UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Modelling Temperature-Induced Performance Losses in Rooftop Solar Panels Using Machine Learning

Prastika Indriyanti*
Faculty of Data Science and Computing
Universiti Malaysia Kelantan
Kota Bharu, Malaysia
Faculty of Computer Science
Universitas Mercu Buana
Jakarta, Indonesia
Email: s24e0494f@siswa.umk.edu.my, prastika@mercubuana.ac.id

Hasyiya Karimah Adli
Faculty of Data Science and Computing
Universiti Malaysia Kelantan
Kota Bharu, Malaysia
Email: hasyiya@umk.edu.my

Nur Dinie Faqihah Binti Jasmi
Faculty of Data Science and Computing
Universiti Malaysia Kelantan
Kota Bharu, Malaysia
Email: s22b0003@siswa.umk.edu.my

Afiyati
Faculty of Computer Science
Universitas Mercu Buana
Jakarta, Indonesia
Email: afiyati.reno@mercubuana.ac.id

Submitted: 15/9/2025. Revised edition: 28/11/2025. Accepted: 5/1/2026. Published online: 28/7/2026
DOI: https://doi.org/10.11113/ijic.v16n1-2.695

***Abstract*—This study investigates the critical impact of temperature in the performance of rooftop solar photovoltaic (PV) systems. Particularly focusing on energy generation efficiency losses when temperatures exceed 28°C. High temperatures increase the internal resistance of PV cells, reducing their ability to convert sunlight into electricity, the research identifies an optimal operating temperature range for solar panels between 24°C and 26°C, where efficiency is maximized. To predict these temperature-dependent performance losses, machine learning models, are employed. A comparation between Linear regression in capturing the complex, non-linear relationships between temperature, irradiance, and energy output. While Linear Regression showed a perfect R2 score of 1.0 indicative of overfitting. The Random Forest model achieved a robust R2 of 0.8987, demonstrating superior generalizability and accuracy of real- world applications. This research validates the effectiveness of machine learning techniques for more reliable solar energy forecasting and optimization and highlights the necessity of considering thermal management in solar system design, especially in hot climates.**

***Keywords*—PV Performance, Temperature Dependences, Machine Learning, Linear Regression, Energy Loss**

I. Introduction

Solar energy has emerged as one of the most promising renewable energy sources due to its sustainability and minimal environmental impact. Photovoltaic (PV) systems, which harness solar energy through panels, are increasingly adopted across various sectors. However, the efficiency of solar panels, in converting sunlight into electricity is affected by several environmental factors, such as temperature, irradiance, and humidity (Ali *et al*. 2020).

Temperature is one of the most significant environmental factors influencing the performance of solar panels. High temperature led to thermal losses in the materials of the solar cells, reducing their ability to generate electricity. Solar panels typically operate optimally at the temperatures between 24ºC and 26ºC, but when temperatures exceed 28ºC, the efficiency begins to decrease substantially. This decline in energy performance in a major concern, particularly in tropical, desert, and subtropical climates. As a result, solar energy systems in these areas may not perform at their maximum potential, which could have long-term implications for energy generation, cost-effectiveness, and sustainability.

Recent research has explored the potential of using computational techniques, particularly machine learning models, to predict energy losses caused by temperature variations. While the negative impact of high temperatures on solar panel performance is well-documented, many studies focus primarily on general temperature fluctuations without providing detailed analysis. Furthermore, traditional models, such as those based on linear regression, are limited in their ability to fully capture the complex, non-linear relationships between temperature and energy generation. The efficiency of a PV system can reduce by 0.5% for each 1ºC more than standard over the ideal operating temperature (Santos *et al.* 2022).

II. Literature Reviews

The operational efficiency of photovoltaic (PV) systems is fundamentally limited by thermal dynamics, making high ambient temperature the most influential environmental factor driving performance loss. The core physical mechanism underpinning this degradation involves the increase of internal electrical resistance within the PV cells as their operating temperature rises. This increased internal resistance directly causes a measurable reduction in the open-circuit voltage and the maximum power point voltage.

Consequently, the ability of the semiconductor material to efficiently convert incident solar irradiance into usable electrical energy is compromised, resulting in a quantifiable reduction in overall energy output and long-term thermal stress on the module materials. This established thermodynamic and electrical principle forms the critical necessity for detailed performance modeling in temperature-variable environments (Onaifo *et al*. 2020; Segbefia 2023).

