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

M.Sc. Data Science Thesis

By

Ayush Uday Bhole

SCHOOL OF DATA SCIENCE
SYMBIOSIS SKILLS AND PROFESSIONAL UNIVERSITY PUNE

August-2025

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

A THESIS

Submitted in partial fulfilment of the
requirements for the award of the degree
of
Master of Data Science

by

Ayush Uday Bhole

SCHOOL OF DATA SCIENCE
SYMBIOSIS SKILLS AND PROFESSIONAL UNIVERSITY PUNE
August-2025

SYMBIOSIS SKILLS AND PROFESSIONAL UNIVERSITY PUNE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **Solar Radiation Prediction Using Machine Learning Models with AutoML Optimization** in the partial fulfilment of the requirements for the award of the degree of **MASTER OF DATA SCIENCE** and submitted in the **SCHOOL OF DATA SCIENCE, SYMBIOSIS SKILLS AND PROFESSIONAL UNIVERSITY PUNE**, is an authentic record of my own work carried out during the time period from 01/02/2025 to 01/07/2025 under the supervision of Prof. Prashant Kulkarni School Of Data Science.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

Ayush

Ayush Uday Bhole

This is to certify that the above statement made by the candidate is correct to the best of my/our knowledge.

Prof. Prashant Kulkarni
Assistant Professor,
School of Data Science, SSPU, Pune

Director, School of Data Science,
SSPU, Pune

Ayush Uday Bhole has successfully given his M.Sc. Oral Examination held on **Date :**

ACKNOWLEDGEMENT

I am deeply grateful for the invaluable support and guidance I have received throughout my journey of completing my master's thesis at Symbiosis Skills and Professional University. This accomplishment would not have been possible without the constant encouragement, motivation, and expert mentorship I received along the way.

First and foremost, I would like to express my sincere gratitude to my thesis guide, Mr. Prashant Kulkarni, for his unwavering support, insightful feedback, and consistent encouragement. His technical expertise and dedication have played a pivotal role in shaping my research work. Mr. Kulkarni's clarity of thought, constructive criticism, and timely suggestions pushed me to explore deeper perspectives and refine the quality of this research. I am truly fortunate to have had the opportunity to work under his guidance.

I would also like to extend my heartfelt thanks to my co-guide, Ms. Shubhangi Tidake, whose valuable insights and academic support greatly enriched my research. Her encouragement, patience, and thoughtful feedback have contributed immensely to this work. Her mentorship helped broaden my understanding and approach toward complex concepts, and her support made this journey smoother.

I would also like to acknowledge the faculty members, staff, and my peers at the School of Data Science for their assistance, shared knowledge, and collaborative spirit during the course of my postgraduate studies.

Lastly, I am profoundly thankful to my family, friends, and loved ones for their unwavering support, patience, and motivation. Their belief in me gave me the strength and confidence to persevere through every challenge.

This thesis is not just the outcome of my effort alone but a reflection of the collective contributions and encouragement I received from many people. Thank you all for being a part of this significant milestone in my academic journey.

Abstract

Data science has grown rapidly in recent years, enabling innovative applications across diverse fields, including renewable energy. One critical area is the accurate prediction of solar radiation, which plays a key role in optimizing the performance and planning of solar power systems. Precise solar radiation estimates are essential for improving energy generation efficiency, grid integration, and overall sustainability.

This study explores the application of advanced machine learning techniques for solar radiation prediction using historical meteorological data. Models such as XGBoost, Random Forest, LightGBM, Neural Networks, and a Stacked Ensemble are implemented and evaluated. The dataset includes essential weather variables such as temperature, humidity, wind speed, and surface pressure. To enhance model performance, manual feature engineering is applied, generating lag values, rolling averages, and time-based features like day, hour, and month. These additions help the models better capture short-term dependencies and periodic trends in solar radiation.

Model performance is assessed using standard regression evaluation metrics: R^2 (coefficient of determination), MAE (mean absolute error), MSE (mean squared error), and RMSE (root mean square error). The findings demonstrate that ensemble and deep learning methods, particularly when combined with effective feature engineering, significantly improve predictive accuracy. This research contributes to the growing intersection of data science and renewable energy, aiming to support smarter solar energy planning and more reliable integration of renewable resources into energy systems.

To further enhance model robustness, the study incorporates automated machine learning (AutoML) using the H2O.ai platform. This approach streamlines hyperparameter tuning and feature selection, ensuring optimal model configurations with minimal manual intervention. The integration of both manual and automated feature engineering techniques allows for a comprehensive understanding of variable importance, improving interpretability and reducing overfitting. Visual tools such as scatter plots and time series graphs are used to compare predicted versus actual values, validating the models' ability to accurately track variations in solar radiation over time.

LIST OF PUBLICATIONS

[1] Ayush Bhole, *"Solar Radiation Prediction Using Machine Learning: A Comparative Analysis with AutoMl Optimization"*, **International Journal of Innovative Research in Technology (IJIRT)**, Volume 12, Issue 2, July 2025.

Paper ID : 182060
Link : https://ijirt.org/Article?manuscript=182060

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 : Introduction

As the shift toward renewable energy continues, precise prediction of solar radiation plays a crucial role in the effective planning and functioning of solar energy systems. Meteorological factors such as temperature, humidity, pressure, and wind speed significantly influence solar radiation, making its prediction a complex task. Traditional methods often lack precision and adaptability. This complexity has led to the use of machine learning to improve prediction accuracy. This research explores the use of machine learning techniques to forecast solar radiation based on actual meteorological data, such as temperature, humidity, atmospheric pressure, wind speed, and time-related solar parameters. 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.

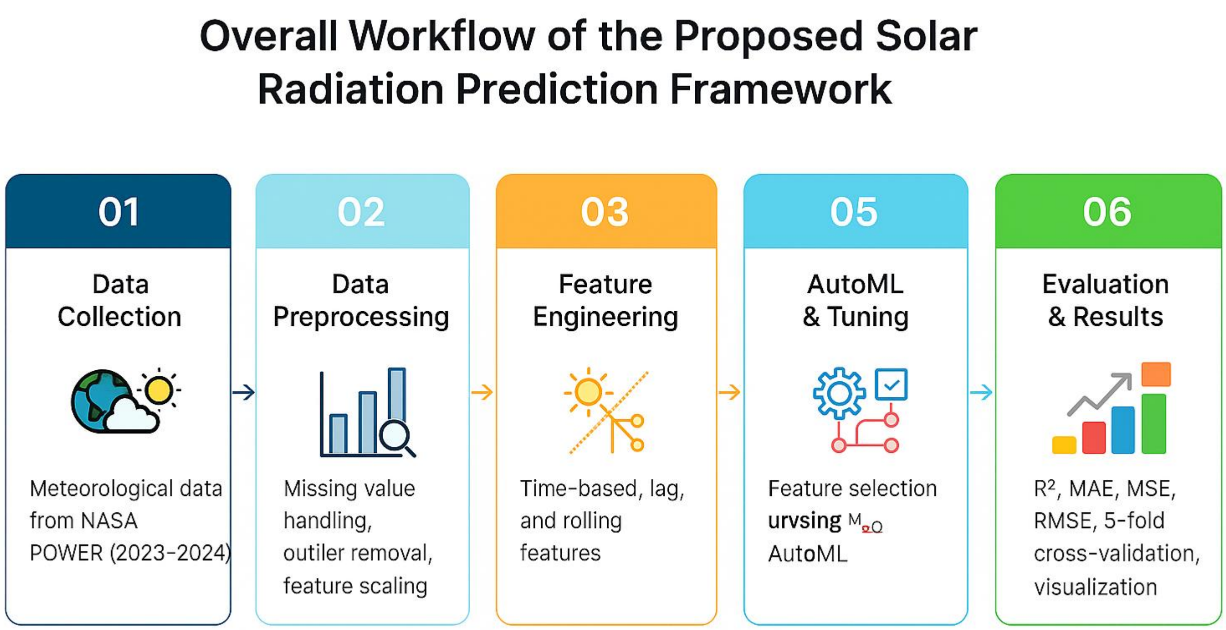

Fig 1 : Overall workflow of the proposed solar radiation prediction framework, outlining the key stages from data collection to model evaluation. This visual summary helps contextualize the structured approach adopted in the study.

