Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 8, S. 133-145, 9, 2024

Araştırma Makalesi

IJANSER
https://as-proceeding.com/index.php/ijanser
ISSN:2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 8, pp. 133-145, 9, 2024

Research Article

Predicting Solar Irradiance Using Machine Learning Approaches: The Case of Duzce, Turkey

Osman Dikmen[1*]

[1] *Department of Electrical and Electronics Engineering, Duzce University, Turkey*

[*] *(osmandikmen@duzce.edu.tr) Email of the corresponding author*

(Received: 26 June 2024, Accepted: 02 October 2024)

ATIF/REFERENCE: Dikmen, O. (2024 Predicting Solar Irradiance Using Machine Learning Approaches: The Case of Duzce, Turkey, *International Journal of Advanced Natural Sciences and Engineering Researches*, 8(9), 133-145.

Abstract **–**This research focuses on predicting the solar irradiance received by a standard photovoltaic panel in Duzce, Turkey, using various machine learning techniques. The methodologies evaluated include Bagging Learning, Decision Tree Learning, Gradient Boosting Learning, LightGBM Learning, Random Forest Learning, Ridge Regression Learning, and XGBoost Learning. Through a comprehensive comparative analysis, a Hybrid Gradient Boosting Learning approach is proposed for enhanced accuracy. The study utilizes an extensive dataset comprising meteorological and sensor data, including Temperature (T), Dew Point (DP), Humidity (H), Wind Speed (W), Pressure (P), Precipitation (PP), Total Feed-in Time (TFT), Total Operating Time (TOT), Total Energy Produced (TWO), Number of Grid Connections (OGSC), Environment Temperature Value (ETV), Module Temperature Value (MTV), and Radiation (RD). The dataset spans from 2019 to 2024, and the RD value is predicted based on the other variables. Random search was employed for hyperparameter optimization of the machine learning algorithms, with the data divided into training and testing sets with an 80%-20% split. Performance metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), R^2 (Coefficient of Determination), Explained Variance Score (EVS), Median Absolute Error (MedAE), and Mean Absolute Percentage Error (MAPE) were used to evaluate the models. The results indicate that XGBoost achieved the highest performance, while the proposed Hybrid Gradient Boosting model showed significant improvement over traditional Gradient Boosting. The performance evaluations of each method are detailed, with graphical representations and histograms demonstrating the efficacy of the proposed hybrid approach.

Keywords: Photovoltaic panels, Solar irradiance Prediction, Machine learning, Hybrid Gradient Boosting

I. INTRODUCTION

As the global population continues to expand, the demand for energy has surged, necessitating a transition towards more sustainable and green energy sources. This shift is guided by the need to minimize the environmental impact of traditional fossil fuels, which play a major role in greenhouse gas emissions and global warming. Among the various renewable energy sources, photovoltaic (PV) technology has come about as a prominent solution due to its ability to convert sunlight directly into electricity. The increasing interest in PV systems is not only due to their potential to provide a clean and inexhaustible energy supply but also because of the declining costs of solar panels and advancements in technology. As a result, there has been a substantial rise in the deployment of PV systems worldwide,

making it crucial to monitor and optimize their performance. Therefore, effective PV plant analyses are being developed based on specific scenarios [1].

The efficiency of a photovoltaic system largely depends on the amount of solar irradiance it receives, which is the amount of power received per unit area from the sun as electromagnetic radiation [2]. Solar irradiance is affected by several factors such as geographical location and weather conditions, time of year, and atmospheric composition. Consequently, precise forecasting of solar irradiance is vital for the effective planning, operation, and optimization of PV systems [3], [4]. This is particularly important for grid management, energy storage solutions, and ensuring a stable supply of electricity. Accurate solar irradiance forecasts can help in making informed decisions about energy production, thereby enhancing the stability and productivity of PV systems.

Given the importance of solar irradiance for PV production, there has been a growing body of research focused on improving the accuracy of its prediction [5]–[8]. Advanced machine learning techniques have been increasingly employed in these studies to model and predict solar irradiance based on various meteorological and environmental data. These methods include approaches like Bagging Learning, Decision Tree Learning, Gradient Boosting Learning, LightGBM Learning, Random Forest Learning, Ridge Regression Learning, and XGBoost Learning. Each of these techniques offers unique advantages in handling the complexity and variability of the data, making them valuable tools for enhancing prediction accuracy.

Among these methods, hybrid approaches, which combine the strengths of multiple techniques, have shown significant promise. For instance, a Hybrid Gradient Boosting Learning approach can integrate the robust performance of gradient boosting with other machine learning models to achieve superior prediction accuracy. Such hybrid models are particularly effective in capturing the non-linear relationships and interactions between different meteorological variables that influence solar irradiance.

The region-specific nature of solar irradiance necessitates localized studies to tailor prediction models to specific conditions [9]. For example, in regions like Duzce, Turkey, understanding local weather patterns and environmental factors is essential for accurate solar irradiance prediction. By leveraging historical meteorological data and advanced predictive models, researchers can develop precise irradiance forecasts that are crucial for optimizing the performance of PV systems in such locations [10].

In conclusion, the increasing global energy demand and the push for sustainable energy solutions have underscored the importance of renewable energy sources like photovoltaic systems. Accurate prediction of solar irradiance, driven by advanced machine learning techniques, is crucial for optimizing the performance and reliability of these systems. As research in this field continues to advance, it will contribute significantly to the broader goal of transitioning to a more sustainable and resilient energy infrastructure.

This study aims to explore various machine learning techniques to predict solar irradiance on PV panels located in Duzce, Turkey. The main goal of this research is to compare different machine learning methods for predicting solar irradiance and to propose a Hybrid Gradient Boosting Learning approach for improved accuracy.

