



OPEN Reliable and efficient solar radiation estimation with the insights of XAI

M. K. Nallakaruppan^{1,7}, Joji Johnson^{2,7}, Shrikant Mapari^{3,7}✉, Anupam Dasgupta^{4,7}, Swapnil Udgir^{5,7} & C. Narmatha^{6,7}

The unpredictability of solar energy has led to reliability and integration problems that require costly and technically complex solutions in the electrical grid. Solar resource availability and energy generation are highly influenced by local climate variables, like atmospheric temperature, humidity, wind and pressure. If there is generational uncertainty, it is challenging to calculate economic criteria such as energy costs and returns, which impacts the feasibility study of a solar power plant. Also, it is very difficult and costly to maintain pyranometers at the locations. To forecast solar irradiance and the power generation at any location, machine learning (ML) techniques can be used. The present work deals with determining the influence of different parameters in predicting the solar radiation by ML models. A comparison study of six regression models: Ada Boosting Regressor (ABR), Gradient Boosting Regressor (GBR), Random Forest Regressor (RFR), Decision Tree Regressor (DTR), Linear Regression (LR), and Extreme Gradient Boosting Regressor (XGBR), shows that RFR gave the highest regression score of 0.9028. This also recorded the Mean Absolute Error (MAE) of 0.6198 and Mean Squared Error (MSE) of 1.348. The next best regression score produced by the Gradient Boosting Regressor (GBR) with value of 0.891. This is 1.18% lower than the RFR. For the RFR regression analysis, an Explainable AI (XAI) model used to interpret the results using Local Interpretable Model-agnostic Explanations (LIME) for local surrogacy and Shapely for global surrogacy. Both the LIME and Shapely interpretations shows that the parameter temperature has the highest correlation with the radiation. The paper would benefit from a more explicit statement of what is new compared to prior studies.

For Earth, the radiation emitted by the Sun is the primary source of energy. The Earth is in an orbit with an average distance of 149 million kilometres from the Sun. Even though the Earth takes an elliptical path around the Sun, the solar radiation incident on the earth's outer atmosphere is relatively constant. The World Radiometric Centre (WRC) recommends a value of 1367 W/m², the Solar Constant, which corresponds to the total radiation incident per unit area per unit time on a surface kept perpendicular to sun's rays just outside the atmosphere¹. However, the radiation incident on the earth's surface at ground level varies drastically at any location. It is due to the effects of parameters like earth's axial tilt and rotation, angle of incidence and air mass, and atmospheric composition and climatic conditions². In addition, parameters like earth's temperature, wind speed, atmospheric pressure, relative humidity, and daily air temperature also affects the value of solar radiation incident³.

Conventionally, the solar radiation was predicted from the statistical data collected from various weather stations, satellite data and Geographic Information System (GIS)^{4,5}. Climate change in recent times due to global warming has led to deviations in solar radiation values predicted from these statistical data. Global warming has led formation of more aerosols and clouds in the atmosphere which contributes to significant reduction in solar radiation incident called solar dimming⁶. Solar dimming is adversely affecting the solar radiation prediction and it is proposed to use the ten most recent years data rather than the longest possible period to predict the radiation values⁷. Recent advances made in the solar photovoltaic technologies have led to steady increase with

¹Parul Institute of Computer Application, Parul University, Limda, Vadodara, Gujarat 391760, India. ²Department of Mechanical Engineering, Faculty of Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu 603203, India. ³Symbiosis Institute of Computer Studies and Research, Symbiosis International (Deemed University), Pune, India. ⁴School of Electronics and Communication Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India. ⁵Balji Institute of Modern Management, Sri Balaji University, Pune, Maharashtra 411033, India. ⁶Faculty of Computers and Information Technology, University of Tabuk, Tabuk, Kingdom of Saudi Arabia. ⁷M. K. Nallakaruppan, Joji Johnson, Shrikant Mapari, Anupam Dasgupta, Swapnil Udgir and C. Narmatha contributed equally to this work. ✉email: mshreek@gmail.com

1,281 GW installed capacity of photovoltaics throughout the world^{8,9}. Solar dimming is a serious concern for these installations as it will lead to lower power production at various sites which in turn negatively affects the economics of a solar power plants⁷. Also, as renewable energy's installed capacity rises globally, the site-specific, regional, and systemic changes in radiation availability will impact renewable energy system performance and economics¹⁰.

In addition, majority of solar forecasting techniques were developed for centralized solar power plants, which only impact a few sites. Due to the rapidly growing installation of PV systems, particularly distributed solar generating systems over large areas, solar forecasting and grid integration are encountering new challenges¹¹. Being closer to customers, having lower transmission loss, removing obstacles to investment, stabilizing the local grid, and providing flexible installation that maximizes land use are some advantages, which has led to more installation of grid tied distributed solar generating systems compared to centralised solar power plants¹². Poor solar forecast will also lead to significant expenses for scaling up or down, little adaptability to real-world operations, and complexity in optimising the installation parameters of distributed solar generating systems. The need of the hour for power system planning to coordinate, integrate, control, and oversee solar energy production in a wide region is accurate and reliable solar forecasting models^{13,14}. Comparative analysis of regression models are already available in the existing research. The novelty of the proposed work is the integration of XAI for interpretability which increases the reliability and trustworthiness of the solar radiation prediction.

Literature review

Machine learning (ML) algorithms use computational techniques can produce more accurate and reliable predictions for solar energy by capturing dynamic interactions between factors and adapting to changing conditions¹⁶. The best machine learning models have a balance between model complexity and accuracy^{15,16}. Feature selection helps determine the best input combination for prediction models; by removing irrelevant or redundant information and keeping the most crucial features. Feature selection can lower computing costs, improve overfitting issues, and solve multicollinearity issues in the models, feature selection can lower computing costs, improve over-fitting issues, and solve multicollinearity issues in the models^{17,18}. At present these models forecast sun radiance using either time series data or sky pictures using different network topologies. Conventionally, there are three types of solar irradiance forecasting models like image-based models¹⁹, hybrid models²⁰, and time series-based models^{21,22}. The three different kinds of solar irradiance forecasting models are all relatively short-term methods that include predicting irradiance levels for a short time frame, often a few seconds to a few minutes in advance²³. They are not appropriate for long-term forecasting since the variables influencing sun irradiance change with time²⁴. Time series models assess a clearness score based on the relationship between incident solar radiation and weather conditions. Time series models have the advantage that it can be easily trained using accurate climate data from weather stations and implemented to predict radiation at nearby locations with similar climatic conditions²⁵.

The overall framework for the solar radiation, the prediction and interpretable framework are represented in the process diagram represented in Fig. 1. The data acquisition is done from sun source and meteorological data sources. The data is pre-processed and applied to ML models. The ML regression scores are later used by XAI models for providing interpretability of four parameters Temperature, Humidity, Pressure and Wind speed.

ML techniques like Linear Regression (LR), Decision Tree Regressor (DTR), Random Forest Regressor (RFR), Gradient Boosting Regressor (GBR), Ada Boosting Regressor (ABR), Extreme Gradient Boosting Regressor (XGBR), Support Vector Machine (SVM), artificial neural network (ANN), Gaussian process (GP), etc are time series-based regression or artificial intelligence (AI) models used to predict solar radiation^{26,27}. ANN was used to determine the most important input parameters affecting solar radiation out of latitude, longitude, temperature, altitude, sunshine hours, maximum temperature, and minimum temperature for various Indian cities. Temperature, altitude, daylight hours, maximum temperature, and minimum temperature are the most important input variables, while latitude and longitude have the least impact on solar radiation prediction, as per

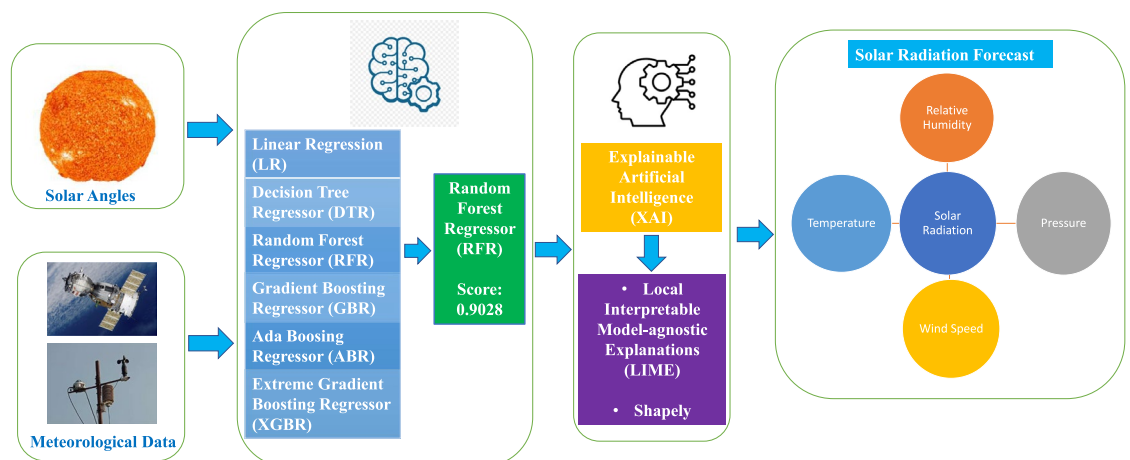


Fig. 1. Overall solar radiation framework.

