Machine Learning Framework for Short-term Photovoltaic Power Forecasting with Feature Selection and Power Range Error Analysis

Nguyen Bao Quoc[1], Bui Vu Minh[2], Ho A Lil[2], Lam Hoang Cat Tien[1], Tran Ngoc Huy Thinh[2,*]
[1]Faculty of Electrical and Electronic Engineering, Cao Thang Technical College, Ho Chi Minh City, Viet Nam
[2]Faculty of Engineering and Technology, Nguyen Tat Thanh University, Ho Chi Minh City, Viet Nam
[*]tnhthinh@ntt.edu.vn

Abstract

This paper presents a machine learning framework for short-term photovoltaic power forecasting, aiming to develop a tiny model suitable for low-cost edge devices while maintaining accuracy. The methodology integrates data preprocessing, time-series feature engineering, feature selection based on correlation and feature importance, and performance evaluation of ANN, Random Forest, SVR, and XGBoost. Forecasting errors are analyzed across different photovoltaic power ranges, while SHAP is used to interpret feature contributions. Results show that reducing input features from 15 to 7 (53.3%) maintains comparable forecasting performance while reducing model complexity. ANN achieves the best forecasting accuracy and trade-off between accuracy and model size, with its trained model size decreasing from 82.80 KB to 70.37 KB. The proposed framework provides a practical approach for developing compact and accurate photovoltaic forecasting models for resource-constrained Edge platforms.

® 2026 Journal of Science and Technology - NTTU

Received 10/04/2026
Accepted 08/09/2026
Published 28/10/2026

Keywords

TinyML; Photovoltaic power forecasting; smart grid; time-series analysis.

1 Introduction

In recent years, renewable energy, especially solar power, has been widely used to reduce greenhouse gas emissions and the use of fossil fuels [1]. However, the output power of photovoltaic (PV) systems depends greatly on weather conditions and solar irradiance. As a result, PV power varies continuously, making power system operation more difficult [2]. For this reason, short-term PV power forecasting is important for many power system tasks, such as power scheduling, reserve planning, and energy management in smart grids. With the development of artificial intelligence and machine learning, many models, including Artificial Neural Networks (ANN), Support Vector Regression (SVR), Random Forest (RF), Gradient Boosting, and Long Short-Term Memory (LSTM), have been used for PV power forecasting. Besides improving forecasting accuracy, many recent studies have focused on running machine learning models directly on edge devices installed at PV systems [3, 4]. To process data and execute machine learning models directly on resource-constrained devices such as microcontrollers, models are optimized in terms of size and parameters to reduce memory and computational requirements. This approach is known as Tiny Machine Learning (TinyML). In this study, TinyML enables PV power forecasting directly at the edge device, thereby limiting data transmission to a remote control center.

The authors implemented a TinyML model to predict solar power output on Raspberry Pi 4 hardware.

https://doi.org/10.55401/vj84sp67

Model pruning and quantization were employed to reduce the model size [5]. An Arduino Nano 33 BLE was used to collect and process environmental sensor signals locally [6]. TinyML enables farmers to optimize irrigation schedules and the utilization of solar and wind energy without continuous data transmission to a cloud center. A lightweight Long Short-Term Memory model was deployed on an ESP32 to forecast solar farm power output [7]. The model was optimized using Diversity Guided PSO. On-site power prediction using TinyML allows PV inverters to receive rapid control references for reactive power dispatch without delays associated with a central control unit. Therefore, it is important to build small machine learning models that still give good forecasting accuracy. One way to do this is to find the most important input features and remove unnecessary ones. To address these issues, we introduce a TinyML model for short-term photovoltaic power generation prediction with the following contributions: (i) Data preprocessing, building time-series features for the dataset, establishing correlations between input features and power generation to eliminate less important features, and training and testing results across different machine learning models. (ii) Using SHapley Additive exPlanations (SHAP) techniques to evaluate the impact of input features on power generation, identifying important features to select for developing a suitable TinyML model for deployment on low-cost edge devices. (iii) Analyzing the performance of the models based on different power generation ranges.

2 Methodology and Framework

2.1 Dataset and Feature Engineering

2.1.1 Data Preprocessing

The training dataset for the machine learning model in this study was obtained from Alice Springs Solar Energy Centre (ASSEC), Australia [8]. This dataset includes time-series PV power along with other meteorological parameters, such as Global Horizontal Radiation (GHR), Ambient Temperature (AT), and Relative Humidity (RH). The dataset used in our paper contains 105,000 samples at a resolution of 5 minutes/sample. After performing the verification steps, we found that the dataset had missing data and anomalies. These steps include interpolation to supplement missing data, removal of duplicate data records, and identification of outliers compared to the physical range of meteorological variables to replace them if necessary. The next step involves normalizing the data to the range [0, 1] using the Min-Max scaling method. Finally, the dataset will be divided into subsets including training, validation, and testing sets to prevent data leakage during model training.

2.1.2 Feature Engineering and Feature Selection

To improve the learning ability of the solar power forecasting model, the features of the dataset are built, including components such as temporal features, meteorological features, and time-series features. To implement temporal features, hour, day, and month variables are encoded using sine-cosine transformations to represent the natural cyclicity of time. The purpose of this method is to avoid interruption between the beginning and end of each cycle [9]:

$$\mathrm{h}_{sin} = \sin\left(\frac{2\pi \mathrm{h}}{24}\right)$$
$$\mathrm{h}_{cos} = \cos\left(\frac{2\pi \mathrm{h}}{24}\right) \quad (1)$$

Equation (1) shows how the hour feature is encoded using the sine and cosine functions. In the same way, the day and month features are encoded using the same method with their own time cycles. The features from past PV power data are used to help the model learn how the data changes over time. These features include lag features, such as $P_{t-1}, P_{t-2}, \ldots, P_{t-k}$ and the moving average feature:

$$MA_t = \frac{1}{n}\sum_{i=0}^{n-1} P_{t-i} \quad (2)$$

Equation (3) shows how the ramp rate is used to describe the recent change in PV power [10]:

$$RR_{t-1} = P_{t-1} - P_{t-2} \quad (3)$$

Where P_{t-1} and P_{t-2} are the PV power values at the two time steps before the prediction time. The Pearson correlation coefficient is used to measure the linear relationship between the input features and PV power [11], as shown in Equation (4).

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (4)$$

In Equation (4), the correlation coefficient r ranges from −1 to 1. The value close to 1 or −1 indicates a strong positive or negative relationship, while a value close to 0 indicates a weak or no linear relationship. Features with a high absolute correlation value are selected for the forecasting model, while features with low correlation can be removed to reduce data size and avoid unnecessary information. In addition to the Pearson correlation coefficient, Random Forest is used to measure the importance of each feature [12]. The importance score is calculated from the impurity reduction of a feature over all trees, as shown in Equation (5):

$$FI_j = \frac{1}{T}\sum_{t=1}^{T}\Delta I_{(j,t)} \quad (5)$$

Where FI_j is the importance of feature j, T is the total number of trees, and $\Delta I_{(j,t)}$ is the impurity reduction of feature j in tree t. Therefore, both methods are used to select the input features. Figure 1 shows that GHR has a strong positive correlation with PV power, while features such as rainfall and humidity have only a weak correlation. Figure 2 illustrates the time series characteristics, showing that the moving average (AP_MA3, AP_lag) and lagged PV power characteristics have the highest correlation with the target variable. The results from Figures 1-3 show that historical PV power features are the most important for forecasting PV power, followed by solar radiation characteristics. Other meteorological characteristics play a less significant role. These results suggest that we can prioritize important characteristics and discard less contributing ones to reduce the complexity of the machine learning model.

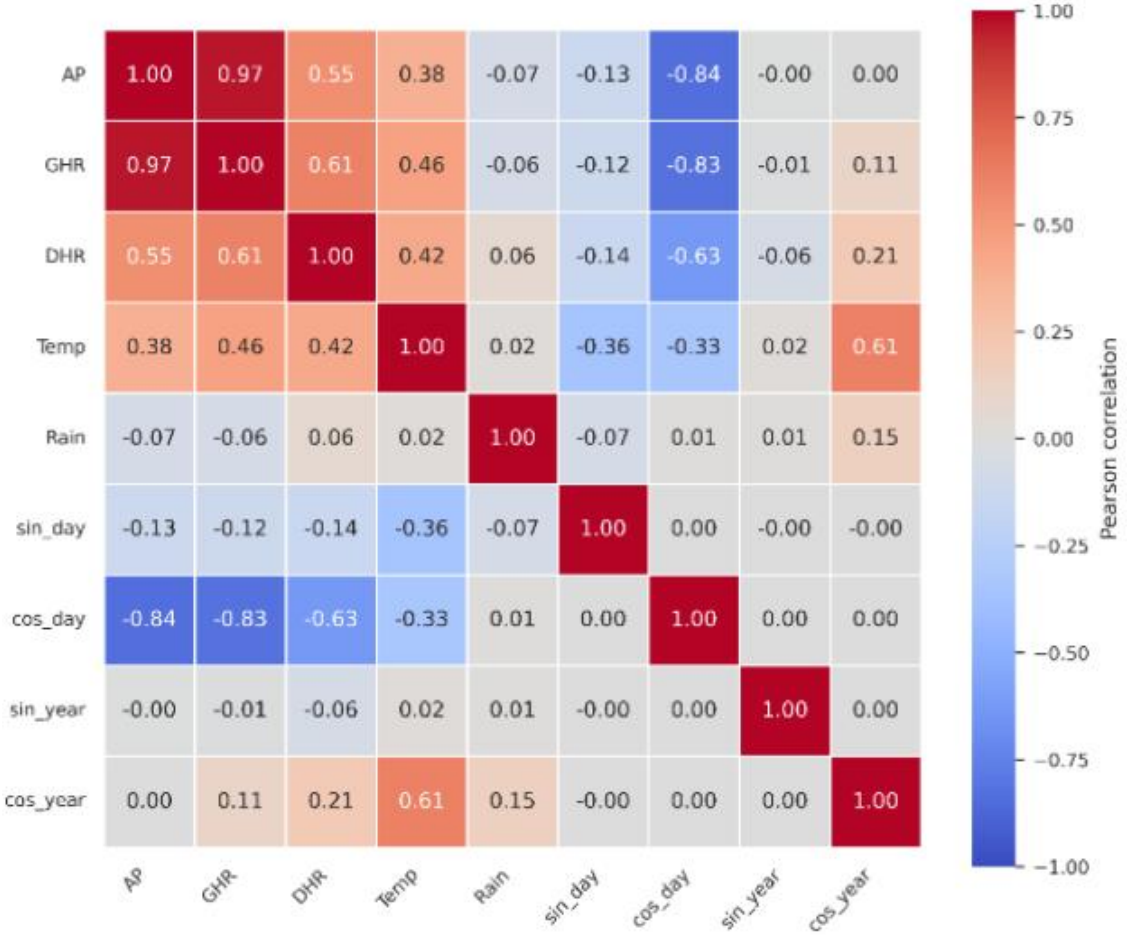


Figure 1 Pearson correlation heatmap of the selected engineered features and PV power

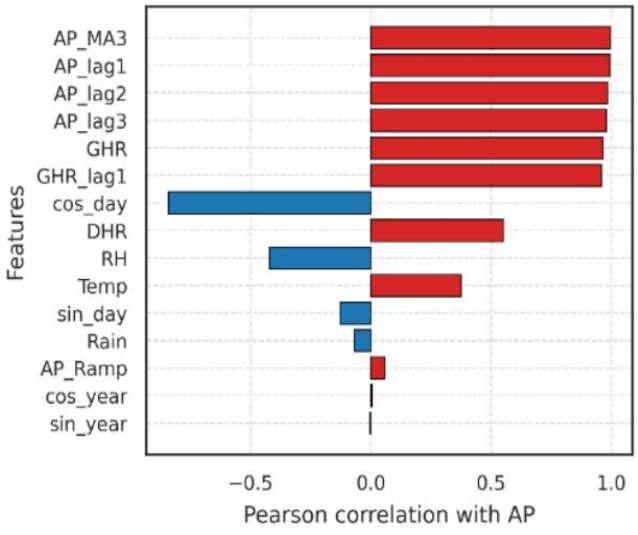


Figure 2 Pearson correlation coefficients between the engineered input features and PV power

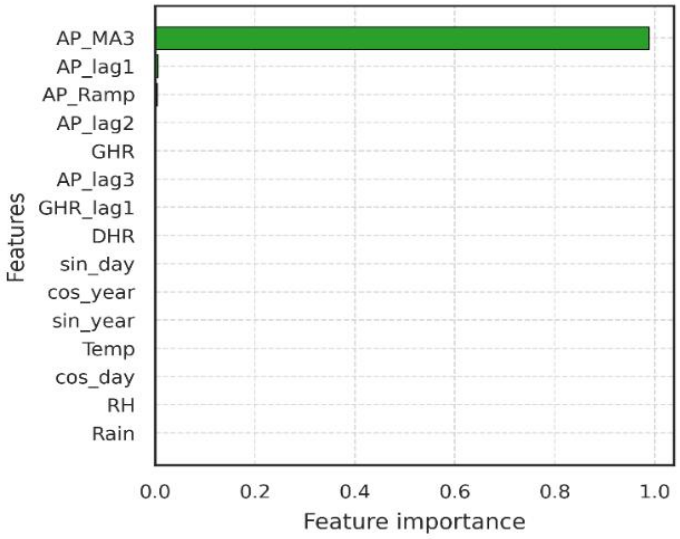


Figure 3 Feature importance ranking of the engineered variables based on the Random Forest model

Based on the Pearson correlation analysis and feature importance results presented in Figures 1-3, the authors compared two configurations: one utilizing all 15 input variables and another employing the 7 selected features. The comparison encompassed the feasibility of hardware implementation and an evaluation of prediction accuracy metrics for the models under these two configurations. Table 1 lists the input features for both configurations of the training dataset.

Table 1 Two configurations with training inputs for the PV power prediction model.

