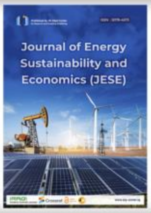
Al-Sibd Center
for Research and Scholarly Publishing

Journal of Energy Sustainability and Economics (JESE)

Journal Homepage : https://jese.srp-center.iq/



Experimental Data–Driven Machine Learning Prediction of Photovoltaic Panel Temperature

Baydaa Adnan Husseina[1] ,Ahmed Ameen Alib[2] , Firas Abdulamir Radhic[3]

[1]Petroleum Engineering Department, College of Engineering, University of Thi-Qar, Thi-Qar 64001, Iraq.
[2]Mechanical Engineering Department, College of Engineering, University of Thi-Qar, Thi-Qar 64001, Iraq.
[3]Solar Energy Research Institute, Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia.

Corresponding Author : ahmedameen@utq.edu.iq

Keywords:

Photovoltaic panel temperature; Machine learning; Solar irradiance; solar cell; weather conditions;

Abstract

Accurate prediction of photovoltaic (PV) panel surface temperature is essential for evaluating thermal behaviour and minimizing efficiency losses, particularly in hot and arid regions. This study presents an experimental and data-driven approach for predicting PV surface temperature using locally measured meteorological data from Nasiriyah City, southern Iraq. Field measurements were conducted over five consecutive days in early July under real outdoor conditions. The PV surface temperature was measured directly using thermocouples fixed on the front surface of the panels, while solar irradiance and ambient temperature were recorded using a solar irradiance meter and a data logger.

The experimentally measured dataset was used to develop and evaluate three regression-based machine learning models: Linear Regression, Random Forest, and Gradient Boosting. A time-based validation strategy was adopted to reflect realistic operating conditions, and model performance was assessed using the coefficient of determination (R^2) and mean absolute error (MAE). The results showed that Linear Regression achieved the highest prediction accuracy, with an R^2 value of 0.9999 and an MAE of 0.028 °C, indicating a strong linear dependency between PV surface temperature and the selected meteorological variables. Although Random Forest and Gradient Boosting models also demonstrated good predictive capability, their higher error values suggest that increased model complexity did not improve performance for the investigated thermal regime.

The findings confirm that PV thermal behavior under steady outdoor conditions is predominantly governed by linear heat transfer mechanisms. Accordingly, simple and physics-consistent models can provide reliable and efficient temperature prediction for PV systems operating in hot-climate environments.

Introduction

Photovoltaic (PV) systems are widely recognized as a key renewable energy technology; however, their performance is strongly influenced by operating temperature [1]. Elevated PV surface temperatures increase thermal losses and lead to a reduction in electrical efficiency, a challenge that is particularly severe in hot and arid regions such as southern Iraq [2]. Under such climatic conditions, PV modules are frequently exposed to high solar irradiance and extreme ambient temperatures, making accurate prediction of PV surface temperature essential for evaluating thermal behavior and supporting effective system operation[3].

Traditionally, PV temperature has been estimated using empirical and semi-empirical thermal models, including NOCT- and Faiman-type formulations, which relate module or cell temperature to meteorological parameters such as solar irradiance, ambient temperature, and wind speed. Recent assessments have shown that incorporating wind effects significantly improves steady-state temperature prediction, and that international standards have gradually evolved toward formulations closely related to Faiman-type models under specified boundary conditions[4]. These approaches remain attractive due to their simplicity and clear physical interpretation; however, their accuracy may be limited when applied to specific local environments without proper calibration.

In parallel, data-driven and machine learning (ML) techniques have gained increasing attention for PV temperature prediction. Keddouda et al. (2024) evaluated several ML algorithms for PV module temperature prediction using ambient temperature, solar radiation, wind speed, and relative humidity, reporting high predictive accuracy and emphasizing the dominant role of meteorological variables[5]. Similarly, Bailek et al. (2020) developed regression-based correlations to estimate PV back-surface temperature under hot and dry outdoor conditions, demonstrating the feasibility of reliable temperature prediction in arid climates[6]. Beyond purely data-driven approaches, hybrid strategies have also been proposed; for instance, Pombo et al. (2022) combined physics-based thermal formulations with neural network corrections to improve PV cell temperature prediction, highlighting the potential advantages of integrating physical insight with machine learning models [7].