WEKA analysis²⁸. Certain models show a positive correlation between the values and the Pearson correlation coefficient.

Prediction accuracy is lower for weakly correlated parameters and greater for highly correlated parameters as inputs in RFR, Extra Trees, Bagging, Decision Tree, and XGB²⁹. In both data-rich and data-poor settings, the use of regressive approaches can benefit by taking advantage of the correlative nature of the irradiance observations³⁰. Since large number of parameters leads to high computational requirements and complexity, it is imperative to reduce the number of parameters and identify the best parameters that influence the prediction. XAI is also used for predicting the Air Quality Index, which is a time series data³¹. In the Era of Industry 4.0–5.0 XAI plays vital role in all the industry sectors³². It is also used in autonomous transportation systems also³³.

This article attempts to find the influence of different independent climatic parameters on solar radiation by a correlation analysis on Six ML models. Linear Regression (LR), Decision Tree Regressor (DTR), Random Forest Regressor (RFR), Gradient Boosting Regressor (GBR), Ada Boosting Regressor (ABR), and Extreme Gradient Boosting Regressor (XGBR) were developed to evaluate and compare the solar radiation prediction. These models incorporate solar angles and atmospheric conditions (wind direction, humidity, wind speed and temperature) as input to predict solar radiation values. The LIME model was used to interpret the influence of the local surrogates, while the SHAPLY model was used to interpret the global surrogacy. The models were trained to forecast and predict from the data collected from a solar radiation resource. The comparative performance analysis of all models was done using metrics such as Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) to identify the variable selection strategies and best performing model.

Traditional shadow-box AI models often come up with good results without making it clear how those results were reached. The viability of solar energy as an investment cannot be determined by the consumers due to the lack of solar energy data at any location and the climate change. In addition, the lack of understanding about the ML's reasoning in these situations can make consumers not trust AI and make them hesitant to trust the choices and suggestions that are made by AI or ML systems. The Explainable Artificial Intelligence (XAI) platform is responsible for the development of AI models that produce results that are comprehensible to consumers and that contribute valuable information regarding the decisions made by AI systems. XAI facilitates the transparency, interpretability, and accountability of AI decisions^{34,35}. The interpretability of the intricate AI models used in significant wind energy applications, such as wind power forecasting, defect detection, predictive maintenance, wind farm optimization, and SCADA data analysis, is enhanced using XAI in wind energy systems³⁶. LIME based XAI was also successfully implemented for efficient monitoring and fault detection of solar photovoltaic panels³⁷. However, XAI has not been used in radiation prediction models. The novelty of the current article is the integration of ML regression scores by XAI models for providing interpretability of the parameters like Temperature, Humidity, Pressure and Windspeed to predict the solar radiation.

Methods

This section of the paper presents the system architecture, description of the dataset³⁸, models and methods used in the proposed work with the equations related to the solar radiation measurements.

Architecture

The architecture of our estimation model is designed to efficiently compute the data using machine learning models. The Fig. 1, illustrates how data processing works in our research by demonstrating each step in the process from input to output. The process involves collection of key parameters such as wind direction, speed, temperature and humidity. The data is then preprocessed to handle discrepancies in the form of missing values, outliers, and normalised, resulting in a clean and standardised input to train the model. Correlation analysis aids in feature selection, while domain expertise supports the development of additional relevant characteristics. Factors like time of day and solar angle capture the cyclic nature of solar radiation patterns.

The radiation experiences at a specific location is greatly influenced by the angle and position of the sun in the sky. Understanding these angles is critical for anticipating the pattern of solar radiation. We incorporated Ada Boost, Random Forest, Decision Tree, Gradient Boost and Extreme Gradient Boosting Regressor models for predicting solar radiation. The dataset is divided into sets to train and evaluate the model's performance. Upon training, our models are tested on a separate dataset to determine its real-world performance. The models are then evaluated on metrics such as Mean Square Error (MSE), R2-Score, and Mean Absolute Error (MAE). These metrics provide information about the model's accuracy and capacity to capture solar radiation fluctuation. After successful evaluation, the model is ready for deployment (Fig. 2).

Dataset

The features that are important to the prediction of the solar radiation are carefully selected from the dataset. There are 32686 instances of data for four attributes as independent variables such as Temperature, Pressure, Wind Speed, and Humidity. The target dependent attribute is temperature. The data types are numerical and they have relatively high correlation with the radiation. The data was recorded in the Moscow, Russia for 4 months. This data was presented in Space Apps Moscow was held on April 29th and 30th on 2017, where 175 people joined the International Space Apps Challenge at this location. There are no bias or missing values available in the dataset³⁹.

Random forest

The Random Forest Regressor model creates multiple decision trees and combines their predictions. Each decision tree is trained on a random subset of the data formed by taking random feature selections. The final

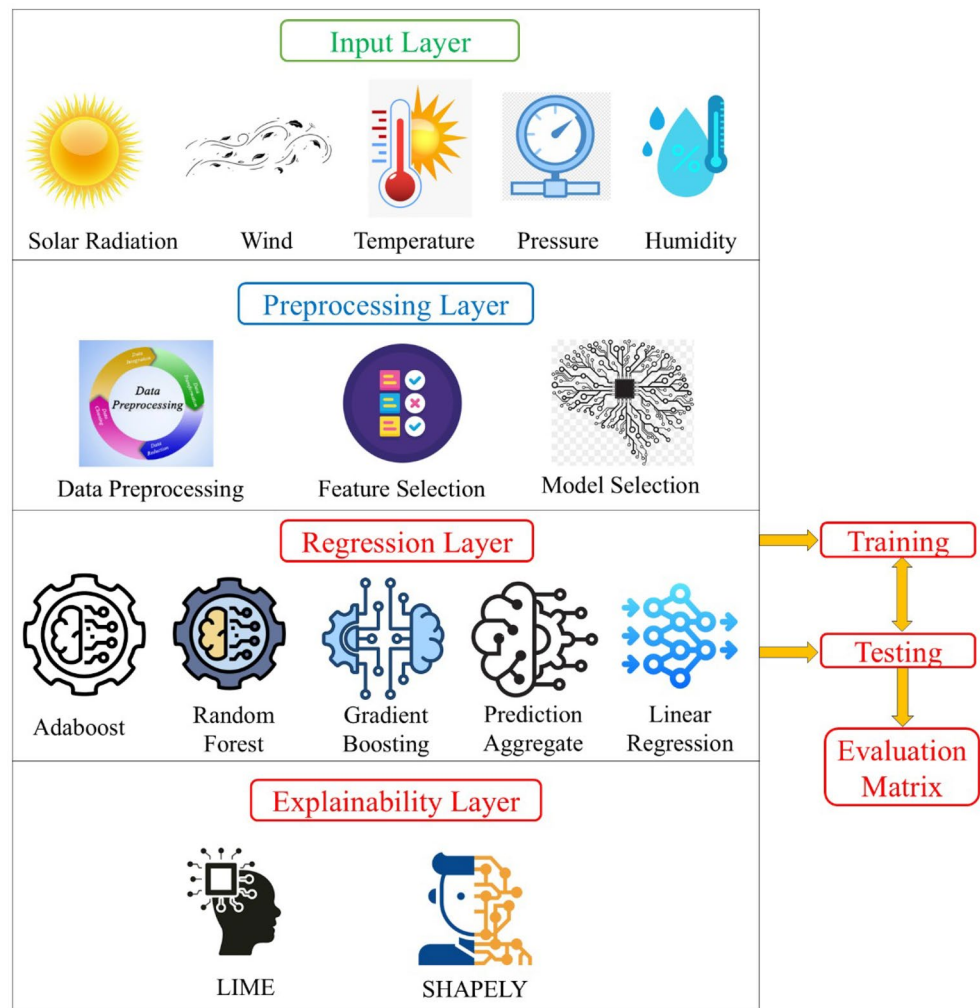


Fig. 2. Architecture flow diagram.

prediction is the average output of the decision trees. This regressor model captures non-linear relationships, making it ideal for handling high-dimensional data.

1. Training

- Bootstrap sample: $X_t, y_t = \text{bootstrap_sample}(X, y)$ This step involves randomly selecting features to create a subset of the original dataset. X_t and y_t
- Train decision tree: $h_t(x) = \text{train_decision_tree}(X_t, y_t)$ This step involves training a decision tree $h_t(x)$ on the bootstrap sample

2. Prediction Aggregate prediction:

$$H(x) = \frac{1}{T} \sum_{t=1}^T h_t(x) \quad (1)$$

where feature matrix (input data) with N samples and M features. y: target variable (solar radiation) with N samples. T: number of trees in the forest. $h_t(x)$: prediction of each decision tree. $H(x)$: final prediction of the model.

Linear regression

The Linear Regression model is commonly used model for continuous prediction. It assumes a linear relationship between the features and the target variable, that is solar radiation. The model maps how atmospheric parameters affect the radiation being experienced at a point on the surface of the Earth. The model fits a hyperplane to the data points and adjusts parameters θ to minimize errors.

1. Training

Model hypothesis: The hypothesis function for linear regression is defined as,

$$h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \cdots + \theta_M x_M \quad (2)$$

where x_1, x_2, \dots, x_M are the features

Cost function: The cost function quantifies how well the linear regression model's predictions match the actual target values. The goal is to minimize this cost function during training.

$$J(\theta) = \frac{1}{2N} \sum_{i=1}^N h_{\theta}((x_i) - y_i)^2 \quad (3)$$

- $J(\theta)$: cost function
- N : i_{th} data point. Each x_i consists of the features x_1, x_2, \dots, x_M for that particular data point.
- $h_{\theta}(x_i)$: predicted value for the i -th input.
- y_i : actual target value (the ground truth) for the i -th data point.

2. Predictions

$$H(x) = h_{\theta}(x) \quad (4)$$

The final prediction is the value computed by the learned linear regression model.

Decision tree regressor

The Decision Tree Regressor maps input features to output values. It creates a tree-like structure where each internal node represents a decision based on feature values, and each leaf node corresponds to a predicted output value. The model recursively partitions the input space into regions where the output is as constant as possible⁴⁰.

1. Training

- Select best feature to split on:

$$\text{Split} = \arg \min_f \left(\frac{1}{N_f} \sum_{i=1}^{N_f} (y_i - \hat{y}_f)^2 \right) \quad (5)$$

This step involves selecting the feature f that minimizes the mean squared error (MSE) within each subset N_f after the split.

- Recursive splitting:

$$h(x) = \text{predict_leaf}(x) \quad (6)$$

The recursive splitting process continues until the stopping criteria are met, such as maximum depth or minimum number of samples per leaf.

2. Prediction The predicted value is:

$$H(x) = \hat{y}_{\text{leaf}} \quad (7)$$

where x : feature matrix (input data) with N samples and M features. \hat{y}_{leaf} : predicted value for the input x based on the leaf node that x falls into.

Gradient boosting regressor

Gradient Boosting Regressor is a more advanced model that sequentially builds multiple decision trees, adjusting errors on each iteration. Unlike Decision Tree which makes predictions by splitting data into branches based on feature values, Gradient Boosting constructs trees in a sequence, learning from previous mistakes to gradually improve the model's accuracy. This "boosting" algorithm makes it more powerful for complex, non-linear relationships, such as those found in solar radiation patterns.

1. Training Initialize the model with an initial prediction:

$$H_0(x) = \frac{1}{N} \sum_{i=1}^N y_i \quad (8)$$

where

- (a) $H_0(x)$: initial prediction
- (b) N : number of samples.

For $t = 1$ to T :

- Compute Residuals:

$$residuals_t = y - H_{t-1}(x) \quad (9)$$

where

- (a) y : actual target values
- (b) $H_{t-1}(x)$: current prediction

- Train Weak Learner: Fit a weak learner to the residuals to capture patterns not yet learned by the model:

$$h_t(x) = \text{train_weak_learner}(X, residuals_t) \quad (10)$$

where $h_t(x)$ is the weak learner trained on the residuals to correct the model's previous errors.

- Update the Model with Boosting:

$$H_t(x) = H_{t-1}(x) + \mu_t h_t(x) \quad (11)$$

where μ_t is the learning rate that scales the contribution of each weak learner $h_t(x)$. This boosting step enables the model to iteratively reduce prediction errors by adding each new learner's corrections to the previous prediction.

2. Prediction

$$H(x) = H_T(x) \quad (12)$$

where $H(x)$ represents the final model's prediction after T boosting iterations.

AdaBoost model

The AdaBoost Regressor model combines multiple weak learners and creates a stronger and more accurate model. It does so by training a series of base models, adjusting the weight of the model in each iteration. It assigns higher weights to misclassified points, thereby highlighting them for subsequent iterations. The integrations of AdaBoost Regressor with XAI, not only offers accurate predictions for solar radiation but makes the prediction easier for human interpretation.

Initialize equal weights for all data points. Let

$$D_1(i) = \frac{1}{N} \quad (13)$$

where

1. $D_1(i)$: weight assigned to i_{th} data of first iteration.
2. N : total number of data points

The weight of the weak learner in the final model (α_t) is calculated as follows:

$$\alpha_t = 0.5 \ln \frac{1 - e^t}{e^t} \quad (14)$$

where

1. e^t : weak learner's error factor at time t .

The weight of the data point at the new iteration is:

$$D_{t+1}(i) = \frac{D_t(i) \cdot e^{(-\alpha_t * y_i * h_t(x_i))}}{Z_t} \quad (15)$$

Upon combining the weighted sum of weak learners, we can the final predicted model as:

$$H(x) = \text{sign} \left(\sum_{i=1}^t \alpha_t * h_t(x) \right) \quad (16)$$

where

1. $h_t(x)$: weak learner at iteration t

Local interpretable model-agnostic explanations (LIME)

In the context of Reliable and Efficient Solar Radiation Estimation, LIME (Local Interpretable Model-agnostic Explanations) is a valuable tool for explaining and interpreting the predictions made by machine learning models. LIME explains individual predictions by approximating the behavior of complex models around a specific prediction. For each data point, it generates a set of similar samples and learns how small changes in feature values impact the prediction. LIME therefore creates a model that represents behavior around that particular instance which is easy for human interpretation. LIME explains a model's predictions as weights of features in the local surrogate model. The model attempts to approximate what is observed by a more complicated black box model in the neighborhood of the point of interest for the data⁴¹.

LIME helps to interpret the models used in the estimation of solar radiation by approximating the behavior of each model around specific data points, thus improving transparency and reliability⁴². For example, for linear regression, LIME identifies the most important features, especially if the features are high dimensional or have multi col-linearity. For decision trees, LIME explains influential feature splits for any given particular prediction. It reveals the important factors in the complex dataset for a random forest. LIME reveals Ada Boost's iterative feature emphasis as a reflection of the season or sensor impact. For gradient boosting, LIME demystifies complex feature contributions that confirm how variables like humidity and cloud cover affect predictions.

Local perturbations

To generate explanations, LIME produces a series of perturbed samples around the target instance, slightly modifying feature values to observe changes in the model's output. This perturbation allows LIME to construct a local, interpretable model that captures the black-box model's behavior in the immediate vicinity of the data instance. By making assumptions and approximation of the model around specific predictions, LIME reveals which features are most influential in the specific prediction.

