Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD29100

Research on Solar Power Generation Prediction Model Based on Ensemble Learning and Feature Engineering

Wei Liu

School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo, China
m13784084897@163.com

Abstract. With the transformation of the global energy structure, solar power generation prediction confronts technical challenges in extracting higher-order cyclical characteristics and enhancing seasonal adaptability. This paper proposes a solar power generation prediction model based on enhanced feature engineering and stacked ensemble learning. In terms of feature engineering, we comprehensively extract Fourier features of daily and annual cycles along with quadratic harmonics. We quantify the dynamic fluctuations of temperature and radiation through rolling statistical indicators and introduce seasonal correction factors to accurately characterize the illumination climate patterns of different seasons. On the model level, we employ stacked ensemble learning that integrates four regression models: Random Forest, Extremely Randomized Trees, Gradient Boosting Trees, and Support Vector Regression, with Ridge Regression as the meta-learner. Experimental results show that the proposed model performs excellently in solar power generation prediction, achieving a MAE of 9.30, an RMSE of 19.48, and an R^2 of 0.9897, significantly outperforming the comparison models. Feature importance analysis indicates that high-order Fourier features, rolling statistical features, and seasonal correction factors are key components for performance improvement. This method provides an effective technical solution for large-scale solar power generation prediction.

Keywords: Fourier Features, Quadratic Harmonics, Rolling Statistics, Ridge Regression, Ensemble Learnin

1. Introduction

As global energy structures transition toward sustainability, solar power generation is increasingly recognized as a vital renewable energy source. However, solar power output is influenced by complex meteorological and seasonal factors, resulting in significant forecasting challenges that are essential for optimizing supply and grid integration [1].

Traditional forecasting methods, including linear regression and ARIMA models, often fail to capture the nonlinearity and randomness inherent in solar power generation [2]. Machine learning techniques such as random forests and neural networks have emerged as effective alternatives [3], yet single-algorithm models can still struggle with seasonal variability and long-term stability [4].

To address these limitations, this study focuses on advanced feature engineering and ensemble learning methods. Sophisticated time series features, including Fourier components and rolling

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD29100

statistics, can significantly enhance model performance [5]. Additionally, ensemble methods, particularly stacking, optimize predictions by integrating multiple algorithms, reducing overfitting and improving robustness [6].

This study aims to develop a robust solar power forecasting model by combining advanced feature engineering with stacking ensemble learning. The main contributions include: (1) modeling complex temporal patterns through advanced cyclical and statistical features; (2) integrating multiple regression models with a ridge regression meta-learner; and (3) evaluating seasonal performance to support improved energy planning and grid integration.

2. Dataset description

The dataset related to the European power system is sourced from Open Power System Data. We extracted solar power generation data and weather data for a specific location in Austria from 2015 to 2019, including the following parameters: solar power generation (kWh), global horizontal irradiance (W/m²), direct normal irradiance (W/m²), and temperature (°C).

3. Proposal of seasonal correction factors

In this paper, we choose to standardize the data to reduce training time and improve prediction accuracy. The feature values are:

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

The target value is:

$$y' = \frac{y - \bar{y}}{\sigma_y} \tag{2}$$

Based on the standardized data, we first establish a baseline linear model using the least squares method:

$$\hat{y}_i = w^T x_i + b \tag{3}$$

Where $\hat{y}_i$ is the predicted value, and x_i includes meteorological features such as temperature and direct and scattered solar radiation, as well as time features represented by sine and cosine transformations.

Considering the significant differences in solar power generation across different seasons, we further introduce the seasonal correction factor SF_s to adjust the predicted values. The correction factor is determined by minimizing the prediction bias of the baseline model for each season:

$$SF_s = \frac{\sum_{i:season=s} (y_i \cdot \hat{y}_i)}{\sum_{i:season=s} (\hat{y}_i)^2} \tag{4}$$

To enhance the correction effect, we assign weights to each season based on sample size and prediction error, reflecting the reliability of seasonal data and the urgency of correction:

$$weight_s = \frac{n_s}{N} \cdot \left(1 + \frac{MAE_s}{\sum_{j=0}^{3} MAE_j}\right) \tag{5}$$

DOI: 10.54254/2755-2721/2025.LD29100

Where n_s is the sample size for that season, N is the total sample size, and MAE_s represents the mean absolute error for that season.

