Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Prediction of Photovoltaic Power Output Using Environmental Parameters: A Comparative Study of Linear Regression and Random Forest

Aarya Sawant
Computer Engineering Department
Fr. Conceicao Rodrigues College of Engineering
Mumbai, India

Prof. Kalpana Deorukhkar
Computer Engineering Department
Fr. Conceicao Rodrigues College of Engineering
Mumbai, India

***Abstract* - Accurate prediction of photovoltaic power output is essential for efficient energy management and reliable operation of smart grids. This study compares Linear Regression and Random Forest models for predicting photovoltaic AC power output using environmental parameters from a public solar dataset with more than 100,000 records. The input variables include solar irradiance, module temperature, ambient temperature, and wind speed, while direct electrical variables were excluded to avoid target leakage. The model performance was assessed using RMSE, MAE, and R^2. The results show that Random Forest achieved better performance than Linear Regression, with lower prediction errors and a higher R^2 value of 0.9831. The findings indicate that environmental-parameter-based machine learning models can provide accurate photovoltaic power prediction, and that Random Forest is better suited for capturing nonlinear relationships in the data.**

I. INTRODUCTION

In recent years, a growing interest in sustainable energy sources has increased the use of solar photovoltaic (PV) systems worldwide. The advantages of using solar energy include its sustainability and the continually decreasing cost of installing PV systems. One of the main challenges in PV systems, however, is that their electrical power output will vary with respect to the environment (solar irradiance, temperature, and wind speed). This inherent variability creates significant operational difficulties when using smart grids as a result of voltage fluctuations, frequency instability, and uncertainty concerning the management of power flow. Accordingly, accurate prediction of the power output of PV systems is critical to the ongoing stability and reliability of present-day electrical energy systems.

The quantity of solar irradiance has the greatest influence on the generation of electricity from the PV module compared to other environmental factors that affect the performance of the PV module. However, other temperature-related factors also affect the electricity generation by a PV module (i.e., both the module and the ambient temperatures). Wind speed is another environmental variable that directly impacts the efficiency of PV systems because a PV module experiences cooling as ambient air passes over it. In turn, these environmental factors have direct impacts on the power output of photovoltaic modules and are typically incorporated into solar power prediction models as input.

Machine learning has played a vital role in solar forecasting and can represent complex relationships from data without being entirely reliant on physical equations. Various algorithms have been effectively used for the prediction of PV power generation.

Some examples of successful algorithms include: Support Vector Machines (SVM), Artificial Neural Networks (ANN), Decision Trees (DT), and ensemble methods (e.g., Random Forest).

Ensemble models provide an effective approach because they can model nonlinear interactions in the environmental variables better than linear models.

Even though many studies have researched machine learning based frameworks for solar forecasting, defined and practical model comparisons would still be helpful to the field. This research compares Linear Regression (LR) and Random Forest (RF) models for predicting photovoltaic power output using environmental parameters from operational data. As the goal of this study was to demonstrate practical applicability, direct electrical variables were excluded from the feature set in an effort to minimise target leakage. The respective model performance was evaluated through the use of Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R^2 [1]. The objective of this research study was to investigate the effectiveness of two models' abilities to predict photovoltaic power output, as well as to emphasise the trade-off between the simplicity of a model and its predictive accuracy.

II. DATASET AND PARAMETERS

The dataset used in this study was obtained from a publicly available photovoltaic system dataset provided by Kanagolkar [7]. The dataset contains more than 100,000 records, all from a grid-connected solar photovoltaic system. It contains environmental data as well as the corresponding AC power output, which was used in the forecasting model. Solar irradiance (W/m^2) is the main input variable for the model, reflecting the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

solar energy hitting the photovoltaic panel surface. Solar irradiance is likely the most important variable in the model, as it is the photovoltaic system's main energy source for producing electricity. Apart from solar irradiance, module temperature and ambient temperature are also used in the model. As the temperature of a module rises, the semiconductor material within the module becomes less efficient, which causes a decrease in both voltage and power output. Wind speed is also used in the model as an environmental parameter, affecting the module's temperature. The variable of interest for this particular problem is AC power output, which is denoted as 'W'. The electrical current-related variables were excluded while choosing the feature set to avoid data leakage, as they are directly associated with power output. These features were selected to enable the model to learn the relationship between environmental conditions and PV power output.