Configuration	Input Features	Number of Features	Reduction
Full Feature	GHR, DHR, Temp, RH, Rain, sin_day, cos_day, sin_year, cos_year, AP_lag1, AP_lag2, AP_lag3, AP_MA3, AP_Ramp, GHR_lag1	15	–
Selected Feature	AP_MA3, AP_lag1, AP_lag2, AP_lag3, GHR, GHR_lag1, cos_day	7	53.3%

2.2 Forecasting Models

Four machine learning models, including ANN, SVR, RF, and XGBoost, were employed for short-term PV power forecasting. These models were selected for their ability to capture nonlinear relationships between meteorological variables, time series features, and PV power. The ANN was implemented as a Multilayer Perceptron with two hidden layers containing 64 and 32 neurons, using ReLU activation and a linear output layer. The model was trained using the Adam optimizer with a learning rate of 0.001, a batch size of 64, and 100 epochs. The SVR model employed an RBF kernel with a regularization parameter of 100, an epsilon value of 0.01, and gamma set to scale. The Random Forest model consisted of 200 trees with a maximum depth of 15 and a minimum of 2 samples per leaf. XGBoost was also employed as a tree-based boosting model for comparison. All four models were trained and evaluated using both the Full Feature and Selected Feature configurations to compare their forecasting performance and suitability for deployment on resource-constrained edge devices.

2.3 Evaluation Methodology

The dataset includes input features x_t mapping to the output label PV power y_t. The prediction result is $\hat{y}_t = f(x_t)$. Therefore, short-term solar power prediction is built as a regression problem. The performance results of the models will be evaluated based on standard regression metrics. These metrics include Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), nMAE, and nRMSE:

To assess model reliability under different operating conditions, errors are further analyzed across PV power ranges:

- Low: $0 \le y_t < 0.3P_{\max}$
- Medium: $0.3P_{\max} \le y_t < 0.7P_{\max}$
- High: $0.7P_{\max} \le y_t \le P_{\max}$

Evaluating model performance across different power ranges helps determine model stability under varying operating conditions. This is particularly important for machine learning models in the high-power range, where the predictive model plays a pivotal role in the precise control of the system. Discrepancies between actual and predicted values over time are also assessed using the absolute difference $|y_t - \hat{y}_t|$ to identify periods characterized by significant deviations or errors. The SHAP method is employed to identify input features that significantly influence the forecasting model's accuracy. SHAP is a commonly used tool for interpreting machine learning model predictions; it relies on cooperative game theory to quantify the contribution of each feature. The prediction equation is presented in Equation (6):

$$f(\mathbf{x}) = \phi_0 + \sum_{i=1}^{M} \phi_i \qquad (6)$$

In Equation (6), the contribution of each feature to the prediction equation is denoted by ϕ_i. This quantity provides a quantitative explanation of the predictive model's behavior.

3 Results and Discussion

The models were evaluated using MAE, RMSE, nMAE, and nRMSE across two input configurations: 15 features (Full Feature) and 7 features (Selected Feature). Figure 4 shows that reducing the number of input features from 15 to 7, corresponding to a 53.3% reduction in input dimensionality, changed MAE, RMSE, nMAE, and nRMSE, while the relative performance trends remained consistent. ANN achieved the lowest forecasting errors, followed by XGBoost and Random Forest, whereas SVR exhibited the highest errors. These results demonstrate that the selected feature set retained essential information for PV power forecasting despite the substantial reduction in input variables. Consequently, the Selected Feature configuration offers a favorable balance between forecasting performance and input complexity. These findings provide a basis for evaluating TinyML model size and deployment feasibility on Edge platforms with limited memory and computational resources.