The correction factor is fine-tuned based on the weights:

$$SF_s^* = SF_s \cdot (1 - 0.05 \cdot (weight_s - 1)) \tag{6}$$

Finally, to prevent extreme corrections, the correction factor is limited to the range of [0.4, 1.8] to ensure the stability of the predicted results.

4. Periodic encoding

4.1. Fourier series expansion

To capture the temporal periodicity of solar power generation, we use Fourier transformation to map time information into periodic features.

Basic periodic features:

The daily periodic component is:

$$\sin\left(2\pi \frac{hour_of_day(t)}{24}\right) \tag{7}$$

$$\cos\left(2\pi \frac{hour_of_day(t)}{24}\right) \tag{8}$$

The annual periodic component is:

$$\cos\left(2\pi \frac{day_of_year(t)}{L(t)}\right) \tag{9}$$

$$\sin\left(2\pi \frac{day_of_year(t)}{L(t)}\right) \tag{10}$$

Where $L(t)$ represents the length of the year L, with 365 days in a common year and 366 days in a leap year.

4.2. Higher-order harmonic components

To capture the semi-annual periodic phenomena (such as similar lighting patterns in spring and autumn), this paper establishes the annual second harmonic:

$$\cos\left(4\pi \frac{day_of_year(t)}{L(t)}\right) \tag{11}$$

$$\sin\left(4\pi \frac{day_of_year(t)}{L(t)}\right) \tag{12}$$

To analyze the intra-day bimodal phenomenon (such as the two peaks of radiation in the morning and afternoon), this paper establishes the daily second harmonic:

$$\cos\left(4\pi \frac{day_of_year(t)}{L(t)}\right) \tag{13}$$

DOI: 10.54254/2755-2721/2025.LD29100

$$\sin\left(4\pi\frac{\mathrm{day_of_year(t)}}{\mathrm{L(t)}}\right) \tag{14}$$

5. Interaction features

To capture the nonlinear relationships between variables, key interaction features are constructed to enhance the model's expressive power.

Temperature-radiation interaction feature:

$$\mathrm{F_{interaction}} = \mathrm{AT_{temperature}} \times \mathrm{AT_{radiation_{d}irect_{h}orizontal}} \tag{15}$$

This feature combines the two main factors that influence solar power generation, thereby better reflecting the dynamic relationship under actual generation conditions.

Radiation ratio feature:

$$\mathrm{F_{ratio}} = \frac{\mathrm{AT_{radiation_{d}iffuse_{h}orizontal}}}{\mathrm{AT_{radiation_{d}irect_{h}orizontal}}+1\times10^{-5}} \tag{16}$$

Through these interaction features, the model is able to more accurately capture the complex nonlinear relationships between meteorological conditions and solar power generation.

6. Rolling statistics

To capture the dynamic changing patterns of meteorological data, rolling statistical features based on a sliding window are designed and constructed.

6.1. Rolling mean

By calculating the moving averages of temperature and radiation data over the past year, short-term fluctuations are effectively smoothed out, allowing for the capture of long-cycle change trends. The formula is:

$$\mathrm{RollingMean_t} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i=t-n+1}}^{\mathrm{t}}\mathrm{x_i} \tag{17}$$

6.2. Rolling standard deviation

By quantifying the degree of fluctuation of data within the time window, it reflects the uncertainty and variability of meteorological conditions. The formula is:

$$\mathrm{RollingStd_t} = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i=t-n+1}}^{\mathrm{t}}\left(\mathrm{x_i}-\bar{\mathrm{x}}_{\mathrm{t}}\right)^2} \tag{18}$$

These rolling statistical features enable the model to better understand the model understand the historical fluctuation patterns of meteorological parameters, thereby enabling more accurate predictions of solar power generation performance under different climatic conditions.