To ensure consistent performance measurement across the industry, PV modules are rated under Standard Test Conditions (STC), which typically define the cell temperature at 25°C and the solar irradiance at 1000 W/m2 Empirical studies and industry consensus have further narrowed the optimal operational window for solar panels, suggesting maximized efficiency when module temperatures are maintained specifically between 24°C and 26°C (Skoplaki and Palyvos 2009; Walker and Desai 2021)

Quantification of efficiency loss in the literature traditionally relies on a negative temperature coefficient. Foundational studies, including the comprehensive review by Skoplaki and Palyvos (2009), consistently confirmed a consensus rate: PV system efficiency generally drops by approximately 0.4% to 0.5% for every increase in temperature above the 25°C benchmark. For example, the average temperature coefficient of efficiency has been found to be around - 0.5%/°C . This linear degradation rate provides a baseline for predicting mild performance losses in moderate climates (Noor 2024; Skoplaki and Palyvos 2009)

The use of machine learning in renewable energy systems has traditionally been towards predicting solar irradiances and managing energy production (Munawar and Wang 2020). Decision trees, SVMs and artificial neural networks achieve high prediction accuracy of solar power under dynamic conditions (Widodo *et al*. 2021; Elsaraiti and Merabet 2022). Furthermore, Alvarez *et al*. in 2020 applied ML models for the evaluation of power systems performance with non-linear and dynamic variable interactions. Random Forest models have demonstrated high accuracy in several works, approximately 94.01% for solar energy forecasting (Anuradha *et al*. 2021; Lee *et al*. 2021; Subramanian *et al*. 2023; Shah *et al*. 2024). Nevertheless, a gap exists in the development of comprehensive models that account for external environment and system-level thermal management.

Several research gaps have been identified in literature from ScienceDirect (https://www.sciencedirect.com) platform shows a search query for research articles related to "solar energy forecasting using machine learning heat dissipation," returning 420 results. While this indicates a growing interest in the field, the relatively low number of studies on the specific topic of heat dissipation in solar energy forecasting using machine learning highlights a critical research gap. Although many studies investigate temperature impacts on the PV performance, it remains scarce literature that can predict the impacts of heat dissipation.

Hence, this study seeks to bridge this gap by investigating the relationship between temperature and the energy performance of rooftop solar systems, focusing on high temperature conditions above 26°C by comparing the performance of machine learning models. The findings will enhance the understanding of temperature-driven performance losses and help optimize solar systems in high temperature regions for better predictive models for solar energy strategies.

III. Methodology

The machine learning (ml) forecasting process utilized in this study is illustrated in Fig. 1. This project focuses on analyzing data and exploring the application of two models of machine learning techniques to predict solar generation. The implementation of the study follows the framework outlined in Fig. 1.

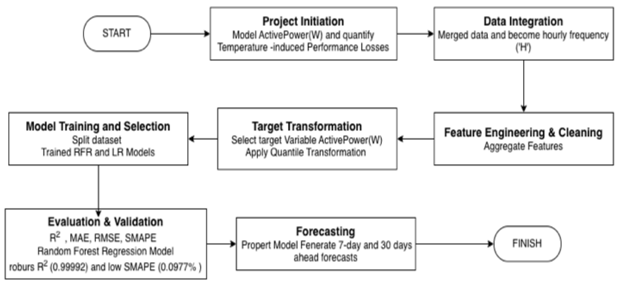

Fig. 1. The Overall work flow of the study

A. Data Collection and Tools

The dataset used in this study originates from the Hong Kong University of Science and Technology (HKUST), which collected comprehensive time-series data on rooftop solar panels over a three-year period (2021–2023) (Lin *et al*., 2025). It includes key environmental variables—temperature, solar irradiance, and humidity—alongside the dependent variable, energy output (kWh). Covering regions with diverse climatic conditions, the dataset enables analysis of solar performance across varying temperature ranges, with particular attention to the critical threshold above 28 °C. The inverter data comprises 137,520 records of solar energy output, while the environmental datasets contain larger volumes: 1,568,712 temperature records, 1,567,665 humidity records, and 1,567,792 irradiance records. Specifically, temperature, humidity, and irradiance data were collected annually with approximately 520,000 entries per variable, representing detailed measurements in degrees Celsius, percentage relative humidity, and watts per square meter (W/m²), respectively.

B. Units

Data preprocessing was carried out to ensure the dataset was clean, consistent, and properly structured for machine learning analysis. The raw data, comprising approximately 1.57 million high-frequency records, were resampled to an hourly ('H') resolution using mean aggregation. A significant proportion of missing values was identified, primarily within the inverter electrical metrics, amounting to 15,264 hourly records (around 58% of the total time series). These gaps were imputed using time-based interpolation (method='time'), which estimates values based on the temporal proximity of adjacent observations, an approach particularly effective for continuous physical variables such as voltage and temperature. To ensure completeness, forward-fill (ffill) and backward-fill (bfill) operations were subsequently applied, resulting in a dataset with no null values.

