IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Solar Energy Prediction And Influential Feature Analysis Through Explainable AI

Archa Gopan, Aneesh S Perumprath

PG Scholar, Assistant Professor
Department of ECE (VLSI and Embedded System)
Musaliar College Of Engineering and Technology, Pathanamthitta

Abstract: The increasing demand for clean and sustainable energy sources has underscored the importance of solar energy in modern power systems. Accurate prediction of solar energy output is essential for efficient energy management. This study proposes the design and development of a machine learning (ML) based predictive model for solar energy forecasting. This system also analysed by Explainable Artificial Intelligence (XAI) techniques to enhance model transparency and trustworthiness. A Random Forest Regressor and XGboost algorithm were selected to capture complex, non-linear relationships between environmental factors and solar power output. The model was trained on real-world datasets .To address the interpretability in ML models, SHapley Additive exPlanations (SHAP) were employed to analyze and visualize the contribution of individual input features to the model's predictions. This enabled the identification of key influencing factors such as angle of incidence, wind direction, solar positioning, temperature, humidity, wind speed, and cloud cover, providing critical insights into the system's decision-making process. The integration of XAI not only improves the credibility of the forecasting model but also supports informed decision-making for stakeholders. This work demonstrates that the combination of robust machine learning algorithms with explainability frameworks offers a powerful tool for solar energy prediction that is both accurate and interpretable.

Index Terms **- Component, formatting, style, styling, insert.**

Introduction

Renewable energy is the power, we harness from natural sources that are constantly replenished by the Earth's own processes. These include sunlight, wind, flowing water, geothermal heat, and organic materials like plants and waste. What sets renewable energy apart from conventional fossil fuels is that it doesn't run out and it doesn't pollute the environment when used. As the world faces growing environmental challenges and a pressing need to reduce carbon emissions, renewable energy has become not just a promising alternative, but a necessary part of our future [1].

Figure 1 : Harnessing Renewable Energy: A Sustainable Future in Our Hands

The basic idea behind renewable energy is simple: we use technology to convert natural elements into usable forms of energy. For example, solar panels capture sunlight and turn it into electricity, while wind turbines use the motion of the wind to generate power. Hydroelectric dams use the force of moving water, and geothermal systems tap into heat stored deep underground. These technologies are already helping to power homes, schools, businesses, and entire cities[2].

There are many reasons why renewable energy is gaining importance across the globe. One major benefit is its role in fighting climate change. Unlike coal or oil, renewable sources release little to no greenhouse gases. They also make countries less dependent on imported fuel, improve energy security, and help bring electricity to remote or rural communities that might otherwise have no access. On top of that, the growth of the renewable energy sector creates millions of jobs worldwide [3].

Renewable energy isn't without its challenges. Some sources depend heavily on weather conditions, which means they don’t always produce energy consistently. Installing renewable systems can also be costly upfront, and some projects, such as large wind farms or dams, can raise environmental or social concerns. Despite these hurdles, advances in energy storage, smart grids, and efficient design are helping to overcome many of these limitations. With ongoing innovation and global cooperation, renewable energy is playing a key role in building a cleaner, healthier, and more sustainable future for everyone [4].

Figure 2 : Solar and Wind Energy Working Together for a Greener Tomorrow

Solar Energy

Solar energy is one of the most important and widely used forms of renewable energy today. It is the energy we receive from the sun in the form of light and heat. Every day, the Earth receives a massive amount of solar radiation more than what the entire world consumes in energy. By using modern technology like solar panels and solar thermal systems, we can convert this sunlight into electricity and heat, making it a clean, sustainable, and environmentally friendly energy source [5].
As concerns about climate change and the depletion of fossil fuels grow, solar energy is being recognized globally as a key part of the solution. It is free, abundant, and available in most parts of the world. With the help of solar technology, we can now power homes, schools, offices, and even entire cities without polluting the environment or relying on non-renewable resources [6].

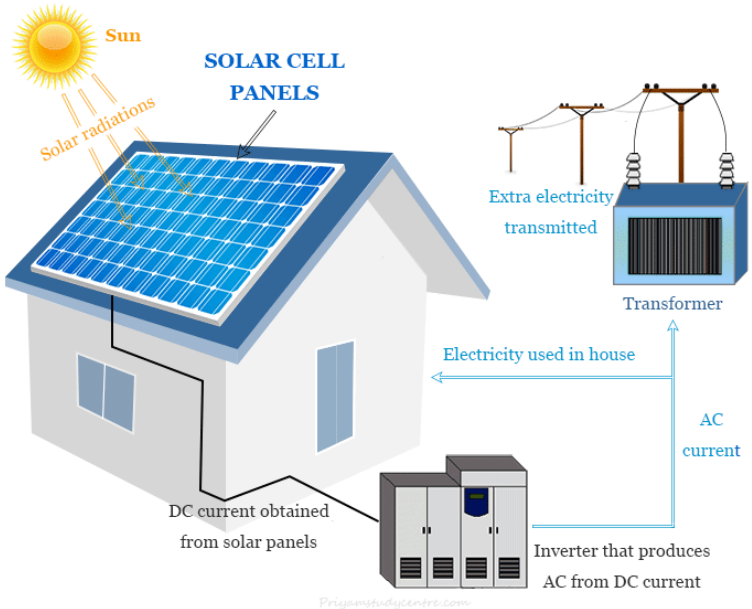

Figure 3 : Implementation of the Solar panel system

Solar energy can be captured and used in two main ways[7]:

Photovoltaic (PV) Systems: These use solar panels made of semiconductor materials (usually silicon) to convert sunlight directly into electricity. This electricity can be used immediately, stored in batteries, or fed into the power grid.