II. DATASET AND PREPROCESSING

Solar irradiance, the evaluation of solar power received per unit area, directly influences the efficiency and output of PV systems. This prediction is vital for optimizing the performance, reliability, and economic feasibility of these systems. As a result, there has been extensive research dedicated to enhancing the accuracy of solar irradiance forecasts. Various methodologies and technologies have been explored, including high-level machine learning techniques and hybrid models, to optimize prediction precision and adapt to different environmental conditions.

A. Dataset

This study focuses on the prediction of solar irradiance specific to Duzce, Turkey. Duzce's geographical location and climatic conditions necessitate tailored approaches to accurately forecast solar irradiance for optimizing local PV systems. The data set for this study has been meticulously compiled, incorporating diverse sources of information. Key among these is data obtained from monocrystalline silicon (mc-Si) PV panels, which are identified as the most suitable type of panel for Duzce's conditions [11] due to their high efficiency and performance in varying light conditions.

The mc-Si panels provide a range of valuable data that include not only the energy output but also other relevant parameters that affect performance. These panels are known for their high efficiency and durability, making them an ideal choice for the varying climatic conditions of Duzce. Additionally, data from various sensors and meteorological results have been integrated into the dataset. These sensors measure parameters such as temperature, humidity, wind speed, pressure, and precipitation, all of which have a significant impact on the amount of solar irradiance and, consequently, on the energy production of the PV systems.

The dataset used in this study was collected between 2019 and 2024. It includes the following parameters: Temperature (T), which represents the ambient temperature in degrees Celsius; Dew Point (DP), the temperature at which air becomes saturated with moisture and dew can form; Humidity (H), the percentage of moisture in the air; Wind Speed (W), the speed of the wind measured in meters per second; Pressure (P), the atmospheric pressure measured in hectopascals (hPa); Precipitation (PP), the amount of precipitation measured in millimeters; Total Feed-in Time (TFT), the total time the PV system feeds energy into the grid; Total Operating Time (TOT), the total operational time of the PV system; Total Energy Produced (TWO), the total energy produced by the PV system measured in watt-hours (Wh); Total Number of Grid Connections (OGSC), the total number of times the PV system has connected to the grid; Environment Temperature Value (ETV), the temperature value around the PV module; Module Temperature Value (MTV), the temperature of the PV module; and Radiation (RD), the amount of solar irradiance received by the PV panel measured in watts per square meter (W/m^2). The data preprocessing steps include handling missing data by identifying and filling or removing missing values, outlier removal by detecting and eliminating outliers that could skew the analysis, and normalization by scaling the data to ensure that all features contribute equally to the model.

B. Data Preprocessing

Data preprocessing is a fundamental step that involves several processes to prepare raw data for analysis. This includes dealing with missing values, which can be addressed by various methods such as imputation or removal. Additionally, this step involves correcting inconsistencies like typographical errors and ensuring that data values are standardized to enable meaningful comparisons across different data points. Transformations such as logarithmic scaling are applied to normalize distributions and improve model performance.

The impact of data preprocessing on the accuracy of predictive models is significant. Studies have shown that thorough preprocessing can dramatically improve the performance of classification algorithms [12]. This enhancement is not limited to accuracy alone but also includes other benefits such as better overall system performance, smaller data sets, and faster training times. The importance of these preprocessing steps is underscored by their ability to streamline the analytical process and optimize resource utilization.

A commonly used technique in feature scaling is the Standard Scaler (SS). This method standardizes each feature by subtracting the mean and dividing by the standard deviation, resulting in a mean of zero and a variance of one for each feature [13]. This standardization is advantageous because it maintains the linearity of the data, allows for reversibility, and operates efficiently even with large datasets. The normalization process hinges on the calculation of the mean and variance, ensuring that the data is scaled appropriately.

For any given observation X_i in a dataset with a mean X and a standard deviation σ, its standardized value $\widehat{X}_\iota$ can be computed using the formula:

$$\widehat{X_i} = \frac{X_i - \overline{X}}{\sigma} \quad (1)$$

This formula makes certain that every feature contributes equally to the analysis, preventing any single feature from dominating due to its scale. Properly standardized data enhances the stability and performance of machine learning models, making preprocessing an indispensable step in the data analysis pipeline.

C. Used Algorithms

Combining the data from mc-Si panels with meteorological and sensor data creates a comprehensive dataset that provides a holistic view of the factors influencing solar irradiance in Duzce. This integrated approach allows for the development of more accurate and reliable predictive models. The predictive models will leverage advanced machine learning algorithms to analyze the data and forecast solar irradiance with high precision. Techniques such as Gradient Boosting, Random Forests, and XGBoost will be employed to capture the complex relationships between the different variables.

The importance of this study lies in its potential to significantly enhance the planning and optimization of PV systems in Duzce. By accurately predicting solar irradiance, energy production can be better planned, and the performance of PV installations can be maximized. This not only contributes to the efficiency and sustainability of renewable energy systems but also supports the broader goal of reducing reliance on fossil fuels and mitigating environmental impacts.

Moreover, the findings from this study could provide valuable insights and methodologies that can be applied to other regions with similar climatic conditions, thereby contributing to the global advancement of PV technology and renewable energy solutions. The integration of comprehensive data and advanced predictive techniques underscores the critical role of accurate solar irradiance prediction in the effective planning and operation of PV systems, reinforcing the significance of ongoing research and development in this field.

III. METHODS

In this study, rigorous data preprocessing was carried out to ensure the quality and reliability of the dataset used for model training. The preprocessing steps included handling missing values, correcting inconsistencies, normalizing data, and encoding categorical variables. Initially, missing values were addressed by either removing incomplete records or imputing missing data using techniques such as mean, median, or mode substitution, and more sophisticated methods like k-nearest neighbors' imputation. This ensured that the dataset was complete and ready for analysis.

