

Short-term solar power forecasting via extreme gradient boosting (XGBoost): experimental validation using real-time data from the Heipang hybrid solar–wind test rig

Ishaku Borok Maren *, Joseph Ajiya Japhet and Bildad Dakwal Lengs

Department of Mechanical Engineering, University of Jos, Plateau State, Nigeria.

World Journal of Advanced Engineering Technology and Sciences, 2026, 20(01), 098–112

Publication history: Received on 25 May 2026; revised on 04 July 2026; accepted on 08 July 2026

Article DOI: <https://doi.org/10.30574/wjaets.2026.20.1.0347>

Abstract

Accurate short-term solar power forecasting is essential for ensuring efficient operation, stability, and intelligent energy management in hybrid renewable energy systems. This study presents an Extreme Gradient Boosting (XGBoost)-based framework for short-term solar power forecasting using real-time experimental data acquired from the Heipang Hybrid Solar–Wind Test Rig, located in Heipang community, Barkin Ladi Local Government Area of Plateau State, Nigeria. The proposed model utilizes meteorological and operational variables, including solar irradiance, ambient temperature, relative humidity, wind speed, and historical photovoltaic power output, to predict short-term solar energy generation under dynamic environmental conditions. Data preprocessing techniques such as normalization, outlier filtering, and feature engineering were employed to improve model robustness and predictive capability. The XGBoost algorithm was optimized through hyperparameter tuning to effectively capture nonlinear relationships and complex interactions among weather-dependent variables influencing photovoltaic performance. Model evaluation was conducted using standard statistical performance metrics, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and coefficient of determination (R^2). Experimental results demonstrated that the developed XGBoost model achieved high forecasting accuracy, fast computational performance, and strong generalization capability under fluctuating meteorological conditions. The findings indicate that gradient boosting techniques provide a reliable and scalable solution for real-time solar power forecasting and smart grid applications. This study contributes to the advancement of data-driven renewable energy forecasting approaches for sustainable hybrid energy system optimization and improved power dispatch planning in developing regions.

Keywords: Short-term solar power forecasting; Extreme Gradient Boosting; Machine learning; Renewable energy forecasting; Hybrid solar–wind system; Photovoltaic power prediction

1. Introduction

The negative impact of climate change on the globe due to the increase in the use of fossil fuels for energy generation necessitated the paradigm shift to clean energy generation systems [1] such as solar and wind. Many countries have and some will develop policies for individual homes and investors to integrate their generated solar energy to the grid and other developing countries such as Nigeria may soon key into that. But solar energy generating systems depend on weather factors such as temperature, humidity and solar irradiance etc., which are unstable thereby making the energy generation unpredictable. Feeding the varying solar energy generated to the grid will create serious problems since the grid is not designed to integrate large amount of varying solar power into it [2] also, as stated by [3] the randomness and uncertainty of the weather parameters that solar power depend upon could affect the stability of the power grid, when they are connected to the grid. In order to effectively control the excess power fed to the grid, utility will need to estimate the solar power to be added to the grid. If the solar power output connected to the grid can be predicted at any

* Corresponding author: Maren, I. B

time, this will enable the utility maximize the use of the resources and maintain the stability of the grid. As stated by [4] that solar power prediction is very significant for efficient planning and management of the resources. The stability of grid can be achieved using power prediction techniques which helps utilities design controllers for a hybrid power plant [5]. Accurate solar performance prediction is essential for optimizing power production and reducing maintenance costs [6].

Machine learning approaches have been increasingly used for the prediction of solar power output from a generation plant. Machine learning approach discovers the connections between input data and output data through some training processes that result to a machine-learning model. The forecasting input data can be fed into the well-trained machine learning model for predictions [7]. Thus, in this study, a prediction models for the solar power outputs for Heipang community were developed. XGBoost, a scalable tree-based ensemble technique is being used in developing complex and nonlinear correlation between meteorological data like wind speed, ambient temperature, solar radiation and relative humidity, and solar power output [8]. It has been popular for its precision, resilience, and speed in structured prediction tasks [9].

In this study, real-time data obtained from a hybrid solar-wind power test rig installed at Heipang Airport in Bakin-Ladi, Plateau State, Nigeria, was used to evaluate the performances of XGBoost model for short-term solar power forecasting. The geographic area of Heipang community gives a distinctive highland microclimate with diverse irradiance patterns, makes it a good site for investigating solar energy forecasting. A designed and calibrated multi-channeled datalogger with sensors integrated to the hybrid solar-wind power test rig records the power output and meteorological characteristics at precise time resolutions, thereby making it possible to create and assess data-driven forecasting models that are adapted to the local climatic conditions.

This research is motivated by the tremendous increase in the use of solar photovoltaic (PV) systems for off-grid and grid-tied hybrid power as there is paradigm shift in the world's energy sector to clean and sustainable sources, especially in developing nations like Nigeria [10] Fluctuations of solar power generation dependent variables pose a challenge in the efficient distribution of power, energy storage management and system stability within the microgrids system. Also, as reported in [11], traditional statistical models such as persistence and ARIMA do not have the capacity to model the nonlinear and dynamic patterns in solar power outputs, particularly in areas with extremely changeable weather and little historical data. Even though [12], confirmed the proficiency of machine learning and deep learning model such as XGBoost in predicting solar power generation based on meteorological variables, not many researchers have examined the accuracy of these models by using real-world data collected from hybrid power plants in elevated regions of sub-Saharan Africa, including Heipang. The specific objectives of this research are to compile and preprocess real-time meteorological and solar power output data from the Heipang hybrid solar-wind power test rig. This includes variables such as temperature, wind speed, humidity, solar irradiance, and the solar PV system's voltage and current output; To design and implement XGBoost model for short-term solar power forecasts using the gathered multi-feature time-series data and MATLAB software; To evaluate the predictive performance of the XGBoost model using standard performance metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2); To analyze the computational efficiency, sensitivity and robustness the model under different weather scenarios, including clear sky, partially cloudy and overcast conditions common in the Heipang highland region. This research is significant as it addresses the gap in the development of XGBoost model for short-term solar power forecasting using real-time weather and power data from the hybrid solar-wind power test rig at Heipang. The use of local test rig in this study motivates contextualization and localization of machine learning models, which is essential for sub-Saharan African policy planning, microgrid optimization, and rural electrification projects [13]. Finally, researchers, utility managers, and designers of energy systems may use this model which may help and encourage the development of smart energy management systems and real-time forecasting dashboards for off-grid and grid-connected solutions in Nigeria and elsewhere, thereby advancing human capital development in smart and artificial intelligence technology.

2. Literature review

Short-term forecasting of solar power generation with accuracy is now greatly better with the application of machine learning and deep learning techniques into renewable energy systems. This is especially significant for hybrid energy systems, such as the Heipang solar-wind power test rig, where energy dispatch and system stability may be affected by real-time weather variability [14]. A potent ensemble learning method, XGBoost, combines gradient boosting with decision tree-based models to give effective forecasting. Its great speed, accuracy, and adaptability have made it a popular choice for solar power prediction [15]. [16] achieved low error rates even in complicated datasets by using XGBoost optimized with Particle Swarm Optimization (PSO) to estimate short-term PV production. In a comparative analysis, Nguyen et al. (2025) demonstrated that XGBoost outperformed traditional regression models and shallow

neural networks in both hourly and sub-hourly forecasts. Additionally, [18] applied a fusion of XGBoost with LSTM and CNN in a hybrid solar-wind setup, reporting improved accuracy and model robustness under partly cloudy and overcast conditions. These findings emphasize XGBoost's suitability for structured data and its strong generalization performance in multi-feature environments. For the case of hybrid prediction models, recent studies increasingly favor hybrid and ensemble models to exploit the strengths of different algorithms. [19] conducted a comprehensive review of deep learning models and highlighted the advantages of combining LSTM with tree-based methods like Random Forest and XGBoost. Similarly, [20] introduced a CNN-LSTM-XGBoost model that achieved lower forecasting error under noisy and weather-variant datasets. Despite significant advancements in machine learning and deep learning models for renewable energy forecasting, several critical research gaps remain, particularly in the context of short-term solar power prediction using real-time data from hybrid solar-wind systems test rig in developing regions like Heipang, Nigeria. There is limited use of localized, real-time data in Sub-Saharan Africa as most solar forecasting models are developed using datasets from Europe, North America, or Asia, where weather stations and solar irradiance monitoring are well established [21]. However, models trained in temperate or subtropical regions may perform poorly in high-altitude, tropical climates like Heipang, which experience rapid changes in solar irradiance, wind speed, and cloud dynamics. There is also a lack of forecasting studies using sensor-based real-time data from hybrid solar-wind test rigs in sub-Saharan Africa, limiting the application of global models in local settings [22]. There is an underexplored model performance under varying sky conditions. Most forecasting models report average performance over a long-time horizon but rarely account for performance variation under specific weather scenarios, such as clear-sky, partly cloudy, and overcast conditions [23]. Few studies assess the sensitivity and robustness of forecasting models across different sky conditions, which is vital for practical deployment in hybrid systems [24].

