Journal of Thermal Engineering
https://jten.yildiz.edu.tr/
https://doi.org/10.47481/jten.0063

RESEARCH ARTICLE

Predictive modeling of solar energy generation using supervised machine learning algorithms

Bhushan Behede[1,*], Makarand Shahade[2], Siddharth Chakrabarti[3], Uday Wankhede[4], Yogesh Sonawane[5], Dattatray Doifode[6], Rahul Patil[7]

[1]School of Technology Management and Engineering, SVKM NMIMS Global University, Dhule, Maharashtra 424001, India
[2]School of Technology Management and Engineering, SVKM NMIMS Global University, Dhule, Maharashtra 424001, India
[3]Dr. Vishwanath Karad MIT World Peace University, Pune, 411038, Maharashtra, India
[4]Government College of Engineering, Nagpur, Maharashtra 441108, India
[5]School of Technology Management and Engineering, SVKM NMIMS Global University, Dhule, Maharashtra 424001, India
[6]School of Technology Management and Engineering, SVKM NMIMS Global University, Dhule, Maharashtra 424001, India
[7]Nutan Maharashtra Institute of Engineering and Technology, Talegaon, Maharashtra 410507, India

Abstract

The nonlinear and dynamic nature of solar energy generation makes effective energy management, operational planning, and grid stability challenging. Purely numerical and statistical approaches are often unable to pick up the nonlinearity in the photovoltaic power generation. The present work tests the predictive capability of supervised regression algorithms for photovoltaic energy forecasting. This research is a comparative analysis of five algorithms, which are Linear Regression, k-Nearest Neighbor, Support Vector Regression, Decision Tree, and Random Forest. Unlike earlier studies that just targeted irradiation parameters and focused on short-term datasets, this study incorporates a long-term dataset, consisting of 95,950 hourly observations collected from a rooftop photovoltaic plant found in the Western Australia region from 1990 to 2014. The dataset includes multiple variables that are categorized into solar-irradiance and meteorological variables. Solar-irradiance variables are Direct-normal, Global-horizontal, and Diffused-horizontal. Meteorological variables are Wet-bulb temperature and Dew-point temperature. The photovoltaic energy produced is the output variable. The model is trained strategically with hyperparameter tuning performed using a hold-out k-fold cross-validation. From the first correlation heatmap, it is found that Direct-normal irradiance affects the photovoltaic energy the most, followed by Global-horizontal, and Diffused-horizontal irradiance. Wet-bulb temperature and the Dew-point temperature are the least influential parameters. The performance metrics evaluated to assess the developed models are Mean-Squared Error, Root Mean-Squared Error, Mean Absolute Error, and the Coefficient of Determination. Overall, the ensemble models showed superior performance to the regression-based models. Ensemble models pick up the non-linearity among the variables well. Compared to other models, Random Forest achieves the highest value of the Coefficient of Determination. Also, the lowest values of Mean Absolute Error, Mean-Squared Error, and Root Mean-Squared Error. Compared to the linear regression model, the random forest model has reduced the Mean-Squared Error by 45.51%, Mean Absolute Error by 41.95%, and increased the Coefficient of Determination by 1.34%. The average reduction in error metrics in all the models stays below 38%. The proposed framework can support utility operators, photovoltaic system engineers, long-term energy planners, and policymakers in efficient photovoltaic energy monitoring and forecasting to enhance societal energy security.

Keywords: Correlation heatmap, learning curve analysis, machine learning, Random Forest, solar irradiation

Cite this article as: Smith, Behede, B., Shahade, M., Chakrabarti, S., Wankhede, U., Sonawane, Y., Doifode, D., & Patil, R. (2026). Predictive modeling of solar energy generation using supervised machine learning algorithms. Journal of Thermal Engineering, 12(5), 1748–1759. https://doi.org/10.47481/jten.0063

*Corresponding Author
E-mail Adress: bhushan.behede@svkm.ac.in

Submitted: : 27 December 2025; **Accepted:** 19 June 2026

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkılıç

1. Introduction

Energy plays a vital role in human development. Conventional and renewable are the two major forms of energy that can be converted into electrical energy. Fossil fuel-driven conventional methods of energy conversion have directly or indirectly led to environmental pollution and hazards to human health [1]. The urgent attention to reducing carbon footprints is possible through the widespread use of renewable energy resources for electricity generation [2,3]. The Government of India is curtailing the use of conventional resources and boosting the use of renewable sources [4]. Solar energy is the most prominent, abundantly available, free form of renewable energy on Earth. [5–8]. Over the past two decades, the solar energy installation capacity of India has almost doubled, while it has still been stagnant for non-renewable energy [4]. In recent times, the country has appeared as a growing leader in adopting solar energy for power generation. The National Solar Mission of India has set the renewable energy production target of 300 GW, which is to be achieved by 2030 [9]. To meet the goal, the government is aiding the renewable energy sector by introducing beneficiary schemes and subsidies [4,7].

Solar Photovoltaic (PV) cells convert incident solar radiation directly into electricity without leaving any carbon footprint [5,10]. Rooftop PV plants are the most adopted choice amongst other commercial and large-capacity PV plants [11]. It offers many advantages, such as affordability, scalability, reliability, effective use of unused roof area, and lower transmission costs [12]. However, technology still has one of the crucial problems, which is the integration of PV energy with the expanding widespread electrical grid. The reason behind this problem is the intermittent availability of solar energy throughout the day. One of the ways to tackle this problem lies in the correct prediction of PV energy output. This will help engineers to achieve better scheduling of energy, management of grid infrastructure, design of storage systems, and decision-making by the government [13]. The solution can be further used to reduce the dependency of the industries on the costly conventional energy sources [14,15].

Predicting the total PV energy harnessed from a rooftop PV plant in advance is a challenging task due to the widespread use of technology. Additionally, the PV energy varies significantly due to non-linear meteorological and environmental conditions [16]. Machine Learning (ML) models have appeared recently as one of the most powerful tools across various scientific solutions, which can provide a structured solution to complex and non-linear problems. It enables correct prediction along with efficient optimization and deeper analysis [17]. ML methods provide added advantages over conventional statistical tools, such as integration of large datasets and the power of combining multiple models [18]. Therefore, ML models are being used recently by many researchers to accurately predict the PV energy generation [17,19–21].