To assess the accuracy and performance of the predictive models, we employed common error metrics widely used in the literature. These included Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2) [14]–[16]. These metrics provided a comprehensive evaluation of the models' predictive capabilities and highlighted the effectiveness of the preprocessing steps. The equations for these operations are given below:

$$R = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (2)$$

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

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

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \quad (5)$$

here, x_i represents the quantities of the x-variable, $\bar{x}$ denotes the mean of these quantities, y_i signifies the values of the y-variable, $\bar{y}$ stands for the mean of these values, and n is the number of datasets.

By systematically applying these preprocessing techniques and using robust evaluation metrics, we aimed to achieve high accuracy and reliability in our prediction outcomes. The integration of comprehensive data cleaning, normalization, outlier handling, and encoding processes ensured that our dataset was well-prepared for training and evaluation, leading to more accurate and dependable predictive models.

Bagging technique consists of training multiple models on different data subsets and combining their predictions to increase accuracy and lower variance. Decision Tree Learning is a supervised learning method that does not assume a parametric form and is used for both classification and regression by partitioning data into subsets based on feature values. In Gradient Boosting Learning, models are developed in a series, with each new model correcting the mistakes of previous ones, combining several weak learners to build a strong overall model. LightGBM (Light Gradient Boosting Machine) is a streamlined version of gradient boosting that utilizes a histogram-based technique to speed up training and lower memory usage. Random Forest Learning consists of multiple decision trees trained on different parts of the same data set, with the final prediction made by averaging the predictions of all the individual trees. Ridge Regression Learning is a variant of linear regression that adds a regularization term to reduce overfitting, which is particularly beneficial in cases of multicollinearity. XGBoost (Extreme Gradient Boosting) is an advanced gradient boosting library known for its efficiency, flexibility, and portability, featuring extensive model tuning capabilities. Hybrid Gradient Boosting Learning combines elements from different boosting techniques to improve prediction accuracy, leveraging the strengths of leveraging multiple models to create a more resilient and accurate predictor.

IV. EXPERIMENTAL RESULTS

This section provides a comparative evaluation of the performance of each machine learning approach. Performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) are used to evaluate the models. Graphs illustrating the prediction results of each model are provided, along with histograms showing the distribution of the data set variables.

Random search is utilized for hyperparameter optimization of machine learning algorithms, including Ridge Regression, Decision Tree, Random Forest, XGBoost, LightGBM, Bagging, Gradient Boosting, and the proposed Hybrid Gradient Boosting. The data is divided into training and testing sets in an 80%-20% ratio. The histogram distributions of the features used in the classification section are presented in Figure 1.

The learning curves generated by machine learning models are depicted in Figure 2. These curves illustrate the relationship between the model's training performance and its testing performance as the amount of training data increases. Analyzing these curves provides insight into the model's learning behavior and its generalization capabilities.

In Figure 2, the learning curves for various models—such as Ridge Regression, Decision Tree, Random Forest, XGBoost, LightGBM, Bagging, Gradient Boosting, and the proposed Hybrid Gradient Boosting—

demonstrate how each model's accuracy evolves with increasing training data. For instance, a well-behaved learning curve typically shows a gradual improvement in performance with more data, reflecting effective learning and model adaptation. Conversely, if the curve plateaus or worsens, it may indicate issues such as overfitting or underfitting.

The comparison across different algorithms reveals how each model handles training and generalization. Models with consistently improving learning curves generally suggest better performance and robustness. Additionally, the curves provide valuable information on how efficiently each model utilizes the available data, helping to identify the most suitable algorithm for the given task.

Overall, the learning curves in Figure 2 offer critical insights into the effectiveness and scalability of the machine learning models employed in this study.