Despite these advances, several gaps remain in the existing literature. First, there is a limited number of studies based on experimentally measured PV surface temperature data collected under real outdoor conditions in Iraqi cities, particularly during extreme summer periods. Second, many studies focus on improving prediction accuracy using increasingly complex algorithms, while fewer investigations critically assess whether such complexity is justified when the underlying thermal behavior of PV panels is predominantly linear. Consequently, the practical benefit of advanced ensemble-based machine learning models compared to simple, physics-consistent linear approaches remains insufficiently explored, especially at the city scale under extreme climatic conditions.

This study aims to develop and evaluate machine learning models for predicting the surface temperature of photovoltaic panels using experimentally measured meteorological data from Nasiriyah City, southern Iraq, during the early part of July. Specifically, Linear Regression, Random Forest, and Gradient Boosting models are implemented and compared using solar irradiance, ambient temperature, and wind speed as input variables. The study seeks to assess the prediction accuracy of each model and to examine whether increased model complexity provides tangible performance improvements over simple, physics-consistent regression models under hot-climate operating conditions.

Methodology

Study Area and Data Description

The experimental setup used for data acquisition is illustrated in Figure 1. The study was conducted in Nasiriyah City, Thi-Qar Governorate, southern Iraq, during a five-day period at the beginning of July, corresponding to peak summer conditions[8]. Three photovoltaic panels were installed in an outdoor environment and exposed to real climatic conditions. The surface temperature of the photovoltaic panels was measured directly using thermocouples fixed on the front surface of the panels to ensure an accurate representation of the operating temperature. Temperature measurements were continuously recorded using a data logger at -15minute. In addition, solar irradiance was measured using a dedicated solar radiation meter placed in close proximity to the photovoltaic panels to capture the incident solar energy accurately. This experimental arrangement provided reliable field data for evaluating the thermal behavior of photovoltaic panels under extreme climatic conditions.

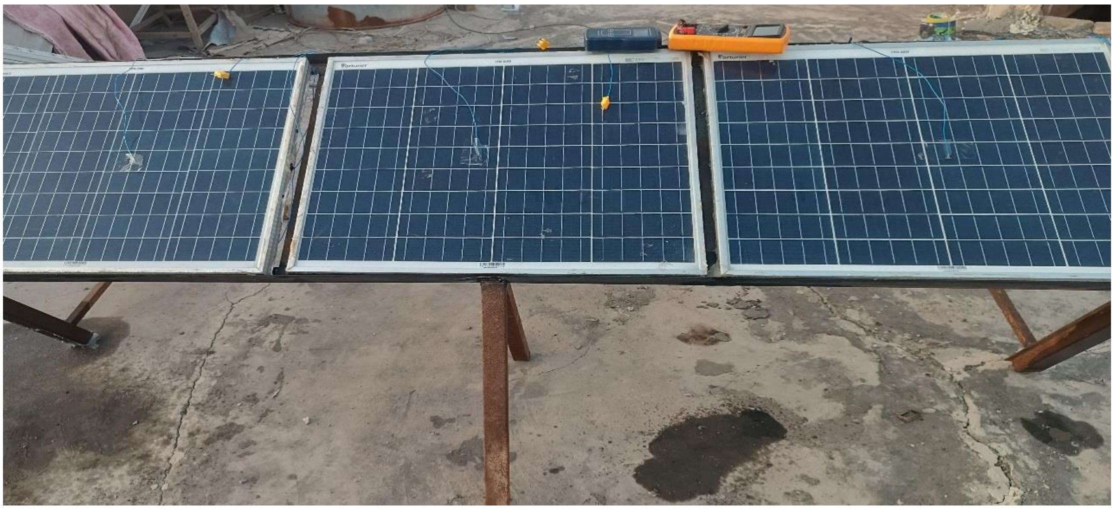


Figure 1: Outdoor experimental setup for photovoltaic panel temperature measurement[8]

Uncertainty Analysis

An uncertainty analysis was performed to assess the reliability of the measured data. The accuracy of the measurement instruments, including the thermocouples, data logger, and solar irradiance meter, was specified by the manufacturers as ±2.5%. Assuming independent measurement errors, the combined uncertainty associated with the measured variables was considered within this range. Given the relatively small uncertainty compared to the observed variations in photovoltaic surface temperature and solar irradiance during the experimental period, the impact of measurement uncertainty on the overall analysis is considered limited. Therefore, the measured data are deemed sufficiently accurate for developing and validating the proposed machine learning models.

