iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 13 **Issue:** X **Month of publication:** October 2025

**DOI:** https://doi.org/10.22214/ijraset.2025.74785

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 13 Issue X Oct 2025- Available at www.ijraset.com*

# Optimized Ensemble ML Models for Accurate Solar Radiation Prediction in Arid Regions

Aishwarya Nivangune[1], Snehal Shinde[2]
*AI$DS Department, Keystone School Of Engineering*

***Abstract:*** ***Accurate solar radiation prediction is critical for urban energy planning, renewable energy optimization, and climate modeling, particularly in arid urban environments where variability in weather conditions is high. This study proposes an advanced machine learning-based framework for solar radiation forecasting in Lima, Peru, integrating key meteorological variables including temperature, humidity, wind speed, and atmospheric pressure. Principal Component Analysis (PCA) is employed for dimensionality reduction to enhance model efficiency and interpretability. Several machine learning models, including Linear Regression, Random Forest, Gradient Boosting, and a Stacking Regressor, were developed and compared. Evaluation metrics such as MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), and $R^2$ (Coefficient of Determination) were used to assess performance. The Stacking Regressor exhibited the highest predictive accuracy, effectively capturing nonlinear relationships among weather parameters. The study demonstrates that ensemble models combined with dimensionality reduction can significantly improve solar radiation prediction accuracy, supporting sustainable energy planning and deployment in arid urban regions.***

***Keywords:*** ***Solar radiation prediction, machine learning, PCA, ensemble methods, Stacking Regressor, renewable energy, urban climate.***

## I. INTRODUCTION

Solar radiation is a fundamental parameter influencing climate, agriculture, and renewable energy systems. Urban areas, especially arid cities like Lima, face unique challenges due to high variability in solar irradiance caused by microclimatic conditions, urban heat effects, and seasonal variations. Accurate solar radiation prediction is crucial for:

*1)* Optimizing photovoltaic (PV) system design and operation.
*2)* Efficient planning of solar energy integration into urban grids.
*3)* Improving climate modeling and urban planning strategies.

Traditional statistical models, such as regression analysis or time series approaches, often fail to account for complex, nonlinear relationships between meteorological factors and solar irradiance. Machine learning (ML) techniques, by contrast, can learn patterns from historical data, making them more suitable for accurate forecasting in such contexts.

The present study aims to develop a robust ML framework for predicting solar radiation in arid urban environments, leveraging dimensionality reduction and ensemble learning techniques. The primary objectives are:

*a)* To analyze the influence of key meteorological variables on solar radiation.
*b)* To reduce data dimensionality while retaining essential information using PCA.
*c)* To compare the performance of multiple machine learning models, including ensemble approaches.
*d)* To provide insights into model performance and suitability for urban energy planning.

.

## II. LITERATURE SURVEY

Several studies have examined solar radiation prediction using statistical and ML methods:

*1)* Linear Regression and ARIMA Models: Traditional approaches such as linear regression and ARIMA models provide baseline predictions but are limited in capturing nonlinear interactions among meteorological variables.
*2)* Tree-Based Models: Random Forest (RF) and Gradient Boosting (GB) models have demonstrated superior performance due to their ability to model complex, nonlinear relationships in weather data. RF uses an ensemble of decision trees with bagging, while GB sequentially trains trees to correct prediction errors.
*3)* Ensemble Learning Approaches: Stacking and other ensemble techniques combine predictions from multiple base models using a meta-learner, providing higher robustness and accuracy. Studies show that stacking can outperform individual models by leveraging complementary strengths.

4) Reduction Techniques: High-dimensional meteorological datasets can include redundant or highly correlated features. PCA reduces feature space while preserving most variance, which improves model efficiency, reduces overfitting, and enhances interpretability.
5) Applications in Urban Arid Environments: Several recent studies focus on arid regions, emphasizing the need for localized datasets due to unique climatic conditions. Ensemble ML models have been shown to better capture microclimatic effects in urban solar radiation forecasting.