Weighted regression model

The objective function for LIME, which balances model fidelity and interpretability, can be expressed as:

$$\text{Objective Function} = \arg \min_g (L(f, g, \pi_x) + \Omega(g)) \quad (17)$$

where:

- $L(f, g, \pi_x)$: Denotes the fidelity loss between the black-box model f and the interpretable model g , measured within a local neighborhood π_x around the instance x .
- π_x : A local neighborhood around the target instance x , created by perturbing x to generate similar samples.
- $\Omega(g)$: A regularization term penalizing the complexity of g , promoting simpler models that enhance interpretability⁴³.

Interpretation of results

The simplified model g reveals the importance of each feature, highlighting which ones most strongly influence the prediction for the specific instance. Visual tools in LIME, such as bar charts, display positive and negative feature influences, helping users understand the role of each feature in the final prediction. LIME has been extensively applied in fields such as healthcare and finance, enabling transparency in high-stakes decision-making with black-box models.

SHAPLEY

SHAP (SHapley Additive exPlanations) is a powerful approach to model outputs with an explanation by assigning an importance value to every feature using cooperative game theory. IT helps to solve the challenge of quantifying individual contributions of certain environmental factors, such as take temperature, cloud cover, and humidity, toward improving predictions and, thus, making models transparent. It can give rise to interactions among features, which is particularly useful especially for attempts at trying to understand complex models like gradient boosting and random forests, or validate predictions with respect to expected physical processes.

The SHAP methodology is grounded in the following principles:

- **Fairness**: SHAP ensures fairness by assigning each feature a contribution value, considering its interactions with other features.

- **Additivity:** A baseline prediction typically constitutes the expected value for that particular model. This measures just how much each feature contributes to taking the prediction away from this baseline, making clear exactly how each feature influences the model's decision.
- **Consistency:** If the contribution of a feature to the prediction increases when another feature is removed, the SHAP value of the feature will also increase⁴⁴. Equation:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (18)$$

ϕ_i : The SHAP value for feature i , S : A subset of features excluding the feature i , N : total set of all features in the model. $f(S)$: The model's output when only the features in S are considered. $f(S \cup \{i\})$: The model's output when the feature i is included along with the features in S . $\frac{|S|!(|N| - |S| - 1)!}{|N|!}$: A combinatorial factor that accounts for the possible feature orderings in the cooperative game framework.

One of SHAP's advantages is its broad applicability across various domains, such as healthcare and finance, where transparency in decision-making is critical. It allows for the identification of influential features, which is essential for high-stakes applications where trust in the model's decisions is paramount. SHAP's ability to explain complex models, such as ensemble learning methods, by breaking down their predictions into feature contributions, makes it invaluable for creating interpretable AI systems⁴⁵.

Moreover, SHAPELY provides clear and consistent explanations for complex machine learning models and has been widely applied across multiple fields due to its ability to break down a model's output in terms of contributions from individual features. SHAP interprets model predictions in fields ranging from healthcare to finance to guarantee transparency and fairness and help better make more effective decisions with actionable insights. This importance of trust and understanding in machine learning systems is reflected in its applications and usability in real-world tasks such as fraud detection, risk assessment, and personalized health.

Solar radiation pattern equations

Solar Declination (δ) It is the angle between the sun's rays and equator.

$$\delta = 23.45 \sin \frac{360}{365} (N + 284) \quad (19)$$

where N is the day of the year

Hour angle (H) The hour angle represents the time since solar noon, measured in degrees.

$$H = 15 \deg(12 - LocalSolarTime) \quad (20)$$

Solar Elevation angle (β) The solar elevation angle indicates the height of the sun above the horizon.

$$\sin(\beta) = \sin(\Phi)\sin(\delta) + \cos(\beta)\cos(\delta)\cos(H) \quad (21)$$

where ϕ is the latitude of the point of reference.

Azimuth angle (θ) The azimuth is the angle between the north vector and the star's vector on the horizontal plane. Azimuth is usually measured in degrees, in the positive range 0° to 360° or in the signed range -180° to $+180^\circ$.

$$\tan(\theta) = \frac{\sin(H)}{\tan(\delta)\cos(\phi) - \sin(\phi)\cos(H)} \quad (22)$$

It's worth noting that the value obtained from the formula may need to be adjusted depending on the position of the sun. The adjustment is typically done to ensure the azimuth angle falls within the correct compass quadrant (e.g., between 0° and 360°).

Metrics for comparative study

When evaluating machine learning models for solar radiation prediction, several regression metrics are commonly used to assess performance:

1. Mean Squared Error (MSE) measures the average squared differences between the predicted (\hat{y}_i) and actual (y_i) values. A smaller MSE indicates better model accuracy.

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (23)$$

2. Mean Absolute Error (MAE) calculates the average of the absolute differences between predicted and actual values. MAE is less sensitive to large errors compared to MSE.

$$MAE = \frac{1}{N} \sum |y_i - \hat{y}_i| \quad (24)$$

3. R-squared (R^2) indicates the proportion of variance in the target variable that is explained by the model. A higher R^2 suggests a better model fit.

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (25)$$

Results

This section describes the detailed experimentation analysis and results of the statistical significance of data attributes, AI results of the regression models such as DTR, RFR, GDR, ABR and XGBR. Later the results are presented for the local surrogates with LIME model. The global surrogacy results are presented using SHAPLEY model.

The dataset contains independent attributes such as Humidity, Temperature, Speed, Wind Direction, Pressure and the dependent attribute is the target Solar radiation. The Pearson correlation model is applied to test the statistical significance of the attributes. The correlation matrix is presented in the Fig. 3. The correlation matrix shows the Radiation has the highest correlation with the Temperature, followed by the Pressure, the Wind Direction and Humidity work against the regression score of the Radiation, showing the negative correlation towards the dependent variable.

The regression scores of the various models are presented in the Table 1. The RFR tops the table with the regression score of 0.9028, followed by the GBR with the score of 0.8910. Thus the XAI models are built, based on the regression score and values of the RFR.

R score is often preferred in the context of Explainable AI (XAI) because it provides a normalized, relative measure of the proportion of variance explained by the model, making it easier to interpret in an absolute sense than MAE or MSE. MAE and MSE are scale-dependent error metrics whose values alone do not inherently indicate a good or bad fit. That is why RF is chosen in this context for explainability.

The models selected for solar radiation regression prediction—Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, AdaBoost Regressor, and Extreme Gradient Boosting (XGBoost) Regressor—were chosen because they collectively provide a strong balance of accuracy, robustness, and interpretability required for modeling the highly non-linear and variable nature of solar radiation. Decision Trees offer a simple and interpretable baseline for capturing fundamental patterns, while Random Forests enhance stability and generalization through bagging, making them resilient to noise and missing data.

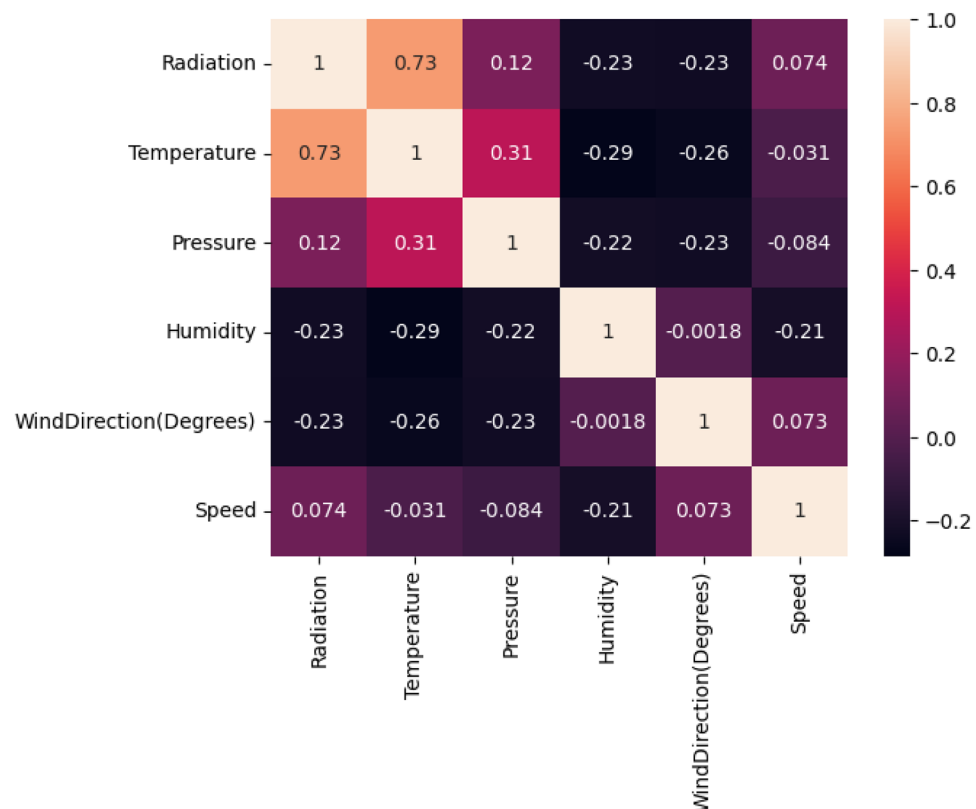


Fig. 3. Correlation map of the solar radiation attributes.

Model	R2-score	MAE	MSE
DTRr	0.7882	0.3152	1.348
RFR	0.9028	0.6198	1.828
GBR	0.891	0.2156	0.6956
ABR	0.8289	0.2792	1.091
XGBR	0.7098	0.3710	1.857

Table 1. Performance comparison of various regression models.

Feature	Weight	Nature
Temperature	0.82	Positive
Pressure	0.60	Positive
Speed	− 0.48	Negative
Humidity	− 0.73	Negative

Table 2. Feature analysis with LIME.

commonly found in environmental datasets. Gradient Boosting methods further improve predictive accuracy by sequentially correcting errors, enabling the model to capture fine-grained atmospheric variations. AdaBoost complements this by focusing learning on difficult-to-predict samples, thus strengthening overall performance. XGBoost, as a state-of-the-art boosting algorithm, brings computational efficiency, built-in regularization, and superior handling of complex feature interactions, making it particularly effective for solar radiation forecasting. Together, these models provide a comprehensive and well-justified suite capable of delivering robust, high-precision solar radiation predictions.

LIME and SHAP decides the weight of the contributing feature and its polarity, which features may increase radiation or temperature. These diagrams show the impact of the features in determining the magnitude of the target variable, their polarity (nature) , importance and weights in determining the regression score of a target variable, it interprets the impact and importance of feature in determining the regression score in both local and global surrogacy.

In regression analysis, each SHAP plot provides a distinct perspective on how input features contribute to the model's continuous output predictions. The SHAP summary plot displays all data points as colored dots, where each dot represents an individual prediction and its SHAP value indicates the magnitude and direction of that feature's influence on the target variable. For instance, in the solar dataset, higher SHAP values for temperature correspond to increased predicted solar radiation, with the color gradient (from blue to red) showing low to high feature values. The dependence plot further isolates one feature to show how its SHAP value changes across the data range, revealing both linear and nonlinear relationships (e.g., temperature rising with radiation until a saturation point). Meanwhile, force and waterfall plots visualize individual predictions by showing how each feature pushes the output above or below the average model prediction. Collectively, these plots allow practitioners to interpret not only which features matter most but also how and to what extent they influence solar radiation predictions across different conditions.

The next presented model is LIME. This model brings about the local surrogacy and the dependency with the features, based on lasso, which is a linear relationship model, that analyses the surrogacy between the dependent and the independent variables on the dataset. The LIME model is presented with two plots such as PyPlot and a Notebook. Pyplot is basically presents the nature of the features with bar chart. The green indicates the feature that contains weight that are positive towards the prediction and red indicates the weights that are lesser the better for the prediction. The four features towards the target prediction with the nature and corresponding weights are presented in the Table 2. These results are with respect to the test data for a particular instance X_test[0]. The pictorial representation of the same is presented in the Fig. 4. The next representation is the Notebook. This shows the prediction score of 50.86 for the instance. This shows the order of importance of the features with respect to the determination of the regression score. This shows 82% importance for the Temperature,73% for the speed, 8.57% importance for the Pressure and 4.25% for the Pressure. The corresponding weights and order of importance are also presented by the notebook, which is represented in Fig. 5.

The statistical significance test is provided in the Table 3. This shows that the *p*-value is 0.0000 and the Correlation is 0.7349, The correlation is positive and the *p*-value is less than <0.05 and hence the data points are statistically significant

The next XAI representation is SHAPELY. This model illustrates several plots such as Force plot, Box plot, Waterfall plot, Decision and Dependency plot. These plots provide the results in both local and global surrogacy based on their nature. The Dependency plot for the global surrogacy, for the complete dataset is presented in Fig. 6. This shows the dependency between two features such as Solar Radiation and Pressure, the blue dotted lines shows strong correlation and red dotted lines show weaker dependency. The density of the dependency is higher during the center of the dataset. There is another plot in XAI for local surrogacy namely known as Partial Dependency Plot(PDP) plot. This shows the dependency between the attributes in the local surrogacy.

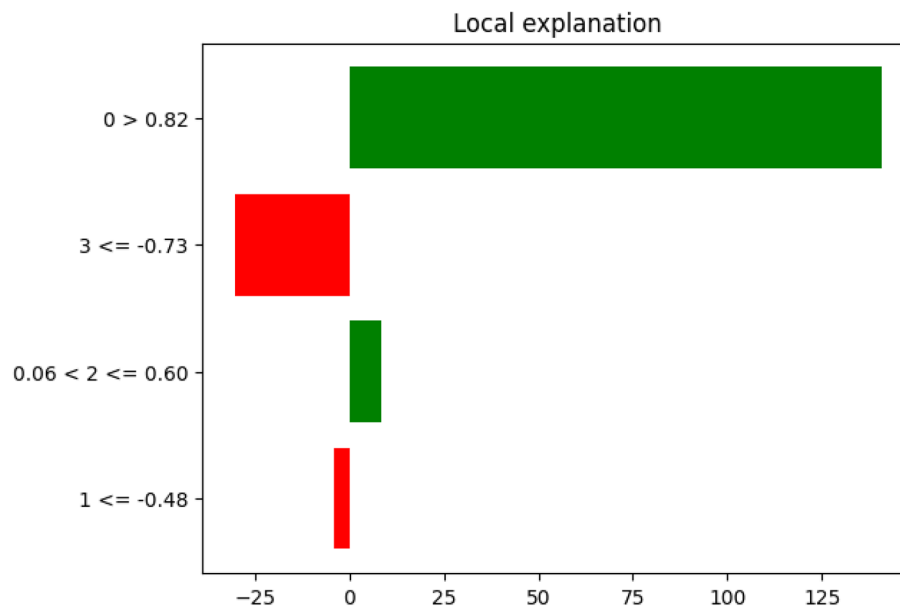


Fig. 4. PyPlot of the solar radiation attributes.

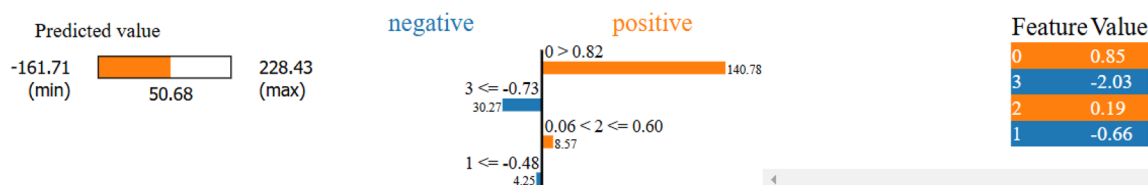


Fig. 5. LIME Notebook of the solar radiation attributes.

Feature	p-value	Correlation
Radiation versus temperature	0.0000	0.7349

Table 3. Statistical significant test.

The next plot presented is the Force plot. This is also a global surrogacy explanation. The force plot has two regions such as low to high and high to low. The regions under blue color represented as low and red colored regions are represented as high. The Force plot is presented in the Fig. 7. The Feature 0 which is the Temperature and Feature 2 which is Speed taking highest priorities as per the Fig. 7, since they share the boundary of red and blue colored regions. The feature weights are also represented for each of the feature.

Feature Importance Plot of the SHAPLEY is presented in Fig. 8. This plot shows the order of importance of the feature in global surrogacy. The red dots towards the maximum the better prediction and blue dots represents minimum the better for the determination of the magnitude of the output. Feature 0 takes the highest priority as per the order of importance for the determination of the magnitude of the output.

The Box plot of SHAPLEY is presented in Fig. 9. This plot shows the weight and nature of the feature for a particular instance. As per the given instance of data, Feature 2, which is Pressure takes the highest positive weight of 14.57, takes the highest priority. The next priority is taken by Feature 0 which takes -1.97 as a negative weight. The same features are also represented by the Waterfall plot of SHAPLEY, with the same data instance and similar weights for all features. This plot is represented by the Fig. 10.

The Decision plot is the final representation of this experimentation which shows the contribution of each of the features in determining the target regression score. The score is distributed over 0 to 1. The features that are connected with blue and red color lines indicate the high and low correlation with the regression score for a data instance in the dataset. The plot is represented by the Fig. 11.

The LIME and SHAP plots play a vital role in interpreting how each feature in the solar dataset influences the predicted solar radiation, thereby enhancing both the transparency and practical usability of the predictive model. Through LIME, individual predictions can be explained locally by identifying which factors most strongly affected a specific outcome. For example, if the model predicts a solar radiation for a certain hour,

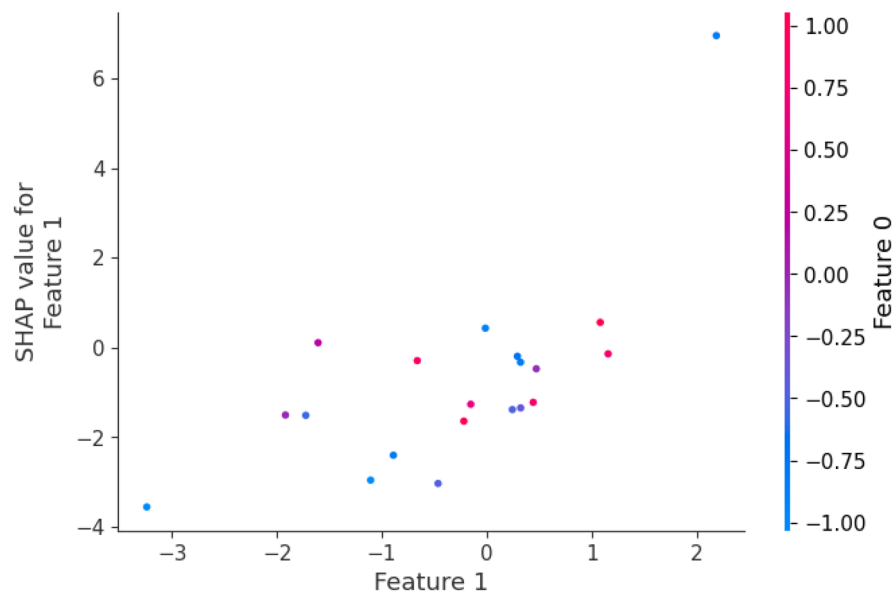


Fig. 6. SHAPLEY dependency of the solar radiation attributes.

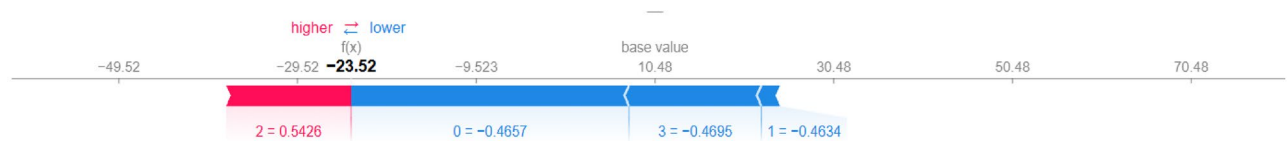


Fig. 7. SHAPLEY force plot for the solar radiation attributes.

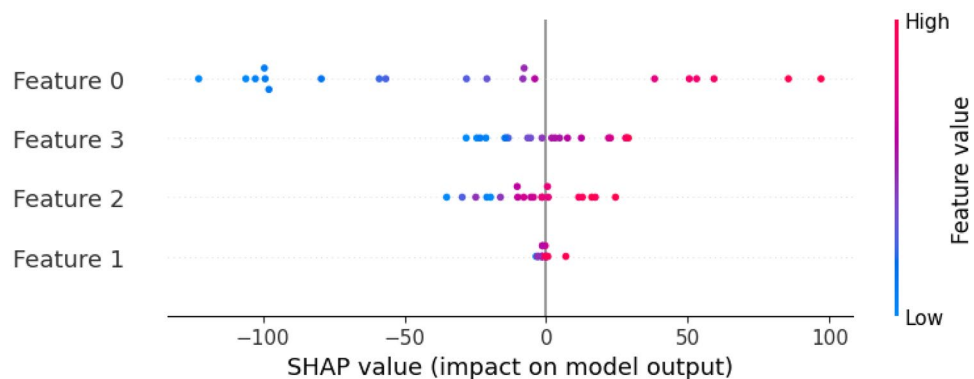


Fig. 8. SHAPLEY Summary Plot for the Solar Radiation attributes.

LIME might show that this is primarily due to elevated temperature and reduced humidity, while wind speed or pressure contribute less. This helps researchers and engineers understand why a model behaves in a certain way under specific environmental conditions, which is critical when applying predictions for real-world solar energy forecasting, environmental monitoring, or microclimate analysis.

On the other hand, SHAP offers both local and global interpretability by quantifying each feature's contribution to every prediction using Shapley values from cooperative game theory. In practical terms, SHAP plots reveal that features such as temperature has the strongest positive impact on solar radiation, while humidity and pressure typically exhibit negative or moderating effects. This global view enables decision-makers to validate whether the model's reasoning aligns with real-world physical relationships, improving trust and reliability. Additionally, SHAP's dependence plots can highlight non-linear effects—such as diminishing solar radiation gains at high

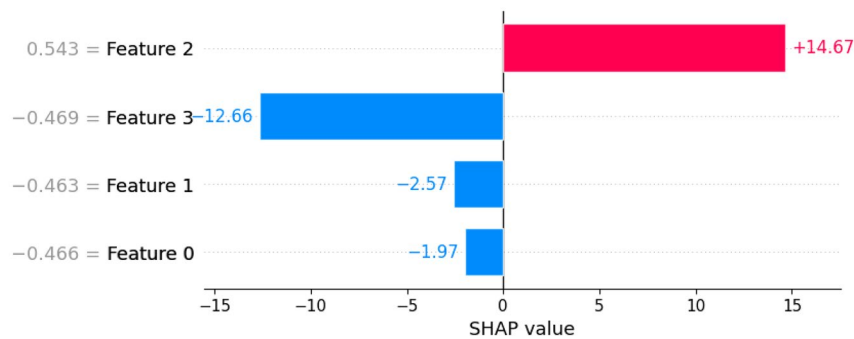


Fig. 9. SHAPLEY box plot for the solar radiation attributes.

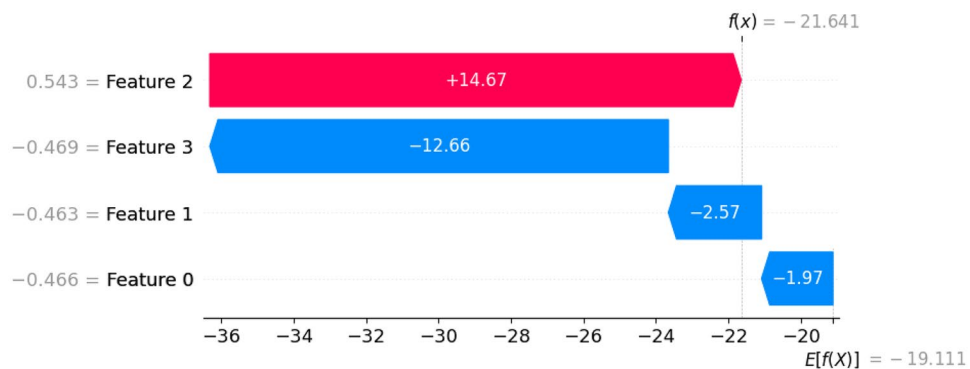


Fig. 10. SHAPLEY waterfall plot for the solar radiation attributes.

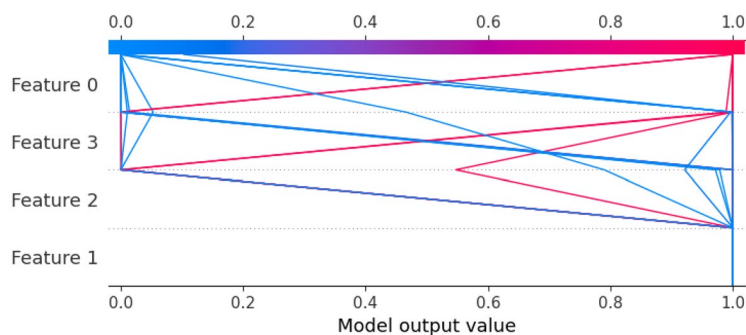


Fig. 11. SHAPLEY decision plot for the solar radiation attributes.

radiation levels—helping practitioners optimize sensor placement, calibrate solar panels, or adjust operational parameters in energy systems for better temperature regulation and forecasting accuracy.

Discussion

This study aims to develop a methodical guide for feature selection, to identify the most important input parameters that affect radiation for forecasting by regression ML models. Accurate forecasting by ML techniques helps eliminate the need for installation of pyranometers at the planned location in advance for solar radiation forecasting. Also, ML regression helps in cost effective and remote radiation forecast even at remote and inaccessible sites, if there are training data sets available for similar locations. This in turn helps to identify locations suitable for solar power plant installations and to correctly size the solar power plants and avoid grid disparity due to fluctuations power production. Further, feature selection can help reduce computation costs, improve over-fitting problems, and resolve multi col-linearity difficulties in the models by eliminating redundant or unnecessary information while retaining the most important features. By identifying the most significant variables, the redundant parameters can be eliminated, reducing the complexity, computation requirements and cost of solar radiation forecasting. Using only four independent variables, the RFR was able to predict

Model	Key parameters	Optimization strategies
Decision tree	max_depth, min_samples_leaf, criterion	1. Limit tree depth to reduce overfitting. 2. Use grid search with cross-validation.
Random forest	n_estimators, max_features, max_depth	1. Increase n_estimators until OOB error stabilizes. 2. Tune max_features for bias–variance control.
AdaBoost	n_estimators, learning_rate	1. Balance n_estimators and learning_rate. 2. Use shallow base learners (depth 1–3).
Gradient boosting	learning_rate, n_estimators, subsample	1. Use small learning_rate with more estimators. 2. Apply early stopping for generalization.
XGBoost	eta, max_depth, subsample, lambda	1. Tune regularization (lambda, alpha). 2. Use early stopping and balanced learning rate.

Table 4. Optimization points for regression models.

radiation with the regression score of 0.9028. The statistical significance of the parameters is tested using the Pearson correlation model shows that temperature has the strongest link to radiation followed by Pressure. Wind direction, and humidity all act against the radiation’s regression score, indicating a negative correlation with the radiation. The study cements the directly proportional relation between radiation and temperature. The four independent parameters can be easily obtained for any location from GIS based satellite data and can be used for solar radiation forecast. Additionally, there may be some possible limitations of the study. The parameters of the optimization for various models used in the proposed work is presented in Table 4.

Advantages of the proposed framework

The proposed framework helps the consumers identify the most productive locations for new solar farms by solar radiation forecasts, which maximize the future energy yield and return on investment. It also helps to carry out proactive maintenance and optimize cleaning schedules of a solar power plant. In addition, radiation prediction can assist in the identification of fluctuations in solar power output, which can lead to the optimization of energy storage, the promotion of grid stability and reliability, and the management of strategic reserves.

- Some unpredictable non-linear parameters such as cloud cover, sunshine hours, and aerosols are neglected and these variables can impact the accuracy of solar radiation prediction. Their correlation needs to be studied in the future to further improve the accuracy of the prediction.
- The model developed is very short-term irradiance prediction techniques and cannot be used for long-term irradiance prediction. Long-term irradiance prediction may requires training the model on GIS based satellite data of the independent parameters of very large data sets over a period of 10+ years and may be computationally intensive.
- The cost effectiveness of the system would highly be reduced using XAI insight for the solar energy estimation. This reduces the need of the sophisticated physical devices, and predicts the factors that influences the radiation with parametric analysis and evaluation.
- There is lesser influence of hardware malfunction, physical availability and personnel requirement to monitor the radiation. The proposed framework builds an automated system that generates the reports for the entire set of people involved.
- The parameters that are changing during idling, down-time, rain-fall and outage can be recorded and their variations can also be mapped in accordance with the radiation estimation.
- This proposed framework uses both local and global surrogacy models, and hence they can work on small samples of data or on an entire dataset also.

Challenges

- Like any ML model, the ability of the model to predict radiation accurately from independent variables depends on the quality of the training data set.
- Some independent parameters like sunshine hours, time of the day, etc are not considered in the study and are neglected due to their very low corelations to radiation as given by previous literatures.
- The effect of atmospheric pollution on radiation forecast is not included in the study.
- The model developed is very short-term irradiance prediction techniques and cannot be used for long-term irradiance prediction .

Conclusion

The proposed work showcases the importance of using an interpretable framework for solar radiation prediction. The interpretable LIME and SHAPELY models are built based on the RFR since the regression score is 0.9026 which is higher when compared with the competing models such as GBR, with regression score of 0.891 and ABR, with the regression score of 0.8289. The LIME and SHAPELY are tracking the influence of the features in the target estimation. The proposed framework also depicts the variations of these parameters under challenging climatic conditions, rain-fall, pollution and other natural calamities and how it impacts the solar radiation and the production of the energy. The proposed work assures timely and reliable prediction of parameters, which

enhances the quantum and quality of the renewable energy production, that is one of the prime sustainable goal in the era of Industry 5.0. The proposed work can also be extended for similar alternative energy production system, to read the influence of various parameters in the generation of the alternative energy.

The XAI adds reliability and interpretability for the prediction of solar radiation and voltage generation. It adds the confidentiality to the user, so that they can anticipate a certain amount of power generation in the near future, because they understand the behavior and influence of the features with XAI plots. Thus, the XAI makes solar power generation process more predictable and dependable in the end user perspective.

Data availability

The datasets used and/or analysed during the current study are available in Kaggle, in the link <https://www.kaggle.com/datasets/dronio/SolarEnergy>

Code availability

Sample codes used in the proposed study is available as a supplementary material.

Received: 13 September 2025; Accepted: 19 December 2025

Published online: 12 January 2026

References

- Li, H., Lian, Y., Wang, X., Ma, W. & Zhao, L. Solar constant values for estimating solar radiation. *Energy* **36**, 1785–1789 (2011).
- Mohammadi, K., Shamshirband, S., Kamsin, A., Lai, P. & Mansor, Z. Identifying the most significant input parameters for predicting global solar radiation using an ANFIS selection procedure. *Renew. Sustain. Energy Rev.* **63**, 423–434 (2016).
- Mousavi, S. M., Mostafavi, E. S., Jaafari, A., Jaafari, A. & Hosseinpour, F. Using measured daily meteorological parameters to predict daily solar radiation. *Measurement* **76**, 148–155 (2015).
- Anselmo, S., Safaeianpour, A., Moghadam, S. T. & Ferrara, M. GIS-based solar radiation modelling for photovoltaic potential in cities: A sensitivity analysis for the evaluation of output variability range. *Energy Rep.* **12**, 4656–4669 (2024).
- Idrovo-Macancela, A., Velecela-Zhindón, M., Barragán-Escandón, A., Zalamea-León, E. & Mejía-Coronel, D. GIS-based assessment of photovoltaic solar potential on building rooftops in equatorial urban areas. *Heliyon* **11**, e41425 (2025).
- Jadhav, A. V., Bhawar, R. L., Dumka, U. C., Rahul, P. & Kumar, P. P. Impacts of meteorological conditions on the plummeting surface-reaching solar radiation over a sub-tropical station-Pune, India. *Energy Sustain. Dev.* **80**, 101444 (2024).
- Müller, B., Wild, M., Driesse, A. & Behrens, K. Rethinking solar resource assessments in the context of global dimming and brightening. *Sol. Energy* **99**, 272–282 (2014).
- Pourasl, H. H., Barenji, R. V. & Khojastehnezhad, V. M. Solar energy status in the world: A comprehensive review. *Energy Rep.* **10**, 3474–3493 (2023).
- Irena. *Renewable Energy Statistics* (accessed 10 November 2024); https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2024/Jul/IRENA_Renewable_Energy_Statistics_2024.pdf (2024).
- Kumler, A. et al. Potential effects of climate change and solar radiation modification on renewable energy resources. *Renew. Sustain. Energy Rev.* **207**, 114934 (2025).
- Xin-gang, Z. & Zhen, W. Technology, cost, economic performance of distributed photovoltaic industry in China. *Renew. Sustain. Energy Rev.* **110**, 53–64 (2019).
- Zakeri, B., Gisse, G. C., Dodds, P. E. & Subkhankulova, D. Centralized versus distributed energy storage-benefits for residential users. *Energy* **236**, 121443 (2021).
- Chu, Y., Li, M., Pedro, H. T. & Coimbra, C. F. A network of sky imagers for spatial solar irradiance assessment. *Renew. Energy* **187**, 1009–1019 (2022).
- Chu, Y., Wang, Y., Yang, D., Chen, S. & Li, M. A review of distributed solar forecasting with remote sensing and deep learning. *Renew. Sustain. Energy Rev.* **198**, 114391 (2024).
- Liu, Q. et al. A review and guide on selecting and optimizing machine learning algorithms for daylight prediction. *Build. Environ.* **244**, 110822 (2023).
- Zhou, Y., Liu, Y., Wang, D., Liu, X. & Wang, Y. A review on global solar radiation prediction with machine learning models in a comprehensive perspective. *Energy Convers. Manag.* **235**, 113960 (2021).
- Yeom, J.-M., Park, S., Chae, T., Kim, J.-Y. & Lee, C. S. Spatial assessment of solar radiation by machine learning and deep neural network models using data provided by the coms mi geostationary satellite: A case study in South Korea. *Sensors* **19**, 2082 (2019).
- Demirhan, H. The problem of multicollinearity in horizontal solar radiation estimation models and a new model for turkey. *Energy Convers. Manag.* **84**, 334–345 (2014).
- Feng, C. & Zhang, J. Solarnet: A sky image-based deep convolutional neural network for intra-hour solar forecasting. *Sol. Energy* **204**, 71–78 (2020).
- Abdallah, M. et al. Daily global solar radiation time series prediction using variational mode decomposition combined with multi-functional recurrent fuzzy neural network and quantile regression forests algorithm. *Energy Rep.* **10**, 4198–4217 (2023).
- Ağbulut, Ü., Gürel, A. E. & Biçen, Y. Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. *Renew. Sustain. Energy Rev.* **135**, 110114 (2021).
- Allal, Z., Noura, H. N. & Chahine, K. Machine learning algorithms for solar irradiance prediction: A recent comparative study. *e-Prime-Adv. Electric. Eng. Electron. Energy* **7**, 100453 (2024).
- Zang, H. et al. Short-term global horizontal irradiance forecasting based on a hybrid CNN-LSTM model with spatiotemporal correlations. *Renew. Energy* **160**, 26–41 (2020).
- Kumari, P. & Toshniwal, D. Deep learning models for solar irradiance forecasting: A comprehensive review. *J. Clean. Prod.* **318**, 128566 (2021).
- Kim, S., Jung, J. & Sim, M. A two-step approach to solar power generation prediction based on weather data using machine learning. *Sustainability* **11**(5), 1501 (2019).
- Huang, L. et al. Solar radiation prediction using different machine learning algorithms and implications for extreme climate events. *Front. Earth Sci.* **9**, 596860 (2021).
- Essam, Y. et al. Investigating photovoltaic solar power output forecasting using machine learning algorithms. *Eng. Appl. Comput. Fluid Mech.* **16**, 2002–2034 (2022).
- Yadav, A. K., Malik, H. & Chandel, S. Selection of most relevant input parameters using weka for artificial neural network based solar radiation prediction models. *Renew. Sustain. Energy Rev.* **31**, 509–519 (2014).
- Jiang, C. & Zhu, Q. Evaluating the most significant input parameters for forecasting global solar radiation of different sequences based on informer. *Appl. Energy* **348**, 121544 (2023).

30. Superchi, F., Moustakis, A., Pechlivanoglou, G. & Bianchini, A. Can machine learning enhance renewable power forecasts? A study on data-driven methods for solar photovoltaic and wind turbines. *Energy* **336**, 138396 (2025).
31. Chandrashekar, S. V. et al. A hybrid physics-informed neural and explainable AI approach for scalable and interpretable AQI predictions. *MethodsX* **15**, 103597. <https://doi.org/10.1016/j.mex.2025.103597> (2025).
32. Krishnasamy, L., Dhanaraj, R. K., Pamucar, D. & Ouaisa, M. In *Distributed Deep Learning and Explainable AI (XAI) in Industry 4.0* (Springer, 2025).
33. Jothi, A. M. S., Rajarathinam, V. D. R. K. & Nayyar, A. Explainable AI-driven intrusion detection for securing IoT-enabled autonomous transportation systems. *Cluster Comput.* <https://doi.org/10.1007/s10586-025-05617-1> (2025).
34. Wang, Y., Azad, M. A., Zafar, M. & Gul, A. Enhancing AI transparency in IoT intrusion detection using explainable AI techniques. *Internet Things* **33**, 101714 (2025).
35. Johannssen, A., Qiu, P., Yeganeh, A. & Chukhrova, N. Explainable AI for trustworthy intelligent process monitoring. *Comput. Ind. Eng.* **209**, 111407 (2025).
36. Dandamudi, J. T., Kandula, R., Raj, R. D. A., Yanamala, R. M. R. & Prakasha, K. K. Explainable ai for wind energy systems: State-of-the-art techniques, challenges, and future directions. *Energy Convers. Manag.* **X28**, 101277 (2025).
37. Adib, A. U. R., Islam, M., Abid, M. S. & Ahshan, R. A deep learning based framework for solar panel segmentation and fault classification enhanced with explainable ai. *Sol. Energy* **302**, 114058 (2025).
38. Andrey (2017).
39. Andrey. Solar radiation prediction (2017).
40. Bertsimas, D., Dunn, J. & Paschalidis, A. Regression and classification using optimal decision trees. In *2017 IEEE MIT Undergraduate Research Technology Conference (URTC)* 1–4, <https://doi.org/10.1109/URTC.2017.8284195> (2017).
41. Salih, A. et al. Commentary on explainable artificial intelligence methods: Shap and lime. *arXiv:2305.02012v3* (2018).
42. Akshitha, K., Roopashree, R., Kodipalli, A. & Rao, T. Utilizing explainable AI methodologies: Lime and shap, for the classification of natural disasters through machine learning algorithms, in *2024 Parul International Conference on Engineering and Technology (PICET)* 1–5. <https://doi.org/10.1109/PICET60765.2024.10716119> (2024).
43. Ng, C. H., Abuwala, H. S. & Lim, C. H. Towards more stable lime for explainable AI, in *2022 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, 1–4, <https://doi.org/10.1109/ISPACS57703.2022.10082810> (2022).
44. Lundberg, S. M. & Lee, S. A unified approach to interpreting model predictions. *CoRR* **abs/1705.07874** (2017). [arXiv:1705.07874](https://arxiv.org/abs/1705.07874).
45. Bugaj, M., Wrobel, K. & Iwaniec, J. Model explainability using shap values for lightgbm predictions, in *2021 IEEE XVIIIth International Conference on the Perspective Technologies and Methods in MEMS Design (MEMSTECH)* 102–106. <https://doi.org/10.1109/MEMSTECH53091.2021.9468078> (2021).

Author contributions

M.K.N Perforated Formal Analysis, Investigation, Visualization , J.J performed conceptualization, validation, and wrote the article, S.M wrote and project management, supports funding and provided software, A.G revised the original article and methodology, S.U does the project management and supervision. S.M did the revision of the article and provided resources. N.C wrote the section of the original article.

Funding

Open access funding provided by Symbiosis International (Deemed University).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-33604-4>.

Correspondence and requests for materials should be addressed to S.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026