Machine Learning Modelling Approach

Three meteorological parameters were selected as input features based on their dominant influence on photovoltaic thermal behavior: solar irradiance (W/m²), ambient temperature (°C), and wind speed (m/s). The output variable considered in this study is the photovoltaic panel surface temperature (°C). All data points were examined for missing values, and only complete records were retained to ensure data consistency. Three regression-based machine learning models were developed and compared. Linear Regression was employed as a baseline model due to its simplicity and consistency with the underlying physical principles governing photovoltaic thermal behavior. In addition, a Random Forest Regressor was implemented as an ensemble learning approach based on multiple decision trees, enabling the modeling of potential non-linear relationships and interactions among the input variables. A Gradient Boosting Regressor was also utilized as a sequential ensemble technique designed to minimize prediction error and enhance predictive accuracy. To preserve the temporal structure of the dataset and prevent data leakage, a time-based data splitting strategy was adopted, in which the first 70% of the data were used for model training and the remaining 30% for testing, reflecting realistic prediction scenarios. Model performance was evaluated using the coefficient of determination (R^2), the mean absolute error (MAE), and the root mean square error (RMSE), which together provide a comprehensive assessment of prediction accuracy and reliability. All data preprocessing, model training, prediction, and evaluation procedures were carried out using Python within a Jupyter Notebook environment. Predicted and measured photovoltaic surface temperature values were exported to Excel files to ensure transparency and reproducibility, and graphical comparisons between measured and predicted temperatures were generated for each model with the corresponding performance indicators embedded within the figures.

Mathematical Formulation

Let the input feature vector at the time step $i$be defined as[9]:

$$\mathbf{x}_i = [G_i, \; T_{a,i}, \; V_{w,i}] \qquad (1)$$

where $G_i$is the solar irradiance $(\mathrm{W/m^2})$, $T_{a,i}$is the ambient temperature (°C), and $V_{w,i}$is the wind speed (m/s). The output (target) variable is the PV surface temperature:

$$y_i = T_{pv,i} \; (^\circ\mathrm{C}) \qquad (2)$$

The machine learning models approximate a regression function $f(\cdot)$that maps meteorological inputs to PV surface temperature:

$$\hat{T}_{pv,i} = \hat{y}_i = f(\mathbf{x}_i) \qquad (3)$$

where $\hat{y}_i$denotes the predicted PV temperature.

For the Linear Regression (LR) model, the mapping can be expressed explicitly as[10]:

$$\hat{y}_i = \beta_0 + \beta_1 G_i + \beta_2 T_{a,i} + \beta_3 V_{w,i} \qquad (4)$$

where $\beta_0$is the intercept and $\beta_1, \beta_2, \beta_3$are the regression coefficients estimated from the training data using least squares minimization:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \qquad (5)$$

For Random Forest (RF) and Gradient Boosting Regressor (GBR), the prediction is obtained through an ensemble of regression trees. In general form:[11], [12]

$$\hat{y}_i = \frac{1}{M}\sum_{m=1}^{M} h_m(\mathbf{x}_i)(\text{RF}) \qquad (6)$$

$$\hat{y}_i = \sum_{m=1}^{M} \gamma_m \; h_m(\mathbf{x}_i)(\text{GBR}) \qquad (7)$$

where $h_m(\cdot)$denotes the m-th regression tree and $M$is the number of trees.
The prediction error at each sample is defined as:

$$e_i = \hat{y}_i - y_i \qquad (8)$$

Model performance was quantified using the following metrics[13][14]:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \qquad (9)$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} \mid y_i - \hat{y}_i \qquad (10)$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \qquad (11)$$

where $\bar{y}$is the mean of the measured PV surface temperature values and $N$is the number of test samples.

Results and Discussions

The developed machine learning models were evaluated using experimentally measured data collected under real outdoor conditions in Nasiriyah City during a five-day period in early July. The dataset represents peak summer operation of photovoltaic (PV) panels, characterized by high solar irradiance and elevated ambient temperatures. Figures 2–4 present the comparison between measured and predicted PV surface temperatures for the Linear Regression, Random Forest, and Gradient Boosting models, respectively.