Solar Thermal Systems: These systems use mirrors or collectors to capture the sun's heat. This heat is used to warm water or air for homes and industries, or even to produce steam that drives turbines to generate electricity.

Modern solar panels are efficient and reliable. They are often installed on rooftops, solar farms, and open lands, and can last 20–30 years with minimal maintenance.

Advantages of Solar Energy [8]

Clean and Green: Solar energy doesn't produce harmful emissions, making it one of the cleanest sources of power.

Renewable and Inexhaustible: The sun shines every day and will continue to do so for billions of years. Thus Solar energy truly sustainable.

Reduces Electricity Bills: Once installed, solar systems can significantly reduce electricity costs for homes and businesses.

Low Maintenance: Solar panels require very little upkeep. Occasional cleaning and basic checks are usually enough.

Energy Independence: By generating power locally, communities and countries can reduce their dependence on imported fossil fuels.

I. Related Works

The accurate forecasting of solar energy is critical for the efficient planning and management of renewable energy systems. Recent research has focused on leveraging machine learning (ML), and ensemble methods, to improve the precision of solar energy predictions.

Mishra et al. [10] proposed an ensemble learning framework combining artificial neural networks (ANN), support vector machines (SVM), and random forests (RF) for global solar radiation forecasting. Their framework demonstrated better accuracy over individual models. Similarly, Rahimi et al. [11] provided a comprehensive review of ensemble algorithms. This includes bagging, boosting, and stacking. This highlights their effectiveness and practical applicability in solar power forecasting.\\

Ghimire et al. [12] introduced a CNN-LSTM-MLP hybrid model for optimizing feature selection . Predicting daily solar radiation shows improved accuracy and robustness. Khan et al. [13] also explored deep ensemble

stacking to forecast photovoltaic (PV) energy. This is achieved by combining the strengths of various deep learning models to manage the complex non-linearities in PV systems.

II.PROPOSED METHOD

In this work, a machine learning model is proposed to accurately predict solar energy output. This is developed from real-world solar and weather datasets. The model leverages historical solar power generation data and various environmental features Explainable AI (XAI) techniques are employed to analyse the features influenced in the ML model. This will enhance the transparency and trustworthiness of the model. These techniques help analyze and visualize the contribution of individual features.

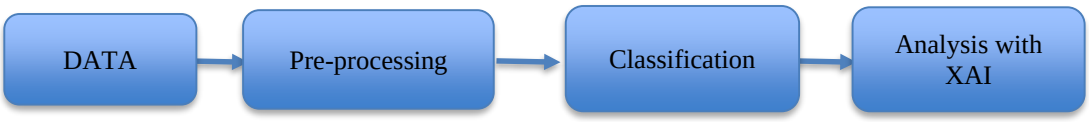

Figure 4 : Block diagram of the proposed system

A DATASET

Two types of datasets were used. The First dataset provides a detailed record of solar energy generation over time, along with various environmental conditions. It contains over 116,000 entries, capturing data at hourly or near-hourly intervals. The primary focus of this dataset is to help analyse and forecast solar power production, using both time-based patterns and weather-related information. Each record includes a timestamp indicating when the data was collected, as well as the corresponding solar power output, measured in megawatts (MW). These two columns are fully populated, making them reliable for performing time-series analysis. The solar output values vary significantly, ranging from around 1,253 MW to over 21,000 MW, reflecting the natural variability in solar energy production across different times of day and weather conditions.

The second dataset contains a rich collection of information that links weather conditions to solar power generation. With over 4,200 records and 21 variables, it offers a detailed snapshot of how various environmental factors influence the performance of a solar energy system. This makes it especially valuable for studies related to renewable energy forecasting and climate-aware power management. At its core, the dataset records how much solar power was generated at specific times, measured in kilowatts. The power output ranges from 0 kW (likely during night-time or very cloudy conditions) to just over 3,000 kW during peak sunlight. On average, the system produced around 1,134 kW, reflecting a wide variation depending on the time of day and weather. The dataset includes several meteorological features. Temperature is recorded at 2 meters above the ground and ranges from chilly sub-zero values to hot summer-like conditions. Humidity levels span the full spectrum, from very dry (7\%) to completely saturated air (100\%). The pressure data, close to the standard atmospheric level, gives insight into weather stability. While most entries report little to no rainfall or snowfall, there are occasional instances of measurable precipitation, adding depth to the dataset. Cloud cover is another critical aspect. The data captures cloud presence at low, medium, and high altitudes, as well as a total cloud cover score. Many records show clear skies, especially at higher cloud levels, which is ideal for solar generation. This kind of cloud information is key to understanding how much sunlight actually reaches the solar panels.

B PRE-PROCESSING

(i) Outlier Detection Using Z-Score

When working with data, it's common to come across values that don't quite fit in with the rest. These unusual values, known as outliers, can have a big impact on analysis results and even throw off machine learning models. One of the simplest ways to spot these outliers is by using something called a Z-score, or standard score. This score tells us how far a particular value is from the average (mean) of the dataset, in terms of standard deviations. The formula for the Z-score is: $Z = (x - \mu) / \sigma$, where x is the data point, μ is the mean, and σ is the standard deviation.