3. Materials and methods

3.1. Data Collection

3.1.1. Design and installation of a hybrid solar-wind power system test rig

A hybrid solar-wind power system test rig was designed and installed on the premises of the Nigerian Meteorological Agency (NIMET), within Yakubu Gowon Airport, Heipang, Plateau State, Nigeria, as shown in Fig. 1. This location's high altitude of 1,200 m above sea level [25], variable irradiance and wind patterns offer an ideal testing ground for model validation. The system comprises a 300W monocrystalline photovoltaic panel with a charge controller, a 300W wind turbine rectifier, and 4 units of 12V, 100Ah battery capacity configured with 2 parallel and 2 in series connections to give 24V, 200Ah battery system. A load of 400 W was connected to the battery. Integrated into the test rig is an Encoder Controlled Multichannel (ECM) datalogger, which was designed, developed, calibrated and validated using standard instruments to record the ambient temperature, solar panel temperature, relative humidity, atmospheric pressure, solar panel current and voltage, and wind turbine current and voltage, whereas the wind cup anemometer records wind speed and the pyranometer records solar radiation.

Item C shown in Fig. 1 is the pyranometer, which is a standard instrument bought to measure the solar radiation falling on the solar panel. As indicated in the operational manual [26], the pyranometer (or solar power meter model SPM 1116SD) has a spectral response between 400 and 1100 nm and a solar power range between 0 W/m² and 2000 W/m² with a resolution of 0.1 W/m² and power accuracy of $\pm 5\%$ reading. It has an SD card capacity range of 1 GB to 16 GB and is powered by UM3/AA (1.5 V) x 6 batteries or a DC 9 V adapter. The pyranometer stores the logged data in the SD card in Excel format with file names SPM01001.XLS and SPM01002.XLS, which can be downloaded from the SD card and saved in the computer for analysis. It has a sampling time setting range of 1 to 3600 seconds.

Similarly, the wind cup anemometer (model AM-4257SD), indicated in Fig. 1 as item B, was purchased to measure the wind speed at the standard meteorological wind speed measurement height of 10 m [27] in the location of the installation of the designed hybrid solar-wind power system. From the operational manual of the cup anemometer [28], it has a measurement range of 0 to 35 m/s with a resolution of 0.1 m/s and accuracy of $\pm (2\%+0.2 \text{ m/s})$. The pyranometer and the wind cup anemometer were both synchronized to log their respective data the same time with the designed and constructed ECM datalogger at 1minute time interval. The ECM datalogger also has embedded in it a control system by means of relays and programme codes to engage and disengage the load from the system.

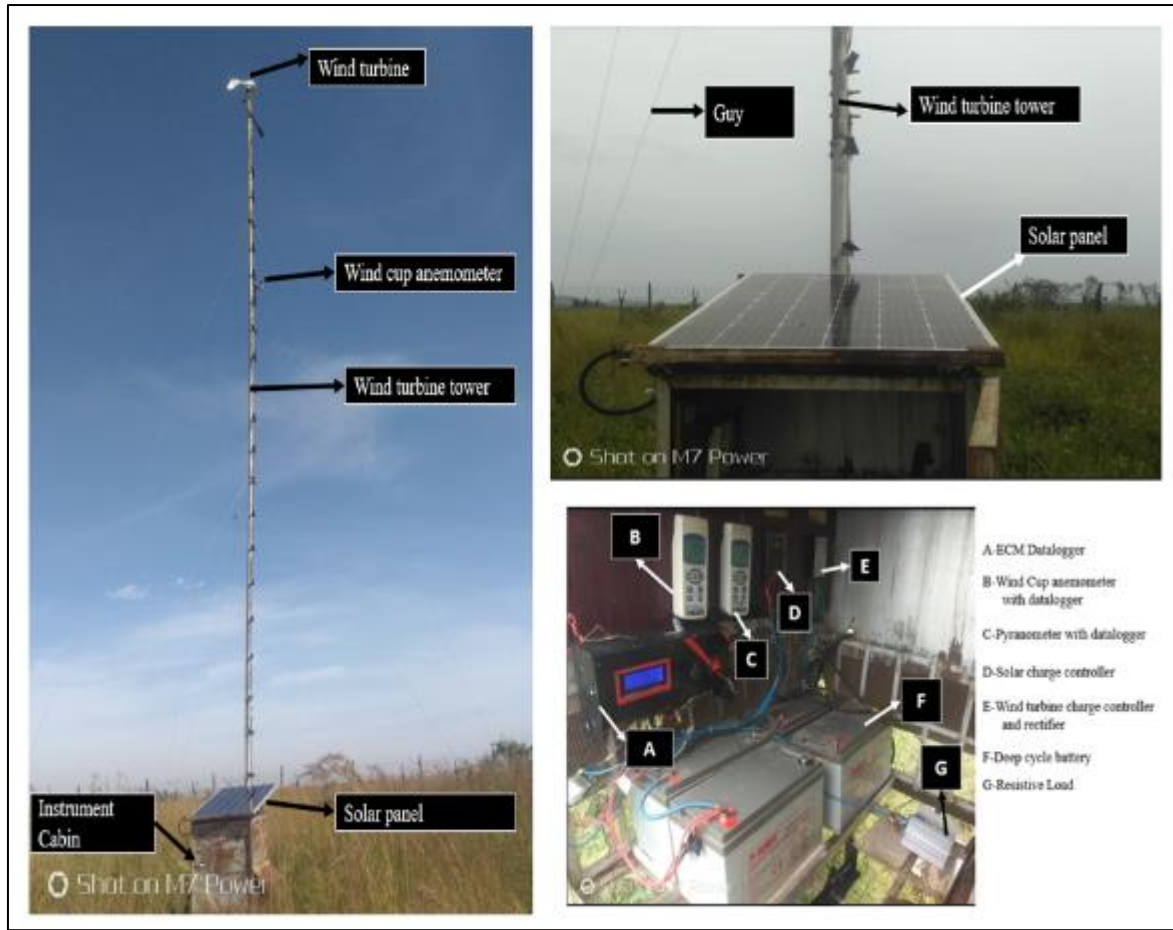


Figure 1 Installed Hybrid Solar-Wind Power System Test Rig

3.1.2. Measured variables

The data from the hybrid solar-wind power system test rig were continuously recorded at 1-minute intervals including solar irradiance (W/m^2), module temperature ($^{\circ}\text{C}$), ambient temperature ($^{\circ}\text{C}$), wind speed (m/s), Relative humidity (%), PV system current (A), voltage (V), and the solar power output determined as the product of the PV system current and voltage. All sensors were calibrated per manufacturer specifications and validated; data were stored with time stamps and exported as CSV files for analysis.

3.2. Data Preprocessing

The data collected from the system were preprocessed. Missing values ($< 1\%$ of the dataset) were handled using linear interpolation for short gaps (< 10 min) and seasonal mean substitution for larger gaps, following best practices in time-series handling [29]. The interquartile range (IQR) method was used to detect and handle outliers in the dataset adopted from [30], given as Eq. 1.

$$IQR = Q_3 - Q_1 \quad (1)$$

Where, Q_1 (25th percentile): 25% of the data falls below this value, Q_3 (75th percentile): 75% of the data falls below this value. Values below the lower bound or above the upper bound are treated as outliers [30] as given by Eq. 2 and 3.

$$\text{Lower Bound} = Q_1 - 1.5 \times IQR \quad (2)$$

$$\text{Upper Bound} = Q_3 + 1.5 \times IQR \quad (3)$$

All features were standardized (zero mean, unit variance) to aid model convergence, per guidelines in LSTM applications [29].

