# MACHINE LEARNING-BASED ESTIMATION OF POWER OUTPUT IN SOLAR PHOTOVOLTAIC SYSTEMS UNDER REAL-WORLD CONDITIONS

**Kim Anh Nguyen*, Doan Tuan Hung Do, Hoang Khoa Trinh, Ngoc Khai Nguyen, Ngoc Bao Doan**

*The University of Danang - University of Science and Technology, Vietnam*

*Corresponding author: nkanh@dut.udn.vn

(Received: May 05, 2025; Revised: June 11, 2025; Accepted: June 20, 2025)

DOI: 10.31130/ud-jst.2025.23(9D).569E

**Abstract -** Precise prediction of DC power output from photovoltaic (PV) systems under real conditions is essential for improving efficiency and detecting degradation. This paper presents a machine learning framework to forecast PV power using electrical signals (e.g., voltage and current) and environmental data (irradiance and temperature). Three regression models -eXtreme Gradient Boosting (XGBoost), Support Vector Regression (SVR), and Artificial Neural Networks (ANN) - were trained on 13,923 samples from a 455.4 kWp solar PV plant in central Vietnam. The XGBoost model delivered the best performance with $R^2 = 0.9998$, mean absolute error (*MAE*) = 1.62 kWh, and root mean square error (*RMSE*) = 2.209 kWh, outperforming conventional methods. Additionally, the low computational demand of the developed model allows implementation on affordable hardware platforms, such as Raspberry Pi 4, enabling practical real-time monitoring and timely detection of PV performance degradation due to factors like panel defects, natural aging, and dust accumulation.

**Key words -** Solar PV system; power output estimation; time-series data; real-world operating condition; machine learning

## 1. Introduction

Solar energy has become a viable alternative to fossil fuels, with photovoltaic (PV) systems offering clean electricity generation via the photovoltaic effect [1]. Although PV modules degrade at a rate of approximately 0.5% per year under standard test conditions, the real-world performance is influenced by internal (e.g., material degradation and hotspots) and external (e.g., dust and weather) factors. Among these, dust accumulation is the dominant environmental factor; however, it remains difficult to quantify owing to its variability.

Several studies have reported substantial efficiency losses due to dust. At a density of 10 g/m², the efficiency dropped by up to 30% in China [2], while under extreme desert conditions in Iran, a 98.13% decline was observed at 330 g/m² [3]. Mahnoor *et al.* [4] recorded dust-related power losses of 15.08% and 25.42% in subtropical and desert climates, respectively, in Pakistan. The physical and chemical characteristics of dust, such as grain size, color, and absorption properties, also play a role [5], [6], although many studies have not considered outdoor weather dynamics.

Hanchicha *et al.* [7] noted that water-soluble dust components can absorb moisture, forming adhesive layers that crystallize under sunlight. However, the scale of this study is limited. Therefore, regular cleaning is essential, determining the optimal intervals is complex without accurate degradation estimates.

Recent studies have attempted to quantify the impact of dust. Shen *et al.* [8] proposed a hybrid mathematical model using a Poisson process and a Markov chain to capture dust and temperature dynamics. However, it lacks validation and ignores dust–temperature interactions. Machine learning approaches, such as ELM and ANN, have been tested [9] but are constrained by small datasets and indirect dust metrics. Elamin *et al.* [10] analyzed natural accumulation over a year via I–V curves, although the conclusions were limited by sparse data points.

Other modeling strategies include theoretical output estimation [11], long-term extrapolation using regression and ANN [12], and binary dust detection via ANN-triggered cleaning systems [13], all of which present key limitations in realism, variable inclusion, or practicality. More recent studies, such as Ma's overlap model [14] and Mohammed's third-diode modeling [15], have made theoretical advances, but lack real-world integration and field data validation.

Given these challenges, machine learning algorithms (ML) have emerged as promising tools in solar analytics [16], often outperforming physics-based models [17]. This study applies several advanced machine learning models to estimate the DC power output of a solar PV system with high accuracy and robustness, using a dataset that includes both electrical and environmental parameters. The results of this study can serve as input for a system designed to quantify dust-induced power losses by comparing the predicted power with the actual power.

## 2. Methodology

### *2.1. Data collection*

In the study by Long *et al.* [18], three regression algorithms - linear regression, pace regression, and support vector machine regression (SVM regression) - were employed to assess the influence of environmental parameters on photovoltaic (PV) power estimation. The results consistently indicated that ambient temperature and solar irradiance were the most critical factors affecting forecasting accuracy, whereas other variables had negligible impacts on model training performance. Based on these findings, our study adopts ambient temperature and solar irradiance in time series as the primary input features for model training.

### *2.2. Support vector machine model*

Support vector regression (SVR) is a machine learning method developed by Smola and Schölkopf [19], based on the SVM framework originally proposed by Vapnik [20].