In a typical bell-shaped (normal) distribution, most values tend to cluster around the average. Around 68\% of the values fall within one standard deviation from the mean, 95\% fall within two, and nearly all—99.7\%—fall within three standard deviations. So, if we come across a value with a Z-score above 3 or below -3, it's considered an outlier because it lies far outside the usual range. Of course, depending on the situation, we might use a slightly lower or higher threshold.

The Z-score approach is popular because it's straightforward and quick to use, especially when working with data that's fairly normally distributed. But it's not perfect. If the data has a lot of extreme values or isn't shaped like a bell curve, the Z-score might not work as well. Also, a single extreme value can affect the mean and standard deviation so much that the method loses accuracy.

(ii) Min-Max Scaling (Normalization)}

Min-max scaling, also known as normalization, is a widely used technique in data preprocessing that transforms numerical features to a common scale, typically ranging from 0 to 1. This is particularly important when working with machine learning algorithms that are sensitive to the magnitude of data values, such as K-Nearest Neighbors and Support Vector Machines. The main idea behind min-max scaling is to rescale each feature based on its minimum and maximum values, using the formula:

$x' = (x - x_{min}) / (x_{max} - x_{min})$

where x is the original value, x_{min} and x_{max} are the minimum and maximum values of the feature, and x' is the scaled value. This transformation ensures that all values are proportionally mapped to a fixed range, preserving their relative differences while eliminating scale differences between features.

Min-max scaling is especially beneficial when features have different units or scales, as it ensures that no single feature dominates the learning process due to its magnitude. For example, in a dataset containing the values [50, 20, 30, 80, 100], applying min-max scaling would transform the value 50 to 0.375, using the minimum of 20 and maximum of 100. The technique is simple, intuitive, and effective, particularly when the data is not normally distributed.

(iii) Handling Missing Values

Missing values are a common issue in real-world datasets and can arise for various reasons, such as data entry errors, equipment malfunction, or intentional non-responses in surveys. Handling missing data effectively is crucial, as improper treatment can lead to biased results, reduce model accuracy, and affect the overall reliability of the analysis.

There are several techniques for handling missing values, and the choice of method depends on the nature of the data and the extent of the missingness. The most straightforward approach is deletion, where rows (or sometimes columns) with missing values are removed. This method is effective when the proportion of missing data is small and the data is missing completely at random (MCAR). However, it can lead to significant data loss and reduced statistical power if missing values are widespread.

Another common strategy is imputation, where missing values are replaced with estimated ones. Simple imputation methods include replacing missing values with the mean, median, or mode of the available data. While easy to implement, these methods may distort the data distribution, especially when the data is not normally distributed.In time-series data, forward fill and backward fill methods are often used, which propagate the last known or next known value to fill in gaps. Ultimately, the strategy for handling missing values should be guided by the data's structure, the pattern of missingness (MCAR, MAR, or MNAR), and the goals of the analysis. Careful treatment ensures that the resulting dataset is as complete and unbiased as possible, enabling more accurate and meaningful modeling.

C CLASSIFICATION}

Random Forest is a machine learning method that helps computers make better decisions by combining the results of many smaller models, called decision trees. Instead of using just one tree to make a prediction, it uses a group (or "forest") of trees to get a more accurate answer.Each decision tree in the forest is trained using a random sample of the data. It's like asking the same question to many people who each saw different pieces of the full picture. Also, when each tree is being built, it doesn't look at all the available information—it only uses a few random parts. This helps make each tree slightly different, which improves the overall result when they're combined.

When it's time to make a prediction, each tree gives its answer. If it's a yes-or-no question (like "Is this spam?"), the most common answer among the trees is chosen. If the answer is a number (like predicting the price of a house), all the answers are averaged to get the final result. Random Forest has many good sides. It works well even when there are a lot of features or noisy data, and it's less likely to make big mistakes compared to just one tree. It can also deal with missing data better than some other methods. However, it can be slower because it needs to build many trees, and it's harder to understand exactly how it came up with a result. It also might not be the best option for text data unless it's carefully prepared.

XGBoost stands for Extreme Gradient Boosting. It's a powerful and fast machine learning method that's often used to win competitions and build strong prediction models, especially when accuracy is very important.At its core, XGBoost builds a series of decision trees, one after another. Unlike Random Forest, where each tree works independently, in XGBoost, each new tree learns from the mistakes of the previous trees. This process is called boosting—it boosts the model's performance step by step.

The model starts with a simple prediction. It then looks at what it got wrong and builds a new tree to correct those errors. This keeps going, with each tree trying to fix the problems left by the ones before it. In the end, all the trees work together to make a better and more accurate prediction.

Explainable AI (XAI) is a way to help people understand how and why artificial intelligence systems make decisions. As AI models become more powerful and complex. They also become harder to interpret. This is why XAI is important: it opens up these "black box" systems so that humans can see what's going on inside. One of the main reasons we need explainable AI is trust. When we can understand why a model made a certain prediction or decision, we're more likely to trust it. This is especially important in areas like healthcare, finance, or criminal justice, where decisions can have serious consequences. XAI also helps when AI systems make mistakes. By explaining what went wrong, it allows developers and data scientists to fix problems and improve the model.