III. METHODOLOGY

A. Data Preprocessing

Before building the model, it was important to check the quality of the data. For instance, non-physical values, such as negative irradiance and power, were removed. There were no missing values in the dataset. Thus, no imputation was necessary.

To balance the influence of each feature on the model and to improve numerical stability, it was important to normalise all variables. Normalisation ensures that no variable dominates the learning process because of its numerical value. The dataset was then split into two sets: a training set and a testing set. The data was split using an 80/20 ratio. The training set was used to train the model, while the testing set was used to evaluate its performance.

B. Model Selection

Two models were implemented for prediction: Linear Regression and Random Forest.

Linear Regression was chosen as a baseline model to test the extent to which the photovoltaic power output can be described in terms of linear relationships between the environmental factors and the photovoltaic output. Linear Regression estimates the parameters that best fit the data by minimising the squared error.

Random Forest is a type of ensemble learning algorithm developed by Breiman [6], which consists of a set of decision trees built using randomised subsets of the data. The predictions are made by averaging the outputs of the individual decision trees. It was chosen for the purpose of the task because it is capable of handling non-linear relationships and complex interactions among the environmental factors.

C. Evaluation Metrics

The performance of the models was evaluated using three widely adopted forecasting metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). These metrics provide complementary perspectives on prediction accuracy.

The RMSE measures the square root of the average squared differences between predicted and actual values:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (1)$$

The MAE computes the average magnitude of prediction errors:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

The coefficient of determination (R^2) represents the proportion of variance explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

where y_i represents the actual power output, $\hat{y}_i$ represents the predicted value, $\bar{y}$ is the mean of actual values, and n is the total number of observations.

D. Experimental Setup

The experiments were conducted using MATLAB R2023b. The data was split into training and testing sets using an 80/20 split. All features were normalised to improve numerical stability during model training.

For the Random Forest model, 100 decision trees were created using bootstrap aggregation. The number of predictors used to split each tree was automatically determined by MATLAB's heuristic function. There was no limit to tree depth to allow for nonlinear interactions between variables.

The Linear Regression model was implemented using ordinary least squares. The coefficients of the model were determined by minimising the sum of the squared differences between predicted and actual power output.

All performance metrics were computed on the unseen test data to provide a fair comparison of all models.

IV. RESULTS AND DISCUSSION

A. Model Performance Comparison

TABLE I
PERFORMANCE COMPARISON OF MODELS

Model	RMSE (W)	MAE (W)	R^2
Linear Regression	12858	7491.7	0.98021
Random Forest	11870	6582.5	0.98313

The predictive accuracy of the Linear Regression and Random Forest models was assessed using RMSE, MAE, and R^2. The comparative results are presented in Table 1.

The Linear Regression model had an RMSE of 12,858 W, an MAE of 7,492 W, and an R^2 value of 0.9802. These results indicate that a large proportion of the variation in photovoltaic power output can be explained by linear relationships between environmental factors and power output.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

The Random Forest model showed better accuracy than the Linear Regression model. It achieved a lower RMSE of 11,870 W, a lower MAE of 6,583 W, and a higher R^2 value of 0.9831. This better performance suggests that nonlinear relationships among environmental factors also play an important role in power prediction.

The difference in model performance between the two models is moderate. The results show that the incorporation of nonlinear relationships in the model improves predictive accuracy.

Though the increase in R^2 from 0.9802 to 0.9831 seems modest, the decrease in RMSE and MAE shows that Random Forest is more adept at detecting non-linear environmental relationships that affect photovoltaic performance. This may suggest that while irradiance explains much of the variation in output, other factors such as temperature and wind speed also contribute in nonlinear ways.

B. Prediction Analysis

Figure 1 demonstrates the correlation between actual and predicted values of AC power, as obtained through the Random Forest model. The cluster of points along the diagonal of the plot shows a high level of agreement between the actual and predicted values. The spread of points at higher levels of power may reflect the variability of environmental conditions.

