Theoretical and Applied Climatology
https://doi.org/10.1007/s00704-024-04951-5

RESEARCH

# Advanced ensemble machine-learning and explainable ai with hybridized clustering for solar irradiation prediction in Bangladesh

**Muhammad Samee Sevas[1] · Nusrat Sharmin[1] · Chowdhury Farjana Tur Santona[1] · Saidur Rahaman Sagor[1]**

Received: 22 December 2023 / Accepted: 25 March 2024


**Abstract**
The solar revolution in Bangladesh stands as a symbol of hope and self-reliance, illuminating communities and steering the nation towards a more sustainable future. Highlighting the crucial importance of solar irradiance forecasting in attaining sustainable energy objectives, this study utilizes advanced machine learning techniques to enlighten solar irradiance prediction in Bangladesh. Ensemble machine learning has become a vital tool that significantly enhances the precision and reliability of solar radiation predictions by amalgamating diverse model outputs, providing a more robust and accurate forecasting frame- work. Previous research efforts have overlooked the exploration of this diverse range in solar radiation prediction using ensemble machine learning. This study addresses this gap by conducting a detailed experiment on ensemble artificial intelligence techniques. Furthermore, the research introduces Explain- able AI (XAI) with ensemble machine learning, shedding light on factors influencing predictions and providing valuable insights for decision-makers in the solar energy sector. Notably, the study pioneers an XAI-based analysis specific to Bangladesh, marking a significant stride in solar radiation predic- tion. Additionally, a novel hybridized approach incorporating various clustering techniques and the LightGBM algorithm is introduced, offering an efficient framework for solar radiation prediction. As a result, this study contributes to our understanding and optimization of solar irradiance prediction by offer- ing a comprehensive method that integrates XAI, ensemble approaches, and machine learning. We have developed an autoML tool based on XAI and Ensem- ble as a further contribution. We have validated our result with the low-code PyCaret machine learning package to see that, among all the methods, lightGBM has shown promising results in terms of solar irradiance prediction. Ensem- ble machine learning boosting techniques, notably LightGBM and CatBoost, exhibited superior performance, achieving high $R^2$ scores of 0.91 and demonstrat- ing remarkable accuracy. Through feature importance analysis, it was discerned that"Sunshine (Hours)" emerged as the most impactful factor influencing the predictive models. Employing a hybridized clustering approach, incorporating K- means-LightGBM, MiniBatchKMeans-LightGBM, Fuzzy C-Means-LightGBM, and GaussianMixture-LightGBM models, excelled in forecasting solar radiation into distinct clusters, particularly excelling in the"Very Cloudy" category.

## 1 Introduction

Solar irradiance in Bangladesh has a profound impact, harnessing the nation's abun- dant sunlight for economic growth, environmental sustainability, and resilience against climate change. Situated in the solar belt, the sun provides clean energy, reduces carbon emissions, and fosters a green conscience. In a land prone to climate change impacts, solar energy serves as a dependable lifeline, powering homes, schools, and hos- pitals. Beyond its practical applications, the solar revolution in Bangladesh symbolizes the nation's dedication to a cleaner and brighter future, fostering unity

✉ Nusrat Sharmin
nusrat@cse.mist.ac.bd

Muhammad Samee Sevas
samee.sevas@gmail.com

Chowdhury Farjana Tur Santona
cftsantona@gmail.com

Saidur Rahaman Sagor
sagorislam01799@gmail.com

[1] Department of Computer Science and Engineering, Military Institute of Science and Technology, Dhaka, Bangladesh

Published online: 17 April 2024

among com- munities and improving livelihoods. The transformative power of the sun underscores the resilience and adaptability of the people, making precise solar radiance forecasting crucial for harnessing this sustainable energy source in the context of Bangladesh.

Forecasting solar irradiance plays a pivotal role in our quest for sustainable energy. It enables us to seamlessly integrate solar power into our energy grid, ensuring a consistent supply of eco-friendly electricity and decreasing our dependence on fossil fuels. This foresight involves deciphering atmospheric circumstances, cloud formations, and solar rhythms, harmonizing scientific knowledge with the natural world. It stands as a testament to human inventiveness, offering a pathway to tap into the universe's energy reservoir. This capacity to predict the sun's energy emissions resembles a celes- tial crystal ball, steering us towards a more eco-friendly and robust tomorrow. Solar radiation forecasting through machine learning is a cutting-edge approach that blends the power of AI with the sun's unpredictable nature. By analyzing historical data and real-time atmospheric conditions, machine learning models provide highly accurate predictions of solar radiation, enabling optimal energy production for solar systems. This technology transcends traditional weather forecasting, offering significant benefits for solar farms, businesses, and households, reducing costs and minimizing environ- mental impact. It represents a harmonious union of nature and innovation, driving us toward a cleaner, more efficient future where every ray of sunlight is maximized for energy generation. Ensemble machine learning is a crucial technique in AI and data science, as it combines the predictions of multiple models to improve accuracy and model robustness. By leveraging the strengths of diverse algorithms, ensembles reduce overfitting and enhance generalization. Popular ensemble methods like Random Forests and Gradient Boosting assemble weak learners into strong ones, making them adaptable to complex real-world data. In the realm of solar radiance prediction, ensemble machine learning also plays a significant role in achieving successful outcomes. Ensembles are vital for feature importance and model interpretability, benefiting fields from finance to healthcare. They are pivotal in achieving more accurate and reliable predictive models in today's data-rich world. In the realm of solar radiance prediction in the context of Bangladesh, ensemble machine learning also plays a significant role in achieving successful outcomes. The ability of solar radiation (Rs) to support life on earth, regulate the climate and weather, and supply a renewable energy source is all dependent on it (Hissou et al. 2023). Prediction and reliable energy production are hampered by its erratic and intermittent nature. To estimate Rs, (Hissou et al. 2023) suggests a novel framework that combines different machine learning models and takes into account overlooked factors. Recursive Feature Elimination (RFE) is employed by the framework in conjunction with algorithms like random forest, logistic regression, decision trees, Pearson correlation, and gradient boosting models. The models perform admirably, with the logistic regression model standing out for its exceptional capabilities (Hissou et al. 2023). The authors in Alam et al. (2023) employed a combination of ensemble machine-learning models, including Gradient-Boosting Regressor, Adaboost Regressor, Random Forest Regressor, and Bagging Regressor to forecast solar irradiation in Bangladesh. The study (Mishra et al. 2023) emphasizes how crucial a solar power forecasting model is for integrating renewable energy sources into the current electrical grid and lessening the effects of solar intermittency. For solar radiation prediction, ensemble methods such as random forest (RF) and extreme gra- dient boosting (XGBoost) are found to be effective, outperforming other models and increasing forecast accuracy. The study emphasizes the necessity of data-driven algo- rithms and precise solar energy generation prediction to accomplish this. Additionally, it highlights how well ensemble approaches work to increase forecast accuracy for solar radiation prediction. McCandless et al. devised a model based on decision trees to predict the temporal and spatial variations in solar irradiance (McCandless et al. 2015). Their predictive framework incorporated meteorological data such as cloud cover and temperature as input parameters. Significantly, their model outperformed a climatology-based model in terms of precision. An ensemble feature selection method is proposed (Solano et al. 2022) to select pertinent input parameters and their past observation values. The paper compares the performance of various machine-learning algorithms for solar radiation forecasting. Forecasting accuracy is increased by the suggested ensemble feature selection method, and the Voting-Average algorithm outperforms other algorithms across all prediction time horizons. Hirata and Aihara employed an infinite-dimensional delay coordinates time series model for solar irradiance prediction (Hirata and Aihara 2017). Their research highlighted the effectiveness of this method, particularly in forecasting irradiance levels during the post-sunrise period. Frimane et al. proposed an innovative approach using a Dirich- let process Gaussian mixture model to generate synthetic time-series data for solar global horizontal irradiance (GHI) at resolutions as fine as 1 min, utilizing input data with resolutions higher than 10 min (Frimane et al. 2019). In (Qing and Niu 2018), an advanced machine-learning model has been utilized to anticipate solar irradiation based on weather measurements. This model relies on long short-term memory (LSTM) networks, renowned for their extended capability to capture temporal dependencies in time series data. Notably, it exhibited commendable predictive performance in contrast to traditional

backpropa- gation neural networks. Hourly predictions of solar irradiation are delivered by this model. Bae et al. explored the utilization of a support vector machine (SVM) in con- junction with k-means clustering to predict solar irradiance one hour ahead (Bae et al. 2016). Their study incorporated diverse meteorological factors, such as cloud cover, as inputs. The findings underscored the superior performance of the SVM regression model com- pared to both artificial neural network (ANN) and nonlinear autoregressive (NAR) approaches for solar irradiance forecasting. Explainable AI (XAI) is a transformative concept in artificial intelligence that aims to enhance the transparency and interpretability of AI systems. It strives to provide human users with a clear understanding of how AI models reach their decisions, moving away from the traditional"black-box" approach. XAI utilizes various methods, includ- ing visualizations and feature analysis, to uncover the rationale behind AI predictions. This transparency is vital for applications where decision-making affects people's lives, such as healthcare and finance, and it aids in addressing biases and ensuring compli- ance with ethical standards. XAI plays a crucial role in building trust, accountability, and ethical responsibility in AI systems. A comprehensive examination of explainable artificial intelligence (XAI) was presented in Arrieta, et al. (2020), covering aspects such as concepts, taxonomies, opportunities, challenges, and the adoption of XAI tools. The authors (Lee et al. 2020) utilized XAI methodologies to interpret the results of load forecasting generated by an XGBoost model. They then demonstrated their analysis using SHAP technique. The literature reveals a multitude of AI applications in the smart grid domain. Explain- able Artificial Intelligence (XAI) has garnered increased attention in fields where the transparency and interpretability of a model's operations are paramount. For instance, it has been extensively explored in healthcare for stroke detection (Prentzas et al. 2019), cybersecurity for intrusion Detection Systems (IDS) (Marino et al. 2018), military applications for target classifi- cation (Pannu et al. 2020), and finance for risk management (Adams and Hagras 2020). Additionally, XAI-based models have found application in smart grid scenarios. In a study outlined in Pierrot and Goude (2011), researchers conducted short-term electricity load forecasting using generalized additive models. This approach facilitated the amalgamation of a regressive component incorporating explanatory variables (such as weather, calendar variables, and global trends) and an auto-regressive component encompassing lagged loads. XAI in solar irradiance prediction introduces transparency and interpretability to the forecasting of solar energy output. It aims to clarify why AI models make specific solar irradiance predictions, providing valuable insights for decision-makers in the solar energy sector. XAI techniques reveal the factors influencing predictions, helping users understand the impact of meteorological variables and historical data on accuracy. This knowledge is instrumental in refining models and optimizing energy generation. XAI promotes trust, informed decisions, and the wider adoption of solar energy as a sustainable power source.

In the exploration of explainability in solar irradiance forecasting, the study by Wang et al. (2020) delves into the application of an intrinsic model. The authors employ a direct explainable neural network characterized by a feedforward architecture that facili- tates mapping the output to the input through a non-linear model. Meanwhile, (Bahani et al. 2020) introduces the Fuzzy Rule Learning through Clustering (FRLC) model for solar irra- diation prediction. Notably, this FRLC model is elucidated using a knowledge-based approach, where interpretation involves the utilization of membership function plots for various features and analogous linguistic methods. In a different investigation, (Chaibi et al. 2021) employs permutation feature importance, SHAP, and feature dependency analysis to explicate the predictions of a solar irradiance forecast model. In (Sevas et al. 2023), the authors uti- lize five machine learning models for solar radiation prediction and SHAP XAI for feature importance.

Clustering weather data serves the purpose of uncovering meaningful patterns, relationships, and variations within meteorological information. By grouping simi- lar weather conditions, clustering enables meteorologists and researchers to identify regional and temporal trends, seasonal variations, and anomalies in the data. This process aids in understanding the complex dynamics of climate, allowing for more accu- rate predictions, targeted resource management, and improved responses to extreme weather events. Furthermore, clustering supports the identification of distinct cli- mate zones, the spatial analysis of geographical regions with similar weather patterns, and the customization of forecasts for specific areas. Overall, clustering weather data is instrumental in extracting valuable insights that contribute to more effec- tive decision-making in fields such as meteorology, climate science, and emergency preparedness.

The existing literature or study indicates a notable absence of concrete efforts based on Bangladeshi datasets, a critical gap, especially as the country advances toward the Fourth Industrial Revolution (4IR). There is a pressing need for substantial work on ensemble machine learning accompanied by thorough interpretation. In this study, we explore the domain of XAI with ensemble machine learning. Additionally, integrating season-wise clustering with regression presents an opportunity to make a significant impact on the context of seasonal data analysis and decision-making. Furthermore, we have created a solar irradiance forecasting tool, and we are confident that this contribution holds substantial importance for the nation of Bangladesh.

The main contributions and novelties of this work are as follows:

1. Solar irradiation data from 32 stations was gathered, and an in-depth investigation of ensemble learning algorithms was carried out. This investigation encompassed diverse techniques, including various averaging, boosting, bagging, stacking, and blending ensemble algorithms. To the best of our knowledge, such a comprehensive investigation has not been undertaken previously. Notably, recent studies (Alam et al. 2023) have only examined four machine-learning algorithms in their analysis.
2. We incorporated three distinct explainable AI models, namely SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanations), and Eli5 (Explain Like I'm 5), into our analysis. The data adaptation process for explainable AI yielded new insights, with a specific focus on identifying the most vital features from the Bangladesh perspective for predicting solar power output. As far as our knowledge extends, this marks the initial XAI-based analysis of solar radiation prediction utilizing a dataset from Bangladesh.
3. We have developed a hybridized method for forecasting solar radiation, which inte- grates clustering algorithms with the LightGBM algorithm. Our approach begins with the application of clustering algorithms, followed by the utilization of the LightGBM Regressor on each cluster. Specifically, we have assessed four clustering algorithms for this purpose: K-means, Mini-batch K-means, Fuzzy C-means, and Gaussian Mixture.
4. We have developed an interactive tool for predicting solar radiation based on weather data as well.
5. We employed the PyCaret machine learning library, a low-code technique, to iden- tify the most suitable model for precise prediction of solar irradiance within the specific context of Bangladesh. Additionally, we conduct a comparative analysis with the outcomes of ensemble machine learning models.

In light of the aforementioned, the following sections delineate the structure of the remainder of the paper. In Section 2, we provide an overview of the dataset, machine learning models, explainable AI, and clustering. Moving to Section 3.4, our proposed methodology encompasses four phases: Phase 1 involves leveraging ensem- ble machine-learning techniques for solar radiation prediction, followed by Phase 2, where Explainable Artificial Intelligence (XAI) is integrated to enhance interpretabil- ity. Phase 3 incorporates hybridized clustering methods for forecasting solar radiation, and in Phase 4, we develop an interactive tool for predicting solar radiation based on weather data. In Section 4.3.1, we present the results and analysis, conducting a comprehensive examination of performance evaluation metrics and assessing feature importance. Section 5 delves into summarizing the key findings. Finally, Section 6 concludes the study by summarizing the insights gained throughout the research, and potential avenues for future work in the field are highlighted.

## 2 Materials and background

In this section, we present an overview of the dataset, various machine learning mod- els including Decision Tree Regressor, Linear Regression, Ridge Regression, Lasso Regression, and Elastic Net. Additionally, we discuss Explainable AI techniques such as SHAP, LIME, and ELI5, as well as clustering methods including K-means, MiniBatchKMeans, Fuzzy C-Means, and Gaussian Mixture Model.

### 2.1 Dataset

We have collected a comprehensive set of meteorological data variables spanning the period from 1961 to 2017 from the website of Bangladesh Agricultural Research Council (BARC) (http://apps.barc.gov.bd/climate/dashboard 2023). These variables encompass Solar Radiation, Maximum Temperature, Minimum Temperature, Humidity, Cloud Coverage, Wind Speed, Sunshine (Hours), and Rainfall. These datasets were procured from 32 different weather sta- tions located throughout Bangladesh and were graciously provided by the Bangladesh Meteorological Department (BMD).

#### 2.1.1 Data preparation and preprocessing

The data is reported every month, except Solar Radiation, which is documented every ten days. To make the Solar Radiation data consistent, we have computed the monthly average by calculating the mean of three decades' worth of values for each month. Subsequently, we employed a one-to-one mapping approach that utilizes key elements such as the year, month, and station. This merged dataset, resulting from the map- ping process, has been meticulously stored as a CSV file. In total, this consolidated dataset contains 15,063 training samples and is structured with eight columns. Next, we engage in Exploratory Data Analysis (EDA) to gain insights, understand the under- lying patterns within the dataset, and check null values. Subsequently, the data is divided into an 80:20 ratio, creating separate training and testing datasets. Finally, our experiment is conducted on these distinct sets of training and testing data. Figure 1 depicts a scatter plot designed to facilitate the comprehension of trends within solar irradiance data in Bangladesh. In Fig. 2, a correlation matrix plot is presented, offering a visual exploration of the relationships among diverse sets of data. Table 1 represents the Description of the dataset and Fig. 2 presents the correlation.

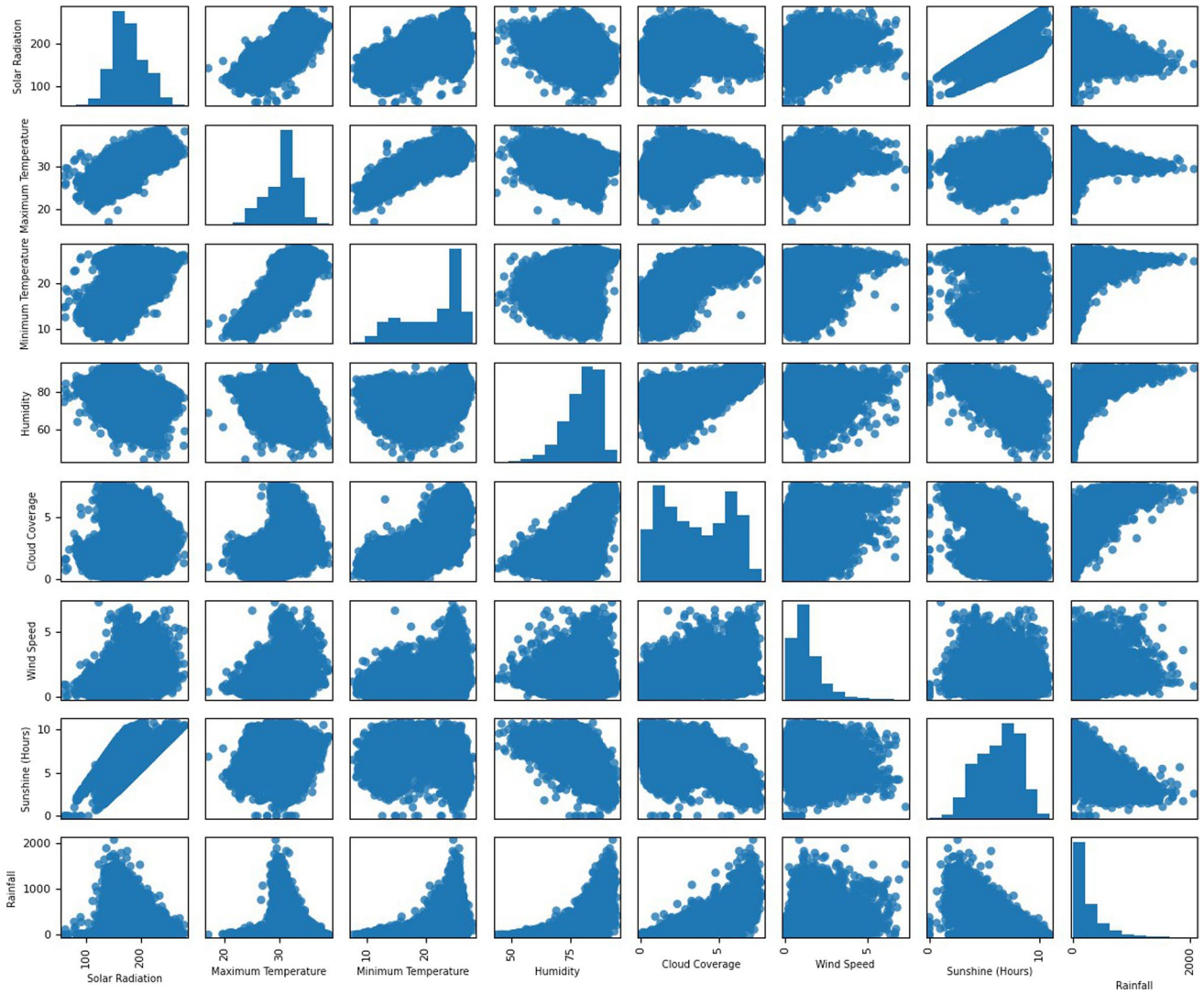



**Fig. 1** Scatter plot of Solar Irradiation dataset

## 2.2 Machine-learning model

### 2.2.1 Decision tree regressor

Decision Tree Regressor is a versatile machine-learning algorithm used for solving regression problems. It is a crucial learning algorithm that is primarily used for data analysis (Kushwah et al. 2022). It is capable of solving problems with regression and classification. It constructs a tree-like structure in which each internal node serves as a point of deci- sion or testing based on the input features, while each leaf node yields a continuous numerical prediction. It follows a recursive process of partitioning the dataset based on the values of these features to minimize prediction errors (Kushwah et al. 2022). Decision trees offer remarkable transparency, enabling users to comprehend the decision-making process and the importance of various features. However, they are vulnerable to overfitting, a situation where they fit the training data too closely, and this can be mitigated by applying pruning techniques and restricting the depth of the tree (Myles et al. 2004). Decision Tree Regressors are frequently employed as fundamental building blocks in ensemble meth- ods and have a wide range of applications in domains like finance, healthcare, and environmental science. They excel in capturing complex relationships and delivering accurate numerical forecasts (Myles et al. 2004).

### 2.2.2 Linear regression

Linear Regression is a type of supervised machine learning model that determines the linear relationship between the independent and dependent variables by construct- ing the best fit linear line between them. Mathematically, we can represent a linear regression as (Groß 2003):

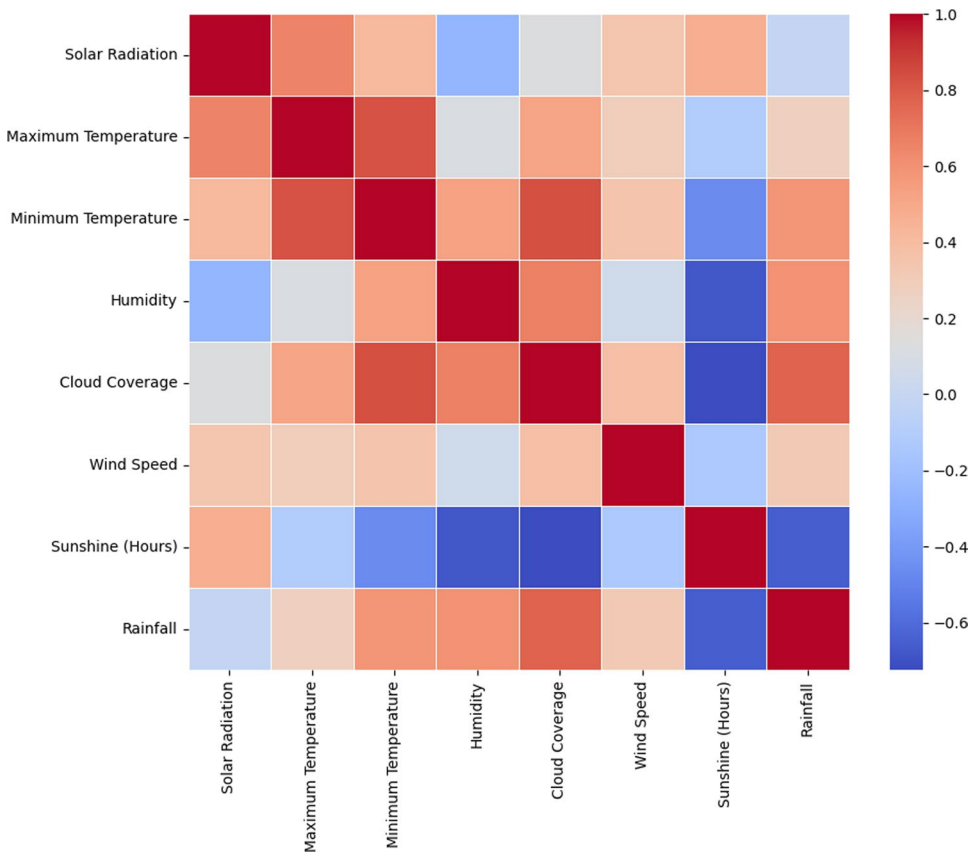



**Fig. 2** Correlation plot of Solar Irradiation dataset

**Table 1** Description of the dataset

| | Year | Month | Solar Radia- tion | Maximum Temper- ature | Minimum Temper- ature | Humi- dity | Cloud Cover- age | Wind Speed | Sunshine (Hours) | Rainfall |
|---|---|---|---|---|---|---|---|---|---|---|
| Count | 15063.00 | 15063.00 | 15063.00 | 15063.00 | 15063.00 | 15063.00 | 15063.00 | 15063.00 | 15063.00 | 15063.00 |
| Mean | 1997.32 | 6.50 | 177.22 | 30.56 | 21.31 | 79.90 | 3.57 | 1.29 | 6.34 | 205.70 |
| Std | 13.32 | 3.45 | 29.99 | 2.80 | 4.89 | 7.06 | 2.09 | 0.89 | 1.87 | 251.08 |
| Min | 1961.00 | 1.00 | 59.27 | 16.95 | 7.52 | 43.94 | 0.00 | 0.00 | 0.00 | 0.00 |
| 25% | 1989.00 | 4.00 | 156.63 | 29.12 | 17.17 | 75.74 | 1.60 | 0.69 | 4.87 | 9.00 |
| 50% | 1999.00 | 6.00 | 173.39 | 31.21 | 23.43 | 81.13 | 3.40 | 1.09 | 6.56 | 115.00 |
| 75% | 2008.00 | 10.00 | 197.26 | 32.42 | 25.48 | 85.32 | 5.63 | 1.64 | 7.81 | 318.00 |
| Max | 2017.00 | 12.00 | 281.19 | 39.16 | 28.15 | 94.52 | 7.77 | 7.28 | 10.98 | 2072.00 |

$$S_{\text{pred}} = \beta_{\text{o}} + \beta_1 \text{fea}_1 + \beta_2 \text{fea}_2 + \cdots + \beta_n \text{fea}_n + \varepsilon \quad (1)$$

where:

| | |
|---|---|
| $S_{\text{pred}}$ | the predicted or dependent variable |
| β0 | the y-intercept, |
| $\beta_1, \beta_2, \ldots, \beta_n$ | the regression coefficients for the independent variables $\text{fea}_1$, |
| fea2, . . . , fean | are the independent variables (features), |
| $\varepsilon$ | the error term, representing the unexplained variation in $S_{\text{pred}}$ that is not accounted for by the independent variables. |

Linear Regression derives its name from its emphasis on showcasing a direct link between independent variables and dependent variables. It essentially unveils how alterations in independent variable values impact the dependent variable (Groß 2003). This association is illustrated by a straight-line slope. The core premise of linear Regression hinges on the assumption of a linear connection between these variables. It presupposes minimal to no multicollinearity among independent factors (Seber and Lee 2012). In the context of linear Regression, the error term should exhibit a typical bell-shaped distribution. Deviating from this distribution could lead to challenges in estimating coefficients,

resulting in excessively wide or narrow confidence intervals (Seber and Lee 2012). Linear Regression doesn't assume autocorrelation within error terms. Any hint of error term correlation can substantially diminish the model's accuracy, typically stemming from residual errors depending on each other (Seber and Lee 2012).

**Ridge regression** Ridge Regression, also referred to as Tikhonov regularization or L2 regularization, is an advanced method in linear regression that extends the traditional least squares approach by incorporating a new dimension. Addressing the complexities arising from multicollinearity, which involves high correlations among independent variables, Ridge Regression serves as a stabilizing force, enhancing the robustness and dependability of the model in the presence of intricate data relationships (McDonald 2009). In machine learning, ridge regression reduces the standard error by incorporating a penalty term into the regression approximations. It helps to obtain estimates that are more accurate. By penalizing the weights of the feature's coefficients and lowering the difference between the actual and predicted observations, this Regression carries out L2 regularization. Besides that, it keeps the model from overfitting and makes it simpler. When the number of predictors in the data set is greater than the number of observations, it is especially advantageous (McDonald 2009).

**Lasso regression** The acronym LASSO represents the Least Absolute Shrinkage and Selection. It is a method of regularization. It is preferred over regression methods for more precise pre- diction. This model makes use of shrinkage. Shrinkage is the process by which data values are shrunk towards a central point known as the mean. The lasso procedure encourages the use of simple, sparse models (those with fewer parameters) (Wang et al. 2007). This specific kind of Regression works well for models that exhibit high levels of multi- collinearity or when you wish to automate specific steps in the model selection process, such as parameter elimination and variable selection (Wang et al. 2007). L1 regularization is used by Lasso Regression. The reason it is used is that it automatically selects features when we have more features.

**Elastic net** Elastic net linear Regression regularizes regression models by combining penalties from the lasso and ridge techniques (Liang and Jacobucci 2020). By taking into account the shortcomings of both lasso and ridge regression methods, the technique improves statistical model regularization. So Elastic Net is (Liang and Jacobucci 2020):

$$\text{Elastic Net Loss} = \text{Lasso Regression} + \text{Ridge Regression} \tag{2}$$

Elastic Net Regression is useful when a dataset has a large number of features and the goal is to prevent overfitting and perform feature selection. It is a balanced combi- nation of Lasso and Ridge Regression, making it a suitable choice when dealing with datasets where many features may be irrelevant or multicollinear. By adjusting the two regularization parameters, $\lambda_1$ and $\lambda_2$, one can fine-tune the trade-off between fea- ture sparsity and coefficient size. This method is particularly effective for handling high-dimensional data and for building models that select essential features while controlling the magnitudes of the coefficients (Liang and Jacobucci 2020).

**Bagging regressor** The baggingRegressor model functions as an ensemble of distinct predictors, working in unison to create a robust and dependable forecasting system. Think of it as a diverse group of experts, each offering a unique perspective, collaborating to provide you with precise predictions (Kadiyala and Kumar 2018).

BaggingRegressor constructs multiple foundational regression models, often uti- lizing Decision Trees, although other regressors can be employed as well. Each of these models learns from a slightly different view of the data. These base models are trained on varying subsets of the dataset, allowing them to capture different subtleties and idiosyncrasies within the data. When you require a prediction, all these indi- vidual models contribute their predictions, and the BaggingRegressor amalgamates their insights to produce a final, comprehensive prediction that mitigates the risk of overfitting and enhances prediction accuracy (Amin et al. 2023).

**Random forest regressor** Random Forest Regressor is a potent predictive tool that employs decision tree prin- ciples to boost precision and trustworthiness (Biau and Scornet 2016). It acts as a diverse committee of decision-makers, with each tree offering a distinct viewpoint. Together, they collabo- rate like a council, voting to determine the most probable outcome, reducing the risk of mistakes, and yielding precise predictions. This method benefits from diversity, as each tree is trained on different data subsets and examines random sets of features. This diversity prevents excessive reliance on specific data patterns and increases resilience to noise and outliers (Biau and Scornet 2016). Furthermore, it can gauge feature importance, revealing influential variables and providing insights into predictive factors.

Random Forest Regressor operates as a collective of experts, pooling their wisdom to provide robust forecasts. This makes it a popular choice in fields requiring accu- rate predictions, showcasing its exceptional ability to harness individual decision trees' strengths for superior predictive performance. The equation for Random Forest Regressor can be expressed as follows:

$$\hat{y}(\text{inp}) = \frac{1}{\text{tot\_dt}} \sum_{d1}^{\text{tot\_dt}} f_d(\text{inp}) \tag{3}$$

where:

$y\hat{}(\text{inp})$ the predicted output for the input inp.
tot dt the total number of decision trees (estimators) in the random forest.
$f_d(\text{inp})$ the prediction made by the $d$-th decision tree in the forest.

**Gradient boosting regressor** Gradient Boosting Regressor stands out as a formidable tool in the realm of machine learning, particularly suited for tackling regression tasks. Its underlying mechanism revolves around the iterative amalgamation of predictions from numerous feeble learn- ers (Natekin and Knoll 2013). With each iteration, the model zeroes in on rectifying the mistakes made in previous rounds, making adjustments to its forecasts in an effort to minimize those errors. This dynamic process persists until a predetermined number of iterations is reached or until the model's performance levels off (Natekin and Knoll 2013). The Gradient Boosting Regres- sor is renowned for its exceptional accuracy, resilience, and proficiency in handling intricate data relationships, rendering it a sought-after choice for predictive modeling across diverse domains.

**XGBoost regressor** The XGBoost Regressor, also known as Extreme Gradient Boosting, is a potent machine-learning algorithm designed for regression tasks (Chen and Guestrin 2016). It operates as an ensem- ble learning technique that amalgamates the forecasts generated by multiple decision tree models, resulting in precise and resilient predictions. What sets XGBoost apart is its exceptional efficiency, speed, and aptitude for handling intricate, non-linear asso- ciations within data. It leverages the principles of gradient boosting, which entails the continuous reduction of the error from prior models by training new models on the residual errors from their predecessors (Chen and Guestrin 2016). XGBoost enjoys widespread popularity in diverse fields of data science and machine learning, largely owing to its exceptional performance and adaptability.

**LightGBM regressor** LightGBM Regressor is a state-of-the-art machine learning model tailored for regres- sion tasks. It is well-known for its remarkable speed and precision, resembling a high-performance sports car in the field of machine learning (Ke et al. 2017). What sets LightGBM apart is its ability to efficiently process data with minimal memory usage, enabling it to handle large datasets and complex feature spaces adeptly. It operates akin to a tur- bocharged engine for regression, swiftly and accurately navigating through the data. However, its appeal extends beyond swiftness; LightGBM is also a finely tuned instru- ment. It optimizes decision trees, adapting to the intricacies of the data to construct a model that comprehends intricate relationships. LightGBM shines in handling both numerical and categorical features, showcasing its versatility across a wide spectrum of regression tasks. Additionally, its parallel computing capacity further elevates its efficiency, delivering a seamless user experience (Ke et al. 2017).

**AdaBoost regressor** AdaBoost Regressor stands out as a potent ensemble learning technique tailored for building resilient regression models. Its operation can be likened to an orchestra, where individual weak learners, akin to musicians, contribute their part (Schapire 2003). In this analogy, AdaBoost assumes the role of a conductor, orchestrating and synchronizing these elements. Notably, it emphasizes the learners that have previously made errors, resembling a conductor rectifying discordant notes. The ensemble hones its perfor- mance through iterative refinement, aiming to minimize prediction inaccuracies (Schapire 2003). AdaBoost amalgamates the strengths of these weak learners, crafting a robust regres- sion model adept at capturing intricate data patterns. The ultimate model mirrors a seamless blend of diverse contributions, enabling precise predictions for the target variable. AdaBoost's proficiency in handling intricate data structures and delivering precise regression outcomes positions it as an invaluable asset in the realm of machine learning.

**CatBoost regressor** CatBoost Regressor is a cutting-edge machine learning algorithm tailored specifically for regression tasks. It stands out for its seamless handling of categorical features, eliminating the need for extensive preprocessing and enhancing user-friendliness, especially for mixed data types (Prokhorenkova et al. 2018). Inspired by a cat's adaptability, it strikes a balance between bias and variance, ensuring reliable predictions and preventing over- fitting by learning iteratively through gradient boosting, akin to a cat mastering a skill through practice.

This algorithm excels in uncovering intricate patterns within large and complex datasets, offering valuable insights that might elude other methods. Its speed and accuracy make it ideal for real-time applications where quick and precise predictions are crucial (Prokhorenkova et al. 2018). CatBoost Regressor acts as a vigilant cat, meticulously exploring data and using its analytical"claws" to extract meaningful information effortlessly. Its robustness and versatility have made it a preferred choice among data scientists, ensuring reliable results across various regression tasks. In essence, CatBoost Regres- sor combines adaptability, precision, and efficiency, simplifying the complexities of regression modeling and data analysis for professionals and enthusiasts alike.

## 2.3 Explainable artificial intelligence (XAI)

Explainable Artificial Intelligence (XAI) is a field of AI research focused on develop- ing machine learning models and algorithms that can provide clear and interpretable explanations for their decisions and predictions (Sevas et al. 2023). The primary objective of XAI is to bridge the gap between the inherent intricacy of machine learning models and the demand for clarity and responsibility in AI systems (Pierrot and Goude 2011). Its purpose is to enable users to grasp the rationale behind a specific AI system's choices. This amplified transparency holds immense importance for adhering to regulations, fostering user confidence, and advancing collaboration between humans and AI.

### 2.3.1 SHAP (shapley additive explanations)

SHAP is a powerful framework rooted in cooperative game theory, designed to unravel the inner workings of complex machine learning models, including black-box ones like deep neural networks and ensembles (Gillies 2013). It introduces a unified metric for assessing feature importance, employing the Shapley value concept from cooperative game the- ory. Features are treated as players, and the model's prediction is distributed fairly among them. Various techniques, such as Kernel SHAP, Tree SHAP, and Linear SHAP, are used to compute SHAP values based on the model type (Jay and Cockett 1994). These values quan- tify the influence of individual features, aiding in model interpretability. Visualization methods like summary plots, force plots, and dependence plots are utilized to repre- sent SHAP values visually, providing crucial insights into the model's behavior. This enhances transparency and interpretability, making it easier to understand and trust the model's predictions, particularly in sensitive applications. Overall, SHAP con- tributes significantly to improving the transparency and interpretability of machine learning models, promoting understanding and confidence in their outcomes.

The Shapley value for feature *ft* is calculated as (Sushanth et al. 2023):

$$\phi_{\mathrm{ft}} = \sum \frac{\mathrm{V}\left(\mathrm{S}_{\mathrm{shap}} \cup \{\mathrm{ft}\}\right) - \mathrm{V}\left(\mathrm{S}_{\mathrm{shap}}\right)}{\left(\left|\mathrm{S}_{\mathrm{shap}}\right| + 1\right)!} \tag{4}$$

where:

| | |
|---|---|
| $\phi_{ft}$ | the Shapley value for feature *ft*. |
| $V\,(S_{shap} \cup \{ft\})$ | the prediction when feature *ft* is included in the subset $S_{shap}$. |
| $V\,(S_{shap})$ | the prediction when feature *ft* is not included in the subset $S_{shap}$. |
| $\lvert S_{shap} \rvert$ | the number of features in subset $S_{shap}$. |
| $(\lvert S_{shap} \rvert + 1)!$ | the number of all possible permutations of subsets that include feature *ft*. |

### 2.3.2 LIME (local interpretable model-agnostic explanations)

LIME is a powerful tool in machine learning interpretability, designed to reveal the decision-making process of complex"black box" models. Instead of explaining the entire model, LIME focuses on specific instances, employing a model-agnostic approach to create a clear and understandable local model that mimics the black box model's behavior for a particular data point (Biparva and Materassi 2023). The challenge lies in finding the right balance between local model complexity and the quality of explanation. LIME is valuable for data scientists and decision-makers, providing insight into the"why" of specific pre- dictions, fostering trust, and promoting transparency in machine learning applications (Biparva and Materassi 2023).

The mathematical representation of LIME is as follows (Zhang et al. 2022):

$$\mathrm{Exp}(z) = \underset{gn}{\operatorname{argmin}} \left[ L(fn, gn, z) + \lambda(gn) \right] \tag{5}$$

where:

| | |
|---|---|
| Exp($z$) | the explanation provided by LIME for a specific instance $z$. |
| arg min | the interpretable model *gn* that minimizes the following expression *gn*. |
| $L(fn, gn, z)$ | the loss function quantifying the difference between the predictions of the black-box model *fn* and the interpretable model *gn* for the instance z. |
| $\lambda(gn)$ | a complexity term assessing the complexity of the interpretable model *gn*. |

#### 2.3.3 ELI5 (explain like i'm 5)

ELI5 is a widely used internet acronym that people employ when they want a simple and easily understandable explanation for complex topics (El-Sappagh et al. 2021). Essentially, when some- one uses "ELI5," they're asking for information to be broken down to a level that even a 5-year-old child could grasp. This method is particularly useful for making intricate subjects like science, technology, law, or economics accessible to a broad audience (El-Sappagh et al. 2021). The primary objective of "ELI5" is to promote inclusive understanding by encourag- ing clear and straightforward communication. When individuals use this term, they typically want answers that steer clear of technical jargon and complicated termi- nology (Chadaga et al. 2023). The aim is to make even the most challenging ideas understandable to anyone, regardless of their level of expertise in a given field. In response to a request for an "ELI5" explanation, people often aim to provide brief and relatable answers. It's a collaborative effort within the community to bridge the knowledge gap and enhance understanding among individuals with varying levels of expertise (Chadaga et al. 2023). This approach has gained popularity as a means to enhance online discussions, making them more informative and engaging while facilitating the sharing of knowledge in a more accessible manner (Kuzlu et al. 2020).

### 2.4 Clustering

#### 2.4.1 K-means

K-means clustering is a key unsupervised machine learning algorithm that partitions a dataset into K clusters based on similarities (Ayodele et al. 2019). It starts by randomly selecting K cluster centroids and assigns each data point to the nearest centroid. The centroids are then updated by computing the average position of data points within each cluster, and this process repeats iteratively until convergence (Ayodele et al. 2019). K-means is known for its computational efficiency and scalability, making it widely used in various applications (Sinaga and Yang 2020). However, selecting the appropriate value for K can be challenging, often requiring domain knowledge or techniques like the elbow method. Despite assuming spherical, equal-sized, and non-overlapping clusters, K-means remains a powerful tool for uncov- ering patterns in data across diverse domains, such as marketing, image analysis, and natural language processing.

The equation for K-means clustering is represented as (Sinaga and Yang 2020):

$$\text{DistanceCalculation} : d\left(\text{point}_i, \text{cen}_z\right) = \sqrt{\sum_{k=1}^{n} \left(\text{point}_{ik} - \text{cen}_{zk}\right)^2} \tag{6}$$

$$\text{CentroidUpdate} : \text{cen}_z' = \frac{1}{|\text{Set}_z|} \sum_{\text{point}_i \in \text{Set}_z} \text{point}_i \tag{7}$$

where:

- $n$ the number of features or dimensions in the data
- $\text{point}_i$: a data point.
- $\text{cen}_z$ a cluster centroid.
- $\text{Set}_z$ the set of data points assigned to cluster $z$.
- $|\text{Set}_z|$ the number of data points in cluster $z$.

#### 2.4.2 MiniBatchKMeans

MiniBatchKMeans serves as a clustering algorithm for segregating a dataset into distinct clusters based on the similarity between data points (Chavan et al. 2015). It represents a modification of the conventional KMeans clustering method, specifically tailored to efficiently manage sizable datasets. In contrast to the standard KMeans, which pro- cess the entire dataset during each iteration, MiniBatchKMeans operates on randomly selected subsets or mini-batches of the data (Peng et al. 2018). This approach enhances computa- tional efficiency and enables the algorithm to handle large datasets that may exceed available memory constraints. The algorithm iteratively updates cluster centroids with each mini-batch, progressively converging towards a solution that accurately reflects the inherent structure of the data (Peng et al. 2018).

#### 2.4.3 Fuzzy C-means

Fuzzy C-Means (FCM) clustering is a mathematical approach employed in data anal- ysis and pattern recognition to segment a dataset into clusters based on similarity (Bezdek et al. 1984). Unlike conventional crisp or"hard" clustering techniques, FCM introduces the concept of degrees of membership, allowing each data point to be associated with mul- tiple clusters to varying extents (Bezdek et al. 1984). In FCM, data points are assigned membership values for each cluster, reflecting the degree to which the point belongs to that cluster (Cannon et al. 1986). These membership values are expressed as fuzzy sets, offering greater flexibility compared to traditional clustering methods. The algorithm iteratively adjusts cluster centers and membership values until convergence, striving to minimize an objective function that measures the overall fuzziness of the clustering arrangement (Cannon et al. 1986).

#### 2.4.4 Gaussian mixture model

Gaussian Mixture Model (GMM) clustering is a probabilistic approach for cluster- ing and density estimation (Weber et al.

2022). Unlike traditional methods, GMM views a dataset as a mixture of Gaussian distributions, each associated with a cluster. The model aims to estimate parameters like mean, covariance, and mixing coefficients for these distributions (Weber et al. 2022). The Expectation–Maximization (EM) algorithm is often used for fitting a GMM by iteratively computing the likelihood of data points belonging to clusters and updating distribution parameters. GMM is valuable for complex datasets with unclear cluster boundaries or overlapping clusters (Weber et al. 2022). It finds applications in diverse fields, such as image processing and speech recognition, where capturing the underlying probability distribution is essential.

## 3 Proposed methodology

We divided our proposed methodology into four parts:

- Phase 1: Ensemble Machine-Learning Techniques for Solar Radiation Prediction
- Phase 2: Explainable Artificial Intelligence (XAI) in the context of Solar Radiation Prediction
- Phase 3: Hybridized Clustering to Forecast Solar Radiation.
- Phase 4: An interactive tool for predicting solar radiation based on weather data.

During Phase 1, we employ five ensemble machine-learning techniques: averaging, stacking, blending, bagging, and boosting. These methods combine multiple models to improve prediction accuracy. Moving on to Phase 2, we delve into the inner workings of machine learning models to understand how each feature influences the model's pre- dictions. To achieve this, we leverage three Explainable AI techniques: SHAP, LIME, and ELI5. In Phase 3, we segment the dataset into clusters and make predictions on the clustered data using the LightGBM model. Finally, in Phase 4, we have developed a tool for forecasting solar radiation based on weather data. Figure 3 illustrates our proposed methodology.

### 3.1 Phase 1: ensemble machine-learning techniques for solar radiation prediction

In this section we will describe five ensemble machine-learning techniques: boosting, blending, bagging, stacking, and averaging. These approaches combine several models in an effort to improve forecast precision.

#### 3.1.1 Averaging

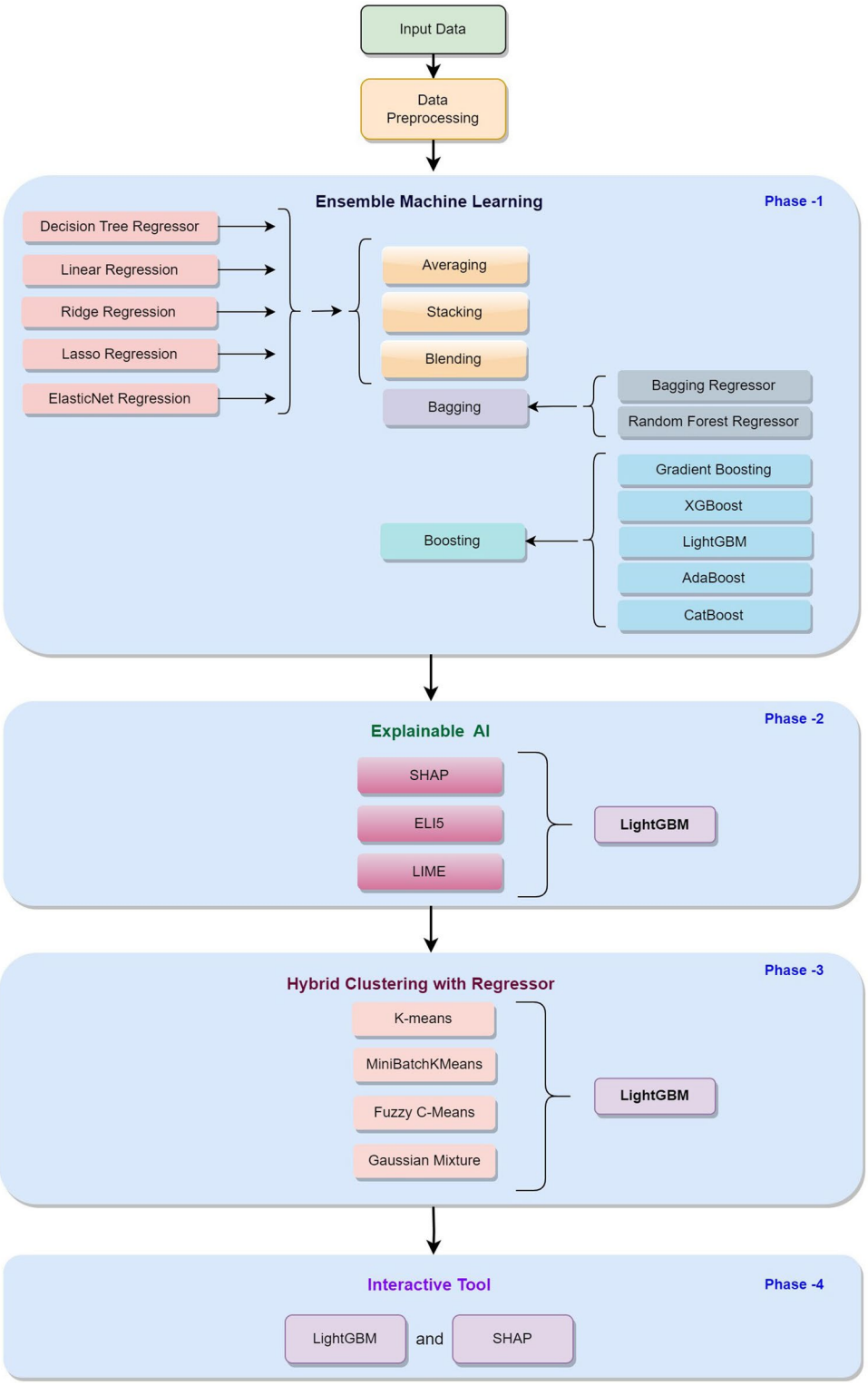
An averaging ensemble can be likened to the harmonious convergence of knowledge within the domain of machine learning. Picture a council composed of a multitude of diverse experts, each bringing their unique perspectives to the table, united in their effort to make a collective and more well-informed decision. In this scenario, these experts are individual machine learning models, and their decision-making process involves consolidating their viewpoints through averaging to reach a final predic- tion that is more precise (Polikar 2012). In our approach, we utilize a combination of five distinct models for Averaging Ensemble Machine-Learning Technique. These mod- els include Decision Tree Regressor $M_1$, Linear Regression$M_2$, Ridge Regression$M_3$, Lasso Regression$M_4$, and ElasticNet Regression $M_5$ and we have also calculated the weighted average of these five methods.

Figure 4 depicts the averaging ensemble learning for solar irradiation prediction. Algorithm 1 represents pseudo code for solar irradiance prediction using Averaging Ensemble Machine-Learning Technique.

**Algorithm 1** Solar irradiance prediction using averaging ensemble machine- learning technique

Step 1: Load the dataset of Solar Radiation, *df*
Step 2: Perform Exploratory Data Analysis on the dataset with feature values denoted as $F = f_1, f_2, f_3, f_4, \ldots, f_n$
Step 3: Apply feature selection and split the dataset into *df_train* and *df_test*
Step 4: Training the train dataset *df_train* with different machine-learning models, $M_1, M_2, \ldots, M_k$
Step 5: Averaging the predictions, $pred_{final} = (pred_1 + pred_2 + \ldots\ldots + pred_k)/k$.
Step 6: set the weighted $weights = w1, w2, w3, w4, w5$
Step 7: Calculate the weighted average of predictions
Step 8: Calculate and store three different error values using $pred_{final}$ and *df_test*, $E_{R^2}$, $E_{MAE}$ and $E_{RMSE}$.

**Fig. 3** Proposed Methodology




### 3.1.2 Bagging

Bagging is a powerful ensemble machine learning technique that combines ”bootstrap” and ”aggregating” to improve predictive models by leveraging diversity (Dietterich et al. 2002). Its key strength is mitigating variance by training multiple instances of a base model on dif- ferent subsets of training data with replacement. This introduces heterogeneity and prevents fixation on training data idiosyncrasies and noise. During prediction, these independently trained models

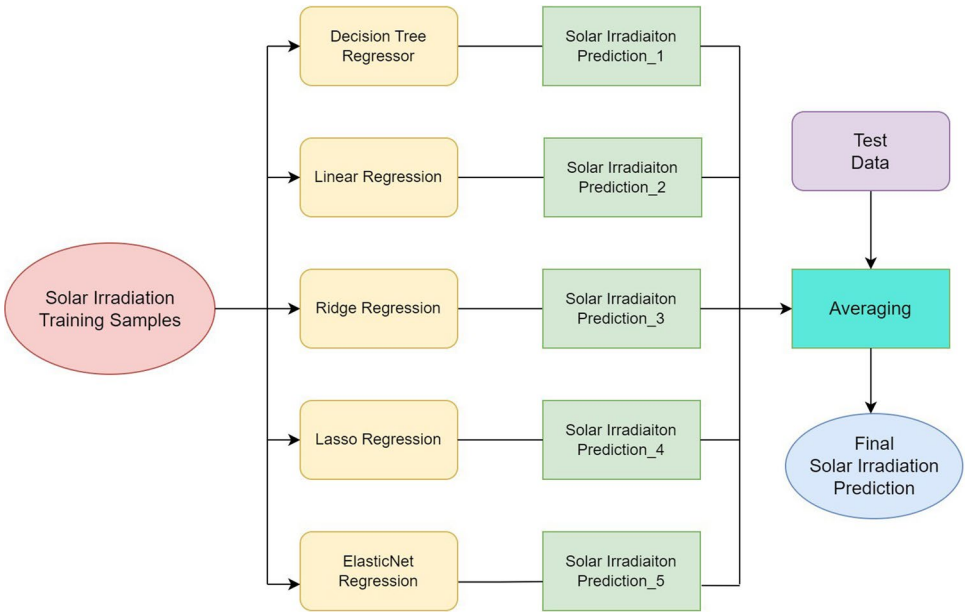



**Fig. 4** Averaging ensemble learning for solar irradiation prediction

contribute their opinions, resulting in a more precise and reliable forecast than a single model could provide. Bagging democratizes the influence of individual models, reducing the risk of overfitting and enhancing general- ization to new data (Dietterich et al. 2002).

Bagging is valued for its simplicity and efficiency and is a fundamental tool in ensem- ble learning, with applications ranging from random forests, a collection of bagged decision trees, to other models benefiting from its stabilizing and predictive potential. Despite its unassuming name, bagging elevates the predictive capabilities of machine learning models. Figure 5 depicts the Bagging Ensemble Learning for Solar Irradiation Prediction.

In our approach, we employ five different models as the base estimators for bagging: decision tree regression, linear regression, ridge regression, Lasso regression, and Elas- tic-Net regression. Algorithm 2 represents pseudo code for solar irradiance prediction using Bagging Regressor.

Algorithm 3 presents pseudo code for forecasting solar irradiance utilizing a Random Forest Regressor.

**Algorithm 2** Solar irradiance prediction using bagging regressor

Step 1: Load the dataset of Solar Radiation, $df$
Step 2: Perform Exploratory Data Analysis on the dataset with feature values denoted as $F = f_1, f_2, f_3, f_4, \ldots, f_n$
Step 3: Apply feature selection and split the dataset into $df\_train$ and $df\_test$
Step 4: Training the train dataset $df\_train$ with different base models for $base\_estimator$ used in the bagging approach, $M_1, M_2, \ldots, M_k$
Step 5: Calculating errors using $df\_test$ for $M_1, M_2, \ldots, M_k$ models, $pred_1, pred_2, \ldots, pred_k$.
Step 6: Store three different error values, $E_{R^2}$, $E_{MAE}$ and $E_{RMSE}$ for each model.

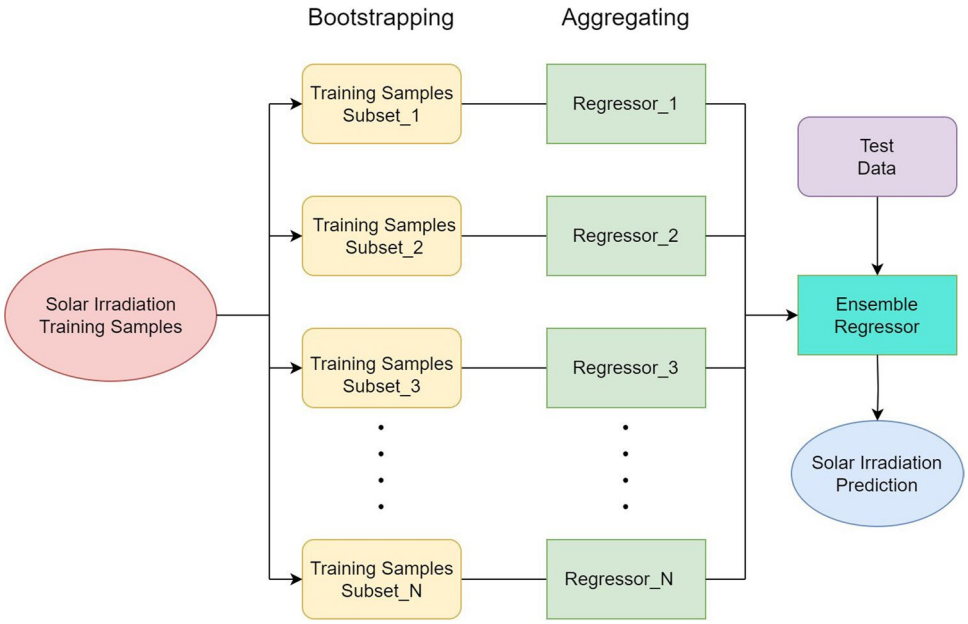



**Fig. 5** Bagging ensemble learning for solar irradiation prediction

**Algorithm 3** Solar irradiance prediction using random forest regressor

Step 1: Load the dataset of Solar Radiation, *df*
Step 2: Perform Exploratory Data Analysis on the dataset with feature values denoted as $F = f_1, f_2, f_3, f_4, \ldots, f_n$
Step 3: Apply feature selection and split the dataset into *df train* and *df test*
Step 4: Training the train dataset *df train* with Random Forest Regressor, *RF*
Step 5: Calculating errors using *df test*.
Step 6: Store three different error values, $E_{R^2}$, $E_{MAE}$ and $E_{RMSE}$.

### 3.1.3 Boosting

Boosting is an ensemble machine learning technique that combines predictions from multiple weak learners to create a robust and accurate predictive model (Sagi and Rokach 2018). Unlike traditional algorithms, boosting trains weak learners sequentially, with each learner focused on correcting the errors of its predecessors. In each iteration, boosting assigns greater importance to misclassified data points, compelling subsequent learners to pay more attention to challenging instances. This iterative process efficiently captures complex patterns in the data. Boosting utilizes a weighted voting system in the final model, where each weak learner contributes to predictions based on their precision, effectively elevating them to strong predictors. (Sagi and Rokach 2018).

Figure 6 illustrates the Boosting Ensemble Learning method for Solar Irradiation Prediction and and Algorithm 4 represents pseudo code for solar irradiance prediction using Boosting Regressor.

### 3.1.4 Stacking

Stacking stands out as an advanced ensemble learning method that fuses forecasts from a variety of base models through the use of a meta-model (Pavlyshenko 2018). This amalgama- tion leverages their strengths and mitigates their shortcomings. Unlike conventional ensemble techniques, Stacking doesn't simply take an average of predictions; instead, it learns how to assign weights to inputs from various models wisely. It assembles a diverse group of models, including decision trees, support vector machines, and neural networks, each adept at capturing distinctive data patterns (Pavlyshenko 2018). These varied view- points are harmonized by the meta-model, resulting in a well-balanced and precise predictive system.

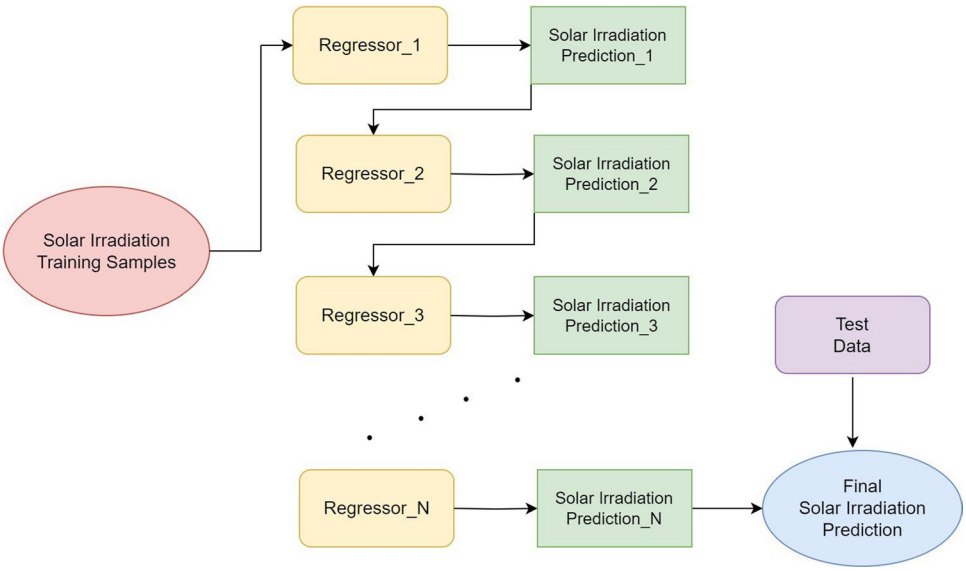



**Fig. 6** Boosting ensemble learning for solar irradiation prediction

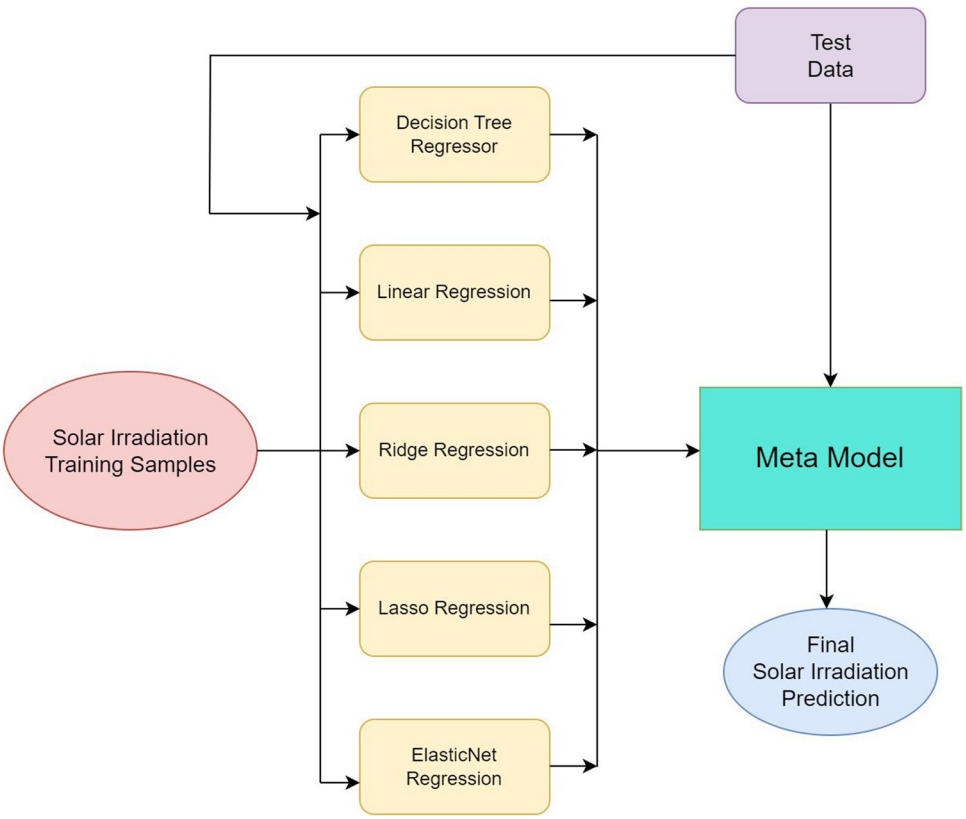



**Fig. 7** Stacking ensemble learning for solar irradiation prediction

**Algorithm 4** Solar irradiance prediction using boosting regressor

Step 1: Load the dataset of Solar Radiation, $df$
Step 2: Perform Exploratory Data Analysis on the dataset with feature values denoted as $F = f_1, f_2, f_3, f_4, \ldots, f_n$
Step 3: Apply feature selection and split the dataset into $df\_train$ and $df\_test$
Step 4: Training the train dataset $df\_train$ with different models for the Boosting approach, $M_1, M_2, \ldots, M_k$
Step 5: Calculating errors using $df\_test$ for $M_1, M_2, \ldots, M_k$ models, $pred_1, pred_2, \ldots, pred_k$.
Step 6: Store three different error values, $E_{R^2}$, $E_{MAE}$ and $E_{RMSE}$ for each model.

Stacking enables the utilization of specialized models tailored to different sets of features, ensuring that no valuable information is left unexplored (Kwon et al. 2019). However, metic- ulous model selection and fine-tuning are imperative for its success, resembling the process of assembling an impeccable ensemble cast for a movie. Ultimately, stacking exemplifies the idea that the collaborative intelligence of diverse models can surpass individual endeavors, rendering it a potent tool in the realm of machine learning.

In our methodology, we employ a fusion of five models to implement the Stack- ing Ensemble Machine Learning approach. These models comprise the Decision Tree Regressor, Linear Regression, Ridge Regression, Lasso Regression, and ElasticNet Regression. Finally, we utilize these five models as the meta-model and find the best model for Stacking. Figure 7 illustrates the Stacking Ensemble Learning approach for predicting Solar Irradiation. Algorithm 5 represents pseudo code for solar irradiance prediction using the Stacking Ensemble Machine-Learning Technique.

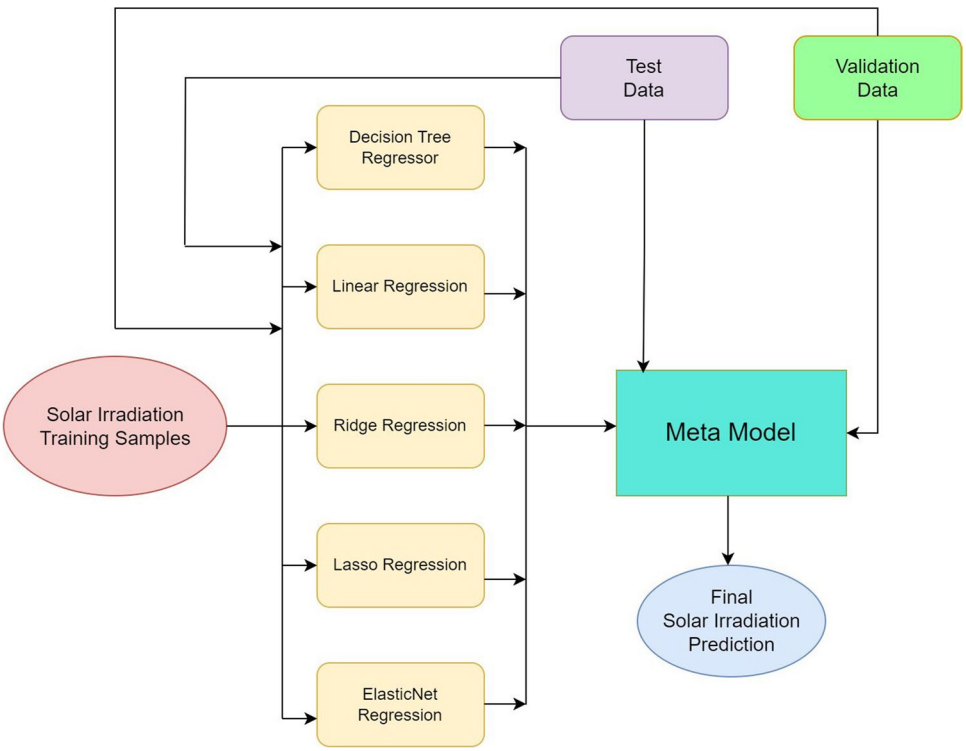



**Fig. 8** Blending ensemble learning for solar irradiation prediction

**Algorithm 5** Solar irradiance prediction using stacking ensemble machine- learning technique

Step 1: Load the dataset of Solar Radiation, *df*
Step 2: Perform Exploratory Data Analysis on the dataset with feature values denoted as $F = f_1, f_2, f_3, f_4, \ldots, f_n$
Step 3: Apply feature selection and split the dataset into *df_train* and *df_test*
Step 4: Selecting Meta Model, $M_{meta}$ for *final_estimator*.
Step 5: Training the train dataset *df_train* with different machine-learning models, $M_1, M_2, \ldots, M_k$ with Meta Model, $M_{meta}$.
Step 6: Calculate and store three different error values using $pred_{final}$ and *df_test*, $E_{R^2}$, $E_{MAE}$ and $E_{RMSE}$.

### 3.1.5 Blending

Blending ensemble regression is a sophisticated technique in machine learning used to improve prediction accuracy in regression tasks. It begins by splitting the training data into a training set and a validation set, where multiple regression models, such as linear regression and decision trees, are trained on the training data. These models generate predictions for both the validation and test sets, and the actual target values of the validation set are retained. A meta-model, often a regression model like linear regression or ridge regression, is then trained using the validation set's true target val- ues and the predictions from the base models on the validation set. The meta-model learns how to optimally combine the base models' predictions by determining the best coefficients (Wu et al. 2021).

Once the meta-model is trained, it is used to make the final predictions on the test set, leveraging the strengths of the base models to provide more accurate and robust results for continuous target variables. This approach is particularly valuable for handling complex and noisy datasets and helps improve model generalization while mitigating overfitting (Wu et al. 2021). Blending ensemble regression offers a powerful means to enhance pre- dictive accuracy by drawing upon the collective insights of diverse regression models, resulting in more reliable and accurate predictions in regression tasks.

In our method, we employ a blend of five unique models to create a Blending Ensemble Machine-Learning Technique. These models comprise the Decision Tree Regressor, Linear Regression, Ridge Regression, Lasso Regression, and ElasticNet Regression. Finally, we utilize these five models as the Meta Model and find the best model for Blending. Figure 8 demonstrates the Blending Ensemble Learning method for Solar Irradiation Prediction. Algorithm 6 represents pseudo code for solar irradiance prediction using the Blending Ensemble Machine-Learning Technique.

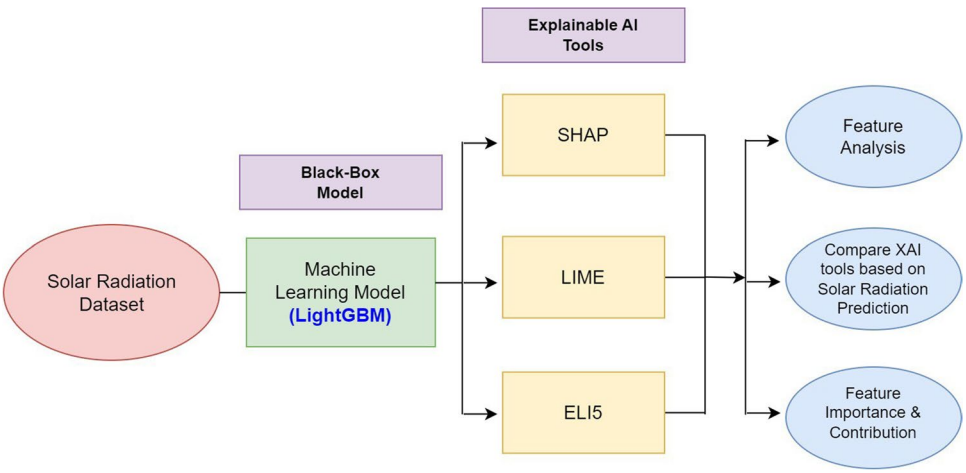



**Fig. 9** Explainable artificial intelligence (XAI) in the context of solar radiation prediction

**Algorithm 6** Solar irradiance prediction using blending ensemble machine- learning technique

Step 1: Load the dataset of Solar Radiation, *df*
Step 2: Perform Exploratory Data Analysis on the dataset with feature values denoted as $F = f_1, f_2, f_3, f_4, \ldots, f_n$
Step 3: Apply feature selection and split the dataset into *df train* , *df_valid*, *df_test*.
Step 4: Training the train dataset *df train* with different machine-learning models, $M_1, M_2, \ldots, M_k$.
Step 5: Make predictions on *df_valid* and *df test* for each model and combine their results in *com_ valid* and *com_test*.
Step 6: Selecting Meta Model, $M_{meta}$.
Step 7: Make predictions on *com_ valid* and *com_test* using $M_{meta}$.
Step 8: Calculate and store three different error values using $pred_{final}$ and *df_test*, $E_{R^2}$, $E_{MAE}$ and $E_{RMSE}$.

### 3.1.6 PyCaret

PyCaret (Ali 2020) is a groundbreaking Python library designed for data scientists and machine learning practitioners, simplifying the journey from raw data to predictive models. As an open-source, low-code framework, PyCaret efficiently manages the entire machine-learning pipeline, offering a user-friendly interface for tasks like data preprocessing, feature engineering, model selection, hyperparameter tuning, evaluating multiple machine-learning models, and deployment. By automating routine tasks such as missing value imputation and categorical variable encoding, PyCaret allows users to focus on core aspects of model creation. Its standout feature is automated model selec- tion, facilitating quick comparison and selection of the most effective algorithms for specific datasets. PyCaret goes beyond conventional libraries by incorporating model interpretation and providing insightful visualizations of the importance of features and SHAP values. It extends its utility to anomaly detection, ensuring a comprehensive approach to data analysis. In essence, PyCaret is more than a library; it is an inno- vation catalyst, democratizing access to machine learning by offering a user-friendly yet robust toolkit. It makes the art and science of predictive modeling accessible to both novices and seasoned practitioners, emphasizing simplicity in navigating the complexities of machine learning tasks.

## 3.2 Phase 2: explainable artificial intelligence (XAI) in the context of solar radiation prediction

In this section, we begin to understand the complex workings of machine learning models and analyze how each feature affects predictions. We use three Explainable AI approaches—SHAP, LIME, and ELI5—to achieve this objective.

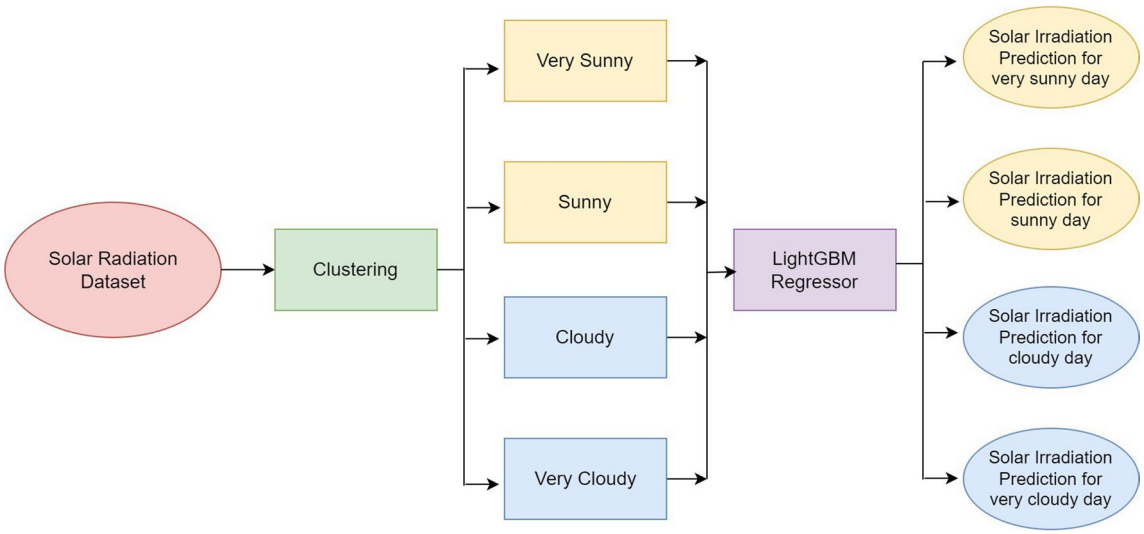



**Fig. 10** Hybridized clustering to forecast solar radiation

We employed three unique Explainable Artificial Intelligence (XAI) methodologies, specifically SHAP, LIME, and ELI5, in combination with LightGBM to forecast Solar Irradiance to gain insights into the logic behind predictions made by ML models. This effort aims to improve clarity, trustworthiness, and accountability, thereby facilitating more informed decision-making processes. Figure 9 illustrates Explainable Artificial Intelligence (XAI) to predict Solar Irradiance. Algorithm 7 represents pseudo code for XAI for Solar Radiation Prediction.

**Algorithm 7** Explainable artificial intelligence (XAI) in the context of solar radiation prediction

Step 1: Load the dataset of Solar Radiation, *df*
Step 2: Perform Exploratory Data Analysis on the dataset with feature values denoted as $F = f_1, f_2, f_3, f_4, \ldots, f_n$
Step 3: Apply feature selection and split the dataset into *df_train* and *df_test*
Step 4: Training the train dataset *df_train* with LightGBM
Step 5: Apply XAI tools (SHAP, LIME, ELI5)
Step 6: Analysis features importance and contribution.

### 3.3 Phase 3: hybridized clustering to forecast solar radiation

As an additional contribution of this study, this section focuses on the clustering aspect. Here, we detail the process of partitioning the dataset into clusters and employing the LightGBM model to generate predictions based on the clustered data. Figure 10 demonstrates Hybridized Clustering for Solar Radiation Prediction. We initially segment the solar dataset into four clusters using the K-means clustering algorithm, categorizing them as"Very Sunny,""Sunny,""Cloudy," and"Very Cloudy."

Subsequently, we proceed to partition each cluster data into separate training and testing sets. Following this, we employ LightGBM Regressor to generate predictions for each cluster's data.

Algorithm 8 represents pseudo code for Hybridized Clustering to forecast Solar Radiation.

**Algorithm 8** Solar irradiance prediction using hybridized clustering

Step 1: Load the dataset of Solar Radiation, *df*
Step 2: Perform Exploratory Data Analysis on the dataset with feature values denoted as, $F = f_1, f_2, f_3, f_4, \ldots, f_n$
Step 3: Apply feature selection and set k=4 for k-means clustering
Step 4: Apply Clustering Algorithm
Step 5: Split the dataset into 4 clusters, $C_1, C_2, C_3, C_4$
Step 6: Split the 4 clusters dataset into $train_1, train_2, \ldots, train_k$ and $test_1, test_2, \ldots, test_k$
Step 7: Training the train datasets $train_1, train_2, \ldots\ldots, train_k$ with LightGBM
Step 8: Calculating errors using $test_1, test_2, \ldots\ldots, test_k$
Step 9: Store three different error values, $E_{R^2}$, $E_{MAE}$ and $E_{RMSE}$ for each cluster.

### 3.4 Phase 4: an interactive tool for predicting solar radiation based on weather data

We have developed an interactive tool with Streamlit (Khorasani and Abdou 2022) that forecasts solar radiation using provided weather data. We employed the LightGBM machine learning model for making predictions. Additionally, we have incorporated SHAP to visualize the importance of features and contributions in the analysis of the predictions.

**Table 2** Evaluation metrics for averaging approach

| Ensemble Machine-Learning | $R^2$ Score | MAE | RMSE |
|---|---|---|---|
| Linear Regression | 0.83 | 9.96 | 12.28 |
| Ridge Regression | 0.83 | 9.96 | 12.28 |
| Lasso Regression | 0.83 | 10.16 | 12.44 |
| ElasticNet Regression | 0.80 | 10.95 | 13.55 |
| Decision Tree Regressor | 0.80 | 8.05 | 13.41 |
| Averaging | **0.86** | 9.11 | 11.47 |
| Weighted Average | 0.85 | 9.14 | 11.51 |

## 3.5 Evaluation metrics

### 3.5.1 R-squared ($R^2$) Score

$R^2$ score, also referred to as the coefficient of determination, is a statistical metric employed to evaluate the effectiveness of a regression model. It quantifies the pro- portion of the variability in the dependent variable that can be accounted for by the independent variables included in the model. Spanning a scale from 0 to 1, an $R^2$ score of 1 signifies a perfect fit of the model to the data, explaining all the variability. Con- versely, an $R^2$ score of 0 implies that the model does not contribute any explanatory power. In practical terms, $R^2$ serves as a valuable tool for assessing how accurately a regression model captures and anticipates the underlying data patterns, making it a valuable measure in the context of regression analysis and model evaluation.

The equation for $R^2$ score is represented as follows:

$$R^2 = 1 - \frac{\sum_{obs=1}^{total} \left(z_{obs} - \widehat{z_{obs}}\right)^2}{\sum_{obs=1}^{total} \left(z_{obs} - \bar{z}\right)^2} \tag{8}$$

where,

*total* the collective count of observations,
$z_{obs}$ the actual value of the *obs*-th observation,
$\hat{z}_{obs}$ for the forecasted value of the *obs*-th observation,
$\bar{z}$ the average of the actual values.

### 3.5.2 Root mean squared error (RMSE)

Root Mean Squared Error (RMSE) serves as the maestro in assessing the precision of a predictive model by calculating the square root of the average squared vari- ances between predicted and actual values. Within the realm of regression analysis, RMSE orchestrates a symphony that balances accuracy and imperfections, produc- ing a distinct tune that reflects how closely predictions align with reality. This metric transforms errors into a succinct measurement, articulating the performance rhythm through a harmonious fusion of squares and roots. Positioned as the pinnacle of model evaluation, RMSE composes a numerical climax, guiding practitioners through the intricate arrangement of predictive analytics.

The equation for Root Mean Squared Error (RMSE) can be expressed as follows:

$$\text{RMSE} = \sqrt{\frac{1}{total} \sum_{obs=1}^{total} \left(z_{obs} - \widehat{z_{obs}}\right)^2} \tag{9}$$

where,

*Total* the collective count of observations,
$z_{obs}$ the actual value of the *obs*-th observation,
$\hat{z}_{obs}$ for the forecasted value of the *obs*-th observation.

### 3.5.3 Mean absolute error (MAE)

Mean Absolute Error (MAE) stands as a commonly employed metric within statistics and machine learning to gauge the precision of a predictive model. It calculates the average of the absolute differences between the predicted values and the actual values in a given dataset. MAE offers a straightforward means to evaluate how well a model performs, treating all errors as equally important

**Table 3** Evaluation Metrics of different base models used in the Bagging approach

| Base Models | $R^2$ Score | MAE | RMSE |
|---|---|---|---|
| Decision Tree Regressor | 0.89 | 6.61 | 9.88 |
| Linear Regression | 0.83 | 9.96 | 12.28 |
| Ridge Regression | 0.83 | 9.97 | 12.28 |
| Lasso Regression | 0.83 | 10.17 | 12.45 |
| ElasticNet Regression | 0.80 | 10.96 | 13.55 |

**Table 4** Evaluation metrics for random forest regressor

| Model | $R^2$ Score | MAE | RMSE |
|---|---|---|---|
| Random Forest Regressor | **0.90** | 6.29 | 9.44 |

**Table 5** Evaluation metrics of different boosting models

| Boosting Models | $R^2$ Score | MAE | RMSE |
|---|---|---|---|
| LightGBM Regressor | **0.91** | 6.35 | 9.15 |
| CatBoost Regressor | **0.91** | 6.05 | 8.81 |
| XGBoost Regressor | 0.90 | 6.52 | 9.49 |
| Gradient Boosting Regressor | 0.89 | 7.23 | 10.15 |
| AdaBoost Regressor | 0.74 | 13.98 | 15.43 |

**Table 6** Evaluation Metrics of different estimator and meta-models used in the Stacking approach

| Estimator | Meta Model | $R^2$ Score | MAE | RMSE |
|---|---|---|---|---|
| Ridge Regression<br>Lasso Regression<br>ElasticNet Regression<br>Decision Tree Regressor | Linear Regression | 0.87 | 8.11 | 10.97 |
| Linear Regression<br>Lasso Regression<br>ElasticNet Regression<br>Decision Tree Regressor | Ridge Regression | 0.87 | 8.11 | 10.98 |
| Linear Regression<br>Ridge Regression<br>ElasticNet Regression<br>Decision Tree Regressor | Lasso Regression | 0.87 | 8.12 | 11.00 |
| Linear Regression<br>Ridge Regression<br>Lasso Regression<br>Decision Tree Regressor | ElasticNet Regression | 0.87 | 8.13 | 11.02 |
| Linear Regression<br>Ridge Regression<br>Lasso Regression<br>ElasticNet Regression | Decision Tree Regressor | 0.76 | 10.52 | 14.79 |

**Table 7** Evaluation Metrics of different estimator and meta-models used in the Blending approach

| Estimator | Meta Model | $R^2$ Score | MAE | RMSE |
|---|---|---|---|---|
| Ridge Regression<br>Lasso Regression<br>ElasticNet Regression<br>Decision Tree Regressor | Linear Regression | 0.86 | 8.36 | 11.27 |
| Linear Regression<br>Lasso Regression<br>ElasticNet Regression<br>Decision Tree Regressor | Ridge Regression | 0.86 | 8.24 | 11.15 |
| Linear Regression<br>Ridge Regression<br>ElasticNet Regression<br>Decision Tree Regressor | Lasso Regression | 0.86 | 8.30 | 11.20 |
| Linear Regression<br>Ridge Regression<br>Lasso Regression<br>Decision Tree Regressor | ElasticNet Regression | 0.86 | 8.30 | 11.17 |
| Linear Regression<br>Ridge Regression<br>Lasso Regression<br>ElasticNet Regression | Decision Tree Regressor | 0.77 | 10.00 | 14.19 |

and disregarding their direction. A smaller MAE reflects a superior model fit to the data, with values approaching zero, indicating a higher level of accuracy. MAE's ease of comprehension and interpretation renders it a pragmatic choice for appraising regression models and evaluating their predictive prowess.

The equation for Mean Absolute Error (MAE) can be calculated as follows:

$$\text{MAE} = \frac{1}{total}\sum_{obs=1}^{total}|z_{obs} - \widehat{z_{obs}}| \tag{10}$$

where,

$total$ the collective count of observations,
$z_{obs}$ the actual value of the $obs$-th observation,
$\hat{z}_{obs}$ for the forecasted value of the $obs$-th observation.

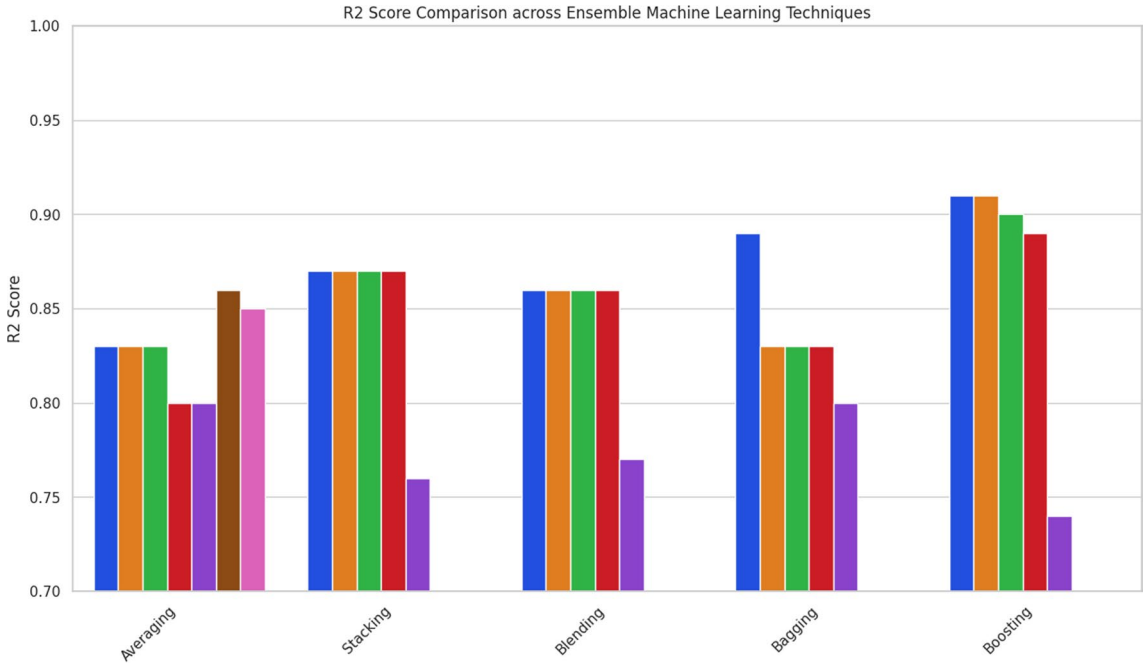



**Fig. 11** $R^2$ Score Comparison across Ensemble Machine Learning Techniques

**Table 8** Evaluation metrics of machine learning models using PyCaret

| ML Models | $R^2$ Score | MAE | RMSE |
|---|---|---|---|
| Extra Trees Regressor | **0.91** | 6.17 | 8.91 |
| Light Gradient Boosting Machine | **0.91** | 6.30 | 9.03 |
| Random Forest Regressor | 0.90 | 6.26 | 9.24 |
| Extreme Gradient Boosting | 0.90 | 6.56 | 9.41 |
| Gradient Boosting Regressor | 0.89 | 7.18 | 9.86 |
| Linear Regression | 0.83 | 9.93 | 12.18 |
| Ridge Regression | 0.83 | 9.93 | 12.18 |
| Lasso Regression | 0.83 | 9.94 | 12.19 |
| Bayesian Ridge | 0.83 | 9.93 | 12.18 |
| Huber Regressor | 0.83 | 9.89 | 12.26 |
| Lasso Regression | 0.83 | 10.10 | 12.33 |
| Lasso Least Angle Regression | 0.83 | 10.10 | 12.33 |
| Decision Tree Regressor | 0.80 | 8.01 | 13.19 |
| Elastic Net | 0.80 | 10.81 | 13.37 |
| K Neighbors Regressor | 0.74 | 11.42 | 15.31 |
| AdaBoost Regressor | 0.73 | 13.87 | 15.39 |

## 4 Experimental setup and results

In this segment, we delve into presenting our findings and analysis. This entails a thorough review of performance evaluation metrics and the assessment of feature importance.

### 4.1 Ensemble machine-learning techniques

#### 4.1.1 Averaging

Table 2 represents evaluation metrics for the Averaging approach. The ensemble machine learning approach using averaging performs well according to the evaluation metrics. It achieves an $R^2$ score of 0.86, indicating that 86% of the variance in the dependent variable is explained. The Mean Absolute Error (MAE) is 9.11, suggesting a relatively small average prediction error, and the Root Mean Square Error (RMSE) is 11.47, which is also a reasonably low error.

#### 4.1.2 Bagging

**Bagging regressor** Table 3 represents different base models employed in the bagging approach. The Bag- ging Regressor utilizes multiple base models, with the Decision Tree Regressor standing out as the top performer, achieving an impressive $R^2$ score of 0.89 along with low MAE and RMSE, indicating exceptional predictive accuracy. While the other base mod- els employed in the Bagging approach demonstrate reasonable performance, with $R^2$ scores hovering around 0.83, they fall short of the outstanding results produced by the Decision Tree Regressor. In contrast to Averaging, Stacking, and Blending, the Bag- ging Regressor, especially when combined with the

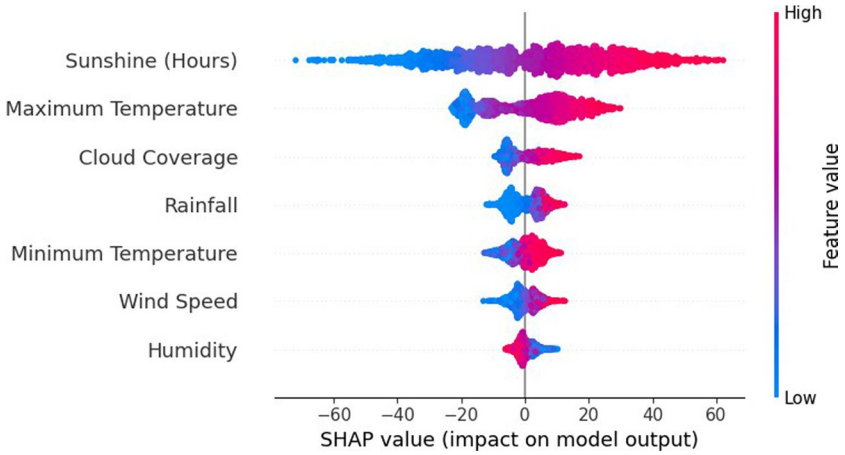



**Fig. 12** SHAP explanation of the Beeswarm plot

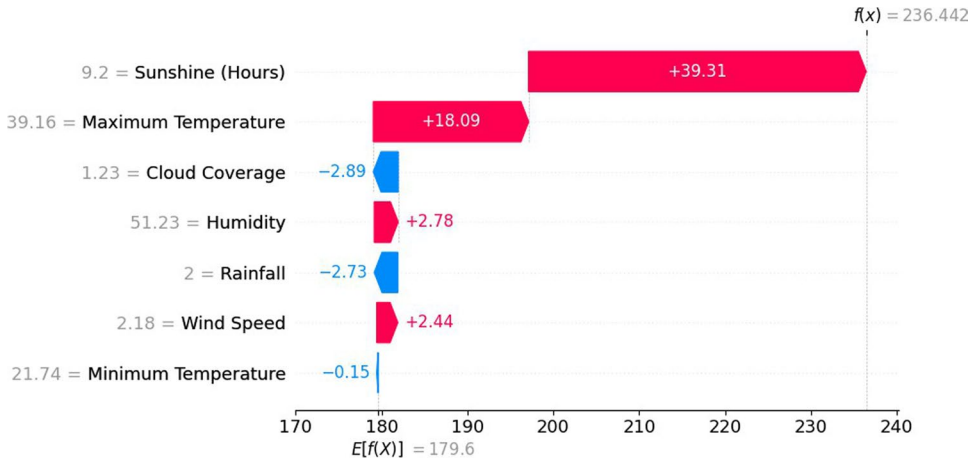
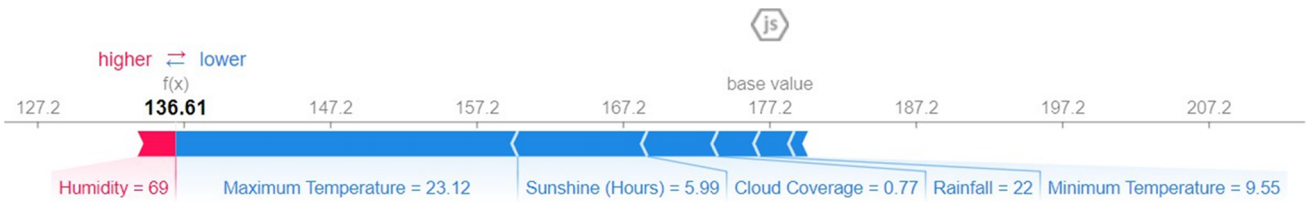
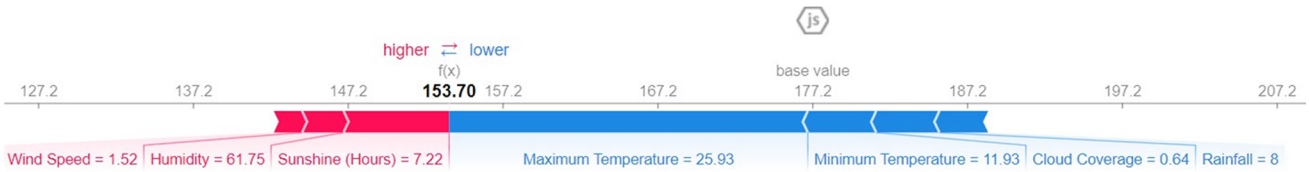



**Fig. 13** SHAP explanation of the waterfall plot with the boundary value (maximum) of maximum temperature




(a)




(b)

**Fig. 14** SHAP explanation of (**a**) 0-th instance, (**b**) 1-st instance

Decision Tree base model, delivers superior outcomes, boasting higher $R^2$ values and lower MAE and RMSE metrics. It's evident that the choice of the base model within the Bagging Regressor can have a profound impact on its performance when compared to other ensemble methods.

**Random forest regressor** Table 4 represents evaluation metrics using a Random Forest Regressor. The Random Forest Regressor demonstrates outstanding performance, with an impressive $R^2$ score of 0.90, a remarkably low MAE of 6.29, and an RMSE of 9.44. These metrics collec- tively illustrate its exceptional

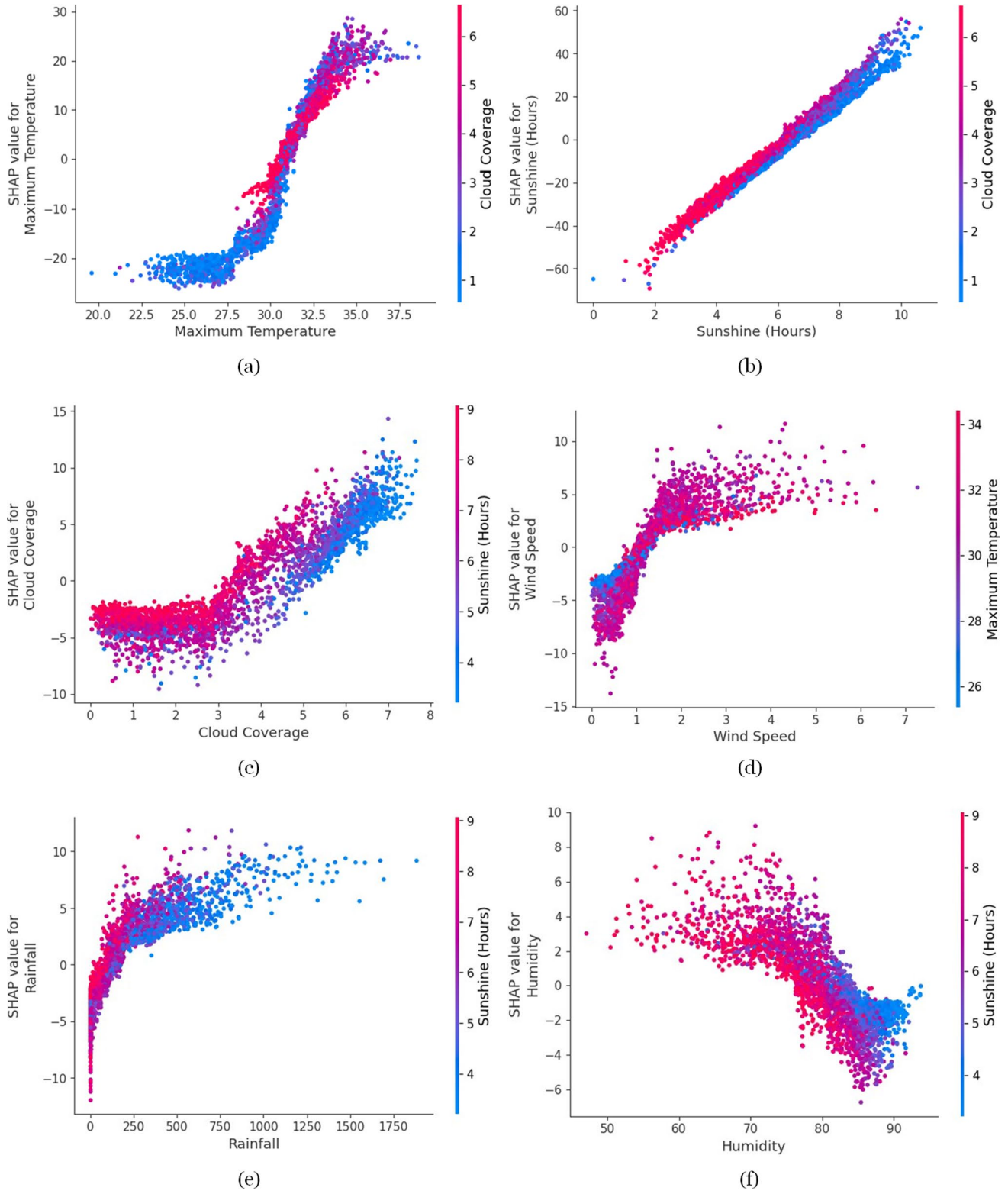



**Fig. 15** SHAP feature dependence plot with interaction visualization (**a**) maximum temperature, (**b**) sunshine(hours), (**c**) cloud coverage, (**d**) wind speed, (**e**) rainfall, and (**f**) humidity

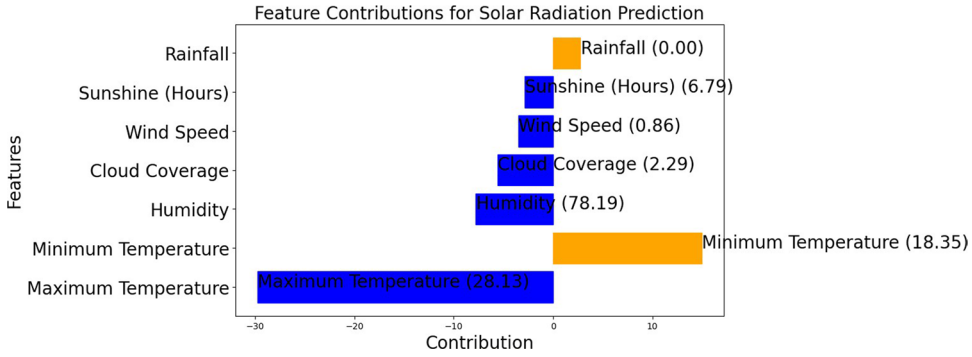



**Fig. 16** LIME explanation of feature contribution for solar radiation prediction with all features

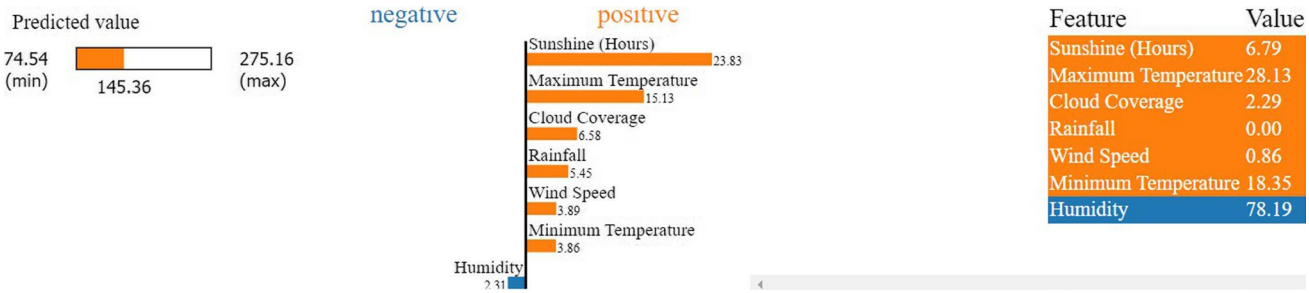



**Fig. 17** LIME explanation of the first instance from the test dataset where actual: 145.203, predicted: 145.355

**Fig. 18** With all six features, ELI5 explanation of feature (**a**) weights and (**b**) con- tribution on the prediction

| Weight | Feature |
|---|---|
| 1.5421 ± 0.0343 | Sunshine (Hours) |
| 0.4519 ± 0.0192 | Maximum Temperature |
| 0.1037 ± 0.0090 | Cloud Coverage |
| 0.0685 ± 0.0067 | Rainfall |
| 0.0671 ± 0.0041 | Minimum Temperature |
| 0.0401 ± 0.0032 | Wind Speed |
| 0.0378 ± 0.0039 | Humidity |

(a)

| Contribution? | Feature |
|---|---|
| +177.184 | <BIAS> |
| -0.774 | Minimum Temperature |
| -0.841 | Humidity |
| -1.716 | Sunshine (Hours) |
| -2.778 | Wind Speed |
| -4.850 | Cloud Coverage |
| -7.765 | Rainfall |
| -13.103 | Maximum Temperature |

(b)

explanatory power and precise prediction capabilities. When compared to the Bagging Regressor with a Decision Tree base model, the Ran- dom Forest Regressor surpasses it by achieving a higher $R^2$ score and lower MAE and RMSE. This highlights the Random Forest's superior performance in this con- text. In contrast to Averaging, Stacking, and Blending, the Random Forest Regressor outshines these ensemble methods by offering superior $R^2$ and lower MAE and RMSE values.

### 4.1.3 Boosting

Table 5 represents evaluation metrics of different Boosting models. Boosting models, particularly LightGBM and CatBoost Regressors, consistently exhibit exceptional performance, as evidenced by their high $R^2$ scores of 0.91 and low MAE values of 6.35 and 6.05, along with minimal RMSE values of 9.15 and 8.81, respectively. These models not only excel in their ability to provide accurate predictions but also in their capacity to elucidate the underlying data relationships, rendering Boosting an appealing choice for ensemble learning.

Compared to other ensemble techniques like Averaging, Stacking, Blending, and Bagging, Boosting methods, such as XGBoost, CatBoost, and LightGBM, outperform the competition consistently. They consistently yield higher $R^2$ scores and lower MAE and RMSE values, underscoring their superiority in terms of accuracy and explanatory power. Hence, in the context of ensemble learning, Boosting methods emerge as the preferred options, offering superior performance.

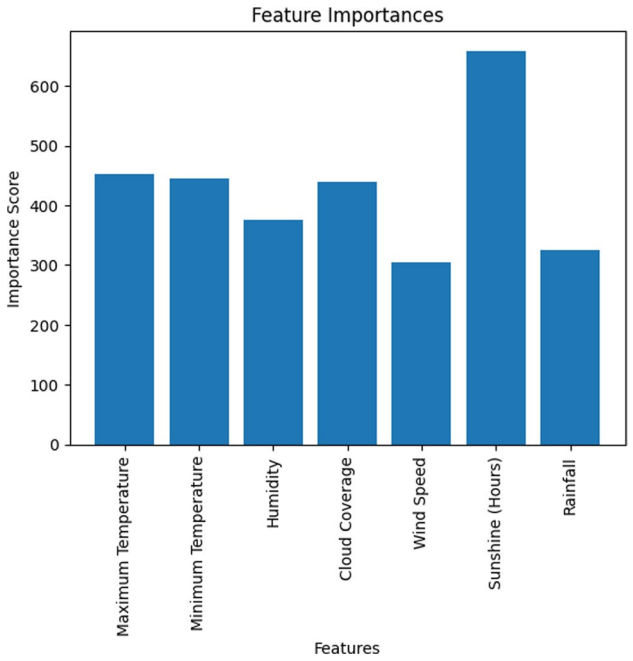



**Fig. 19** ELI5 explanation of feature importance on the prediction

**Table 9** Evaluation metrics for K- means-LightGBM

| Cluster | $R^2$ Score | MAE | RMSE |
|---|---|---|---|
| Very Sunny | 0.90 | 6.05 | 8.47 |
| Sunny | 0.86 | 5.05 | 7.58 |
| Cloudy | 0.84 | 7.83 | 12.06 |
| Very Cloudy | 0.93 | 3.98 | 5.18 |

**Table 10** Evaluation metrics for MiniBatchKMeans-LightGBM

| Cluster | $R^2$ Score | MAE | RMSE |
|---|---|---|---|
| Very Sunny | 0.88 | 6.67 | 9.42 |
| Sunny | 0.85 | 8.02 | 12.15 |
| Cloudy | 0.86 | 5.20 | 7.75 |
| Very Cloudy | 0.93 | 4.15 | 5.20 |

**Table 11** Evaluation Metrics for Fuzzy C-Means-LightGBM

| Cluster | $R^2$ Score | MAE | RMSE |
|---|---|---|---|
| Very Sunny | 0.89 | 6.06 | 8.62 |
| Sunny | 0.83 | 7.79 | 10.97 |
| Cloudy | 0.85 | 6.69 | 10.15 |
| Very Cloudy | 0.94 | 3.94 | 5.26 |

**Table 12** Evaluation Metrics for GaussianMixture-LightGBM

| Cluster | $R^2$ Score | MAE | RMSE |
|---|---|---|---|
| Very Sunny | 0.85 | 6.22 | 8.71 |
| Sunny | 0.86 | 7.33 | 11.05 |
| Cloudy | 0.86 | 5.40 | 8.21 |
| Very Cloudy | 0.94 | 3.86 | 4.87 |

#### 4.1.4 Stacking

Table 6 represents evaluation metrics for the Stacking approach. Compared to Aver- aging, Stacking demonstrates a marginal improvement in the $R^2$ score, reaching 0.87 for meta-models, including Linear Regression, Ridge Regression, Lasso Regres- sion, and ElasticNet Regression. Additionally, Stacking yields reduced Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for these meta-models, indicat- ing improved predictive accuracy compared to the Averaging technique. However, it's noteworthy that employing a Decision Tree Regressor as the meta-model leads to a decrease in performance, resulting in an $R^2$ score of 0.76, MAE of 10.52, and MSE of 14.79.

#### 4.1.5 Blending

Table 7 represents evaluation metrics for the Blending approach. The Blending approach showcases strong performance, achieving an $R^2$ score of 0.86 for meta mod- els including Linear Regression, Ridge Regression, Lasso Regression, and ElasticNet Regression. However, its efficacy diminishes to 0.77 when employing the Decision Tree Regressor as the meta-model. In contrast to Averaging, Blending delivers comparable results with marginally lower Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) while maintaining the same $R^2$ score. When juxtaposed with Stacking, Blending exhibits a slightly lower $R^2$ score alongside higher MAE and RMSE val- ues. Consequently, Blending occupies an intermediate position between Averaging and Stacking in terms of performance, striking a balanced compromise between simplicity and accuracy.

### 4.2 Comparison of various ensemble techniques

Figure 11 illustrates $R^2$ score comparison across ensemble machine learning techniques, and it shows that LighGBM gives the better result. We further continue the experiment with LightGBM for explainable AI and clustering techniques.

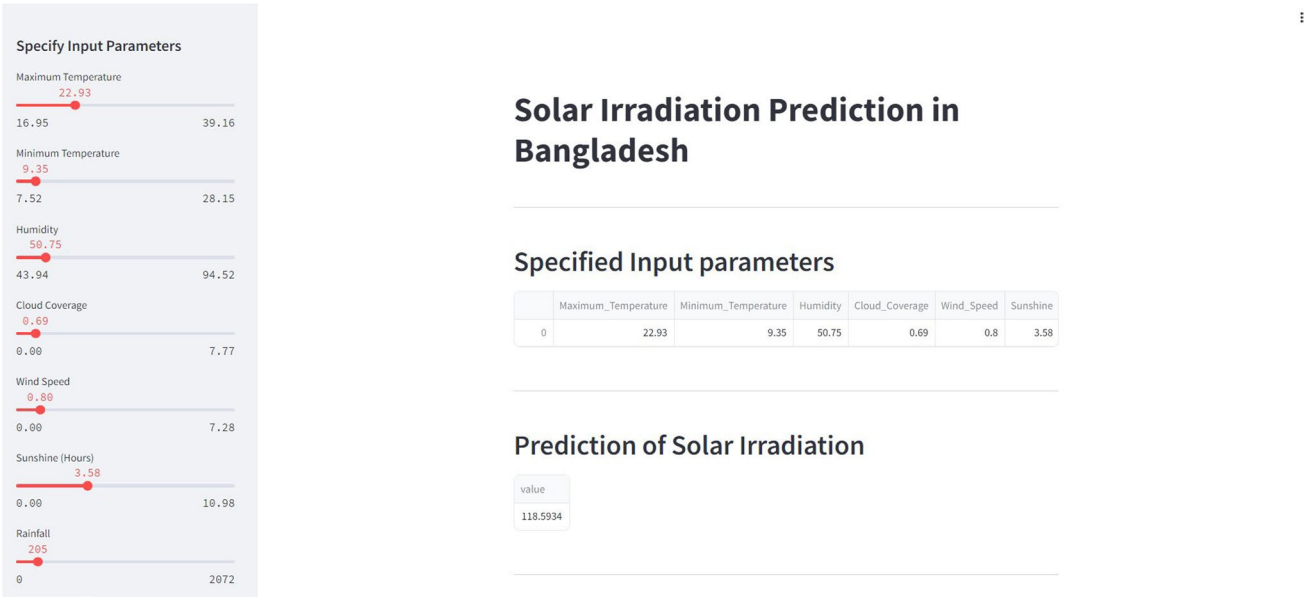



**Fig. 20** Interactive tool with solar prediction

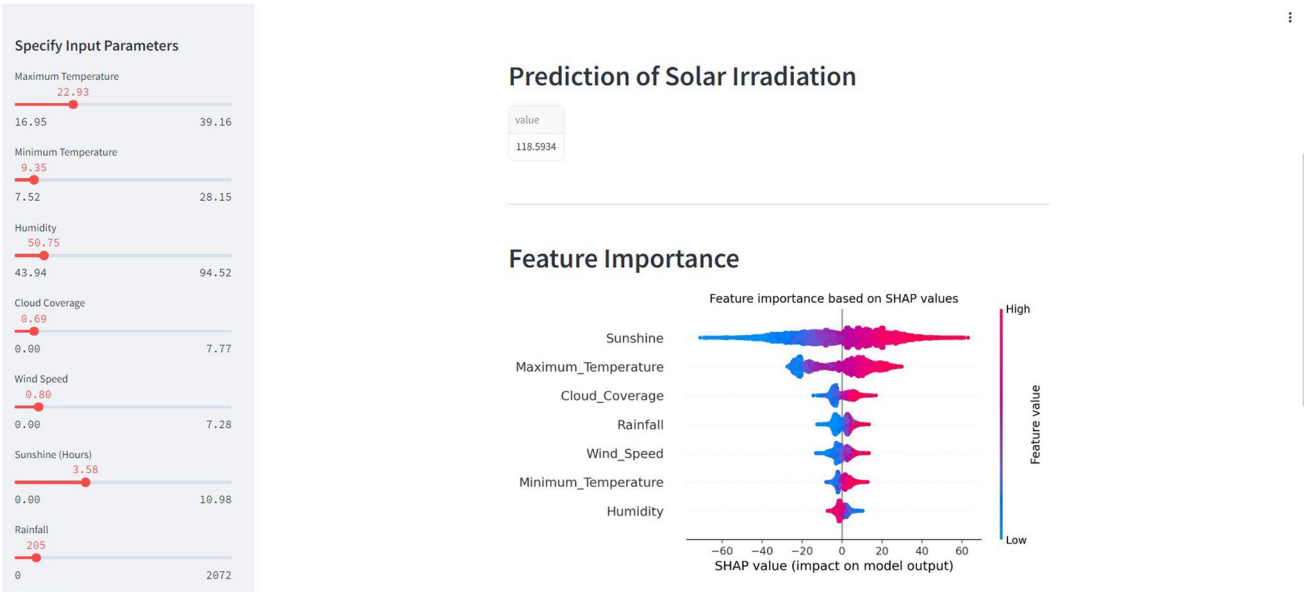



**Fig. 21** Interactive tool with SHAP value

#### 4.2.1 PyCaret

Table 8 represents evaluation metrics of machine learning models using PyCaret. Using PyCaret, the top-performing models, such as Extra Trees Regressor and Light Gra- dient Boosting Machine, achieve high $R^2$ scores (around 0.91), indicating a strong correlation between predicted and actual values. These models also exhibit rela- tively low MAE and RMSE values, suggesting accurate predictions. On the other hand, models like K Neighbors Regressor and AdaBoost Regressor have lower $R^2$ scores, indicating comparatively weaker performance in capturing the variance of the data. Linear regression-based models, while providing reasonable performance, exhibit slightly lower $R^2$ scores compared to some ensemble models.

### 4.3 Explainable AI

#### 4.3.1 SHAP

The SHAP explanation shows that sunshine(hours) has the greatest impact following the maximum temperature. Figure 12 shows the SHAP explanation of the beeswarm plot,

and Fig. 13 shows the waterfall plot with the maximum boundary value of the maximum temperature feature. In the force plot of Fig. 14, we have used the zeroth and the first instance of the test dataset. We predicted 136.61 and 153.70, where the actual value was 131.48 for the 0-th instance and 151.42 for the first instance. Within the figure's dependency plot in Fig. 15, the plot's shape offers information about how the features affect the prediction, indicating whether the relationship is monotonic, linear, or characterized by more intricate patterns. From the figure, it can be clearly visible that all other features are dependent on cloud coverage, maximum temperature, and sunshine(hours).

#### 4.3.2 LIME

Figure 16 represents the LIME explanation of feature contribution on solar radiation prediction of Bangladesh region. The influence of these features on the prediction can be understood very clearly from this explanation. The figure shows that the maximum temperature has the highest negative contribution, which indicates that the influences of all features have a stronger impact and influence on the prediction. The prediction probability for the model is 145.36, including all six features.

Local interpretation of the 0-th instance has been shown in Figure 17 which is obtained through the LIME explain instance() function. According to the results, the predicted value is 145.355, and the actual value is 145.203 for the first instance.

#### 4.3.3 ELI5

Figure 18(a) shows the feature weights obtained with all its features using the ELI5 show weights() function. We run eli5.explain prediction() function to see how the model works with all of its six features. The forecasting shows the sum of the feature contribution in addition to the"BIAS," as shown in Fig. 18(b). Figure 19 shows the ELI5 explanation of feature importance on the prediction and their feature importance score. It is seen that sunshine (hours) gives the greatest score, while temperature (maximum, minimum) and cloud coverage are second.

### 4.4 Hybridized clustering

#### 4.4.1 K-means-LightGBM

Table 9 represents evaluation metrics for K-means-LightGBM. The K-means- LightGBM methodology excels in weather prediction, particularly in the"Very Sunny" and"Very Cloudy" clusters. It achieves an outstanding $R^2$ score of 0.90 for very sunny days and an exceptional score of 0.93 for heavily overcast days. While still perform- ing well in"Sunny" and"Cloudy" clusters, it demonstrates versatility and remarkable accuracy in predicting varying weather conditions.

#### 4.4.2 MiniBatchKMeans-LightGBM

Table 10 represents evaluation metrics for MiniBatchK-Means-LightGBM. The MiniBatchKMeans-LightGBM approach also exhibits exceptional performance in weather prediction in"Very sunny" and"Very cloudy" scenarios. However, it main- tains strong predictive capabilities in"Sunny" and"Cloudy" clusters, its versatility and remarkable accuracy shine through when forecasting diverse weather conditions.

#### 4.4.3 Fuzzy C-means-LightGBM

Table 11 represents evaluation metrics for Fuzzy C-Means-LightGBM. Like the K-means-LightGBM and MiniBatch-KMeans-LightGBM methodologies, the Fuzzy C- Means-LightGBM approach demonstrates outstanding efficacy in predicting weather patterns, particularly excelling in scenarios characterized by"Very Sunny" and"Very Cloudy" conditions and slightly reduced performance in clustering of"Sunny" and"Cloudy" conditions.

#### 4.4.4 Gaussian mixture-LightGBM

Table 12 represents evaluation metrics for GaussianMixture-LightGBM. Similar to K-means-LightGBM, MiniBatchK-Means-LightGBM, and Fuzzy C-Means-LightGBM, the GaussianMixture-LightGBM approach demonstrates outstanding performance in weather prediction under"Very Cloudy" conditions. However, its effectiveness diminishes when confronted with"very sunny" clustering scenarios. Interestingly, the clustering performance for"sunny" and"cloudy" conditions remains relatively consistent.

### 4.5 Interactive tool

In the Interactive Tool, Figs. 20 and 21 depict the forecast for solar radiation and the corresponding feature contributions. The left sidebar allows users to modify input parameter values under"Specify Input Parameters," with the updated settings reflected in the"Specified Input Parameters" section. The LightGBM machine learn- ing model predicts solar radiation, and the results are displayed in the"Prediction of Solar Irradiation" section. Additionally, the"Feature Importance" section utilizes SHAP to illustrate the significance and contribution of different features.

## 5 Discussion

In our research, we utilized a dataset specific to Bangladesh to forecast solar irradiance and tried to enlighten Bangladesh's solar energy future.

In the first phase of our study, we evaluated various ensemble machine learn- ing methods, including Averaging, Stacking, Blending, Bagging, Random Forest, and Boosting, to assess their predictive performance. We found that Averaging and Stack- ing achieved strong $R^2$ scores of 0.86 and 0.87, striking a balance between simplicity and accuracy. Stacking performed slightly better than Averaging. Blending had a sim- ilar $R^2$ score (0.86) to Averaging but showed a slight edge in terms of lower Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). However, it fell slightly behind Stacking in $R^2$ and had higher MAE and RMSE, offering a trade-off between simplicity and accuracy. The Bagging Regressor, especially with a Decision Tree as a base model, achieved an impressive $R^2$ score of 0.89, indicating strong predictive accuracy. Other base models in Bagging performed reasonably with $R^2$ scores around 0.83. The choice of the base model significantly influenced Bagging's performance. The Random Forest Regressor outperformed Bagging with the Decision Tree base model, achieving an exceptional $R^2$ score of 0.90, low MAE, and RMSE. Boosting methods, such as LightGBM and CatBoost Regressors, consistently demonstrated outstanding performance with high $R^2$ scores (0.91) and low MAE, indicating accu- rate predictions and strong data relationship elucidation. XGBoost, CatBoost, and LightGBM consistently outperformed the competition, providing higher $R^2$ scores and lower MAE and RMSE values, highlighting their accuracy and explanatory power. Finally, PyCaret highlights strong performance from models like Extra Trees Regres- sor and Light Gradient Boosting Machine, yielding high $R^2$ scores around 0.91, while K Neighbors Regressor and AdaBoost Regressor show comparatively weaker results. Linear regression-based models perform reasonably but with slightly lower $R^2$ scores compared to ensembles.

The second phase of our study sheds light on the key drivers behind solar radi- ation prediction in Bangladesh. Three Explainable AI methods were employed to gain insights into how specific features impact the model's predictions. SHAP Expla- nations have shown that sunshine hours and maximum temperature are the most significant factors influencing the prediction. Meanwhile, ELI5's Feature Importance Scores underscore the importance of sunshine hours, closely followed by maximum and minimum temperatures, as well as cloud coverage. These scores provide a numer- ical measure of how each feature affects the prediction, confirming the importance of sunshine hours. LIME Explanations, on the other hand, delved deeper into the con- tributions of each feature and revealed that maximum temperature has the strongest negative impact, indicating its significant influence on the prediction. Taking all six features into account, the prediction probability was calculated to be 145.36.

Moving to our study's third phase, we used Hybridized Clustering to forecast Solar Radiation in Bangladesh. Notably, the K-means-LightGBM, MiniBatchKMeans- LightGBM, and Fuzzy C-Means-LightGBM models excelled in forecasting solar radiation levels within both"Very Sunny" and"Very Cloudy" clusters. However, the effectiveness of the GaussianMixture-LightGBM approach declined notably in scenar- ios characterized by"Very Sunny" clusters. Interestingly, the clustering performance remained relatively stable when predicting solar radiation in"Sunny" and"Cloudy" conditions.

Finally, In the fourth phase of our study, we've developed an interactive tool with Streamlit that forecasts solar radiation using provided weather data. We've utilized the LightGBM machine learning model for prediction, and to enhance interpretability, we've integrated SHAP to visualize the importance of features and contributions in the analysis of the predictions.

## 6 Conclusion

In this study, we assessed the predictive performance of various ensemble machine- learning methods using a dataset specific to Bangladesh. Following that, we ran this dataset through three XAI models: SHAP, LIME, and ELI5. Finally, our research concludes with a combined prediction based on hybridized clustering. In conclusion, our study used a multifaceted approach to forecasting solar irradiance in Bangladesh, employing a methodology aimed at understanding, predicting, and leveraging the region's solar resources. From employing collective machine learning strategies to incorporating Explainable AI approaches each stage has made a significant contribu- tion to illuminating the path toward a more sustainable and energy-secure future for Bangladesh. This all-encompassing strategy, which combines environmental awareness with technological advancements, represents an ongoing commitment to maximizing the sun's potential for a greener, more sustainable world.

Future studies on solar irradiance forecasting in Bangladesh might look into a num- ber of approaches to improve precision and usefulness. This includes developing more sophisticated temporal and spatial models that take geographic specifics into account, incorporating climate change factors, exploring real-time predictive models for prac- tical applications, field testing for model validation, analyzing policy and economic implications, and developing community engagement initiatives to raise awareness. Also included are further refinements of ensemble machine learning techniques and the integration of advanced AI methodologies beyond the XAI models used. These options present viable means of enhancing solar irradiance forecasts, assisting Bangladesh in building a more sustainable and energy-secure future.

**Author contributions** SS did the experiment and wrote the initial draft. NS planned the methodology, revised the manuscript, and supervised the whole work. FS and SRS write the partial experiment and draft.

**Funding** None.

**Data availability** None.

## Declarations

**Competing interests** The authors declare that they have no conflict of interest.

## References


Adams J, Hagras H (2020) A type-2 fuzzy logic approach to explainable ai for reg- ulatory compliance, fair customer outcomes and market stability in the global financial sector. In: 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp 1–8. IEEE

Alam MS, Al-Ismail FS, Hossain MS, Rahman SM (2023) Ensemble machine- learning models for accurate prediction of solar irradiation in bangladesh. Processes 11(3):908

Ali M (2020) Pycaret: An open source, low-code machine learning library in python. PyCaret version 2

Amin MN, Iftikhar B, Khan K, Javed MF, AbuArab AM, Rehman MF (2023) Prediction model for rice husk ash concrete using ai approach: Boosting and bagging algorithms. In: Structures, vol 50, pp 745–757. Elsevier

Arrieta AB, D´ıaz-Rodr´ıguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garc´ıa S, Gil-L´opez S, Molina D, Benjamins R et al (2020) Explainable artifi- cial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Inf Fusion 58:82–115

Ayodele TR, Ogunjuyigbe ASO, Amedu A, Munda JL (2019) Prediction of global solar irradiation using hybridized k-means and support vector regression algorithms. Renew Energy Focus 29:78–93. https://doi.org/10.1016/j.ref.2019.03.003

Bae KY, Jang HS, Sung DK (2016) Hourly solar irradiance prediction based on support vector machine and its error analysis. IEEE Trans Power Syst 32(2):935–945

Bahani K, Ali-Ou-Salah H, Moujabbir M, Oukarfi B, Ramdani M (2020) A novel interpretable model for solar radiation prediction based on adaptive fuzzy clus- tering and linguistic hedges. In: Proceedings of the 13th international conference on intelligent systems: theories and applications, pp 1–6

Bezdek JC, Ehrlich R, Full W (1984) Fcm: The fuzzy c-means clustering algorithm. Comput Geosci 10(2–3):191–203

Biau G, Scornet E (2016) A random forest guided tour. TEST 25:197–227

Biparva D, Materassi D (2023) Interpretation of explainable ai methods as identification of local linearized models. IFAC-PapersOnLine 56(2):2383–2388. https://doi.org/10.1016/j.ifacol.2023.10.1211. 22nd IFAC World Congress

Cannon RL, Dave JV, Bezdek JC (1986) Efficient implementation of the fuzzy c- means clustering algorithms. IEEE Trans Pattern Anal Mach Intell 2:248–255

Chadaga K, Prabhu S, Bhat V, Sampathila N, Umakanth S, Chadaga R (2023) A decision support system for diagnosis of covid-19 from non-covid-19 influenza-like illness using explainable artificial intelligence. Bioengineering 10(4):439. https://doi.org/10.3390/bioengineering10040439

Chaibi M, Benghoulam EM, Tarik L, Berrada M, Hmaidi AE (2021) An inter- pretable machine learning model for daily global solar radiation prediction. Energies 14(21):7367

Chavan M, Patil A, Dalvi L, Patil A (2015) Mini batch k-means clustering on large dataset. Int J Sci Eng Technol Res 4(07):1356–1358

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp 785–794

Dietterich TG et al (2002) Ensemble learning. The handbook of brain theory and neural networks 2(1):110–125

El-Sappagh S, Alonso JM, Islam SR, Sultan AM, Kwak KS (2021) A multi- layer multimodal detection and prediction model based on explainable artificial intelligence for alzheimer's disease. Sci Rep 11(1):2660. https://doi.org/10.1038/s41598-021-82098-3

Frimane Â, Soubdhan T, Bright JM, Aggour M (2019) Nonparametric bayesian-based recognition of solar irradiance conditions: Applica- tion to the generation of high temporal resolution synthetic solar irradiance data. Solar Energy 182:462–479. https://doi.org/10.1016/j.solener.2019.02.052

Gillies S (2013) The shapely user manual. URL https://pypi.org/project/Shapely. Accessed 20 Aug 2023

Groß J (2003) Linear Regression vol 175. Springer

Hirata Y, Aihara K (2017) Improving time series prediction of solar irradiance after sunrise: Comparison among three methods for time series prediction. Solar Energy 149:294–301. https://doi.org/10.1016/j.solener.2017.04.020

Hissou H, Benkirane S, Guezzaz A, Azrour M, Beni-Hssane A (2023) A novel machine learning approach for solar radiation estimation. Sustainability 15(13):10609–10609. https://doi.org/10.3390/su151310609

http://apps.barc.gov.bd/climate/dashboard (2023) [Online; Accessed 2023–08–11]

Jay CB, Cockett JRB (1994) Shapely types and shape polymorphism. In: European Symposium on Programming, pp. 302–316. Springer

Kadiyala A, Kumar A (2018) Applications of python to evaluate the performance of bagging methods. Environ Prog Sustain Energy 37(5):1555–1559

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y (2017) Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30

Khorasani M, Abdou M, Hern´andez Fern´andez J (2022) Streamlit basics. 31–62

Kushwah JS, Kumar A, Patel S, Soni R, Gawande A, Gupta S (2022) Comparative study of regressor and classifier with decision tree using mod- ern tools. Mater Today: Proc 56:3571–3576. https://doi.org/10.1016/j.matpr.2021.11.635. First International Conference on Design and Materials

Kuzlu M, Cali U, Sharma V, Guler O (2020) Gaining insight into solar photo- voltaic power generation forecasting utilizing explainable artificial intelligence tools. IEEE Access 8:187814–187823. https://doi.org/10.1109/ACCESS.2020.3031477

Kwon H, Park J, Lee Y (2019) Stacking ensemble technique for classifying breast cancer. Healthcare Inform Res 25(4):283–288

Lee Y, Oh J, Kim G (2020) Interpretation of load forecasting using explainable artifi- cial intelligence techniques. Trans Korean Inst Electr Eng 69(3):480–485

Liang X, Jacobucci R (2020) Regularized structural equation modeling to detect measurement bias: Evaluation of lasso, adaptive lasso, and elastic net. Struct Equ Model: Multidiscip J 27(5):722–734. https://doi.org/10.1080/10705511.2019.1693273

Marino DL, Wickramasinghe CS, Manic M (2018) An adversarial approach for explainable ai in intrusion detection systems. In: IECON 2018–44th Annual conference of the IEEE industrial electronics society, pp 3237–3243. IEEE

McCandless TC, Haupt SE, Young GS (2015) A model tree approach to forecasting solar irradiance variability. Solar Energy 120:514–524. https://doi.org/10.1016/j.solener.2015.07.020

McDonald GC (2009) Ridge regression. Wiley Interdiscip Rev: Comput Stat 1(1):93–100

Mishra DP, Jena S, Senapati R, Panigrahi A, Salkuti SR (2023) Global solar radiation forecast using an ensemble learning approach. Int

J Power Electron Drive Syst 14(1):496–496. https://doi.org/10.11591/ijpeds.v14.i1.pp496-505
Myles AJ, Feudale RN, Liu Y, Woody NA, Brown SD (2004) An introduction to decision tree modeling. J Chemom 18(6):275–285
Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Front Neurorobot 7:21
Pannu HS, Malhi A et al (2020) Deep learning-based explainable target classifica- tion for synthetic aperture radar images. In: 2020 13th International Conference on Human System Interaction (HSI), pp. 34–39. IEEE
Pavlyshenko B (2018) Using stacking approaches for machine learning models. In: 2018 IEEE second international conference on data stream mining & processing (DSMP), pp 255–258. IEEE
Peng K, Leung VC, Huang Q (2018) Clustering approach based on mini batch kmeans for intrusion detection system over big data. IEEE Access 6:11897–11906
Pierrot A, Goude Y (2011) Short-term electricity load forecasting with generalized additive models. Proceedings of ISAP power 2011
Polikar R (2012) Ensemble learning. Ensemble machine learning: Methods and applications 1–34
Prentzas N, Nicolaides A, Kyriacou E, Kakas A, Pattichis C (2019) Integrating machine learning with symbolic reasoning to build an explainable ai model for stroke prediction. In: 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), pp 817–821. IEEE
Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A (2018) Catboost: unbiased boosting with categorical features. Adv Neural Inf Proc Syst 31
Qing X, Niu Y (2018) Hourly day-ahead solar irradiance prediction using weather forecasts by lstm. Energy 148:461–468. https://doi.org/10.1016/j.energy.2018.01.177
Sagi O, Rokach L (2018) Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8(4):1249
Schapire RE (2003) The boosting approach to machine learning: An overview. Nonlinear estimation and classification, 149–171
Seber GA, Lee AJ (2012) Linear regression analysis. Wiley
Sevas MS, Tur Santona CF, Sharmin N (2023) Ensemble machine-learning model for solar radiation prediction using explainable ai. In: 2023 14th International Conference on Computing Communication and Networking Technologies (ICC-CNT), pp 1–6. https://doi.org/10.1109/ICCCNT56998.2023.10307694
Sinaga KP, Yang M-S (2020) Unsupervised k-means clustering algorithm. IEEE Access 8:80716–80727
Solano ES, Dehghanian P, Affonso CM (2022) Solar radiation forecasting using machine learning and ensemble feature selection. Energies 15(19):7049–7049. https://doi.org/10.3390/en15197049
Sushanth K, Mishra A, Mukhopadhyay P, Singh R (2023) Near-real-time forecast- ing of reservoir inflows using explainable machine learning and short-term weather forecasts. Stochastic Environmental Research and Risk Assessment 1–21. https://doi.org/10.1007/s00477-023-02489-y
Wang H, Li G, Tsai C-L (2007) Regression coefficient and autoregressive order shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol 69(1):63–78. https://doi.org/10.1111/j.1467-9868,2007.00577.x
Wang H, Cai R, Zhou B, Aziz S, Qin B, Voropai N, Gan L, Barakht-enko E (2020) Solar irradiance forecasting based on direct explainable neural network. Energy Convers Manage 226:113487
Weber CM, Ray D, Valverde AA, Clark JA, Sharma KS (2022) Gaussian mixture model clustering algorithms for the analysis of high-precision mass mea- surements. Nucl Instrum Methods Phys Res, Sect A 1027:166299
Wu T, Zhang W, Jiao X, Guo W, Hamoud YA (2021) Evaluation of stacking and blending ensemble learning methods for estimating daily reference evapotranspi- ration. Comput Electron Agric 184:106039
Zhang Z, Damiani E, Al Hamadi H, Yeun CY, Taher F (2022) Explainable artificial intelligence to detect image spam using convolutional neural network. In: 2022 International Conference on Cyber Resilience (ICCR), pp. 1–5. https://doi.org/10.48550/arXiv.2209.03166. IEEE