is used to ensure that the training and validation sets are treated in the same way.

3. Research process and results

3.1. PV power prediction model based on machine learning

The aim of this experiment is to build a PV power prediction model and visualize the output results through different machine learning methods to explore the most suitable prediction model for this model.

Figure 2 illustrates the exact flow of this experiment. First, it is proposed to carry out the necessary data cleaning for outliers and missing values in the observation data to ensure the completeness and accuracy of the data. The correlation between the variables is then analyzed through the visualization

of basic information about the statistics of the data, providing a basis for subsequent model construction. After the data normalization process, the data is divided into training and test sets and the hyperparameter network is defined to optimize the hyperparameter set using grid search. Subsequently, the Support Vector Regression (SVR), Catboost and Random Forest models were initialized and the corresponding hyperparameters were set, respectively. The models were trained under the optimized parameter set and interpretive analysis was performed by SHAP (SHapley Additive exPlanations) method to fit the contribution of each feature to the model prediction results. Finally, the results of each model are output and visualized, including the prediction results of Catboost, Random Forest and SVR, to provide a basis for model performance evaluation and comparison.

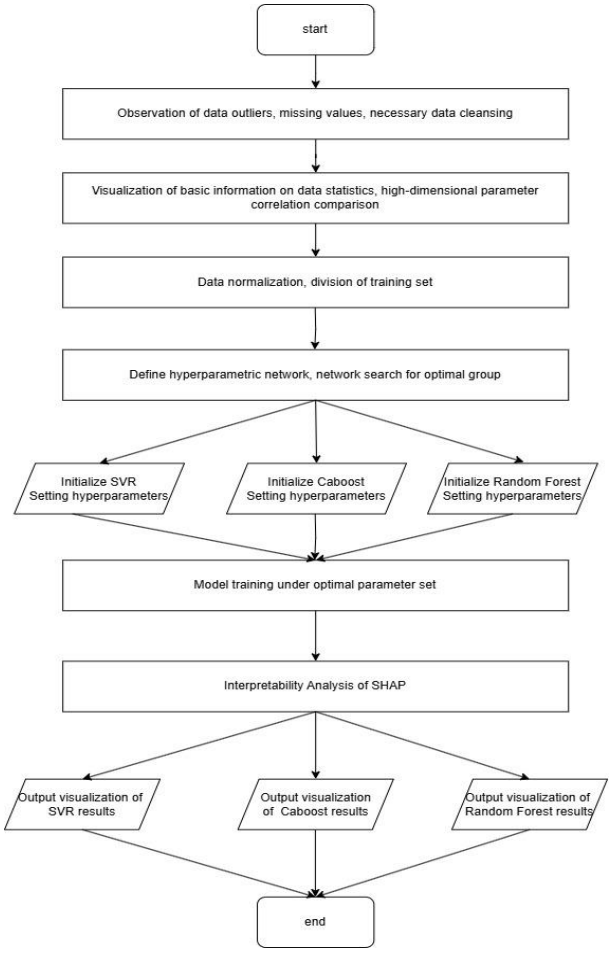

Figure 2. Flowchart of PV power prediction

3.1.1 CatBoost model

CatBoost uses the information of target variables for goal coding, reduces the dimensionality of category features to avoid the dimension explosion problem, and optimizes the model with the statistical information of target variables. Therefore, CatBoost becomes an ideal candidate algorithm when dealing with tasks involving a large number of class features, such as PV power prediction. The core idea of target coding is to code categories by calculating the average of the target values for each category. For a category feature X, its target code can be calculated by the following equation:

$$X\_\{encoded\} = 1/N\_k\backslash sum\_\{i \in C\_k\} y\_i \quad (2)$$

where C_k is all the samples corresponding to category k; N_k is the number of samples in category k; and y_i is the true value of sample i in category k.