Another big reason for XAI is fairness. Sometimes, AI systems learn biases from the data they are trained on, which can lead to unfair treatment of certain groups. With explainable AI, we can check whether a model is making biased decisions based on things like gender, race, or age. In addition, there are legal rules in some countries (like the European Union's GDPR) that require AI decisions to be explainable to users.

There are two general types of AI models when it comes to explainability. Some models, like decision trees or linear regression, are naturally easy to understand. Others, like neural networks or ensemble methods, are more complex and need special tools to explain how they work.

SHAP (SHapley Additive exPlanations)

SHAP is a tool that helps us understand how a machine learning model makes decisions. When a model gives a result—like guessing the price of a house or whether someone will pay back a loan—SHAP explains which input features (like income, age, or location) affected that result and by how much. The idea comes from a fair sharing method used in teamwork. Imagine a group of people working together and earning a reward. SHAP works out how much each person (or in our case, each feature) contributed to the success. In machine learning, the "team" is made up of the features, and the "reward" is the model's prediction. SHAP tells us how much each feature helped push the result up or down.

SHAP helps turn complicated AI decisions into something people can understand. It's a helpful tool when we need to explain, trust, or check the fairness of AI models.

III.RESULTS AND DISCUSSIONS

This project focuses on predicting solar energy production using a dataset that includes various features related to solar radiation and other environmental factors. The prediction model utilizes machine learning techniques, specifically the Random Forest Regressor, and integrates Explainable AI (XAI) with SHAP values to interpret the model's decisions.

The dataset was loaded from a CSV file, with the following steps performed to clean and prepare the data for modelling:

Removing Unnecessary Columns: The 'Datetime' column was removed as it was deemed irrelevant for prediction purposes.

Target and Features Separation: The target variable, solar_mw (solar energy output in megawatts), was isolated from the rest of the dataset, which comprises the input features (X).

Handling Categorical Data: Columns containing string values, such as 'am', were identified. These non-numeric values were replaced with numeric equivalents (e.g., 'am' was replaced with 0), and the data was converted to a float type for compatibility with the machine learning model.

Outlier detection using the Z-score method revealed a total of 194 records in the dataset that exhibit values significantly different from the norm. The Z-score technique identifies any data point that lies beyond three standard deviations from the mean in any of the numerical columns. These outliers could correspond to unusual or extreme weather conditions, potential sensor errors, or unexpected solar power generation behavior. For example, one record showed unusually high atmospheric pressure and a completely clear sky, yet recorded very low power generation — possibly indicating early morning or late evening hours. Other instances included high-pressure days with minimal cloud cover and unexpected spikes or drops in energy output. Such data points, while rare, are important to consider. Depending on the objectives of the analysis, they may either be excluded to improve model accuracy or examined further to gain insights into rare but impactful scenarios in solar energy production.

The dataset was divided into training and testing sets using the train_test_split method from sklearn.model_selection. This step ensures that the model is evaluated on data it has not seen during training, allowing for an unbiased performance assessment. The training set comprised 80% of the data, while the testing set accounted for the remaining 20%. Random Forest Regressor and XGBoost algorithms were chosen for the prediction task due to its ability to model complex relationships and provide robust performance with minimal tuning. The model was initialized with 100 trees and a fixed random state (42) to ensure reproducibility of results. The model was trained using the training data (X_train and y_train) by calling the fit() method.

5. Model Evaluation

After training, the model was evaluated on the test set (X_test, y_test) using the following performance metrics:

Mean Squared Error (MSE): Measures the average squared difference between the predicted and actual values.

Root Mean Squared Error (RMSE): The square root of MSE, providing a more interpretable error metric in the original units of the target variable.

Mean Absolute Error (MAE): The average absolute difference between predicted and actual values, offering a more direct sense of prediction error.

R-squared (R^2): Indicates the proportion of variance in the target variable explained by the model.

These metrics offer valuable insights into the accuracy and generalization ability of the trained model.

TABLE I : Performance Measures on DATASET 1 without Pre-processing

Mean Squared Error (MSE)	5721710.11
Root Mean Squared Error (RMSE)	2392.01
Mean Absolute Error (MAE)	1895.81
R-squared (R2)	0.02

