International Journal of Innovations & Research Analysis (IJIRA)
ISSN :2583-0295(Online), Impact Factor: 7.662, Volume 05, No. 04(I), October- December, 2025, pp 197-208

Performance Analysis of Artificial Intelligence Models for Solar Radiation Forecasting

Vinay Gupta[1*] | Dr. Shyam Sunder Kaushik[2]

[1]Research Scholar, Department of Electrical Engineering, Shri Krishna University, Chattarpur M.P.
[2]Associate Professor, Department of Electrical Engineering, Shri Krishna University, Chattarpur M.P.

[*]Corresponding Author: vinnu56@gmail.com

Citation: Gupta, V., & Kaushik, S. (2025). Performance Analysis of Artificial Intelligence Models for Solar Radiation Forecasting. International Journal of Innovations & Research Analysis, 05(04(I)), 197–208

ABSTRACT

The solar radiation prediction is one of the conditions of successful introduction of photovoltaic systems into modern power grids. The nonlinearity and intermittency nature of the solar irradiance necessitated by complexes of meteorological forces and time have a problem with traditional statistical and physical forecasting models. This paper involves a detailed performance assessment of the machine learning models of artificial intelligence based on solar radiation prediction, on the basis of a real meteorological dataset. The use of a systematic methodology entailing data cleaning, time-based feature extraction, feature engineering and feature selection helps to improve data quality and model learning potential. A number of machine learning models, namely K-Nearest Neighbors (KNN), Extra Trees, Random Forest Regressor, and XGBoost are found and tested through standard regression measurements of MAE, MSE, RMSE and R 2 score. The experimental findings reveal that the overall level of the nonlinear and ensemble-based models is far much better than the conventional ones, and the best predictive accuracy (R 2 = 0.946) is observed to be obtained with the KNN model. The results indicate the suitability of the instance-based and ensemble learning method in describing intricate solar radiation patterns, which is likely to provide enhanced forecasting accuracy when planning a renewable energy resource and grid stability.

***Keywords**: Solar Radiation Forecasting, Artificial Intelligence, Machine Learning, Ensemble Learning, K-Nearest Neighbors, XGBoost, Renewable Energy Prediction.*

Introduction

Proper forecasting of the solar radiation was an important ingredient in the successful integration of renewable energy systems into contemporary power grids. With the growing trend in the world towards an environmentally friendly energy solution, solar energy is one of the most prospective sources because it is abundant and has minimal environmental impact. The intermittent and unpredictable quality of solar irradiance, however, is a serious problem to energy planners, grid operators and to photovoltaic (PV) system designers[1].

Accurate prediction of solar radiation is thus necessary in order to maximize the PV system performance, grid stability, energy storage, and facilitate dependable planning of energy. The time-tested classical forecasting techniques, such as statistical models and physical-based ones, have long been in use over the decades[2]. Although they are simple to implement and interpret, they are frequently difficult to represent the nonlinearities and dynamical variations inherent in the data on solar radiation, especially when weather conditions change rapidly or in an area with complex geographical and climatic variations. The problems related to solar radiation prediction are also enhanced by the fact that the underlying data is

highly fluctuating[3]. The factors affecting solar irradiance are numerous and they include cloud cover, atmospheric aerosols, change in humidity, atmospheric temperature and geography factors like latitude and terrain.

All these cause a lot of nonlinearity, noise and seasonality to solar radiation data sets which in most cases leads to missing or incomplete records[4]. As a result, traditional models may often be less accurate and less adaptable to these circumstances, which explains the necessity of intelligent and adaptive forecasting models that are able to discover complex temporal and spatial patterns using large and heterogeneous data. Artificial intelligence (AI) has become a new method to tackle these issues, providing strong solutions to the nonlinear and high-dimensional character of solar radiation data[5]. The methods of AI, such as Extreme gradient boosting (XGBoost), Extra Tree, random forests, deep learning structures, hybrid models have demonstrated the enormous potential of finding latent patterns and dependencies that are largely neglected by conventional methods. Using historical data and real-time measurements, the accuracy, robustness, and adaptability of these models can deliver more accurate forecasts of energy-related decisions and grid operations to provide improved decision-making[6]. The increased number of studies on this topic indicates the significance of analyzing and comparing AI models in order to determine the most successful and efficient ways to predict solar radiation, which will eventually lead to more trustworthy and efficient renewable energy systems globally.