1.1 Motivation

The increasing demand for sustainable energy solutions has made solar energy a crucial component in the global transition toward cleaner power generation. However, the intermittent and variable nature of solar radiation poses significant challenges for planning, efficiency, and integration into modern energy systems. This has fueled my motivation to develop a robust solar radiation prediction framework that supports the optimization of solar power utilization. As a data science enthusiast, I am driven by the objective of using intelligent systems to improve the accuracy of forecasting models that directly influence energy planning and grid stability.

This study is motivated by the need to leverage cutting-edge machine learning algorithms and feature engineering techniques to handle complex, real-world meteorological datasets. Conventional forecasting methods frequently struggle to accurately model the non-linear dynamics and time-based patterns present in solar radiation data. By integrating models such as XGBoost, Random Forest, LightGBM, Neural Networks, and ensemble stacking, the goal is to enhance predictive performance and ensure adaptability to seasonal and environmental variations.

Furthermore, the use of AutoML tools like H2O.ai in this research highlights the motivation to automate and optimize the model selection and tuning process. AutoML helps identify the most relevant features and model configurations, reducing manual bias and enhancing reproducibility. The inclusion of engineered features such as lag variables, rolling averages, and time-based attributes adds further depth to the modeling process, enabling the system to learn from both short-term trends and periodic patterns.

In an era where data availability is expanding rapidly and energy systems are growing increasingly complex, relying on traditional static models is no longer sufficient. My motivation stems from the desire to contribute to the evolving field of renewable energy forecasting by developing a machine learning-based framework that is both data-driven and scalable. The ultimate aim is to aid stakeholders whether researchers, energy planners, or engineers in making more informed, real-time decisions to maximize solar power potential and support sustainable energy development.

1.2 Purpose and Research Question

This study aims to explore and evaluate the effectiveness of advanced machine learning algorithms for predicting solar radiation using real-world meteorological data. With the rising importance of solar energy in sustainable power generation, accurate radiation prediction plays a critical role in optimizing solar power system design, improving grid reliability, and enhancing energy planning. The study leverages a combination of models including XGBoost, Random Forest, LightGBM, Neural Networks, and Stacked Ensembles, along with AutoML tools like H2O.ai, to automate model selection, feature importance evaluation, and hyperparameter tuning.

By incorporating time-based features, lag values, and rolling averages into the dataset, the study also investigates how engineered temporal features impact the learning and generalization ability of prediction models. The goal is to develop a scalable, data-driven solar radiation prediction framework that can generalize well across varying meteorological conditions and improve decision-making in renewable energy systems.

Main Research Question:

How can machine learning and AutoML techniques improve the accuracy of solar radiation prediction, and to what extent do feature engineering and ensemble methods enhance model performance under different meteorological conditions?

Sub-questions:

1. What are the limitations of traditional statistical approaches in solar radiation prediction, and how do machine learning models overcome these challenges?
2. How do different machine learning models (XGBoost, Random Forest, LightGBM, Neural Networks) compare in terms of predictive accuracy and computational efficiency?
3. What is the impact of manual feature engineering—such as lag variables, rolling means, and temporal decomposition—on model performance?
4. How does the use of AutoML (H2O.ai) enhance feature selection and model tuning compared to manual approaches?
5. Does the integration of ensemble learning methods (like Stacked Ensembles) significantly improve model robustness and accuracy over single-model approaches?
6. What are the challenges and trade-offs involved in implementing machine learning-based solar radiation prediction models in real-world energy systems?

1.3 Target Group

The intended audience for this study includes energy analysts, data scientists, researchers in renewable energy, and environmental engineers who are engaged in the fields of solar energy forecasting, smart energy systems, and sustainable energy planning. These professionals typically possess a background in data analysis, machine learning, atmospheric science, or meteorology, and are interested in applying advanced predictive models to enhance the accuracy and efficiency of solar power generation, resource management, and grid optimization.

This study is also highly relevant for government agencies, energy policy planners, utility companies, and solar energy developers seeking reliable and accurate solar radiation predictions to support strategic decision-making, energy resource allocation, infrastructure development, and real-time energy integration. These stakeholders rely on predictive analytics to improve the operational flexibility of renewable energy grids and meet regulatory and sustainability goals.

Additionally, academic researchers, postgraduate scholars, and students in the domains of data science, energy informatics, environmental science, and renewable energy engineering may benefit from the insights offered in this research. It serves as a valuable reference for those exploring the intersection of artificial intelligence, environmental modelling, and energy forecasting technologies, and contributes to the ongoing evolution of intelligent energy systems.

By providing a comparative analysis of various machine learning models and incorporating AutoML tools for automated model optimization and selection, this study addresses the specific needs of stakeholders aiming to improve the efficiency, adaptability, reliability, and scalability of solar forecasting systems. It also provides practical guidance for the deployment of data-driven solutions in real-world energy applications, making it relevant for both research and industrial implementation.

1.4 Outline

Chapter 2 presents a comprehensive literature review to understand the background and previous research related to solar radiation prediction using machine learning. Chapter 3 describes the methodology in detail, including the dataset, feature engineering techniques, model selection, and the proposed framework. It also compares existing models with the proposed stacked ensemble approach. Chapter 4 discusses the experimental results and outcomes obtained from the models and evaluates their performance using statistical metrics. Finally, Chapter 5 summarizes the key findings of the study and provides an outlook for future research directions and possible enhancements to the current work.

Chapter 2 : Literature Review

The following chapter reviews key studies that have contributed to the field of solar radiation prediction using machine learning. This review provides a foundation for understanding the strengths and limitations of current approaches and positions the present research in that context.

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

This study applies Gradient Boosting, Random Forest, and Decision Trees for predicting solar radiation over the UTEQ campus. With R^2 scores of 0.76 and 0.74 for Gradient Boosting and Random Forest respectively, the models proved effective for practical implementation. The paper also introduced a real-time forecasting tool, emphasizing the application of ML in operational energy systems. This aligns closely with the present research, which extends this work by exploring additional models such as LightGBM, Neural Networks, and ensemble stacking, while incorporating AutoML for automated optimization.

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

This comparative study evaluated ANN, SVM, and k-NN for daily solar radiation prediction. ANN outperformed the others in terms of R^2 and RMSE, highlighting the strength of neural networks in modeling non-linear relationships in meteorological data. While this study focuses on daily data, the current research advances this by using hourly data and exploring feature engineering and ensemble learning to enhance prediction granularity and model accuracy.

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

This review offers a broad comparison of machine learning and time-series models such as ANN, SVM, Random Forest, regression trees, and ARIMA. It stresses the importance of accurate forecasting for grid stability and energy planning. The review also recommends hybrid models and integration of multiple data sources. Inspired by this direction, the present study builds a hybrid framework that combines manual feature engineering with AutoML-based model selection and tuning, helping to balance interpretability and performance.

Zhihong Pang et al. (2020) – Solar Radiation Prediction Using Recurrent Neural Network and Artificial Neural Network: A Case Study with Comparisons

This study compares RNN and ANN for building energy systems and finds RNN slightly more accurate but computationally expensive. The trade-off between accuracy and efficiency is a core theme echoed in the current research. Instead of deep sequence models like RNN, the present work focuses on tree-based models and shallow neural networks, which are computationally efficient and suitable for real-time prediction when enhanced with AutoML techniques.