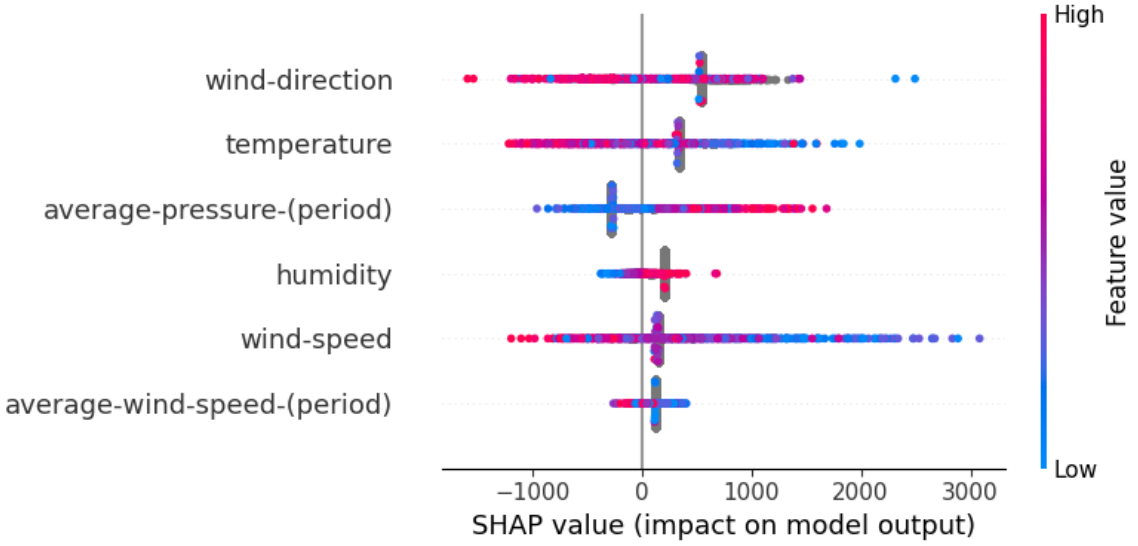

Figure 7: Explainable AI Analysis on DATASET 1 without pre-processing

TABLE II : Performance Measures on DATASET 2 without Pre-processing

Mean Squared Error (MSE)	166252.44
Root Mean Squared Error (RMSE)	407.74
Mean Absolute Error (MAE)	255.60
R-squared (R2)	0.81

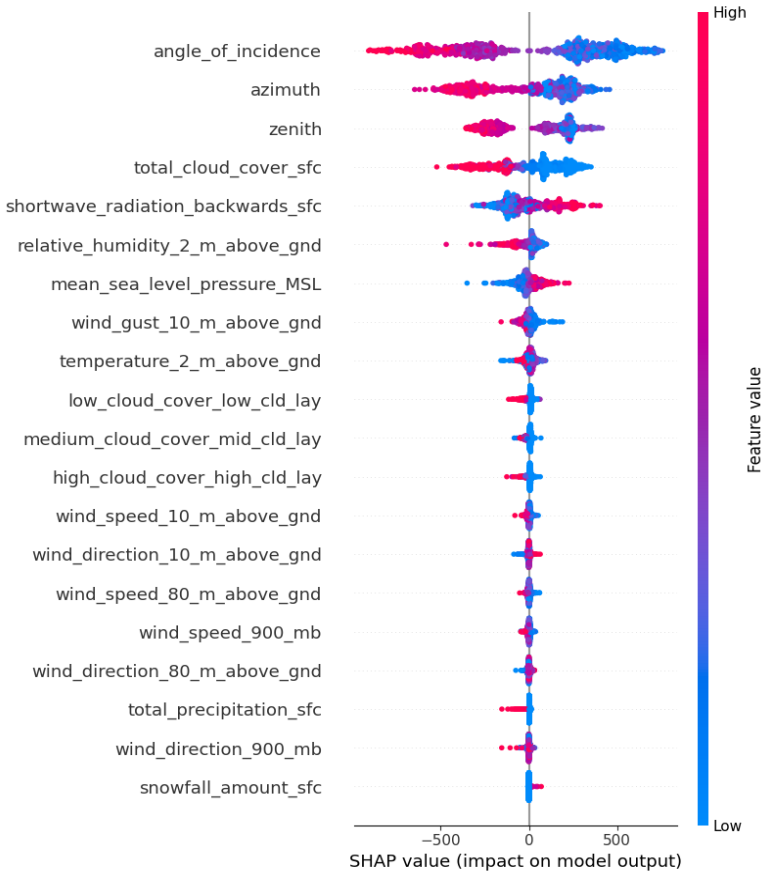

Figure 8: Explainable AI Analysis on DATASET 2 without pre-processing

TABLE III : Performance Measures on DATASET 1 with Pre-processing and Random Forest Algorithm

Mean Squared Error (MSE)	0.02
Root Mean Squared Error (RMSE)	0.12
Mean Absolute Error (MAE)	0.10
R-squared (R2)	0.69

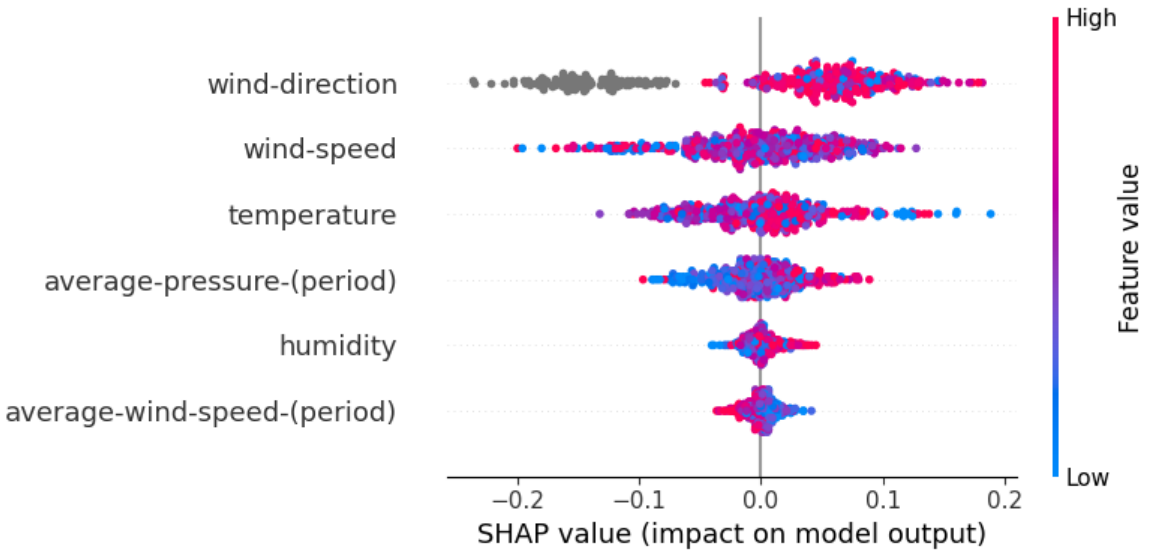

Figure 9: Explainable AI Analysis on DATASET 1 with Pre-processing and Random Forest Algorithm