Unlike SVM, which focuses on classification tasks, SVR is designed for regression and aims to construct an accurate model capable of mapping complex input data to continuous output values. SVR performs well even on non-linear datasets by maintaining high prediction accuracy.

$$\left|y_i - f\left(x_i\right)\right|_{\varepsilon} = \max\left(0, \left|y_i - f(x_i)\right| - \varepsilon\right) \tag{1}$$

SVR constructs a regression model by defining an ε-insensitive zone (ε-tube) around the true output values. In this tube, prediction errors between the actual values. The SVR model begins with a linear prediction function in the feature space:

$$f\left(x\right) = \left\langle \omega, x \right\rangle + b \tag{2}$$

or in higher-dimensional feature space:

$$f\left(x\right) = \left\langle \omega, \phi\left(x\right) \right\rangle + b \tag{3}$$

where, $\omega$ is the weight vector in feature space; $\phi\left(x\right)$ is a non-linear mapping from input space to feature space; $b$ is the bias term used to shift the separating hyperplane. To train the model, SVR solves a convex optimization problem to minimize the trade-off between model complexity and prediction error:

$$\min \frac{1}{2}\left\|\omega\right\|^2 + C\sum_{i=1}^{n}\left|y_i - f(x_i)\right|_{\varepsilon} \tag{4}$$

Subject to $y_i - \left\langle \omega, x_i \right\rangle - b \le \varepsilon + \xi_i^*$, $\left\langle \omega, x_i \right\rangle + b - y_i \le \varepsilon + \xi_i$, and $\xi_i, \xi_i^* \ge 0$. Where, $C$ is the regularization parameter controlling the trade-off between model complexity $\frac{1}{2}\left\|\omega\right\|^2$ and tolerance to deviation ($\xi_i + \xi_i^*$); $\varepsilon$ defines the width of the $\varepsilon$-tube within which no penalty is assigned to errors; $\xi_i$ and $\xi_i^*$ are slack variables allowing data points to fall outside the ε-tube with penalties, thus improving robustness against noise or outliers.

Kernels are a fundamental component enabling SVR to effectively address problems involving non-linear relationships between input and output data. A kernel operates by mapping the original input data into a higher-dimensional feature space - where non-linear relationships can be represented linearly. In SVR, instead of directly computing the complex mapping into this high-dimensional space, the kernel trick is employed to compute inner products between feature vectors implicitly via a kernel function and defined as follows:

$$k\left(x_i, x_j\right) = \left\langle \phi(x_i), \phi(x_j) \right\rangle \tag{5}$$

### 2.3. eXtreme Gradient Boosting Model

XGBoost (eXtreme Gradient Boosting), developed by Chen and Guestrin [21], is an advanced machine learning algorithm based on the Gradient Tree Boosting framework. It is designed to solve regression and classification problems with high computational efficiency and prediction accuracy. The algorithm focuses on optimizing a regularized learning objective by leveraging second-order gradient information. Thanks to its capability to handle large datasets, mitigate overfitting, and maximize computational efficiency, XGBoost has become a powerful tool in machine learning - well-suited for a wide range of tasks beyond conventional time-series forecasting.

XGBoost constructs an ensemble of decision trees in an additive manner, where each successive tree refines the predictions made by the previous trees. The predicted value $\hat{y}_i^{(t)}$ for a sample $x_i$ at time $t$ is calculated as:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k\left(x_i\right) = \hat{y}_i^{(t-1)} + f_t\left(x_i\right) \tag{6}$$

where, $f_k\left(x_i\right)$ is the prediction function of the $k$ tree. XGBoost optimizes the following objective function at step $t$:

$$G^{(t)} = \sum_{i=1}^{n} l\left(\hat{y}_i^{\,t}, y_i\right) + \sum_{k=1}^{t}\Omega(f_k) \tag{7}$$

where, $n$ is the number of training samples; $y_i$ is the actual target value; $l\left(\hat{y}_i, y_i\right)$ is the loss function; $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\left\|w\right\|^2$ is the regularization term for the $k$ tree. In this expression, $T$ the number of leaves in the tree, reflecting its structural complexity; $w$ a vector of prediction scores assigned to each leaf; $\gamma$ a regularization parameter penalizing model complexity; $\gamma$ the regularization coefficient on the leaf weights $w$, used to prevent overfitting.

At each iteration $t$, a new tree $f_t$ is added to minimize the second-order Taylor expansion of the objective function:

$$G^{(t)} = \sum_{i=1}^{n}\left[g_i f_t\left(x_i\right) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t) \tag{8}$$

where, $g_i = \partial_{y_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}$ is first-order gradient and $h_i = \partial^2_{y_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ is second-order gradient.

### 2.4. Model evaluation metrics

In this study, three commonly used statistical indicators are employed to evaluate the performance of machine learning models [22], [23] such as *RMSE* (root mean square error), *MAE* (mean absolute error), and $R^2$ (coefficient of determination).

*RMSE* measures the average magnitude of the squared differences between predicted values and actual values, followed by taking the square root of the result. It emphasizes larger errors due to the squaring operation, making it particularly sensitive to outliers and is defined as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(H_{i,m} - H_{i,c}\right)^2} \tag{9}$$

Where, $n$ denotes number of observations; $H_{i,c}$ is predicted value; and $H_{i,m}$ denotes actual (measured) value.