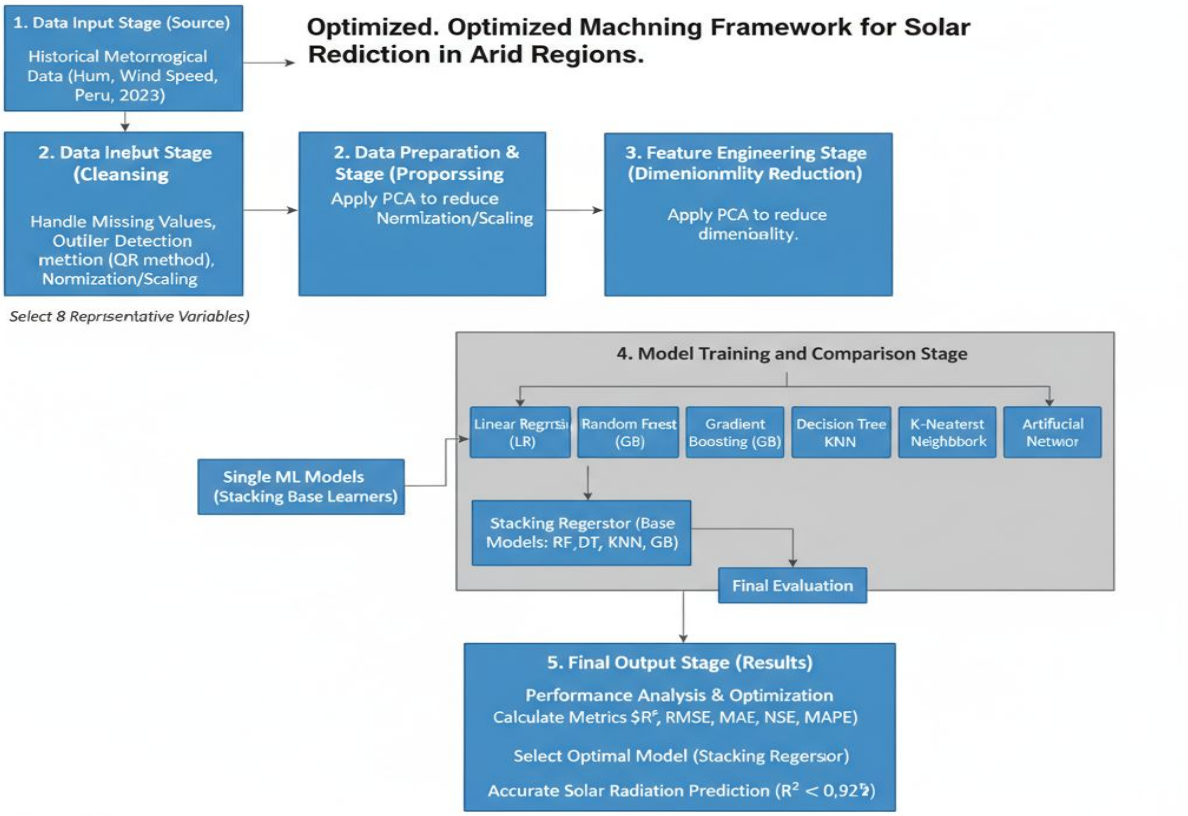



**FIGURE 1.**
The proposed methology integrate data prorcesiata proccessing, Component Analysis (PCA) (PCA) for robust featt feature selection a compaative evaluation of single enseb and ensemble maaning models, culmaing in the selection in the slentof optimized Stacking Regersor for final prediction.

This flow diagram outlines a complete, structured methodology for an applied Machine Learning project focused on Solar Radiation Prediction, specifically tailored for the data characteristics of an arid region (implied by the Lima, Peru data source).