3.3. XGBoost Architecture

XGBoosts is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. The XGBoosts architecture is shown in Fig. 2. It provides a parallel tree boosting that solve many data science problems in fast and accurate way. The XGBoost model predicts by combining the outputs of many decision trees trained sequentially, where each new tree corrects the residual errors of the previous ones using gradient and hessian information, while regularization terms control the complexity of the model to improve generalization [31]. The Mathematical representations of XGBoosts as adopted from [32] are as follows:

Overall prediction function:

$$\hat{y}_i = \sum_{t=1}^M f_t(x_i) \quad (4)$$

Where, $\hat{y}_i$ = predicted output for sample i ; M = total number of trees; f_t = decision tree added at iteration t .

The Objective Function:

XGBoost minimizes a regularized objective at each boosting step:

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (5)$$

Where, x_i feature vector of sample i ; y_i true label of sample i ; $l()$: differentiable loss function (e.g. squared error, logistic loss); $\Omega(f_t)$: regularization term that penalizes complex trees.

Tree Structure:

Each f_t is a regression tree:

$$f_t(x) = w_j q(x_i) \quad (6)$$

Where $q(x)$ index of the leaf where x_i falls; w_j weight/score of leaf j ; t number of leaves in the tree

Gradient and Hessian (Optimization):

To find the best tree, XGBoost uses a second-order Taylor expansion of the loss:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (7)$$

$g_i = \frac{\partial l}{\partial \hat{y}_i^{(t-1)}}$ is the first-order gradient (error slope); $h_i = \frac{\partial^2 l}{\partial (\hat{y}_i^{(t-1)})^2}$ is the second-order gradient (curvature).

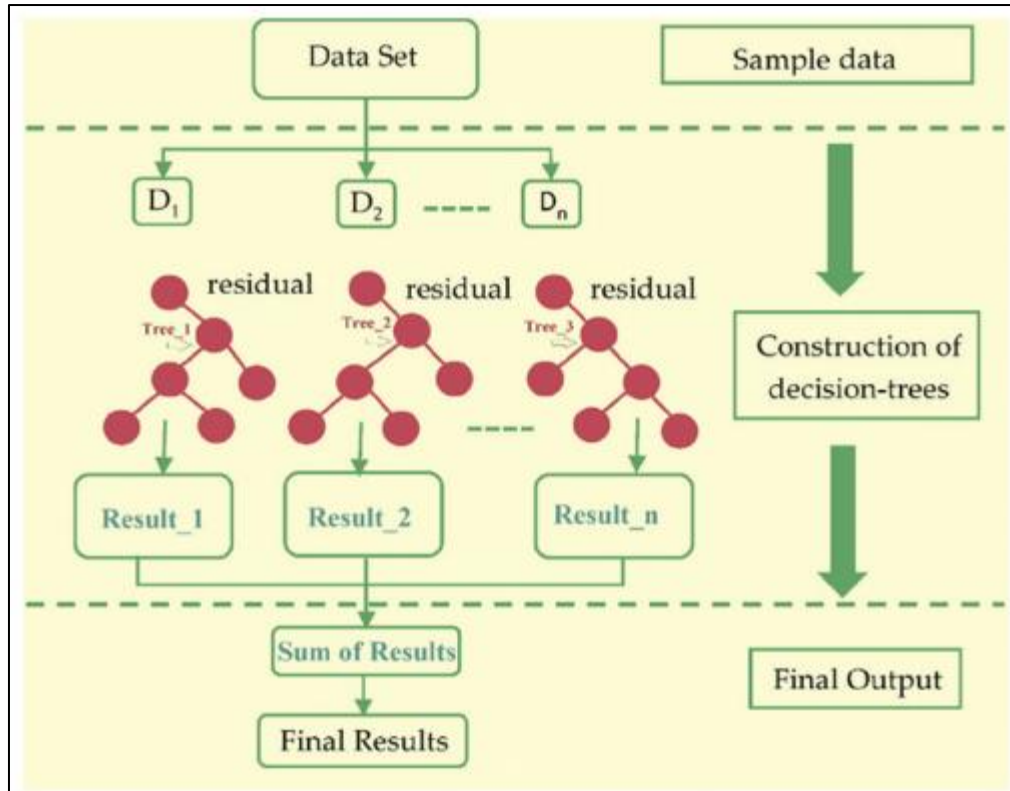


Figure 2 General Architecture of XGBoots Algorithm [32]

These values guide how the tree splits data and assigns weights.

Regularization:

To prevent over-fitting, XGBoost penalizes complex trees:

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (8)$$

Where, T = number of leaves; w_j = weight (score) of leaf j ; γ = penalty for adding a leaf; λ = L2 regularization on leaf weights

3.4. Model Training & Evaluation

3.4.1. Dataset partitioning

The dataset was split into 70% for training, 15% for validation, 15% for testing (chronological split) [33] to ensure distribution of weather scenario labels across splits.

3.4.2. Forecast horizon

Accurate 1-minute (ultra-short-term) solar power forecasting is essential for real-time grid management and maximizing the integration of photovoltaic (PV) energy. The rapid variability of solar output, mainly due to cloud movement, makes this a technically challenging but increasingly important area of research. 1-minute forecasts are primarily used for real-time microgrid control, dispatch, and maximizing PV hosting capacity in off-grid and grid-connected systems[34]. Models were trained for 1-minute ahead predictions.

3.4.3. Performance metrics

The performance of the prediction models was evaluated using MAE (Mean Absolute Error), RMSE (Root Mean Square Error), MAPE (Mean Absolute Percentage Error) and R^2 (Coefficient of Determination). Metrics computed for both global dataset and weather-conditioned subsets (clear, partly cloudy, overcast) [35].

3.4.4. Sensitivity and robustness analysis of forecasting models under variable weather scenarios

The performance of solar power forecasting models can vary significantly depending on prevailing weather conditions. In the highland region of Heipang, Plateau State, rapid fluctuations in cloud cover, irradiance, and wind speed introduce nonlinear and stochastic variability in solar PV output. Therefore, assessing the sensitivity (measure of the responsiveness of a model to changing inputs) and robustness (measure of the consistency of the accuracy of the model across varying conditions) of machine learning models such as XGBoost is critical for real-world deployment in hybrid systems [36]. Perturbation Sensitivity Analysis (PSA) involves slightly disturbing (perturbing) each input variable (e.g., solar irradiance, temperature, wind speed, humidity, etc.) and observing how much the model output (predicted solar power) changes. If small changes in an input variable cause large fluctuations in the prediction, it means the model is sensitive to that variable. However, if the prediction barely changes, then the model is robust with respect to that variable.

3.4.5. Classification of sky conditions using random forest

In order to assess the performance of the prediction models under different sky condition, the study stratified the solar dataset into three distinct meteorological categories: (i) Clear-Sky (ii) Partly Cloudy, and (iii) Overcast, using Random Forest (RF) classifier. The classifier distinguished between these conditions based on Global Horizontal Irradiance (GHI), power output, and time of day features.

4. Results and discussion

4.1. Extreme Gradient Boosting (XGBoost) Model Prediction Visualization

The XGBoost prediction model for solar power output is shown in Fig. 3 with predicted curve in red color closely following the actual curve in blue throughout the time series. The daily solar power generation trends, peak irradiance periods and transient fluctuations are adequately captured by the model with minimum deviations confirming that the ensemble boosting model performed well. The performance results of the XGBoost model shows that the training dataset gives the metrics RMSE = 6.09 W, MAE = 2.91 W, and $R^2 = 0.99$, and the testing dataset gives RMSE = 8.10 W, MAE = 4.05 W, and $R^2 = 0.98$. The results showed the highest prediction performance among all tested models, with very low errors and very high accuracy. The XGBoost model maintained strong prediction stability even during fluctuating irradiance conditions and peak solar generation periods as illustrated by the testing plot. The plots show very small deviations between the actual and predicted outputs, indicating excellent generalization capability and resistance to overfitting.

