Journal of Computation Physics and Earth Science
ISSN: 2776-2521 (online)
Volume 5, Number 2, February 2026, Page 232 - 243
https://journal.physan.org/index.php/jocpes/index

Solar Radiation Computation from Satellite Weather Data in Batam Using Linear Regression, Random Forest, and Decision Tree

Prayoga Pandapotan Simanjuntak[1], Marzuki Sinambela[2]
[1,2]Undergraduate Program Applied Instrumentation Meteorology Climatology and Geophysics (STMKG)

Article Info

Article history:

Received December 31, 2025
Revised February 28, 2026
Accepted February 28, 2026

Keywords:

Solar Radiation
Weather Forecast
Regression Model
Machine Learning
Linear Regression
Random Forest
Decision Tree
Satellite Data
Renewable Energy
Batam City

A B S T R A C T

This study addresses the necessity of evaluating solar radiation as a renewable energy source in tropical regions, specifically focusing on the challenges of estimation in Batam. The objective is to model daily solar radiation levels using satellite-derived weather data to overcome the lack of surface observation stations. Daily meteorological variables, including air temperature, relative humidity, rainfall, surface pressure, and wind speed, were sourced from the NASA POWER platform for the period January 1, 2020, to July 2, 2025. To ensure robust model generalization and prevent data leakage, the dataset was partitioned chronologically, utilizing data from 2020–2024 for training and the year 2025 for independent testing. Three computational models Linear Regression (LR), Random Forest (RF), and Decision Tree (DT) were applied to the processed data. The evaluation results indicate that the Random Forest model achieved the highest relative performance among the tested algorithms, recording a Mean Squared Error (MSE) of 19.61, a Mean Absolute Error (MAE) of 3.42, and a coefficient of determination R^2 of 0.20. In comparison, the Linear Regression model produced an R^2 of 0.19, while the Decision Tree showed significantly lower predictive accuracy. Despite being the most viable model, an R^2 of 0.20 reveals that the current predictors explain only 20% of the variance in solar radiation, highlighting the inherent complexity of tropical atmospheric dynamics. These findings suggest that while machine learning offers a promising framework for energy planning in Batam, further research incorporating additional explanatory features, such as cloud cover or aerosol indices, is required to improve model reliability.

Corresponden Author:

Prayoga Pandapotan Simanjuntak
Undergraduate Program Appllied Instrumentation Meteorology Climatology and Geophysics (SMKG)
Tangerang, Indonesian
Email: prayogapandapotan@gmail.com

1. INTRODUCTION

Solar radiation serves as the primary source of energy for the Earth and plays an essential role in many environmental and climatic processes [1]. Even slight variations in the amount of solar energy reaching the Earth's surface can lead to significant climatic impacts. Therefore, obtaining accurate estimates of solar radiation is crucial for designing, planning, and evaluating systems used to harness solar energy resources [2]. However, in many regions, particularly in developing countries, ground-based solar radiation measurement stations are still limited in number and spatial coverage [3]. This limitation has encouraged the development of alternative estimation approaches, including physical empirical models and data-driven methods that utilize other meteorological parameters to estimate solar radiation [3].

Indonesia receives relatively high solar radiation, with an average potential of approximately 4.8 kWh/m²/day, making it a promising region for the development of renewable solar energy [4]. Despite this potential, the utilization of solar energy in Indonesia remains relatively low. In areas with limited observational infrastructure, researchers often rely on secondary datasets such as NASA POWER. This satellite-based dataset

provides long-term, consistent meteorological and radiation data and has become an important data source for estimating solar energy potential in locations lacking direct radiation observations [3].

Various computational approaches have been developed to predict solar radiation. In recent years, machine learning techniques have gained increasing attention due to their capability to capture complex and non-linear relationships among meteorological variables [5]. Numerous algorithms have been applied in solar radiation studies, including Random Forest, Neural Networks, Support Vector Regression, Decision Trees, Bagging methods, as well as deep learning architectures such as CNN, RNN, and LSTM, often combined with boosting techniques [5]. For instance, Fan et al. reported that machine learning models can achieve better predictive accuracy than traditional empirical approaches for daily solar radiation estimation [2]. Similarly, studies conducted in Türkiye utilized NASA POWER data to train and compare several deep learning and boosting models such as LSTM, GRU, XGBoost, and Random Forest for solar radiation prediction. However, many of these studies focus on specific geographic locations and do not fully consider the performance of prediction models in complex tropical maritime environments [6].

Although machine learning models such as Linear Regression, Decision Tree, and Random Forest have been widely applied in solar radiation studies, their performance in tropical coastal regions such as Batam remains relatively underexplored. Batam is characterized by high humidity, frequent cloud formation, and strong seasonal variability, which may significantly influence solar radiation patterns. Understanding the capability of machine learning models in such environments is important for supporting renewable energy planning in regions with limited observational data.

Therefore, this study aims to evaluate the performance of several classical machine learning models Linear Regression, Decision Tree, and Random Forest in approximating satellite derived solar radiation in Batam using meteorological variables obtained from the NASA POWER dataset. Batam, located in the tropical region of Indonesia, is estimated to have high solar radiation potential, yet direct radiation measurements in this area remain limited. In this study, NASA POWER meteorological data are processed using Python to develop predictive models, and the performance of each method is evaluated using standard regression metrics. The results of this research are expected to provide insights into the applicability and limitations of machine learning approaches for solar radiation prediction in tropical coastal environments, while also contributing useful information for solar energy development and climate mitigation planning in the region.