TABLE IV : Performance Measures on DATASET 1 with Pre-processing and XGBoost Algorithm

Mean Squared Error (MSE)	0.05
Root Mean Squared Error (RMSE)	0.21
Mean Absolute Error (MAE)	0.18
R-squared (R2)	0.09

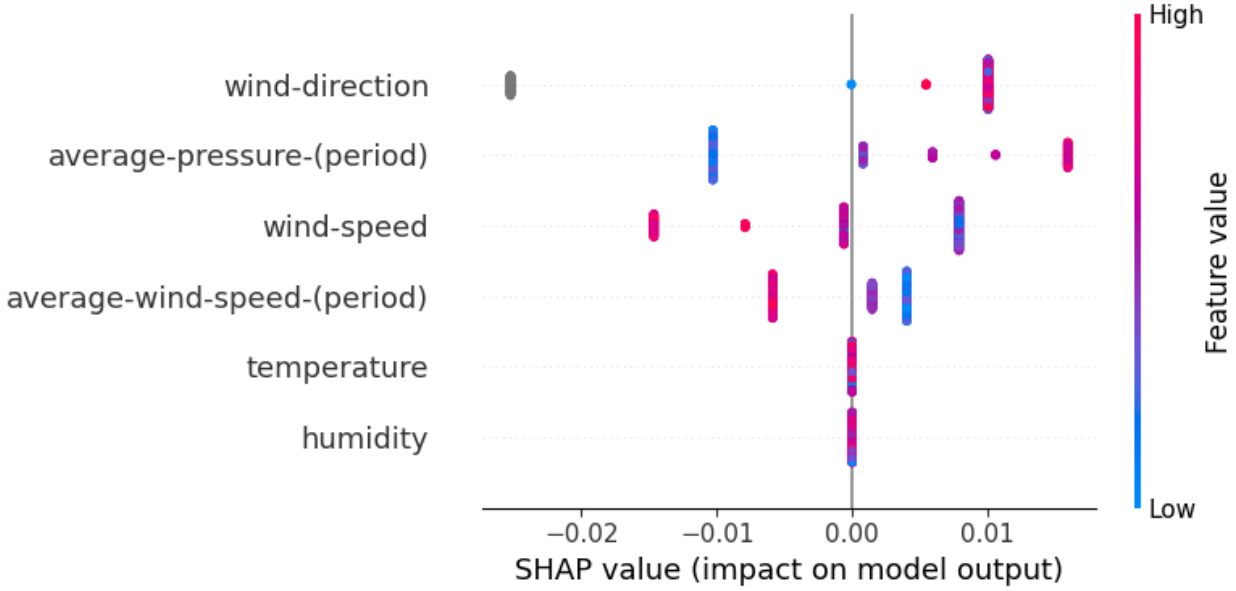

Figure 10: Explainable AI Analysis on DATASET 1 with Pre-processing and XGBoost Algorithm

TABLE V : Performance Measures on DATASET 2 with Pre-processing and Random Forest Algorithm

Mean Squared Error (MSE)	0.01
Root Mean Squared Error (RMSE)	0.11
Mean Absolute Error (MAE)	0.07
R-squared (R2)	0.86