Outlier treatment was performed in two stages. First, physical clipping was applied to mitigate sensor-induced anomalies, constraining irradiance values between 0 and 1200 W/m² and temperature values between –5 °C and 60 °C. Second, statistical detection using Z-score analysis (|Z| > 3) identified 645 soft outliers within the *totalActivePower* (W) variable. These outliers were flagged but retained to preserve the model's sensitivity to legitimate high-variability operational patterns.

a) To optimize the dataset for machine learning analysis, two key preprocessing steps were implemented. First, phase metric aggregation was performed to address redundancy within the three-phase AC measurements (L1, L2, L3). Twelve related columns were consolidated into three representativescalar features: acCurrent_mean, acVoltage_mean, and activePower_sum, the latter representing the total active power across all phases. This aggregation effectively reduced multicollinearity and feature dimensionality while preserving the essential electrical characteristics of the system. Second, temporal feature encoding was applied by extracting time-based attributes: hour, dayofweek, and month from the hourly index.

The target variable, `totalActivePower` (W), exhibited extreme right-skewness and substantial zero-inflation due to nighttime zero production. To address these issues, a **Quantile Transformation (QT)** was applied (Fig. 2). This non-parametric transformation maps the distribution into a near-Gaussian form, enhancing statistical stability and improving model performance under non-linear conditions. Following the Quantile Transformation, all features including the transformed targetewere scaled using **Min–Max normalization** to a fixed range (typically [0, 1]). This ensured that all variables contributed proportionally to the learning process without bias toward features with larger numeric magnitudes (Müller & Guido, 2017).

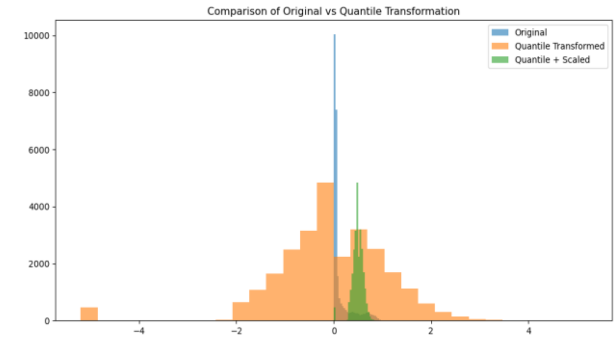

Fig. 2. Comparison of the original totalActivePower distribution and the data after

C. Data Split and Model Selection

The final pre-processed dataset was divided into training and testing subsets to evaluate the model's ability to generalize to unseen data. 70% of the data was allocated for training and the remaining 30% for testing. The training set was used to fit the models, while the testing set served to assess predictive performance and robustness on unseen samples. The selection of machine learning models was guided by their suitability for predicting solar energy output and capturing complex, temperature-dependent performance dynamics. Two primary models were implemented:

Linear Regression (Baseline Model):
Linear regression was used as a benchmark to establish a reference level of predictive performance. Its interpretability provides a clear first-order understanding of the relationships between features and the target variable. The model estimates parameters by minimizing the residual sum of squares (RSS) using Ordinary Least Squares (OLS) (Hastie, Tibshirani & Friedman, 2017). The linear model can be expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon \tag{1}$$

where β_0 is the intercept, β_0 are the coefficients quantifying the magnitude and and direction of the impact of each feature (X_i), and is ϵ the error term algorithms (Hastie, Tibshirani, and Jerome Friedman 2017; Muller and Guido 2017).

Random Forest (Ensemble Model):
Given the non-linear and complex relationships among environmental parameters affecting solar performance (e.g., the steep decline in efficiency above 28 °C), a Random Forest model was employed. As an ensemble method, it constructs multiple decision trees using bootstrapped samples and random feature subsets at each split, thereby capturing intricate dependencies and interactions between variables (Müller & Guido, 2017). The final prediction is the average of all tree outputs, which significantly reduces variance and overfitting,

resulting in more robust and generalizable predictions compared to single decision trees or linear models.

Prophet Time-Series Model:
For multi-horizon forecasting, the Prophet algorithm (developed by Meta) was utilized to exploit strong daily and annual seasonality in the solar dataset, enabling accurate 7-day and 30-day production forecasts.
Finally, hyperparameter tuning was conducted using a combination of Grid Search and Random Search methods to systematically identify the optimal model configurations that maximized predictive performance.

D. Model Evaluationn

Model evaluation is a critical step following model training, conducted to ensure that the developed machine learning models generalize effectively to unseen data and are not overfitted to the training set (Hastie *et al*., 2017; Müller & Guido, 2017). The performance of each model was assessed using a separate testing dataset and evaluated through standard regression metrics, including **Mean Absolute Error** (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2).