The algorithm updates the model by employing a loss function minimization strategy in each round of iterations, and its optimization process is implemented through gradient descent. Each round of

update adjusts the parameters of the learner model by calculating the current loss function in a gradient. The residual calculation follows equation:

$$r\_i = y\_i - F(x\_i) \quad (3)$$

where r_i is the residual of the i-th data point, y_i is the true value, and F(x_i) is the current prediction of the model.

Repeatedly calculating the fitting residuals with updating the model, from which multiple base learners are iteratively trained until a predetermined number of iteration terminations are reached. Each round of iteration generates a new base learner, which are combined in a weighted manner to form a robust final model. The current model is updated using the newly trained base learners. In CatBoost, the update is usually a weighted update, i. e:

$$F\_m(x) = F\_\{m - 1\}(x) + \eta \cdot h\_m(x) \quad (4)$$

where F_{m-1}(x) is the model of the previous step (m-1st model), h_m(x) is the predicted value of the current base learner (mth decision tree), and η is the learning rate, which controls the magnitude of the contribution of each base learner.

3.1.2 Parameters of the random forest model

Random forest is an integrated learning model that consists of multiple decision trees. It constructs decision trees through random resampling and node random splitting techniques and integrates the results using a voting mechanism. Random forests can effectively analyze complex feature interactions, are robust to noise and missing data, and learn faster. Its unique variable importance measure provides a powerful tool for feature selection in high-dimensional data and performs well in datasets with high feature relevance.

After data preprocessing, the dataset is divided into training set and validation set, and each decision tree is trained by Bootstrap sampling a randomly selected subset from the training set and constructed based on the feature random selection strategy. Although individual decision trees may perform weakly, the overall model performance improves significantly after integration.

In this model, the original training set has 6887 samples. Each sample is randomly selected by Bootstrap sampling with a probability of 1/6887 to be selected by sampling, and put back after each sampling so that each tree will get a training subset with an upper limit of 6887.

Using the Filter method means that only a portion of the features are considered when the tree is split, and each tree makes decisions based on a different subset of features, thus reducing the overfitting that results from the accumulation of highly correlated feature factors that are constantly selected.

The training set for this model has 11 features. In a traditional decision tree, all features are usually evaluated to select the best split feature, whereas in a random forest, the split only considers the k features that are randomly selected.

3.1.3 Support vector regression

The core idea of SVR is to map data from a low-dimensional space to a high-dimensional space by means of a kernel function, and to find a linear regression model that predicts successive values such that the vast majority of the deviations from the training data points to the function are within a predetermined threshold. With this mapping, SVR is able to find a suitable function to fit the data in high dimensional space without the need for complex nonlinear fitting directly in the original space.

First, SVR needs to define the regression model in the original space i. e. find a function f(x)=w^Tx+b such that the error between the predicted value f(x) and the actual value y is as small as possible. Then, to reduce the error, SVR introduces an ε-insensitive loss function that penalizes the model when its prediction error exceeds the ε threshold. In order for the function f(x) to be as smooth as possible, the penalty parameter C will determine how hard the error is penalized by controlling the support vector of the model. To deal with nonlinear regression problems, SVR uses a kernel function to map the data to a high-dimensional space in order to find the optimal regression

function using linear regression methods assuming that the data in the high-dimensional space may become linearly differentiable.

The optimization objective of SVR is to minimize the following loss function:

$$\min_{w,b,\varepsilon_i,\varepsilon_i^*} \left(\frac{1}{2}||w||^2 + C\sum_{i=1}^{m}(\varepsilon_i + \varepsilon_i^*)\right) \quad (5)$$

Where $||w||^2$ is the weight of the regression function; ε_i and ε_i^* are slack variables that measure those errors that fall outside the range of ε; and C is a penalty parameter that controls the balance between the complexity of the model and the error.

Finally, SVR obtains the optimal parameters w and b by solving a convex optimization problem to obtain a regression function that can fit the data optimally.