2. RESEARCH METHOD

This research was conducted through six main steps that were structured to make predictions. The first step involved collecting and cleaning data consisting of air temperature, relative humidity, rainfall, surface pressure, wind speed, and time data. The data was examined to ensure that there were no missing values, duplicates, or outliers that could affect the model results. During the data pre-processing phase, we transitioned from a random split to a chronological data splitting strategy to respect the temporal and seasonal characteristics of the meteorological variables [8]. Specifically, data from January 2020 through December 2024 was used for model training, while the period from January 2025 to July 2025 was reserved for testing. This approach prevents data leakage and ensures the model is evaluated on its ability to generalize to future conditions. In the subsequent data exploration phase, visualizations such as scatter plots and correlation matrices were employed to analyze the relationship patterns between features and the target variable. The analysis confirmed that temperature has a positive relationship with solar radiation, while humidity and rainfall exhibit negative correlations [8].

The next step is machine learning modelling, applying three algorithms: Linear Regression as the base model, Decision Tree to capture non-linear patterns, and Random Forest to improve accuracy using ensemble techniques. Model evaluation was performed using the MAE, MSE, and R^2 metrics. To strengthen the analysis, visualization of the prediction results was performed, including a comparison graph between the predicted and observed values. This served to assess how well the model could reflect the actual data. The last step was the interpretation and analysis of the results, which compared the performance of each model scientifically [2].

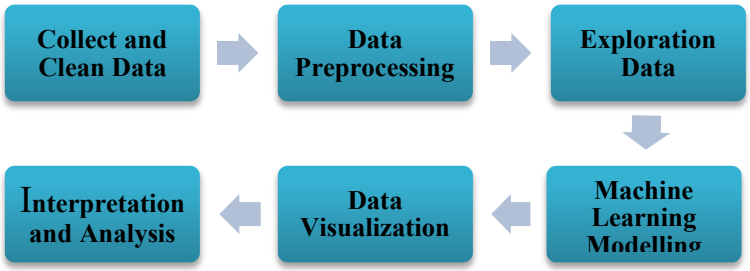

Fig 1. Research Flow

Data Source

The data used in this study was obtained from NASA's POWER (Prediction of Worldwide Energy Resource) platform, which is a global database for reprocessing meteorological and solar radiation information [7]. The NASA POWER dataset provides daily data with international coverage, including variables related to solar radiation and important weather elements. For the case study in Batam City (Riau Islands Province, Indonesia), information was downloaded for coordinates 1.1288° N, 104.0403° E. BT (Batam is at 1.13°N, 104.04°E) for the period from January 1, 2020, to July 2, 2025. The primary variables included are daily solar radiation irradiance at the surface in units of MJ/m²/day, as well as other weather variables such as air temperature, relative humidity, precipitation, surface pressure, and wind speed [7]. One advantage of NASA POWER data is that it provides data for specific locations like Batam. However, it is important to note that model-based reanalysis data (such as MERRA-2 and CERES) may have gaps or inaccuracies, especially in tropical regions with limited field observations [7]. Therefore, before modelling, data cleaning and transformation are necessary to ensure the consistency and completeness of the dataset.

Data processing was performed using Jupyter Notebook with Python programming. Some of the main steps in data processing include:

- Data Download and Division: The large NASA POWER dataset was downloaded based on annual periods (2020–2025) to facilitate file management. Each CSV file stores daily data from that year.
- Initial Reading and Inspection: CSV files were opened and inspected using the panda's library in Python. The column structure was checked, including the header rows and extreme values (-999 indicates missing data as noted in the metadata).
- Date Conversion: The original data displayed the time in the "YEAR" and "DOY" (day-of-year) columns. We use a Python function to combine YEAR and DOY into a standard date format (YYYY-MM-DD). Then, the time information is expanded by adding separate columns for Month and Day so that seasonal or daily analysis can be performed in more detail.
- Missing Data Handling: Missing or invalid values in the NASA POWER dataset are represented by the placeholder value −999. In this study, these values were first identified and converted into missing entries (NaN). Subsequently, all records containing missing values were removed from the dataset to maintain data consistency and avoid potential bias in the machine learning models. After the cleaning process, the dataset was re-examined to ensure that no missing values remained before proceeding to the modelling stage.
- Data Reorganization: After the data is cleaned and formatted, it is reorganized based on time ranges (e.g., monthly) according to the model's requirements. This process enables more efficient visualization and modelling.

These data preparation steps within the Python ecosystem in Jupyter Notebook provide significant flexibility. The final processed data is then used in the development of predictive models such as linear regression, Random Forest, and Decision Tree. Thus, the data processing and transformation workflow ensures high-quality input for the model, in line with recommendations in the literature on the use of satellite data for renewable energy analysis [7].

Feature Selection

Feature selection is an especially crucial step in developing models for predicting solar radiation, as it can improve the accuracy and effectiveness of the model. By selecting relevant features and removing fewer valuable ones, the complexity of the model can be reduced, thereby reducing the risk of overfitting and speeding up processing time. Good feature selection also ensures that the model only learns from attributes that truly contribute to solar radiation, resulting in improved prediction accuracy. Conversely, irrelevant, or noisy features can degrade model performance. Therefore, the feature selection process is crucial for selecting input variables before model creation [8].

This step aims to convert temporal data into numerical features that can be used by the model. For example, information about the month can capture the seasonal impact of sunlight throughout the year. All the features mentioned earlier are then entered into the model without any being removed, as initial analysis shows that each feature contributes useful information for predicting solar radiation. No features were completely removed from the dataset, as no features were statistically proven to be irrelevant after correlation checks and feature importance assessments [8].