Chapter 3 : Methodology

In this chapter, we explore the background, recent developments, and related work surrounding solar radiation prediction using data-driven approaches. Solar energy has emerged as a central component of the global shift toward sustainable power generation. However, the irregular and location-dependent nature of solar radiation introduces challenges for power system design, energy demand planning, and integration into the grid. Traditionally, models such as empirical regression or physical-statistical hybrids have been used for solar prediction, but these approaches often lack adaptability and accuracy under dynamic weather conditions.

Recent advances in machine learning have enabled more accurate and adaptive solar radiation prediction models. Algorithms such as Random Forest, XGBoost, LightGBM, and Neural Networks can capture complex non-linear patterns in meteorological data, including features like temperature, humidity, wind speed, and atmospheric pressure. These models have outperformed traditional methods in terms of accuracy and generalization. Researchers have also emphasized the importance of feature engineering, which plays a crucial role in improving model performance by extracting temporal patterns from weather data. Lag variables, rolling averages, and time-based features (e.g., hour of the day or month of the year) help capture periodicity and short-term variations in solar radiation.

In recent years, ensemble learning methods such as XGBoost, Random Forest, and LightGBM have gained popularity due to their robustness and ability to handle noise and overfitting. Studies such as those by Pang et al. (2020) and Voyant et al. (2017) have shown that ensemble techniques outperform individual models in both short-term and long-term solar radiation prediction tasks.

Additionally, the use of AutoML platforms like H2O.ai and TPOT has enabled automatic feature selection, hyperparameter tuning, and model comparison. These tools minimize manual effort, reduce bias in model development, and improve reproducibility. For instance, Li et al. (2021) utilized H2O AutoML to evaluate multiple models and found that ensemble models with stacking architectures consistently performed best across seasonal data.

Although several studies have examined solar radiation prediction regionally, limited work has focused on combining manual feature engineering with AutoML-driven workflows. Moreover, research is still evolving in understanding model generalizability across various timeframes, seasons, and geographic locations.

This study extends prior research by proposing a hybrid framework that integrates engineered temporal features with multiple machine learning models and AutoML-based optimization. The objective is to build a reliable and scalable solar radiation prediction system adaptable to diverse meteorological conditions while maintaining high accuracy.

Let us now examine each component of this research methodology in greater detail.

3.1 About the Dataset

The dataset used in this study was obtained from the NASA POWER (Prediction of Worldwide Energy Resources) database, a reliable and widely used platform that offers global meteorological and solar radiation data designed to support applications in renewable energy, agriculture, and climate-related studies. The platform provides satellite-derived, high-resolution, and hourly datasets, making it highly suitable for research that requires detailed temporal granularity and spatial consistency.

For this study, data was collected for the years 2023–2024, focusing on hourly measurements relevant to solar radiation prediction. Key environmental parameters include solar irradiance, temperature, humidity, wind speed, and surface pressure factors known to significantly influence solar radiation. The selection of this dataset ensures that the model is trained on up-to-date and realistic weather conditions, enabling a more accurate representation of the current atmospheric dynamics. The comprehensiveness and accessibility of NASA POWER make it an ideal source for developing data-driven models aimed at improving the performance and reliability of solar energy systems.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. Higher humidity can notably reduce atmospheric clarity, thereby limiting the amount of solar radiation that reaches the Earth's surface.
Hour: Represents the time of day and allows the model to understand daily variations in solar radiation, which change naturally from sunrise to sunset.
Surface_Pressure: Surface pressure represents the force exerted by the atmosphere at ground level and is commonly measured in kilopascals (kPa). It affects air density and correlates with different weather conditions such as clear skies or storms that can directly influence solar radiation levels.
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.

3.2 Data Cleaning and Preprocessing

Effective data preprocessing is essential for ensuring high-quality input to machine learning models, particularly when working with environmental datasets containing real-world weather and solar radiation parameters. The raw dataset used in this study included key features such as solar radiation (ALL_SKY_SFC_SW_DWN), temperature, relative humidity, surface pressure, wind speed, hour, and derived variables like previous hour radiation, rolling averages, month, and wind power.

The cleaning process began with handling missing values, which are common in satellite-derived or sensor-based datasets. Missing or null entries in continuous variables like temperature, humidity, and radiation were addressed using linear interpolation or forward-fill methods to maintain temporal continuity. This was crucial for preserving the time series structure, especially for engineered features.

To ensure data consistency, columns with irregular formatting or spacing (e.g., ALL_SKY _SFC_ SW_ DWN) were renamed for clarity and ease of access during modeling. Additionally, duplicate rows and irrelevant columns, if any, were removed after verifying data uniqueness.

Outliers in features like wind_speed, temperature, and solar radiation were identified using statistical thresholds such as the interquartile range (IQR). Outliers were either removed or capped using winsorization to avoid skewing the training process.

Min-Max scaling was applied to numerical variables like temperature, pressure, humidity, and wind speed to standardize their values within the 0 to 1 range. This ensured compatibility with models sensitive to feature scales, especially neural networks.

Categorical features like month and hour were already in numerical format but were further validated for encoding consistency. These time-based features helped capture seasonal and diurnal patterns in solar radiation. The engineered features such as prev_hour_radiation and rolling_radiation_3 were created using lag-based transformations, helping the model recognize short-term trends and temporal dependencies in solar irradiance.

This comprehensive preprocessing pipeline ensured that the dataset was clean, consistent, and well-structured ready to be used effectively by both traditional and ensemble-based machine learning algorithms.

3.3 Data Visualization

Various data visualization methods were used to extract valuable insights and reveal underlying patterns in the dataset. 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. These visualizations provided a foundation for feature selection and model development by highlighting correlations, trends, and anomalies in the data.

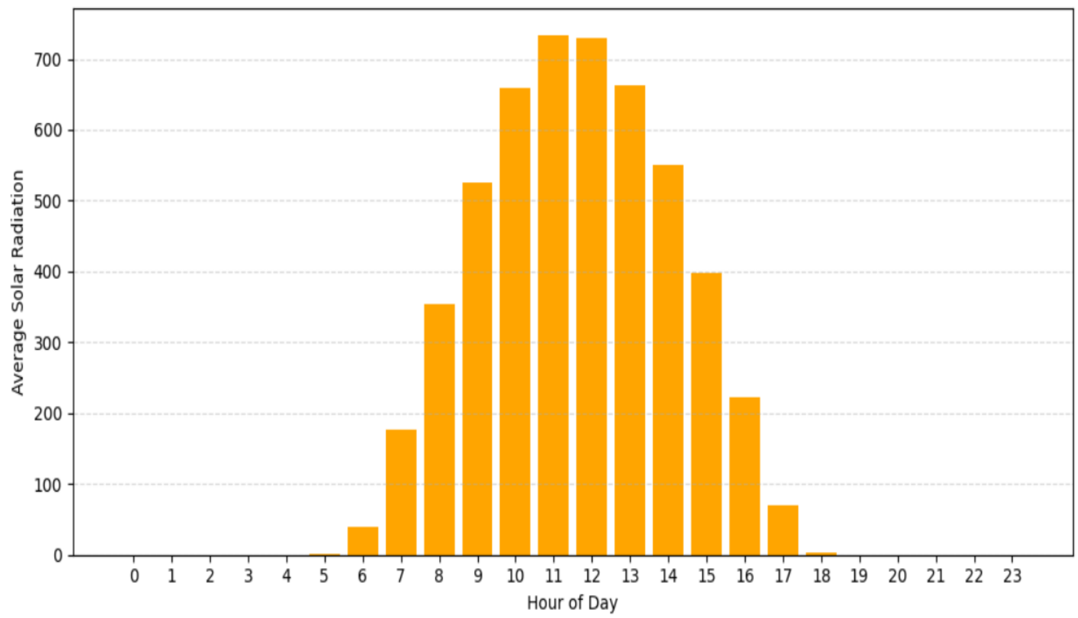

Fig 2 : Average Solar Radiation By Hour

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. The sharp rise in radiation during the morning and the gradual decline toward sunset indicate consistent solar availability during midday. Understanding this hourly distribution is crucial for optimizing photovoltaic system performance and scheduling energy usage in solar-powered applications.