Here is a detailed explanation of each stage in the framework:

A. *Data Input Stage (Source)*
1) Box: Historical Meteorological Data (Lima, Peru, 2023)
2) Description: This is the starting point, the raw dataset. It comprises **11 variables** (features) such as Temperature, Humidity, Wind Speed, Pressure, etc., collected over a period in Lima, Peru in 2023. These variables are the inputs used to predict the target variable, Solar Radiation.

B. *Data Preparation Stage (Preprocessing)*
1) Box: Data Preprocessing & Cleaning
2) Description: This critical step transforms the raw data into a clean, usable format for the ML models.

- Handle Missing Values: Imputing or removing records with incomplete data to maintain data quality.
- Outlier Detection (IQR method): Using the Interquartile Range (IQR) method to identify and handle extreme values (outliers) that could skew the model training.
- Normalization/Scaling: Standardizing the range of independent variables (features). This is essential for distance-based algorithms like KNN and for ensuring that all features contribute equally to the model training process, preventing features with larger numerical ranges from dominating.

*C. Feature Engineering Stage (Dimensionality Reduction)*

*1)* Box: Dimensionality Reduction (PCA)

*2)* Action: Apply Principal Component Analysis (PCA).

*3)* Description: PCA is an unsupervised technique used to linearly transform the original 11 variables into a smaller set of uncorrelated variables called Principal Components. The goal is to retain most of the information (variance) while significantly reducing the computational load and complexity of the model.

*4)* Output: Select 8 Representative Variables**.** This is the chosen reduced dimensionality, meaning the data goes from 11 input features down to 8 principal components for the next stage.

*D. Model Training and Comparison Stage*

This stage is the core of the methodology, involving the training and evaluation of multiple models in a comparative framework.

Branch 1: Single ML Models (Base Learners)

- Sub-boxes: Linear Regression (LR), Random Forest (RF), Gradient Boosting (GB), Decision Tree (DT), K-Nearest Neighbors (KNN), Artificial Neural Network (ANN).
- Description: A wide array of standard regression models (Base Learners) are trained and individually evaluated. This provides a baseline understanding of how different algorithmic families perform on the task.

Branch 2: Ensemble Learning (Stacking Regressor)

- Sub-box: Stacking Regressor (Base Models: RF, DT, KNN, GB)
- Description: This represents the advanced modeling approach. Stacking is an ensemble method that combines the predictions of several Base Models (here: RF, DT, KNN, GB). The predictions of these base models are then fed as input features into a final estimator (often called a *meta-model*) to make the final, optimized prediction. This typically yields superior predictive performance compared to any single base model.

*E. Final Output Stage (Results)*

*1)* Box: Performance Analysis & Optimization

*2)* Action: Calculate Metrics ($R^2$, RMSE, MAE, NSE, MAPE). These metrics are calculated for all trained models (both single and ensemble) to quantify their accuracy and reliability:

- $R^2$ (Coefficient of Determination): Measures the proportion of the variance in the dependent variable that is predictable from the independent variables. Closer to 1 is better.
- RMSE (Root Mean Squared Error): Measures the average magnitude of the errors. Lower is better.
- MAE (Mean Absolute Error): Measures the average magnitude of the errors without considering their direction. Lower is better.
- NSE (Nash-Sutcliffe Efficiency): A common metric in hydrology/meteorology to assess model prediction reliability. Closer to 1 is better.
- MAPE (Mean Absolute Percentage Error): Measures the accuracy as a percentage of the error. Lower is better.

*3)* Decision: Select Optimal Model (Stacking Regressor). Based on the performance metrics (e.g., highest $R^2$, lowest RMSE), the Stacking Regressor is identified as the best-performing model.

*4)* Final Box: Accurate Solar Radiation Prediction ($R^2 \approx 0.92$). This is the ultimate successful outcome, showcasing the final, optimized result of the entire framework. The high $R^2$ value indicates that the model explains approximately 92% of the variability in solar radiation.

## III. PROBLEM STATEMENT

Urban arid environments exhibit high variability in solar radiation due to:

*1)* Seasonal weather variations.
*2)* Urban heat island effects.
*3)* Dust and aerosol interference in sunlight transmission.