Table 1. Main features considered for the proposed models.

Feature	Significance	Impact
Date (YEAR/DOY)	The combination of year and day-to-n (Day of Year) serves as a time marker that captures seasonal and annual radiation trends.	The addition of time features allows the model to recognize periodic patterns (dry season, rainy season) and long-term trends such as climate change, improving the accuracy of temporal predictions.
Solar Radiation (MJ/m²/day)	Main target/output. Indicates the amount of daily irradiance received by the surface in units of MJ/m²/day.	This is the variable predicted by the model. Its value is strongly influenced by atmospheric factors such as clouds, water vapor, and particles.
Air Temperature (°C)	Describes the average daily air temperature at a height of 2 meters. Affects atmospheric dynamics and surface warming.	Temperature is an important indicator in convection processes and cloud formation, which influence radiation reduction or reflection. This helps the model relate thermal energy fluctuations with daily radiation.
Relative Humidity (%)	Air humidity at 2 meters above the surface. Reflects the water vapor content that affects light transmission and scattering.	High humidity leads to more cloud formation and radiation scattering effects, especially in tropical regions. This feature helps the model predict radiation decreases on cloudy or rainy days.
Rainfall (mm/day)	Corrected precipitation amount in mm/day. Precipitation is strongly correlated with cloud formation and reduced radiation reaching the surface.	Rainfall acts as an indirect predictor of cloud density. This feature allows the model to capture extreme events such as heavy rainfall that significantly reduce radiation values.
Surface Pressure (kPa)	Atmospheric pressure at the surface in kPa. Influences air density and its ability to absorb and reflect radiation.	Pressure fluctuations indicate weather changes (e.g., storms or local convergence). This helps the model detect unstable atmospheric conditions affecting radiation intensity.
Wind Speed (m/s)	Wind speed measured at a height of 2 meters. Wind plays a role in distributing clouds, water vapor, and aerosols in the atmosphere.	Strong winds can disperse clouds or transport moisture to other areas, influencing sky clarity. This feature helps detect sudden weather variations that affect solar radiation.

Research Model

Linear Regression (LR)

Linear regression is a statistical method for analysing and modelling the relationship between one dependent variable and one or more independent variables. The basic purpose of this analysis is to predict the value of the dependent variable using the values of the independent variables. This is achieved by identifying the linear equation that best fits the observed data, reducing the difference between the predicted values and the actual values [8]. This study uses multiple linear regression with the multiple regression equation:

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_pX_p \quad (1)$$

Explanation:
Y = dependent variable
b_0 = constant
$b_1 = b_2 = \cdots = b_P$ = regression coefficient
$X_1 = X_2 = \cdots = X_P$ = independent variable

Where the prediction variable is solar radiation as the dependent variable, and the predictor or independent variables are air temperature, relative humidity, rainfall, surface pressure, and wind speed.

Decision Tree Regressor

The Decision Tree Regressor builds the model by partitioning the data into subsets that minimize the variance of the target variable. The splitting criterion used is the reduction in Mean Squared Error (MSE), defined as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2 \quad (2)$$

where N is the number of samples in the node, y_i is the actual value, and $\bar{y}$ is the mean of the target variable within that node.

Random Forest Regressor

The random forest model is the most popular technique for regression and classification in decision tree learning. This model is very efficient, and, at the same time, its regression accuracy is better than other regression methods. This random model builds many interconnected decision trees in the training phase. After developing several decision trees, the output of the model is obtained by averaging the output values of all individual trees [11].

$$Q = \frac{1}{n}\sum_{i=1}^{n}(y_i - y) \tag{3}$$

Where:

$$y = \frac{1}{n}\sum_{i=1}^{n}(y_i) \tag{4}$$

Explanation:
n = number of data points
y_i = response value
y = predicted value at the node (average response value)

Hyperparameter Configuration

A machine learning algorithm's ability to identify patterns in the data is influenced by hyperparameters, which are crucial parameters that are set prior to training. Appropriate hyperparameter selection enhances predictive models' capacity for generalization and helps balance the bias-variance trade-off [12].

The Scikit-learn Python library was used to implement the machine learning models in this study. The models were run using the default hyperparameter configuration supplied by the Scikit-learn implementation in order to provide a consistent and equitable comparison among the assessed algorithms. This method is frequently used in baseline machine learning research to assess various algorithms' intrinsic prediction power without requiring a lot of hyperparameter adjustment.

As a foundational statistical model, the Linear Regression model was used. In the meantime, potential non-linear correlations between solar radiation and meteorological variables including temperature, humidity, rainfall, pressure, and wind speed were captured using tree-based algorithms like Decision Tree and Random Forest. While the Decision Tree model offers a clear framework for comprehending the connection between predictor variables and solar radiation, the Random Forest model functions as an ensemble learning technique that integrates several decision trees to increase prediction stability through averaging.

Table 2. Hyperparameter configuration of the machine learning models

Models	Hyperparameter	Value	Description
Linear Regression	parameters	default (Scikit-learn)	Baseline linear regression model
Decision Tree	parameters	default (Scikit-learn)	Tree-based regression model
Random Forest	parameters	default (Scikit-learn)	Ensemble regression model

Model Evaluation Criteria

First, compare solar radiation data with air temperature, relative humidity, rainfall, surface pressure, and wind speed data. Plot graphs to help understand and visualize exactly how the available data sets are related. To maximize model performance, minimize the predefined loss function and achieve better results with fewer errors by tuning hyperparameters [12].

