Vol. 8, No.2, Juli 2025, Halaman 431-444 **ISSN 2622-7398**
DOI:https://doi.org/10.30596/rmme.v8i2.26011 http://jurnal.umsu.ac.id/index.php/RMME
Jurnal Rekayasa Material, Manufaktur dan Energi

Machine Learning-Based Regression Model for Predicting Global Horizontal Radiation and Global Horizontal Irradiance: A Case Study in Banda Aceh

M Salamul Fajar S*[1], Ikramullah Muhammad[2], Akbar Rizqullah[3] Thaharul Fikri[4], Nural Fajri[5], & Faris Ahmad Mizanus S[6]
[1,2,3,4,5]Department of Mechanical and Industrial Engineering, Universitas Syiah Kuala, Indonesia
[6] Department of Mechanical Engineering, Universitas Samudera, Indonesia
*Email: salamulfajar@usk.ac.id

ABSTRACT
Global Horizontal Radiation (GHR) and Global Horizontal Illumination (GHI) are critical environmental parameters that play a vital role in solar energy development, precision agriculture, and sustainable urban planning. However, their prediction remains challenging due to the high variability caused by atmospheric conditions. This study evaluates the performance of various machine learning models in predicting GHR and GHI using a comprehensive dataset comprising 29 environmental features. The models tested include Linear Regression, Random Forest Regressor, XGBoost Regressor, LightGBM Regressor, Support Vector Regressor (SVR), and Artificial Neural Network (ANN). The results consistently show that ensemble-based models, particularly LightGBM Regressor, provide the best predictive performance for both target variables, achieving very high R-squared values (approaching 0.999). XGBoost and Random Forest also demonstrate highly competitive performance. ANN performs well, while Linear Regression and SVR show lower accuracy. These findings underscore the significant potential of advanced machine learning models in predicting environmental parameters with high accuracy, which has important implications for renewable energy optimization, smart agriculture, and sustainable urban planning.

Keywords: *Global Horizontal Radiation, Global Horizontal Illumination, Machine Learning, Environmental Prediction.*

INTRODUCTION

Solar radiation and global horizontal illumination are critical environmental parameters with extensive implications across various sectors, including renewable energy, agriculture, and urban planning. Solar radiation, as the Earth's main energy source, plays a vital role in climate regulation and ecological balance, while global horizontal illumination affects visibility, natural lighting, and building energy efficiency [1–3]. Accurate prediction of these parameters is essential for optimizing photovoltaic (PV) systems, planning sustainable agriculture, and designing energy-efficient urban infrastructure.

Conventional measurement methods, such as ground-based pyranometers or illuminance meters, although precise, are costly and geographically limited, especially in developing regions. This limitation drives the demand for alternative predictive techniques capable of utilizing more accessible meteorological data, such as temperature, humidity, atmospheric pressure, wind speed, and cloud cover [4]. In recent years, machine learning (ML) has emerged as a powerful tool for modeling the complex and nonlinear interactions between environmental variables and solar energy indicators, offering scalability and generalization across locations [5].

Numerous studies have demonstrated the effectiveness of ML algorithms in solar radiation forecasting. Park et al. [6] demonstrated the use of Light Gradient Boosting Machine (LightGBM) for multi-step-ahead forecasting of solar radiation in Jeju Island, South Korea. Their results revealed not only high accuracy but also the advantages of fast training and handling large datasets with minimal computational overhead. Solano and Affonso [7] utilized ensemble voting that combined multiple algorithms such as Random Forest and XGBoost, which significantly improved the accuracy of short-term solar irradiation prediction in Brazil. Additionally, Wu et al. [3]

Copyright@2025 Jurnal Rekayasa Material, Manufaktur dan Energi. This is an open acces article under the CC-BY-SA lisence (https://creativecommons.org/licenses/by-sa/4.0/).

employed visible all-sky imaging integrated with machine learning for predicting GHI in Tibet, effectively addressing challenges posed by frequent cloud cover.

The applicability of ML-based prediction has also been validated in arid and tropical climates. Nematchoua et al. [8] evaluated six machine learning algorithms to predict daily global solar radiation across 27 European countries. They concluded that ensemble-based methods such as Random Forest and Gradient Boosting consistently outperformed conventional linear models. Imam et al. [2] investigated the performance of six different regressors including ANN, Random Forest, and SVR—for solar irradiance prediction in Northern Saudi Arabia, highlighting the robustness of nonlinear models under arids climatic conditions. Buster et al. [1] integrated physical laws with ML to enhance prediction accuracy in the National Solar Radiation Database (NSRDB), showcasing the potential of hybrid approaches. In addition, Zhou et al. [9] compared various tree-based ensemble models LightGBM, XGBoost, and CatBoost across several Chinese meteorological stations. Their study emphasized the importance of feature selection and hyperparameter optimization for maximizing prediction accuracy.

These studies collectively suggest that machine learning, particularly ensemble methods and hybrid physical data approaches, holds great promise for predicting solar radiation and illumination across diverse climatic zones. However, further research is needed in tropical settings like Banda Aceh, where cloud dynamics and atmospheric variability present unique challenges to accurate forecasting.

Solar radiation and illumination prediction have been an active area of research for several decades, with various approaches developed to improve accuracy and reliability. Broadly, prediction methods can be categorized into physical models, statistical models, and machine learning-based models [3]. Physical models are based on atmospheric physics equations and radiation transfer, often requiring extensive data input and intensive computation. Statistical models, on the other hand, use historical relationships between radiation and other meteorological variables but may struggle to capture complex non-linear patterns. With the advent of big data and increased computational power, machine learning-based models have shown superior performance in many cases [2].

Given the high solar potential and dynamic tropical weather in Banda Aceh, Indonesia, a reliable data-driven approach is critical for regional energy and sustainability planning. This study aims to evaluate and compare the performance of several ML regression algorithms Linear Regression, Random Forest, XGBoost, LightGBM, SVR, and ANN for predicting both Global Horizontal Radiation and Global Horizontal Illumination. A dataset comprising 29 environmental features is used, and models are assessed using MAE, MSE, RMSE, and R^2. The results are expected to inform future solar energy deployment strategies and support sustainable development in tropical urban contexts.

METHODOLOGY

This section outlines the methodological approach used in this study to predict GHR and GHI. It includes a description of the dataset, data pre-processing steps, the machine learning models applied, and the evaluation metrics used to assess model performance. Figure 1 illustrates the overall methodology adopted in this research, from data loading and pre-processing to model training, evaluation, and result analysis. This diagram illustrates a common machine learning methodology workflow, from initial data handling to result analysis.

Copyright@2025 Jurnal Rekayasa Material, Manufaktur dan Energi. This is an open acces article under the CC-BY-SA lisence (https://creativecommons.org/licenses/by-sa/4.0/).

