ISSN 2186-5647

Comparative Analysis of Machine Learning Models in Solar Power Generation Forecasting

－Applications of Random Forest, Gradient Boosting, and Ridge Regression－

Engineering Faculty, Mahasarakham University, Thailand Sarinya Sala-ngam
Sakon Nakhon Rajabhat University, Thailand Prakaykaew Boottarat
Engineering Faculty, Mahasarakham University, Thailand Chonlatee Photong
Industrial Technology of Nihon University, Japan * Jun Toyotani

1 Introduction

Solar power's growing role in global energy sustainability emphasizes the need for accurate forecasting to optimize photovoltaic (PV) systems and ensure grid stability [1]. The variability in solar energy, influenced by solar irradiance and ambient temperature, poses challenges for precise prediction [2]. Machine learning models, particularly ensemble techniques such as Random Forest (RF) and Gradient Boosting (GB), have shown promise in capturing complex relationships between these environmental factors and solar output. Ridge Regression (RR) offers a simpler, interpretable alternative for forecasting [3].

With the increasing integration of solar energy into the global grid, accurate forecasting is crucial for efficient PV system operation and grid stability. Effective models are needed to manage resources like spinning reserves and frequency response. This study aims to compare the performance of RF, GB, and RR models using data from a solar PV rooftop system in Thailand to identify the most effective approach for different forecasting scenarios.

Studies show Random Forest is highly effective for solar forecasting, thanks to its ability to handle complex data patterns [4][5]. Gradient Boosting, particularly XGBoost, has been reported to slightly outperform Random Forest in terms of root mean square error (RMSE) [6]. Other models, including Support Vector Regression, Linear Regression, and ARIMA, have been explored but with varied success [3][4][6]. The impact of environmental factors on PV output is well-documented, adding complexity to forecasting tasks.

There is a lack of comprehensive studies comparing advanced ensemble techniques like Gradient Boosting and Random Forest with simpler models like Ridge Regression. Existing research often evaluates models in isolation rather than providing a detailed comparative analysis under varying conditions or practical scenarios. Moreover, there is limited research on the performance of these models in specific geographic contexts.

This study fills these gaps by offering a detailed comparison of RF, GB, and RR models using data from a solar PV rooftop system at Sakon Nakhon Rajabhat University, Thailand, spanning from 2019 to 2022. The evaluation, based on mean absolute error (MAE), reveals that Gradient Boosting achieves the highest accuracy, followed by Ridge Regression and Random Forest. These findings enhance understanding of model performance for solar forecasting and provide a framework that can be adapted to other renewable energy sources, supporting improved energy management and grid stability.

2 Research Methodology

2.1 Data Collection

This study utilizes historical data from a

solar PV rooftop system located at Sakon Nakhon Rajabhat University, Thailand, spanning from May 2019 to December 2022. The dataset comprises solar irradiance and ambient temperature as the primary features, with the following specifics about the PV system: situated at 17.19° N, 104.09° E, with a metal sheet building structure, a tilt of 5° and azimuth of 31°, and a total of 454 PV modules, each rated at 320 Wp (Fig.1). The PV module array includes 18 and 19 modules in series and 21 parallel strings, supported by 6 grid-tied inverters, each with a capacity of 25 kW, totaling 145.28 kWp. Data collection included measurements of solar irradiance, ambient temperature, and solar power generation shown in Table 1.

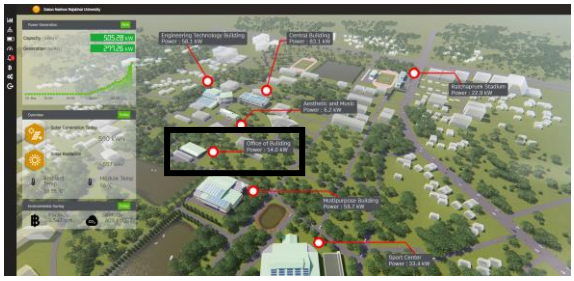

Fig.1 Location and geographical details

Fig.2 The details of installed the solar PV system

Table 1 Data collection

No.	Production (kWh)	Solar irradiance	Ambient Temp.
1	5213.80	271.31	33.79
2	18538.00	302.36	34.10
3	19279.40	286.26	32.85
4	17111.70	271.08	31.63
5	18343.50	260.78	31.71
…	…	…	…
37	16993.50	283.97	32.00
38	16607.00	274.47	32.00
39	10696.10	274.12	31.00
40	16221.50	250.34	32.90
41	9078.20	267.61	32.40

2.2 Model Development

Three machine learning models were selected for this study: Random Forest (RF), Gradient Boosting (GB), and Ridge Regression (RR) with Python code. The Random Forest model, an ensemble technique that builds multiple decision trees and aggregates their predictions, was set up with 30, 50 and 100 trees as hyperparameters. Gradient Boosting, implemented through XGBoost, creates sequential models that correct previous errors, using a maximum depth of 2, 3, 5, and 7, a learning rate of 0.05, 0.1, and 0.2 boosting rounds. Ridge Regression, a linear model with L2 regularization, was applied to mitigate overfitting by penalizing large coefficients, with regularization strength (alpha: 0.2, 0.3, 0.5 and 1) optimized via cross-validation.

