Ensemble machine-learning model for solar radiation prediction using explainable AI

Muhammad Samee Sevas, Chowdhury Farjana Tur Santona, Nusrat Sharmin
Department of Computer Science and Engineering
Military Institute Of Science and Technology (MIST), Dhaka-1216, Bangladesh
Email: samee.sevas@gmail.com, cftsantona@gmail.com, nusrat@cse.mist.ac.bd

***Abstract*—Solar radiation prediction is crucial in many contexts, including solar energy production, climate and weather research, and agricultural planning. Accurate and timely solar radiation estimation opens up opportunities for effective solar energy planning and optimization. Even though a number of machine learning algorithms have been put forth in the literature to predict solar radiation, explainable AI (XAI) has not yet been thoroughly investigated. In this study, we are focusing on filling some research gaps in the performance analysis of ensemble methods for predicting solar radiation based on explainable AI. The research aims to achieve a balance between model interpretability and prediction accuracy by exploring ensemble techniques that go beyond conventional approaches. Additionally, by exploring ensemble-based methods and incorporating explainable AI techniques, we intend to improve prediction accuracy while providing interpretable insights for decision-making in solar energy applications. The results have been reported in terms of R2 Score, mean absolute error (MAE), and root mean squared error (RMSE).**

***Index Terms*—Solar radiation prediction, Ensemble methods, Explainable AI, Performance analysis.**

I. Introduction

The word ”solar radiation” refers to all of the electromagnetic radiation that the sun emits [3]. It is the measurement of solar energy that reaches the earth's surface and critical factor in determining solar energy output. A higher level of solar irradiance indicates that solar power installations have a greater energy potential. Thus predicting solar radiance is very crucial.

Accurate and precise prediction of solar radiation aids scientists and other concerned individuals in understanding and forecasting climate changes, regional weather patterns, and long-term climate conditions in a region. Scientists and farmers can also assess the availability of sunlight in different areas by measuring solar radiation, which aids in crop selection, agricultural planning, and ecological studies. So, for the work to be done perfectly, forecasting solar radiation is required. The accurate and timely estimation of solar radiation makes it possible to plan and optimize solar energy.

Numerous studies have presented and examined various machine learning and deep learning models in order to forecast solar radiation. ANN, machine learning, and SVM/SVR, all three terms are more and more used in literature [3]. ANN is the method most commonly used in global radiation forecasting [3]. ANN, ARIMA, and naive methods are frequently used [3]. (SVM, SVR, k-mean) begin to be utilized more regularly and Boosting, regression trees, random forests, and other methods are rarely used [3]. LSTM, RNN, and DNN are used more often while DBN, ESN, and CNN are rarely used [4]. They put a heavy workload on CPU, memory, or both but give highly accurate predictions [4]. Explainable AI (XAI) makes predictions from machine learning models simpler to comprehend. The significance of XAI lies in the comprehension and trust embedded in the outcomes of machine learning algorithms. In this study, we will discuss the contribution and importance of different input features in influencing the prediction of solar radiation.

Despite extensive publications in the area of solar radiation prediction or forecasting and explainable AI, there is still a limited exploration of ensemble-based methods and inadequate comparison of explainable AI techniques specifically for solar radiation prediction. The unique contribution of this paper can be summarized as follows:

1) Investigated various ensemble machine-learning models by exploring the feature importance.
2) Using explainable AI to improve the interpretability of the solar radiation prediction problem.

In light of this, the following sections outline the structure of the remaining paper. Section II presents a comprehensive literature review. Section III covers the formulation of the solar radiance problem, the framework for solar radiation, ensemble algorithms used for solar radiance prediction, and an introduction to the explainable AI SHAP method. Moving on to Section IV, it presents the results and analysis, including a thorough examination of performance evaluation metrics and an assessment of feature importance. Lastly, in Section V, the study reaches its conclusion by summarizing the key findings and insights discussed throughout the research, while also highlighting potential avenues for future work in the field.