Mean square error (MSE) is the simplest function that can be calculated in machine learning. MSE takes the difference between the model's prediction and the actual data or ground truth, squares it, and applies the average across the entire dataset [13][14]. MSE will never be negative with the formula (2)
Explanation:
N = number of data points
y_i = actual data value
y = predicted data value

The mean absolute error (MAE) is a model prediction performance indicator, which is achieved by observing how close the predicted variable is to the observed variable [13]:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y| \quad (5)$$

Explanation:
N = number of data points
y_i = actual data value
y = predicted data value

R^2 is a value that shows how much the independent variable (exogenous) affects the dependent variable (endogenous) [12]. With the formula:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y_i})^2}{\sum(y_i - y)^2} \quad (6)$$

Explanation:
y_i = actual observed value
$\hat{y_i}$ = predicted value
$\bar{y}$ = mean of observed values

R^2 represents the proportion of variance in the dependent variable that can be explained by the independent variables.

3. RESULT AND DISCUSSION

Data Analysis

Table 3 presents a summary of descriptive statistics for the meteorological variables used in modelling daily solar radiation predictions. The parameters shown include maximum, minimum, average, and standard deviation values, which describe the distribution and diversity of data within the observation range [13].

Table 3. Model Evaluation Summary

Parameters	Max	Min	Mean	Sdt
Solar radiation (MJ/m²/day)	24.96	1.15	17.39	4.85
Air Temperature (°C)	30.18	25.61	28.13	0.88
Relative Humidity (%)	92.43	70.64	81.95	3.61
Rainfall (mm/day)	255.88	0.00	12.01	17.61
Surface Pressure (kPa)	101.30	100.50	100.84	0.13
Wind Speed (m/s).	6.77	0.66	3.02	1.04

The target variable, Solar Radiation, shows a maximum value of 24.96 MJ/m²/day and a minimum value of 1.15 MJ/m²/day, with an average of 17.39 MJ/m²/day and a standard deviation of 4.85. The significant standard deviation indicates a notable daily variation in the intensity of solar radiation received by the Earth's surface. This reflects that the dataset encompasses a range of weather conditions, from very sunny days to days with heavy cloud cover or overcast skies. Air temperature exhibits a more consistent distribution, with an average value of 28.13°C and a standard deviation of only 0.88. The temperature ranges from 25.61°C to 30.18°C indicates that the observation location is in a tropical climate with small daily temperature variations.

This temperature consistency may influence the stability of solar radiation, as temperature is related to atmospheric conditions such as cloud density and sky clarity.

Relative humidity is also within a narrow range, with an average of 81.95% and a standard deviation of 3.61. This indicates that humidity tends to be high and stable, as is typically the case in tropical regions with high air humidity throughout the year. On the other hand, the Rainfall variable shows a very uneven distribution, with a minimum value of 0 mm/day and a maximum of 255.88 mm/day, and a standard deviation of 17.61. This indicates extreme conditions in the dataset, ranging from dry days to heavy rainfall. This is important to note because high rainfall is negatively correlated with solar radiation intensity due to increased cloud cover.

Air surface pressure shows extraordinarily slight variation, with an average of 100.84 kPa and a standard deviation of only 0.13. This indicates that atmospheric pressure changes in the study area are stable, reflecting geographical conditions without extreme topographical or climatic fluctuations. Meanwhile, wind speed has an average of 3.02 m/s with a standard deviation of 1.04. The range of values from 0.66 to 6.77 m/s indicates moderate variation in this parameter. Although not a direct factor influencing radiation values, wind speed can affect cloud distribution and thus influence the amount of radiation reaching the surface. Overall, this descriptive statistical information provides important insights into the characteristics of the data used in modelling. Variability among variables should be considered during the preprocessing stage and in selecting a model capable of addressing high data variability, especially for variables with large fluctuations such as rainfall and solar radiation itself.

Feature-Target Relationship Analysis

To gain insight into how each input variable affects solar radiation values, scatter plots were created to show the relationship between each feature and the target. The results of these graphs provide an initial understanding of the type of interaction (linear, non-linear, or irregular) between the input variables and the predicted results.