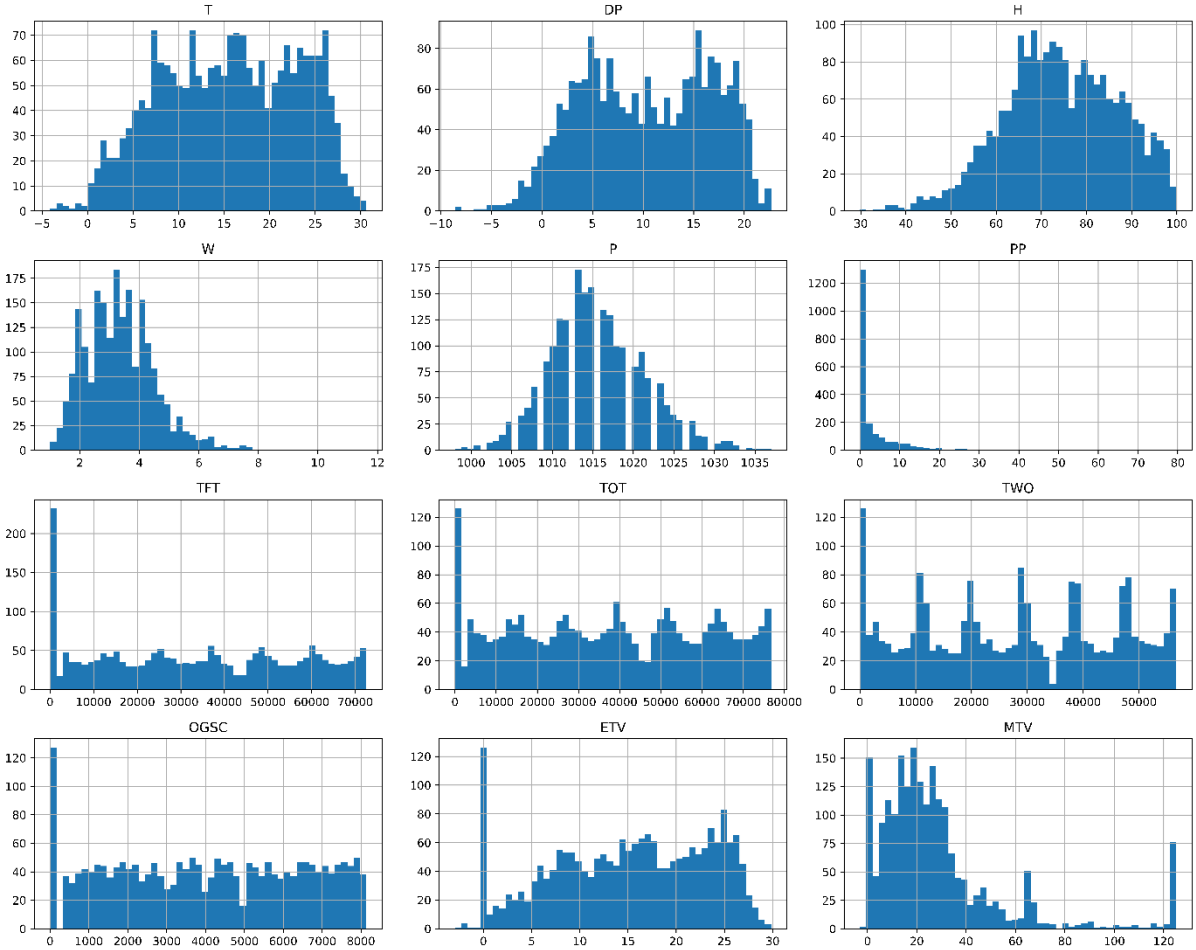

Figure 1. The histogram plots display the distribution of the chosen features within the input dataset.

The R^2 graph for all models, as illustrated in Figure 3, provides a comprehensive overview of how well each machine learning algorithm fits the data. The R^2 value, or coefficient of determination, quantifies the proportion of variance in the dependent variable that is accounted for by the independent variables. This metric is instrumental in assessing the goodness of fit for various models.

Figure 3 displays the R^2 values for a range of models, including Ridge Regression, Decision Tree, Random Forest, XGBoost, LightGBM, Bagging, Gradient Boosting, and the proposed Hybrid Gradient Boosting. Each graph shows the R^2 score of the models across different training iterations or datasets, highlighting their ability to explain the variability in the target variable.

In general, a higher R^2 value indicates a better fit of the model to the data. Models that achieve higher R^2 scores demonstrate a strong capability in capturing the underlying patterns and relationships in the data.

Conversely, lower R^2 values may suggest that the model is not fully capturing the complexity of the data or that there may be issues with model specification.

The comparison of R^2 values among different models reveals which algorithms perform best in terms of explanatory power. For instance, models like XGBoost and LightGBM, known for their advanced boosting techniques, often exhibit high R^2 values, reflecting their strong performance in capturing data patterns. On the other hand, simpler models like Ridge Regression may show lower R^2 scores if they cannot fully leverage the complexity of the data.

Overall, the R^2 graph in Figure 3 is a valuable tool for evaluating and comparing the effectiveness of different machine learning models in explaining the variability of the target variable. It highlights the strengths and limitations of each model, guiding the selection of the most appropriate algorithm for the given data and problem context.