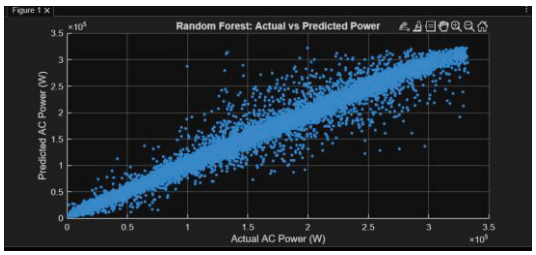

Fig. 1. Actual vs Predicted AC Power

C. Feature Importance

Figure 2 presents the feature importance obtained from the Random Forest model. The feature importance analysis from the Random Forest model shows that solar irradiance is the primary factor that affects photovoltaic power output. This is consistent with the fundamental principles of photovoltaic operation, as it is the primary factor that determines the input to the system. The other factors, such as temperature-related parameters and wind speed, are of secondary importance, which is likely due to efficiency and cooling effects.

The dominance of irradiance as a factor further suggests that the model is capturing relationships that are consistent with physical behavior rather than arbitrary statistical ones.

D. Implications for Sustainable Energy Systems

The precise forecasting of photovoltaic power generation can enable better scheduling, demand management, and renewable power integration in a smart grid infrastructure. The high accuracy of the forecasting model in this study indicates

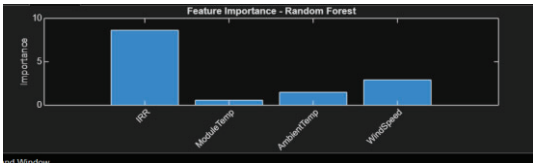

Fig. 2. Feature Importance from Random Forest

that data-driven models based on environmental measurements can effectively provide accurate power forecasts. Although the Random Forest method had better predictive accuracy, it had relatively higher computational complexity when compared to the Linear Regression method. The Linear Regression method is better suited for applications that require faster inference and better interpretability, such as in embedded systems or real-time systems with limited computational power. The Random Forest method is better suited for applications with sufficient computational power, such as in centralised systems in the smart grid domain.

V. CONCLUSION

In order to ensure efficient smart grid operation in sustainable energy infrastructures, precise short-term forecasting of photovoltaic power output is vital. This study compared Linear Regression and Random Forest for predicting photovoltaic power output using environmental parameters.

The study indicated that photovoltaic power output can be predicted effectively based on environmental factors. Although the Linear Regression algorithm was able to effectively model the primary linear relationship between solar irradiance and power output, the performance of the Random Forest algorithm was superior, with lower RMSE and MAE values, and a higher R^2 value of 0.9831. This suggests the existence of nonlinear relationships between environmental factors affecting photovoltaic power output.

Photovoltaic power generation is influenced predominantly by solar irradiance, which is consistent with the physical behaviour of photovoltaic systems. The findings support the use of machine learning models for solar power forecasting, ultimately allowing for the reliable integration of renewable energy into sustainable energy systems. Future research can include model generalisation in different geographical settings and the development of adaptive models.

REFERENCES

[1] K. J. Iheanetu, “Solar Photovoltaic Power Forecasting: A Review,” *Sustainability*, vol. 14, no. 24, p. 17005, 2022, doi: 10.3390/su142417005.

[2] C. Wan, J. Zhao, Y. Song, Z. Xu, J. Lin, and Z. Hu, “Photovoltaic and Solar Power Forecasting for Smart Grid Energy Management,” *CSEE Journal of Power and Energy Systems*, vol. 1, no. 4, pp. 38–46, 2015.

[3] K. Barhmi, C. Heynen, S. Golroodbari, and W. van Sark, “A Review of Solar Forecasting Techniques and the Role of Artificial Intelligence,” *Solar*, vol. 4, no. 1, pp. 99–135, 2024, doi: 10.3390/solar4010005.

[4] Y. Ledmaoui, A. El Maghraoui, M. El Aroussi, R. Saadane, A. Chebak, and A. Chehri, “Forecasting Solar Energy Production: A Comparative Study of Machine Learning Algorithms,” *Energy Reports*, vol. 10, pp. 1004–1012, 2023, doi: 10.1016/j.egyr.2023.07.042.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

[5] R. Vennila *et al.*, "Forecasting Solar Energy Production Using Machine Learning," *International Journal of Photoenergy*, vol. 2022, pp. 1–12, 2022, doi: 10.1155/2022/7797488.
[6] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[7] A. Kanagolkar, "SolarGeneration," Kaggle Dataset, 2024. doi: 10.34740/kaggle/dsv/8700050.