As shown in Figure 2, the Linear Regression model exhibits an excellent agreement between the measured and predicted PV surface temperatures over the daytime operating period (06:00–18:00). The predicted values closely follow the measured temperature profile, with minimal deviation throughout the day. This behavior reflects the strong linear dependence of PV surface temperature on solar irradiance and ambient temperature under steady-state outdoor conditions. From a physical perspective, the absorbed solar radiation constitutes the primary heat input to the PV module, while convective heat transfer driven by ambient air temperature and wind speed governs heat dissipation. The high accuracy achieved by the linear model ($R^2 = 0.9999$, MAE = 0.028 °C) confirms that the dominant thermal mechanisms governing PV temperature in this experimental setup are predominantly linear.

The prediction results obtained using the Random Forest model are illustrated in Figure 3. Although a strong correlation between measured and predicted temperatures is still observed, noticeable scatter appears around the ideal agreement line, particularly during periods of high irradiance near midday. This behavior can be attributed to the tendency of ensemble tree-based models to introduce local fluctuations when modeling smooth physical relationships. In the present case, the PV thermal response is governed by continuous heat transfer processes, and the introduction of excessive model flexibility leads to moderate prediction errors (MAE = 0.85 °C). These results indicate that Random Forest models may be less effective when the underlying physical system exhibits limited nonlinearity.

The Gradient Boosting model results, presented in Figure 4, demonstrate improved performance compared to the Random Forest model, with reduced scatter and a closer alignment to the measured data. The sequential learning strategy employed by Gradient Boosting allows for a more refined approximation of the thermal behavior, resulting in a lower MAE of 0.58 °C and an R^2 value of 0.996. Nevertheless, slight deviations are still observed at higher temperature levels, suggesting that while Gradient Boosting partially captures subtle variations, it does not outperform the physics-consistent linear formulation for the investigated conditions.

A quantitative comparison of the three models is summarized in Table 1, which reports the corresponding error metrics. The results clearly indicate that the Linear Regression model outperforms the more complex machine learning approaches in terms of prediction accuracy. This outcome highlights a key physical insight: when PV surface temperature is primarily governed by linear heat balance mechanisms—namely solar heat absorption and convective heat losses simple regression models aligned with physical principles can outperform advanced nonlinear algorithms.

Overall, the experimental results confirm that model complexity should be selected based on the physical characteristics of the system rather than algorithmic sophistication alone. While ensemble-based machine learning models remain valuable for highly nonlinear or transient thermal systems, the present study demonstrates that for outdoor PV panels operating under stable hot-climate conditions, linear models provide a robust, interpretable, and computationally efficient solution for surface temperature prediction.

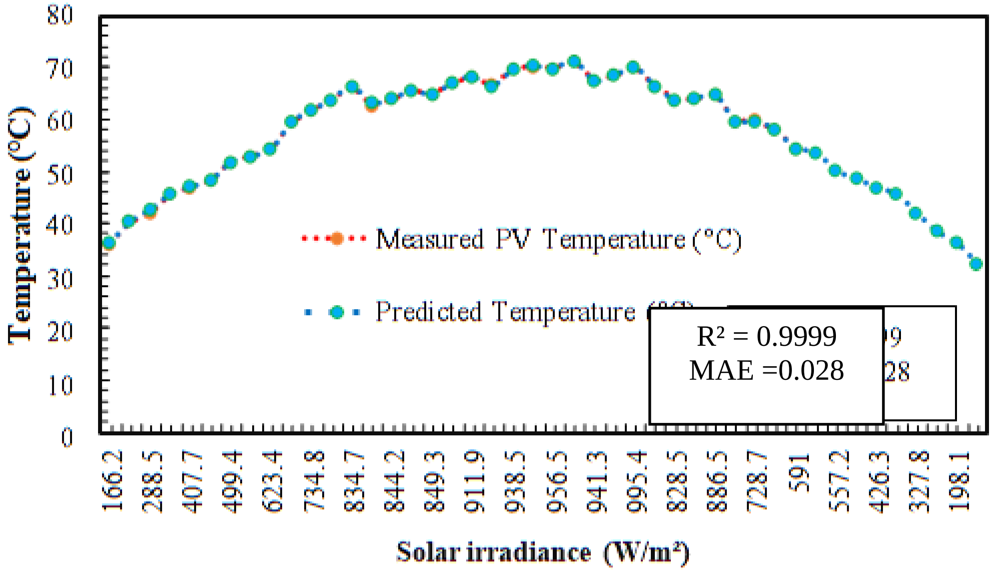


Figure 2: Comparison between Measured and Predicted Photovoltaic Surface Temperature Using Linear Regression