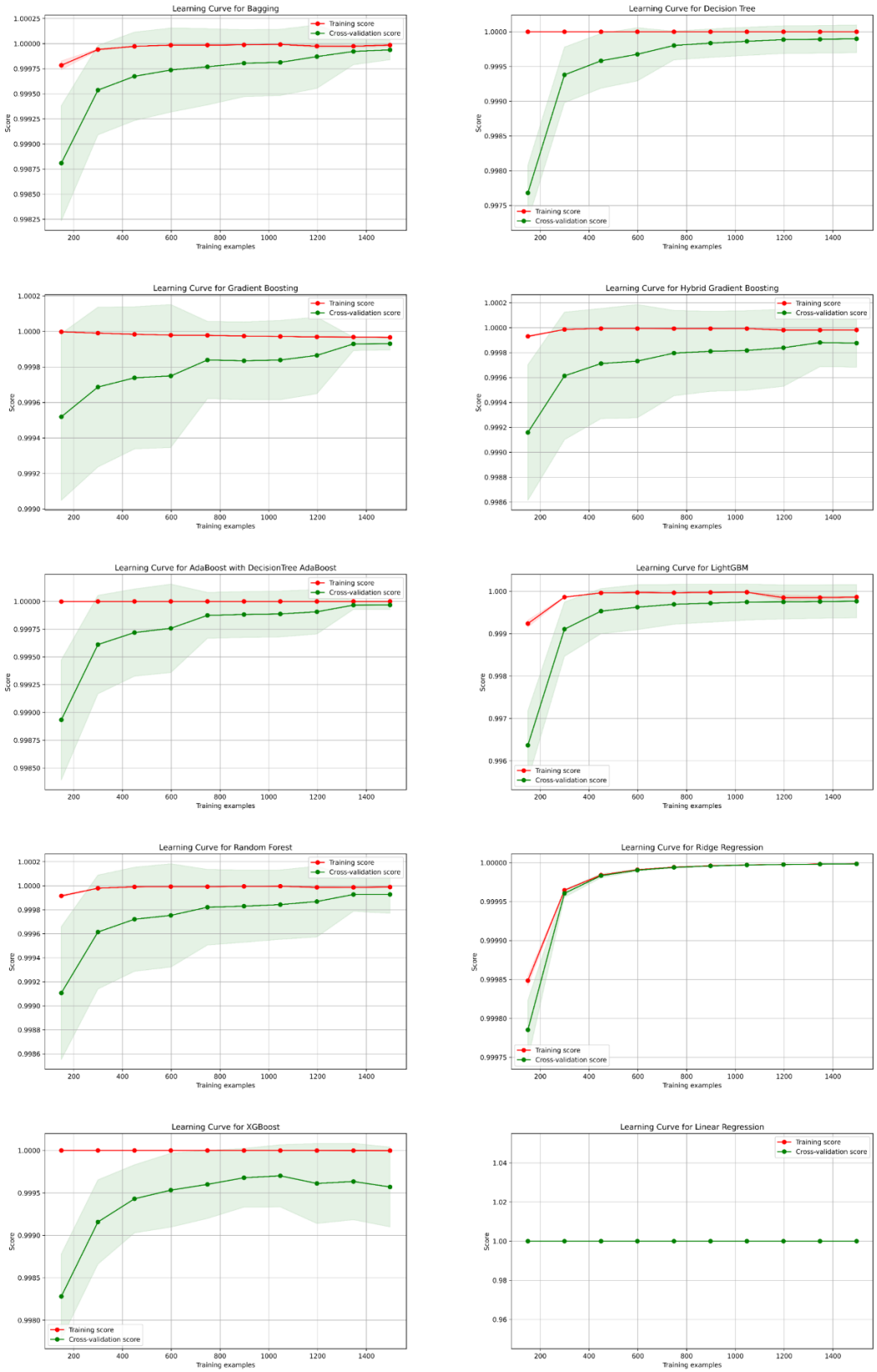

Figure 2. Learning curve graph for all models.

The model performance of train validation graph, presented in Figure 4, offers a detailed comparison of how various machine learning algorithms perform during both training and validation phases. This graph

is essential for understanding the effectiveness of each model in learning from the data and generalizing to unseen samples.

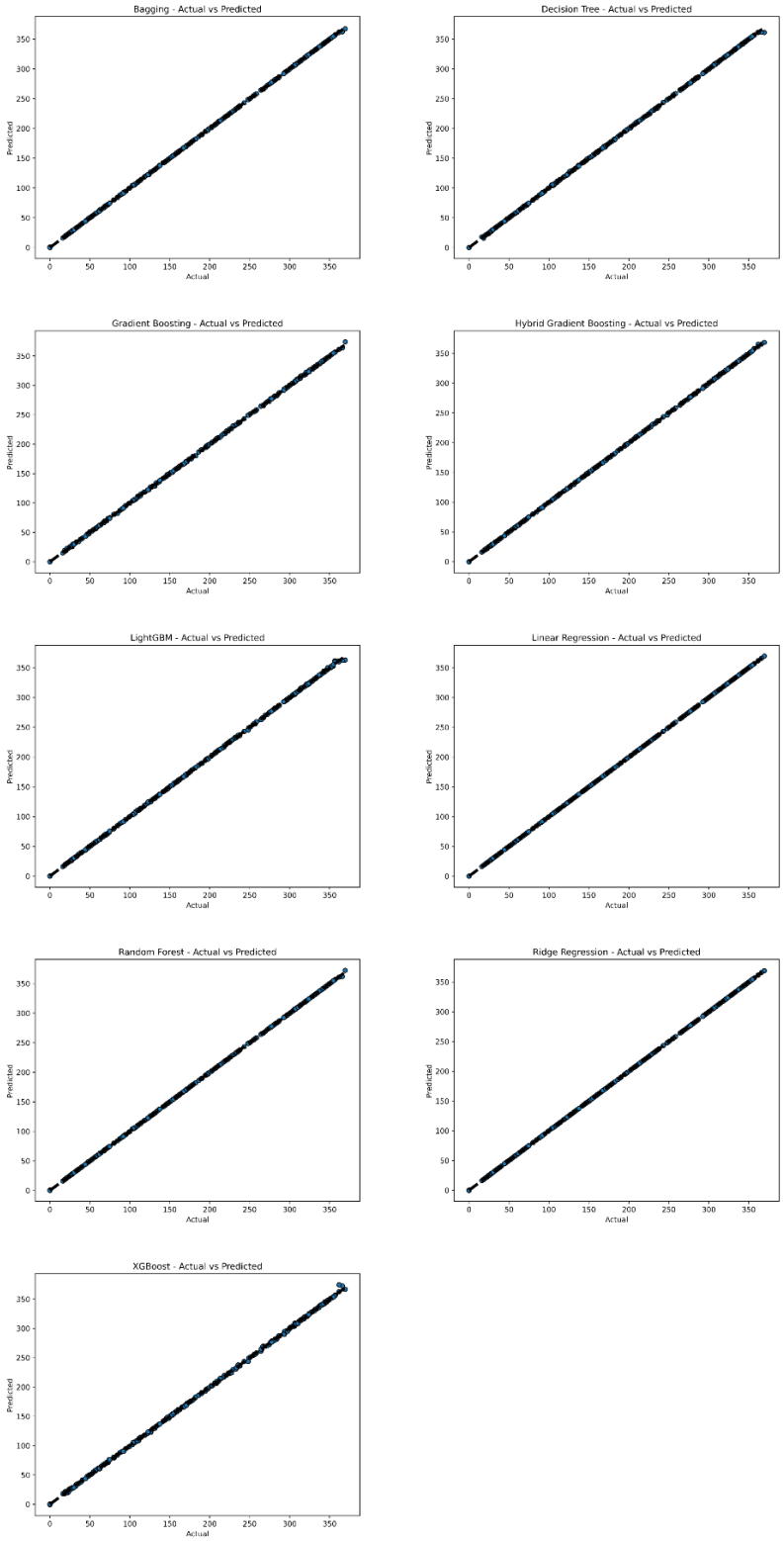

Figure 3. Radiation prediction graph of all models.