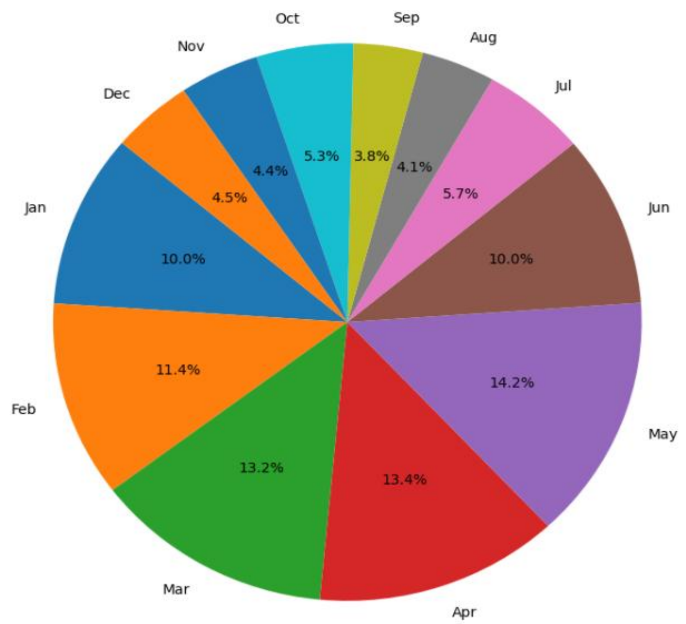

Fig 3 : Monthly Distribution of solar radiation

The pie chart illustrates the monthly distribution of total solar radiation throughout the year. It is evident that April (13.4%) and May (14.2%) contribute the highest share, indicating peak solar intensity during these months. This is followed by March (13.2%) and February (11.4%), which also show substantial solar radiation levels. The lowest contributions are observed in September (3.8%), August (4.1%), and November (4.4%), suggesting reduced solar exposure during the monsoon and early winter periods.

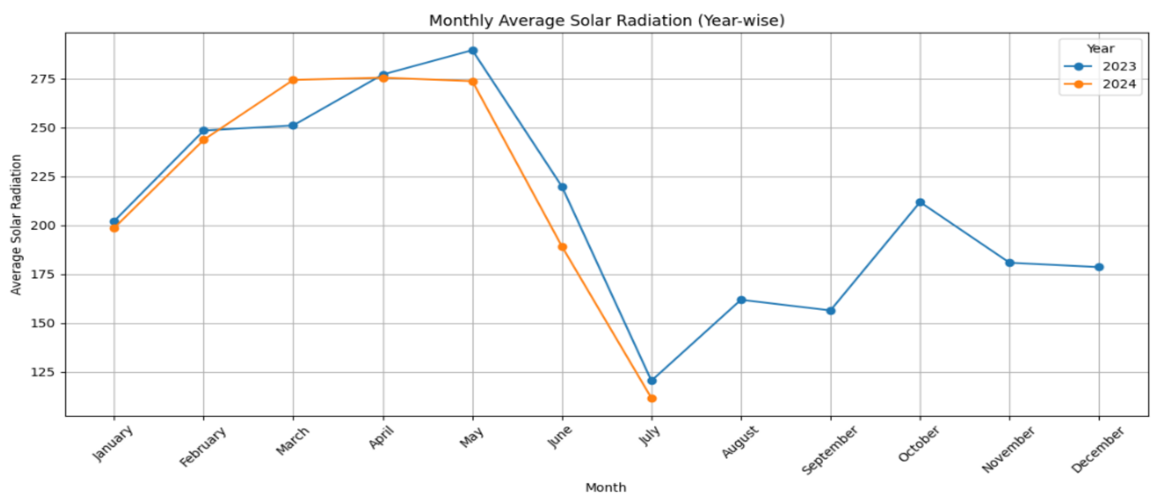

Fig 4 : Line Plot for Monthly Average Solar Radiation

The line plot above presents a comparative analysis of the monthly average solar radiation for the years 2023 and 2024. The X-axis denotes the months of the year, while the Y-axis represents the average solar radiation (measured in W/m²). The plot provides valuable insights into seasonal variations in solar energy availability, which are critical for designing and optimizing solar power systems.

In both years, solar radiation levels exhibit a typical seasonal pattern, peaking during the pre-monsoon months (March to May) and declining significantly during the monsoon period (June to August). For instance, May 2023 recorded the highest average radiation, while July 2024 experienced the lowest due to heavy cloud cover and rainfall typical of monsoon seasons in tropical regions.

Notably, 2024 shows slightly higher radiation in the early months (January–March) compared to 2023, while the mid-year values (June and July) are comparatively lower. These inter-annual differences could be attributed to dynamic atmospheric factors such as cloud cover, aerosol concentration, and temperature fluctuations.

The visualization highlights the importance of temporal variability in solar radiation, which must be considered when forecasting energy output or planning installations. Such comparative plots aid in validating prediction models, understanding annual weather dynamics, and improving the robustness of renewable energy systems.

3.4 Feature Engineering and Initial Variable Selection

This study adopted a two-step method for feature selection. Initially, manual feature engineering was performed to generate time-related and lag-based features such as month and previous hour radiation to capture recurring and time-dependent 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.

3.5 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. The best-performing models were Stacked Ensembles, with RMSE ranging from 21.80 to 22.09 and MAE between 9.95 and 10.21, followed closely by Gradient Boosting Machines (GBMs), which showed RMSE values between 22.31 and 23.79 and MAE from 10.14 to 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.

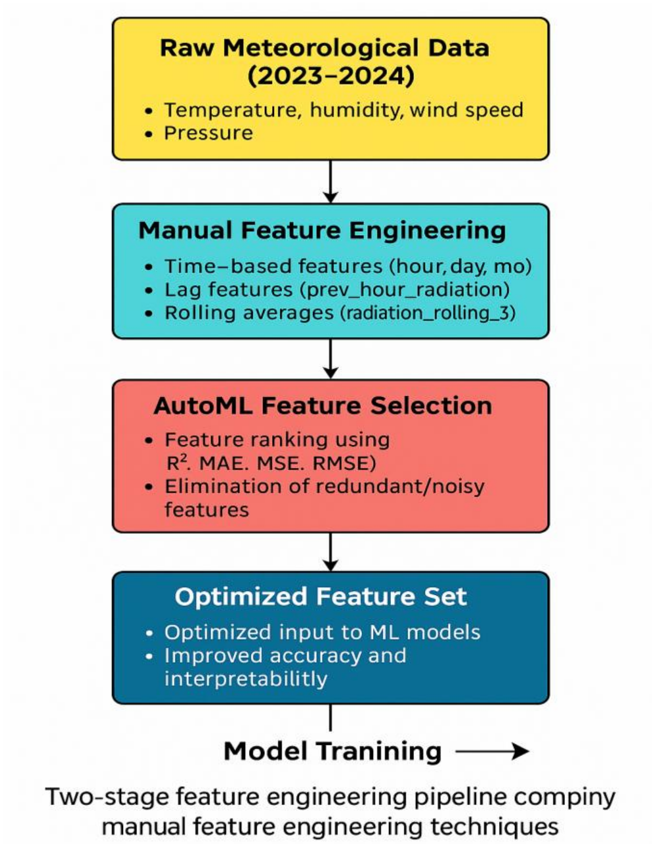

Fig 5 : Workflow of Manual and Automated Feature Selection

3.6 Machine Learning Algorithms

The research utilized advanced machine learning models to estimate solar radiation based on key meteorological variables. Models like XGBoost, Random Forest, LightGBM, and a multi-layer Neural Network with the Adam optimizer were evaluated. A Stacking Regressor was implemented, using Linear Regression as the meta-model, to improve the overall generalization of predictions. 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 Ensemble Stacking achieved the best accuracy, effectively capturing the temporal and environmental patterns of solar radiation.