Let $\hat{Y}_i$ denote the predicted value, Y_i the actual value, $\bar{Y}$ the mean of the actual values, and n the number of observations. The metrics are defined as follows:

Mean Absolute Error (MAE):
MAE measures the average magnitude of prediction errors. A lower MAE indicated better performance. It is easy to interpret and is robust to outliners (Brownlee 2017). A lower MAE indicates better predictive performance.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | Y_i - Y_i| \tag{1}$$

Root Mean Squared Error (RMSE):
RMSE calculates the square root of the average squared prediction errors, penalizing larger errors more heavily than MAE. It is particularly useful when large deviations are critical (Chai & Draxler, 2014).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i' - Y_i)^2} \tag{2}$$

Mean Squared Error (MSE):
MSE represents the average squared differences between predictions and actual values. While sensitive to large errors, it is less interpretable than RMSE.

$$MSE = \frac{1}{n} \sum_{i-1}^{n}(Y_i - Y_i)^2 \tag{3}$$

Mean Absolute Percentage Error (MAPE):
MAPE expresses prediction error as a percentage of the actual value, providing a scale-independent measure. However, it can be unstable when actual values are near zero.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} | \frac{Y_i - Y_i}{Y_i} | \, x\ 100 \tag{4}$$

R-squared (R^2):
R^2 quantifies the proportion of variance in the dependent variable that is explained by the model. A value close to 1 indicates strong explanatory power, though it should be interpreted alongside error-based metrics (Bergstra & Bengio, 2012).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - Y_i)^2}{\sum_{i=1}^{n}(Y_i - \underline{Y})^2} \tag{5}$$

By employing this combination of metrics, the study comprehensively evaluates the models' accuracy, robustness, and generalization ability. The ultimate objective is to determine which algorithm, either Linear Regression or Random Forest is the most effectively predicts solar energy performance based on temperature variations.

E. Energy Loss Calculation

To quantify the specific thermal impact on system efficiency, the metric 'Energy Loss per Degree Celsius (kWh)' was employed. Energy loss is defined as the discrepancy between the theoretical maximum energy that a system could produce under optimal conditions and its actual observed output. The energy loss is measure as percentage:

$$Energy\ Loss(\%) = \frac{(actual\ Energy - Ideal\ Energy}{Ideal\ Energy} * 100 \tag{6}$$

where actual energy refers to the energy generated under teal environmental conditions (temperature, irradiance, humidity), and ideal energy is the theoretical amount the panels should generate under optimal conditions. The calculation helps quantify the efficiency losses of solar panels due to environmental factors such as temperature (Walker and Desai 2021). It allows for a clear understanding of how much energy is being lost as temperature increases and deviates from ideal conditions. Besides, it is possible to track how well the solar panels are performing over time, especially during hot weather when temperature-induced losses are more likely to occur.

IV. Resuls and Discussion

Fig. 3 illustrates the average daily solar pattern across different hours of the day, showing how key environmental variables and solar power output vary on average. The plot shows the diurnal behavior of key environmental and PV system variables. Irradiance (W/m², blue line with circles) represents the average solar radiation received at the panel surface, peaking around midday (10–14h) when the sun is highest and dropping to near zero at night (0–5h and 18–23h), as expected. Temperature (°C, orange line with circles) indicates the average ambient temperature, rising gradually from morning, reaching its peak in the early afternoon, and decreasing in the evening, without dropping to zero at night, reflecting realistic ambient conditions. Relative humidity (RH %, green line with x markers) remains relatively steady throughout the day with minor fluctuations, typically higher in the early morning and late evening and lower around midday due to solar heating.

Meanwhile, solar power (W, red line with squares) represents the actual PV energy output, following a trend like irradiance but scaled to the panel's capacity. It peaks sharply around 11–13h and drops to near zero at night, with a slightly asymmetric curve likely caused by temperature effects and slight efficiency reductions at higher temperatures. Overall, the plot highlights that solar power closely follows irradiance, temperature peaks slightly after irradiance and can influence PV efficiency, and nighttime zero output is clearly visible. These diurnal trends make the plot valuable for feature engineering in machine learning models (including hour, irradiance, temperature, and relative humidity), understanding daily energy production patterns, and planning energy storage or grid integration strategies. The average active power produced by the system throughout the day. It mirrors the patterns seen in the irradiance and temperature plots, peaking at around 12:00 PM. As irradiance increases, more power is generated, leading to a peak in energy production at midday. After this peak, power generation decreases as the sun starts to set, reducing the available solar energy.