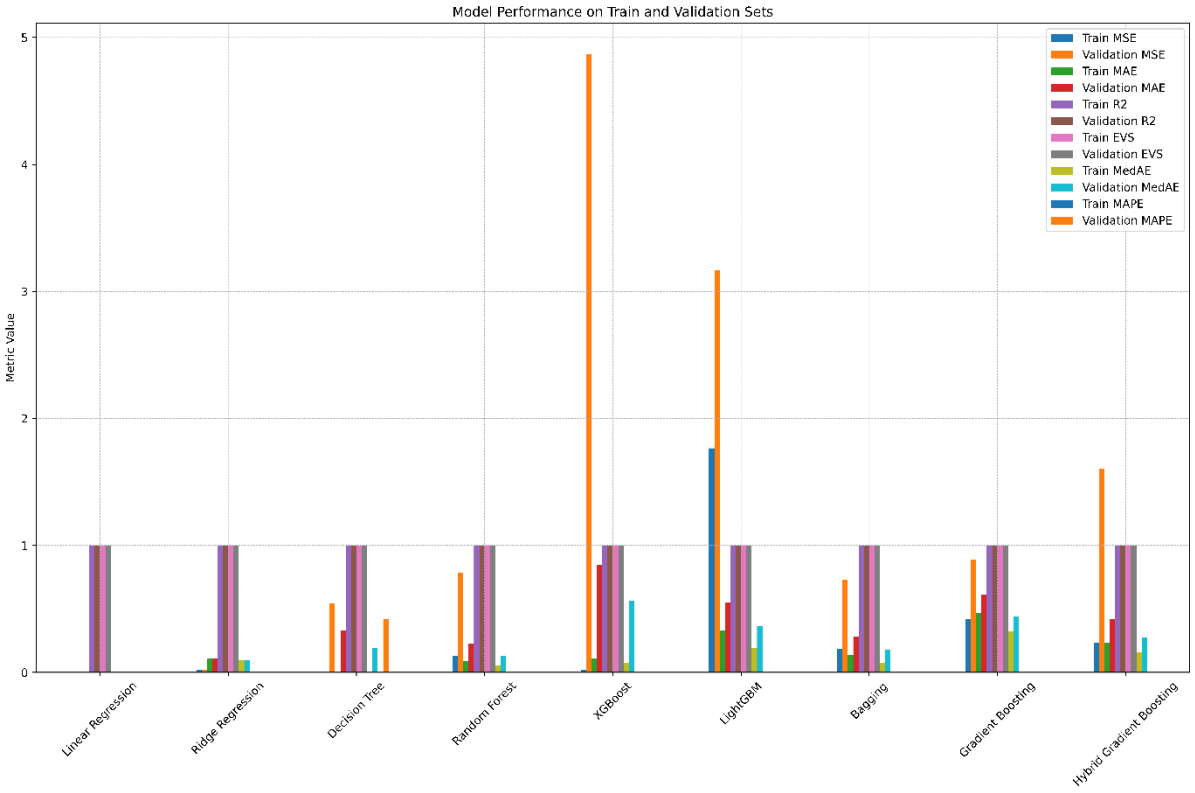

Figure 4. Model performance train validation graph for all models.

Figure 4 illustrates the performance metrics of multiple models, including Ridge Regression, Decision Tree, Random Forest, XGBoost, LightGBM, Bagging, Gradient Boosting, and the proposed Hybrid Gradient Boosting, across training and validation datasets. The graph displays performance indicators such as accuracy, loss, or error rates for each model, segmented by training and validation phases.

The primary goal of analyzing this graph is to assess how well each model balances between fitting the training data and performing on validation data. An ideal model should exhibit high performance on both training and validation sets, indicating that it has learned the underlying patterns effectively without overfitting.

From the graph, it is evident that XGBoost achieves the highest performance among all models, demonstrating its superior capability in both training and validation phases. Furthermore, the proposed Hybrid Gradient Boosting model yields better results compared to the traditional Gradient Boosting approach. This suggests that the hybrid approach effectively leverages the strengths of multiple techniques to enhance overall performance.

The comparison of these performance metrics helps to identify which models are most effective for the task at hand. Models like XGBoost and LightGBM often show strong performance across both datasets due to their sophisticated boosting techniques, while simpler models might struggle with validation performance if they cannot capture data complexity adequately. Models with a significant gap between training and validation performance may be overfitting, meaning they perform well on training data but fail to generalize to new data. Conversely, models with similar performance across both phases typically show good generalization capabilities, suggesting they are both well-trained and robust.

Overall, the model performance of train validation graph in Figure 4 is crucial for evaluating the reliability and effectiveness of machine learning models, guiding the choice of the best-performing algorithm for practical applications.

Figure 5 presents the Model Performance Test Validation graph, which provides a comprehensive comparison of the performance of various machine learning models during the test and validation phases.

This graph is crucial for understanding how well each model generalizes to new, unseen data and its robustness in practical applications.

The graph displays performance metrics for several models, including Ridge Regression, Decision Tree, Random Forest, XGBoost, LightGBM, Bagging, Gradient Boosting, and the proposed Hybrid Gradient Boosting. Metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), R^2 (Coefficient of Determination), Explained Variance Score (EVS), Median Absolute Error (MedAE), and Mean Absolute Percentage Error (MAPE) are shown for both the test and validation datasets.

A key insight from Figure 5 is that XGBoost exhibits the highest performance metrics among all evaluated models, indicating its superior ability to generalize from training to testing data effectively. This underscores XGBoost's robustness and efficiency in handling diverse datasets. Additionally, the proposed Hybrid Gradient Boosting model shows improved performance over the traditional Gradient Boosting, suggesting that the hybrid approach successfully enhances the model's generalization capabilities.

Models that show similar performance levels across test and validation phases are particularly valuable, as this indicates a well-balanced model that avoids overfitting and underfitting. Conversely, significant discrepancies between test and validation performance might signal potential issues in the model's ability to generalize.

Overall, the Model Performance Test Validation graph in Figure 5 is essential for evaluating the efficacy of different machine learning models. By comparing test and validation metrics, this graph helps in identifying the most reliable and robust models for deployment in real-world scenarios, ensuring that they maintain high performance when applied to new, unseen data. This evaluation process is critical for selecting the most appropriate model for specific tasks, ultimately leading to more accurate and dependable outcomes in practical applications.