DOI: 10.54254/2755-2721/2025.LD29100

7. Stacked models

This paper constructs a stacked ensemble framework based on four regression models, enhancing prediction performance by combining the strengths of different algorithms.

7.1. Base learners

Random Forest and Extremely Randomized Trees: These algorithms adopt Bootstrap sampling to construct multiple decision trees. Random Forest selects the optimal features during node splitting, while Extremely Randomized Trees randomly choose split points. The ensemble prediction formula is:

$$\hat{y}\left(x\right) = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t\left(x\right) \tag{19}$$

Gradient boosting trees gradually correct prediction errors through an iterative process. Residual calculation:

$$r_{i,m} = y_i - F_{m-1}(x_i) \tag{20}$$

Cumulative model updates are as follows:

$$F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x) \tag{21}$$

Support Vector Regression employs kernel functions to achieve nonlinear regression, with the regression function given by:

$$f\left(x\right) = \sum_{i=1}^{N}\left(\alpha_i - \alpha_i^*\right)K\left(x_i, x\right) + b \tag{22}$$

7.2. Stacked ensemble strategy

Utilizing the training set cross-validation predicted values of four base models: Random Forest Regression, Extremely Randomized Trees Regression, Gradient Boosting Regression, and Support Vector Regression, a new feature matrix $Z \in R^{N\times 4}$ is formed:

$$Z = \begin{bmatrix} \hat{y}_1^{(rf)} & \hat{y}_1^{(etr)} & \hat{y}_1^{(gr)} & \hat{y}_1^{(svr)} \\ \hat{y}_2^{(rf)} & \hat{y}_2^{(etr)} & \hat{y}_2^{(gr)} & \hat{y}_2^{(svr)} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{y}_N^{(rf)} & \hat{y}_N^{(etr)} & \hat{y}_N^{(gr)} & \hat{y}_N^{(svr)} \end{bmatrix} \tag{23}$$

Our final model is Ridge Regression:

$$f_{Ridge}\left(Z\right) = Z\beta + b, Z \in R^{N\times 4}, \beta \in R^4, b \in R \tag{24}$$

When training the final model, the optimal β and b are calculated by minimizing the objective function:

DOI: 10.54254/2755-2721/2025.LD29100

$$\min_{\beta,b} \frac{1}{N} \sum_{i=1}^{N} \left[y_i - (Z_i \beta + b)^2 + \| \beta \|^2 \right] \tag{25}$$

For the feature data $X_{new} \in R^{M \times d}$, predictions are made using all four trained base models separately:

$$\hat{y}_{new}^{(rf)} = f_{rf}(X_{new}) \tag{26}$$

$$\hat{y}_{new}^{(etr)} = f_{rf}(X_{new}) \tag{27}$$

$$\widehat{y}_{new}^{(gbr)} = f_{rf}(X_{new}) \tag{28}$$

$$\hat{y}_{new}^{(svr)} = f_{rf}(X_{new}) \tag{29}$$

Using the four base models for prediction, we obtain new predicted values and form a new feature matrix $Z_{new} \in R^{N \times 4}$:

$$Z_{new} = \begin{bmatrix} \hat{y}_{new,1}^{(rf)} & \hat{y}_{new,1}^{(etr)} & \hat{y}_{new,1}^{(gbr)} & \hat{y}_{new,1}^{(svr)} \\ \hat{y}_{new,2}^{(rf)} & \hat{y}_{new,2}^{(etr)} & \hat{y}_{new,2}^{(gbr)} & \hat{y}_{new,2}^{(svr)} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{y}_{new,M}^{(rf)} & \hat{y}_{new,M}^{(etr)} & \hat{y}_{new,M}^{(gbr)} & \hat{y}_{new,M}^{(svr)} \end{bmatrix}, Z_{new} \in R^{M \times 4} \tag{30}$$

Finally, predictions are made using the trained Ridge model:

$$\hat{y}_{final,new} = Z_{new} \beta + b \tag{31}$$

After outputting the model predictions, the values are de-standardized and then multiplied by the seasonal factors to obtain the final predicted values. This stacking strategy fully utilizes the complementarity of each base model and combines the seasonal adjustment factors to output the final predictions.