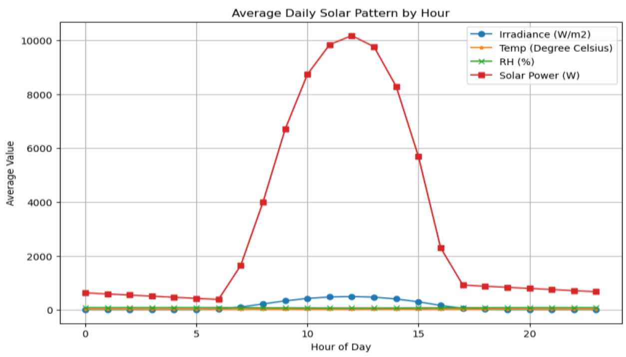

Fig. 3. Average Daily trends of irradiance, temperature, relative humidity and solar power by hour.

Fig. 4 represents the temperature-dependent PV (photovoltaic) energy loss, showing how energy loss in a solar energy system changes with temperature. The data illustrates the percentage energy loss per degree of temperature, spanning from 25.5°C to 29.0°C. Initially, at lower temperatures (around 25.5°C), there is only a slight increase in energy loss, remaining close to zero, indicating that the system operates efficiently with little to no energy loss. However, as the temperature rises to 28.0°C, a sharp spike in energy loss is observed, exceeding 1,000%. This significant increase suggests that the efficiency of the system drops drastically as temperature rises, which is a typical occurrence in solar energy systems. Higher temperatures induce thermal losses, reducing the system's overall performance and energy conversion efficiency (Onaifo *et al*. 2021; Skoplaki and Palyvos 2009).

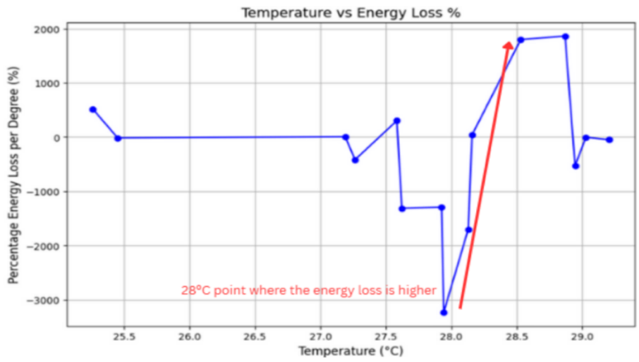

Fig. 4. Temperature vs Energy Loss %.

Between 28.0°C and 29.0°C, the plot shows fluctuations in energy loss, with some data points dipping into negative values (down to -3,000%), while others show increases. These irregularities could stem from various factors like system inefficiencies, or environmental conditions such as sudden cloud cover, shading, or other atmospheric phenomena that influence the performance of photovoltaic cells in an unpredictable manner (Onaifo *et al*. 2021). The overall trend of the plot aligns with the well-established observation that higher temperatures lead to increased energy loss. As the temperature approaches 29.0°C, the energy loss becomes more severe, likely due to overheating, which diminishes the efficiency of the photovoltaic cells (Skoplaki & Palyvos, 2009). This illustrates the importance of temperature monitoring in solar energy systems, as excessive heat can significantly compromise the performance and energy yield of solar panels (Walker and Desai 2021b).

The fluctuations observed in the temperature range between 28.0°C and 29.0°C might indicate issues related to the system's temperature regulation or inherent data variability. In real-world applications, solar panels are influenced by environmental factors such as wind, varying irradiance levels, or local microclimates, all of which can lead to inconsistent energy loss measurements (Onaifo *et al*. 2021). In conclusion, the plot highlights the inverse relationship between temperature and energy efficiency in photovoltaic systems. As temperatures rise, energy loss escalates sharply, particularly above 28.0°C, consistent with the known thermal degradation mechanisms of solar cells (Skoplaki & Palyvos, 2009).

The monthly plot shows the energy loss per degree Celsius (Fig. 5) across the 12 months of the year. It reflects significant seasonal trends, with high values of energy loss in April, May, June, and July. For example, April shows the highest recorded energy loss of 2.62 kWh per degree Celsius, which could correlate with warmer temperatures in the spring and summer months. The negative values observed in some months, such as April and December, indicate that energy loss was lower during colder periods. This suggests that in the colder months, temperature changes did not lead to significant energy loss, possibly because lower temperatures result in less thermal stress on the system.