3.7 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 (Extreme Gradient Boosting)

XGBoost is a powerful and efficient implementation of gradient boosting designed for speed and performance. It creates a series of decision trees in sequence, with each tree aiming to reduce the mistakes made by the one before it. XGBoost incorporates regularization techniques (L1 and L2) to reduce overfitting and supports parallel computation, handling missing data, and efficient memory usage. Its strong performance and ability to scale make it a popular choice for solving structured data challenges. In this research, XGBoost was fine-tuned using both Grid Search and Randomized Search to optimize key parameters such as the number of estimators, maximum depth, and learning rate. This tuning helped the model effectively capture complex patterns in solar radiation data, leading to improved prediction accuracy.

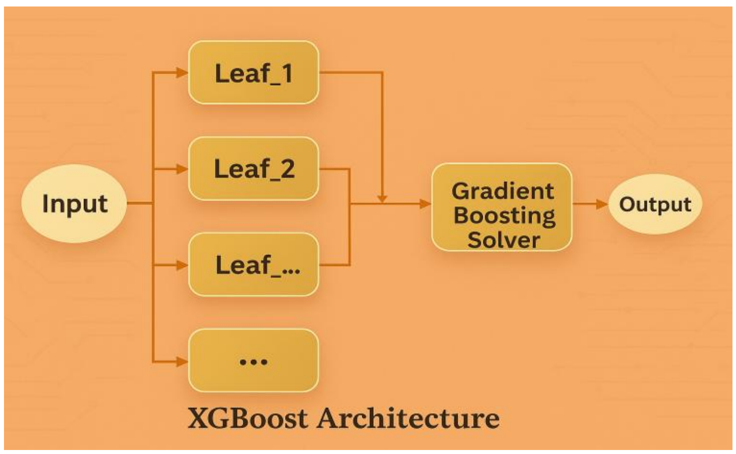

Fig 6 : XG-Boost Architecture

Random Forest Regressor

The Random Forest algorithm performs regression by building several decision trees and producing the final output based on the average prediction from all trees. Unlike single-tree models, Random Forest reduces variance through this ensemble approach, resulting in more stable and reliable predictions. It is particularly well-suited for datasets with complex feature interactions and non-linear relationships, such as those found in meteorological conditions influencing solar radiation.

This study focused on tuning essential hyperparameters such as the number of trees, the maximum depth allowed for each tree, and the minimum number of samples needed to split a node and all these parameters were optimized using Randomized Search. This method allowed for efficient exploration of the hyperparameter space, enhancing the model's generalization ability while preventing overfitting. Random Forest proved effective in handling noisy data and capturing diverse patterns across weather conditions, contributing to improved solar radiation prediction accuracy in the comparative analysis.

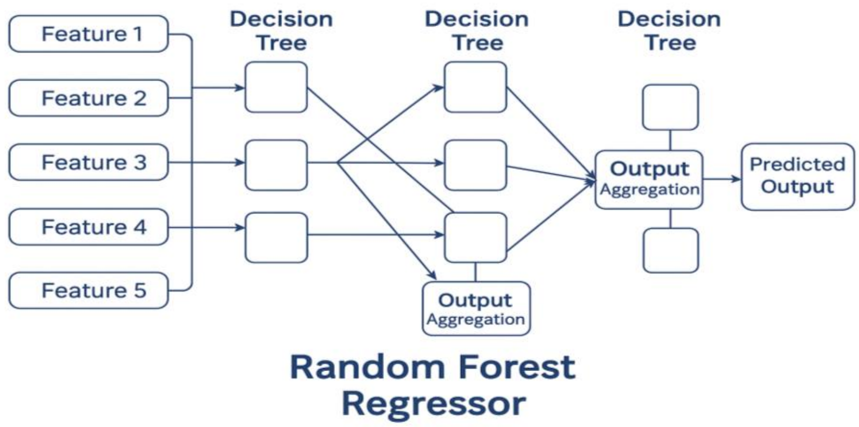

Fig 7 : Random Forest Architecture

LightGBM Regressor

LightGBM is a high-performance gradient boosting framework optimized for speed and low memory usage, making it suitable for handling large and complex datasets. Unlike traditional boosting methods, LightGBM grows trees leaf-wise, which enables it to converge faster and achieve better accuracy with fewer iterations. This makes it particularly suitable for solar radiation prediction, where fine-grained meteorological features play a significant role.

Key parameters such as the number of leaves, learning rate, maximum tree depth, and feature fraction were carefully adjusted to strike a balance between speed, accuracy, and overfitting control. LightGBM's native support for categorical features and built-in regularization further contributed to its robustness when modeling complex patterns in weather data. Its performance in this study confirmed its strength as a lightweight yet highly accurate alternative to traditional ensemble methods.

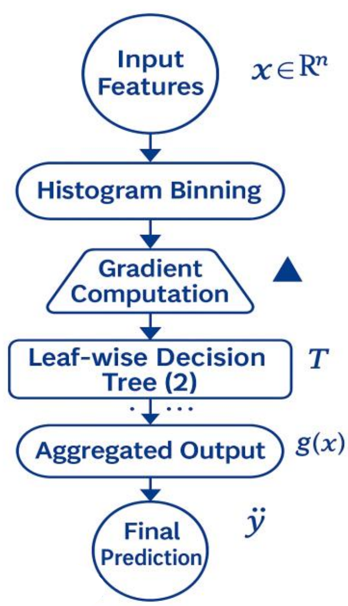

Fig 8 : LightGBM Architecture

Neural Network :

A feedforward Neural Network architecture was employed to model the complex, non-linear relationships present in the solar radiation dataset. The network consisted of three hidden layers with 128, 64, and 32 neurons respectively, each activated using ReLU functions. The model was trained using standardized inputs, optimized with the Adam optimizer, and evaluated using Mean Squared Error (MSE) as the loss function. Training was carried out over 100 cycles (epochs), with data processed in batches of 32 samples at a time.

This setup allowed the neural network to effectively learn complex patterns within the data, especially those influenced by time-based and weather-related variations. Its flexibility and ability to generalize well made it a competitive performer in predicting solar radiation when compared to tree-based models.

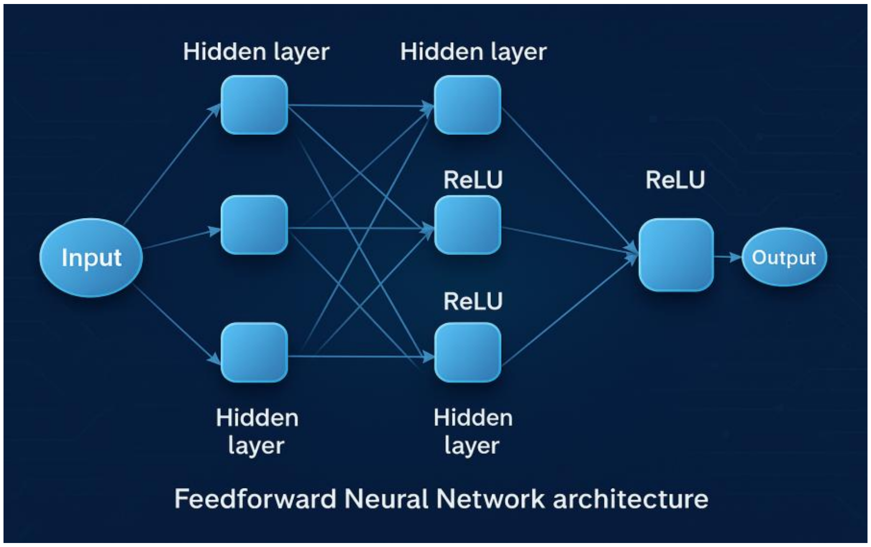

Fig 9 : Architecture of Neural Network

Stacking Regressor :

The stacking ensemble approach combines the predictive strengths of multiple base learners—namely XGBoost, Random Forest, LightGBM, and a Neural Network. These models individually learn from the training data and generate predictions, which are then fed into a meta-learner, in this case, a Linear Regression model. The meta-learner learns to best combine the outputs of the base models to produce the final prediction. Each base model was fine-tuned using hyperparameter optimization prior to integration. Cross-validation was employed to validate the ensemble's performance, ensuring that it generalizes well across different subsets of the data. This technique effectively leverages the diversity among models, reducing bias and variance for improved solar radiation prediction accuracy.

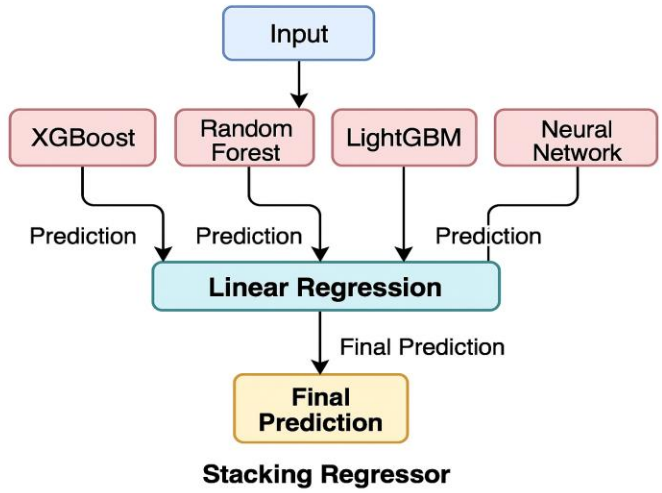

Fig 10 : Architecture of Stacking Regressor

3.8 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:

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

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 is particularly effective in penalizing larger errors more than smaller ones, making it useful for understanding how well a model handles significant deviations.

A lower MSE indicates that the model makes predictions that are closer to the actual values, reflecting better overall accuracy. Due to its sensitivity to large errors, MSE is a crucial metric in solar radiation prediction, where precise estimations are essential for reliable energy planning.
Its formula is:

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

Root Mean Squared Error (RMSE) : RMSE serves as a crucial metric for evaluating the performance of regression models in this study. It represents the square root of the average squared differences between predicted and actual values, thereby expressing prediction error in the same unit as the target variable. As a result, RMSE is easier to understand and apply in practical scenarios than MSE.

By giving greater weight to larger errors, RMSE is particularly useful for identifying models that perform poorly under extreme conditions. It helps assess both the accuracy and reliability of predictions, especially in applications like solar radiation forecasting, where larger deviations can impact energy planning and grid operations. Its formula is:

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

Coefficient of Determination (R^2 Score):

This study employs the Coefficient of Determination (R^2) to assess the overall performance of the predictive models. R^2 indicates how much of the variation in the target variable solar radiation is accounted by the model's input features. A value approaching 1 suggests that the model accurately captures most of the variability within the data.

R^2 is particularly useful for comparing different models, as it provides a normalized score indicating how well the model captures the underlying data patterns. In the context of solar radiation prediction, a high R^2 value reflects a model's ability to generalize across varying weather conditions and time-based fluctuations. It is given by the formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi - \widehat{yi})^2}{\sum_{i=1}^{n}(yi - \overline{yi})^2}$$

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.

3.9 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.

3.10 Result

To evaluate the generalization capability and robustness of machine learning models in predicting solar radiation, this study employed a 5-fold cross-validation approach. In this technique, the training dataset was partitioned into five equal subsets (folds). During each run, the dataset was split into five segments, with the model trained on four segments and tested on the fifth to assess its performance.

This approach provided a more reliable and unbiased estimate of model performance by mitigating the risk of overfitting and reducing the dependency on a single train-test split. By averaging performance across all folds, the method ensured that the models were evaluated under varied conditions, improving their generalization to unseen data. The results from cross-validation revealed that the stacked ensemble model and neural network consistently outperformed other models across all folds. These models demonstrated superior predictive accuracy and minimized error metrics, highlighting their reliability for real-world solar energy forecasting applications. This validation strategy was crucial in selecting models with strong and stable performance, ultimately enhancing the credibility and applicability of the study's findings.

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 and consistent results in solar radiation prediction, demonstrating its ability to capture complex relationships within the meteorological data. It recorded a mean R^2 of 0.981, indicating that the model was able to explain 98.1% of the variance in the target variable. In terms of error metrics, it achieved a mean MSE of 1165.64, a mean MAE of 20.00, and a mean RMSE of 33.77, reflecting solid predictive accuracy with relatively low error.

These results highlight Random Forest's strength in handling non-linear dependencies and reducing overfitting through its ensemble of decision trees. Its robustness, scalability, and ability to handle high-dimensional data make it a reliable choice for solar radiation modeling, especially when interpretability and performance are both essential. The model's inherent feature importance mechanism also aids in identifying the most influential meteorological parameters contributing to solar radiation. Furthermore, its performance remained consistent across cross-validation folds, showcasing its stability under varying data distributions. Random Forest's resistance to noise and overfitting makes it especially suitable for complex real-world datasets with non-linear interactions. Its parallel processing capability also enhances computational efficiency, enabling faster training on large datasets.

LightGBM :

LightGBM achieved a mean R^2 score of 0.966, indicating that it was able to explain 96.6% of the variance in solar radiation values. It recorded a mean MSE of 2303.17, a mean MAE of 36.99, and a mean RMSE of 47.77, reflecting slightly higher prediction error compared to Random Forest and XGBoost. Despite this, the model effectively captured patterns within the meteorological data and maintained strong performance.

Its gradient boosting framework, coupled with efficient leaf-wise tree growth and native support for categorical features, makes LightGBM a powerful model for structured data. Despite the slightly lower accuracy, its speed, scalability, and ability to handle large datasets efficiently highlight its value in solar radiation prediction tasks where computational performance is also a factor.

Neural Network :

The Neural Network model exhibited strong predictive performance, achieving a mean R^2 score of 0.988, meaning it explained 98.8% of the variability in the actual solar radiation values. It recorded a mean MSE of 698.85, a mean MAE of 13.52, and a mean RMSE of 25.87, reflecting its ability to deliver accurate and consistent predictions across all folds of cross-validation.

The model's deep architecture, featuring multiple hidden layers with non-linear activation functions, enabled it to learn intricate dependencies within the meteorological data. Its strength lies in its capacity to automatically capture complex patterns and interactions between features such as temperature, humidity, wind speed, and pressure making it highly effective for solar radiation modeling. The Neural Network consistently ranked among the top-performing models in this study, demonstrating excellent generalization and minimal overfitting.

Stacked Model :

The stacked model achieved strong results, recording an average R^2 score of 0.993, along with a mean MSE of 495.41, MAE of 11.45, and RMSE of 22.25. This model, combining predictions from multiple models, demonstrated superior accuracy and robustness in solar radiation forecasting.

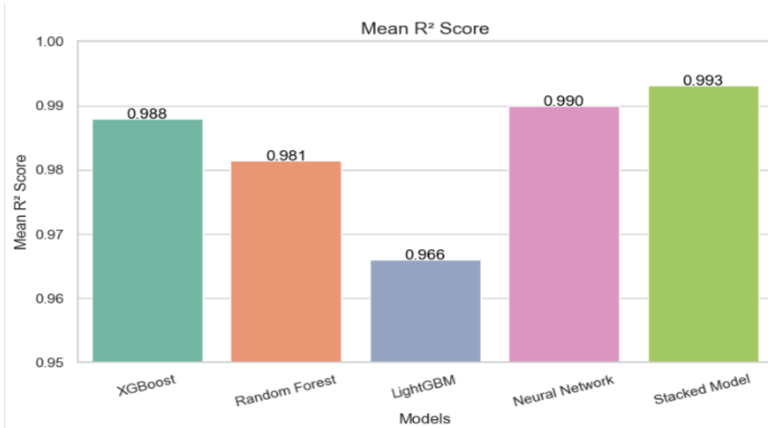

Fig 11 : 5-Fold Cross-Validated R^2 Scores

This bar chart compares the average R^2 scores achieved by five different models: XGBoost, Random Forest, LightGBM, Neural Network, and the Stacked Model. Among them, the Stacked Model demonstrated the best overall performance, achieving the highest R^2 score of 0.993, indicating its superior ability to generalize across varying data patterns.

The chart clearly highlights that while Neural Network (0.988) and XGBoost (0.988) also performed exceptionally well, the ensemble-based Stacked Model was able to leverage the strengths of multiple algorithms to deliver even higher predictive accuracy. This visualization reinforces the effectiveness of model ensembling in improving the robustness and performance of solar radiation prediction systems.

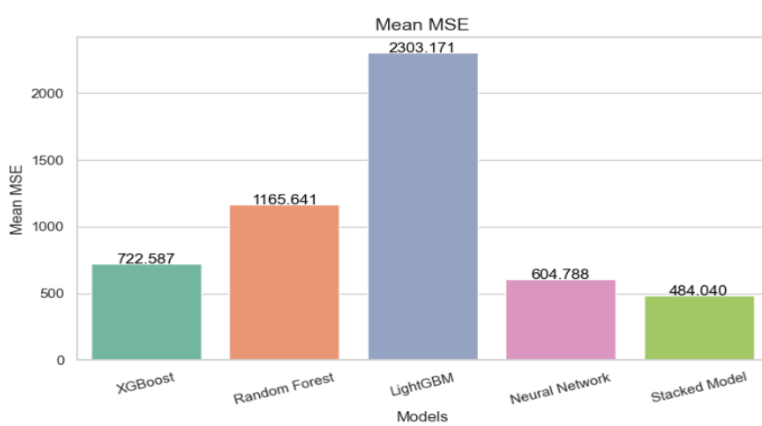

Fig 12 : 5-Fold Cross-Validated Mean Squared Error (MSE) Scores

This bar chart illustrates the average MSE for each model, reflecting the squared error between actual and predicted values. The Stacked Model achieved the smallest MSE, reflecting its strong effectiveness, whereas LightGBM showed the highest MSE, indicating potential for enhancement.

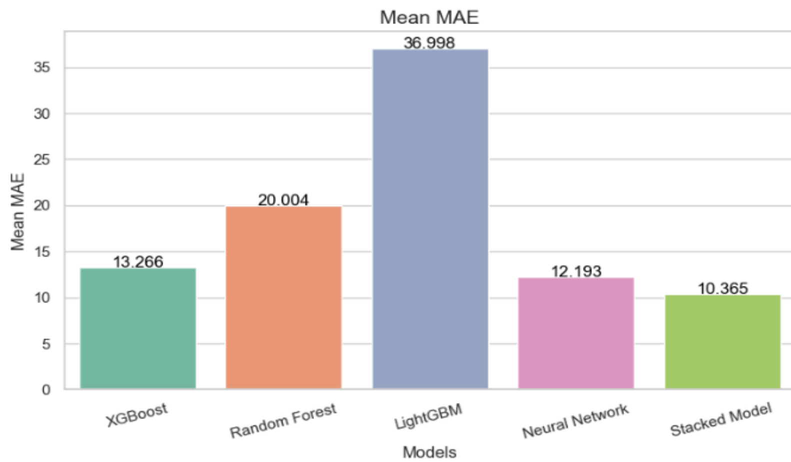

Fig 13 : 5-Fold Cross-Validated Mean Absolute Error (MAE) Scores

This bar chart illustrates the average MAE across different models, where a lower value indicates better prediction accuracy. The Stacked Model achieved the lowest MAE, demonstrating the highest precision, followed by the Neural Network and XGBoost, which also performed competitively. This highlights how ensemble and deep learning techniques are capable of reducing prediction errors effectively.

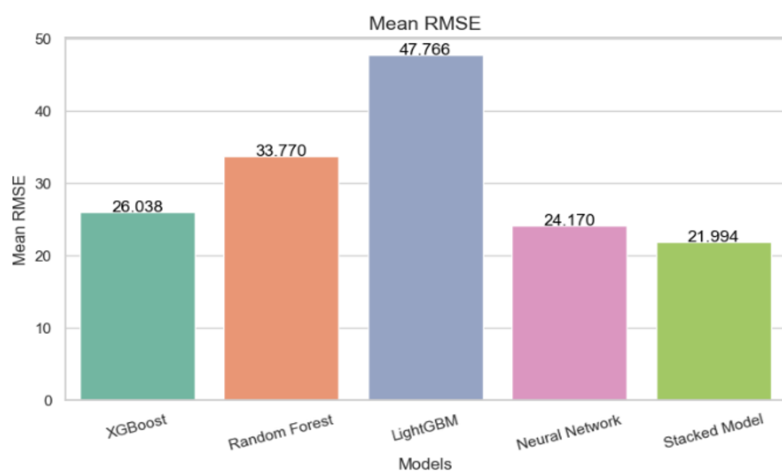

Fig 14 : 5-Fold Cross-Validated Root Mean Squared Error (RMSE) Score

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 achieved the lowest RMSE value, highlighting its high prediction accuracy, with the Neural Network and XGBoost models performing next best.

3.11 Overfitting Check

To assess overfitting, R^2 scores were calculated for both the training and testing sets across all models Random Forest, XGBoost, LightGBM, Neural Network, and the Stacked Ensemble. The R^2 gap, defined as the difference between training and testing scores, was used as a quantitative measure of model generalization. A threshold of 0.05 was set to identify overfitting risks. All models demonstrated 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. These results validate the robustness and reliability of the models in accurately predicting solar radiation without simply memorizing patterns from the training data.

To support this analysis, a custom Python function was developed using sklearn.metrics.r2_score, which systematically calculated and printed the R^2 values for each model's training and testing phases. The function identified possible overfitting when the difference in R^2 scores went beyond the specified threshold. This automated approach ensured a consistent and efficient evaluation across all models, highlighting the reliability of ensemble and deep learning models, particularly the Stacked Ensemble and Neural Network, which showed the smallest generalization gaps.

3.12 Month-Based Temporal Split Evaluation

To assess how well the model generalizes to unseen future data, an independent month-based evaluation was performed. The complete dataset, spanning from January to July 2024, was split into two distinct periods:

- **Training Set:** January 2023 to June 2024
- **Testing Set:** July 2024

This approach reflects a real-world deployment scenario, where models are trained on historical data and tested on a future period to evaluate their temporal robustness and forecasting capability. The Stacked Ensemble model, which combines XGBoost, Random Forest, LightGBM, and Neural Network as base learners, was trained using the preprocessed features from the January–June dataset. The model was then tested on the July data to predict solar radiation (ALLSKY_SFC_SW_DWN) and evaluate its generalization ability.

The model's performance was assessed using standard regression metrics:

- **R^2 Score (Coefficient of Determination):** Measures how well the predicted values match the actual values.

  MAE (Mean Absolute Error): This metric shows how far predictions are from actual values on average, treating all errors equally regardless of whether they are positive or negative.

- **RMSE (Root Mean Squared Error):** Gives more weight to large errors and is sensitive to outliers.

Metric	Training Set (Jan 2023 – Jun 2024)	Testing Set (July 2024)
R^2 Score	0.9962	0.9578
RMSE	16.30	26.90

Table 2 : Model Evaluation Result

The table presents a summary of evaluation metrics for both the training and testing phases of the model. It highlights the consistency of model performance across different time periods, demonstrating stability when applied to new data.

Two plots were generated to further evaluate the model's predictions:

1. **Scatter Plot – Actual vs. Predicted (Test Set):**

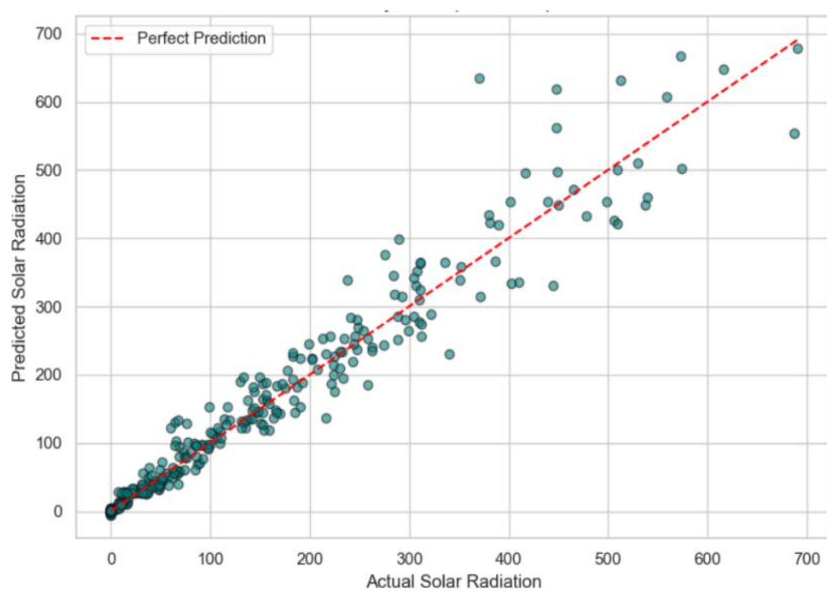

Fig 15 : Scatter Plot of Actual vs Predicted Solar Radiation for July 2024

The scatter plot shows the relationship between the actual and predicted solar radiation values for July 2024 using the Stacked Ensemble model. Each point represents a prediction for a specific time instance, and the red dashed line indicates the ideal case where predictions perfectly match the actual values. Most of the points are closely clustered around this line, demonstrating the model's strong predictive capability.

The high R^2 score of 0.96 further supports the model's accuracy. Although a few outliers are visible, especially at higher radiation values, the overall distribution confirms a strong correlation and low prediction error, making the model reliable for forecasting solar radiation.

2. **Time Series Plot – Predicted vs. Actual Over July:**

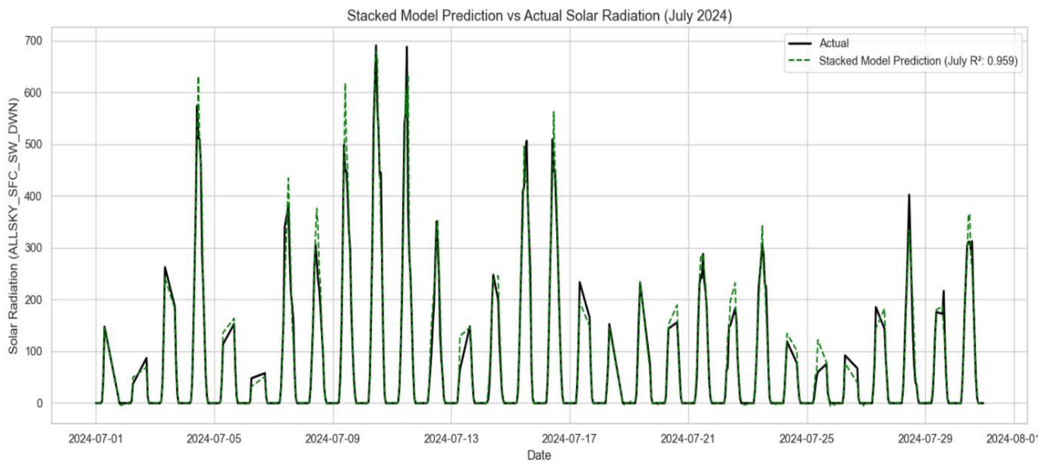

Fig 16 : Time Series Plot of Predicted vs Actual Solar Radiation Using Stacked Model

The time series plot compares the actual and predicted solar radiation values for July 2024 using the Stacked Ensemble model. The black line represents the observed solar radiation, while the green dashed line shows the model's predictions. The close alignment between the two curves across the entire month demonstrates the model's ability to capture the daily solar radiation trends with high accuracy. The model successfully tracks the characteristic rise and fall of solar radiation values, reflecting its effectiveness in learning temporal patterns.

Despite minor deviations on some days, particularly during peak radiation periods, the overall prediction remains consistent and reliable. The high R^2 value of 0.959 further confirms the model's strong generalization capability. This result validates the suitability of the Stacked Ensemble approach for forecasting solar radiation in real-world scenarios where accuracy and temporal consistency are critical.

Chapter 4 : Conclusion

This study focused on accurately predicting solar radiation using advanced 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 the H2O AutoML framework, the models effectively captured complex meteorological and temporal dependencies in the data. Key engineered features included time-based variables (such as hour, day, and month), lag values, and rolling statistical measures (like moving averages), which significantly enhanced the model's sensitivity to short-term fluctuations and seasonal trends. This hybrid approach of combining domain-driven feature construction with automated optimization enabled the creation of highly accurate and generalizable models.

Model performance was assessed using standard evaluation metrics including R^2, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). Among all the models tested, the Stacked Ensemble model delivered the best results, achieving an R^2 score of 0.9932, MAE of 10.35, and RMSE of 21.91, indicating its exceptional ability to minimize both bias and variance in predictions. The Neural Network also showed competitive performance, with an R^2 of 0.9883, MAE of 13.52, and RMSE of 25.87, suggesting its strength in capturing non-linear relationships in the data. Comparatively, LightGBM and Random Forest, while efficient and interpretable, showed lower predictive accuracy in this specific context, likely due to their reduced capacity to model temporal dependencies without deep learning architectures or stacking enhancements.

These findings emphasize the importance of ensemble learning and automated model tuning in achieving high-accuracy solar radiation prediction. The superior performance of the Stacked Ensemble model highlights how combining multiple base learners can mitigate individual model weaknesses and yield robust predictions. Moreover, the study underscores the practical benefits of using ML for solar radiation prediction, enabling smarter grid integration, optimized energy planning, and data-driven decision-making in the context of renewable energy systems. The results demonstrate how integrating AutoML pipelines with advanced learning models not only boosts predictive performance but also simplifies deployment workflows for use in real-world solar energy applications.

Chapter 5 : Future Work

Building upon the findings of this study, several directions can be pursued to further enhance the accuracy and practicality of solar radiation prediction. One promising extension involves the inclusion of additional meteorological and environmental variables, such as cloud index, solar zenith angle, aerosol optical depth, and dew point temperature, which can help capture more nuanced atmospheric conditions affecting solar radiation.

Moreover, advanced deep learning architectures such as Long Short-Term Memory (LSTM) networks or hybrid CNN-LSTM models could be explored to better capture complex temporal dynamics and long-term dependencies in time-series weather data. These architectures may provide even greater accuracy by learning sequential patterns more effectively than conventional models.

Another valuable direction involves expanding the dataset both spatially and temporally—by including multiple geographical locations and extended time periods—to evaluate the generalizability and robustness of the models under diverse environmental conditions. This would not only improve the model's adaptability but also make it suitable for broader deployment.

From an automation standpoint, future studies could investigate the use of other AutoML platforms such as TPOT, Google Cloud AutoML, or AutoKeras. These platforms could offer different optimization strategies, model pipelines, and hyperparameter tuning mechanisms that may yield superior results in specific contexts.

Lastly, deploying the models in real-time solar energy monitoring and control systems could greatly enhance operational decision-making. Integrating these models into IoT-based smart grid environments or renewable energy management platforms would allow for dynamic energy forecasting and efficient resource allocation, thus supporting the transition to a more sustainable energy future.

REFERENCES

[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