2.3 Model Training and Validation

The dataset (41 datasets) was divided into training and testing sets, with 80% of the data used for training and 20% for testing. To enhance the robustness of the model

evaluation, a 5-fold cross-validation, optimizing for Mean Absolute Error (MAE), approach was employed. This procedure was repeated five times, and the performance metrics were averaged to ensure reliability.

2.4 Performance Evaluation

The performance of each model was assessed using MAE. MAE measures the average magnitude of prediction errors, reflecting the model's accuracy in solar power forecasting. This metric was calculated for each model on the test set to determine their effectiveness and accuracy.

2.5 Comparison and Analysis

The results from each model were compared to identify the most effective approach for forecasting solar power generation. This comparative analysis revealed the strengths and limitations of Random Forest, Gradient Boosting, and Ridge Regression, providing insights into which model best captures the complex relationships between environmental factors and solar power output. The findings contribute valuable knowledge to the field of renewable energy forecasting and offer practical guidance for selecting the most suitable model for solar power prediction.

4. Results and discussion

The performance of the three machine learning models, Random Forest, Gradient Boosting, and Ridge Regression was evaluated using MAE on both the training and test datasets (Table 1).

Random Forest exhibited strong performance on the training set, with a MAE of 801.72 However, it showed a significant decline in performance on the test set, where the MAE increased to 2,421.47. This suggests that the model may be overfitting, effectively capturing patterns in the training data but struggling to generalize to new, unseen data.

Gradient Boosting, on the other hand, delivered more balanced performance between the training and test datasets. With a MAE of 1,336.89 on the training set and a MAE of 2,064.67 on the test set, it demonstrated better generalization compared to Random Forest. This model provided the most reliable predictions, making it the strongest performer overall.

Ridge Regression, while the simplest of the three models, showed consistent performance with a MAE of 1,991.89 on the training set and a MAE of 2,065.82 on the test set. Although it maintained stability across both sets, it lagged behind Gradient Boosting and Random Forest in predictive accuracy.

Table 2 The results of model performance evaluation using MAE

Models	MAE	
	Train	Test
Random Forest	801.72	2421.47
Gradient Boosting	1336.89	2064.67
Ridge Regression	1991.89	2065.82

In summary, Gradient Boosting outperformed the other models, striking a balance between accuracy and generalization. Random Forest, while powerful on the training data, experienced a sharp drop in test performance, indicating overfitting. Ridge Regression, though stable, lacked the predictive strength of the ensemble models. These findings underscore the advantages of Gradient Boosting in solar power generation forecasting, offering a robust solution for optimizing photovoltaic system operations and improving grid management.

5.Conclusions

This study presents a comparative analysis of Random Forest, Gradient Boosting, and Ridge Regression for forecasting solar power generation. Gradient Boosting demonstrated superior

performance with balanced accuracy across training and test sets, while Random Forest showed signs of overfitting. Ridge Regression, although stable, was less accurate compared to the ensemble models. These findings highlight the effectiveness of Gradient Boosting in capturing the non-linear relationships in solar generation data.

A key contribution of this research is filling the gap in existing studies by providing a detailed comparison of these machine learning models. Prior research lacked comprehensive evaluations of ensemble methods for solar forecasting, and this study demonstrates their superiority, particularly in terms of generalization. However, the study is limited by a relatively small dataset and the use of only two environmental features. Future research should incorporate additional variables and expand to other geographic regions to enhance model robustness.

Future research should focus on expanding the dataset to include more diverse environmental factors and geographical locations to enhance model accuracy and robustness. Researchers are encouraged to explore advanced ensemble techniques and address overfitting in Random Forest models to further improve prediction performance in solar energy forecasting.

References

1) J. Lee, J. Ko, C. J. Park, and G.-L. Park, "A Prediction Model For Solar Energy Generation Built Upon Status Monitoring," *IJCA*, vol. 9(8), pp. 349–358.
2) P. Boottaraja, and N. Phuangpornpitak, "Performance Analysis of 325 kW Solar PV Rooftop System Using PVsyst Program", *Int. Journal of Environmental and Rural Development* (2019), vol. 10(2), pp. 40-45.
3) N. Sultana, and T. Ahmed, "Performance Analysis of Machine Learning Models in Solar Energy Forecasting", *Int. Journal of Machine Learning* (2023), Vol. 13(3), pp.131-135.
4) K. Anuradha, D. Erlapally, G. Karuna, V. Srilakshmi, and K. Adilakshmi, "Analysis Of Solar Power Generation Forecasting Using Machine Learning Techniques," *E3S Web Conf.*(2021), vol. 309, p. 01163.
5) K. Mahmud, S. Azam, A. Karim, S. Zobaed, B. Shanmugam, and D. Mathur, "Machine Learning Based PV Power Generation Forecasting in Alice Springs," *IEEE Access* (2021), vol. 9, pp. 46117–46128.
6) M. M. Sucharitha, S. Sowjanya, K. N. Sumalatha, S. Ayesha, and M. S. A. Basha, "Predictive Modeling of Solar Energy Production: A Comparative Analysis of Machine Learning and Time Series Approaches," *IEEE Int. Conf. for Women in Innovation, Technology & Entrepreneurship (ICWITE)* (2024), pp. 235-241.