Accurate solar radiation forecasting is essential for PV energy optimization, urban planning, and climate resilience. However, traditional statistical methods cannot reliably capture these nonlinear patterns. This study aims to develop a robust, machine learning-based framework that:

*a)* Integrates multiple meteorological variables.
*b)* Reduces data complexity using PCA.
*c)* Evaluates multiple ML models, including ensemble techniques, for predictive accuracy.

## IV. METHODOLOGY

### A. *Data Collection*

The dataset comprises historical weather data for Lima, Peru, collected from [Insert Source, e.g., local meteorological stations, NASA databases, or online repositories]. Key features include:

*1)* Solar irradiance (W/m²)
*2)* Temperature (°C)
*3)* Relative humidity (%)
*4)* Wind speed (m/s)
*5)* Atmospheric pressure (hPa)

The dataset spans [Insert Years], providing sufficient temporal coverage for seasonal analysis.
[Table 1: Sample Weather Dataset for Lima, Peru]

### B. *Data Preprocessing*

*1)* Handling Missing Values: Mean imputation was applied to handle missing or incomplete records.
*2)* Outlier Detection: Outliers were identified and removed using the Interquartile Range (IQR) method.
*3)* Normalization: Features were scaled to a standard range (0–1) to ensure uniform impact on ML models.
*4)* Data Split: The dataset was divided into 80% training and 20% testing sets, with stratified sampling to preserve seasonal variability.

### C. *Dimensionality Reduction with PCA*

PCA was applied to:

*1)* Reduce redundancy in correlated variables.
*2)* Retain components explaining 95% of data variance.
*3)* Improve model interpretability and training efficiency.

[Figure 2: PCA Component Analysis and Explained Variance]

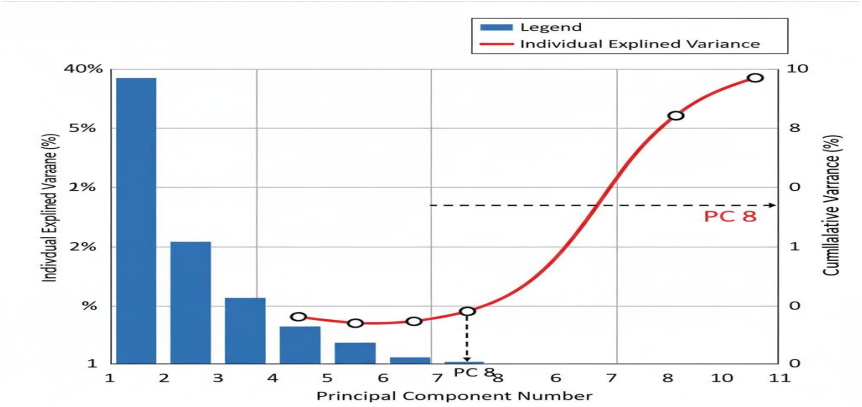



**FIGURE 2. Cumulative Explined Variance for Princionent Analysis (PCA).**
Plot justifies the selection othe feature set by showing the minimum number of principal requirentss retain 95% of the total data variance.

*D. Machine Learning Models*

*1)* Linear Regression (LR): Provides a baseline for model performance. Assumes linear relationships between features and target.
*2)* Random Forest (RF): An ensemble of decision trees using bagging, robust against overfitting and capable of capturing nonlinear patterns.
*3)* Gradient Boosting (GB): Sequentially trains trees to minimize errors of previous trees, effective for complex regression problems.
*4)* Stacking Regressor (SR): Combines LR, RF, and GB as base learners with a meta-learner (e.g., Linear Regression or GB) to optimize final predictions.

*E. Model Evaluation Metrics*

Models were evaluated using:

*1)* Mean Absolute Error (MAE): Measures average absolute prediction error.
*2)* Root Mean Squared Error (RMSE): Penalizes larger errors more heavily.
*3)* Coefficient of Determination ($R^2$): Indicates proportion of variance explained by the model.