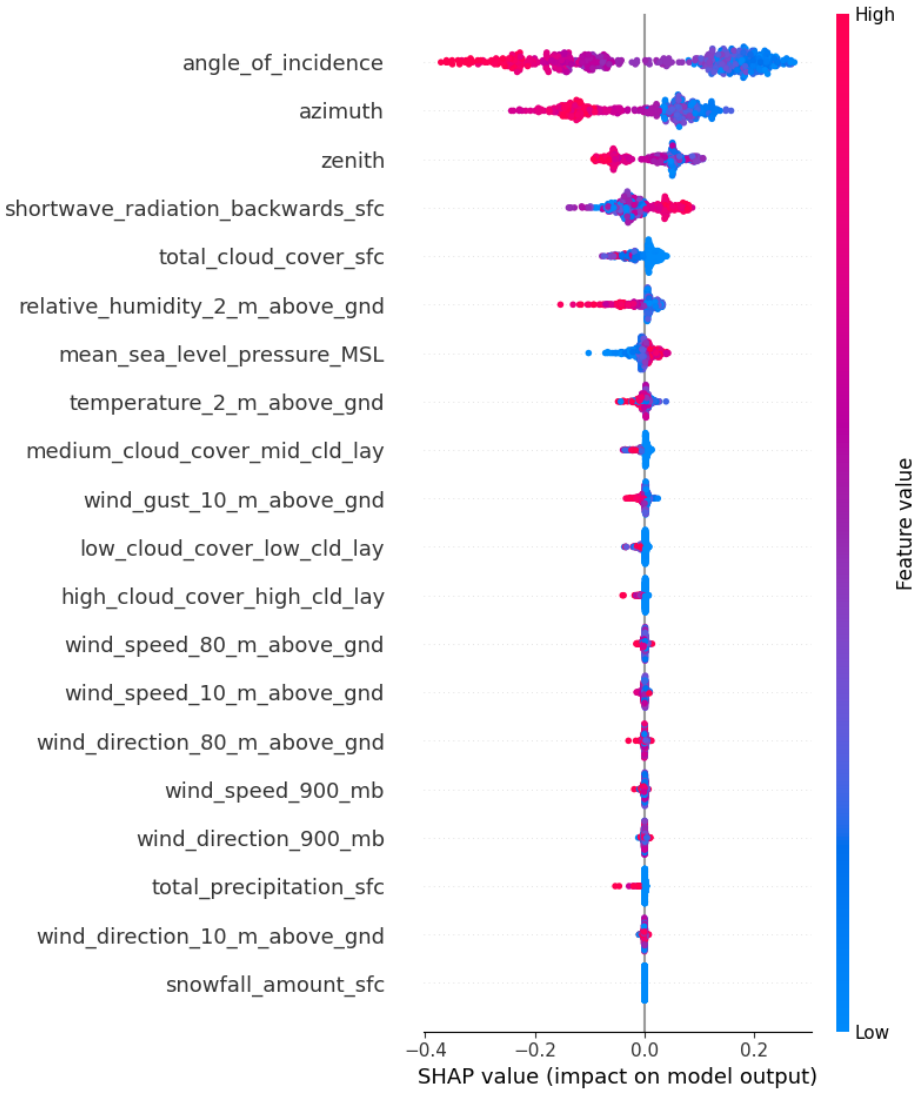

Figure 11: Explainable AI Analysis on DATASET 2 with Pre-processing and Random Forest Algorithm

TABLE VI : Performance Measures on DATASET 2 with Pre-processing and XGBOOST Algorithm

Mean Squared Error (MSE)	0.05
Root Mean Squared Error (RMSE)	0.212
Mean Absolute Error (MAE)	0.19
R-squared (R2)	0.52

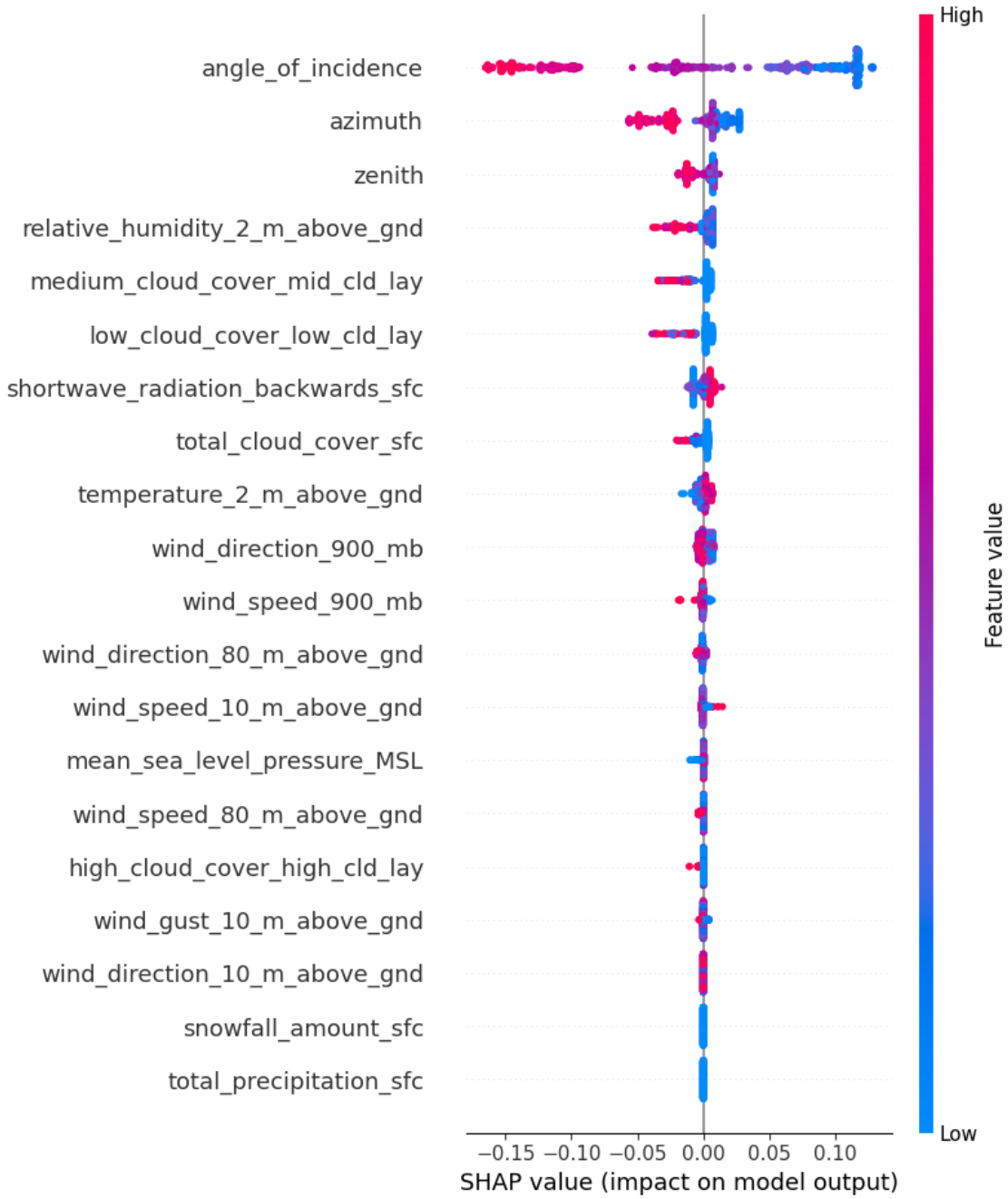

Figure 12 : Explainable AI Analysis on DATASET 2 with Pre-processing and XGBoost Algorithm