Motivation and Contributions of the Study

The research has been driven by the rising reliance on solar energy as a clean energy source and the rising concern on precise solar radiation forecasting in order to assist grid reliability, energy storage control and optimization of photovoltaic systems. The time-series data of solar radiation are nonlinear and very noisy, and forecasting them in an effective way is a difficult task due to dynamic weather conditions affecting time-series predictability through conventional statistical and physical models. In addition, seasonal changes and atmospheric disturbances also contribute to the variability that increases the problem of forecasting accuracy. Machine learning models that are based on artificial intelligence have high potential to discover more complex correlations and temporal variations directly on data, which makes them more reliable and adaptive in forecasting solar radiation. The significant contributions of the study are the following:

Used a publicly available high-resolution dataset of solar radiation to be sure that it is reproducible and has real-world relevance; the available data included meteorological variables and temporal variables.

Conducted extensive data preprocessing, such as failures to record data, outliers, time-wise feature extraction and feature engineering to boost model strength.

Applied and tested several machine learning algorithms, such as KNN, Extra Trees, Random Forest and XGBoost, within a single experimental environment.

Performed an in-depth comparative study of performance based on MAE, MSE, RMSE, and R2.

Established that instance-based and ensemble models are more effective compared to traditional methods in the nonlinear dynamics of solar radiation.

Structure of the Paper

The paper is structured as follows: Section II discusses the recent advancements. Section III explains the dataset, preprocessing techniques, and machine learning models used. Section IV presents model evaluations and comparative performance analysis. Finally, Section V summarizes key findings and proposes directions for future research.

Literature Review

Recent works use machine learning, ensemble, deep learning and hybrid models to predict solar radiation, with enhanced accuracy, they however lack unified benchmarking, generalizability analysis, and balanced performance complexity assessment.

Kaplan and Kaplan (2026), offer three novel ML models for making GSR predictions using GSR data and comparing their performance to other GSR prediction models commonly used in the literature. Support Vector Regression (SVR), Robust Linear Regression (RLR), and Gaussian Process Regression (GPR) techniques were employed in the development of the prediction models. ML was used in the MATLAB software to create the new models in this investigation. The analysis of variance (R2) was used

to compare the obtained results. The GPR technique (R2: 0.85) beat all examined models in terms of accuracy, according to the results on all employed statistical indicators[7].

Vijay Babu *et al.* (2025), study proposes a comprehensive data-driven framework for solar energy forecasting using multiple machine learning (ML) techniques, including Multiple Linear Regression, Ridge, Lasso, Decision Tree Regression, Support Vector Regression, and ensemble-based models such as Random Forest, AdaBoost, Bagging, and Gradient Boosting Regressors. Historical solar power and weather datasets were used to train and evaluate the models across multiple performance metrics. Among the models, the Gradient Boosting Regressor demonstrated the best performance, achieving an R2 of 0.827, RMSE of 399.44, and MAE of 253.62, marking a significant improvement over baseline models[8].

Zhang *et al.* (2025), propose a multivariable solar radiation prediction model based on TVFEMD, FE, RF, TDCS, and Pyraformer algorithms. The number of data sequences is reduced via fuzzy entropy-based aggregation, which, when combined with different features, creates a multivariable input feature matrix. The TDCS-RF-TVFEMD-FE-Pyraformer multivariate model's prediction metrics are examined in this study in comparison to nine other multivariate benchmark models. TVFEMD, FE, and RF boost models correctness, according to the results. Pyraformer's RMSE and MAE exceed the baseline models by 10% to 50% after TDCS tuning, while R and SMAPE also outperform the baseline models[9].

Tanoli *et al.* (2024), examines the connection between solar radiation output characteristics and input parameters such as date, temperature, pressure, precipitation, and aerosol. The CERES dataset, which spans the years 2001 to 2021, provided the data used in this investigation. The study takes into account thirty-seven glaciers in Gilgit and eight glaciers in KPK. With an MSE of 598.326, MAE of 18.9685, nRMSE of 0.06973, and R2 score of 0.916399, the findings for the KPK location show that the FFNN method had the best accuracy. With an MSE of 738.78, MAE of 20.6887, nRMSE of 0.08071, and R2 score of 0.886703, the FFNN algorithm also fared better than other models for the Gilgit location[10].

Sevas *et al.* (2024), advances knowledge and optimization of solar irradiance prediction by providing a thorough strategy that combines machine learning, ensemble techniques, and XAI. have further contributed by creating an autoML tool based on XAI and Ensem-ble. have verified the results using the low-code PyCaret machine learning program and found that, of all the techniques, lightGBM has demonstrated the most promising outcomes in terms of sun irradiance prediction. Superior performance was demonstrated by ensurable machine learning boosting algorithms, particularly LightGBM and CatBoost, which demonstrated amazing accuracy and achieved high R2 scores of 0.91[11].