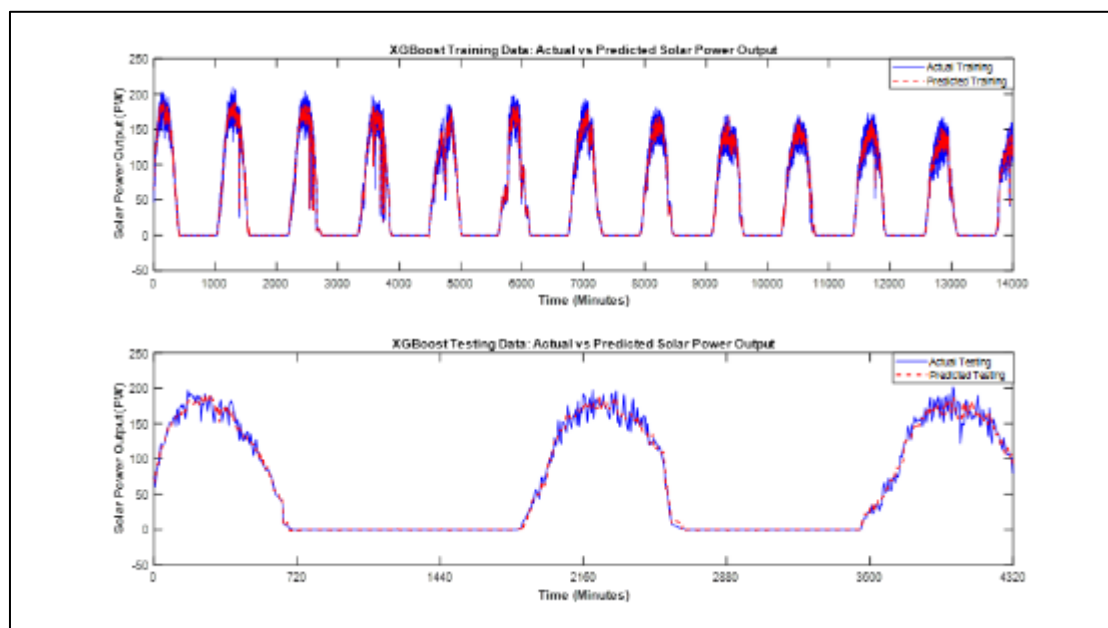


Figure 3 XGBOOST model plot of actual and predicted solar power output for training and testing data

XGBoost prediction model gives good results because of its advanced boosting method, which combines many weak models step by step to reduce errors and improves forecasting accuracy. The model also has regularization mechanisms

and optimized tree learning strategies that enhance predictive robustness while reducing model complexity. According to [37], XGBoost works efficiently and at scale because it controls overfitting and speeds up training using parallel processing.

The low MAE and RMSE values obtained in this study shows that the model effectively minimized both average and large prediction errors. Furthermore, the high R^2 values show that the XGBoost model explained almost all changes in the solar power data. Similar studies have reported that ensemble boosting algorithms such as XGBoost outperform many conventional machine learning and deep learning approaches in photovoltaic forecasting applications because of their ability to model nonlinear feature interactions and noisy environmental data efficiently [38]. The plots and performance metrics demonstrate that the XGBoost model provided high forecasting accuracy and strong generalization capability in model prediction. This suggests that the model is highly suitable for practical renewable energy forecasting applications, including smart grid operation, photovoltaic energy optimization, and intelligent energy scheduling systems.

The XGBoost model perturbation sensitivity analysis is shown in Fig. 4. The XGBoost model relies overwhelmingly on the Global Horizontal Irradiance (GHI) with 98.15% influence serving as the dominant factor driving the predictions, while other weather variables (temperature, humidity, air pressure) play only supporting roles. Ambient temperature (T) has 1.14% influence with very minor influence. Atmospheric Pressure (AP) has 0.41% influence with negligible contribution. Relative humidity (RH) has 0.30% influence with the least impact. This agrees with solar energy principles, since solar radiation directly controls photovoltaic output [39].

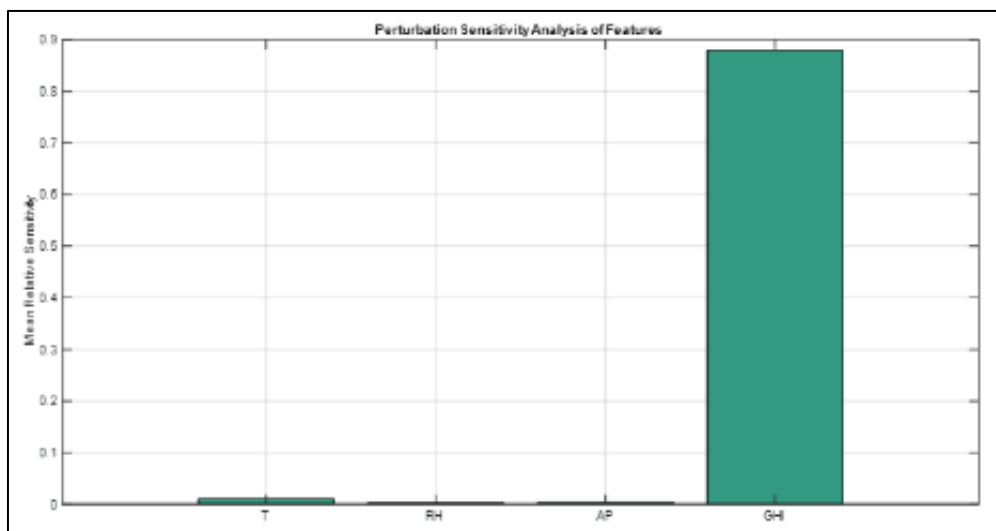


Figure 4 XGBOOST Model Perturbation Sensitivity Analysis

4.2. Sky Condition Classification

4.2.1. Classification accuracy

Fig. 5 shows the confusion matrix of the classified sky condition using the Random Forest model. The model achieved perfect classification accuracy on the analyzed subset. The matrix exhibited strict diagonal dominance with all 5,039 clear-sky, 2,025 partially cloudy, and 2,321 Overcast instances correctly assigned to their respective true classes. The model shows zero classification errors indicated by the absence of off-diagonal entries. This result points out the robustness of the Random Forest algorithm in handling the non-linear decision boundaries innate in environmental data, attesting to its suitability for complex classification tasks without overfitting [40].

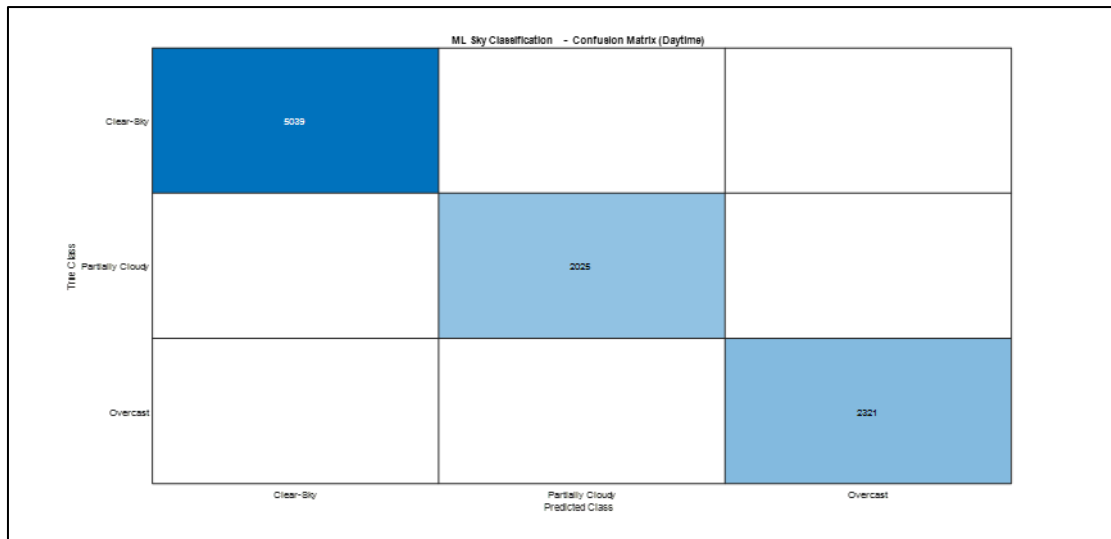


Figure 5 Confusion matrix

4.2.2. Feature and temporal analysis

The physical distinctiveness of the classified clusters was validated through feature space and temporal analysis by plotting the power against GHI as shown in Fig. 6. Clear-sky data (red circles) clustered in the upper-right quadrant, characterized by high GHI (600–1000 W/m² and high-power output (150–250 W, reflecting optimal generation conditions. Overcast data (blue squares) concentrated in the lower-left quadrant with low GHI (0–300 W/m², consistent with the significant attenuation of solar radiation by cloud cover [41]. The Partially Cloudy class (green triangles) occupied an intermediate position but displayed high variance and scatter, indicative of the stochastic irradiance fluctuations and ramping events caused by transient cloud shading [42].