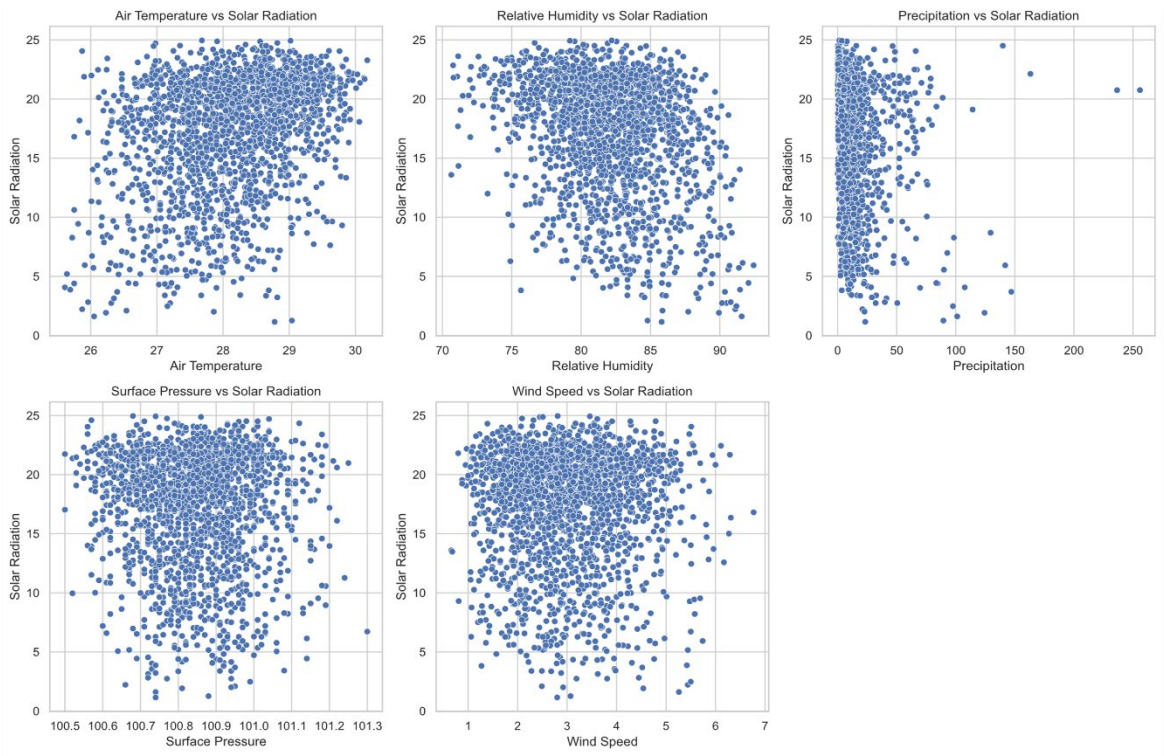

Fig 2. Feature-Target Relationships

Air temperature shows a positive semi-linear relationship with solar radiation, which is consistent with meteorological theory, whereby sunny days tend to have higher temperatures and higher solar radiation. Although this relationship is consistent with a linear model, a non-linear approach may still provide improvements, especially in extreme conditions. Relative humidity shows a non-linear inverse relationship with solar radiation. High humidity, often associated with clouds or fog, is associated with a decrease in solar

radiation reaching the surface. This non-linearity suggests that models such as Decision Tree or Random Forest are more effective in describing this dynamic.

Rainfall exhibits a unique threshold effect: solar radiation remains high under dry conditions but decreases sharply even with minimal rainfall. This discontinuous nonlinear pattern underscores the need for flexible modelling methods capable of addressing sudden changes in response behaviour. In this case, wind speed and surface pressure do not show clear or consistent patterns with solar radiation. While there are some noticeable trends, such as a decrease in radiation under exceptionally low or high wind conditions, these variables have limited predictive capability when considered individually, although their importance in multivariate interactions requires further investigation using model-based importance metrics.

Correlation Analysis

To evaluate the linear relationship between variables, a correlation matrix is constructed and displayed via a heat map. This matrix quantitatively shows the extent to which one variable is linearly related to another, measured using Pearson's correlation coefficient, which varies from -1 to 1. Positive values indicate a direct relationship, while negative values indicate an inverse relationship. The explanation below highlights the main patterns related to solar radiation modelling [15].