3.1.4 Model Hyperparameter Settings

Table. 1. Optimal hyperparameters for three models

model	hyperparameterization	sense	optimum value
CatBoost	depth	Tree depth when training a decision tree	4
	learning_rate	Learning rate for model updating	0. 0475
	n_estimators	Number of trees	19000
Random forest	max_depth	Maximum depth per tree	[34,35]
	n_estimators	Number of trees	[415,418,419,420,421,422,423,424,427,430,]
	min_samples_split	Minimum number of samples required to partition internal nodes	2
	min_samples_leaf	Minimum number of samples required for leaf nodes	1
SVR	C	penalty parameter	150
	epsilon	The model's allowable margin of error	0. 00001
	kernel	Radial Basis Function (RBF) Kernel	rbf

3.2. Analysis of projected results

In this paper, the dataset is divided according to the ratio of 8:2, 80% of the dataset is used for training and 20% of the dataset is used for testing. The performance of the three models in PV power prediction is shown in Figure 3. The horizontal axis of the graph is the true power value (W) and the vertical axis is the predicted power value (W). CatBoost model (Figure 3, (a)), the goodness of fit between predicted and actual values, and most of the data points are distributed on the ideal diagonal, indicate that the model has a high prediction accuracy. Random forest model (Figure 3, (b)), the fit between predicted and true values is also high, but the distribution of data points is slightly dispersed compared to the CatBoost model, indicating that the prediction accuracy of this model is slightly lower than that of the CatBoost model. sVR model (Figure 3, (c)), the fit between predicted and true values is the highest, and the distribution of data points is almost completely on the ideal diagonal, indicating that this model has a very high prediction accuracy. indicating that the model has a very high prediction accuracy.

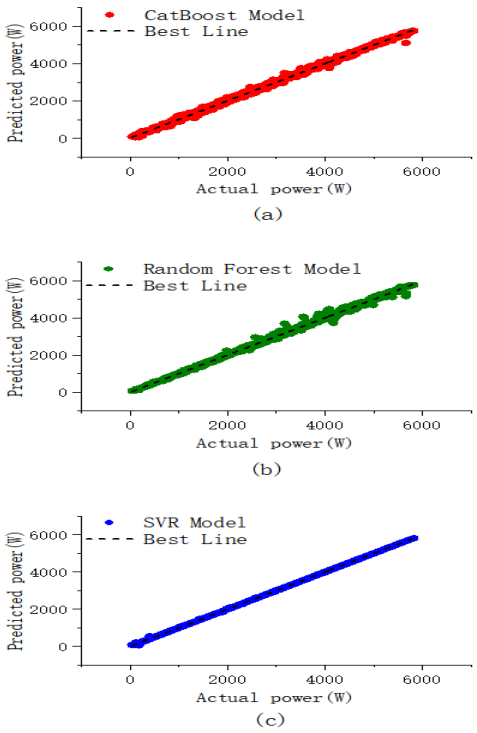

Figure 3. Plot of exact value versus error value

Figure 4 demonstrates the prediction errors of the three models. The horizontal coordinates in the figure indicate the samples in the testing phase, the vertical coordinates indicate the errors, the larger the absolute value of the vertical coordinates indicates the larger the prediction errors of the sample points, and the closer the absolute value of the vertical coordinates of the model as a whole is to zero, it means that the prediction result of the model is relatively good.

The CatBoost model (Figure 4, (a)) has a relatively uniform error distribution, but there are some large error points. The percentage of residual error fluctuates between -100% and 100%, indicating that the model has a large prediction error on some samples.

The random forest model (Figure 4, (b)) has a relatively centralized error distribution, but there are some large error points. The percentage of residual error fluctuates between -100% and 50%, indicating that the model has a large prediction error on some samples.

The SVR model (Figure 4, (c)), on the other hand, has the most concentrated distribution of errors and smaller error values. The percentage of residual error fluctuates between -50% and 50%, indicating that the model has a small prediction error on most samples.