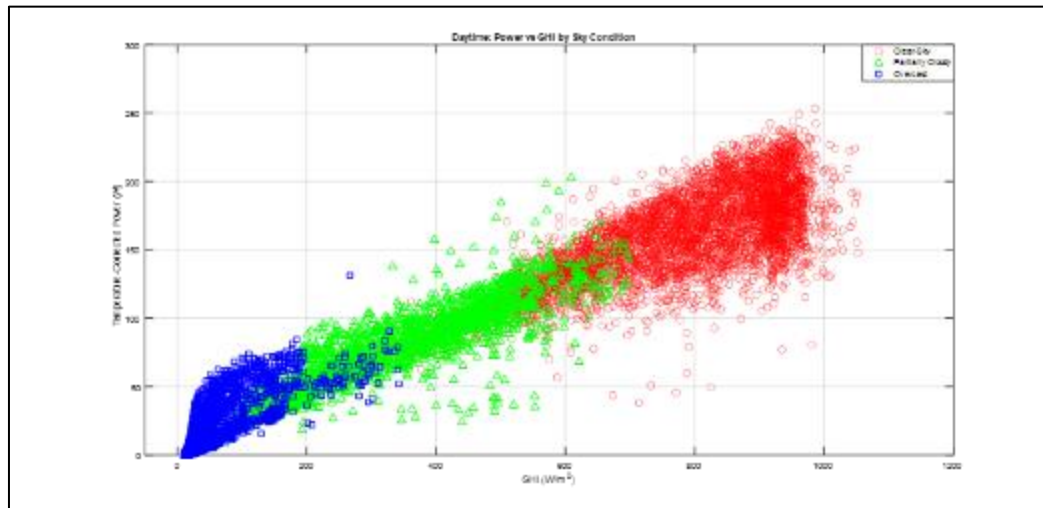


Figure 6 Daily plot of solar power output vs. GHI by sky condition

4.2.3. Daily power output by sky condition

The daily solar power output plot by sky conditions is shown in Fig. 7. The time-series plot revealed distinct diurnal patterns. Clear-sky data followed a symmetric bell-shaped pattern peaking at solar noon corresponding to the upper limit of energy generation.

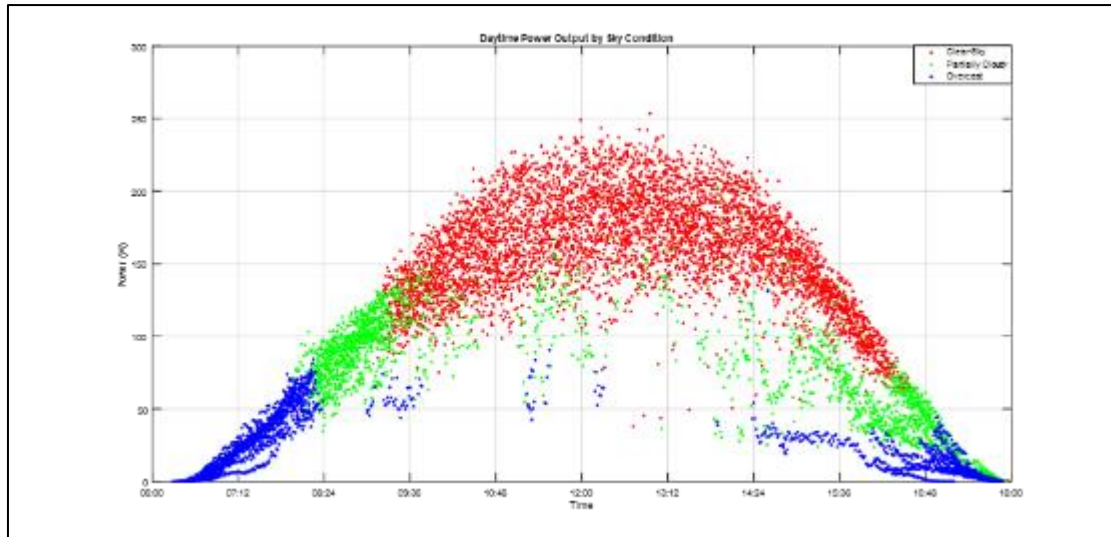


Figure 7 Daily plot of solar power output by sky condition

Overcast data formed a suppressed baseline, while partially cloudy data appeared as intermittent fluctuations beneath the clear-sky envelope. This temporal clustering assisted the model in distinguishing classes, particularly separating the consistent midday peaks of clear-sky conditions from the variable generation patterns of cloudy periods [43]. The Random Forest classifier effectively segregated the solar data into physically meaningful sky conditions. Although 2D projections showed some overlap—particularly for the volatile partially cloudy class—the multi-dimensional feature space allowed for perfect separation. This robust stratification is a critical prerequisite for subsequent forecasting, as different sky conditions necessitate distinct predictive modeling approaches to maximize accuracy [44].

4.3. Performance Across Weather Conditions

Fig. 8 shows the clear sky training plot of the XGBoost model. The prediction curve represented by the red dash line captures the general trend of solar power generation, even though the dataset is a classified across the period of consideration and may not fit into real-time diurnal solar power generation. However, it appears to smooth out the high-frequency noise present in the actual data (blue solid line). In the testing plot, the discrepancy is more apparent; the actual data shows significant volatility and sharp drops (possibly due to minor cloud interference or sensor noise despite the clear sky label), which the model fails to predict accurately, resulting in a smoother prediction curve that lags behind the erratic actual values. The model performance metrics recorded an R^2 of 0.7055, with an RMSE of 15.32 W and MAE of 11.69 W. While an R^2 of 0.70 is generally considered a good fit. This suggests that XGBoost, a tree-based ensemble method, may struggle with the specific high-frequency variance or noise present in this particular clear sky dataset.

Fig. 9 presents the partially cloudy plots demonstrating the model's ability to handle non-linear and intermittent data. In the training phase, the predictions track the sharp fluctuations caused by cloud cover reasonably well. In the testing phase, the model captures the general sawtooth pattern of power generation typical of passing clouds, although it occasionally underestimates peak power outputs (e.g., around sample index 350). The performance improves significantly compared to clear sky, with an R^2 of 0.8257. The RMSE (12.40 W) and MAE (8.20 W) are lower, indicating that the model generalizes better to the stochastic nature of partial cloud cover than it does to the noisy clear sky data.