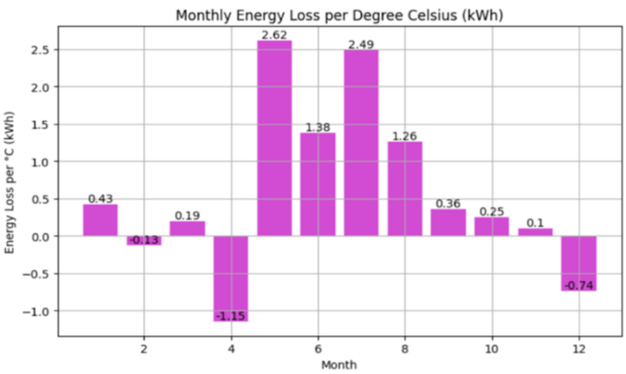

Fig. 5. Monthly Energy Loss Per Degree Celsius (kWh)

A correlation heatmap is a powerful tool used to visually represent the relationships between different variables within a dataset. It allows us to easily identify and interpret the strength and direction of these relationships. The correlation coefficient, which ranges from -1.00 to +1.00, indicates how closely related two variables are. Fig. 6 shows the power (kWh) exhibits a strong positive correlation with energy_kWh (1.00), meaning that the power generated and the energy produced are highly correlated.

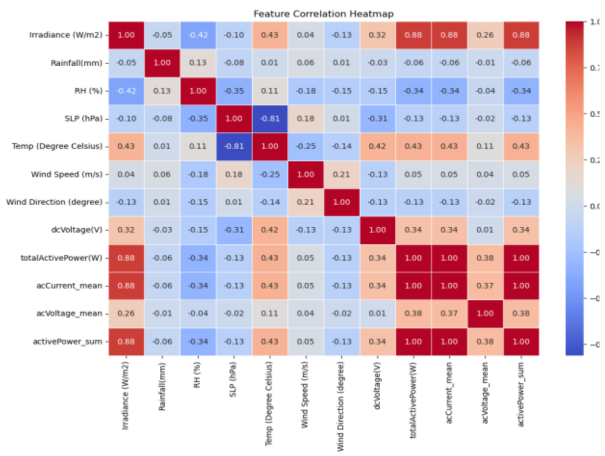

Fig. 6. Heatmap showing the relationship between features.

To evaluate the predictive capability of different machine learning approaches for photovoltaic (PV) power generation, two models were developed and tested: a Random Forest Regressor and a Linear Regression model. The models were trained on preprocessed data incorporating feature engineering and transformations designed to capture the non-linear relationships between environmental variables (such as irradiance, temperature, and humidity) and PV output. Model performance was assessed using standard regression metrics to determine accuracy, robustness, and reliability over a full 24-hour cycle. Model performance was evaluated using the metrics summarized in Table 1.

The initial Linear Regression model produced unrealistically perfect results, with an R^2 of 1.0 and near-zero error metrics (MAE and MSE), which was later traced to data leakage. One predictor, activePower_sum, was directly correlated with the target variable totalActivePower (W), invalidating the model as a genuine predictor. In contrast, the Random Forest Regressor demonstrated robust and reliable performance, achieving a very high R^2 of 0.999998. This outcome underscores the effectiveness of the preprocessing, feature engineering, and the application of quantile transformation in capturing the complex non-linear behavior of photovoltaic (PV) power generation.

TABLE I. SUMMARY OF MODEL PERFORMANCE MATRICS

Matric	**Random Forest Model**	**Linear Regression Model**
More table copy[a]		
MAPE	0.99	1.01e-15
MAE	1.22	1.67e-12
MSE	40.47	2.19e-12
R^2	0.99	1.00
Training time (s)	43.0764	0.011

Additionally, conventional MAPE proved unreliable due to extreme percentage errors during nighttime zero production. To address this, the symmetric mean absolute percentage error (SMAPE) was used, yielding a low value of 0.0977%, confirming that the Random Forest model provides accurate and consistent predictions across the full 24-hour cycle.

The dataset in this study was normalized using a Quantile Transformation to improve data distribution and model stability; however, the research deliberately focused on basic machine learning models—specifically Random Forest and Linear Regression. This decision was made to establish a strong and interpretable baseline performance. Linear Regression provides a simple, transparent understanding of the linear relationships between features and target variables, while Random Forest introduces non-linearity and captures complex feature interactions, serving as a robust comparative model. Using these fundamental algorithms also allowed for a clearer evaluation of how the Quantile Transformation affected predictive performance without the added complexity of advanced models such as neural networks or gradient boosting. Moreover, the performance metrics demonstrated excellent results ($R^2 \approx 1.0$ and minimal error values), indicating that more sophisticated models would likely yield only marginal improvements while increasing computational cost and reducing interpretability. In terms of efficiency, Linear Regression trained in approximately 0.01 seconds and Random Forest in about 43 seconds, highlighting their suitability for iterative testing and use in resource-constrained environments.

The simplicity and reproducibility of these models further strengthen their practical value. Given that both models achieved near-perfect accuracy after normalization, reflected in R^2 values close to 1.0 and extremely low MAPE, MAE, and MSE scoresthe use of basic machine learning methods was considered both efficient and empirically sufficient for this research.