Vol. 8, No.2, Juli 2025, Halaman 431-444 ISSN 2622-7398
DOI:https://doi.org/10.30596/rmme.v8i2.26011 http://jurnal.umsu.ac.id/index.php/RMME
Jurnal Rekayasa Material, Manufaktur dan Energi

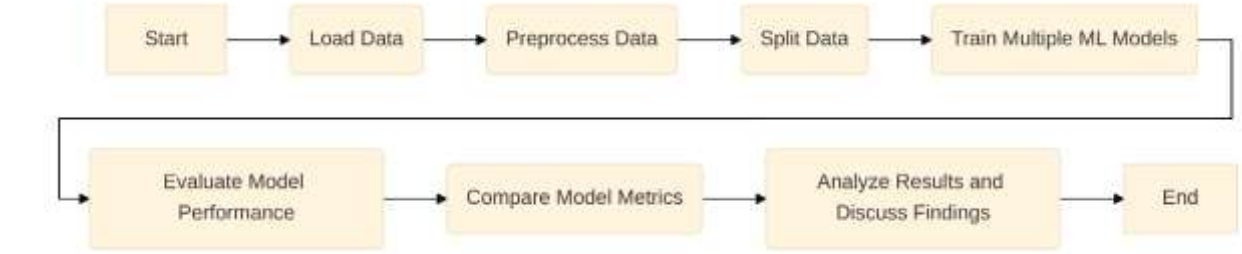


Figure 1. Research Methodology Flowchart

The process begins with loading data, followed by identifying features and targets, handling missing values, normalizing features, and splitting the data into training and testing sets. After data preparation, multiple machine learning models such as Linear Regression, Random Forest Regressor, XGBoost Regressor, LightGBM Regressor, SVR, and ANN Regressor are trained in parallel. Following training, the performance of each model is evaluated, and model metrics are compared to determine the best-performing model. The final stages involve a thorough analysis of the results and discussion of findings, leading to the conclusion of the machine learning process. This flow demonstrates a comprehensive approach to building and evaluating predictive models.

Dataset Description

The dataset utilized in this study was sourced from app.ensims.com, specifically for the Banda Aceh region. It comprises 8,395 records and includes 29 numerical features. These features encompass a wide range of environmental and meteorological variables that are pertinent to the prediction of solar radiation and illumination. Among them are dry bulb temperature, dew point temperature, relative humidity, atmospheric pressure, extraterrestrial radiation, horizontal infrared radiation, global horizontal radiation, direct normal radiation, and diffuse horizontal radiation. Additionally, the dataset includes global horizontal illumination, direct normal illumination, diffuse horizontal illumination, zenith luminance, wind direction and speed, total sky cover, opaque sky cover, visibility, ceiling height, precipitable water, aerosol optical depth, surface albedo, liquid precipitation depth and amount, as well as wind-related radiation features such as direct normal and diffuse horizontal wind energy radiation.

The Figure 2 displays a pairplot (scatterplot matrix) that visualizes the pairwise relationships among several key numerical variables within the dataset employed for predicting solar radiation and illumination in Banda Aceh. The variables examined include *DryBulbTemperature*, *RelativeHumidity*, *WindSpeed*, *GHR*, *DirectNormalRadiation*, and *DiffuseHorizontalRadiation*. Each subplot in the matrix corresponds to a scatter plot between two variables, while the diagonal elements present the individual distributions of each feature. The visualization reveals a clear **positive correlation** between *DryBulbTemperature* and both *GHR* and *DirectNormalRadiation*, indicating that higher air temperatures tend to coincide with increased solar radiation levels. In contrast, *RelativeHumidity* exhibits a **negative association** with *GHR*, suggesting that elevated humidity often associated with cloud formation and atmospheric moisture can diminish the amount of solar irradiance reaching the surface.

Copyright@2025 Jurnal Rekayasa Material, Manufaktur dan Energi. This is an open acces article under the CC-BY-SA lisence (https://creativecommons.org/licenses/by-sa/4.0/).

Vol. 8, No.2, Juli 2025, Halaman 431-444 ISSN 2622-7398
DOI:https://doi.org/10.30596/rmme.v8i2.26011 http://jurnal.umsu.ac.id/index.php/RMME
Jurnal Rekayasa Material, Manufaktur dan Energi

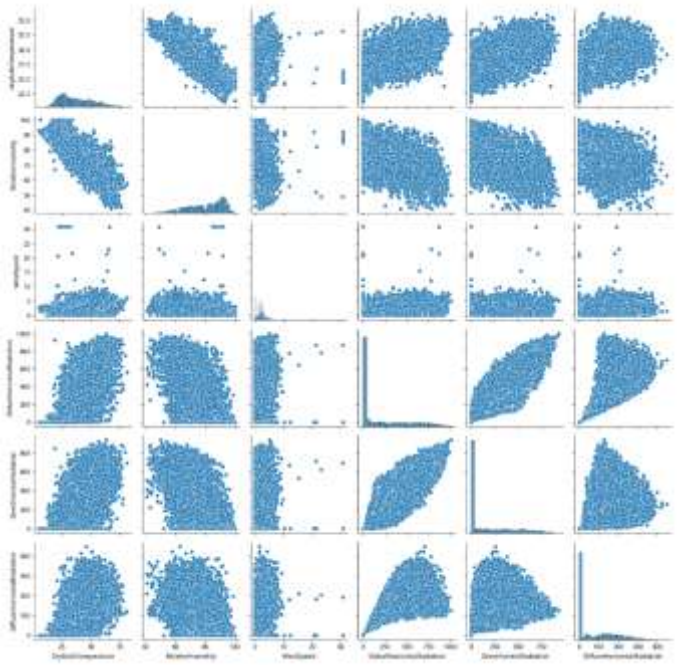


Figure 2. Data relationships scatterplot matrix

Meanwhile, *WindSpeed* appears to have minimal or no significant correlation with the radiation and humidity variables, implying a weaker direct influence in this context. These visual patterns are essential for identifying feature importance and interactions among variables, which is particularly useful during the feature selection phase of machine learning modeling for solar radiation prediction.

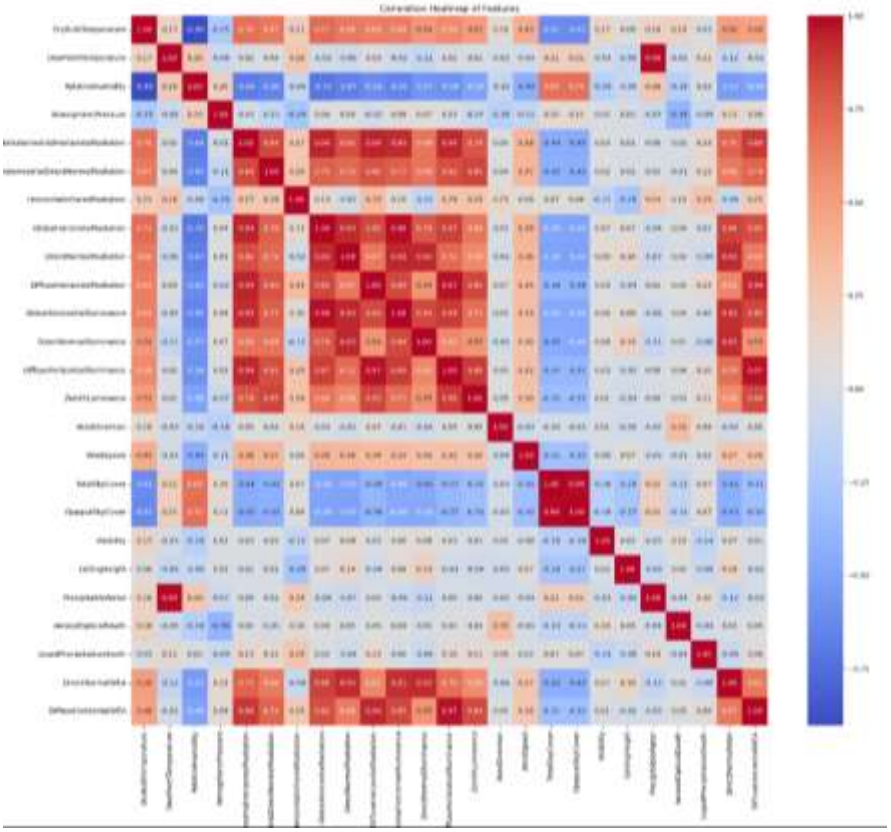


Figure 3. Data Correlation Heatmap

Copyright@2025 Jurnal Rekayasa Material, Manufaktur dan Energi. This is an open acces article under the CC-BY-SA lisence (https://creativecommons.org/licenses/by-sa/4.0/).

Figure 3 presents a correlation heatmap illustrating the Pearson correlation coefficients among 29 numerical environmental and meteorological variables in the Banda Aceh dataset. The color gradient represents the strength and direction of linear relationships, with red shades indicating positive correlations and blue indicating negative correlations. The analysis reveals several strong associations, particularly among radiation and illumination features. For instance, GHR is highly correlated with GHI (r = 0.97) and *Direct Normal Radiation* (r = 0.86), confirming their interrelated nature. Similarly, *Diffuse Horizontal Radiation* shows an almost perfect correlation with *Diffuse Horizontal Illuminance* (r = 0.97). These findings suggest that illumination variables may serve as effective proxies or complementary indicators for solar radiation in predictive modeling. Notably, *Dry Bulb Temperature* demonstrates a moderate-to-strong positive correlation with GHR (r = 0.71), whereas *Relative Humidity* exhibits a strong negative relationship with both temperature (r = –0.70) and radiation-related variables. This inverse association is consistent with the physical phenomenon in which high humidity often associated with increased cloud cover attenuates solar irradiance.

Data Pre-processing

Before training the machine learning models, the dataset underwent several pre-processing steps to ensure data quality and suitability. The Figure 4 flowchart illustrates the data preprocessing steps pipeline for machine learning models, starting with raw data from which features and targets are identified. Subsequently, missing values are checked; if found, they are handled before proceeding to feature scaling. Following this, the data is split into training and testing sets, ultimately yielding pre-processed data ready for use in machine learning models. These steps include:

1. **Feature and Target Identification:** Two main target variables were identified for prediction: GHR and GHI. All other columns in the dataset were considered predictor features (independent) [1].
2. **Missing Value Handling:** Initial analysis showed that the dataset had no missing values in any column, thus no imputation or removal of rows/columns was necessary [10,11].
3. **Feature Normalization:** Since the features had different scales and value ranges, normalization was performed using MinMaxScaler from the scikit-learn library. This normalization rescales features so that all values fall within the [0, 1] range. This process is crucial for models sensitive to feature scaling, such as Support Vector Regressor and Artificial Neural Networks, and to accelerate the convergence of optimization algorithms.
4. **Data Splitting:** The dataset was divided into two subsets: training data and testing data. This split was performed using the train_test_split function from scikit-learn with a fixed random_state to ensure reproducibility of results. The training data was used to train the models, while the testing data was used to evaluate model performance on unseen data [9,12,13].

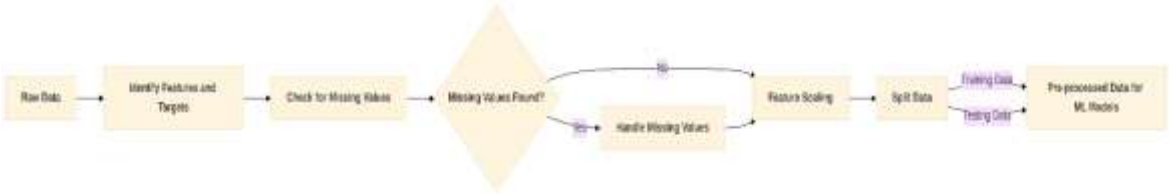


Figure 4. Data Pre-processing Diagram

Machine Learning Models

This study tested six different machine learning regression models to predict global horizontal radiation and illumination. The selection of these models was based on their popularity, proven performance in regression tasks, and the diversity of algorithmic approaches they offer.

Linear Regression: Its main advantages are easy interpretability and computational efficiency. However, its linearity assumption can limit its ability to capture more intricate relationships in environmental data. A basic statistical model that models a linear relationship between independent and dependent variables. Despite its simplicity, this model provides an important performance baseline [14–17].

Copyright@2025 Jurnal Rekayasa Material, Manufaktur dan Energi. This is an open acces article under the CC-BY-SA lisence (https://creativecommons.org/licenses/by-sa/4.0/).

Random Forest Regressor: An ensemble model based on decision trees that builds many decision trees during training and outputs the average of the predictions from each tree. It is effective in reducing overfitting and handling non-linear relationships [2,3,7,9,12,18,19].

Gradient Boosting Regressor (XGBoost): A highly efficient and flexible implementation of gradient boosting. XGBoost is known for its high speed and accuracy, as well as its ability to handle various types of data and problems [9,12,18].

Gradient Boosting Regressor (LightGBM): Another gradient boosting algorithm designed for efficiency and speed, especially on large datasets. LightGBM uses Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) techniques to accelerate training without significantly sacrificing accuracy [6,9,20].

Support Vector Regressor (SVR): An extension of Support Vector Machine for regression tasks. SVR works by finding the best-fit function for the data with a tolerable error margin, using a kernel function (in this case, Radial Basis Function or RBF) to handle non-linearity [2,18,21,22].

Artificial Neural Network (ANN) - MLPRegressor: A biologically inspired model consisting of layers of interconnected neurons. MLPRegressor from scikit-learn is a simple feedforward neural network implementation. In this study, the ANN was configured with two hidden layers (64 and 32 neurons), ReLU activation function, and Adam solver, with a maximum of 500 iterations [18,22,23].

Evaluation Metrics

To evaluate the performance of regression models, several common metrics are used: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared. MAE measures the average absolute difference between predicted and actual values, providing an indication of the average error in the same units as the target variable [2,3,7,17,18,24,25].

- **Mean Absolute Error (MAE):** Measures the average magnitude of errors between predictions and actual values.
- **Mean Squared Error (MSE):** Measures the average squared error, giving more weight to larger errors.
- **Root Mean Squared Error (RMSE):** The square root of MSE, returning the metric to the same units as the target variable.
- **R-squared (R^2):** Indicates the proportion of variance in the dependent variable that can be explained by the model. Higher values indicate a better model fit.

  MSE gives greater weight to larger errors by squaring the differences. RMSE is the square root of MSE, thus returning the metric to the same units as the target variable and being easier to interpret than MSE. R-squared (coefficient of determination) indicates the proportion of variance in the dependent variable that can be predicted from the independent variables, with values close to 1 indicating an excellent model. The interpretation and formula of evaluation metrics can be seen at Table 1.

Table 1. Interpretation and formula for evaluation metric

Evaluation Metric	Interpretation	Formula
Mean Absolute Error (MAE)	Smaller values indicate a better model.	$MAE = \frac{1}{n}\sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert$
Mean Squared Error (MSE)	Smaller values indicate a better model. Gives more weight to large errors.	$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
Root Mean Squared Error (RMSE)	Smaller values indicate a better model. In the same units as the target variable.	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$
R-squared (R^2)	Values close to 1 indicate an excellent model. A value of 0 means the model is no better than the mean.	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

Copyright@2025 Jurnal Rekayasa Material, Manufaktur dan Energi. This is an open acces article under the CC-BY-SA lisence (https://creativecommons.org/licenses/by-sa/4.0/).

All these metrics were calculated on the test dataset to ensure an unbiased evaluation of the model's generalization capabilities. The results from each model will be collected and compared to identify the best-performing model for the global horizontal radiation and illumination prediction tasks.

RESULTS AND DISCUSSION

Prediction of Global Horizontal Radiation (GHR)

From Table 2 is a summary of the result from evaluation metrics for GHR prediction. The ensemble-based models and artificial neural networks show significantly superior performance compared to Linear Regression and SVR. Specifically, the LightGBM Regressor achieved the best performance for global horizontal radiation prediction, with the lowest MAE (0.0042), lowest RMSE (0.0076), and highest R-squared (0.9993). This indicates that LightGBM is capable of explaining almost 99.93% of the variability in global horizontal radiation data, signifying an excellent model fit to the data.

XGBoost Regressor also showed very competitive performance, slightly below LightGBM, with an R-squared of 0.9990. Random Forest Regressor also performed very well with an R-squared of 0.9989. The superior performance of these boosting (XGBoost, LightGBM) and bagging (Random Forest) models can be attributed to their ability to capture complex non-linear relationships and feature interactions within the dataset. These models effectively reduce bias and variance through the combination of many decision trees.

Table 2. Evaluation metric results for GHR

Model	MAE	MSE	RMSE	R-squared
Linear Regression	0.0161	0.0006	0.0244	0.9926
Random Forest Regressor	0.0050	0.0001	0.0095	0.9989
XGBoost Regressor	0.0048	0.0001	0.0089	0.9990
LightGBM Regressor	0.0042	0.0001	0.0076	0.9993
Support Vector Regressor	0.0309	0.0015	0.0385	0.9815
ANN Regressor	0.0089	0.0002	0.0131	0.9978

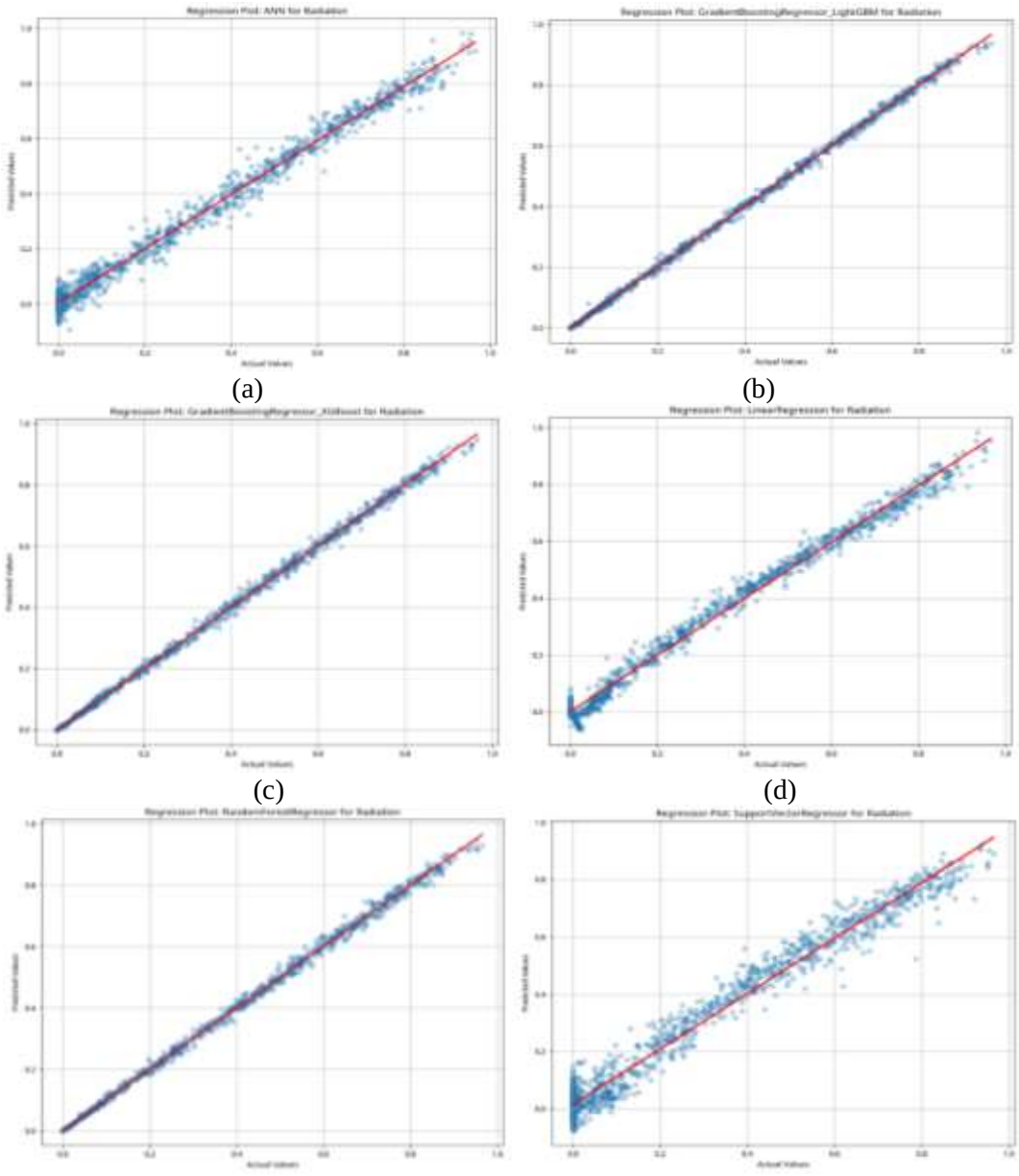
Conversely, Linear Regression, despite yielding a relatively high R-squared (0.9926), had higher MAE and RMSE compared to ensemble models and ANN. This suggests that the linearity assumption may not be entirely adequate to capture all the complexities in environmental data. The SVR showed the weakest performance among all models, with the highest MAE and RMSE and the lowest R-squared (0.9815). This relatively low performance of SVR might be due to its sensitivity to hyperparameters and higher computational complexity, which may require more extensive tuning to achieve optimal performance on this dataset.

The ANN Regressor showed very good performance, ranking fourth after the ensemble models, with an R-squared of 0.9978. This confirms the ability of artificial neural networks to model complex non-linear relationships, although in the basic configuration used in this study, its performance was slightly inferior to optimized gradient boosting models.

Based on the regression plot visualizations of six machine learning models at Figure 5, the Gradient Boosting Regressors particularly XGBoost and LightGBM demonstrated superior predictive performance. This is evident from the distribution of predicted values closely aligned along the identity line (y = x), indicating highly accurate predictions of surface radiation values. These models effectively captured the complex nonlinear relationships inherent in atmospheric radiation dynamics, making them well-suited for this type of problem. Additionally, the Random Forest Regressor showed competitive results, with predictions also clustering near the identity line, suggesting strong performance with relatively lower model complexity compared to boosting techniques. In contrast, the Linear Regression model exhibited significant limitations, with predictions deviating substantially from the identity line. This highlights its inadequacy in modeling the nonlinear and multifactorial nature of surface radiation. The SVR and ANN models yielded moderate performance. While they showed improved alignment compared to linear

Copyright@2025 Jurnal Rekayasa Material, Manufaktur dan Energi. This is an open acces article under the CC-BY-SA lisence (https://creativecommons.org/licenses/by-sa/4.0/).

regression, both exhibited noticeable deviations, especially in extreme radiation values. The suboptimal performance of ANN may be attributed to limited training data or suboptimal hyperparameter tuning. Overall, the results underscore the effectiveness of ensemble-based boosting approaches in modeling surface radiation prediction and their strong potential for implementation in data-driven forecasting systems.



(a) (b) (c) (d) (e) (f)

Figure 5. GHR Regression results: a.) ANN model; b.) LightGBM; c.) XGBoost; d.) Linear Regression; e.) Random Forest; f.) SVM model

Copyright@2025 Jurnal Rekayasa Material, Manufaktur dan Energi. This is an open acces article under the CC-BY-SA lisence (https://creativecommons.org/licenses/by-sa/4.0/).

The set of Figure 6 displays the comparison between actual and predicted values of global horizontal radiation using six different machine learning models. Each graph visualizes how closely each model's predictions align with actual values across the test samples, providing visual insights into model performance. From the visual comparison, it is evident that ensemble learning models such as LightGBM, XGBoost, and Random Forest show a remarkable ability to track the actual radiation patterns, with minimal deviations. The predicted lines in these models closely follow the actual values, indicating high prediction accuracy. The ANN model also demonstrates a strong fit, although with slightly more fluctuations than LightGBM or XGBoost. In contrast, Linear Regression and SVR exhibit more visible discrepancies, with the predicted values frequently underestimating or overestimating the actual values, suggesting weaker performance.

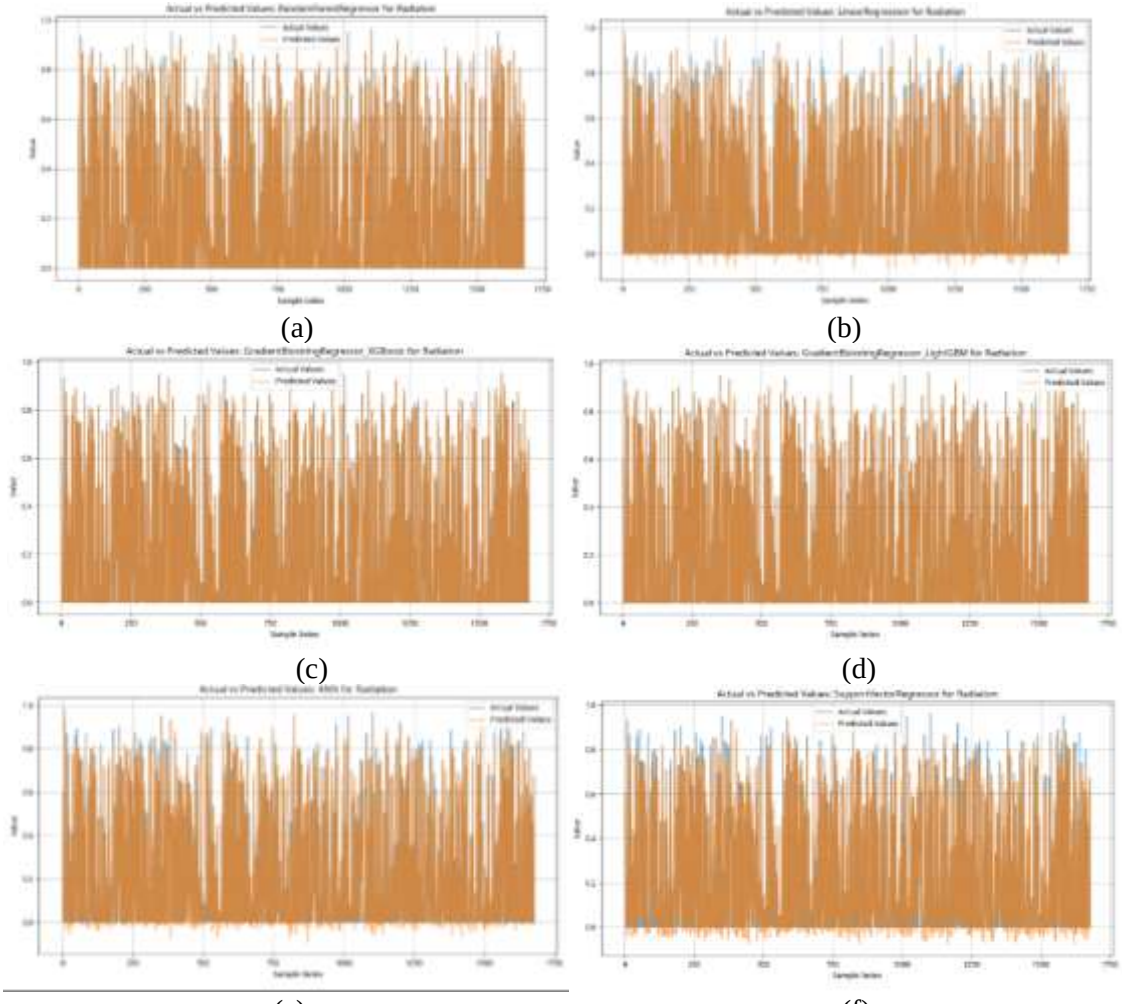


Figure 6. Comparison between actual and predicted result of GHR: a.) ANN model; b.) LightGBM; c.) XGBoost; d.) Linear Regression; e.) Random Forest; f.) SVM model

The visualizations confirm the quantitative results that ensemble models, particularly LightGBM and XGBoost, offer superior predictive capabilities for solar radiation modeling. These findings reinforce the suitability of advanced machine learning techniques for complex environmental prediction tasks, especially when dealing with non-linear and high-dimensional data such as meteorological inputs for solar energy forecasting.

Prediction of Global Horizontal Illumination (GHI)

Similar to radiation prediction, result for GHI at Table 3. the LightGBM Regressor again demonstrated the best performance for global horizontal illumination prediction, with the lowest MAE (0.0047), lowest RMSE (0.0086), and highest R-squared (0.9991). This indicates that

Copyright@2025 Jurnal Rekayasa Material, Manufaktur dan Energi. This is an open acces article under the CC-BY-SA lisence (https://creativecommons.org/licenses/by-sa/4.0/).

LightGBM is also highly effective in predicting illumination, explaining over 99.9% of the data variability.

Table 3. Evaluation metric results for GHR

Model	MAE	MSE	RMSE	R-squared
Linear Regression	0.0179	0.0007	0.0273	0.9908
Random Forest Regressor	0.0054	0.0001	0.0108	0.9986
XGBoost Regressor	0.0056	0.0001	0.0104	0.9987
LightGBM Regressor	0.0047	0.0001	0.0086	0.9991
Support Vector Regressor	0.0333	0.0016	0.0401	0.9802
ANN Regressor	0.0091	0.0002	0.0132	0.9979

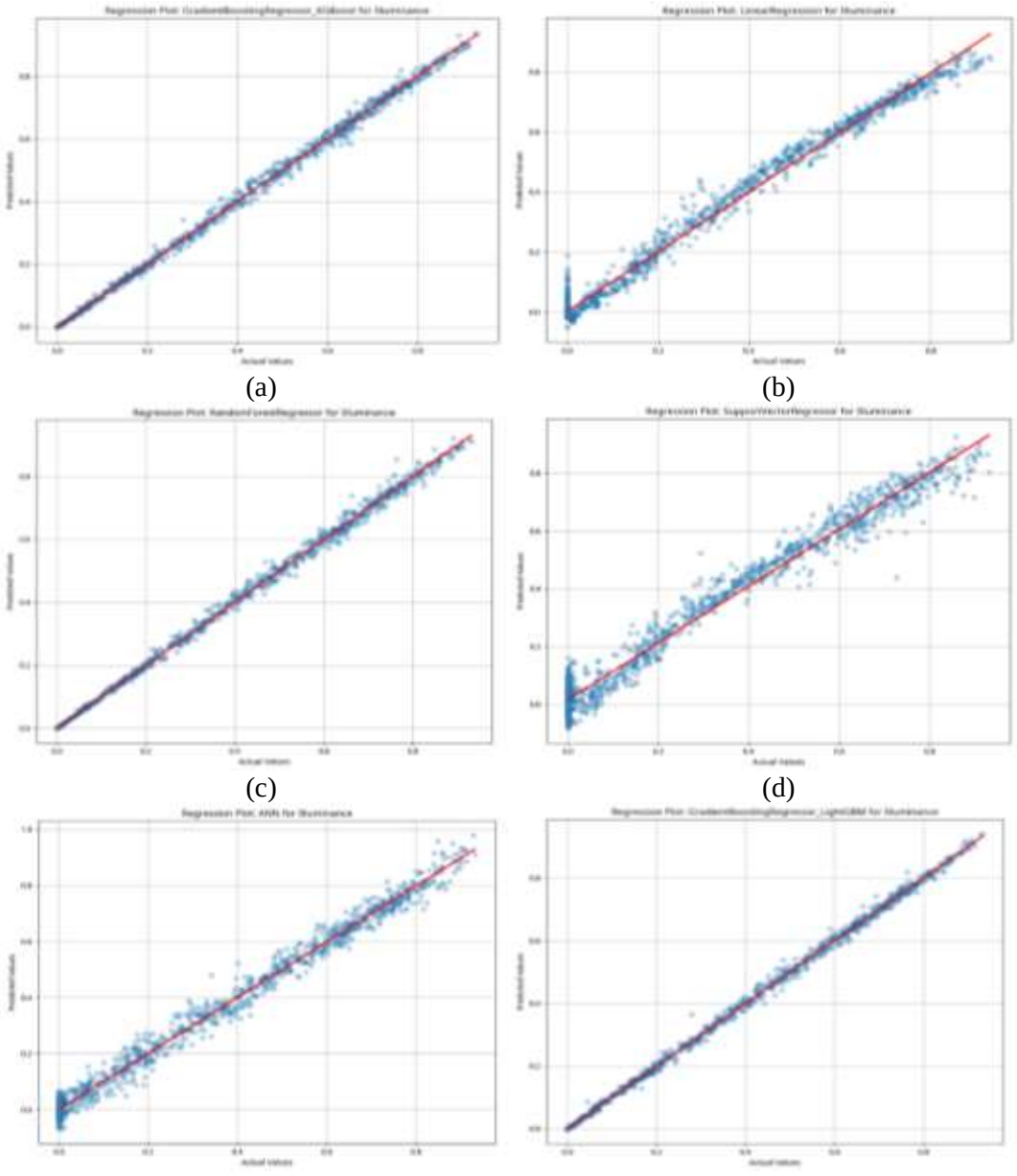
XGBoost Regressor and Random Forest Regressor also maintained very strong performance, with R-squared values of 0.9987 and 0.9986, respectively. The consistency of these ensemble models' performance for both target variables demonstrates their robust capability in handling complex environmental data. Linear Regression and SVR again showed lower performance compared to ensemble models and ANN. SVR once again performed the weakest, with an R-squared of 0.9802. This reinforces the observation that SVR may require more careful tuning or may be less suitable for the specific characteristics of this dataset compared to other models.

The ANN Regressor also showed very good performance for illumination prediction, with an R-squared of 0.9979, comparable to its performance in radiation prediction. This indicates that ANN is a solid choice for this prediction task, although slightly outperformed by LightGBM and XGBoost.

From the observation of the plots at Figure 7, a clear difference in performance among the models is evident. The Linear Regression model consistently shows the widest spread of data points, deviating significantly from the diagonal line. This indicates low accuracy and a fundamental inability of the linear model to capture the complex fluctuations and non-linear relationships inherent in the illuminance data. Its predictions tend to exhibit significant bias and high variability. Meanwhile, the Support Vector Regressor (SVR) demonstrates better performance compared to Linear Regression, with most points following the trend of the diagonal line. Conversely, ensemble-based models such as Random Forest Regressor, Gradient Boosting Regressor (XGBoost), and Gradient Boosting Regressor (LightGBM) consistently exhibit very strong performance. Their plots display data points that are remarkably dense and tightly concentrated around the line of perfect prediction. This density indicates a very high level of accuracy, precision, and consistency in illuminance prediction across the entire range of values. The ANN its regression plot showing a minimal and very tight spread of data points around the diagonal line. This highlights ANN's superior ability to learn and capture highly complex and non-linear data patterns, even including subtle fluctuations that might be missed by other models. Overall, from the visual perspective of the regression plots, more advanced and adaptive models such as ensemble methods (Random Forest, XGBoost, LightGBM) and especially Neural Networks, significantly outperform simple linear regression models in this illuminance prediction task.

Copyright@2025 Jurnal Rekayasa Material, Manufaktur dan Energi. This is an open acces article under the CC-BY-SA lisence (https://creativecommons.org/licenses/by-sa/4.0/).

Vol. 8, No.2, Juli 2025, Halaman 431-444 ISSN 2622-7398
DOI:https://doi.org/10.30596/rmme.v8i2.26011 http://jurnal.umsu.ac.id/index.php/RMME
Jurnal Rekayasa Material, Manufaktur dan Energi



(a) (b) (c) (d) (e) (f)

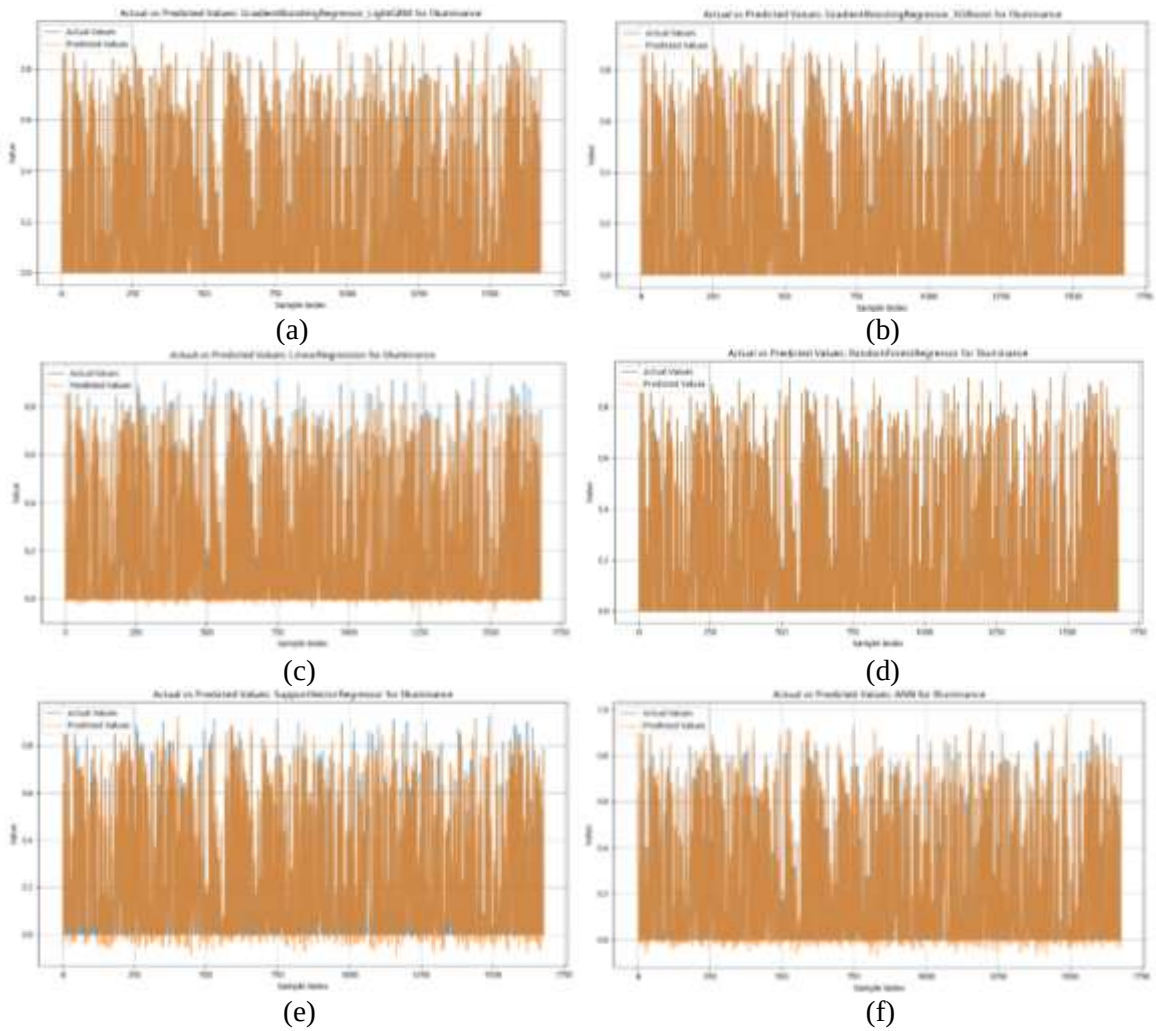
Figure 7. GHI Regression results: a.) ANN model; b.) LightGBM; c.) XGBoost; d.) Linear Regression; e.) Random Forest; f.) SVM model

From actual and predict result at Figure 8, the Linear Regression model clearly demonstrates the weakest performance, failing to capture the complex fluctuations and non-linearity of illuminance data, with predictions that tend to be overly smooth and frequently deviate significantly from actual values. Meanwhile, the Support Vector Regressor (SVR) shows moderate ability, capable of following general trends but occasionally lacking precision in capturing extreme changes.

Conversely, ensemble-based models such as Random Forest Regressor, Gradient Boosting Regressor (XGBoost), and Gradient Boosting Regressor (LightGBM) consistently exhibit strong performance. These three models are highly effective in handling non-linear data patterns and complex interactions, producing relatively accurate predictions that closely follow the variations in illuminance data. Among these models, LightGBM and XGBoost often stand out due to their balance of high accuracy and computational efficiency.

Copyright@2025 Jurnal Rekayasa Material, Manufaktur dan Energi. This is an open acces article under the CC-BY-SA lisence (https://creativecommons.org/licenses/by-sa/4.0/).

Vol. 8, No.2, Juli 2025, Halaman 431-444 ISSN 2622-7398
DOI:https://doi.org/10.30596/rmme.v8i2.26011 http://jurnal.umsu.ac.id/index.php/RMME
Jurnal Rekayasa Material, Manufaktur dan Energi



(a) (b) (c) (d) (e) (f)

Figure 8. Comparison between actual and predicted result of GHI: a.) ANN model; b.) LightGBM; c.) XGBoost; d.) Linear Regression; e.) Random Forest; f.) SVM model

However, ANN visually displays the best performance, demonstrating a superior ability to capture highly complex and non-linear data patterns, including subtle fluctuations. ANN's predictions appear most aligned with actual values, showcasing remarkable adaptive learning capacity. Overall, for the task of illuminance prediction with complex data, advanced models like Random Forest, XGBoost, LightGBM, and especially ANN, are significantly superior and recommended compared to simpler linear approaches.

Overall, the results of this study consistently demonstrate that ensemble-based models such as LightGBM, XGBoost, and Random Forest significantly outperform traditional models like Linear Regression and SVR in predicting global horizontal radiation and illumination. LightGBM, in particular, stands out as the best performing model for both target variables, followed by XGBoost and Random Forest. The superiority of these models likely stems from their ability to handle non-linear relationships, feature interactions, and the inherent complexity of environmental data.

Factors such as data normalization and proper data splitting also played crucial roles in achieving high model performance. Normalization ensures that all features contribute fairly to the model training process, preventing features with larger value scales from dominating. Splitting data into separate training and testing sets ensures that model performance evaluation reflects its generalization capability on unseen data. It should be noted that although SVR showed relatively low performance in this study, this does not mean that SVR is unsuitable for radiation/illumination

Copyright@2025 Jurnal Rekayasa Material, Manufaktur dan Energi. This is an open acces article under the CC-BY-SA lisence (https://creativecommons.org/licenses/by-sa/4.0/).

prediction tasks in general. SVR performance is highly dependent on careful kernel selection and hyperparameter tuning.

CONCLUSION AND SUGGESTIONS

This study successfully explored and compared the performance of various machine learning models in predicting global horizontal radiation and global horizontal illumination using a comprehensive dataset consisting of 29 environmental features. The evaluation results indicate that ensemble-based models, particularly the LightGBM Regressor, consistently provided the best predictive performance for both target variables, followed by the XGBoost Regressor and Random Forest Regressor. These models achieved very high R-squared values (approaching 0.999), demonstrating an exceptional ability to explain data variability and predict target values with high accuracy.

ANN also showed very good performance, confirming their potential in handling complex non-linear relationships in environmental data. On the other hand, Linear Regression, while providing reasonably good results, showed limitations in capturing all data complexities. The SVR showed the weakest performance among the tested models, which may indicate the need for more extensive hyperparameter tuning or that this model is less suitable for the specific characteristics of this dataset.

The implications of these findings are highly significant. The high prediction accuracy for global horizontal radiation and illumination can support various practical applications, including:

- Optimization of Renewable Energy Systems: More accurate predictions enable more efficient planning and operation of solar panel systems, optimizing energy production and reducing operational costs.
- Smart Agriculture: Farmers can leverage these predictions to optimize irrigation schedules, crop management, and protection against extreme weather conditions, ultimately increasing yields.
- Urban Planning and Building Design: Architects and urban planners can use predicted illumination data to design more energy-efficient buildings with optimal natural lighting, as well as create more comfortable and sustainable urban environments.

REFERENCES

[1] Buster G, Bannister M, Habte A, Hettinger D, Maclaurin G, Rossol M, Sengupta M and Xie Y 2022 Physics-guided machine learning for improved accuracy of the National Solar Radiation Database *Sol. Energy* **232** 483–92

[2] Imam A A, Abusorrah A, Seedahmed M M A and Marzband M 2024 Accurate Forecasting of Global Horizontal Irradiance in Saudi Arabia: A Comparative Study of Machine Learning Predictive Models and Feature Selection Techniques *Mathematics* **12**

[3] Wu L, Chen T, Ciren N, Wang D, Meng H, Li M, Zhao W, Luo J, Hu X, Jia S, Liao L, Pan Y and Wang Y 2023 Development of a Machine Learning Forecast Model for Global Horizontal Irradiation Adapted to Tibet Based on Visible All-Sky Imaging *Remote Sens.* **15**

[4] Yesilyurt H, Dokuz Y and Dokuz A S 2024 Data-driven energy consumption prediction of a university office building using machine learning algorithms *Energy* **310** 133242

[5] Zhou Y, Liu Y, Wang D, Liu X and Wang Y 2021 A review on global solar radiation prediction with machine learning models in a comprehensive perspective *Energy Convers. Manag.* **235** 113960

[6] Park J, Moon J, Jung S and Hwang E 2020 Multistep-ahead solar radiation forecasting scheme based on the light gradient boosting machine: A case study of Jeju Island *Remote Sens.* **12**

[7] Solano E S and Affonso C M 2023 Solar Irradiation Forecasting Using Ensemble Voting Based on Machine Learning Algorithms *Sustain.* **15**

[8] Nematchoua M K, Orosa J A and Afaifia M 2022 Prediction of daily global solar radiation and air temperature using six machine learning algorithms; a case of 27 European countries *Ecol. Inform.* **69**

Copyright@2025 Jurnal Rekayasa Material, Manufaktur dan Energi. This is an open acces article under the CC-BY-SA lisence (https://creativecommons.org/licenses/by-sa/4.0/).

[9] Zhou Z, Lin A, He L and Wang L 2022 Evaluation of Various Tree-Based Ensemble Models for Estimating Solar Energy Resource Potential in Different Climatic Zones of China *Energies* **15**
[10] Zhao R C, Wei D, Ran Y B, Zhou G, Jia Y C, Zhu S L and He Y Q 2022 Building Cooling load prediction based on LightGBM *IFAC-PapersOnLine* **55** 114–9
[11] Narvaez G, Giraldo L F, Bressan M and Pantoja A 2021 Machine learning for site-adaptation and solar radiation forecasting *Renew. Energy* **167** 333–42
[12] Demir V 2025 Evaluation of Solar Radiation Prediction Models Using AI: A Performance Comparison in the High-Potential Region of Konya, Türkiye *Atmosphere (Basel).* **16**
[13] Dahmani A, Ammi Y and Hanini S 2024 A Novel Non-Linear Model Based on Bootstrapped Aggregated Support Vector Machine for the Prediction of Hourly Global Solar Radiation *Smart Grids Sustain. Energy* **9** 1–13
[14] Goyal M, Pandey M and Thakur R 2020 Exploratory Analysis of Machine Learning Techniques to predict Energy Efficiency in Buildings *ICRITO 2020 - IEEE 8th Int. Conf. Reliab. Infocom Technol. Optim. (Trends Futur. Dir.* 1033–7
[15] Fan C, Ding Y and Liao Y 2019 Analysis of hourly cooling load prediction accuracy with data-mining approaches on different training time scales *Sustain. Cities Soc.* **51** 101717
[16] Faddel S, Tian G, Zhou Q and Aburub H 2020 On the Performance of Data-Driven Reinforcement Learning for Commercial HVAC Control *2020 IEEE Ind. Appl. Soc. Annu. Meet. IAS 2020*
[17] Yu H, Jiang S, Chen M, Wang M, Shi R, Li S, Wu J, Kui X, Zou H and Zhan C 2024 Machine learning models for daily net radiation prediction across different climatic zones of China *Sci. Rep.* **14** 1–16
[18] Ledmaoui Y, El Maghraoui A, El Aroussi M, Saadane R, Chebak A and Chehri A 2023 Forecasting solar energy production: A comparative study of machine learning algorithms *Energy Reports* **10** 1004–12
[19] Voyant C, Notton G, Kalogirou S, Nivet M L, Paoli C, Motte F and Fouilloy A 2017 Machine learning methods for solar radiation forecasting: A review *Renew. Energy* **105** 569–82
[20] Sevas M S, Sharmin N, Santona C F T and Sagor S R 2024 *Advanced ensemble machine-learning and explainable ai with hybridized clustering for solar irradiation prediction in Bangladesh* vol 155
[21] Saxena N, Kumar R, Rao Y K S S, Mondloe D S, Dhapekar N K, Sharma A and Yadav A S 2024 Hybrid KNN-SVM machine learning approach for solar power forecasting *Environ. Challenges* **14** 100838
[22] Ağbulut Ü, Gürel A E and Biçen Y 2021 Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison *Renew. Sustain. Energy Rev.* **135**
[23] Hemavathi U, Medona A C V, Dhilip Kumar V and Raja Sekar R 2021 Review for the Solar Radiation Forecasting Methods Based on Machine Learning Approaches *J. Phys. Conf. Ser.* **1964**
[24] Opoku R, Mensah G, Adjei E A, Bosco Dramani J, Kornyo O, Nijjhar R, Addai M, Marfo D, Davis F and Obeng G Y 2023 Machine learning of redundant energy of a solar PV Mini-grid system for cooking applications *Sol. Energy* **262** 111790
[25] Nwokolo S C, Obiwulu A U and Ogbulezie J C 2023 Machine learning and analytical model hybridization to assess the impact of climate change on solar PV energy production *Phys. Chem. Earth* **130** 103389

Copyright@2025 Jurnal Rekayasa Material, Manufaktur dan Energi. This is an open acces article under the CC-BY-SA lisence (https://creativecommons.org/licenses/by-sa/4.0/).