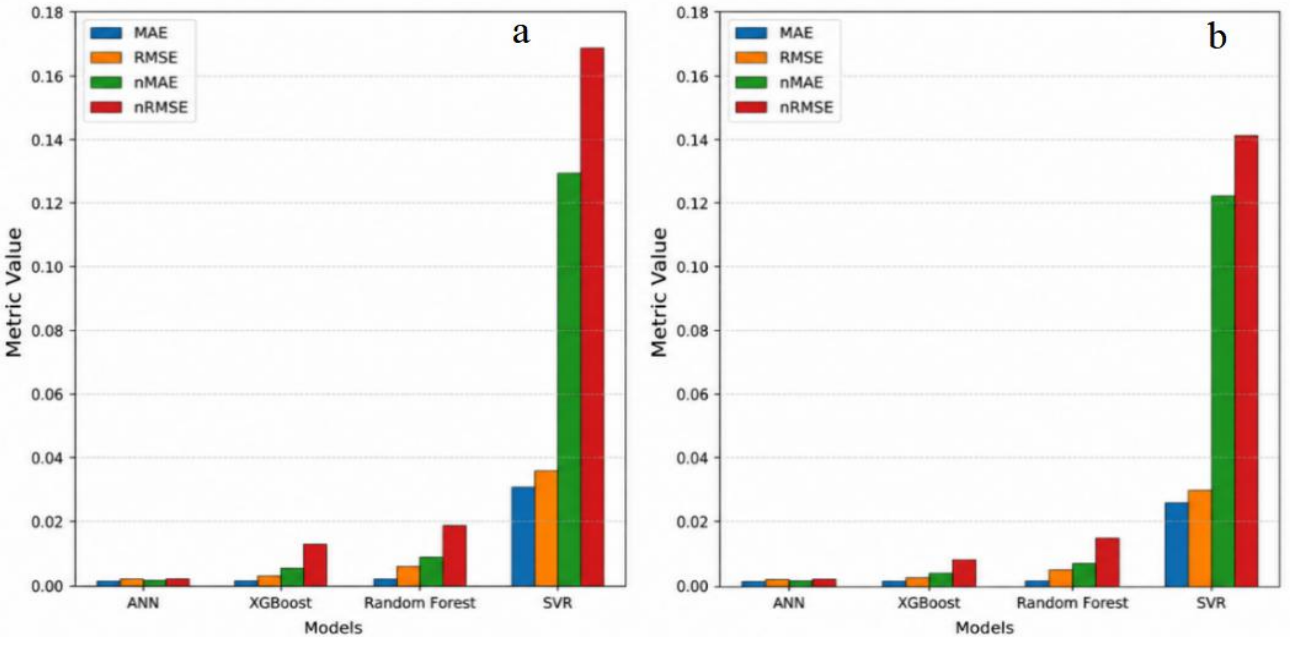


Figure 4 Comparison of MAE, RMSE, nMAE, and nRMSE for ANN, XGBoost, Random Forest, and SVR using (a) the full feature configuration and (b) the selected feature configuration

The authors evaluated the practical deployment feasibility of the models on resource-constrained hardware. The post-training sizes of the ANN, SVR, XGBoost, and Random Forest models were determined for both Full Feature and Selected Feature configurations. These model sizes were compared against the Flash/storage and RAM/SRAM capacities of devices ranging from microcontrollers to single-board computers. The results presented in Table 2 allow for a simultaneous assessment of model size and hardware resource compatibility, thereby facilitating the selection of suitable models for Edge deployment.

Table 2 TinyML model size and deployment feasibility on resource-constrained edge platforms

Model Configuration	**Trained Model Size** (KB)	**Arduino Nano 33 BLE Sense** 1 MB Flash/256 KB RAM	**ESP32-S3** ≥ 4 MB Flash/512 KB SRAM	**STM32H7** 2 MB Flash/~1 MB RAM	**Raspberry Pi Zero** 2 W microSD/512 MB RAM	**Overall Assessment**
ANN – Full	82.80	Suitable*	Suitable	Suitable	Highly suitable	Good
ANN – Selected	70.37	Suitable*	Suitable	Suitable	Highly suitable	Recommended
SVR – Full	8.92	Memory-feasible*	Suitable	Suitable	Highly suitable	High forecasting error
SVR – Selected	4.31	Memory-feasible*	Suitable	Suitable	Highly suitable	Smallest size but high forecasting error
XGBoost – Full	1,252.13	Not suitable	Limited*	Limited*	Suitable	Large model footprint
XGBoost – Selected	1,299.17	Not suitable	Limited*	Limited*	Suitable	No model-size reduction
RF – Full	115,915.39	Not suitable	Not suitable	Not suitable	Suitable*	Excessively large
RF – Selected	104,267.63	Not suitable	Not suitable	Not suitable	Suitable*	Still excessively large

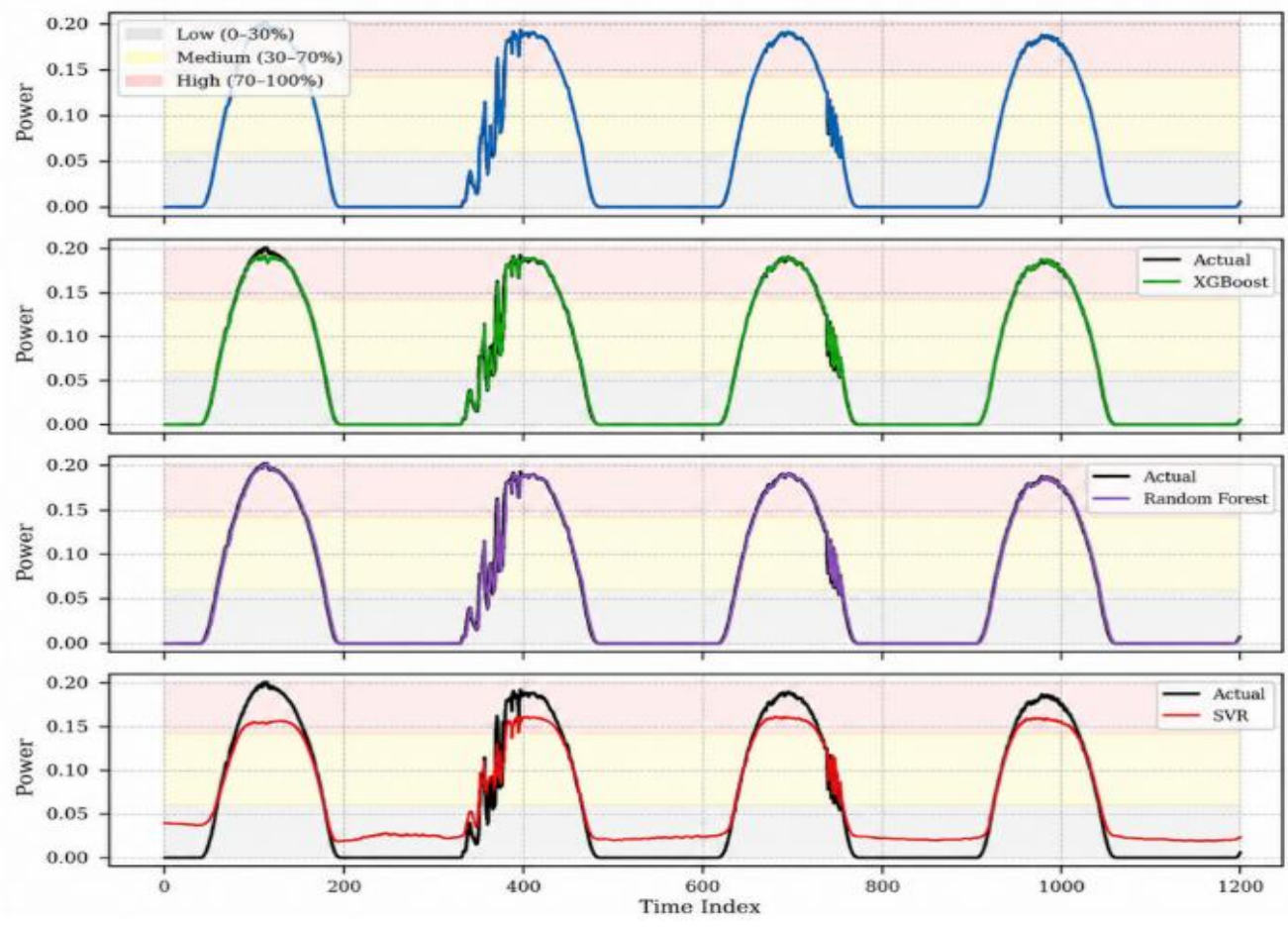


Figure 5 Actual and predicted PV power over time for different models and power levels

The results in Table 2 indicate that, although SVR has the smallest model size, its high prediction error limits its practical applicability. In contrast, XGBoost and Random Forest have significantly larger model sizes, making them less suitable for deployment on resource-constrained microcontroller platforms. The ANN demonstrates the best balance between accuracy and model size; specifically, the Selected Feature configuration occupies only 70.37 KB, making it compatible with the surveyed Edge platforms. The results in Figure 5 demonstrate that the ANN model's predictions closely track actual PV power values, whether under stable operating conditions or during periods of rapid power fluctuation. The Random Forest and SVR models show significant prediction errors.

To more accurately assess the performance of the machine learning models, the power output was divided into three ranges: low power (≤ 30% of peak power), medium power (30% to 70%), and high power (70% to 100%). MAE and RMSE metrics were used to evaluate the models' forecasting performance across these ranges. The results shown in Figure 6 indicate that the ANN maintained low prediction errors across all power ranges, a crucial factor for control systems, whereas both XGBoost and Random Forest exhibited significant errors as power output increased. Finally, SVR demonstrated high error values. These results suggest that aggregate evaluation metrics are insufficient for assessing model performance under varying conditions.

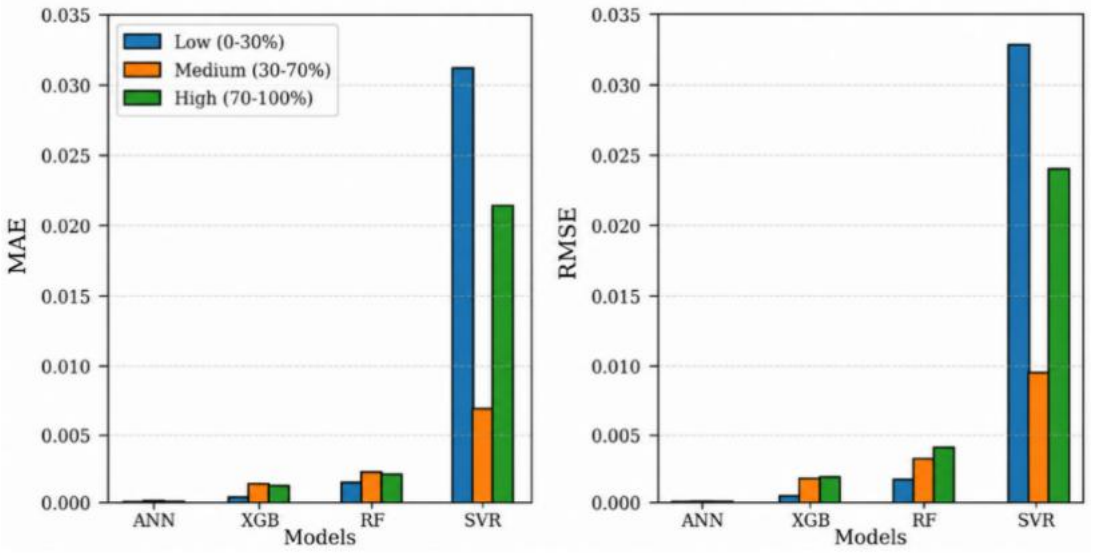


Figure 6 Comparison of MAE and RMSE across different PV power ranges

Figure 7 presents the results of the feature impact assessment for the PV power forecasting model using the SHAP method. The results indicate that time-series features play a crucial role; specifically, moving average features and historical PV power output have a significant influence on forecasting performance. Solar irradiance serves as the next most important input feature for the machine learning model, while weather-related features have a lesser impact. It is concluded that short-term PV power forecasting relies heavily on data regarding the most recent PV power output.

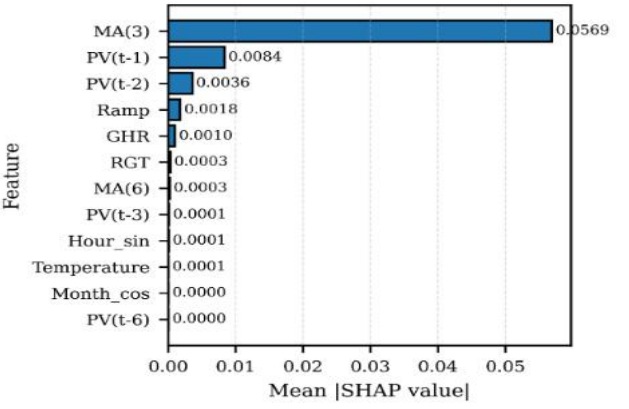


Figure 7 SHAP-based feature importance ranking of the selected forecasting model

The paper presents a combined approach involving feature selection techniques, model performance evaluation across different power output ranges, and error analysis using the SHAP method, thereby providing a comprehensive overview of the performance of all the models.

4 Conclusion

This paper presents a machine learning framework for short-term PV power forecasting. The methodology encompasses data preprocessing, time-series feature engineering, feature selection, and the evaluation of four machine learning models: ANN, SVR, Random Forest, and XGBoost. In addition to using standard error metrics to assess model performance, the study analyzes errors across different power ranges to examine forecasting accuracy under varying power system operating conditions, particularly at high power levels and during periods of rapid power fluctuation. The results indicate that the ANN achieved the highest forecasting accuracy among the tested models. XGBoost and Random Forest also yielded good, stable results, whereas SVR underperformed when the PV power data exhibited high nonlinearity and rapid fluctuations. The application of the SHAP method to analyze the contribution of input features revealed that time-series features, specifically historical PV power data and moving averages, had the greatest impact on forecasting results. Weather and temporal features had a lesser influence but still provided useful information to the models.

Furthermore, reducing the number of input features from 15 to 7, a 53.3% reduction, maintained comparable predictive performance while lowering model complexity. Among the evaluated models, the ANN demonstrated the best balance between prediction accuracy and deployment requirements. Specifically, the post-training size of the ANN model decreased from 82.80 KB (Full Feature) to 70.37 KB (Selected Feature). Although the SVR model was smaller, its significantly higher prediction error limited its practical applicability, whereas XGBoost and Random Forest required substantially more storage space. Consequently, the ANN with the Selected Feature configuration was identified as the preferred choice for PV power forecasting on Edge platforms. Identifying low-contribution input features also provides a basis for potentially eliminating unnecessary measurement channels or sensors in real-world deployments, thereby reducing the complexity and cost of data acquisition systems. Overall, the proposed feature selection method offers a practical approach to developing compact and accurate PV power forecasting models for low-cost edge devices.

References

1. Abdelsalam, H., Souri, A., & İnanç, N. (2025). A time series forecasting approach based on gradient boosting method for IoT-based solar energy production systems. *Energy Storage and Saving, 4(4)*, 413-420. doi:https://doi.org/10.1016/j.enss.2025.04.003.

2. Cosgun, A. E. (2025). Enhancing Photovoltaic Energy Output Predictions Using ANN and DNN: A Hyperparameter Optimization Approach. *Energies, 18(17)*, 4564. doi:10.3390/en18174564.

3. Asghar, R., Fulginei, F. R., Quercio, M., & Mahrouch, A. (2024). Artificial Neural Networks for Photovoltaic Power Forecasting: A Review of Five Promising Models. *IEEE Access, 12*, 90461-90485. doi:10.1109/ACCESS.2024.3420693.

4. Barhmi, K., Heynen, C., Golroodbari, S., & van Sark, W. (2024). A Review of Solar Forecasting Techniques and the Role of Artificial Intelligence. *Solar, 4(1)*, 99-135. doi:10.3390/solar4010005.

5. El-Amarty, N., Fadili, H. E., & Bennani, S. D. (2024). Accurate Short-Term Solar Irradiance Forecasting with TinyML on Edge Device. *Paper presented at the 2024 International Conference on Circuit, Systems and Communication (ICCSC)*.
6. Stojkovic, A., Nikolic, B., Zivkovic, M., & Bacanin, N. (2025). Photovoltaic Farm Production Forecasting: Modified Metaheuristic Optimized Long Short-Term Memory-Based Networks Approach. *IEEE Access*, 13, 25198-25222. doi:10.1109/ACCESS.2025.3537407.
7. Trînc, E.-C., Niţă, V., Stolojescu-Crisan, C., Ancuţi, C., Mihai, R. M., & Sultănoiu, C. P. (2026). Cost-Effective TinyML-Ready Design and Field Deployment of a Solar-Powered Environmental Monitoring Data Collector Using LTE-M Communication. *Applied Sciences*, 16(7), 3237. doi:10.3390/app16073237.
8. Desert Knowledge Australia Solar Centre, Alice Springs. Retrieved from http://dkasolarcentre.com.au/.
9. Gruosso, G., & Gajani, G. S. (2022). Comparison of Machine Learning Algorithms for Performance Evaluation of Photovoltaic Energy Forecasting and Management in the TinyML Framework. *IEEE Access, 10*, 121010-121020. doi:10.1109/ACCESS.2022.3222986.
10. Lappalainen, K., Wang, G. C., & Kleissl, J. (2020). Estimation of the largest expected photovoltaic power ramp rates. *Applied Energy, 278*, 115636. doi:https://doi.org/10.1016/j.apenergy.2020.115636.
11. Zhang, H., Di, X., & Yang, L. (2023). An Improvement of Pearson Correlation Coefficient for Variable Importance Analysis. *Paper presented at the 2023 8th International Conference on Computational Intelligence and Applications (ICCIA)*.
12. Li, Z., Li, J., Ye, X., Xiong, X., & Li, T. (2024). Photovoltaic Power Forecasting Model based on Random Forest and BiLSTM Neural Network. *Paper presented at the 2024 IEEE 7th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*.

Khung học máy cho dự báo ngắn hạn công suất điện mặt trời với lựa chọn đặc trưng và phân tích sai số theo từng vùng công suất

Nguyễn Bảo Quốc[1], Bùi Vũ Minh[2], Hồ A Lil[2], Lâm Hoàng Cát Tiên[1], Trần Ngọc Huy Thịnh[2,*]

[1]Khoa Điện – Điện tử, Trường Cao đẳng Kỹ thuật Cao Thắng, Thành phố Hồ Chí Minh, Việt Nam
[2]Khoa Kỹ thuật – Công nghệ, Trường Đại học Nguyễn Tất Thành, Thành phố Hồ Chí Minh, Việt Nam
[*]tnhthinh@ntt.edu.vn

Tóm tắt Bài báo trình bày một khung học máy cho dự báo công suất quang điện ngắn hạn, nhằm phát triển một mô hình học máy siêu nhỏ phù hợp với các thiết bị biên giá rẻ mà vẫn duy trì độ chính xác. Phương pháp bao gồm tiền xử lý dữ liệu, xây dựng đặc trưng chuỗi thời gian, lựa chọn đặc trưng dựa trên tương quan và mức độ quan trọng, cùng đánh giá hiệu năng của ANN, Random Forest, SVR và XGBoost. Sai số dự báo được phân tích theo các khoảng công suất quang điện khác nhau, trong khi SHAP được sử dụng để diễn giải đóng góp của các đặc trưng. Kết quả cho thấy việc giảm số đặc trưng đầu vào từ 15 xuống 7 (53,3 %) vẫn duy trì hiệu năng tương đương đồng thời giảm độ phức tạp mô hình. ANN đạt độ chính xác tốt nhất và cân bằng tốt nhất giữa độ chính xác và kích thước mô hình, giảm từ 82,80 KB xuống 70,37 KB. Khung đề xuất phù hợp để phát triển các mô hình dự báo quang điện nhỏ gọn, chính xác trên các nền tảng Edge hạn chế tài nguyên.

Từ khóa Học máy siêu nhỏ; dự báo công suất điện mặt trời; lưới điện thông minh phân tích chuỗi thời gian.