Fig. 7 illustrates the Prophet model's forecasting performance for 7-day (left) and 30-day (right) horizons. The model effectively captures the strong daily and annual seasonality inherent in solar time series data, as evident from the recurring fluctuations across both forecast horizons. The blue shaded regions represent the prediction intervals (uncertainty bounds), while the black points correspond to the observed values. For both horizons, the Prophet model generates stable and consistent forecasts, maintaining a clear seasonal structure without abrupt deviations. However, it can be observed that uncertainty increases with the forecast horizon and the prediction intervals widen noticeably in the 30-day forecast compared to the 7-day forecast. This reflects the model's decreasing confidence over longer-term predictions, a common characteristic of time series models when extended beyond short-term horizons.

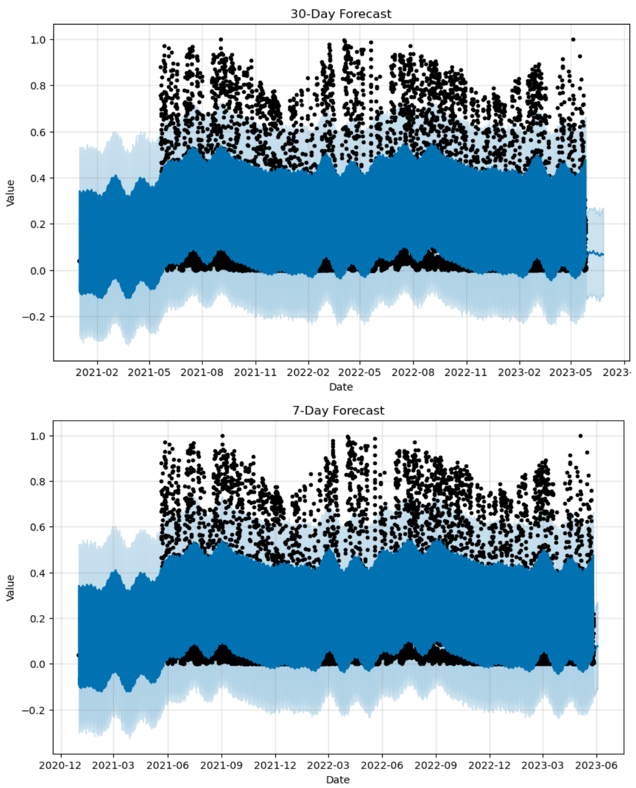

Fig. 7. 7- days ahead prediction (up) and 30- days ahead prediction (down)

Despite its robustness in identifying and projecting seasonality, the figure also highlights the Prophet model's limitations for production-level solar forecasting. The forecasts, while smooth and stable, may not fully capture short-term variability driven by rapidly changing meteorological conditions (e.g., cloud cover, temperature, irradiance). As the discussion notes, integrating reliable meteorological forecasts such as future irradiance and temperature as external regressors could enhance prediction accuracy and better represent real-world fluctuations.

In contrast, Random Forest models, as mentioned, offer complementary strengths. Their ensemble nature aggregating outputs from multiple decision trees allows them to model non-linear interactions and complex dependencies between variables more effectively than Prophet, which primarily models additive seasonality and trends. Moreover, the Random Forest's inherent robustness against overfitting makes it well-suited for handling the stochastic and multivariate nature of solar forecasting, especially when diverse meteorological predictors are available.

V. Conclusion and Future Work

The findings of this study confirm the critical influence of temperature on solar energy system performance, specifically identifying an optimal operating range between 24ºC and 26ºC. Furthermore, the data showed that temperatures exceeding 28ºC lead to a marked reduction in energy generation efficiency due to increased internal resistance and thermal stress within the PV cells. This highlights a significant challenge for solar deployment in consistently hot climates and underscores the necessity for system designs that incorporate effective thermal mitigation strategies.

Based on the evaluation metrics, both models achieved strong predictive performance, with the Linear Regression model showing near-perfect statistical scores (MAPE, MAE, MSE, and $R^2 \approx 1$). However, these exceptionally small error values suggest overfitting or numerical instability, likely due to the model's sensitivity to the specific training data rather than genuine predictive power. In contrast, the Random Forest model achieved realistic and robust results (MAPE = 0.99, MAE = 1.22, MSE = 40.47, R^2 = 0.99), indicating a strong fit with high generalization capability. Although the Random Forest required a longer training time (43.08 s vs. 0.01 s), its ensemble structure effectively captured non-linear relationships and variable interactions, providing more reliable and interpretable forecasts for complex datasets such as solar time series.

Future work will aim to enhance the Random Forest model's generalizability and predictive stability through further optimization and the application of regularization techniques, particularly by exploring the integration of LASSO-based methods. In addition to these internal improvements, an important direction will be the validation of the model's robustness and transferability across diverse datasets, ensuring its applicability to a broader range of solar forecasting contexts.