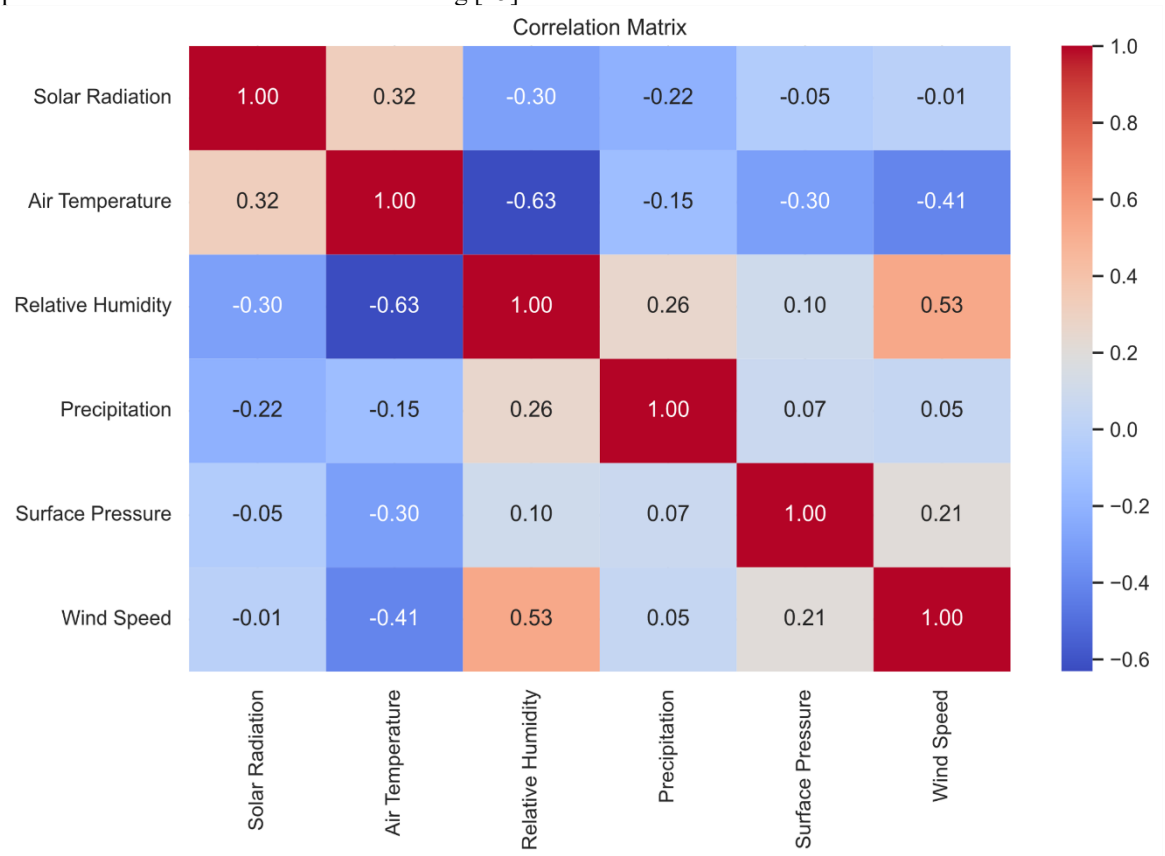

Fig 3. Correlation Matrix

Of all the features available, air temperature shows the highest positive correlation with solar radiation. This is in line with the principle of physics which states that sunny conditions when sunlight reaches its peak often occur simultaneously with higher temperatures. This confirms the role of temperature as the main predictor in solar radiation estimation models. Meanwhile, relative humidity was found to be negatively correlated with solar radiation. This association is intuitive: increased humidity is typically associated with the presence of clouds or rain, both of which contribute to a reduction in the amount of solar energy received. Therefore, relative humidity not only acts as a direct modifier of solar radiation but also as an indicator of atmospheric clarity.

Precipitation also shows a moderate negative correlation with solar radiation. Although the strength of this correlation is lower than that of humidity, there is still an indication that rainy conditions are associated with lower solar radiation. This makes precipitation a valuable element, especially in models designed to manage non-linear changes and sudden shifts. Surface pressure and wind speed have a weak or negligible correlation with solar radiation. However, their role should not be entirely ignored; in the context of specific

local meteorological systems or when interacting with other variables, they can contribute marginally to improving prediction accuracy.

Research Results

This study aims to estimate solar radiation on the surface of Batam City using regression-based machine learning methods. The dataset includes daily meteorological variables such as year, date, air temperature, relative humidity, atmospheric pressure, rainfall, and wind speed, with solar radiation as the target variable. The data underwent a preprocessing and cleaning stage to ensure the quality and consistency of the input variables before model training.

Considering the temporal nature of meteorological data, the dataset was divided using a chronological split rather than random sampling. Data from earlier years were used for model training, while the most recent portion of the dataset was reserved for testing. This approach prevents data leakage and better reflects real-world forecasting scenarios, where models are trained using historical data and evaluated on future observations [1].

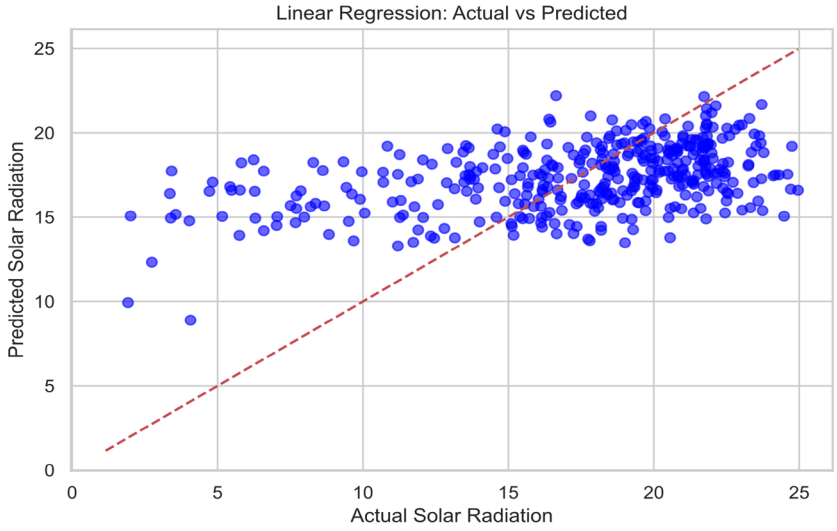

Fig 4. Actual Solar Radiation vs Predicted Solar Radiation Model Linear Regression

The relationship between the actual values and the predicted values in Fig. 4 has an MSE value of 19.78, a MAE value of 3.47, and an R^2 value of 0.19. The points in Fig. 4 appear to be more scattered than the diagonal line. This indicates that Linear Regression tends to underfit, as it is unable to capture the complexity of the non-linear relationship between the input and output variables.

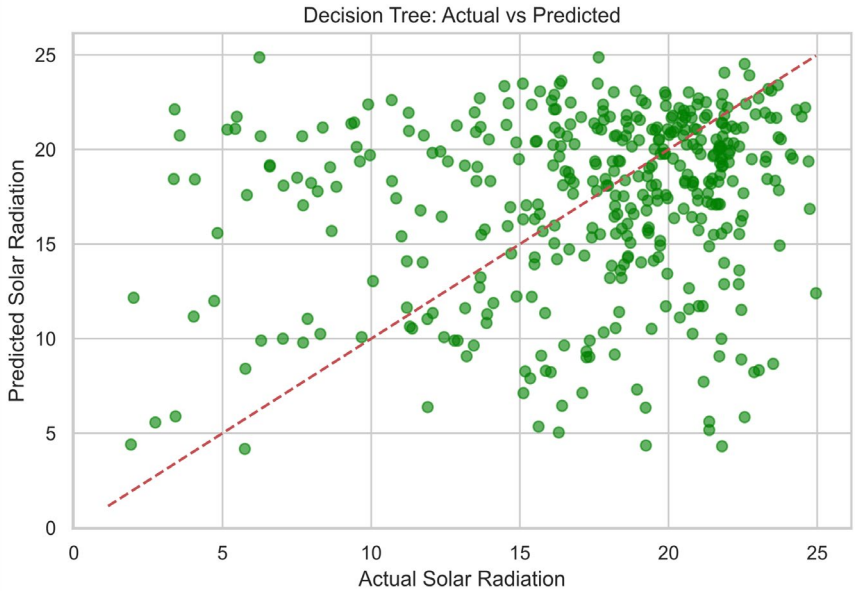

Fig 5. Actual Solar Radiation vs Predicted Solar Radiation Model Decision Tree Regressor