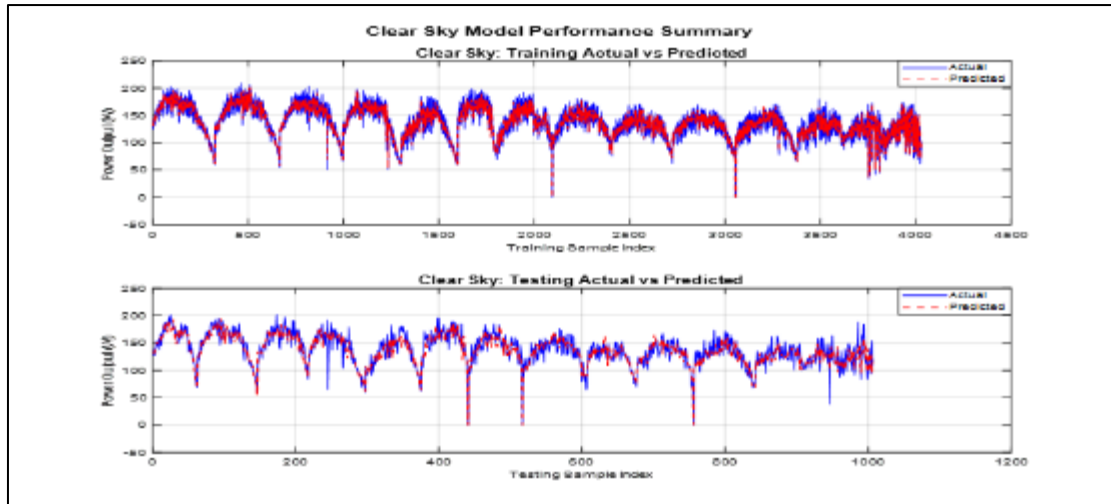


Figure 8 Plot of the XGBoost prediction model for training and testing data under clear sky condition

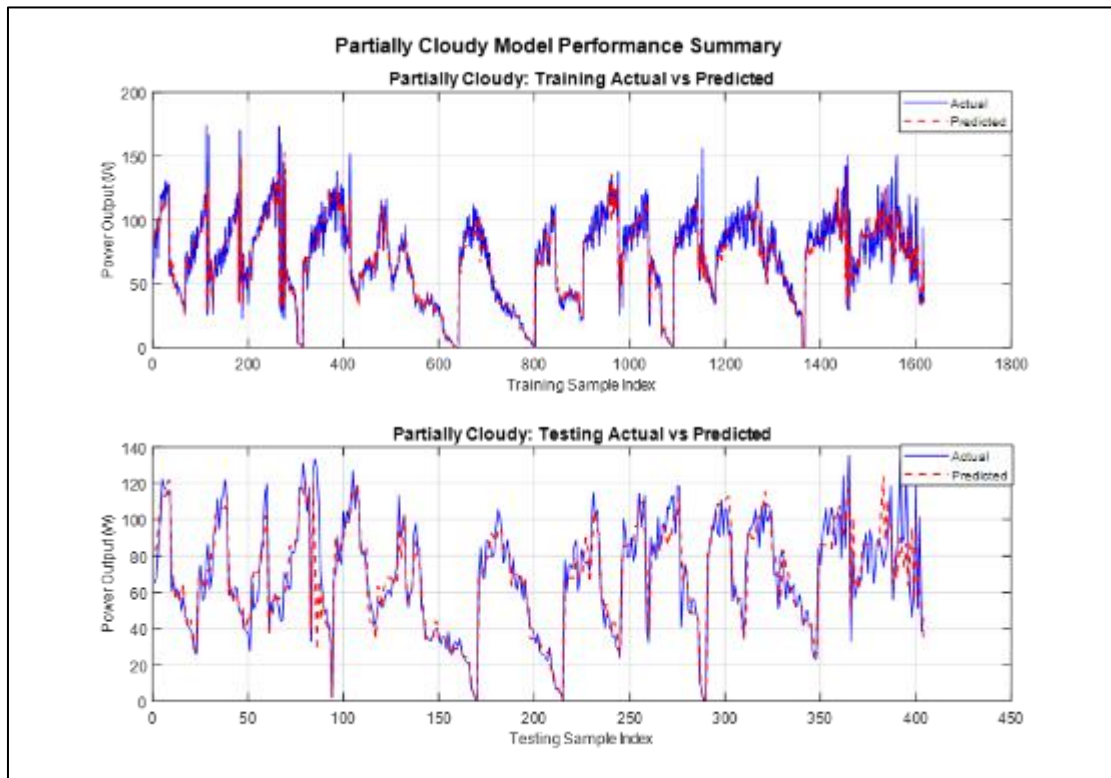


Figure 9 Plot of the XGBoost prediction model for training and testing data under partly cloudy condition

The model achieves its highest accuracy under overcast conditions as shown in Fig. 10. The training and testing plots show an almost perfect overlap between the actual and predicted power output. The smooth, gradual ramps characteristic of diffuse irradiance in overcast weather are easily learned by the gradient boosting trees. The model achieves a very high R^2 of 0.9551, with minimal errors (RMSE = 4.13 W, MAE = 2.58 W). This confirms that XGBoost is highly effective at modeling stable, low-volatility time-series data.

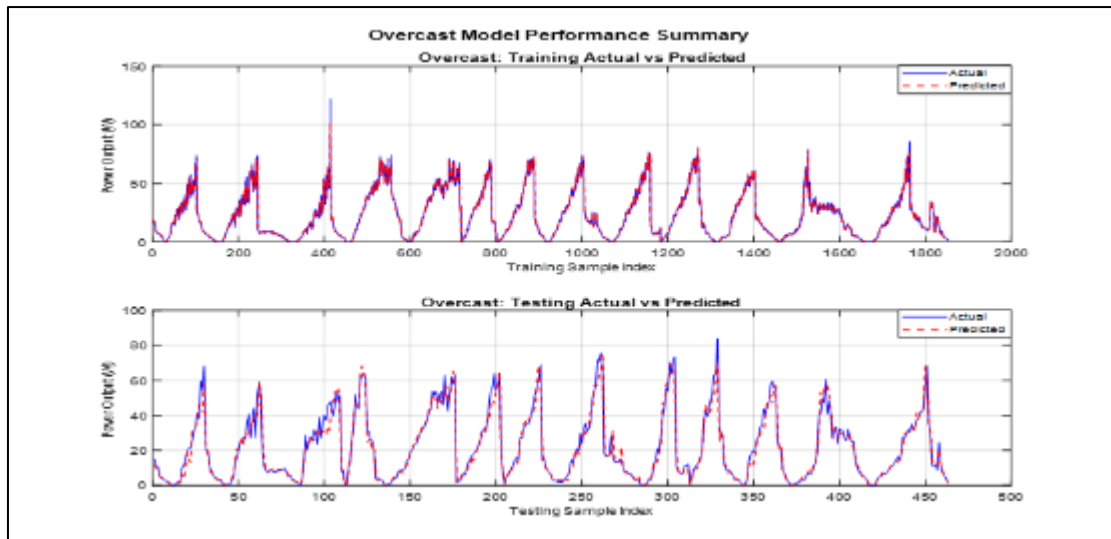


Figure 10 Plot of the XGBoost prediction model for training and testing data under overcast sky condition

The bar charts shown in Fig. 11 provide a comparative summary of the XGBoost model's efficacy across the three sky conditions. Comparing the coefficient of determination, the chart shows a clear progression, overcast ($R^2 \approx 0.96$) > partially cloudy ($R^2 \approx 0.83$) > clear sky ($R^2 \approx 0.71$). This trend is somewhat counter-intuitive, as clear sky conditions are typically the easiest to predict due to their deterministic nature. However, in this specific dataset, the clear sky testing data appears to contain significant noise or anomalies (as seen in the jagged blue lines in the time-series plots) that the tree-based XGBoost model could not fully capture. Conversely, the overcast data is smooth and stable, allowing the model to achieve near-perfect correlation.

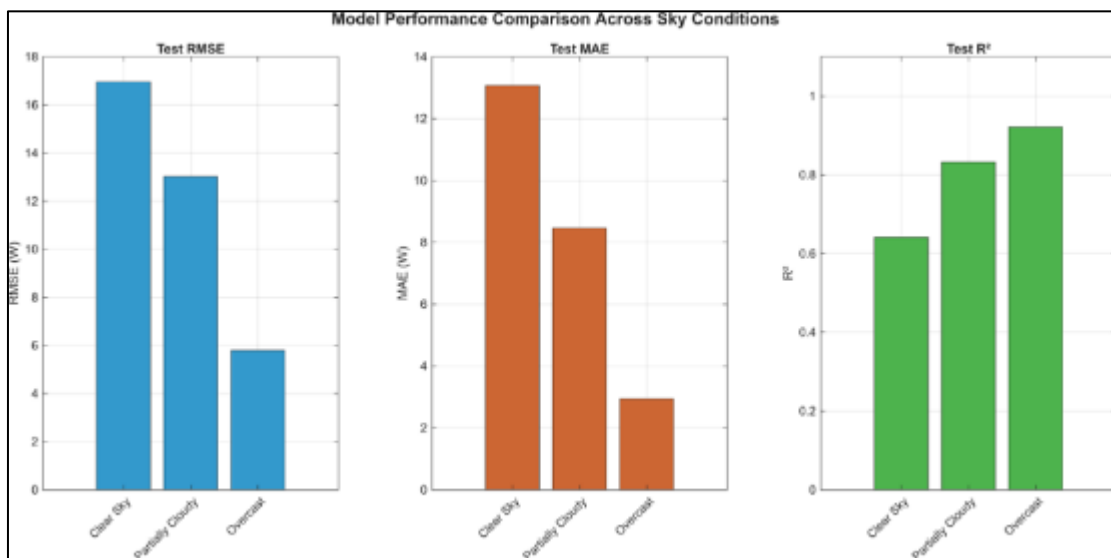


Figure 11 Comparison of performance metrics of XGBoost model for different sky conditions