Samir et al. [20] have developed seven regression-based ML models to predict the PV energy of a standalone system. These ML models are namely Linear, Ridge, Lasso, Bayesian, Decision Tree (DT), Gradient Boosting, and Artificial Neural Network (ANN). Irradiance and meteorological parameters were considered as input variables. DT model outperformed the other models and appeared as the best-suited model with a Coefficient of Determination (R^2) value of 0.99. Abderraouf Bouakkaz et al. [19] have evaluated the performance of different ML models, such as LR, Random Forest (RF), Support Vector Regression (SVR), and ANN. Two distinct datasets are used to train these models; one has meteorological parameters, and another being from on-site measurements. Compared to others, RF has recorded the best performance with the minimum value of Root Mean-Squared Error (RMSE), Mean-Squared Error (MSE), and Mean Absolute Error (MAE). Montaser Abdelsattar et al. [18] have tested six ML models, namely, CatBoost, Gradient-Boost Machine (GBM), Multilayer Perceptron (MLP) regressor, SVR, XGBoost, and RF. To train and test these models, 4213 data points are used, which are collected from one of the PV energy production facilities. The result shows that with an R^2 value of 0.940, RF has delivered the best overall performance. Christian Idogho et al. [22] have developed SVR and ANN models to predict the PV energy. The models are stretched across the different regions of Nigeria. The dataset used to develop the models consists of data points from the past 12 years, including observations of irradiation parameters, cloud cover data, temperature parameters, and humidity values. Out of the developed models, it has been found that regions with hot and dry climate conditions have higher PV energy harnessing potential than regions with high cloud cover and humidity. Aayushi Tondon et al. [23] have evaluated the performance of LR, k-NN, Extreme Gradient Boosting (XGB), DT, and RF. The dataset used to develop the models was obtained from NASA's POWER project, which is particularly for the Indian state of Rajasthan. The dataset encompasses the various parameters, such as specific humidity, surface pressure, wind speed, temperature, and dew point, from 2013 to 2022. The performance of the model was assessed using four performance metrics. With the highest value of R^2 compared to all other models, RF has appeared as the best model. Mohd. Herwan Sulaiman et al. [24] did the research on the use of RF, SVR, and DT models for short-term PV energy prediction under various Malaysian weather conditions. In this study, the dataset used was for seven days from January 15–21, 2024, and it was collected from different geographical areas. The dataset includes observations of hourly irradiation data, relative humidity, atmospheric temperatures, and wind velocity. The result shows that RF outperforms others with the highest performance metrics. Rahul Gupta et al. [25] conducted a study on an explainable data-driven setting for the prediction of GHI using different regression-based ML algorithms. The developed models are DT, RF, Extreme Gradient-Boost (XGB), and Extra Trees (ET). Variance inflation factor (VIF) is applied to select the best feature. The models were evaluated by computing various performance metrics like MAE, RMSE, and R^2. The best performance is obtained for the Extra Trees model. It shows high accuracy and precision. The inclusion of SHAP (Shapley Additive Explanations) values made the model interpretable by quantifying each variable's involvement in accurately predicting the PV energy. Saroj Kumar Panda and Manoj Kumar Panda [26] have developed two ML models, namely Linear Least-Squared Regression and SVR. These models are trained using a dataset that spans ten months. All the data are collected from the weather station of the University of Massachusetts Amherst. This data set consists of input variables such as dry-bulb and dew-point temperatures, wind velocity, sky cover, probability of precipitation, and relative humidity. The output intensity of solar radiation in W/m2 is predicted. From the developed models, it is found that SVR with Radial Basis Function kernel has lower RMSE and higher accuracy. The study by Chaaban et al. [27] presented a comparative evaluation of five ML models, namely, LR, RF, k-NN, ANN, and Long Short-Term Memory (LSTM). Real-world operational data of two years, starting from July 2020 to June 2022, was obtained from the 300 MW Sakaka PV Independent energy Producer plant found in Saudi Arabia. The dataset consists of input variables such as Global-horizontal irradiance from pyranometers, plane-of-array irradiance, module temperature, and other weather parameters. The models have predicted hourly PV energy generation in the output. Compared to other models, the RF has the highest value of R^2 and the lowest values of RMSE and MAE. Abhishek Kumar Tripathi et al. [28] have performed a comparative analysis to find the best-suited ML models amongst Multivariate Regression (MR), SVR, and Gaussian Process Regression (GPR). The dataset used in the study was obtained from an experimental setup of a rooftop PV plant found at Surampalem, Andhra Pradesh, India. The dataset spans 60 days from January to March 2022, with measurements taken at 12 noon every day. The models were trained using Ambient temperature, Relative humidity, Solar irradiation, Panel temperature, Voltage, and Current as input variables and predicted PV energy produced in the output. Compared to other models, SVR has appeared as a superior model with the highest value of R^2 and the lowest values of MAE and MSE.

From the extensive review of recently published research, authors have found that most investigations are done on limited and short-term datasets [22–24,26–29]. Although the trained models have shown good performance, possibilities of long-term seasonal variability, inter-annual climatic fluctuations, and seasonal to annual shifts in irradiation patterns cannot be ignored. Limited and short-term datasets may reduce the generalizability of the regression models. Some of the studies are solely focused on the prediction of irradiation parameters rather than the direct prediction of PV energy [23,25,29]. Though it offers value addition, there is a need for direct prediction of PV energy, which may help to achieve better real-world energy management and grid stability solutions. The present scenario also demands a sufficiently structured benchmarking tool, which is necessary to evaluate the performance of high-end, complex, and computationally intensive ML models.

2. Novelty and contribution of the research

The current research aims to compare five ML models used in the prediction task of PV energy. The identified research gap is addressed in this research using a long-term dataset obtained from a rooftop PV plant found in Albany, Western Australia. The dataset consists of recorded observations of 25 years, from 1990 to 2014. It has hourly observations of solar irradiance, such as Global-horizontal irradiance (GHI), Direct-normal irradiance (DNI), Diffused-horizontal irradiance (DHI), Wet-bulb temperature (WBT), Dew-point temperature (DPT), and PV energy produced. The dataset presents a region that is rich in a temperate maritime climate and provides insights into the coastal atmospheric conditions. This introduces a unique variability in data patterns that are not extensively studied in existing literature. With this dataset, the authors have given sufficient attention to the shortcomings occurring in the correct prediction of PV energy due to influential factors such as variable seasonal cycles, extreme weather variation across more than two decades, and long-term variation in irradiation trends. Additionally, the combination of irradiation-related and weather-related observations will lead the models towards better generalization. The models developed in this research are, namely, LR, k-NN, SVR, DT, and RF. The models are assessed by performing comparative analysis using four key statistical parameters, which are RMSE, MSE, MAE, and R^2.

3. Methodology

To develop the ML models, an organized, structured flow is adopted, as shown in Figure 1. The study begins with buying the dataset. It was downloaded from the Mendeley data repository (https://data.mendeley.com/datasets/c4rn7mtfrf/1) [30]. As previously mentioned, the dataset is of the rooftop PV plant found in Albany, Western Australia. Initially, the dataset is loaded using the panda's library. The informative statistics of the dataset are obtained using the shape, columns, and info () commands. There is a total of 95950 observations available initially in the dataset. Before training the models, the data cleaning and pre-processing steps are performed on the data. To get the information about missing values in the dataset, isnull (). sum () commands are used. No missing values were found in the dataset. Now, duplicated () and sum () commands are used to detect any duplicate rows in the dataset. It is found that the dataset does not have any duplicate rows. The columns, such as reading numbers, dates, time, etc., are dropped from the dataset using the drop command. head () command is used to see the refined columns in the dataset. The first five rows and all columns given by the head () command are shown in Table 1. The clean and structured dataset is now ready for further analysis. The refined dataset with hourly resolution now consists of columns such as GHI, DNI, DHI, WBT, DPT, and Energy. The target variable is set to Energy (kWh), and the remaining attributes are set to input variables. A summary of the input and output variables is shown in Table 2.

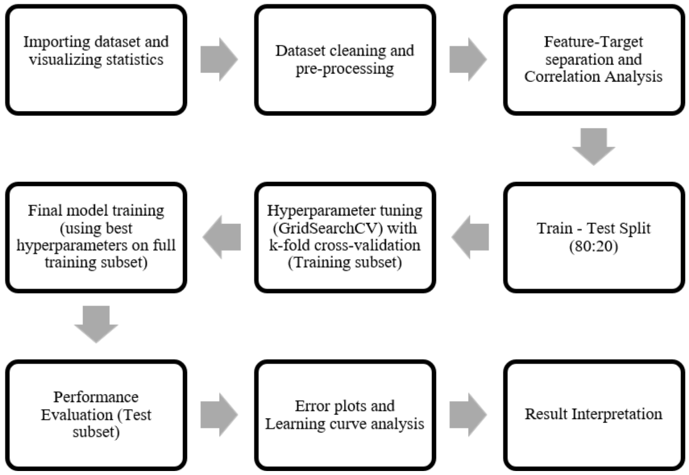


Figure 1. Methodology of research

Table 1. Representative sample of the dataset showing the first five rows and all feature columns

Index	GHI	DNI	DHI	Wet bulb temp	Dew point temp	Energy
Unit	**W/m²**	**W/m²**	**W/m²**	**°C**	**°C**	**kWh**
0	107	19	102	14.5594	13.6	12.646700
1	481	28	454	17.1797	13.9	21.657320
2	358	12	347	16.8141	13.6	11.432791
3	274	10	266	15.9516	13.1	9.527326
4	261	19	248	15.9078	13.1	11.938500

Table 2. Descriptive statistics of variables used in machine learning model development

Parameter Name	Mean Value	Std. Deviation Value	Minimum Value	25th Percentile	Median (50th)	75th Percentile	Maximum Value
GHI (W/m²)	397.44	268.59	0.00	183.48	355.00	602.00	1128.00
DNI (W/m²)	371.62	321.46	0.00	73.00	283.00	583.00	1054.00
DHI (W/m²)	168.92	115.23	0.00	79.00	141.00	240.00	603.00
Wet-Bulb Temp (°C)	13.23	3.03	1.19	13.10	13.10	15.35	24.86
Dew-Point Temp (°C)	10.30	3.49	-12.00	7.98	10.28	12.80	24.00
Energy (kWh)	352.23	253.59	0.95	79.00	258.60	628.49	1015.97

The dataset split is performed with an 80:20 ratio of training to testing. The regular training subset is used to train LR, k-NN, DT, and RF, while a representative training subset is used for SVR. The test subset is exclusively kept for assessing the performance of models. Except for LR, which is assumed to be the baseline model, hyperparameter tuning is employed for all models using GridSearchCV. This approach yields the best hyperparameter configuration. To generalize the models, the k-fold cross-validation technique is applied during training. In this technique, within the training subset, the data is divided into 'k' subsets. 'k' subsets are employed for training, and 'k-1' subsets are used for validation. The value of 'k' is five for LR, k-NN, DT, and RF, while three for SVR. For k-NN and SVR, Feature scaling was performed on the training set using the standardization technique through the StandardScaler function. The scaling was performed before cross-validation and not independently within each fold. Table 3 shows the details of the hyperparameters and cross-validation settings used during the training of the models. The performance of the model is calculated using four performance metrics, namely, MAE, MSE, RMSE, and R^2. Equations used to evaluate performance metrics are given numbers (1), (2), (3), and (4), respectively [23].

$$MAE = \left(\frac{1}{n}\right)\sum_{i=1}^{n} |y_i - \bar{y}_i| \tag{1}$$

$$MSE = \left(\frac{1}{n}\right)\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

$$RMSE = \sqrt{\left[\left(\frac{1}{n}\right)\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right]} \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{4}$$

Table 3. Hyperparameter tuning and cross-validation strategy to select optimal parameters of regression models

Model	Hyperparameters Considered	Cross-Validation Strategy	Best / Final Parameters
LR	None	5-Fold	Default setting
k-NN	n_neighbors: 3, 5, 7, 9, 11 Weights: 'uniform', 'distance' Metric: 'euclidean', 'manhattan'	5-Fold	n_neighbors = 11 weights = 'distance' metric = 'manhattan'
SVR	Kernel: 'rbf' Regularization parameter (C): 1, 10 Gamma: 'scale' Epsilon: 0.1	3-Fold	kernel = 'rbf' C = 10 gamma = 'scale' epsilon = 0.1
DT	max_depth min_samples_leaf min_samples_split	5-Fold	max_depth = 10 min_samples_leaf = 10 min_samples_split = 2
RF	n_estimators max_depth min_samples_split min_samples_leaf	5-Fold	n_estimators = 100 max_depth = None min_samples_leaf = 5 min_samples_split = 2

4. Result and discussion

4.1. Correlation of features

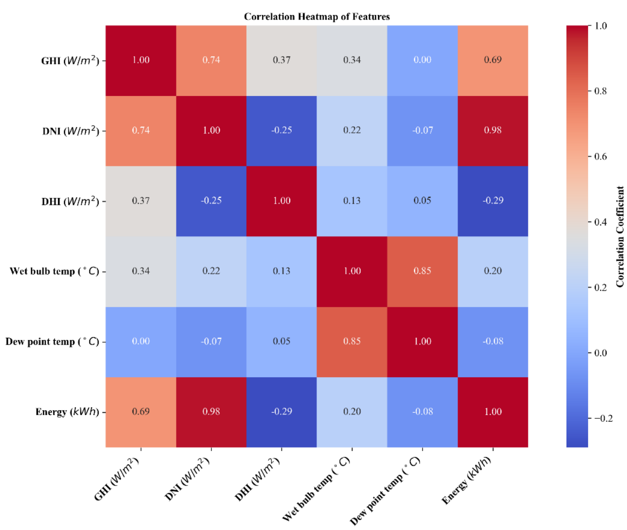
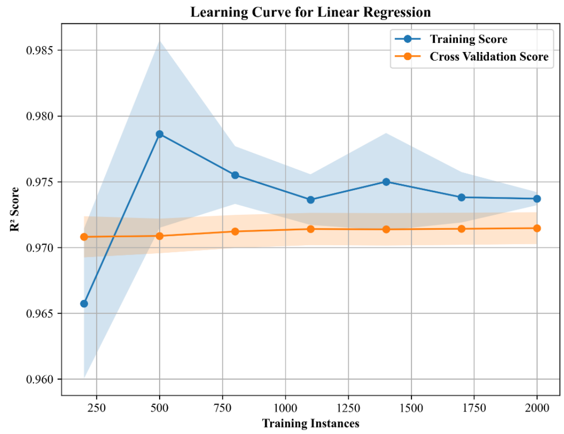


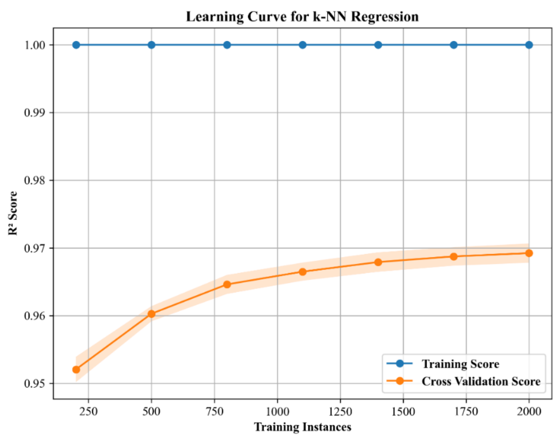
Figure 2. Correlation matrix of input parameters and target variable (Energy in kWh), with numerical Pearson correlation coefficient value shown inside each cell and a color bar showing correlation strength

The heatmap of the correlation matrix is shown in Figure 2. It is the linear relationships of different input features amongst the input and the target variable. The color scale is the values of Pearson's Correlation Coefficient, ranging from -1 to +1, where dark red shades show strong positive correlations and navy-blue shades denote negative correlations. A value must be positive when the target variable is directly proportional to the feature, and negative if having inverse proportion. Among the input features, DNI and GHI show strong positive correlations with coefficient values of 0.98 and 0.69, respectively. DNI shows a very strong positive correlation, showing that small variations in DNI significantly affect the target variable. GHI exhibits a moderate to strong positive correlation. DNI is solar radiation that falls on a normal surface to the solar rays. GHI is solar radiation that falls on a horizontal surface. DNI and GHI values are on the larger side when the sky is clean and clear. Under such conditions, solar radiation reaches the PV surface with minimal atmospheric scattering, and hence the PV energy output is maximized. In contrast, DHI exhibits a negative correlation (-0.29) with the target variable, showing an inverse relationship. Higher DHI values show solar radiation scattered by clouds, dust particles, and water vapor, etc., before reaching the PV surface. This reduces PV energy generation. WBT is a combined indicator based on the temperature as well as the humidity. It is the temperature of the air obtained with water evaporation at constant atmospheric pressure. Higher WBT values are attributed to the humid and warm environmental conditions. A mild positive correlation coefficient (0.20) is seen for WBT, which shows that it is in an indirect relationship with moisture and ambient thermal conditions. Dew-point temperature is the temperature at which atmospheric moisture begins to condense. Higher values of DPT indicate a humid environment. The correlation analysis shows that dew-point temperature negatively affects photovoltaic energy output. This is because reduced atmospheric clarity slightly decreases solar transmittance, and hence, PV energy output reduces. The correlation feature analysis supports the underlying scientific phenomena associated with the selected meteorological variables and offers valuable insights into the fact that consideration of both environmental and atmospheric parameters is important.

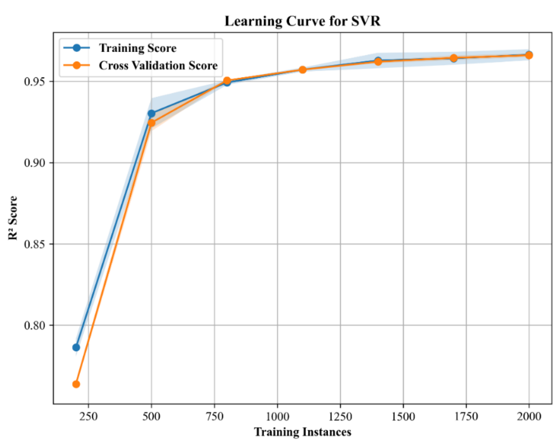
4.2. Learning curve analysis



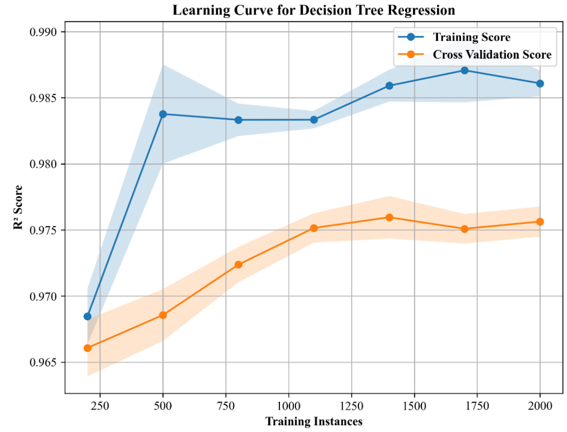
(a)



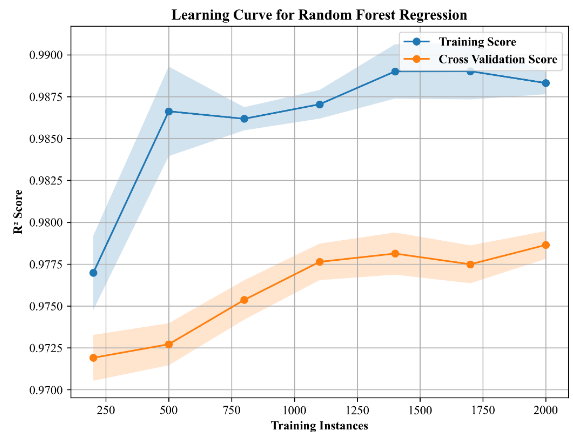
(b)



(c)



(d)



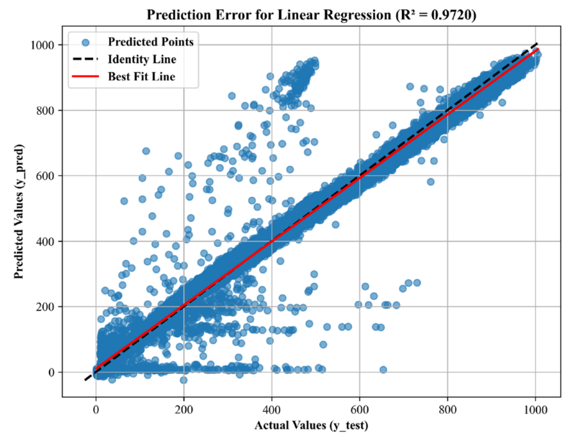
(e)

Figure 3. Learning curves analysis showing training and cross-validation performance lines for five regression algorithms: **(a)** LR, **(b)** k-NN, **(c)** DT, **(d)** SVR, and **(e)** RF

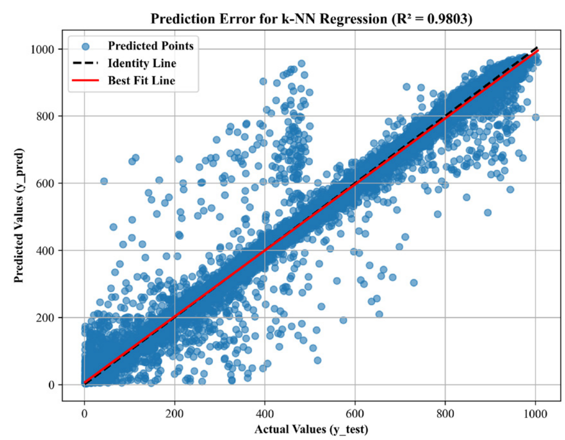
The set of learning curve plots presented in Figure 3 is the training and cross-validation performance of five supervised ML models. The curves show the variation in the R^2 as the number of training instances increases, providing insights into the learning stability of each model, generalization ability, and bias–variance characteristics. In the baseline model LR, as shown in Figure 3 (a), both the training and cross-validation curves converge rapidly as the training instances increase. As the gap between the curves is narrow, it shows low variance, but slight plateauing shows that the model reaches its performance too early. Therefore, moderate bias is present, which limits the model to capture nonlinearity between the variables. The steady rise in the cross-validation score is seen for k-NN against the number of training instances. At lower training instances, a clear gap is visible between the curves, as shown in Figure 3 (b), which shows overfitting. This is attributed to the inherent nature of k-NN. But as the number of training instances increased, the curve reached stability. This shows improvement in the generalization. From the learning curve analysis of SVR as shown in Figure 3 (c), an improvement in both training and validation scores is seen as soon as the training instances are increased. Both the curves converge beyond 800 sample instances, showing a balanced bias-variance characteristic. The DT model, on the other hand, as shown in Figure 3 (d), shows better performance across all training instances. The first gap between the curves at a lower number of training instances shows minute overfitting. However, as soon as the training instances are increased, the gap narrows and validation performance increases. Some variance is still visible, which shows that some non-linearity still needs to be addressed. The learning performance of RF, as shown in Figure 3 (e), shows improvement over DT. It is seen that the training score stays high, yet the cross-validation score steadily improves with increasing training instances. A smaller gap can be seen compared to DT. The convergence of the two curves at large sample instances shows better generalization compared to DT. This improvement is obvious due to the advantage of ensemble averaging in RF compared to DT. The consistent upward trend of the validation curve shows that the model is effectively capturing the non-linear complexities in the variables without substantial overfitting. Overall, RF shows stable convergence among all other models, which shows a good bias-variance trade-off.

4.3. Regression analysis and model fitting

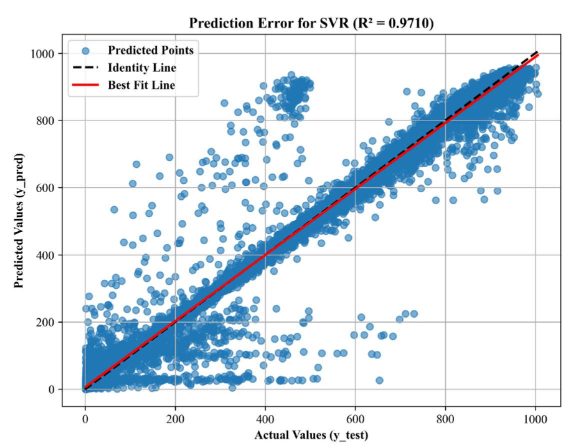
Error plots are used to check the agreement between the actual and predicted target variable. In all error plots, as shown in Figure 4 (a) to (e), the reference line is shown by black dash line. It shows an ideal prediction condition. The red line is the best-fit regression line. Points close to the best-fit regression line represent the better prediction accuracy, while points farther away show the error variance. As shown in Figure 4 (a), the LR model shows a near-perfect fit; however, noticeable dispersion is seen at lower and mid-range energy values. Several points deviating from the regression line show underfitting in some regions and overfitting in some regions. A small amount of bias is also detected from the deviation of the two lines. This is obvious since LR assumes a linear relationship between the variables under consideration. LR does not assume non-linearity, and hence it generalizes weakly. In the k-NN error plot, as shown in Figure 4 (b), the points are closely distributed around the regression line. Less dispersion is visible around the mid-range values. Overall alignment of the ideal line with the best-fit line is improved consistently; some points stay scattered. k-NN generalizes well compared to LR, showcasing its ability to capture certain non-linearities in the dataset. Slightly higher dispersion is visible in SVR, as shown in Figure 4 (c). Deviation is greater in the mid-range values. Due to its sensitive kernel, using a subset instead of the actual dataset and hyperparameter selection makes the model generalize less compared to LR. Though there is less dispersion in DT compared to LR and SVR, some visible dispersion is present at mid-range values. At higher energy values, as shown in Figure 4 (d), the ideal line almost coincides with the best-fit line. Though the DT model predicted the target variable correctly, there is scope for improvement, especially in the mid-range. The DT model typically captures non-linearity by assuming a single tree structure. This can be further improved by taking advantage of the ensemble method, which is implemented in the RF model. In the RF model error plot, as shown in Figure 4 (e), tight clustering is seen over the entire range of values. The ideal line almost overlaps the best-fit line, which shows low bias and variance. RF is superior in capturing non-linearity compared to DT due to its ensemble averaging principle.



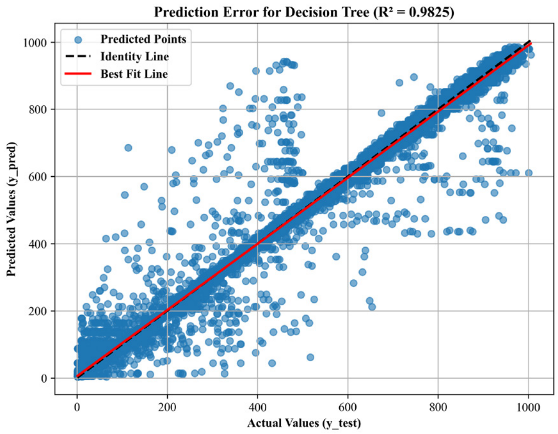
(a)



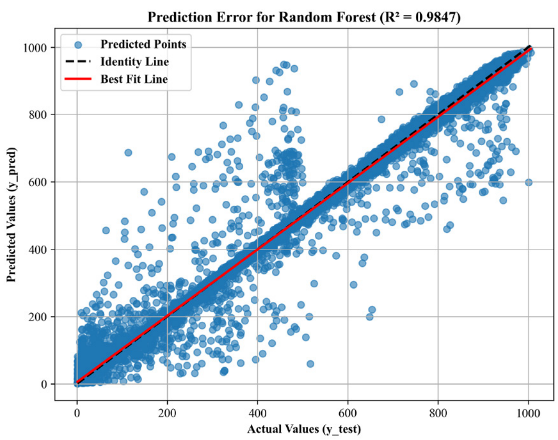
(b)



(c)



(d)



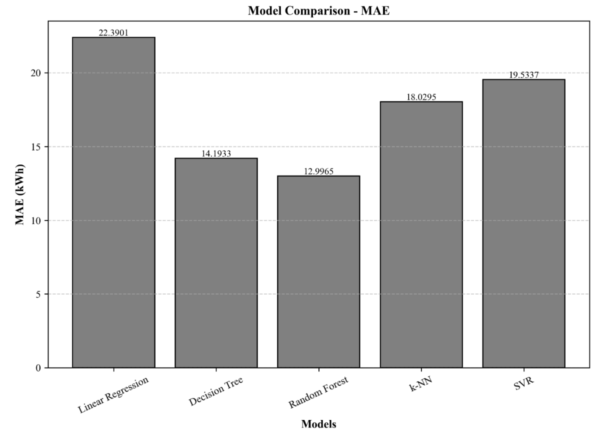
(e)

Figure 4. Predicted versus actual photovoltaic energy: **(a)** for LR, **(b)** for k-NN, **(c)** for DT, **(d)** for RF, and **(e)** for SVR

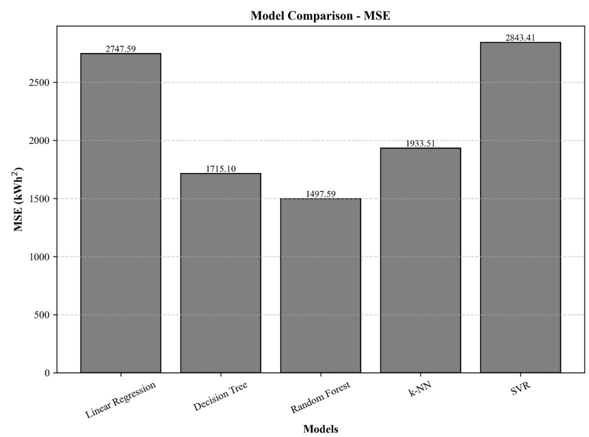
4.4. Statistical performance evaluation and comparative analysis

Table 4. Performance statistics of Machine Learning Models

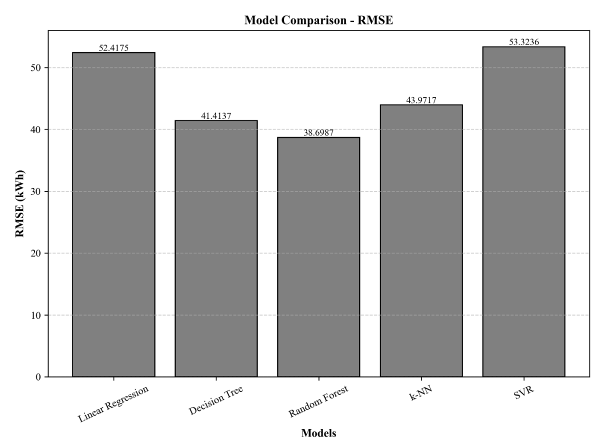
Name of ML Model	MAE value (kWh)	MSE value (kWh^2)	RMSE value (kWh)	R^2 Score
LR	22.3901	2747.59	52.4175	0.9720
DT	14.1933	1715.10	41.4137	0.9825
RF	12.9965	1497.59	38.6987	0.9847
k-NN	18.0295	1933.51	43.9717	0.9803
SVR	19.5337	2843.41	53.3236	0.9710



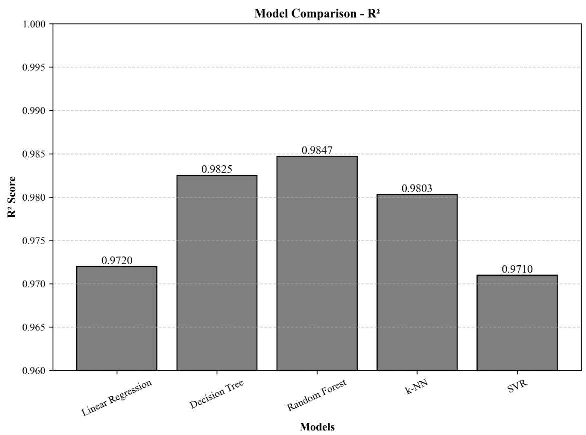
(a)



(b)



(c)



(d)

Figure 5. Comparative performance statistics: (a) MAE, (b) MSE, (c) RMSE, and (d) R^2

Table 4 presents a combined summary of the final test performance metrics (MAE, MSE, RMSE, and R^2) for all the developed machine learning models. Figures 5(a) to (d) present a comparative assessment of these four statistical performance metrics. The RF model has reached the lowest MAE value, which is 12.9965 kWh. The MAE values for DT, k-NN, and SVR are 14.1933 kWh, 18.0295 kWh, and 19.5337 kWh, respectively. The highest value of MAE is 22.3901 kWh, which is obtained for the LR model. With LR as the baseline model, RF has reduced MAE by 41.95 %. The performance of RF is superior to others due to its ability to aggregate multiple decision trees. With this ensemble approach, variance is reduced, and the model generalizes well. On the other hand, LR assumes a linear relationship, which can't predict the non-linearity between the input and target variables. Figure 5 (b) shows the comparison of the MSE values for all the models. Once again, RF has appeared as the superior amongst all. It produced the lowest MSE value, which is around 1497.59 kWh^2. This confirms that RF is performing firmly in curtailing squared prediction errors. Followed by RF, DT, k-NN, and LR have performed better with MSE values of 1715.10 kWh^2, 1933.51 kWh^2, and 2747.59 kWh^2, respectively. SVR has the highest MSE, which is 2843.41 kWh^2. This shows a weak generalization. Though the SVR is better at handling non-linearity than LR, its performance depends upon the proper choice of kernel type, regularization parameter, dataset complexity, and epsilon margin. On the other hand, LR assumes a straight linear relationship, which is sometimes beneficial. RMSE values are an extension of MSE, as shown in Figure 5 (c), which further strengthens the results. RF has the lowest RMSE value, which is 38.6987 kWh. RMSE values for DT, k-NN, and LR are 41.4137 kWh, 43.9717 kWh, and 52.4175 kWh, respectively. SVR has the highest RMSE value, which is 53.3236 kWh. Figure 5 (d) highlights the R^2 score. It shows the model's goodness of fit. An R^2 value close to 1 is preferred because it shows perfect prediction. RF reached the highest R^2 value, which is 0.9847. R^2 values for DT, k-NN, LR, and SVR are 0.9825, 0.9803, 0.9720, and 0.9710, respectively.

Overall, the entire discussion of statistical parameters shows that RF consistently outperformed the other models across all evaluations, offering the best balance of accuracy, stability, and generalization. RF can couple several individual decision trees, which leads to an increase in the generalizability of the model and so accurately. RF was able to effectively predict nonlinearity between the input variables and the output. RF was also seen as less sensitive towards the noise and multicollinearity of the dataset than other machine learning models. Its ability to generalize well on unseen data resulted in lower error values and higher prediction accuracy compared to other machine learning models.

The result shows good alignment with literature. In the present study, RF has outperformed the baseline LR model. A similar kind of dominance is also reported in many recent studies [19,23,24,31]. Recent literature has already confirmed that the ensemble-based learning approach of RF picks the non-linearity in the data more effectively than other single learner models, and current research agrees with it. In the present study, RF has also outperformed SVR, which was also reported recently by Bouakkaz et al. [19]. One of the important studies by Tandon et al. [23] has reported a 19.47% improvement in performance with RF. In comparison, the current study has reported a 26.17% improvement. In the study by Idogho et al [22]. The best performing model, k-NN, has reported R^2 score of 0.95, while in the current study, the k-NN model has achieved R^2 score of 0.98, which is 3.15% greater. A study by Tripathi et al. [28] proved that the SVR model has achieved an R^2 value of 0.99, while in the current study, it is 0.9710, which is lower. This discrepancy may be attributed to dataset characteristics, hyperparameter settings, kernel choice, and epsilon margin. In addition, this also underlines the fact that the prediction of PV energy is a challenging task and needs to be handled effectively in future studies.

5. Limitation

The present research is focused on developing ML models that are trained, tested, and confirmed on a dataset that was obtained from a single-location installation facility. This may not generalize well with the other geographic regions. The data set lacks a few added factors, such as the orientation of the panel, shading effect, dust accumulation, maintenance practices, etc. This may also add some value to the prediction accuracy. The current study does not incorporate training temporal models, such as LSTM and other advanced deep learning approaches that can capture time-dependent complex non-linear patterns amongst the data points and autocorrelation structures. This may further improve the accuracy of prediction.

6. Conclusion

The current study is focused on the comparative analysis of five regression models that are developed to predict photovoltaic energy. Unlike many earlier studies that predict irradiation parameters, this research predicts photovoltaic energy based on the irradiation and environmental parameters. This study also addresses the identified research gap of using a short-term dataset for model training by incorporating a long-term 25-year dataset of a rooftop PV plant found in Albany, Western Australia. The adopted dataset consists of 95950 hourly instances in total, which imposes diverse climatic and seasonal conditions during the training of the models. Five supervised regression models trained are Linear Regression, k-Nearest Neighbor, Support Vector Regression, Decision Tree, and Random Forest. Performance metrics such as Mean Absolute Error, Mean-Squared Error, Root Mean-Squared Error, and Coefficient of Determination are calculated to assess the models. In addition, learning curve analysis and prediction error plots are employed to see the bias-variance characteristics and the behavior of the models in generalization. In the context of the prediction of photovoltaic energy, this study can be treated as a dependable approach. ML models are far better than the traditional numerical and statistical models.

The comparative study shows that the ensemble method implemented in the Random Forest model shows superior performance. But the Support Vector Regression and Linear Regression models show weak generalization. Compared to Support Vector Regression and Linear Regression, Random Forest has been able to reduce the Root Mean-Squared Error by approximately 27%. Prediction of Mean Absolute Error is 33% lower for Random Forest than Support Vector Regression and 42% lower compared to Linear Regression. The highest value of the Coefficient of Determination obtained for Random Forest shows its superiority in prediction. k-Nearest Neighbor and Decision Tree models have also shown better performance than Support Vector Regression and Linear Regression, but do not generalize, in contrast to Random Forest. Both models have lower variance than Support Vector Regression and Linear Regression, but higher than Random Forest.

From the correlation analysis, it is found that Direct Normal Irradiance has the highest impact on the target variable compared to other input variables. It shows 98% positive strength of linear association with the target variable. The positive strength of linear association is 69% and 20% for Global Horizontal Irradiance and Wet Bulb Temperature, respectively. Diffused horizontal irradiance and Dew-point temperature show inverse strength of linear association of 29% and 8%, respectively. This analysis may act as a foundation for PV system engineers and researchers to better plan their scheduling activities. It is advisable that close monitoring of the Direct Normal Irradiance will improve the efficiency of energy production. Forecasting the accuracy of Direct Normal Irradiance will help to solve the issues related to energy planning, delivery management, and grid stability. On the other hand, prediction accuracy also depends on environmental factors. Overall, future research work should not consider irradiance parameters alone to improve prediction accuracy, but the combination of irradiance and environmental parameters.

Acknowledgement

The authors express sincere gratitude to the School of Technology Management and Engineering, SVKM NMIMS Global University, Dhule, Maharashtra, India, for providing the necessary infrastructure and academic environment to successfully carry out this research work. The authors sincerely acknowledge Dr. Hitesh Thakare, Head of the Department, and Dr. Nilesh Salunke, Principal, for their constant support, motivation, and continuous encouragement throughout the course of this research. The constructive comments and insightful suggestions provided by the reviewers and editors are gratefully acknowledged, as they significantly contributed to improving the quality and clarity of this manuscript.

Abbreviations

n	Number of observations
y_i	Actual value
$\hat{y}_i$	Predicted value
$\bar{y}_i, \bar{y}$	Means of values
$(y_i - \hat{y}_i)^2$	Squared error
MAE	Mean Absolute Error
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
R^2	Coefficient of Determination
ML	Machine Learning
LR	Linear Regression
DT	Decision Tree
SVR	Support Vector Regression
k-NN	k-Nearest Neighbor
RF	Random Forest
GHI	Global Horizontal Irradiance
DNI	Direct Normal Irradiance
DHI	Diffuse Horizontal Irradiance
PV	Photovoltaic
kWh	Kilowatt-hour; 1 kWh = 3.6×10^6 joules
MW	Megawatt; 1 MW = 1000 kW
GW	Gigawatt; 1 GW = 10^6 kW

Author contributions

Bhushan Behede: Conceptualization, data acquisition, model development, writing – Original draft. Makarand Shahade: Methodology design, project supervision, project administration. Siddharth Chakrabarti: Technical guidance, method review, understanding reviewer comments, and restructuring the manuscript accordingly. Uday Wankhede: Statistical analysis support, interpretation of results, understanding reviewer comments, and restructuring the manuscript accordingly. Yogesh Sonawane: Data preprocessing, aiding in literature review and gap identification. Dattatray Doifode: Analysis support, data visualization, result processing, editing manuscript. Rahul Patil: Analysis support, model validation, performance evaluation.

Declaration of interests

The authors declare that they have no competing interests.

Data availability statement

The dataset used in the present research work is open source and free to use for everyone. It can be accessed from the data repository of Mendeley. Link: https://data.mendeley.com/datasets/c4rn7mtfrf/1 (DOI: 10.17632/c4rn7mtfrf.1) [30].

Declaration of generative AI and AI-assisted technologies in the writing process

The authors used Grammarly during the preparation of this manuscript solely to assist with spelling, grammar, punctuation, and language refinement.

References

[1] S. Zare, A.R. Tavakolpour-Saleh, T. Binazadeh, Analytical investigation of free piston Stirling engines using practical stability method, Chaos Solitons Fractals 167 (2023) 113082. https://doi.org/10.1016/J.CHAOS.2022.113082.

[2] S. Zare, F. Pourfayaz, A.R. Tavakolpour-Saleh, R.A. Lashaki, A design method based on neural network to predict thermoacoustic Stirling engine parameters: Experimental and theoretical assessment, Energy 309 (2024) 133113. https://doi.org/10.1016/J.ENERGY.2024.133113.

[3] S. Zare, Z. Raeisi, R.A. Lashaki, M. Makki, N. Ghasemi, Evaluation of a thermoacoustic Stirling oscillator using a describing function and a genetic algorithm, Appl. Therm. Eng. 278 (2025) 127125. https://doi.org/10.1016/J.APPLTHERMALENG.2025.127125.

[4] K. Sunil kumar, R. Surakasi, G. kumar, S. Sudha, G. D'mello, B.T. Tharanisrisakthi, Recent trends in the development and growth of solar technologies: a review, Journal of Thermal Engineering 11 (2025) 1231–1244. https://doi.org/10.14744/thermal.0000973.

[5] P. Bhat, A.S. Iyengar, N. Abhilash, P.K. Reddy, Experimental investigation and validation of solar PV cooling for enhanced energy conversion efficiency for Indian climatic conditions, Journal of Thermal Engineering 8 (2022) 711–718. https://doi.org/10.18186/thermal.1195570.

[6] S.H. Pervez, M.A. Kamran, Sallahudin, A. Ahad, M.A.Z. Khan, M. Faiq, Development and performance analysis of hybrid photovoltaic/thermal (PV/T) system, Journal of Thermal Engineering 7 (2021) 1936–1944. https://doi.org/10.18186/THERMAL.1051272.

[7] T.S.D. Singh, B.A. Shimray, S.N. Meitei, Performance Analysis of a Rooftop Grid-Connected Photovoltaic System in North-Eastern India, Manipur, Energies (Basel). 18 (2025). https://doi.org/10.3390/en18081921.

[8] R.K. Donga, S. Kumar, A. Kumar, Performance evaluation of parabolic trough collector with receiver position error, Yildiz Technical University Press, 2021.

[9] Niti ayog report, (n.d.).

[10] R. Dehini, B. Berbaoui, Solar energy control and power quality improvement using multilayer feed forward neural network, Yildiz Technical University Press, 2018.

[11] R.D. Das Rao, B.S. Gowda, Technical and economic feasibility analysis of 1kW Rooftop Solar Photovoltaic System for Mysuru, Karnataka, India, Journal of Thermal Engineering 11 (2025) 1507–1519. https://doi.org/10.14744/thermal.0000999.

[12] D. Sethi, P.G. Kosmopoulos, Rooftop Solar Photovoltaic Potential in Polluted Indian Cities: Atmospheric and Urban Impacts, Climate Trends, Societal Gains, and Economic Opportunities, Remote Sens. (Basel). 17 (2025). https://doi.org/10.3390/rs17071221.

[13] P. Di Leo, A. Ciocia, G. Malgaroli, F. Spertino, Advancements and Challenges in Photovoltaic Power Forecasting: A Comprehensive Review, Energies (Basel). 18 (2025). https://doi.org/10.3390/en18082108.

[14] S.J. Plathottam, A. Rzonca, R. Lakhnori, C.O. Iloeje, A review of artificial intelligence applications in manufacturing operations, J. Adv. Manuf. Process. 5 (2023). https://doi.org/10.1002/amp2.10159.

[15] V. Mohanasundaram, B. Rangaswamy, Photovoltaic solar energy prediction using the seasonal-trend decomposition layer and ASOA optimized LSTM neural network model, Sci. Rep. 15 (2025). https://doi.org/10.1038/s41598-025-87625-0.

[16] A. Sardarabadi, A. Heydarian Ardakani, S. Matrone, E. Ogliari, E. Shirazi, Multi-temporal PV power prediction using long short-term memory and wavelet packet decomposition, Energy and AI 21 (2025). https://doi.org/10.1016/j.egyai.2025.100540.

[17] H.N. Nguyen, Q.T. Tran, C.T. Ngo, D.D. Nguyen, V.Q. Tran, Solar energy prediction through machine learning models: A comparative analysis of regressor algorithms, PLoS One 20 (2025). https://doi.org/10.1371/journal.pone.0315955.

[18] M. Abdelsattar, M.A. Ismeil, M.M.A. Azim Zayed, A. Abdelmoety, A. Emad-Eldeen, Assessing Machine Learning Approaches for Photovoltaic Energy Prediction in Sustainable Energy Systems, IEEE Access 12 (2024) 107599–107615. https://doi.org/10.1109/ACCESS.2024.3437191.

[19] A. Bouakkaz, A. Lahsasna, A.J.G. Mena, S. Haddad, M.L. Ferrari, R. Jiménez-Castaneda, Evaluating and analyzing the performance of PV power output forecasting using different models of machine-learning techniques considering prediction accuracy, International Journal of Renewable Energy Development 14 (2025) 158–167. https://doi.org/10.61435/ijred.2025.60547.

[20] S.A. Hamad, M.A. Ghalib, A. Munshi, M. Alotaibi, M.A. Ebied, Evaluating machine learning models comprehensively for predicting maximum power from photovoltaic systems, Sci. Rep. 15 (2025). https://doi.org/10.1038/s41598-025-91044-6.

[21] N. Saxena, R. Kumar, Y.K.S.S. Rao, D.S. Mondloe, N.K. Dhapekar, A. Sharma, A.S. Yadav, Hybrid KNN-SVM machine learning approach for solar power forecasting, Environmental Challenges 14 (2024). https://doi.org/10.1016/j.envc.2024.100838.

[22] C. Idogho, E.O. Abah, J.O. Onuhc, C. Harsito, K. Omenkaf, A. Samuel, A. Ejila, I.P. Idoko, U.E. Ali, Machine Learning-Based Solar Photovoltaic Power Forecasting for Nigerian Regions, Energy Sci. Eng. 13 (2025) 1922–1934. https://doi.org/10.1002/ese3.70013.

[23] A. Tandon, A. Awasthi, K.C. Pattnayak, A. Tandon, T. Choudhury, K. Kotecha, Machine learning-driven solar irradiance prediction: advancing renewable energy in Rajasthan, Discover Applied Sciences 7 (2025). https://doi.org/10.1007/s42452-025-06490-8.

[24] M.H. Sulaiman, M.S. Jadin, Z. Mustaffa, M.N. Mohd Azlan, H. Daniyal, Short-Term forecasting of floating photovoltaic power generation using machine learning models, Cleaner Energy Systems 9 (2024). https://doi.org/10.1016/j.cles.2024.100137.

[25] R. Gupta, A.K. Yadav, S.K. Jha, Prediction of Global Horizontal Irradiance Using an Explainable Data Driven Machine Learning Algorithms, Electric Power Components and Systems (2024). https://doi.org/10.1080/15325008.2024.2310771.

[26] S. Kumar Panda, M. Kumar Panda, Application of Machine Learning for the Prediction of Solar Power Generation, Cureus Journal of Computer Science (2025). https://doi.org/10.7759/s44389-024.

[27] A.K. Chaaban, N. Alfadl, A comparative study of machine learning approaches for an accurate predictive modeling of solar energy generation, Energy Reports 12 (2024) 1293–1302. https://doi.org/10.1016/j.egyr.2024.07.010.

[28] A.K. Tripathi, M. Aruna, P. V. Elumalai, K. Karthik, S.A. Khan, M. Asif, K.S. Rao, Advancing solar PV panel power prediction: A comparative machine learning approach in fluctuating environmental conditions, Case Studies in Thermal Engineering 59 (2024). https://doi.org/10.1016/j.csite.2024.104459.

[29] O.O. Apeh, N.I. Nwulu, Machine Learning Approach for Short- and Long-Term Global Solar Irradiance Prediction, Journal of Environmental and Earth Sciences 7 (2025) 321–342. https://doi.org/10.30564/jees.v7i1.7060.

[30] V. Beeravalli, Energy Prediction, (2023). https://doi.org/10.17632/c4rn7mtfrf.1.

[31] M. Abdelsattar, M.A. Ismeil, M.M.A. Azim Zayed, A. Abdelmoety, A. Emad-Eldeen, Assessing Machine Learning Approaches for Photovoltaic Energy Prediction in Sustainable Energy Systems, IEEE Access 12 (2024) 107599–107615. https://doi.org/10.1109/ACCESS.2024.3437191.