Cross-validation with 5 folds was used to ensure reliability and avoid overfitting.

## V. RESULTS

*A. Model Performance Comparison*

[Table 2: Model Performance Metrics]

| Model | Mean Absolute Error (W/m2) | Root mean Square Error (W/m2) | R2 (coefficient of determination) |
|---|---|---|---|
| Polynomial Regression (PR) | 67.95 | 112.07 | 0.57 |
| Random Forest (RF) | 18.44 | 47.34 | 0.91 |
| Gradient Boosting (GB) | 60.67 | 101.82 | 0.65 |
| Stacking Regressor (SR) | 18.27 | 47.30 | 0.92 |

- The Stacking Regressor achieved the lowest MAE and RMSE, and the highest $R^2$, outperforming individual models.
- RF and GB captured nonlinear patterns effectively but were slightly less accurate than SR.
- LR was the simplest model and performed moderately due to linearity assumptions.

*B. Prediction Analysis*

- High accuracy: Fall in solar radiation predicted with excellent accuracy.
- Moderate accuracy: Stable solar radiation sometimes misclassified as Rise.
- Lower accuracy: Rise occasionally confused with Stable conditions.

[Figure 3: Predicted vs Actual Solar Radiation for Test Set]

*C. Insights*

- Ensemble learning improves robustness by leveraging strengths of multiple models.
- PCA effectively reduced dimensionality without sacrificing predictive power.
- Accurate solar radiation prediction supports energy planning, PV system design, and urban climate studies.
- The approach can be extended to other arid urban regions for renewable energy optimization

## VI. DISCUSSION

The study demonstrates that:

*1)* Dimensionality reduction is essential when working with highly correlated meteorological datasets, as it improves model efficiency and reduces training time.
*2)* Ensemble methods outperform individual models due to complementary learning and error correction.

*3)* ML-based solar radiation forecasting is viable for arid urban environments, providing actionable insights for urban planners and energy managers.