To enhance the interpretability of the model, SHAP (SHapley Additive exPlanations) values were used. SHAP values quantify the contribution of each feature to a specific prediction, helping to understand the importance and impact of each input variable. The SHAP explainer was initialized with the trained Random Forest model, and SHAP values for the test set were computed. This step provides a deeper understanding of how individual features influence the model's predictions.

The shape of the SHAP values was printed to verify the output structure, with each feature’s contribution to each prediction clearly identified.

IV. CONCLUSION AND FUTURESCOPE

Solar energy is not just an alternative; it is an essential part of the global energy solution. With its countless benefits and rapidly advancing technology, it offers hope for a future powered by clean and renewable sources. Though there are challenges to overcome, the long-term benefits of solar energy make it one of the most promising options for individuals, communities, and nations alike. Solar energy prediction is a critical part of renewable energy management. With the help of advanced data analytics, AI models, and satellite monitoring, it's now possible to forecast solar energy with increasing accuracy. This not only enhances energy efficiency but also supports the global transition to a cleaner and more sustainable power system.

This project demonstrates the use of machine learning for solar energy prediction. The Random Forest Regressor and XGboost were trained and evaluated with strong performance across all metrics, and SHAP values provided crucial insights into how the model makes predictions. The combination of accurate forecasting and interpretability ensures the model is not only effective but also transparent. This makes it suitable for applications where understanding the reasoning behind predictions is essential. Wind direction and angle of incidence are important features in the solar energy prediction.

The future of solar energy is bright—both literally and figuratively. Technological advances are making solar panels more efficient, affordable, and versatile. New innovations like solar tiles, transparent solar panels, and floating solar farms are already in development. Governments and organizations around the world are also offering subsidies, tax benefits, and incentives to encourage the use of solar energy.

As the world moves toward a cleaner and more sustainable future, solar energy will continue to play a vital role in reducing carbon emissions, protecting the environment, and improving access to electricity worldwide. . Future work can explore further model optimization and additional feature engineering to improve performance.

REFERENCES

[1] Dincer I. Renewable energy and sustainable development: a crucial review. Renewable and sustainable energy reviews. 2000 Jun 1;4(2):157-75.

[2] Twidell J. Renewable energy resources. Routledge; 2021 Nov 29.

[3] Bull SR. Renewable energy today and tomorrow. Proceedings of the IEEE. 2001 Aug;89(8):1216-26.

[4] Stram BN. Key challenges to expanding renewable energy. Energy Policy. 2016 Sep 1;96:728-34.

[5] Kannan N, Vakeesan D. Solar energy for future world:-A review. Renewable and sustainable energy reviews. 2016 Sep 1;62:1092-105.

[6] Granqvist CG. Solar energy materials. Advanced Materials. 2003 Nov 4;15(21):1789-803.

[7] Meinel AB, Meinel MP. Applied solar energy: an introduction. NASA STI/Recon Technical Report A. 1977;77:33445.

[8] Lakatos L, Hevessy G, Kovács JJ. Advantages and disadvantages of solar energy and wind-power utilization. World Futures. 2011 Aug 1;67(6):395-408.

[9] Maradin D. Advantages and disadvantages of renewable energy sources utilization. International Journal of Energy Economics and Policy. 2021;11(3):176-83.

[10] Blanco J, Malato S, Fernández-Ibañez P, Alarcón D, Gernjak W, Maldonado MI. Review of feasible solar energy applications to water processes. Renewable and Sustainable Energy Reviews. 2009 Aug 1;13(6-7):1437-45.

[11] Khatib T, Mohamed A, Sopian K. A review of solar energy modeling techniques. Renewable and Sustainable Energy Reviews. 2012 Jun 1;16(5):2864-9.

[12] Blaga R, Sabadus A, Stefu N, Dughir C, Paulescu M, Badescu V. A current perspective on the accuracy of incoming solar energy forecasting. Progress in energy and combustion science. 2019 Jan 1;70:119-44.

[13] Suehrcke H, McCormick PG. A performance prediction method for solar energy systems. Solar energy. 1992 Jan 1;48(3):169-75.

[14]. Mishra, D. P., Jena, S., Senapati, R., Panigrahi, A., & Salkuti, S. R. (2023). Global solar radiation forecast using an ensemble learning approach. International Journal of Power Electronics and Drive Systems, 14(1), 496.

[15]. Ahm NA. Machine Learning Approches for Evaluating the Properties of Materials. Journal of Computational Intelligence in Materials Science. 2023 Jun 10;1:067-76. [3]. You, L., & Zhu, M. (2023). Digital Twin simulation for deep learning framework for predicting solar energy market load in Trade-By-Trade data. Solar Energy, 250, 388-397.

[16]. Krishnan, N., Kumar, K. R., & Inda, C. S. (2023). How solar radiation forecasting impacts the utilization of solar energy: A critical review. Journal of Cleaner Production, 135860.