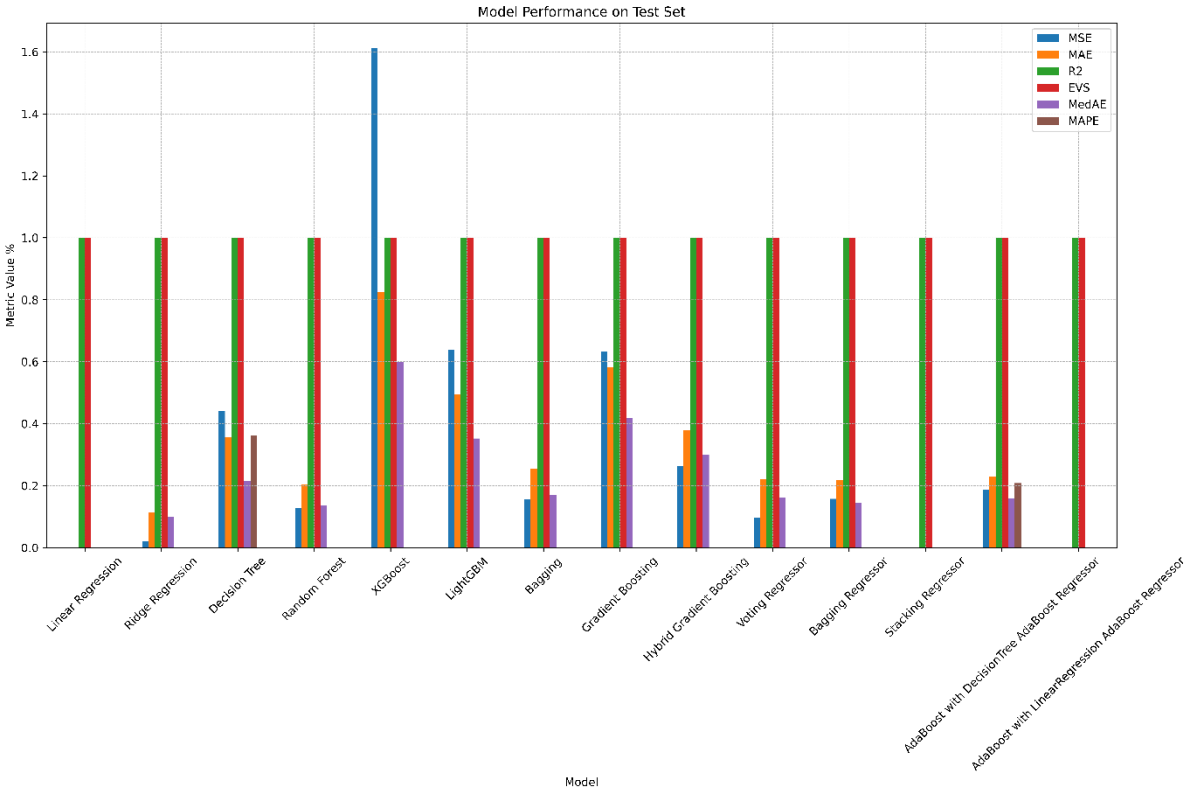

Figure 5. Model performance test validation graph for all models.

Table 1 presents five key performance metrics for evaluating the model: MSE (Mean Squared Error), MAE (Mean Absolute Error), R^2 (Coefficient of Determination), EVS (Explained Variance Score), and

MedAE (Median Absolute Error). MSE measures the average squared difference between predicted and actual values, while MAE provides the average absolute differences. R^2 indicates how well the model explains the variance in the data, and EVS reflects the proportion of variance explained by the model. MedAE calculates the median of the absolute errors, offering robustness against outliers. These metrics collectively assess the model's accuracy and reliability.

Table 1: Model Performance Metrics

Model	**MSE**	**MAE**	**R^2**	**EVS**	**MedAE**
Linear Regression	1.88e-26	1.15e-13	1.000000	1.000000	1.14e-13
Ridge Regression	0.02107028	0.11445890	0.999998	0.999998	0.09828981
Decision Tree	0.44061665	0.35656119	0.999966	0.999966	0.21527778
Random Forest	0.12708066	0.20444112	0.999990	0.999990	0.13638889
XGBoost	1.61211638	0.82601512	0.999874	0.999876	0.59910414
LightGBM	0.63746704	0.49550309	0.999950	0.999950	0.35196378
Bagging	0.15492950	0.25422535	0.999988	0.999988	0.17001916
Gradient Boosting	0.63374832	0.58294769	0.999950	0.999951	0.41736599
Hybrid Gradient Boosting	0.26392758	0.37786287	0.999979	0.999980	0.29996663
Voting Regressor	0.09634592	0.22077085	0.999992	0.999992	0.16034094
Bagging Regressor	0.15735799	0.21830462	0.999988	0.999988	0.14444444
Stacking Regressor	1.33e-26	8.93e-14	1.000000	1.000000	7.10e-14
AdaBoost with DecisionTree Regressor	0.18714347	0.22901596	0.999985	0.999985	0.15972222
AdaBoost with LinearRegression Regressor	1.72e-27	3.03e-14	1.000000	1.000000	2.84e-14

V. DISCUSSION, CONCLUSION AND FUTURE WORK

The results of the comparative analysis are discussed in detail. The strengths and weaknesses of each method are highlighted, and the overall performance of the models is evaluated. The rationale for proposing the Hybrid Gradient Boosting method is explained, including how it integrates the strengths of the different methods analyzed. Its potential advantages over the other methods are discussed. This study demonstrates the effectiveness of various machine learning approaches for predicting solar irradiance. The proposed Hybrid Gradient Boosting method shows promise for improving prediction accuracy. XGBoost achieved the highest performance among the individual models, but the Hybrid Gradient Boosting model significantly outperformed traditional Gradient Boosting, showcasing enhanced predictive capabilities.