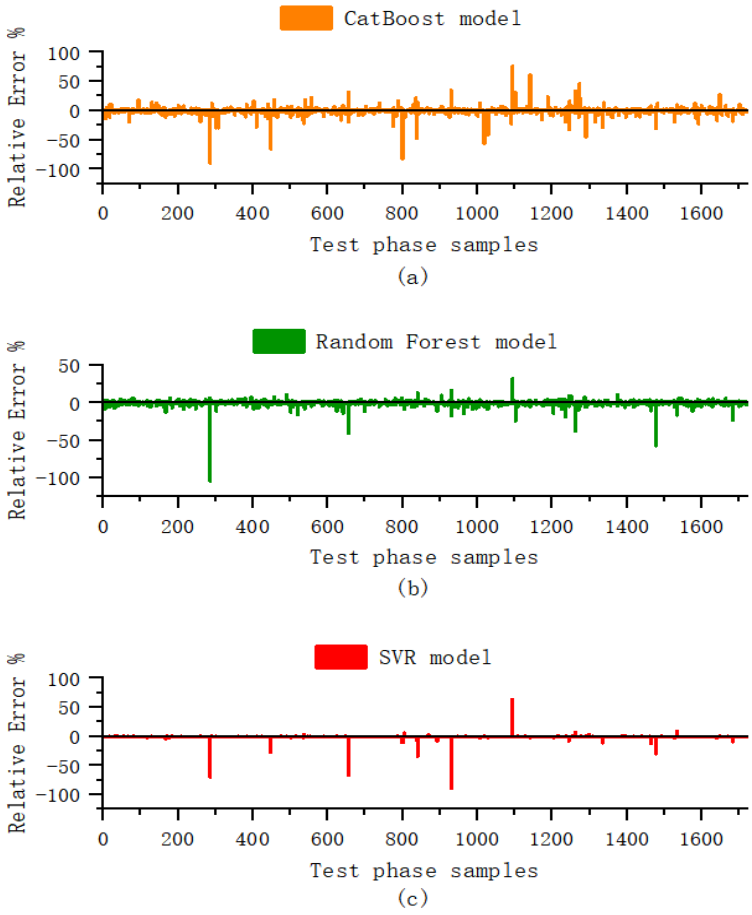

Figure 4. Error plots for the three models

Figure 5 demonstrates the line graphs of the predicted values of the three models compared with the true values, the horizontal coordinate in the graph indicates the number of test samples, and the vertical coordinate indicates the PV power, when the line graph of the true values and the line graph of the model's predicted values have a higher degree of overlap, it means that the model's prediction results are relatively good.

The CatBoost model (Figure 5, (a)) has a relatively high degree of overlap between the predicted and true value folds, and there are some points at some inflection points and on the folds that do not overlap, indicating that there are some sample points that still have a certain degree of error, but the error is relatively small, suggesting that the model has better prediction results.

The predicted values of the random forest model (Figure 5, (b)) also have a relatively high degree of overlap with the true value folds, but there are some large error points, especially in the samples of 10, 80, 170 or so the error is more obvious, in the samples of 80 where the error reaches a maximum, there is about 500 watts of error, overall the error is relatively small, compared to the CatBoost model is worse.

The predicted values of the SVR model (Figure 5, (c)) have a very high overlap with the true value folds, almost no error occurs, and the error range is small, concentrating within 10 watts, which is significantly better than the CatBoost model and the Random Forest model, indicating that the model has a very high prediction accuracy.