The relationship between the actual and predicted values in Fig. 5 yields an MSE of 37.42, an MAE of 4.58, and an R^2 value of −0.53. A negative R^2 indicates that the Decision Tree model performs worse than a simple baseline prediction based on the mean of the target variable. This suggests that the model failed to capture meaningful predictive patterns within the dataset.

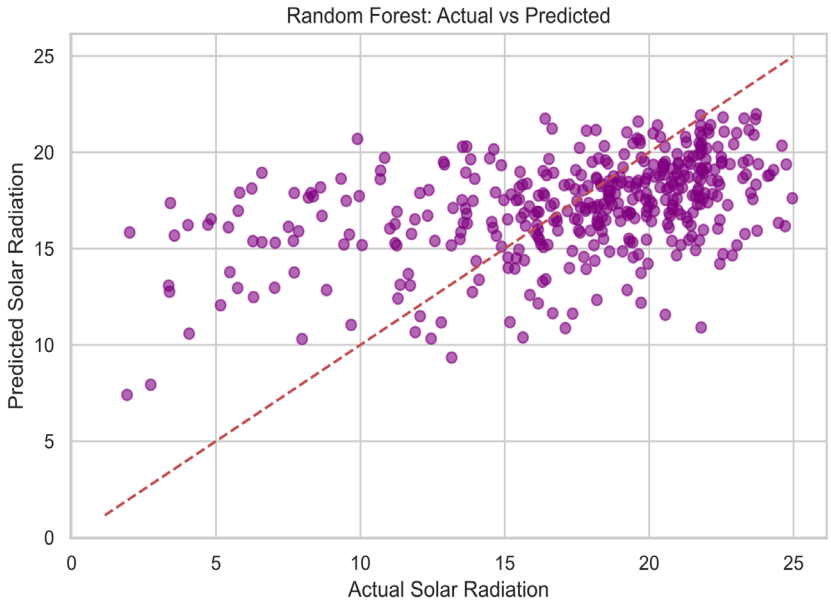

Fig 6. Actual Solar Radiation vs Predicted Solar Radiation Model Random Forest Regressor

The relationship between the actual value and the predicted value in Fig. 6 has an MSE value of 19.61, a MAE value of 3.42, and an R^2 value of 0.20. The points are distributed most densely along the diagonal line, indicating an elevated level of accuracy among the three and good generalization without excessive overfitting.

Discussion

The three machine learning models were trained and evaluated to predict solar radiation based on predetermined variables. The evaluation was conducted using three metrics:

Table 4. Model Evaluation Summary

Model	MSE	MAE	R^2
Linear Regression	19.79	3.47	0.19
Decision Tree	37.42	4.58	-0.53
Random Forest	19.61	3.42	0.20

Table 4 supports the selection of Random Forest achieved the best relative performance among the evaluated models in this study. This model effectively and efficiently balances the bias-variance trade-off. Therefore, it is suitable for predicting solar radiation in the real world even though the data is unstable and the relationship is non-linear.

It is important to note that both the predictor variables and the target solar radiation values were obtained from the NASA POWER dataset. Therefore, this study does not represent a validation against independent ground-based measurements such as pyranometer observations. Instead, the models should be interpreted as learning the internal relationships within the NASA POWER dataset. Thus, the results represent an emulation or approximation of NASA POWER solar radiation estimates rather than an independent validation of solar radiation prediction in Batam.

Implications and Recommendations

This study provides a path for practitioners in energy and environmental management in tropical regions. In practical terms, the Random Forest model obtained is a vital tool for the Batam City government and renewable energy agencies [16]. With an accuracy of ±3.42 MJ/m² in predicting daily solar radiation, the model can enable more precise planning of solar power plant (SPP) infrastructure, thereby maximizing local solar potential [4]. These findings can drive policies for solar power plants in Batam City. However, caution is still needed, as reliance on satellite data risks leaving biases in tropical regions, where dynamic cloud cover

may not be fully recorded. Therefore, collaboration with the Indonesian Meteorological, Climatological, and Geophysical Agency (BMKG) is necessary to validate ground-truth data, which is key to mitigating these limitations.

4. CONCLUSION

This research approach indicates that the synergy between open satellite data and artificial intelligence can offer a viable pathway to overcoming energy data scarcity in remote areas. Among the evaluated models, the Random Forest model produced the best performance for predicting solar radiation in Batam City due to its ability to capture non-linear interactions between air temperature, relative humidity, rainfall, surface pressure, and wind speed.

However, while the model recorded an MSE of 19.61 and an MAE of 3.42, the resulting R^2 value of 0.20 must be interpreted with caution. This result indicates that the current meteorological predictors explain only about 20% of the variance in solar radiation, suggesting that predicting solar irradiance in Batam's coastal tropical climate remains challenging with the current set of variables.