Acknowledgment

The The authors would like to thank the Faculty of Data Science and Computing, Universiti Malaysia Kelantan, and Faculty of Computer Science, Universitas Mercu Buana for providing the data with sufficient technical support.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

REFERENCES

[1] Ali, K. J., Mohammad, A. H., & Hasan, G. T. (2020). An empirical correlation of ambient temperature impact on PV module considering natural convection. *Indonesian Journal of Electrical Engineering and Computer Science, 19*(2), 627–634. https://doi.org/10.11591/ijeecs.v19.i2.pp627-634

[2] Alvarez, L. F. J., Gonzalez, S. R., Lopez, A. D., Delgado, D. A. H., Espinosa, R., & Gutierrez, S. (2020). Renewable energy prediction through machine learning algorithms. In *2020 IEEE Andescon (Andescon 2020)*. https://doi.org/10.1109/ANDESCON50619.2020.9272029

[3] Anuradha, K., Erlapally, D., Karuna, G., Srilakshmi, V., & Adilakshmi, K. (2021). Analysis of solar power generation forecasting using machine learning techniques. *E3S Web of Conferences, 309*. https://doi.org/10.1051/e3sconf/202130901163

[4] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*, 281–305.

[5] Besseau, R., Tannous, S., Douziech, M., Jolivet, R., Prieur-Vernat, A., Clavreul, J., Payeur, M., *et al*. (2023). An open-source parameterized life cycle model to assess the environmental performance of silicon-based photovoltaic systems. *Progress in Photovoltaics: Research and Applications, 31*(9), 908–920. https://doi.org/10.1002/pip.3695

[6] Brownlee, J. (2017). *Machine learning mastery with Python mini-course*.

[7] Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. *Geoscientific Model Development, 7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

[8] Elsaraiti, M., & Merabet, A. (2022). Solar power forecasting using deep learning techniques. *IEEE Access, 10*, 31692–31698. https://doi.org/10.1109/ACCESS.2022.3160484

[9] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

[10] Lee, C. H., Yang, H. C., & Ye, G. B. (2021). Predicting the performance of solar power generation using deep learning methods. *Applied Sciences, 11*(15). https://doi.org/10.3390/app11156887

[11] Lin, Z., Zhou, Q., Wang, Z., Wang, C., Bookhart, D. B., & Leung-Shea, M. (2025). A high-resolution three-year dataset supporting rooftop photovoltaics (PV) generation analytics. *Scientific Data, 12*(1), 1–13. https://doi.org/10.1038/s41597-025-04397-y

[12] Müller, A. C., & Guido, S. (2017). *Introduction to machine learning with Python*. O'Reilly Media.

[13] Munawar, U., & Wang, Z. (2020). A framework of using machine learning approaches for short-term solar power forecasting. *Journal of Electrical Engineering and Technology, 15*(2), 561–569. https://doi.org/10.1007/s42835-020-00346-4

[14] Onaifo, F., Okandeji, A. A., Ajetunmobi, O., & Balogun, D. (2021). Effect of temperature, humidity and irradiance on solar power generation. *Journal of Engineering Studies and Research, 26*(4), 113–119. https://jesr.ub.ro/journal/article/view/243/232

[15] Santos, L. D. O., Carvalho, P. C. M., & Carvalho Filho, C. D. O. (2022). Photovoltaic cell operating temperature models: A review of correlations and parameters. *IEEE Journal of Photovoltaics, 12*(1), 179–190. https://doi.org/10.1109/JPHOTOV.2021.3113156

[16] Shah, A., Viswanath, V., Gandhi, K., & Patil, N. (2024). Predicting solar energy generation with machine learning based on AQI and weather features. *CEUR Workshop Proceedings, 3940*, 1–16.

[17] Skoplaki, E., & Palyvos, J. A. (2009). On the temperature dependence of photovoltaic module electrical performance: A review of efficiency/power correlations. *Solar Energy, 83*(5), 614–624. https://doi.org/10.1016/j.solener.2008.10.008

[18] Subramanian, E., Karthik, M. M., Krishna, G. P., Prasath, D. V., & Kumar, V. S. (2023). Solar power prediction using machine learning. *arXiv preprint*. http://arxiv.org/abs/2303.07875

[19] Walker, A., & Desai, J. (2021). Understanding solar photovoltaic system performance. *U.S. Department of Energy*.

[20] Widodo, D. A., Iksan, N., Udayanti, E. D., & Djuniadi. (2021). Renewable energy power generation forecasting using deep learning method. *IOP Conference Series: Earth and Environmental Science, 700*(1). https://doi.org/10.1088/1755-1315/700/1/012026