II. Literature Review

Solar radiation is affected by some important factors like weather classification and performance evaluation metrics [1]. k-NN showed the worst result compared to RMSE, MABE, and R^2 in most of the cases where deep learning is the only model to pass the t-critic value for prediction [1]. Prior to model training, filter methods were used and benefited from low computational costs [5]. They were able to get the ideal feature subsets [5]. But the computational cost was high in that case [5]. The feature selection process was carried out

by embedded methods, and they were more accurate than filter methods at doing so [5]. Deep learning models like LSTM, RNN, and DNN are used more often, whereas DBN, CNN, and ESN are used less [4]. Some of the DL models are computationally expensive but show good accuracy in result [4]. There is more to explore deep learning models in time-series forecasting problems like solar radiation prediction [4]. Ensemble machine-learning models can provide significant information with high accuracy for evaluating solar energy resources [7]. The GBR performs the best with the default parameters because its standard deviation of errors is the smallest [7]. It is anticipated that by including explainability [13], the ML models will be more reliable and widespread, enabling future advancements in the regularization of these models [9].

The current soft-computing-based methods for forecasting solar irradiance are modeled as "black boxes," typically expressed by standard unintelligible functions like the sigmoid [10]. It is challenging to interpret the prediction results from these functions [10].

Research Gap: Three significant research gaps can be identified from the existing state of the art. *Firstly*, there is a limited exploration of ensemble-based machine-learning methods for solar radiation prediction. *Secondly*, there is an inadequate comparison of explainable AI techniques in the context of solar radiation prediction. *Thirdly*, The explainable AI is yet to explore.

III. Methodoloy

A. Solar Radiation as regression probelm

We require the development of a regression model capable of predicting the S_{rad} at a particular location based on the given parameters: Tem, Pre, Hum, W_{dir}, and W_{spd}.
Here, S_{rad} stands for solar radiation, Tem is for temperature, Pre is for air pressure, Hum is for air humidity, W_{dir} is for direction of wind and W_{spd} is for speed of wind.

The goal is to identify a function $Func$, where $Func$ stands for the prediction model, such that $Func(Tem, Pre, Hum, W_{dir}, W_{spd}) \rightarrow S_{rad}$. $Func$ in the context of regression can be optimized by various machine learning models. The goal is to reduce the discrepancy between the model's predictions and the real values of solar radiance, aiming for minimal difference.

B. Proposed framework for solar radiance prediction

The proposed framework for solar radiance prediction is illustrated in Figure 1. As the framework demonstrates, we first collect the solar radiation data, extract it, and store it in a CSV file. After that, we preprocess the data (check for missing values, feature extraction, etc.) and next divide the dataset into training and testing datasets. Several ensemble machine-learning algorithms have been used after that. In addition, we employ feature importance measures using $SHAP$ values, to analyze the contribution of each feature.

Algorithm 1 represents pseudo code for solar radiance prediction.

Algorithm 1 : Solar Radiance Prediction

Step 1: Load the dataset of solar Radiation, *df*
Step 2: Apply EDA (Exploratory Data Analysis) on the loaded dataset with feature values, $F = f_1, f_2, f_3, f_4, ..., f_n$
Step 3: Calculate feature importance, F_i, using $SHAP$ for each feature
Step 4: Calculate SHAP values, F_s, using $SHAP$ for each feature
Step 5: Apply feature selection and split the dataset into df_train and df_test
Step 6: Training the train dataset df_train with different ensemble models, $M_1, M_2, ..., M_k$
Step 7: Store three different error values, E_{R^2}, E_{MAE} and E_{RMSE} for each model

C. Machine learning approaches

Machine learning (ML) is a tool that employs programmed algorithms to provide a system to analyze input data, learn from it, and optimize its operations by making predictions. Multiple machine-learning approaches have been employed in our work to predict solar radiation. In the subsection, we briefly describe the specific algorithms employed in our research.