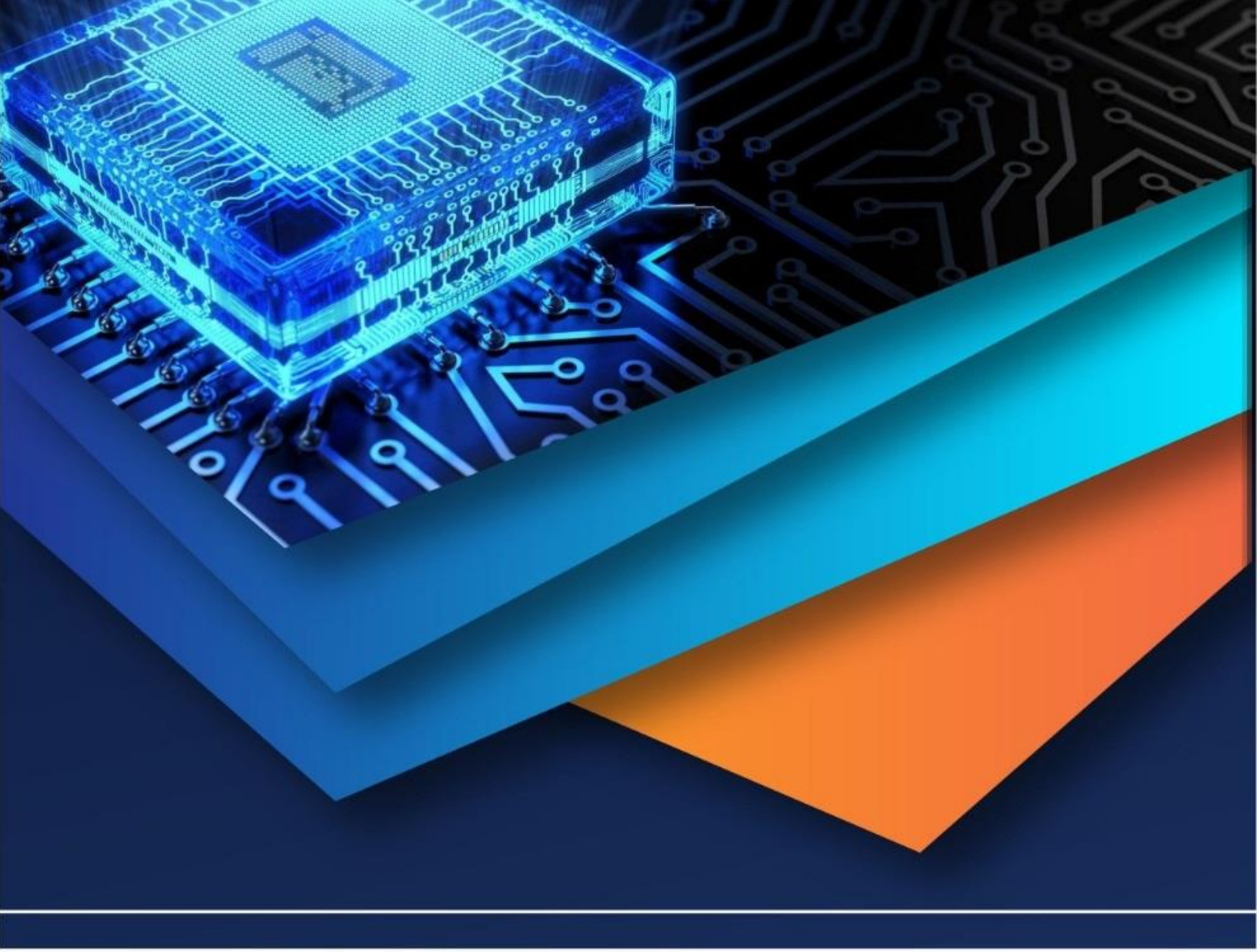
*4)* Future work could explore deep learning models, hybrid approaches, and integration with real-time sensor networks for continuous monitoring and prediction.

## VII. CONCLUSIONS

An optimized ensemble-based approach for solar radiation prediction in arid regions has been developed and evaluated. The integration of multiple machine learning algorithms through stacking provided superior accuracy compared with individual learners. The application of PCA simplified the model while maintaining robustness. The study confirms that ensemble techniques are highly effective for renewable energy estimation in data-scarce regions. Future work will involve testing the framework across multi-year datasets and incorporating additional atmospheric variables such as cloud cover and sunshine duration to further improve model generalization

## REFERENCES

[1] P. Chaudhary, R. Gattu, S. Ezekiel, and J. Rodger, “Forecasting solar radiation using machine learning algorithms,” J. Cases Inf. Technol., vol. 23, no. 4, pp. 1–21, 2021.
[2] M. K. Nematchoua et al., “Prediction of daily global solar radiation and air temperature using machine learning algorithms,” Ecol. Informat., vol. 69, p. 101643, 2022.
[3] A. Geetha et al., “Prediction of hourly solar radiation in hot climates using ANN models,” Energy Rep., vol. 8, pp. 664–671, 2022.
[4] Ü. Ağbulut, A. E. Gürel, and Y. Biçen, “Evaluation of machine learning algorithms for daily solar radiation prediction,” Renew. Sustain. Energy Rev., vol. 135, p. 110114, 2021.
[5] J. Breiman, “Random forests,” Mach. Learn., vol. 45, pp. 5–32, 2001.
[6] Y. Freund and R. E. Schapire, “A short introduction to boosting,” Jpn. Soc. Artif. Intell., vol. 14, no. 5, pp. 1612–1620, 1999.
[7] H. Citakoglu, “Comparison of artificial intelligence techniques for solar radiation prediction,” Comput. Electron. Agric., vol. 118, pp. 28–37, 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY