Solar Radiation Prediction Using Machine Learning and Feature Engineering: A Comparative Analysis with AutoML Optimization

Ayush Uday Bhole[1], Prashant Kulkarni[2], and Shubhangi Tidake[3]
[1]*M. Sc. (Data Science) Student, Symbiosis Skill & Professional University, Pune*
[2]*Assistant Professor, Symbiosis Skill & Professional University, Pune*
[3]*Assistant Professor, Symbiosis Skill & Professional University, Pune*

Abstract - Accurate prediction of solar radiation is essential for optimizing renewable energy generation. This study utilizes machine learning models such as XGBoost, Random Forest, LightGBM, Neural Networks and stacked ensemble model to predict solar radiation. These models are trained and evaluated using historical weather data, including meteorological characteristics such as temperature, humidity, wind speed, and surface pressure.

To improve prediction accuracy, the dataset is enhanced with manually engineered features, including rolling averages, lag values, and additional time-based variables such as day and month, thereby enabling the models to more effectively capture complex temporal patterns in solar radiation dynamics.

Performance metrics such as R^2 (coefficient of determination), MAE (mean absolute error), MSE (mean squared error), and RMSE (root mean square error) are used to evaluate the effectiveness of the model. The goal of this research is to improve the accuracy of solar radiation prediction and contribute to the advancement of renewable energy systems and sustainable energy planning.

Keywords - Cross-Validation, Solar radiation, Machine Learning, Neural Network, Automl.

I. INTRODUCTION

In the transition to renewable energy, accurate solar radiation prediction is vital for efficient solar power system design and operation. Solar radiation is affected by meteorological factors like temperature, humidity, pressure, and wind speed, making its prediction complex. Traditional methods often lack precision and adaptability. This complexity has led to the use of machine learning to improve prediction accuracy.

This study investigates the application of machine learning for predicting solar radiation using real-world meteorological data, including temperature, humidity, pressure, wind speed, and time-based solar variables. It involves data preprocessing, feature engineering, and comparison of models such as Random Forest, XGBoost, LightGBM, Neural Networks, and ensemble stacking. AutoML tools like H2O.ai were used for hyperparameter tuning and feature selection. Models were evaluated using 5-fold cross-validation and metrics including R^2, MAE, MSE, and RMSE. The study also identifies key weather factors influencing solar radiation and demonstrates the value of machine learning and AutoML in enhancing renewable energy prediction.

II. LITERATURE REVIEW

Jordy Anchundia Troncoso (2023) – Solar Radiation Prediction in the UTEQ Based on Machine Learning Models

Findings/Remarks: This study compares machine learning models like Gradient Boosting, Random Forest, and Decision Trees for solar radiation prediction at the UTEQ campus. Gradient Boosting and Random Forest performed best ($R^2 = 0.76$ and 0.74), and the study also introduced a real-time forecasting tool, demonstrating practical use. [1]

Ü Ağbulut (2021) – Prediction of Daily Global Solar Radiation Using Different Machine Learning Algorithms

Findings/Remarks: This comparative study evaluated several algorithms including ANN, SVM, and k-NN. Results showed that Artificial Neural Networks outperformed the others in terms of RMSE and R^2, emphasizing the importance of selecting models based on dataset characteristics and performance metrics.[2]

Zhihong Pang (2020) – Solar radiation prediction using recurrent neural network and artificial neural network: A case study with comparisons

Findings/Remarks: The study explored RNN and ANN models for predicting solar radiation in the context of building energy control. Although RNNs delivered higher prediction accuracy, they demanded greater computational power compared to other models. The trade-off between model complexity and efficiency is emphasized.[3]

Cyril Voyant (2017) – Machine Learning Methods for Solar Radiation Forecasting: A Review

Findings/Remarks: This paper provides a broad review of machine learning models including ANN, SVM, Random Forest, regression trees, and ARIMA. It highlights the relevance of accurate solar radiation forecasting for grid integration, energy planning, and storage system design.[6]

III. RESEARCH AND METHODOLOGY

A. About the Dataset

The dataset used in this study was obtained from the NASA POWER (Prediction of Worldwide Energy Resources) database, which provides global meteorological and solar radiation data specifically tailored for renewable energy applications. The NASA POWER platform offers high-resolution, satellite-based hourly data on various environmental parameters, making it a reliable source for solar energy prediction research. For this project, data was collected for the years 2023-2024, covering hourly measurements.