- Linear Regressor: A machine learning algorithm called Linear Regressor is utilized for regression problems. In order to model the connection between input characteristics and target variables, a linear equation is fitted to the training data. In order to make predictions, it calculates the linear function's defining coefficients. The algorithm is easy to understand, effective for linear connections, and straightforward.
- Support Vector Regression (SVR): SVR is a regression algorithm that identifies a hyperplane to illustrate the relationship between input features and target values. By minimizing prediction errors within a tolerance margin, SVR accommodates non-linear relationships using kernel functions. It successfully handles outliers and generates precise forecasts by managing the trade-off between model complexity and error.
- Gradient Boosting Regressor: Gradient Boosting Regressor is a method for machine learning that creates an ensemble model by adding weak learners (usually decision trees) in a sequential manner to fix the mistakes caused by prior models. Every new model is taught to pay particular attention to situations where the prior models misclassified the situation or left excessive residuals. Gradient Boosting Regressor seeks to build a robust predictive model by integrating the predictions from many learned models. It requires accurate hyperparameter adjustment to get top performance, which might be computationally expensive. [14]
- Light Gradient Boosting Machine Regressor (LGBM Regressor): An effective, high-performing, and reliable machine learning method called LGBM Regressor was

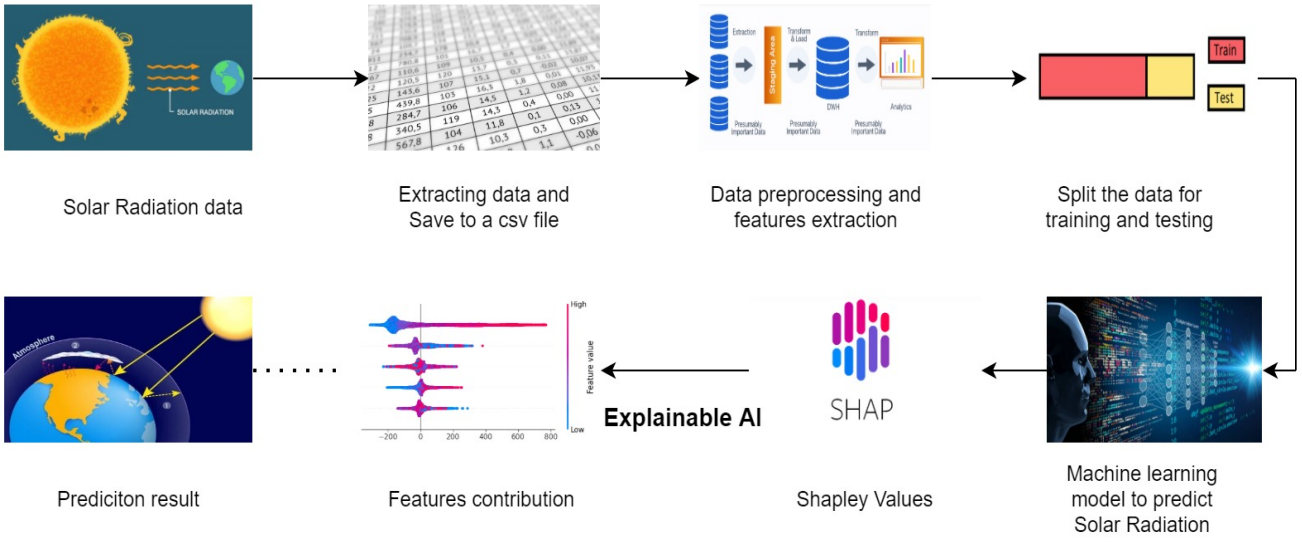

Fig. 1: Framework for solar radiance prediction

created for regression problems. It is based on the LightGBM framework, a gradient-boosting framework that uses methods from tree-based learning. In order to make optimal predictions for continuous target variables, LGBMRegressor successively adds decision trees and minimizes the loss function. Because of its effectiveness and ability to handle huge datasets, it is a popular choice for regression challenges requiring high performance and accuracy. [14]

- Extreme Gradient Boosting Regressor (XGBRegressor): A powerful machine learning algorithm created particularly for regression problems is called XGBRegressor. It builds an ensemble of weak regression models that repeatedly fix mistakes using the XGBoost framework and gradient boosting. The effectiveness, scalability, and capacity for handling large datasets of XGBRegressor are well recognized. It captures intricate relationships, automatically manages missing variables, and makes precise predictions. [14]