*MAE* is an index that measures the average of the absolute errors between the predicted values (calculated/estimated values) and the actual values (measured/actual values). It provides a simple way to evaluate the accuracy level of the prediction model without being affected by the sign of the error (positive or negative) and is calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|H_{i,c} - H_{i,m}\right| \qquad (10)$$

Where, $n$ represents number of observations; $H_{i,c}$ refers to predicted value; and $H_{i,m}$ is actual value.

$R^2$ is a statistical index that measures the extent to which the prediction model can explain the variability of the actual data. It indicates the percentage of variation in the actual values that is explained by the predicted model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(H_{i,m} - H_{i,c})}{\sum_{i=1}^{n}(H_{i,m} - H_{m,ave})^2} n \qquad (11)$$

Where, $n$ is number of observations; $H_{i,m}$ is actual value; $H_{m,ave}$ is mean of the actual values; and $H_{i,c}$ is predicted value. $R^2$ value ranges from 0 to 1. The closer $R^2$ is to 1, the better the model is at explaining the variability of the actual data, and the $R^2$ value closer to 0 indicates that the model fails to explain any variation in the data.

## 3. Experiments and discussion

To enable the training of a model capable of reliably estimating the power output of a specific PV system under given irradiance and environmental conditions, it is essential to collect data under relatively clean-panel conditions. In this study, the PV panels were cleaned regularly, on average once every two weeks, to ensure this requirement is met. The proposed model was trained and tested on a dataset comprising 13,923 samples collected from a 455.4 kWp solar PV plant located in central Vietnam. The system comprises 828 Jinko JKM550M-72HL4 photovoltaic modules, each with a rated power of 550 Wp, resulting in a total nominal capacity of 455.4 kWp. The system is equipped with three Sungrow SG125CX-P2 inverters (each rated at 125 kVA), which are connected to the electric grid.

The environmental parameters at the plant site were collected through a sensor network integrated into the monitoring architecture of the project. Specifically, a Pyranometer SMP10 was employed to measure global solar irradiance, while a TA-EXT-RS485-E sensor was used to record ambient temperature. These sensors were installed coplanar with the PV modules to ensure that the measured data accurately reflected the actual operating conditions of the system.

Monitoring data were acquired using a MOXA UC-8100 data logger, which interfaced with the Sungrow SG125CX-P2 inverters via the Modbus RTU protocol over an RS485 communication line. The collected parameters included voltage, current, output power, power factor, and grid frequency. All data were transmitted to a centralized SCADA system via a 4G LTE RUT955 router, enabling real-time analytics, performance monitoring, fault detection, and predictive maintenance planning.

The data were sampled at 15-minute intervals over a period from January 2024 to September 2024. This dataset included two dominant weather conditions: clear skies and overcast skies.

The preprocessing phase involved removing duplicate records through consistency checks and eliminating entries with zero AC output power, as well as days during which the system was offline. These steps were taken to minimize the influence of anomalous or irrelevant data. The cleaned dataset was subsequently partitioned into two subsets: 80% for training the machine learning models and 20% for evaluating forecasting performance.

In the remainder of this section, the paper focuses on presenting and analyzing the training results of two prominent machine learning models including XGBoost and SVR-linear, which were used to estimate the output power of the photovoltaic (PV) system. The input data include environmental temperature, solar irradiance, and time, with the initial step involving data filtering to ensure consistency. The training process was carried out on actual measured data, and model performance was evaluated using statistical indicators such as *RMSE*, *MAE*, and *R²*.

### *3.1. Training results of XGBoost model*

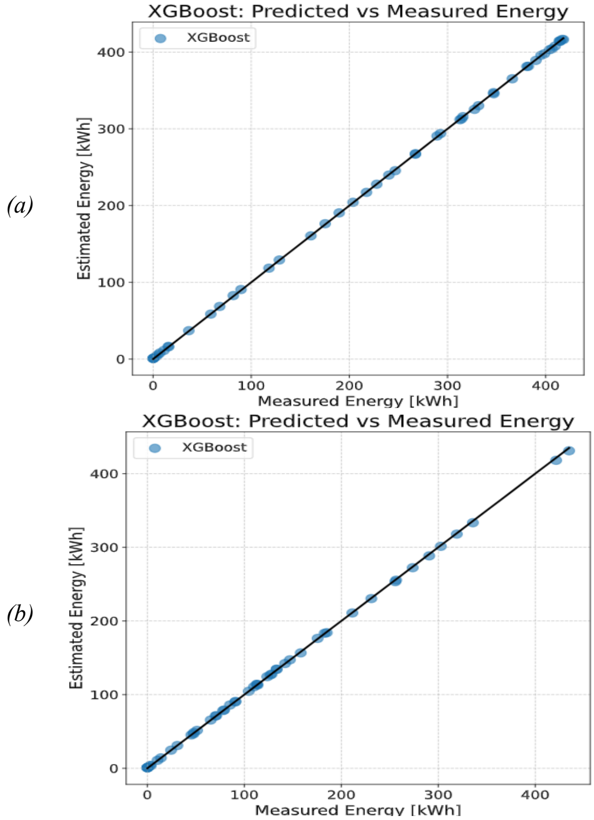


***Figure 1.*** *Scatter plots of the XGBoost model under conditions a) sunny day with high temperature and b) clody day with low temperature*