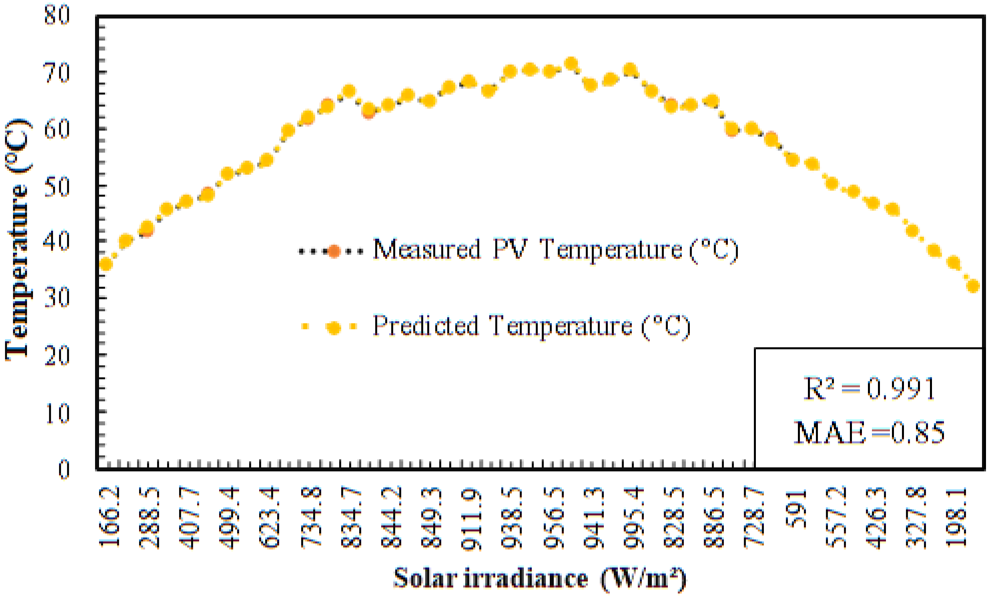


Figure 3: Measured versus Predicted Photovoltaic Surface Temperature Using Random Forest

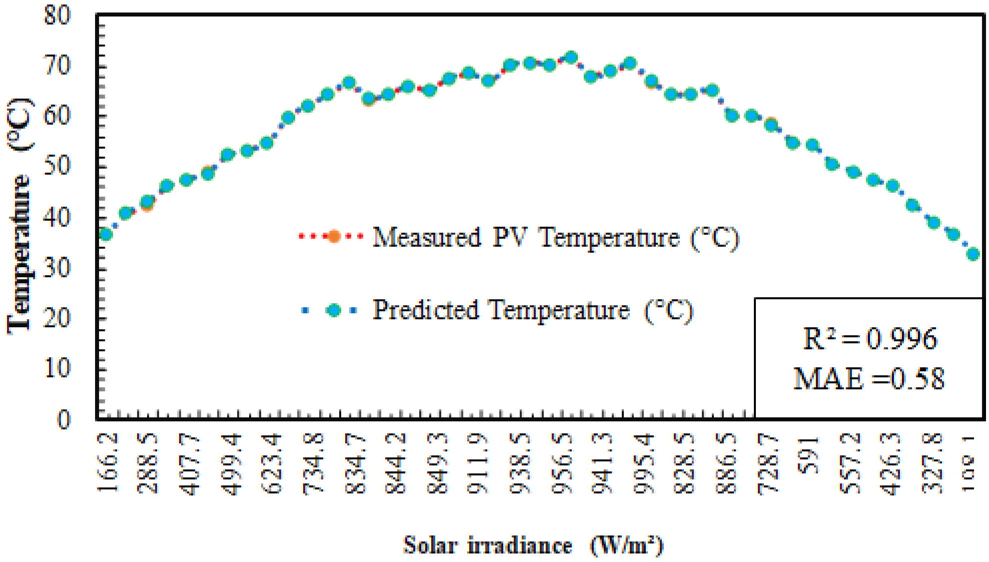


Figure 4: Prediction Performance of Gradient Boosting for Photovoltaic Surface Temperature

Table 1: Comparison of Prediction Errors for Machine Learning Models

Model	R²	MAE (°C)
Linear regression	0.999	0.028
Random forest	0.991	0.85
Gradient boosting	0.996	0.58

Conclusion

This study investigated the prediction of photovoltaic (PV) panel surface temperature using experimentally measured data collected under real outdoor conditions in Nasiriyah City during early July. Field measurements were obtained using thermocouples attached to the front surface of the PV panels, a data logger, and a solar irradiance meter. Three machine learning models—Linear Regression, Random Forest, and Gradient Boosting—were evaluated using the measured dataset.