D. Explainable AI: SHAP

Explainable AI (XAI) focuses on making AI models transparent and interpretable. Shapley Additive Explanations (SHAP)[1] is an Explainable AI technique that provides insights into predictions by attributing a "Shapley value" to each input. These values quantify the individual contributions of inputs toward the model's overall output. SHAP evaluates different feature combinations for evaluating their impact on predictions, giving a consistent knowledge of feature importance. Visualizing SHAP values help provide intuitive explanations. Overall, SHAP is an important tool in the field of explainable AI because it enables us to understand the inner workings of complex models and present human-understandable justifications for their predictions. This builds confidence and makes it easier to deploy AI systems responsibly. [15]
We are using two types of graph plots for this research to understand the feature importance using SHAP. They are:

- Bee swarm plot: The distribution of data points for various categories is shown on a bee swarm plot, a kind of data visualization. Each data point is represented as a "bee" or marker along the category axis, with their vertical position denoting the value of a continuous variable.
- Bar plot: The contribution of several characteristics to the result of a machine learning model is visually represented by a bar plot. It makes use of horizontal bars, each of which stands for a feature and whose length reflects the feature's influence on the prediction. The prediction increases when contributions are positive, whereas it decreases when contributions are negative. The bar plot makes it easier to understand the important features and how they affect the results of the model.

IV. Experiment Setup

A. Dataset

The dataset used for our study was collected from Kaggle [8]. The dataset consists of 32686 training samples and 11 columns. The columns are UNIXTime, Data, Time, Radiation, Temperature, Pressure, Humidity, WindDirection(Degrees), Speed, TimeSunRise, and TimeSunSet. Out of the features listed in the dataset column, we utilize 6 of them for our experiment. Figure 2 represents the visualization of Solar Radiation Prediction data using a heatmap.

[1]https://learn-scikit.oneoffcoder.com/shap.html

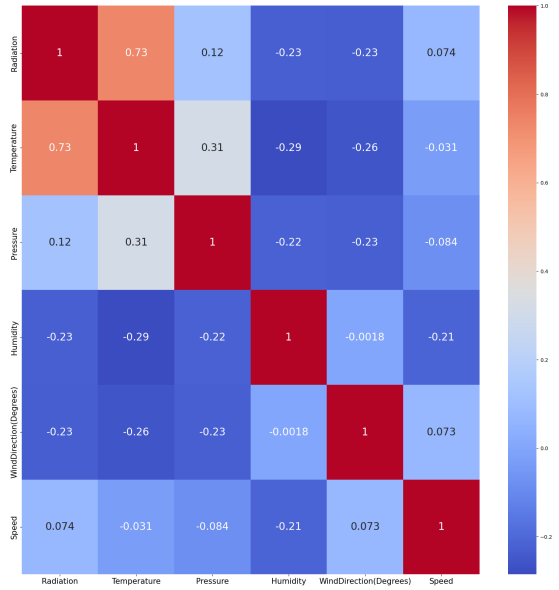

Fig. 2: Visualizing solar radiation prediction data using heatmap

B. Evaluation Metrics

The study takes into account the following metrics while evaluating the models:

- R-squared (R^2) Score: R-squared, a statistical metric with a range from 0 to 1, indicates how well a regression model fits the data by indicating the percentage of variation in the dependent variable that can be accounted for by the independent variables.
- Mean Absolute Error (MAE): The average absolute difference between predicted and actual values is measured by the Mean Absolute Error (MAE), a regression or forecasting statistic. A smaller value of MAE denotes better prediction accuracy and fewer errors.
- Root Mean Squared Error (RMSE): The average magnitude of the discrepancies between predicted and actual values is expressed in regression models using the RMSE measure. It considers both the size and the directional component of the errors. In contrast to measures like MAE, smaller RMSE values indicate better predictive ability, and it is particularly sensitive to outliers, making it a valuable tool to evaluate model accuracy.