The XGBoost algorithm, due to its iterative boosting structure and its ability to model complex nonlinear relationships, was implemented to predict the power output of a solar energy system. After optimizing the hyperparameters, the model achieved superior performance indicators: *RMSE* was 1418.196 Wh, *MAE* was 1030.6 Wh, and $R^2$ reached 0.999875. Figure 1 presents scatter plots comparing predicted and measured energy output (in kWh) over multiple time intervals. The data points demonstrate a strong linear correlation, indicating near-perfect prediction quality. The model maintained this consistency across various operating conditions, two representative days are shown: a sunny day with high temperature in Figure 1a, and a cloudy day with low temperature in Figure 1b. This stability emphasizes the model effectiveness and generalizability in real-world settings, where weather factors are continuously variable and difficult to predict.

A detailed error analysis is presented in Figure 2, corresponding to the sunny day with high temperature in Figure 2a and the cloudy day with low temperature in Figure 2b, highlighting the model's consistent performance over a 24-hour cycle. The top graph in Figure 2 compares actual and predicted energy values (in Wh), where the deviation is relatively small. The model performs reliably across both peak and low-load periods, with small variance in prediction errors. The average absolute error during the analyzed period in Figure 2 is about 580 Wh, a very small figure compared to daily solar energy output of approximatelt 455.4 kWh.

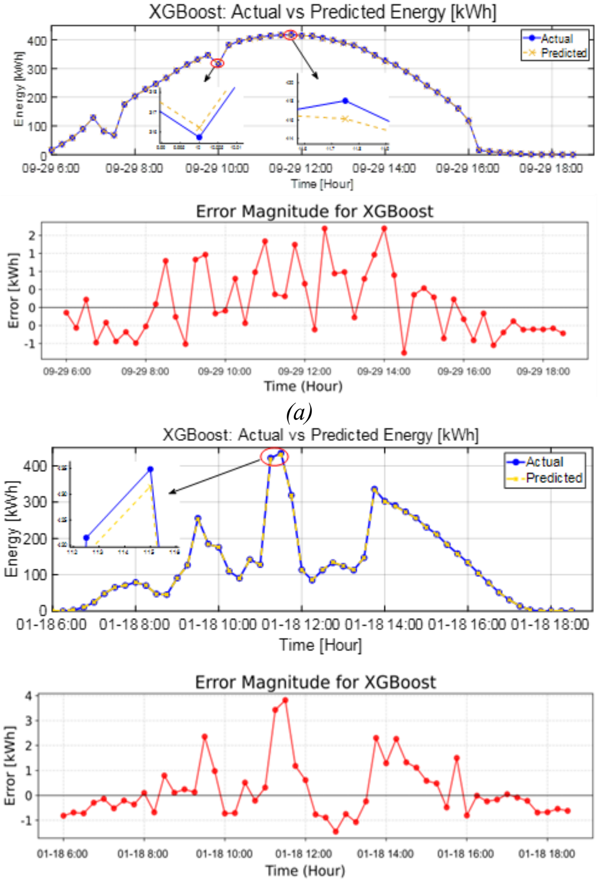


***Figure 2.*** *Estimated power and actual power by the XGBoost model a) sunny day with high temperature and b) clody day with low temperature*

### *3.2. Training results of SVR-linear model*

The SVR-linear model, a variant of SVM using a linear kernel, was implemented to predict energy output from a photovoltaic (PV) system. After being trained and fine-tuned, the model achieved the following notable performance metrics: the *RMSE* was 2259.284 Wh, the *MAE* was 1776.326 Wh, and the Coefficient of Determination $R^2$ reached 0.999675.

Similar to Figure 1, Figure 3 displays scatter plots comparing predicted and actual energy output (in kWh) across multiple distinct datasets for SVR-linear model under conditions sunny day with high temperature (Figure 3a) and clody day with low temperature (Figure 3b). The data points are clustered closely around the $y = x$ line, indicating a high degree of agreement between the predicted and measured values. This consistency was maintained across various operating conditions, from sunny to cloudy days, affirming the adaptability and reliability of the model in real-world scenarios - where environmental factors such as solar irradiance and ambient temperature can vary rapidly.