The error metrics follow the inverse order of R^2 clear sky (highest error) > partially cloudy (medium error) > Overcast (lowest error). The clear sky condition has the highest RMSE (15.32 W) and MAE (11.69 W). This high error is likely due to the model's inability to track the high-frequency fluctuations in the actual data. Partially cloudy with moderate errors (RMSE 12.40 W). The model handles the variability well, but the inherent unpredictability of clouds leads to higher errors than in overcast conditions. Overcast with lowest errors (RMSE 4.13 W). The low magnitude of power generation combined with high signal stability results in the best performance.

5. Conclusion and recommendations

The following conclusions are drawn from this study: The XGBoost model achieved high predictive performance with performance metrics (testing: MAE = 4.05 W, RMSE = 8.10 W, $R^2 = 0.98$), demonstrating excellent generalization capability and resistance to overfitting through its regularization mechanisms and parallelized tree-learning architecture (Zhou et al., 2023). Perturbation sensitivity analysis confirmed Global Horizontal Irradiance (GHI) as the overwhelming driver of predictions (98.15% influence), consistent with fundamental photovoltaic principles wherein solar radiation directly controls module output (Bamisile et al., 2025). The Random Forest classifier achieved 100% accuracy in categorizing sky conditions, validating its robustness for handling non-linear environmental decision boundaries and enabling condition-specific forecasting strategies [40]. Model accuracy followed the order Overcast ($R^2 = 0.9551$, RMSE = 4.13 W) > partially cloudy ($R^2 = 0.8257$, RMSE = 12.40 W) > clear sky ($R^2 = 0.7055$, RMSE = 15.32 W). The counter-intuitive underperformance under clear skies was attributed to high-frequency noise or sensor anomalies in the testing data that tree-based ensembles smoothed rather than replicated.

His study recommends the deployment of explainable feature-weighting systems which implements SHAP-based interpretability tools alongside perturbation analysis to dynamically quantify feature contributions across sky conditions, enhancing operational transparency for grid dispatch decisions. It also recommends multi-temporal horizon extension which expands the model to support 15-min, hourly, and day-ahead forecasting horizons to accommodate diverse grid-operation needs, leveraging XGBoost's proven scalability for multi-step prediction tasks.

Compliance with ethical standards

Acknowledgement

The authors sincerely acknowledge the Nigerian Meteorological Agency (NIMET) office Heipang, where the hybrid solar–wind test rig is located and colleagues for their technical assistance and valuable suggestions.

Disclosure of conflict of interest

The authors further declare that there are no conflicts of interest regarding the publication of this manuscript.

References

- [1] Bo. Yang et al., "Novel bio-inspired memetic salp swarm algorithm and application to MPPT for PV systems considering partial shading condition," J. Clean. Prod., vol. 215, pp. 1203–1222, Apr. 2019, doi: 10.1016/j.jclepro.2019.01.150.
- [2] R. Nageem and R. Jayabarathi, "Predicting the Power Output of a Grid-Connected Solar Panel Using Multi-Input Support Vector Regression," in Procedia Computer Science, Elsevier B.V., 2017, pp. 723–730. doi: 10.1016/j.procs.2017.09.143.
- [3] X. Chen, X. Zhang, M. Dong, L. Huang, Y. Guo, and S. He, "Deep Learning-Based Prediction of Wind Power for Multi-turbines in a Wind Farm," Front. Energy Res., vol. 9, Jul. 2021, doi: 10.3389/fenrg.2021.723775.
- [4] K. Gandhi, A. Shah, V. Viswanath, P. Ranka, U. Dedhia, and N. Patil, "Predicting Solar Energy Generation with Machine Learning based on AQI and Weather Features," 2023, doi: 10.21203/rs.3.rs-3178713/v1.
- [5] D. Van Tai, "Solar photovoltaic power output forecasting using machine learning technique," in Journal of Physics: Conference Series, Institute of Physics Publishing, Oct. 2019. doi: 10.1088/1742-6596/1327/1/012051.
- [6] M. J. Marshall, Durgadevi, C. Gayatri, and R. Deepa, "Predicting Wind Turbine Performance Using Machine Learning Techniques," in E3S Web of Conferences, EDP Sciences, May 2023. doi: 10.1051/e3sconf/202338701003.
- [7] C. Vennila et al., "Forecasting Solar Energy Production Using Machine Learning," International Journal of Photoenergy, vol. 2022, 2022, doi: 10.1155/2022/7797488.
- [8] N. E. Benti, M. D. Chaka, and A. G. Semie, "Forecasting Renewable Energy Generation with Machine Learning and Deep Learning: Current Advances and Future Prospects," Sustainability, vol. 15, pp. 1–33, 2023, doi: 10.3390/su15097087.

- [9] N. Aksoy and I. Genc, "Predictive models development using gradient boosting based methods for solar power plants," *J. Comput. Sci.*, vol. 67, p. 101958, Mar. 2023, doi: 10.1016/J.JOCS.2023.101958.
- [10] O. E. Olabode, T. O. Ajewole, I. K. Okakwu, A. S. Alayande, and D. O. Akinyele, "Hybrid power systems for off-grid locations: A comprehensive review of design technologies, applications and future trends," *Sci. Afr.*, vol. 13, Sep. 2021, doi: 10.1016/j.sciaf. 2021.e00884.
- [11] J. M. Guerrero et al., "Systematic Review on Impact of Different Irradiance Forecasting Techniques for Solar Energy Prediction," 2022, doi: 10.3390/en15176267.
- [12] M. M. Forootan, I. Larki, R. Zahedi, and A. Ahmadi, "Machine Learning and Deep Learning in Energy Systems: A Review," *Sustainability*, vol. 14, no. 8, 2022, doi: 10.3390/su14084832.
- [13] S. A. Muraina, B. P. Akinbamiwa, D. S. Abiola, and S. I. Charles, "Artificial Intelligence-Driven Renewable Energy Solutions for Rural Electrification in Africa," *IJEMT International Journal of Engineering and Modern Technology*, 2025, doi: 10.56201/ijemt.vol.11.no3.
- [14] M. Elsaraiti and A. Merabet, "Solar Power Forecasting Using Deep Learning Techniques," *IEEE Access*, vol. 10, pp. 31692–31698, 2022, doi: 10.1109/ACCESS.2022.3160484.
- [15] S. Rinesh et al., "Prediction and classification of solar photovoltaic power generation using extreme gradient boosting regression model," *International Journal of Low-Carbon Technologies*, vol. 19, pp. 2420–2430, 2024, doi: 10.1093/ijlct/ctae197.
- [16] S. Subramani, S. Hari, K. Asha, V. Francis, R. Densil, and L. P. Shaji, "Enhanced Short-Term Photovoltaic Power Prediction using a Hybrid Improved Whale Optimization Algorithm with XGBoost," in *2024 Second International Conference on Smart Technologies for Power and Renewable Energy (SPECon)*, Ernakulam, India: IEEE, Apr. 2024, pp. 1–4. doi: DOI10.1109/specon61254.2024.10537570.
- [17] H. N. Nguyen, Q. T. Tran, C. T. Ngo, D. Dam Nguyen, and V. Q. Tranid, "Solar energy prediction through machine learning models: A comparative analysis of regressor algorithms," *PLoS One*, vol. 20, no. 1, 2025, doi: 10.1371/journal.pone.0315955.
- [18] M. Ali, J. Gil, K. S. Kim, and S. K. Kim, "Attention-Driven Hybrid Ensemble Approach with Bayesian Optimization for Accurate Energy Forecasting in Jeju Island's Renewable Energy System," *IEEEAccess*, vol. 13, 2025, doi: 10.1109/ACCESS.2025.3526943.
- [19] Q. X. Waqar W, "Data Driven Long Short-Term Load Prediction: LSTM-RNN, XG-Boost and Conventional Models in Comparative Analysis," *Computational Intelligence*, vol. 41, no. 3, 2025.
- [20] R. Amir-Hosseini, S. Javad, and S. Miadreza, "An Advanced Hybrid CNN-LSTM-XGBoost Model with Wavelet Transform and Attention Mechanism for Accurate Short-Term Electricity Price Forecasting," *SSRN*, 2025, [Online]. Available: <https://ssrn.com/abstract=5314801>
- [21] R. Perez et al., "Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe," *Solar Energy*, vol. 94, pp. 305–326, Aug. 2013, doi: 10.1016/J.SOLENER.2013.05.005.
- [22] N. Zwane et al., "A Bibliometric Analysis of Solar Energy Forecasting Studies in Africa," Aug. 01, 2022, MDPI. doi: 10.3390/en15155520.
- [23] N. Yuhao, S. Yuchi, C. Yuanlei, O. Rachel, and B. Adam, "PV power output prediction from sky images using convolutional neural network: The comparison of sky-condition-specific sub-models and an end-to-end model," *Journal of Renewable and Sustainable Energy*, vol. 12, no. 4, p. 046101, 2020.
- [24] P. S. Bhattacharya and D. T. Dimitrios, "Robust model rankings of forecasting performance," *J. Forecast.*, vol. 37, no. 6, pp. 676–690, 2018.
- [25] W. S. Zi and M. D. Hyacinth, "Climate Change, Rainfall Trends and Variability in Jos Plateau," *Journal of Applied Sciences*, vol. 20, no. 2, pp. 76–82, Feb. 2020, doi: 10.3923/jas.2020.76.82.
- [26] LUTRON, "Exposure meter LUTRON SPM-1116SD - Lutron Instruments," lutroninstruments.eu. Accessed: Dec. 08, 2023. [Online]. Available: <https://www.lutroninstruments.eu/exposure-meters/exposure-meter-lutron-spm-1116sd/>
- [27] S. Knoop, P. Ramakrishnan, and I. Wijnant, "Dutch offshore wind atlas validation against cabauw meteomast wind measurements," *Energies (Basel)*, vol. 13, no. 24, Dec. 2020, doi: 10.3390/en13246558.