The performance of each model was evaluated using several metrics, as shown in Table 1. XGBoost demonstrated the highest performance with a Mean Squared Error (MSE) of 1.612116, Mean Absolute Error (MAE) of 0.826015, and R^2 of 0.999874. Additionally, the proposed Hybrid Gradient Boosting model outperformed the traditional Gradient Boosting approach, indicating enhanced generalization capabilities.

Performance metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), R^2 (Coefficient of Determination), Explained Variance Score (EVS), Median Absolute Error (MedAE), and Mean Absolute Percentage Error (MAPE) were used to evaluate the models. The data was divided into training and testing sets with an 80%-20% split. The results indicate the efficacy of the proposed hybrid approach through graphical representations and histograms.

Suggestions for future research include exploring other advanced machine learning techniques, incorporating additional data sources, and applying the methods to other geographical locations. By broadening the scope of data and enhancing model complexities, the prediction accuracy for solar irradiance can be further improved, making these methods more robust and versatile across different environmental conditions. The study underscores the importance of continuous innovation and adaptation in machine learning methodologies to meet the growing energy needs driven by global population growth and the increasing demand for renewable energy sources.

REFERENCES

[1] C. B. Oguz, E. Avci, and S. B. Ozturk, "Analysis of PV power plant performance considering combination of different MPPT algorithms, shading patterns and connection types," *Eng. Sci. Technol. an Int. J.*, vol. 48, p. 101559, Dec. 2023, doi: 10.1016/j.jestch.2023.101559.

[2] Ö. Ayvazoğluyüksel and Ü. B. Filik, "Estimation methods of global solar radiation, cell temperature and solar power forecasting: A review and case study in Eskişehir," *Renew. Sustain. Energy Rev.*, vol. 91, pp. 639–653, Aug. 2018, doi: 10.1016/j.rser.2018.03.084.

[3] A. P. Grantham, P. J. Pudney, L. A. Ward, M. Belusko, and J. W. Boland, "Generating synthetic five-minute solar irradiance values from hourly observations," *Sol. Energy*, vol. 147, pp. 209–221, May 2017, doi: 10.1016/j.solener.2017.03.026.

[4] H. Bouzgou and C. A. Gueymard, "Fast short-term global solar irradiance forecasting with wrapper mutual information," *Renew. Energy*, vol. 133, pp. 1055–1065, Apr. 2019, doi: 10.1016/j.renene.2018.10.096.

[5] M. Ajith and M. Martínez-Ramón, "Deep learning algorithms for very short term solar irradiance forecasting: A survey," *Renew. Sustain. Energy Rev.*, vol. 182, p. 113362, Aug. 2023, doi: 10.1016/j.rser.2023.113362.

[6] B. K. Puah *et al.*, "A regression unsupervised incremental learning algorithm for solar irradiance prediction," *Renew. Energy*, vol. 164, pp. 908–925, Feb. 2021, doi: 10.1016/j.renene.2020.09.080.

[7] M. Golam, R. Akter, J.-M. Lee, and D.-S. Kim, "A Long Short-Term Memory-Based Solar Irradiance Prediction Scheme Using Meteorological Data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2021.3107139.

[8] J. Ramirez-Vergara, L. B. Bosman, W. D. Leon-Salas, and E. Wollega, "Ambient temperature and solar irradiance forecasting prediction horizon sensitivity analysis," *Mach. Learn. with Appl.*, vol. 6, p. 100128, Dec. 2021, doi: 10.1016/j.mlwa.2021.100128.

[9] H. Wen, Y. Du, X. Chen, E. G. Lim, H. Wen, and K. Yan, "A regional solar forecasting approach using generative adversarial networks with solar irradiance maps," *Renew. Energy*, vol. 216, p. 119043, Nov. 2023, doi: 10.1016/j.renene.2023.119043.

[10] X. Hou, C. Ju, and B. Wang, "Prediction of solar irradiance using convolutional neural network and attention mechanism-based long short-term memory network based on similar day analysis and an attention mechanism," *Heliyon*, vol. 9, no. 11, p. e21484, Nov. 2023, doi: 10.1016/j.heliyon.2023.e21484.

[11] E. Elibol and O. Dikmen, "Long-term performance investigation of different solar panels in the West Black Sea Region," *Clean Technol. Environ. Policy*, vol. 26, no. 3, pp. 875–899, Mar. 2024, doi: 10.1007/s10098-023-02658-1.

[12] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Inf. Process. Manag.*, vol. 50, no. 1, pp. 104–112, Jan. 2014, doi: 10.1016/j.ipm.2013.08.006.

[13] P. Ferreira, D. C. Le, and N. Zincir-Heywood, "Exploring Feature Normalization and Temporal Information for Machine Learning Based Insider Threat Detection," in *2019 15th International Conference on Network and Service Management (CNSM)*, IEEE, Oct. 2019, pp. 1–7. doi: 10.23919/CNSM46954.2019.9012708.

[14] L. Han *et al.*, "Modeling maize above-ground biomass based on machine learning approaches using UAV remote-sensing data," *Plant Methods*, vol. 15, no. 1, p. 10, Dec. 2019, doi: 10.1186/s13007-019-0394-z.

[15] V. Joshua, S. M. Priyadharson, and R. Kannadasan, "Exploration of Machine Learning Approaches for Paddy Yield Prediction in Eastern Part of Tamilnadu," *Agronomy*, vol. 11, no. 10, p. 2068, Oct. 2021, doi: 10.3390/agronomy11102068.

[16] J. D. de Guia, R. S. Concepcion II, H. A. Calinao, R. R. Tobias, E. P. Dadios, and A. A. Bandala, "Solar Irradiance Prediction Based on Weather Patterns Using Bagging-Based Ensemble Learners with Principal Component Analysis," in *2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC)*, IEEE, Dec. 2020, pp. 1–6. doi: 10.1109/R10-HTC49770.2020.9356988.