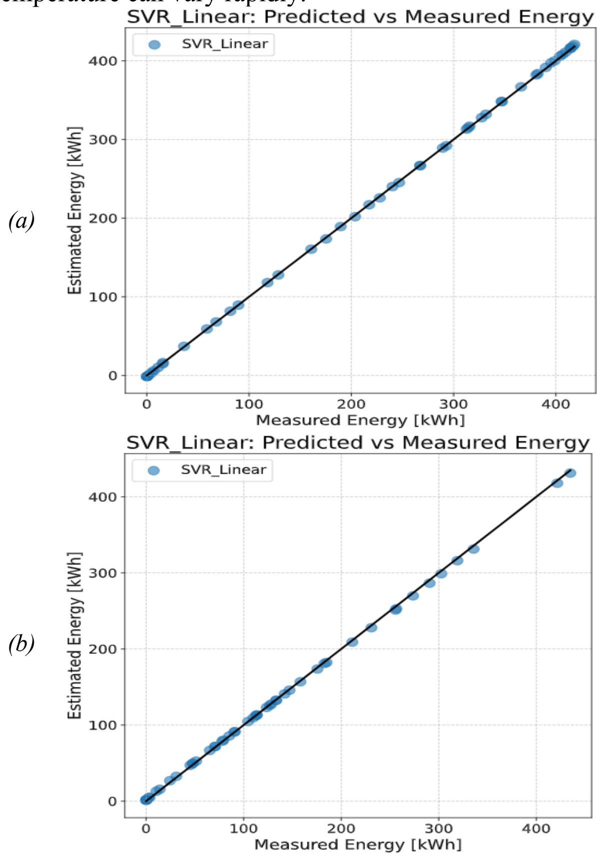
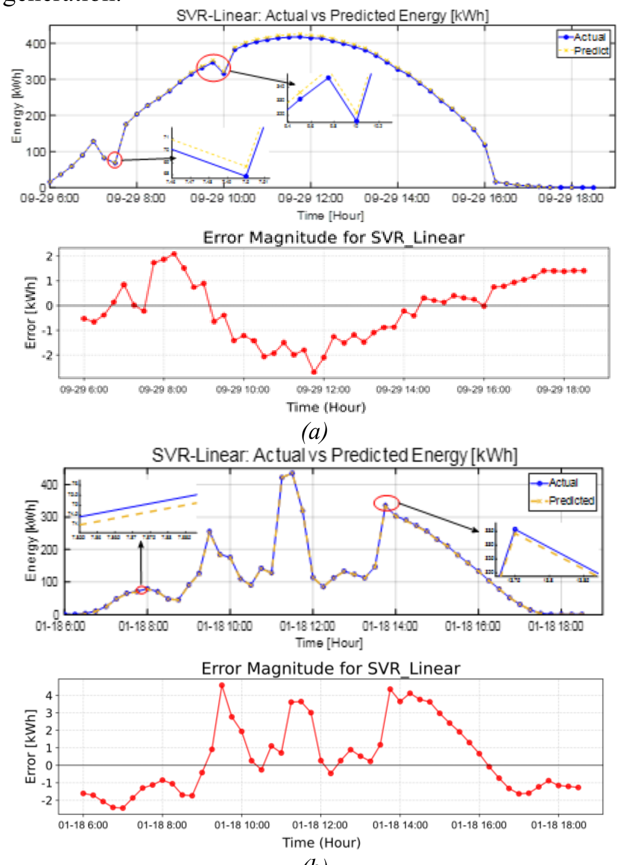


***Figure 3.*** *Scatter plots of the SVR-linear model under conditions a) sunny day with high temperature and b) clody day with low temperature*

A more detailed error analysis is provided in Figure 4, offering deeper insights into the model's performance over time. The top row of each subfigure (e.g., Figures 4a and 4b) compares the predicted and actual energy values (in Wh), while the bottom row displays the absolute prediction errors. The SVR-Linear model demonstrates stable performance across time intervals, with error levels remaining low - even during periods of power output transition, particularly under varying irradiance conditions. The average absolute error during the analyzed time period is negligible compared to the system's total daily generation.



(b)

***Figure 4.*** *Estimated power and actual power by SVR-linear model a) sunny day with high temperature and b) clody day with low temperature*

## 4. Model Validation

In this section, the study conductes to compare the performance of six models used to estimate the power output of the PV system, including two prominent models (XGBoost and SVR-Linear) alongside four other models, such as CatBoost, LightGBM, SVR-RBF, and SVR-Poly. The evaluation is conducted on datasets that represent different weather conditions.

Table 1 presents the performance metrics of the six ML models evaluated on two representative days, randomly selected to reflect contrasting weather conditions: January 30, 2024 (sunny with high temperatures), and September 30, 2024 (cloudy with low temperatures), the following observations can be made:

- **XGBoost** showed almost perfect precision, with its $R^2$ being very close to 1.0 with the lowest RMSE, reflecting good generalization and effective reduction of overfitting, especially for non-linear trends;