B. Data Description

ALL_SKY _SFC_ SW_ DWN: This variable represents the total solar radiation incident on a horizontal surface under all-sky conditions. Measured in watts per square meter (W/m²), it is a critical parameter for solar energy applications as it directly influences the amount of solar power available for generation.

Temperature: The temperature of the surrounding air close to the Earth's surface, recorded in degrees Celsius (°C). It affects the effectiveness of photovoltaic systems and is generally used to assess thermal impacts on solar radiation and outfit performance.

Relative_Humidity: It is defined as the percentage ratio of the current water vapor present in the air to the maximum amount it can retain at a particular temperature. Increased humidity levels can significantly decrease atmospheric transparency, which in turn affects the level of solar radiation reaching the ground.

Hour: Indicates the specific time of day, helping the model recognize fluctuations in solar radiation caused by the natural progression of sunlight from morning to evening.

Surface_Pressure: The pressure exerted by the atmosphere at ground level, typically measured in kilopascals (kPa). It influences air density and is closely linked to weather patterns such as storms or clear skies, which directly impact the availability of solar radiation.

Wind_Speed: The velocity of air movement near the Earth's surface, measured in meters per second (m/s). Wind speed affects the cooling of solar panels, which can influence their efficiency, and plays a key role in understanding overall weather patterns.

Engineered Variables for Solar Radiation Analysis

prev_hour_radiation: The solar radiation intensity observed in the previous hour, which can indicate trends in radiation levels over time.

Rolling_radiation_3: A rolling average or sum of solar radiation intensity over the past three hours, providing a smoothed indication of radiation patterns.

Month: Numeric representation of the month, influencing solar radiation patterns due to seasonal changes in sunlight exposure.

Wind Power: Likely representing the wind energy potential or related metric based on wind speed, influencing renewable energy considerations.

C. Data Visualization

To gain meaningful initial insights and uncover hidden patterns within the dataset, a variety of data visualization techniques were employed. These visual tools played a crucial role in exploring the temporal, seasonal, and statistical characteristics of the data, helping to better understand the distribution, variability, and relationships among key meteorological features and solar radiation.

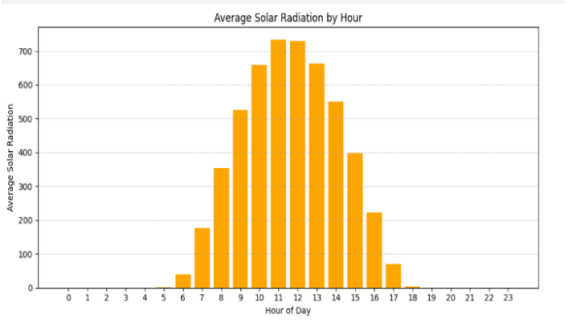

Fig 1 : Average Solar Radiation By Hour

The bar chart shows the average solar radiation at each hour of the day, highlighting the daily solar cycle. Radiation is near zero at night, increases after sunrise, peaks around noon (11 AM–1 PM), and declines in the evening. This pattern reflects typical daylight variation and helps identify the most effective hours for solar energy harvesting.

D. Feature Engineering and Initial Variable Selection

A two-stage approach was used for feature selection in this study. First, manual feature engineering was applied to extract time-based and lag features (e.g., month, prev_hour_radiation) that capture temporal and periodic patterns in solar radiation. Then, automated feature selection using AutoML tools was employed to evaluate and rank features based on performance metrics such as R^2, MAE, MSE, and RMSE. This hybrid method helped reduce noise, eliminate redundant variables, and retain only the most relevant features, ultimately enhancing both model accuracy and interpretability.

E. Feature Selection Using H2O AutoML

To optimize the prediction of solar radiation, an automated machine learning (AutoML) approach using H2O was employed, focusing on feature selection and subsequent model building.

Feature selection was conducted using H2O AutoML, which automatically identified the most relevant predictors for solar radiation forecasting. After converting the dataset into an H2O Frame, AutoML evaluated multiple models based on metrics such as RMSE, MSE, and MAE.

Top-performing models included Stacked Ensembles (RMSE: 21.80–22.09, MAE: 9.95–10.21) and Gradient Boosting Machines (GBMs) (RMSE: 22.31–23.79, MAE: 10.14–11.61). Notably, the Stacked Ensemble and GBM Model 2 achieved the best error metrics, highlighting their superior ability to model complex solar radiation patterns.

F. Machine Learning Algorithms

This study applied advanced machine learning algorithms to predict solar radiation using selected meteorological features. Models like XGBoost, Random Forest, LightGBM, and a multi-layer Neural Network with the Adam optimizer were evaluated. A Stacking Regressor with Linear Regression as the meta-learner was also implemented to enhance generalization. The dataset was split using train_test_split, and model performance was measured using R^2, MAE, MSE, and RMSE. H2O's AutoML was employed for feature selection and tuning. The Neural Network and Stacking Ensemble achieved the best accuracy, effectively capturing the temporal and environmental patterns of solar radiation.

G. Algorithmic Configuration and Optimization Procedures

In this research, various machine learning algorithms were applied to develop robust predictive models for solar radiation. Hyperparameter tuning was conducted using Grid Search and Randomized Search to identify the best-performing configuration for each model. The parameters for each algorithm are described as follows.

XGBoost Regressor : XGBoost was tuned using Grid and Randomized Search to optimize parameters like estimators, depth, and learning rate, improving its ability to capture non-linear patterns in solar radiation data.

Random Forest Regressor : For the Random Forest algorithm, a parameter search was conducted over key hyperparameters including the number of estimators, maximum depth, and minimum samples per leaf. Randomized Search enabled efficient sampling of the parameter space to improve generalization performance.

LightGBM Regressor : The LightGBM model was optimized using Randomized Search due to its wide and flexible hyperparameter space. Parameters such as the number of leaves, learning rate, and feature fraction were evaluated to enhance model speed and accuracy while avoiding overfitting.

Neural Network : A feedforward Neural Network with three hidden layers (128, 64, 32 neurons) was trained using standardized inputs, the Adam optimizer, and MSE loss over 100 epochs (batch size 32). This setup effectively captured nonlinear patterns, yielding accurate solar radiation predictions. This configuration enabled the network to effectively learn nonlinear dependencies in the data, resulting in highly accurate solar radiation predictions.

Stacking Regressor : This ensemble model integrated XGBoost, Random Forest, LightGBM, Neural Network as base learners, with Linear Regression used as the meta-learner. Hyperparameters for each base model were tuned individually prior to integration. The stacking model was validated using cross-validation to ensure stable performance across folds.

H. Metrics Used for Model Evaluation

To assess the performance of the machine learning models for solar radiation prediction, the following evaluation metrics were utilized:

Mean Absolute Error (MAE) : It serves as a key indicator in this research for assessing model's accuracy. MAE calculates how far predictions are from actual values on average, using absolute differences and treating each error with equal importance to provide a straightforward accuracy measure. Its formula is:

$$MAE = \frac{1}{n}\sum_{i=0}^{n}|yi - \hat{y}i|$$

Mean Squared Error (MSE): This research uses Mean Squared Error (MSE) to evaluate model performance by averaging the squared differences between predicted and actual values. MSE gives more weight to larger errors, making it useful for assessing how well models handle significant deviations. Lower MSE indicates better predictive accuracy.

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(yi - \hat{y}i)^2$$

Root Mean Squared Error (RMSE) It serves as a crucial metric for evaluating regression models, representing prediction error in the same unit as the target variable. By emphasizing larger errors, it helps assess the accuracy and reliability of model predictions. It is calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)^2}$$

Coefficient of Determination (R^2 Score) : This study employs the coefficient of determination (R^2) to assess model performance. It indicates the proportion of variation in the target variable that is explained by the model's predictors, and is given as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi - \hat{y}i)^2}{\sum_{i=1}^{n}(yi - \bar{y}i)^2}$$

Values approaching 1 indicate stronger predictive capability, signifying that the model accounts for most data variability. In this research, higher R^2 values reflect more effective explanation of variance by the proposed model.

I. Validating Model Predictions Using Cross-Validation

Cross-validation was employed in this study to evaluate the generalization performance of machine learning models for solar radiation prediction. A 5-fold cross-validation approach was used, where the training data was split into five equal parts. In each iteration, four folds were used for training and one for validation, ensuring that each subset was tested once. This process helped mitigate overfitting and provided a more reliable estimate of model accuracy. The performance outcomes from all folds were combined to evaluate the overall stability and reliability of the models. Based on this analysis, the stacked ensemble model and neural network consistently outperformed other models in terms of predictive accuracy and error minimization.

By using this method, the study was able to identify models that performed well across different data partitions, ensuring dependable predictions for real-world solar energy applications.

J. Result

A comparative analysis was conducted to evaluate the predictive performance of various machine learning models for solar radiation prediction. Each model was assessed using key regression metrics R^2 (coefficient of determination), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Squared Error (MSE) calculated via 5-fold cross-validation. The average results across folds provided insight into each model's generalization capability and predictive accuracy. Here are the key findings for each model.

Models	Mean R^2	Mean MSE	Mean MAE	Mean RMSE
XGBoost	0.988	722.59	13.27	26.04
Random Forest	0.981	1165.64	20.00	33.77
LightGBM	0.966	2303.17	36.99	47.77
Neural Network	0.988	698.85	13.52	25.87
Stacked Model	0.993	495.41	11.45	22.25

Table 1: Summary of 5-Fold Cross-Validation Performance Metrics

XGBoost :
XGBoost demonstrated strong performance with a mean R^2 score of 0.988, indicating its ability to explain 98.8% of the variance in solar radiation data. It achieved a mean MSE of 722.59, mean MAE of 13.27, and mean RMSE of 26.04, highlighting its effectiveness in accurate solar radiation prediction.

Random Forest :
Random Forest achieved strong results with a mean R^2 of 0.981, mean MSE of 1165.64, mean MAE of 20.00, and mean RMSE of 33.77, indicating its effectiveness in modelling solar radiation patterns.

LightGBM :
LightGBM achieved a mean R^2 score of 0.966, a mean MSE of 2303.17, a mean MAE of 36.99, and a mean RMSE of 47.77. Although slightly lower than XGBoost and Random Forest, LightGBM still showed competitive performance in solar radiation prediction.

Neural Network :
The Neural Network model exhibited strong predictive performance with a mean R^2 score of 0.988, mean MSE of 698.85, mean MAE of 13.52, and mean RMSE of 25.87.Its ability to learn complex nonlinear relationships made it effective in modeling solar radiation patterns, yielding highly accurate forecasts across cross-validation folds.

Stacked Model :
The Stacked Model performed exceptionally well with a mean R^2 score of 0.993, mean MSE of 495.41, mean MAE of 11.45, and mean RMSE of 22.25. This model, combining predictions from multiple models, demonstrated superior accuracy and robustness in solar radiation forecasting.

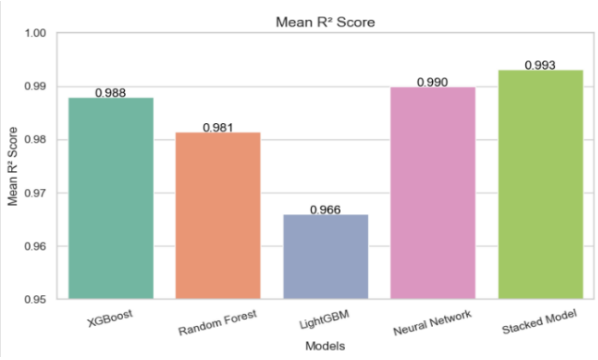

Fig 2: 5-Fold Cross-Validated R^2 Scores

This bar chart compares the average R^2 scores achieved by XGBoost, Random Forest, LightGBM, Neural Network, and the Stacked Model. The Stacked Model outperformed all others, achieving the highest R^2 score of 0.993, indicating superior generalization performance.

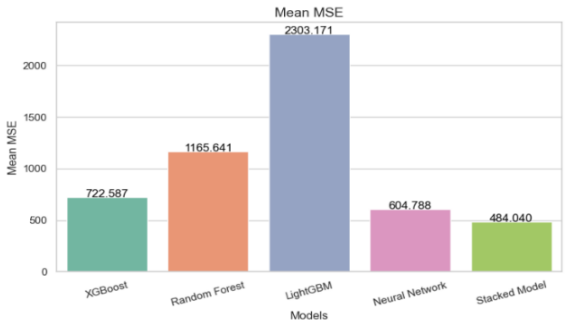

Fig 3: 5-Fold Cross-Validated Mean Squared Error (MSE) Scores

This bar chart illustrates the average MSE for each model, reflecting the squared error between actual and predicted values. The Stacked Model had the lowest MSE, showing strong performance, while LightGBM had the highest, suggesting room for improvement.

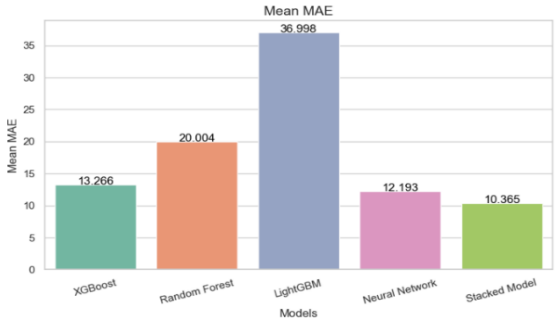

Fig 4: 5-Fold Cross-Validated Mean Absolute Error (MAE) Scores

This bar chart illustrates the average MAE across different models. A lower MAE indicates better predictive accuracy, with the Stacked Model achieving the lowest error, followed by Neural Network and XGBoost.

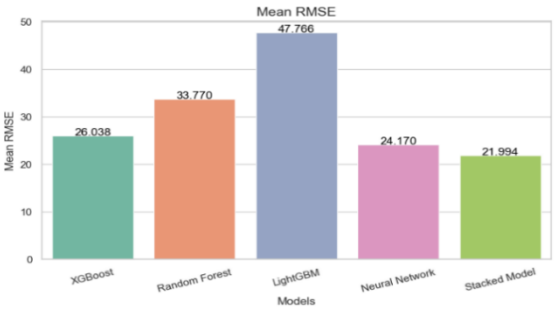

Fig 5: 5-Fold Cross-Validated Root Mean Squared Error (RMSE) Score

This chart compares the RMSE values obtained from each model. RMSE places greater weight on larger errors, making it valuable for assessing a model's effectiveness in minimizing major prediction inaccuracies. The Stacked Model recorded the lowest RMSE, indicating its superior precision, followed by Neural Network and XGBoost.

K. Overfitting Check

To assess overfitting, R^2 scores were calculated for both training and testing sets across all models Random Forest, XGBoost, LightGBM, Neural Network, and Stacked Ensemble. The R^2 gap, defined as the difference between training and testing scores, was used to evaluate model generalization. A threshold of 0.05 was set to flag overfitting. All models exhibited minimal R^2 gaps (ranging from 0.0054 to 0.0133), indicating that they generalized well to unseen data and did not overfit the training set. This confirms the model's robustness and reliability in accurately predicting solar radiation without memorizing patterns from the training data.

IV. CONCLUSION

This study focused on accurately predicting solar radiation using machine learning models such as XGBoost, Random Forest, LightGBM, Neural Networks, and a Stacked Ensemble. By integrating manual feature engineering with automated feature selection through H2O AutoML, the models successfully captured both meteorological and temporal patterns relevant to solar radiation. Key features included time-based variables, lag values, and rolling averages, which significantly enhanced the model's ability to detect short-term fluctuations and periodic trends.

Performance evaluation using R^2, MSE, MAE, and RMSE revealed that the Stacked Ensemble model outperformed all other models, achieving the highest accuracy ($R^2 = 0.9932$, MAE = 10.35, RMSE = 21.91). The Neural Network model also demonstrated strong performance ($R^2 = 0.9883$, MAE = 13.52, RMSE = 25.87), making it a competitive alternative. In contrast, models like LightGBM and Random Forest showed lower predictive accuracy in this context. The results show that ensemble and deep learning models improve the reliability of solar radiation prediction, supporting more effective solar energy planning and contributing to sustainable power generation.

V. FUTURE WORK

For future work, this approach can be extended by incorporating additional weather variables such as the cloud index, the solar zenith angle, or the optical depth of the aerosol to further improve accuracy. Advanced deep learning models like LSTM or hybrid CNN-LSTM can be explored to better capture temporal dependencies. Expanding the dataset across multiple regions and time periods will help assess model generalizability. Furthermore, deploying these models into real-time solar energy management systems and experimenting with other AutoML tools such as TPOT or Google. Future work could investigate other AutoML platforms or ensemble-based tuning methods to further enhance prediction accuracy and operational deployment in renewable energy systems.

REFERENCE

[1] Jordy Anchundia Troncoso, Ángel Torres Quijije, Byron Oviedo, Cristian Zambrano-Vega,- "Solar Radiation Prediction in the UTEQ Based on Machine Learning Models".

[2] Ü. Ağbulut, A. E. Gürel, and Y. Biçen, Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. Sustainable Energy Technologies and Assessments,

[3] Zhihong Pang, Fuxin Niu, Zheng O'Neill, "Solar radiation prediction using recurrent neural network and artificial neural network: A case study with comparisons," Solar Energy, vol. 193, pp. 781–794, 2019.

[4] H. Hissou, S. Benkirane, A. Guezzaz, M. Azrour, and A. Beni-Hssane, "A novel machine learning approach for solar radiation estimation," Sustainability, vol. 15, no. 13, art. no. 10609, Jul. 2023,

[5] Irfan Khan Tanoli, Asqar Mehdi, Abeer D. Algarni, Azra Fazal, Talha Ahmed Khan, Sadique Ahmad, Abdelhamied A. Ateya , "Machine learning for high-performance solar radiation prediction," Energy Reports, vol. 10, pp. 4794–4804, Dec. 2024

[6] Cyril Voyant, Gilles Notton, Soteris Kalogirou, Marie Laure Nivet, Christophe Paoli, Fabrice Motte, Alexis Fouilloy, "Machine learning methods for solar radiation forecasting: A review," Renewable Energy, vol. 105, pp. 569–582, 2017.

[7] Amit Kumar Yadav, S.S. Chandel, “Solar radiation prediction using Artificial Neural Network techniques: A review" Renewable and Sustainable Energy Reviews,2014
[8] K. Bakirci, “Models of solar radiation with hours of bright sunshine: A review”Renewable and Sustainable Energy Reviews