- [28] LUTRON, "Anemometer LUTRON AM-4257SD - Lutron Instruments," [lutroninstruments.eu](https://www.lutroninstruments.eu/mechanical-anemometers/anemometer-lutron-am-4257sd/). Accessed: Dec. 08, 2023. [Online]. Available: <https://www.lutroninstruments.eu/mechanical-anemometers/anemometer-lutron-am-4257sd/>
- [29] T. Niedzielski and M. Halicki, "Improving Linear Interpolation of Missing Hydrological Data by Applying Integrated Autoregressive Models," *Water Resources Management*, 2023, doi: 10.1007/s11269-023-03625-7.
- [30] A. Alabrah, "An Improved CCF Detector to Handle the Problem of Class Imbalance with Outlier Normalization Using IQR Method," *Sensors*, vol. 23, no. 9, 2023, doi: 10.3390/s23094406.
- [31] D. Nielsen, "Tree Boosting with XGBoost Why Does XGBoost Win 'Every' Machine Learning Competition?" Norwegian University of Science and Technology, 2016.
- [32] S. Ahmeda et al., "The Deep Learning ResNet101 and Ensemble XGBoost Algorithm with Hyperparameters Optimization Accurately Predict the Lung Cancer," *Applied Artificial Intelligence*, vol. 37, no. 1, 2023, doi: 10.1080/08839514.2023.2166222.
- [33] V. R. Joseph and A. Vakayil, "SPlit: An Optimal Method for Data Splitting," *Technometrics*, vol. 64, no. 2, pp. 166–176, 2022, doi: 10.1080/00401706.2021.1921037.
- [34] J. Kang, J. Lee, and S. Lee, "Data-Driven Minute-Ahead Forecast of PV Generation with Adjacent PV Sector Information," *Energies (Basel)*, vol. 16, no. 13, Jul. 2023, doi: 10.3390/en16134905.
- [35] S. Poddar, M. Kay, and J. Boland, "A Standardized Sky Condition Classification Method for Multiple Timescales and Its Applications in the Solar Industry," *Energies (Basel)*, 2024, doi: 10.3390/en17184616.
- [36] J. DeMarchi, R. Rijken, J. Melrose, B. Madahar, G. Fumera, and F. Roli, "Evaluation of Robustness Metrics for Defense of Machine Learning Systems," in *2023 International Conference on Military Communications and Information Systems (ICMCIS)*, Skopje, North Macedonia: IEEE, May 2023.
- [37] Z. Arif Ali, Z. H. Abduljabbar, H. A. Tahir, A. Bibo Sallow, and S. M. Almufti, "eXtreme Gradient Boosting Algorithm with Machine Learning: a Review," *Academic Journal of Nawroz University*, vol. 12, no. 2, pp. 320–334, May 2023, doi: 10.25007/ajnu.v12n2a1612.
- [38] B. Zhou, X. Chen, G. Li, P. Gu, J. Huang, and B. Yang, "XGBoost–SFS and Double Nested Stacking Ensemble Model for Photovoltaic Power Forecasting under Variable Weather Conditions," *Sustainability*, vol. 15, no. 17, p. 13146, Sep. 2023, doi: 10.3390/su151713146.
- [39] O. Bamisile, C. Acen, D. Cai, Q. Huang, and I. Staffell, "The environmental factors affecting solar photovoltaic output," *Renewable and Sustainable Energy Reviews*, vol. 208, p. 115073, Feb. 2025, doi: 10.1016/j.rser.2024.115073.
- [40] Q. Wang, T.-T. Nguyen, J. Z. Huang, and T. T. Nguyen, "An efficient random forests algorithm for high dimensional data classification," *Adv. Data Anal. Classif.*, vol. 12, no. 4, pp. 953–972, Dec. 2018, doi: 10.1007/s11634-018-0318-1.
- [41] U. C. Dumka, P. G. Kosmopoulos, S. S. Ningombam, and A. Masoom, "Impact of Aerosol and Cloud on the Solar Energy Potential over the Central Gangetic Himalayan Region," *Remote Sens. (Basel)*, vol. 13, no. 16, p. 3248, Aug. 2021, doi: 10.3390/rs13163248.
- [42] K. Lappalainen and S. Valkealahti, "Fluctuation of PV array global maximum power point voltage during irradiance transitions caused by clouds," *IET Renewable Power Generation*, vol. 13, no. 15, pp. 2864–2870, Nov. 2019, doi: 10.1049/iet-rpg.2019.0085.
- [43] Z. Xie, J. Chen, Y. Yin, and A. Meng, "Ultra Short-Term Photovoltaic Power Forecasting Based on Similar Day Theory and BATCN," in *2023 3rd International Conference on Energy, Power and Electrical Engineering (EPEE)*, IEEE, Sep. 2023, pp. 174–178. doi: 10.1109/EPEE59859.2023.10352024.
- [44] N. Masi and R. Balestriero, "A NOVEL APPROACH TO AI WEATHER EVALU-ATION THROUGH STRATIFIED ASSESSMENTS OF FORE-CASTS OVER EARTH," *SAFE*, pp. 1–24, 2025.
- [45] Y. Zhou, Y. Li, D. Wang, and Y. Liu, "A multi-step ahead global solar radiation prediction method using an attention-based transformer model with an interpretable mechanism," *Int. J. Hydrogen Energy*, vol. 48, no. 40, 2023, doi: 10.1016/j.ijhydene.2023.01.068.
- [46] O. Bamisile, C. Acen, D. Cai, Q. Huang, and I. Staffell, "The environmental factors affecting solar photovoltaic output," *Renewable and Sustainable Energy Reviews*, vol. 208, p. 115073, Feb. 2025, doi: 10.1016/j.rser.2024.115073.