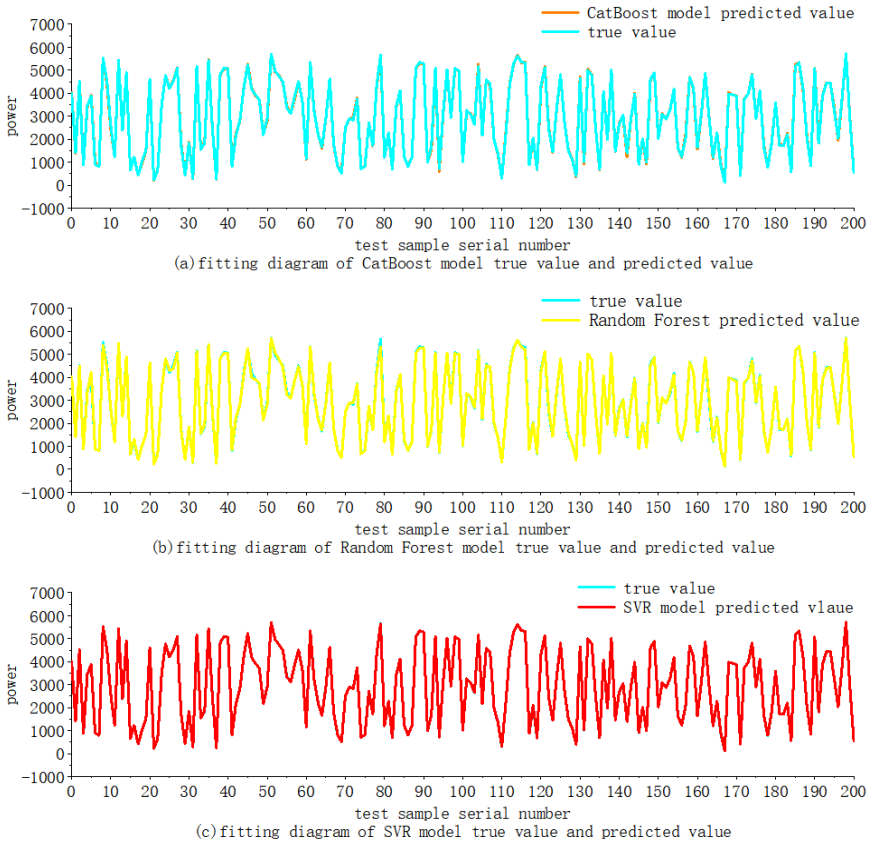

Figure 5. Comparison of the three model predictions with the true value

Therefore, by comparing the prediction results and error distribution of CatBoost, Random Forest and SVR models, it can be found that the SVR model has the best performance in PV power prediction, with the highest fit between predicted and true values, the smallest percentage of residual error, which only fluctuates between -50% and 50%, and the predicted value has a high degree of overlap with the true value. The CatBoost model and the Random Forest model also have high prediction accuracy, but compared to the SVR model, the error distribution is more dispersed and there are some larger error points. Among them, SVR model is the most suitable model for PV power prediction in this experiment.

4. SHAP Model Interpretability Analysis

4.1. Distribution of SHAP values

Figure 6 shows the distribution of SHAP values for each feature, each point represents the SHAP value of a sample, and the color indicates the magnitude of the feature value (red indicates high values, blue indicates low values). From the figure, it can be seen that Illumination Intensity has the greatest impact on the model output, with a wide distribution of SHAP values, and high illumination intensity (red) usually corresponds to positive SHAP values, indicating that the higher the illumination intensity, the greater the predicted PV power. Features such as Conversion Efficiency B, A, and C also have some effect on the model output, Board Temperature has a smaller effect on the model output, and other features such as Voltage C, Ambient Temperature, Voltage A, ID, and Voltage B have an almost negligible effect on the power prediction.