The results showed that Linear Regression achieved the highest prediction accuracy ($R^2 = 0.9999$, MAE = 0.028 °C), indicating that the thermal behavior of PV panels under the investigated conditions is predominantly linear. Although Random Forest and Gradient Boosting models also demonstrated good performance, their higher error values suggest that increased model complexity did not improve prediction accuracy for this application.

It is recommended that simple, physics-consistent regression models be employed for PV surface temperature prediction in hot and arid climates. Future work should focus on extending the measurement period and validating the proposed approach under different seasonal and operating conditions.

References

[1] F. Obeidat, “A comprehensive review of future photovoltaic systems,” *Solar Energy*, vol. 163, pp. 545–551, Mar. 2018, doi: 10.1016/J.SOLENER.2018.01.050.

[2] M. A. Hameed, I. Kaaya, M. Al-Jbori, Q. Matti, R. Scheer, and R. Gottschalg, “Analysis and prediction of the performance and reliability of PV modules installed in harsh climates: Case study Iraq,” *Renew Energy*, vol. 228, p. 120577, Jul. 2024, doi: 10.1016/J.RENENE.2024.120577.

[3] A. K. Tripathi *et al.*, “Advancing solar PV panel power prediction: A comparative machine learning approach in fluctuating environmental conditions,” *Case Studies in Thermal Engineering*, vol. 59, p. 104459, Jul. 2024, doi: 10.1016/J.CSITE.2024.104459.

[4] N. Patel, B. E. Pieters, K. Bittkau, E. Sovetkin, K. Ding, and A. Reinders, "Assessing the accuracy of two steady-state temperature models for onboard passenger vehicle photovoltaics applications," *Progress in Photovoltaics: Research and Applications*, vol. 32, no. 11, 2024, doi: 10.1002/pip.3832.

[5] A. Keddouda *et al.*, "Photovoltaic module temperature prediction using various machine learning algorithms: Performance evaluation," *Appl Energy*, vol. 363, p. 123064, Jun. 2024, doi: 10.1016/J.APENERGY.2024.123064.

[6] N. Bailek, K. Bouchouicha, M. A. Hassan, A. Slimani, and B. Jamil, "Implicit regression-based correlations to predict the back temperature of PV modules in the arid region of south Algeria," *Renew Energy*, vol. 156, pp. 57–67, Aug. 2020, doi: 10.1016/J.RENENE.2020.04.073.

[7] D. V. Pombo, P. Bacher, C. Ziras, H. W. Bindner, S. V. Spataru, and P. E. Sørensen, "Benchmarking physics-informed machine learning-based short term PV-power forecasting tools," *Energy Reports*, vol. 8, pp. 6512–6520, Nov. 2022, doi: 10.1016/J.EGYR.2022.05.006.

[8] A. A. Ali, D. A. Lafta, S. W. Noori, F. Abdulamir, and F. L. Rashid, "Innovative materials integrated with PCM for enhancing photovoltaic panel efficiency: An experimental investigation," *J Energy Storage*, vol. 102, p. 114258, Nov. 2024, doi: 10.1016/J.EST.2024.114258.

[9] E. Skoplaki and J. A. Palyvos, "On the temperature dependence of photovoltaic module electrical performance: A review of efficiency/power correlations," *Solar Energy*, vol. 83, no. 5, pp. 614–624, May 2009, doi: 10.1016/J.SOLENER.2008.10.008.

[10] V. A. Barbur, D. C. Montgomery, and E. A. Peck, "Introduction to Linear Regression Analysis.," *The Statistician*, vol. 43, no. 2, 1994, doi: 10.2307/2348362.

[11] L. Breiman, "Random forests," *Mach Learn*, vol. 45, no. 1, 2001, doi: 10.1023/A:1010933404324.

[12] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann Stat*, vol. 29, no. 5, 2001, doi: 10.1214/aos/1013203451.

[13] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Clim Res*, vol. 30, no. 1, 2005, doi: 10.3354/cr030079.

[14] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature," *Geosci Model Dev*, vol. 7, no. 3, 2014, doi: 10.5194/gmd-7-1247-2014.