8. Results analysis

Table 1. The performance of the prediction model

Prediction Model	MAE	RMSE	R2
Composite Ridge Ensemble Model	9.3000	19.4838	0.9897
LSTM	29.0363	59.1947	0.8947
Random Forest Regression	11.8188	24.3770	0.9708
Support Vector Regression	29.8497	61.0489	0.8761
Extremely Randomized Trees Regression	11.5205	23.8047	0.9707
Gradient Boosting Regression	13.4143	25.7258	0.9678

DOI: 10.54254/2755-2721/2025.LD29100

Table 2. The R^2 of the prediction model for each season

Prediction Model	Spring	Summer	Autumn	Winter
Composite Ridge Ensemble Model	0.9867	0.974	0.9881	0.99
LSTM	0.9540	0.9331	0.8135	0.8878
Random Forest Regression	0.9715	0.9532	0.9727	0.9719
Support Vector Regression	0.8845	0.8865	0.8860	0.8873
Extremely Randomized Trees Regression	0.9726	0.9678	0.9792	0.9796
Gradient Boosting Regression	0.9786	0.9219	0.9676	0.9455

8.1. Experiment results analysis

As shown in Table 1, the proposed stacked ensemble model outperforms all comparison models significantly across all evaluation metrics The ensemble model achieves a MAE of 9.30, an RMSE of 19.48, and an R^2 of 0.9897, demonstrating a noticeable improvement in precision compared to the second-best model, the Extremely Randomized Trees Regression. The prediction errors of LSTM and Support Vector Regression are relatively large, indicating that traditional single models struggle to fully capture the complex characteristics of solar power generation.

The seasonal analysis in Table 2 shows that the ensemble model maintains an R^2 of over 0.97 in all four seasons, demonstrating good seasonal adaptability. In contrast, other models exhibit significant seasonal fluctuations, with the LSTM achievingan R^2 of only 0.8135 in autumn.

8.2. Feature importance analysis

After removing Support Vector Regression, the performance of the ensemble model actually declines with R^2 dropping to 0.9803, indicating that SVR captures nonlinear relationships that other models may overlook, highlighting the complementary advantages of ensemble learning. Feature elimination experiments indicate that removing rolling statistical features significantly degrades model performance with R^2 decreasing from 0.9897 to 0.9491, while the impact of removing Fourier features is relatively small, underscoring the crucial role of rolling statistical features in capturing dynamic changes in time series.