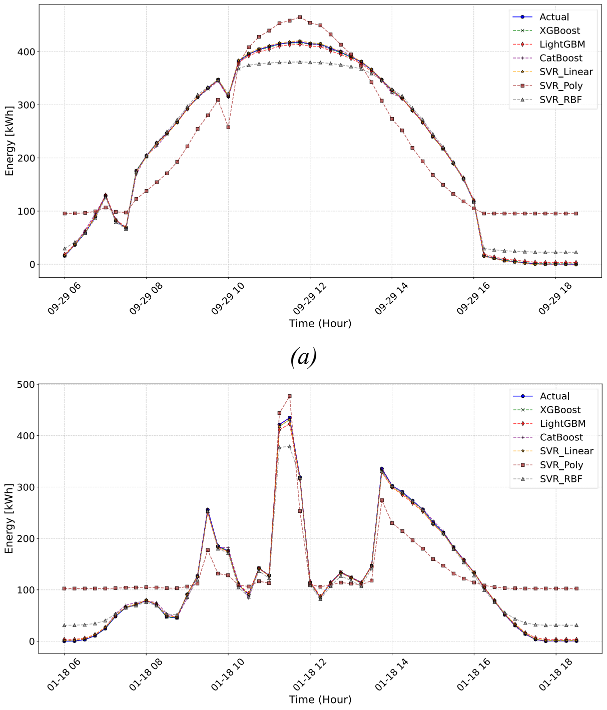
- When comparing **CatBoost** and **LightGBM**, despite being both boosting models, LightGBM showed 1.5–2.5 times larger errors compared to XGBoost, especially on January 30, reflecting greater sensitivity to data variability;

- **SVR-linear** performed steadily, with an $R^2$ above 0.999 and an *MAE* below 2000, ideally matching the linear trends of the dataset;

- **SVR-RBF** and **SVR-Poly**, however, performed extremely badly, with *RMSE* above 19,000 and 60,000, respectively, indicating severe overfitting due to an excessive model complexity.

***Table 1.*** *Performance metrics of predicting models under different weather conditions*

| ML model | Metrics | January 30, 2024 | September 30, 2024 |
|---|---|---|---|
| **XGBoost** | $R^2$ | 0.9998 | 0.9999 |
| | *MAE* | 1623.8674 | 928.2887 |
| | *RMSE* | 2209.5434 | 1416.5921 |
| **LightGBM** | $R^2$ | 0.9986 | 0.9995 |
| | *MAE* | 4429.0955 | 2625.5064 |
| | *RMSE* | 5486.8042 | 3294.4017 |
| **CatBoost** | $R^2$ | 0.9996 | 0.9997 |
| | *MAE* | 2111.964 | 1731.0408 |
| | *RMSE* | 2912.9823 | 2419.6645 |
| **SVR-Linear** | $R^2$ | 0.9994 | 0.9998 |
| | *MAE* | 3158.037 | 1207.684 |
| | *RMSE* | 3601.1452 | 2133.896 |
| **SVR-Poly** | $R^2$ | 0.8165 | 0.8127 |
| | *MAE* | 55988.8487 | 57153.0592 |
| | *RMSE* | 63166.4559 | 63248.0781 |
| **SVR-RBF** | $R^2$ | 0.9883 | 0.9825 |
| | *MAE* | 12559.3511 | 11420.7179 |
| | *RMSE* | 15961.5189 | 19344.2833 |



(b)

***Figure 5.*** *Estimated power versus actual output power across six ML models under conditions a) sunny day with high temperature and b) clody day with low temperature*

Figure 5 shows estimated power versus the PV system's actual power output across six ML models under conditions such as a sunny day with high temperature (Figure 5a) and a clody day with low temperature (Figure 5b). In Figure 5a, the actual power curve exhibits a clear "bell-shaped" pattern with a peak at noon. Boosting models such as XGBoost and CatBoost closely track the peak values, while SVR-RBF and SVR-Poly show phase shift and significant error at peak irradiance hours. In Figure 5b, continuous variations in irradiance pose a challenge to all models, especially those using nonlinear kernels such as SVR-RBF and SVR-Poly, which struggle to adapt to sudden changes.

From the obtained results as shown in Figure 5, three main patterns emerge:

**- XGBoost** and **CatBoost**: Deliver the most stable results in terms of curve shape and peak tracking, demonstrating strong nonlinear learning capacity and superior generalization typical of Boosting models;

**- SVR-linear**: Performs well under low-temperature or stable irradiance conditions, aligning with its assumption of linear data structure. However, its prediction error increases considerably in the presence of data volatility;

**- SVR-RBF** and **SVR-Poly**: Exhibit the highest errors and lowest stability, particularly when irradiance conditions change abruptly. This highlights the suboptimal performance of nonlinear kernels under the current dataset.

To evaluate the effectiveness and novelty of the applied models relative to previous work, we also re-implemented the Artificial Neural Network (ANN) model from the study by Asiedu *et al.* [24], which achieved optimal performance for one-day-ahead PV power forecasting with $R^2$ = 0.8702, *MAE* = 0.3043 kWh và *RMSE* = 0.7477 on a 180 kWp system in Ghana, and retrained it using actual data from January 30 collected from the 455.4 kWp PV system in Quang Nam, Vietnam. The training results are presented in Table 2.

***Table 2.*** *Performance comparison of ANN [24] and proposed models*

| Models | *RMSE* (Wh) | *MAE* (Wh) | $R^2$ |
|---|---|---|---|
| **XGBoost** | 2,209.54 | 1,623.87 | 0.9998 |
| **SVR-Linear** | 3,601.15 | 3,158.04 | 0.9994 |
| **ANN** [24] | 12,391.48 | 8,209.93 | 0.9929 |

Although the ANN model exhibited acceptable performance on its original system, the retraining results showed a significant decline in accuracy when applied to local data: *RMSE* reached 12,391.48 Wh, was 8,209.93 Wh, and the coefficient of determination R2 was 0.9929. In contrast, XGBoost demonstrated superior performance, with an *RMSE* of only 2,209.54 Wh (an 82% reduction compared to ANN) and $R^2$ = 0.9998. SVR-linear also outperformed ANN with an *RMSE* of 3,601.15 Wh and $R^2$ = 0.9994.

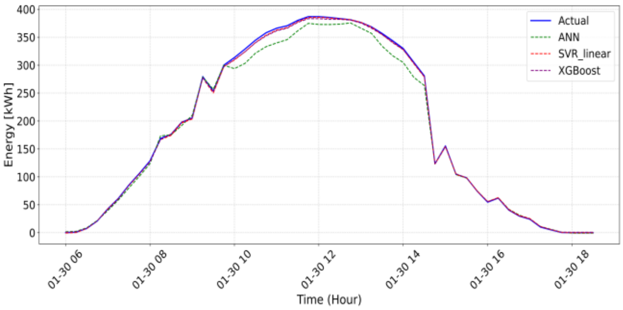


***Figure 6.*** *Estimated power of the ANN model [24] versus the proposed models*

Figure 6 shows that the XGBoost model closely follows the actual power output, especially during rapid rise-and-fall periods in the morning and afternoon. Conversely, ANN tends to underpredict at peak times, reflecting its slower adaptation to rapid fluctuations.

The obtained results affirm the reliability and adaptability of Boosting-based models, particularly XGBoost, when applied to PV power forecasting tasks in dynamic tropical weather conditions.

## 5. Conclusion

This study proposed, trained, and evaluated two promising machine learning models, namely XGBoost and Support Vector Regression (SVR), achieving high predictive accuracy. These models were applied to estimate the DC power output of a solar PV system using a dataset comprising 13,923 samples, which included electrical parameters of the PV system, module temperature, and environmental parameters collected from a real-world 455.4 kWp installation located in central Vietnam.

The results demonstrated that the XGBoost model outperformed other models, achieving a remarkably high coefficient of determination R² of 0.9998, a low MAE of 1.62 kWh, and a RMSE of 2.209 kWh. These findings indicate that XGBoost not only exhibits strong predictive capability but is also sensitive to environmental condition variations. Although not as accurate as XGBoost, the linear SVR model still provided reliable results when the input features exhibited linear characteristics; however, its performance degraded under noisy or nonlinear conditions. Additionally, further comparative experiments showed that both XGBoost and SVR offered better predicted performance and robustness than other models such as LightGBM, CatBoost, and artificial neural networks (ANNs).

Currently, this work is limited to model selection and evaluation, with the aim of achieving the highest possible accuracy in predicting DC power output. The proposed models have not yet been implemented or integrated into a real-time monitoring platform for assessing dust-induced power losses. Consequently, future research will focus on validating and deploying the selected model within embedded systems to enable real-time estimation of performance degradation and dust-related losses in PV systems.

## REFERENCES

[1] D. C. Jordan, S. R. Kurtz, K. VanSant, and J. Newmiller, “Compendium of photovoltaic degradation rates”, *Progress in Photovoltaics: Research and Applications*, vol. 24, no. 7, pp. 978–989, 2016. https://doi.org/10.1002/pip.2744

[2] Y. Chen, Y. Liu, Z. Tian, Y. Dong, Y. Zhou, X. Wang, and D. Wang, “Experimental study on the effect of dust deposition on photovoltaic panels”, *Energy Procedia*, vol. 158, pp. 483–489, 2019. https://doi.org/10.1016/j.egypro.2019.01.139

[3] S. A. Sadat, J. Faraji, M. Naziffard, and A. Ketabi, “The experimental analysis of dust deposition effect on solar photovoltaic panels in Iran’s desert environment”, *Sustainable Energy Technologies and Assessments*, vol. 47, 2021. https://doi.org/10.1016/j.seta.2021.101542

[4] M. Rashid, M. Yousif, Z. Rashid, A. Muhammad, M. Altaf, and A. Mustafa, “Effect of dust accumulation on the performance of photovoltaic modules for different climate regions”, *Heliyon*, vol. 9, no. 12, e23069, 2023. https://doi.org/10.1016/j.heliyon.2023.e23069

[5] Z. A. Darwish, K. Sopian, and A. Fudholi, “Reduced output of photovoltaic modules due to different types of dust particles”, Journal of Cleaner Production, vol. 280, 124317, 2021. https://doi.org/10.1016/j.jclepro.2020.124317

[6] T. M. A. Alnasser, A. M. J. Mahdy, K. I. Abass, M. T. Chaichan, and H. A. Kazem, “Impact of dust ingredient on photovoltaic performance: An experimental study”, *Solar Energy*, vol. 195, pp. 651–659, 2020. https://doi.org/10.1016/j.solener.2019.12.008

[7] A. A. Hachicha, I. Al-Sawafta, and Z. Said, “Impact of dust on the performance of solar photovoltaic (PV) systems under United Arab Emirates weather conditions”, *Renewable Energy*, vol. 141, pp. 287–297, 2019. https://doi.org/10.1016/j.renene.2019.04.004

[8] Y. Shen, M. Fouladjiard, and A. Grall, “Impact of dust and temperature on photovoltaic panel performance: A model-based approach to determine optimal cleaning frequency”, *Heliyon*, vol. 10, e25390, 2024. https://doi.org/10.1016/j.heliyon.2024.e25390

[9] W. Al-Kouz, S. Al-Dahidi, B. Hammad, and M. Al-Abed, “Modeling and analysis framework for investigating the impact of dust and temperature on PV systems’ performance and optimum cleaning frequency”, *Applied Sciences*, vol. 9, no. 7, 1397, 2019. https://doi.org/10.3390/app9071397

[10] A. Elamim, S. Sarikh, B. Hartiti, A. Benazzouz, S. Elhamaoui, and A. Ghennioui, "Experimental studies of dust accumulation and its effects on the performance of solar PV systems in Mediterranean climate", *Energy Reports,* vol. 11, pp. 2346-2359, 2024. https://doi.org/10.1016/j.egyr.2024.01.078

[11] A. A. Babatunde, S. Abbasoglu, and M. Senol, “Analysis of the impact of dust, tilt angle and orientation on performance of PV plants”, *Renewable and Sustainable Energy Reviews*, vol. 90, pp. 1017–1026, 2018. https://doi.org/10.1016/j.rser.2018.03.102

[12] İ. Kayri and M. T. Bayar, “A new approach to determine the long-term effect of efficiency losses due to different dust types accumulation on PV modules with artificial neural networks”, *Journal of Cleaner Production*, vol. 434, 140282, 2023. https://doi.org/10.1016/j.jclepro.2023.140282

[13] S. Hossain, A. M. Arika, I. N. Fahim, J. Uddin, A. Ahmed, H. J. Apon, and M. A. Hoque, “Enhancing solar panel performance: A machine learning approach to dust detection and automated water sprinkle-based cleaning strategy”, *Solar Energy*, vol. 287, 113240, 2025. https://doi.org/10.1016/j.solener.2025.113240

[14] M. Ma, Z. Li, W. Ma, R. Zhang, and X. Zhou, “Comprehensive investigation for power degradation of dust-covered photovoltaic modules based on the overlap model: A case study”, *Solar Energy*, vol. 291, 113389, 2025. https://doi.org/10.1016/j.solener.2025.113389

[15] M. H. Qais, H. M. Hasanein, and S. Alghuwainem, “Identification of electrical parameters for three-diode photovoltaic model using analytical and sunflower optimization algorithm”, *Applied Energy*, vol. 250, pp. 109–117, 2019. https://doi.org/10.1016/j.apenergy.2019.05.013

[16] J. F. Gaviria, G. Narváez, C. Guillen, L. F. Giraldo, and M. Bressan, “Machine learning in photovoltaic systems: A review”, *Renewable Energy*, vol. 196, pp. 298–318, 2022. https://doi.org/10.1016/j.renene.2022.06.015

[17] R. A. A. Ramadhan, Y. R. J. Heatubun, S. F. Tan, and H.-J. Lee, “Comparison of physical and machine learning models for estimating solar irradiance and photovoltaic power”, *Renewable nergy*, vol. 178, pp. 1006–1019, 2021. https://doi.org/10.1016/j.renene.2021.06.079

[18] H. Long, Z. Zhang, and Y. Su, “Analysis of daily solar power prediction with data-driven approaches”, *Applied Energy*, vol. 126, pp. 29–37, 2014. https://doi.org/10.1016/j.apenergy.2014.03.084

[19] A. J. Smola and B. Schölkopf, “A tutorial on support vector regression”, *Statistics and Computing*, vol. 14, pp. 199–222, 2004. https://doi.org/10.1023/B:STCO.0000035301.49549.88

[20] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd edition. New York, NY: Springer-Verlag, 2000. https://doi.org/10.1007/978-1-4757-3264-1

[21] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system”, in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785

[22] M. Jobayer, M. A. H. Shaikat, M. N. Rashid, and M. R. Hasan, “A systematic review on predicting PV system parameters using machine learning”, *Heliyon*, vol. 9, e16815, 2023. https://doi.org/10.1016/j.heliyon.2023.e16815

[23] S. G. Gouda, Z. Hussein, S. Luo, and Q. Yuan, “Model selection for accurate daily global solar radiation prediction in China”, *Journal of Cleaner Production*, vol. 221, pp. 132–144, 2019. https://doi.org/10.1016/j.jclepro.2019.02.211

[24] S. T. Asiedu, F. K. A. Nyarko, S. Boahen, F. B. Effah, and B. A. Asaaga, “Machine learning forecasting of solar PV production using single and hybrid models over different time horizons”, *Heliyon*, vol. 10, No. 7, e28898, 2024. https://doi.org/10.1016/j.heliyon.2024.e28898