The equations for figuring out the three metrics are as follows:

$$R^2 = 1 - \frac{\sum_{ob=1}^{z}(b_{ob} - \hat{b}_{ob})^2}{\sum_{ob=1}^{z}(b_{ob} - \bar{b})^2} \tag{1}$$

$$\text{MAE} = \frac{1}{z}\sum_{ob=1}^{z}|b_{ob} - \hat{b}_{ob}| \tag{2}$$

$$\text{RMSE} = \sqrt{\frac{1}{z}\sum_{ob=1}^{z}(b_{ob} - \hat{b}_{ob})^2} \tag{3}$$

Here,

z represents the collective count of observations,

b_{ob} stands for the actual value of the ob-th observation,

$\hat{b}_{ob}$ stands for the forecasted value of the ob-th observation,

$\bar{b}$ represents the average of the actual values,

$||$ indicates the absolute value.

C. Results and Analysis

In the following, the results have been discussed in two parts: i) Model evaluation based on error, ii) Model interpretation ability based on feature importance.

TABLE I: Evaluation Metrics of Machine Learning Models

Models	Evaluation Metrics		
	R^2 Score	MAE	RMSE
Linear Regression	0.568	157.743	207.396
SVR	0.533	125.669	215.687
Gradient Boosting	0.728	99.230	164.736
XGBoost	0.729	99.980	164.400
LightGBM	0.734	98.421	162.777

1) Error-based Model Evaluation: Evaluation metrics for machine learning models are represented in Table I. Besides, Figure 3 illustrates a bar plot of evaluation metrics for different models. The results indicate that Gradient Boosting, XGBoost, and LightGBM exhibit the lowest error and perform as the best model for this dataset. In contrast, SVR and linear regression both perform far less effectively than the top three models.

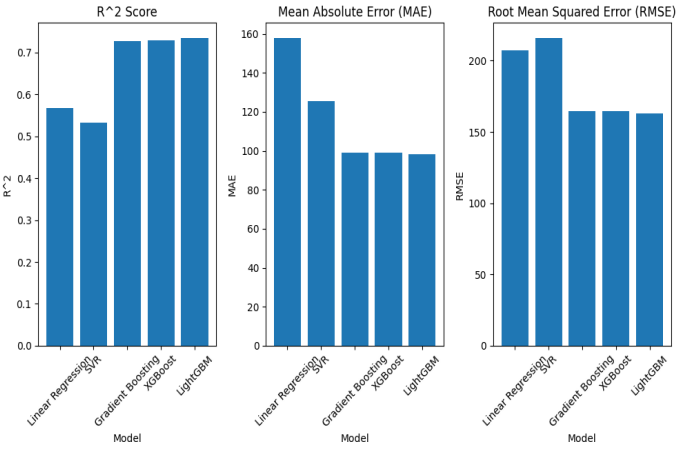

Fig. 3: Bar Plot of evaluation metrics for different models

2) Model Explainability:

- Linear Regressor: According to the Linear Regressor Model, Temperature emerged as the predominant factor, as evident from the analysis of Figure 4 and Figure 5. The next crucial feature was Pressure, then Speed and Wind Direction (Degrees). And the less important feature was Humidity according to the Linear Regressor Model.
- Gradient Boosting Regressor: From Figure 6 and Figure 7, it comes as no surprise that the most crucial feature

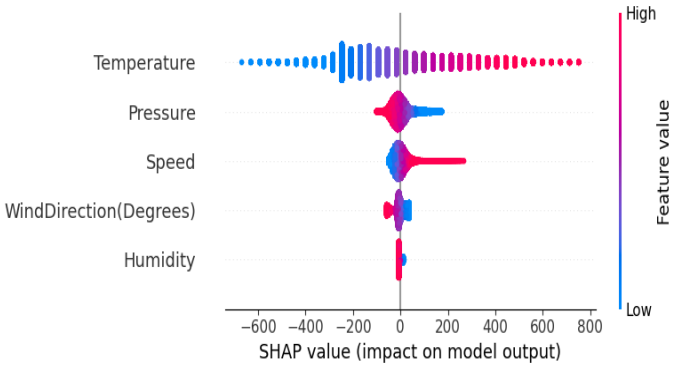

Fig. 4: Bee Swarm plot for Linear Regressor Model

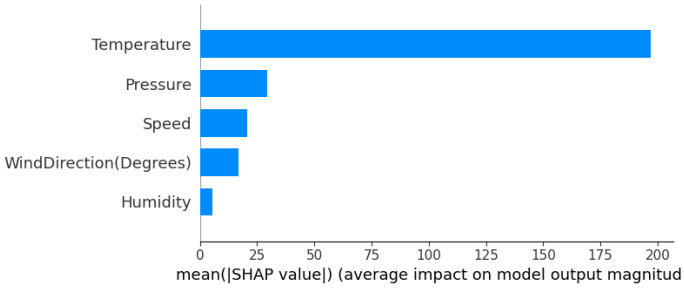

Fig. 5: Bar plot for Linear Regressor Model

was Temperature. But it is worth mentioning that, unlike the Linear Regressor Model, the Gradient Boosting Regressor Model gives higher importance to Wind Direction (Degrees) and Humidity compared to Pressure and Speed.

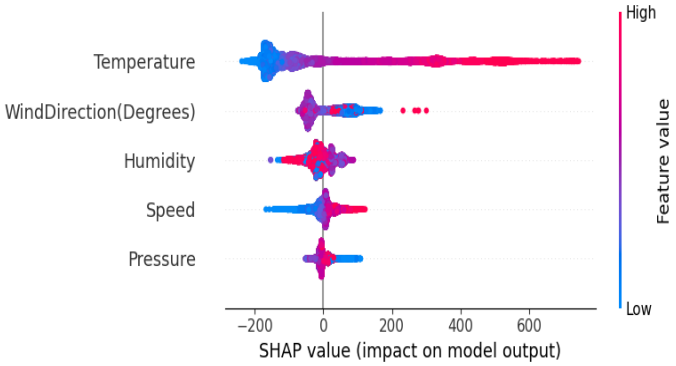

Fig. 6: Bee Swarm plot for Gradient Boosting Regressor Model

- Light Gradient Boosting Regressor (LGBM Regressor): From Figure 8 and Figure 9, we can see that the most important feature was Temperature in this model as well. The significance of the other features was similar to Gradient Boosting Regressor Model.
- Extreme Gradient Boosting Regressor (XGBRegressor): As can be seen from Figure 10 and Figure 11, the feature importance was almost similar to LGBM Regressor Model and Temperature was the most important feature here also.

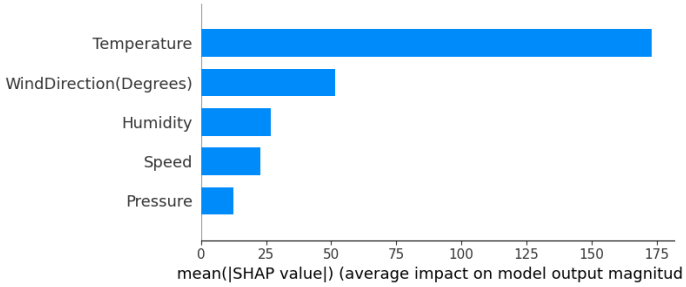

Fig. 7: Bar plot for Gradient Boosting Regressor Model

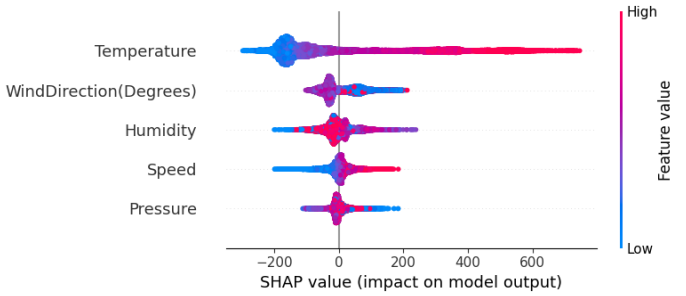

Fig. 8: Bee Swarm plot for LGBM Regressor Model

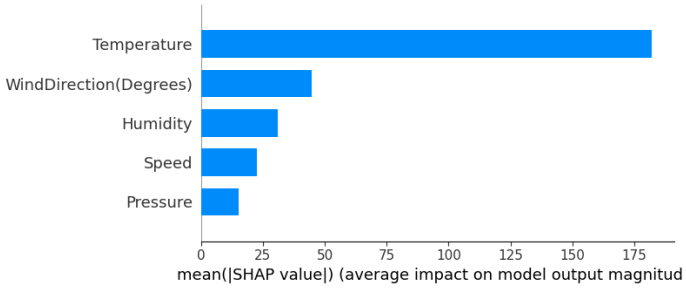

Fig. 9: Bar plot for LGBM Regressor Model

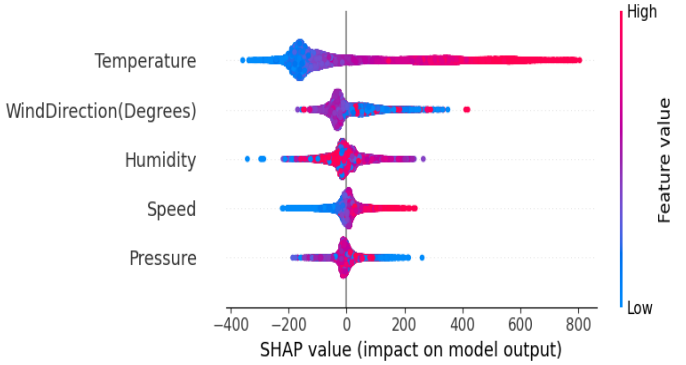

Fig. 10: Bee Swarm plot for XGBRegressor Model

Based on the evaluation metrics, Gradient Boosting, XGBoost, and LightGBM demonstrate superior performance with the

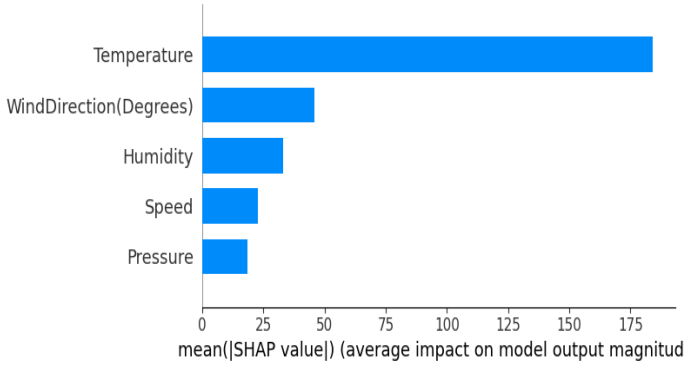

Fig. 11: Bar plot for XGBRegressor Model

lowest error rates among all models evaluated on this dataset. Conversely, SVR and linear regression exhibit significantly less effective performance compared to the top three models.

Furthermore, the analysis of SHAP values reveals that when considering only one feature, Temperature emerges as the most crucial factor for accurately predicting the models' output across all machine learning algorithms. Additionally, Wind Direction (Degrees) emerges as the second most influential feature specifically for Gradient Boosting, XGBoost, and LightGBM.

V. Conclusion

Solar radiation is an important factor considering the recent increase in global warming. The accurate prediction of that will help to correct utilization of it though solar energy consumption as well as have taken the plan to decrease global warming. To balance prediction accuracy and model interpretability, this paper concludes by discussing the performance analysis of ensemble methods for solar radiation prediction based on explainable AI. Linear Regression, SVR, Gradient Boosting, XGBoost, and LightGBM were five ensemble algorithms used in our study to evaluate a dataset. Gradient Boosting, XGBoost, and LightGBM beat the other models with noticeably reduced error rates, according to assessment measures. Additionally, through the computation of SHAP values to determine feature importance, we consistently found Temperature to be the dominant factor in predicting model outcomes across all machine learning algorithms.

Future research could study the integration of ensemble algorithms with other advanced techniques like deep learning or hybrid models to further improve prediction accuracy. In addition, including other geographical and climatic variables could expand the scope of the study and provide a more comprehensive understanding of solar radiance dynamics. Moreover, investigating alternative explainable AI methods and their interpretations could provide further insights into feature importance and model predictions.