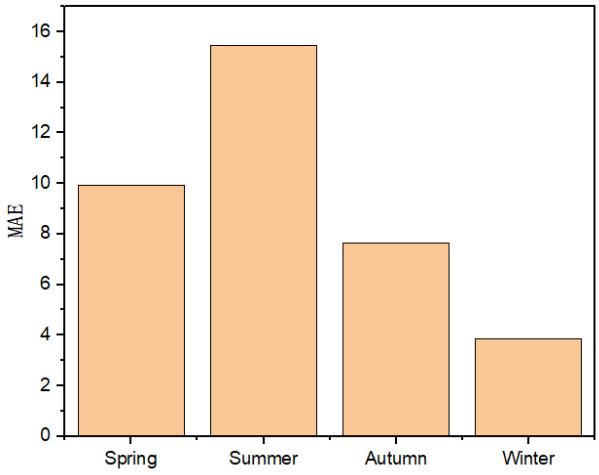


Figure 1. MAE by season

DOI: 10.54254/2755-2721/2025.LD29100

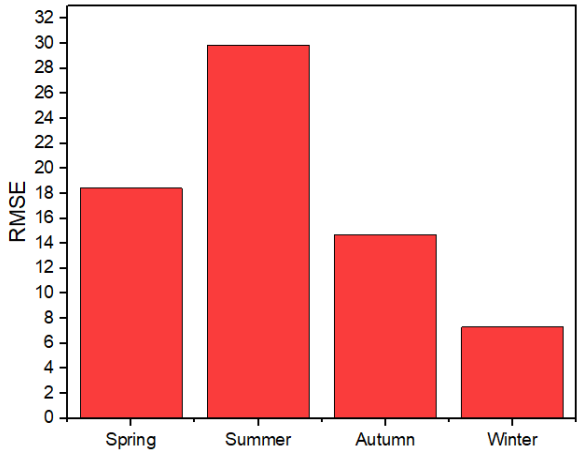


Figure 2. RMSE by season

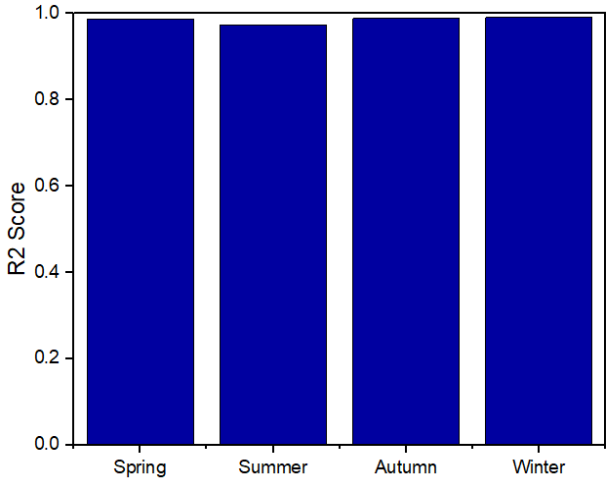


Figure 3. R^2 Score by season

8.3. Seasonal performance analysis

Figures 1-3 show that the model exhibits the largest prediction errors in summer, primarily due to significant fluctuations in power generation during this season. The model achieves the highest fitting accuracy in winter, indicating that it can handle situations with low and relatively stable power generation effectively.

8.4. Limitations and improvements

Despite the good performance of the model presented in this study, there remain some limitations. Firstly, the model has relatively high complexity and computational cost, which may pose efficiency challenges in real-time prediction scenarios. Secondly, the model is mainly trained on historical meteorological data, and its predictive ability for extreme weather events requires further validation.

Future research can explore the following directions: (1) introducing a hybrid architecture that combines deep learning with traditional ensemble learning to enhance computational efficiency while maintaining prediction accuracy; (2) integrating multi-source data such as satellite cloud images and ground observations to improve the model's responsiveness to sudden weather changes;

DOI: 10.54254/2755-2721/2025.LD29100

(3) exploring online learning mechanisms to enable the model to adapt to long-term trend changes caused by climate change and equipment aging; (4) expanding the application range of the model to verify its generalization performance under different geographical locations and climate conditions.

References

[1] IRENA. (2020). Renewable Power Generation Costs in 2019. International Renewable Energy Agency.
[2] Kalogirou, S. A. (2014). Solar Energy Engineering: Processes and Systems. Academic Press.
[3] Poncelet, K., et al. (2016). "A robust approach for short-term solar power forecasting." Solar Energy, 139, 73-83.
[4] Zhao, Q., et al. (2016). "Solar irradiance forecasting using ARIMA model." Renewable Energy, 91, 1-8.
[5] Díaz, J., et al. (2020). "Machine learning techniques for solar power forecasting: A review." IEEE Transactions on Sustainable Energy.
[6] Abdel-Ghany, A. M., et al. (2020). "Predicting solar energy production using machine learning: A review." Journal of Cleaner Production, 258, 120836.