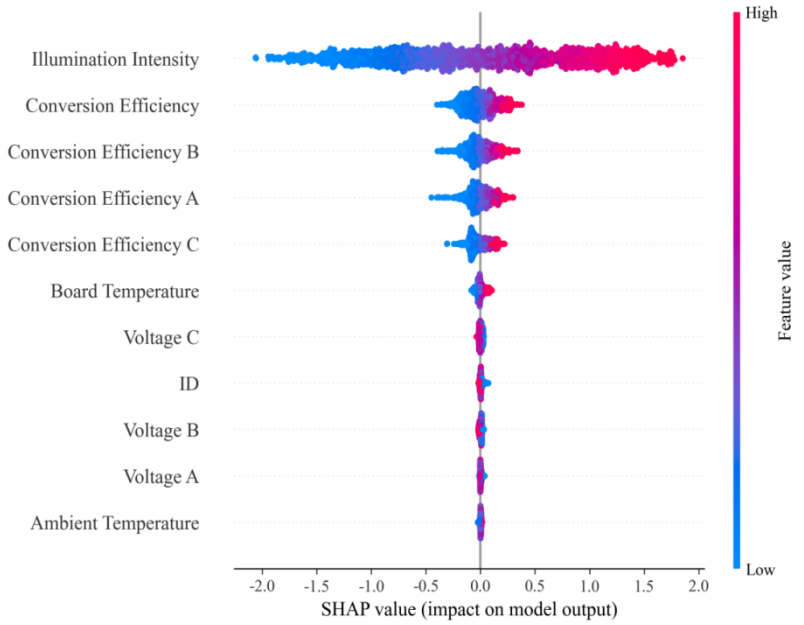

Figure 6. Distribution of SHAP values

4.2. Characteristic Importance Map

Figure 7 demonstrates the absolute value of the average SHAP value for each feature, which can reflect the importance of each feature on the model output. Illumination Intensity is the most important feature, Conversion Efficiency B, A, and C have some effect on the model, Board Temperature has a small effect on the model output, and the other features have almost negligible effects. This agrees with the conclusion drawn from the distribution of SHAP values.

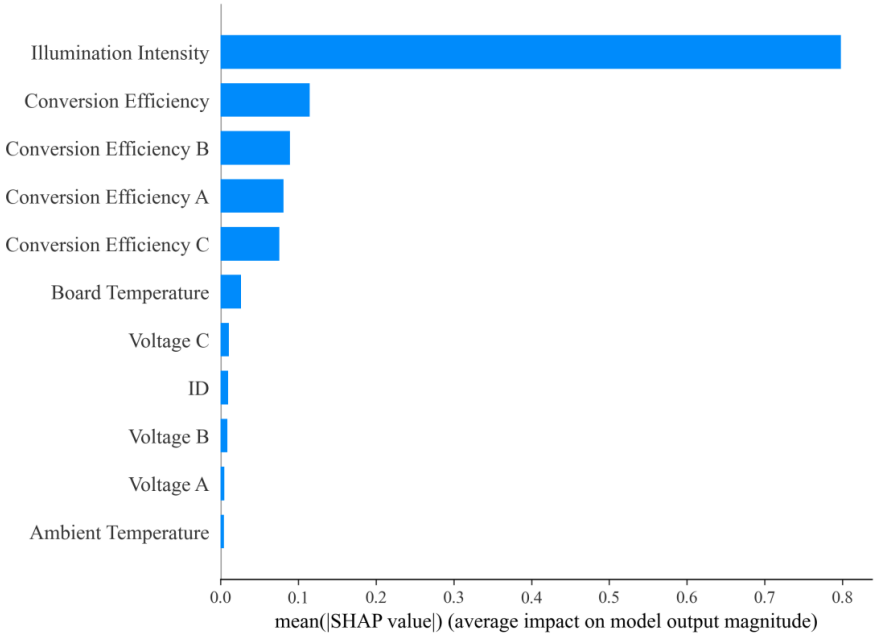

Figure 7. Characteristic importance map

4.3. Interpretability analysis

Figure 8 shows the marginal contribution of each feature to the model's predictions, using color coding to show the effect of each feature on the instances. Illumination Intensity has a significant positive effect on most of the samples, Conversion Efficiency B, A, and C have some positive effect

on most of the samples, and Board Temperature has a negative effect on some of the samples, and the other features have a small effect on most of the samples.

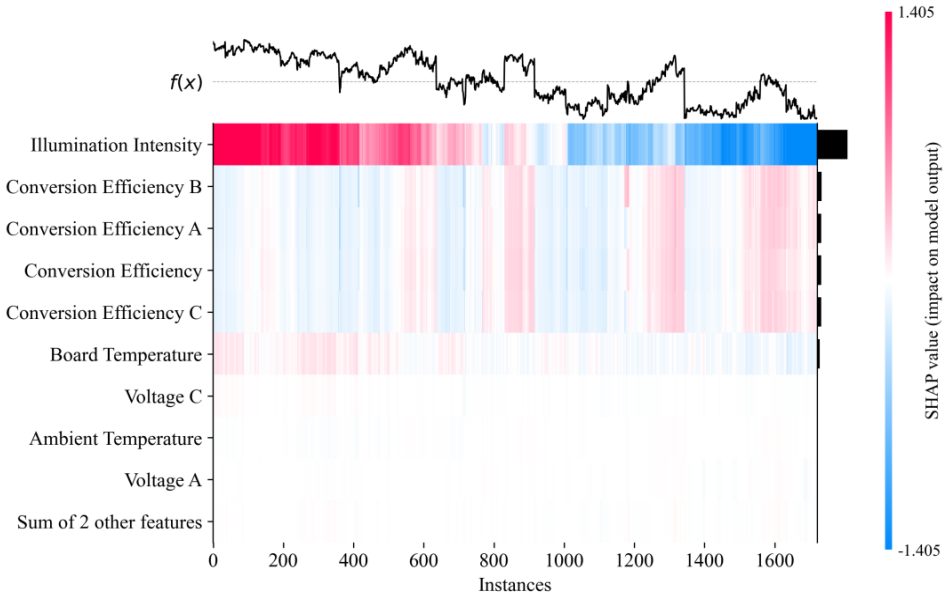

Figure 8. Heatmap

5. Conclusions and prospects

5.1. Conclusions

This study confirms the pivotal role of illumination intensity in photovoltaic (PV) power generation systems. By analyzing experimental data, it is evident that variations in illumination intensity directly influence the light absorption capacity of PV cells, subsequently impacting the power output significantly. The study demonstrates a strong positive correlation between illumination intensity and power generation (with a correlation coefficient of 0.8717), suggesting that optimizing illumination intensity should be a priority in PV system design. Achieving a certain threshold of illumination intensity can enhance the power generation efficiency of photovoltaic cells by 10% to 30%, thereby boosting system output power and reducing the cost per unit of electrical energy generated.

To achieve more accurate PV power predictions, this study employed a Support Vector Regression (SVR) model, which has proven to be highly effective among various prediction models, offering the highest accuracy and lowest error rate (with residual error percentages fluctuating between -50% and 50%). The SVR model is capable of capturing complex nonlinear relationships and maintaining robust prediction performance, especially when dealing with high-dimensional data incorporating multiple meteorological factors and historical data. Through the application of the SHAP (SHapley Additive exPlanations) analysis method, illumination intensity was identified as the most crucial factor influencing prediction outcomes, with a substantial weighting and contribution exceeding 70%. Other meteorological factors, such as temperature, exhibit relatively minor effects on power generation. These findings underscore the strong potential for practical solar power generation applications, particularly in sunny regions.

5.2. Prospects

In this study, experimental variable factors are used as important inputs for PV power prediction, including illumination intensity, conversion efficiency, and voltage. Variations in these factors have a direct impact on the performance of the PV system, and combining these data can help to model it more accurately. And in addition to meteorological factors, historical power generation data also

contains the relationship between past power generation and meteorological conditions. By analyzing this data, machine learning models are able to identify potential patterns and trends, thus providing valuable references for future power generation forecasts. By introducing richer meteorological factors and historical data and applying more advanced deep learning models, this study aims to significantly improve the prediction accuracy of photovoltaic (PV) power and help decision makers better plan and manage the operation of PV power plants.

References

[1] Monica B, Adrián R, Raul G, et al. Photovoltaic Power Generation Forecasting for Regional Assessment Using Machine Learning[J]. Energies, 2022, 15(23): 8895-8895.

[2] J. KI. Solar Photovoltaic Power Forecasting: A Review[J]. Sustainability, 2022,14(24): 17005-17005.

[3] Sameer D A, Jehad A, Osama A, et al. A feature transformation and extraction approach-based artificial neural network for an improved production prediction of grid-connected solar photovoltaic systems[J]. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 2022, 44(4): 9232-9254.

[4] Connor S, Mominul A, Alhussein A. Machine learning for forecasting a photovoltaic (PV) generation system[J]. Energy, 2023, 278.

[5] Huifang F, Chunsheng Y. A novel hybrid model for short-term prediction of PV power based on KS-CEEMDAN-SE-LSTM[J]. Renewable Energy Focus, 2023, 47100497.

[6] Diwaker P, Aanchal K, Prerna G. An Enhanced Drift-Free Perturb and Observe Maximum Power Point Tracking Method Using Hybrid Metaheuristic Algorithm for a Solar Photovoltaic Power System[J]. Iranian Journal of Science and Technology, Transactions of Electrical Engineering, 2023, 48(2): 759-779.

[7] Li D, Zhu D, Tao T, et al. Power Generation Prediction for Photovoltaic System of Hose-Drawn Traveler Based on Machine Learning Models[J]. Processes, 2023, 12(1).

[8] Yuan J, Tang X, Yuan W. Exploring Pathways toward the Development of High-Proportion Solar Photovoltaic Generation for Carbon Neutrality: The Example of China[J]. Processes, 2023, 12(1).

[9] Li G, Ding C, Zhao N, et al. Research on a novel photovoltaic power forecasting model based on parallel long and short-term time series network[J]. Energy, 2024, 293130621.

[10] Tercha W, Tadjer A S, Chekired F, et al. Machine Learning-Based Forecasting of Temperature and Solar Irradiance for Photovoltaic Systems[J]. Energies, 2024, 17(5).

[11] Hou Z, Zhang Y, Liu Q, et al. A hybrid machine learning forecasting model for photovoltaic power[J]. Energy Reports, 2024, 115125-5138.

[12] Wang A, Lin Q, Liu C, et al. Sustainable Energy Development: Reviewing Carbon Emission Reduction in Photovoltaic Power Systems[J]. Sustainability, 2024, 16(23): 10428-10428.

[13] López D R, Rojas L M J, Sevil A S J, et al. Optimising Grid-Connected PV-Battery Systems for Energy Arbitrage and Frequency Containment Reserve[J]. Batteries, 2024, 10(12): 427-427.

[14] Wang Y, Lee S, Li C, et al. Techno-economic evaluation of solar photovoltaic power production in China for sustainable development and the environment[J]. Environment, Development and Sustainability, 2024, (prepublish): 1-30.

[15] Chen R, Gao S, Zhao Y, et al. A hybrid model based on the photovoltaic conversion model and artificial neural network model for short-term photovoltaic power forecasting[J]. Frontiers in Energy Research, 2024, 121446422-1446422.

[16] Pan C, Liu Y, Oh Y, et al. Short-Term Photovoltaic Power Forecasting Using PV Data and Sky Images in an Auto Cross Modal Correlation Attention Multimodal Framework[J]. Energies, 2024, 17(24): 6378-6378.

[17] Mbey F C, Kakeu F J V, Boum T A, et al. Solar photovoltaic generation and electrical demand forecasting using multi-objective deep learning model for smart grid systems[J]. Cogent Engineering, 2024, 11(1).

[18] Runqi Z, Xiangyu W, Xiaochong G, et al.Analysis of Investment Efficiency in New Energy Projects by Traditional Energy Enterprises: A Case Study of Rooftop Photovoltaic Power Projects in the Chongqing Regional Market[J]. Proceedings of Business and Economic Studies, 2024, 7(6): 73-82.