References

[1] Ü. Ağbulut, A. E. Gürel, and Y. Biçen, "Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison," Renewable Sustainable Energy Reviews, vol. 135, p. 110114, Jan. 2021, doi: 10.1016/j.rser.2020.110114.

[2] O. Bamisile, M. Aftab, C. J. Ejiyi, N. Yimen, S. Obiora, and Q. Huang, "Comparison of machine learning and deep learning algorithms for hourly global/diffuse solar radiation predictions," International Journal of Energy Research, vol. 46, no. 8, pp. 10052–10073, Feb. 2021, doi: 10.1002/er.6529.

[3] Shuai, Y., Notton, G., Wongwises, S., Nivet, M. L., Paoli, C., Motte, F., Fouilloy, A. (2017b). Machine learning methods for solar radiation forecasting: A review. Renewable Energy, 105, 569–582. https://doi.org/10.1016/j.renene.2016.12.095

[4] P. Kumari and D. Toshniwal, "Deep learning models for solar irradiance forecasting: A comprehensive review," Journal of Cleaner Production, vol. 318, p. 128566, Oct. 2021, doi: 10.1016/j.jclepro.2021.128566.

[5] Zhou, Y., Liu, Y., Wang, D., Liu, X., Wang, Y. (2021). A review on global solar radiation prediction with machine learning models in a comprehensive perspective. Energy Conversion and Management, 235, 113960. https://doi.org/10.1016/j.enconman.2021.113960

[6] E. D. Obando, S. X. Carvajal and J. Pineda Agudelo, "Solar Radiation Prediction Using Machine Learning Techniques: A Review," in IEEE Latin America Transactions, vol. 17, no. 04, pp. 684-697, April 2019, doi: 10.1109/TLA.2019.8891934.

[7] Md. S. Alam, F. S. Al-Ismail, M. S. Hossain, and S. M. Rahman, "Ensemble Machine-Learning Models for Accurate Prediction of Solar Irradiation in Bangladesh," Processes, vol. 11, no. 3, p. 908, Mar. 2023, doi: 10.3390/pr11030908.

[8] "Solar Radiation Prediction," https://www.kaggle.com/datasets/dronio/SolarEnergy, [Online; accessed 2023-05-14].

[9] Machlev, R., Heistrene, L., Perl, M. L., Levy, K., Belikov, J., Mannor, S., Levron, Y. (2022b). Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. Energy and AI, 9, 100169. https://doi.org/10.1016/j.egyai.2022.100169

[10] Wang, H., Cai, R., Zhou, B., Aziz, S., Qin, B., Voropai, N., Gan, L., Barakhtenko, E. (2020). Solar irradiance forecasting based on direct explainable neural network. Energy Conversion and Management, 226, 113487. https://doi.org/10.1016/j.enconman.2020.113487

[11] Li, J., Ward, J. M., Tong, J., Collins, L., Platt, G. (2016). Machine learning for solar irradiance forecasting of photovoltaic system. Renewable Energy, 90, 542–553. https://doi.org/10.1016/j.renene.2015.12.069

[12] Fouilloy, A., Shuai, Y., Notton, G., Motte, F., Paoli, C., Nivet, M. L., Guillot, E., Duchaud, J. (2018). Solar irradiation prediction with machine learning: Forecasting models selection method depending on weather variability. Energy, 165, 620–629. https://doi.org/10.1016/j.energy.2018.09.116

[13] Sarp, S., Knzlu, M., Cali, U., Elma, O., Guler, O. (2021). An Interpretable Solar Photovoltaic Power Generation Forecasting Approach Using An Explainable Artificial Intelligence Tool. https://doi.org/10.1109/isgt49243.2021.9372263

[14] Bentéjac, C., Csörgő, A., Martínez-Muñoz, G. (2021b). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

[15] Holzinger, A., Saranti, A., Molnar, C., Biecek, P., Samek, W. (2022). Explainable AI Methods - A Brief Overview. In Springer eBooks (pp. 13–38). https://doi.org/10.1007/978-3-031-04083-22