The performance of the model highlights both the potential and the limitations of data-driven approaches in the Indonesian context [3]. On one hand, satellite reanalysis data can partially address the scarcity of ground-based radiation observations. On the other hand, the relatively low explanatory power suggests that additional meteorological or atmospheric variables are required to improve predictive capability.

These findings provide an important baseline for future solar energy modelling studies in tropical regions that experience complex atmospheric dynamics. Future research should integrate additional physically relevant predictors such as cloud cover indicators, aerosol proxies, or atmospheric transparency indices to enhance model reliability. In addition, validation using independent ground-based observations would further strengthen the scientific robustness of solar radiation prediction studies in Indonesia.

Overall, this research contributes to understanding the complexity of solar radiation modelling in tropical coastal environments and highlights the importance of interdisciplinary collaboration among meteorological institutions, energy practitioners, and data scientists to support the development of renewable energy systems in Indonesia.

REFERENCE

[1] L. Huang, J. Kang, M. Wan, L. Fang, C. Zhang, and Z. Zeng, "Prediction of Solar Radiation Using Various Machine Learning Algorithms and Its Implications for Extreme Climate Events," Front Earth Sci (Lausanne), vol. 9, April 2021, doi: 10.3389/feart.2021.596860.

[2] J. Fan et al., "Empirical Models and Machine Learning for Predicting Daily Global Solar Radiation from Sunshine Duration: A Review and Case Study in China," Renewable and Sustainable Energy Reviews, vol. 100, pp. 186–212, Feb. 2019, doi: 10.1016/j.rser.2018.10.018.

[3] M. Jamei et al., "Data-Based Model for Predicting Solar Radiation in Semi-Arid Regions," Computers, Materials and Continua, vol. 74, no. 1, pp. 1625–1640, 2023, doi: 10.32604/cmc.2023.031406.

[4] R. Ismayanti and W. Maulana Baihaqi, "Predicting the Potential of an Area to Become a Solar Power Plant Using Machine Learning," JIITE.

[5] V. Demir, "Evaluation of Solar Radiation Prediction Models Using AI: Performance Comparison in the High-Potential Region of Konya, Türkiye," Atmosphere (Basel), vol. 16, no. 4, Apr. 2025, doi: 10.3390/atmos16040398.

[6] Q. Li, M. Bessafi, and P. Li, "Mapping Surface Solar Radiation Predictions with a Linear Regression Model: A Case Study on Reunion Island," Atmosphere (Basel), vol. 14, no. 9, September 2023, doi: 10.3390/atmos14091331.

[7] L. P. Darman, Januhariadi, M. P. Yudha, and Aslan, "Assessment of NASA POWER Reanalysis Products as an Alternative Data Source for Weather Monitoring in West Sumbawa, Indonesia," in E3S Web of Conferences, EDP Sciences, Feb. 2024. doi: 10.1051/e3sconf/202448506006.

[8] I. K. Tanoli et al., "Machine learning for high-performance solar radiation prediction," Energy Reports, vol. 12, pp. 4794–4804, Dec. 2024, doi: 10.1016/j.egyr.2024.10.033.

[9] P. Setiawati et al., "SOLAR ENERGY GENERATION PREDICTION: A Machine Learning Approach for Network Stability and Efficiency," Jurnal Pilar Nusa Mandiri, vol. 21, no. 1, pp. 34–43, Mar. 2025, doi: 10.33480/pilar.v21i1.6126.

[10] U. V. Teja, M. Sai Kiran, V. Karthikeya, M. E. Murali, and T. Kumanan, "Prediction of Solar Radiation Using Machine Learning and Python."

[11] R. Srivastava, A. N. Tiwari, and V. K. Giri, "Prediction of Solar Radiation Using MARS, CART, M5, and Random Forest Models: A Case Study for India," Heliyon, vol. 5, no. 10, October 2019, doi: 10.1016/j.heliyon.2019.e02692.

[12] D. Ardiansyah, "COMPARISON OF SUNRADIATION PREDICTION MODELS BASED ON MACHINE LEARNING AT THE FATMAWATI SOEKARNO BENGKULU METEOROLOGICAL STATION," Megasains, vol. 14, no. 1, Sep. 2023, doi: 10.46824/megasains.v14i1.129.

[13] C. G. Villegas-Mier, J. Rodriguez-Resendiz, J. M. Álvarez-Alvarado, H. Jiménez-Hernández, and Á. Odry,

"Optimized Random Forest for Solar Radiation Prediction Using Sunlight Hours," Micromachines (Basel), vol. 13, no. 9, September 2022, doi: 10.3390/mi13091406.

[14] M. Attya, O. Abo-Seida, H. Mohamed, and A. Mohammed, "A Hybrid Deep Learning Framework for Solar Radiation Prediction Based on Satellite Images and Regional Data," Neural Comput Appl, July 2025, doi: 10.1007/s00521-025-11197-3.

[15] A. N. Oktaviani et al., "Analysis of Solar Radiation Prediction Using Machine Learning Algorithms and Bayesian Optimization Implementation in the Province of DKI Jakarta," Journal of Information Management & Information Systems (MISI), vol. 8, no. 1, 2025, doi: 10.36595/misi.v5i2.

[16] A. Mujtaba, "Towards Sustainable Energy Policies: Machine Learning Applications in Projecting Bio Solar Consumption in Indonesia," 2024, doi: 10.18196/jsp/v15i1.360.