Mishra *et al.* (2023), investigate automatically developing site-specific prediction models using machine learning to generate solar radiation from meteorological station weather forecast reports. Depending on the features of the solar PV system being used, the corresponding solar power output can be calculated from the predicted solar radiation. Improving forecast accuracy is the difficulty. With R2 values of 0.809494 and 0.645419, respectively, ensemble techniques like random forest (RF) and extreme gradient boosting (XGBoost) outperform most models in the field of solar energy prediction by improving stability and combining multiple machine learning models to reduce variation and bias[12].

The table I provides a summary of datasets, models, performances, limitations, and gaps among studies with a lack of standard comparisons frameworks of evaluating artificial intelligence models in solar radiation forecasting.

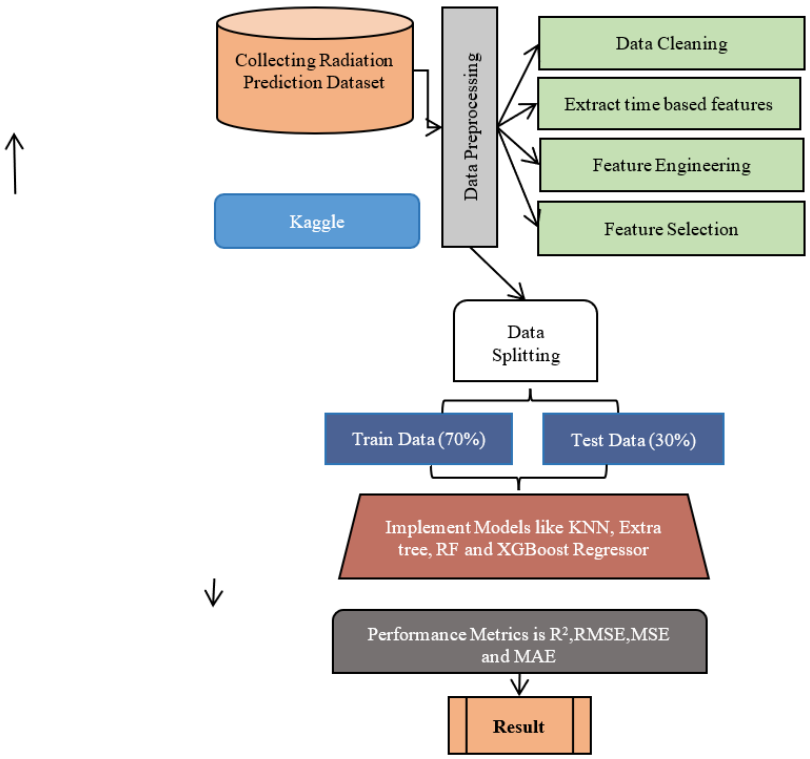
Summary of Related Work on Solar Radiation Forecasting using MI

Reference	Data Source	Models Used	Best Performance	Key Strengths	Limitations
Kaplan and Kaplan, (2026)	GSR sensor data	SVR, Robust Linear Regression, Gaussian Process Regression	0.85 (GPR)	Demonstrates superiority of probabilistic GPR for nonlinear GSR patterns	Limited to traditional ML models; no ensemble, deep learning, or feature engineering analysis
Vijay Babu et al., (2025)	Historical solar power + high-resolution meteorological & solar geometry features	MLR, Ridge, Lasso, DT, SVR, RF, AdaBoost, Bagging, Gradient Boosting	0.827 (GBR)	Strong feature engineering improves forecasting accuracy	No uncertainty quantification; deep learning models not explored

Zhang et al., (2025)	Multivariate meteorological data	TVFEMD, FE, RF, TDCS optimization, Pyraformer Transformer	Not explicitly stated (RMSE & MAE improved 10–50%)	Advanced signal decomposition and attention-based transformer	Extremely complex pipeline; high computational cost; poor interpretability
Tanoli et al., (2024)	Temperature, pressure, precipitation, aerosol, date	FFNN, other ML models	0.916 (FFNN)	High accuracy across multiple geographic regions	Focused on glacier regions only; no ensemble or XAI analysis
Sevas et al., (2024)	Solar irradiance + meteorological data	LightGBM, CatBoost, PyCaret AutoML	0.91 (LightGBM)	Combines explainability and AutoML for model selection	Relies on low-code tools; limited manual model optimization
Mishra et al., (2023)	Weather station forecast reports	Random Forest, XGBoost	0.81 (RF), 0.65 (XGBoost)	Site-specific modeling improves local accuracy	Lower performance than recent ensemble/deep models

Methodology

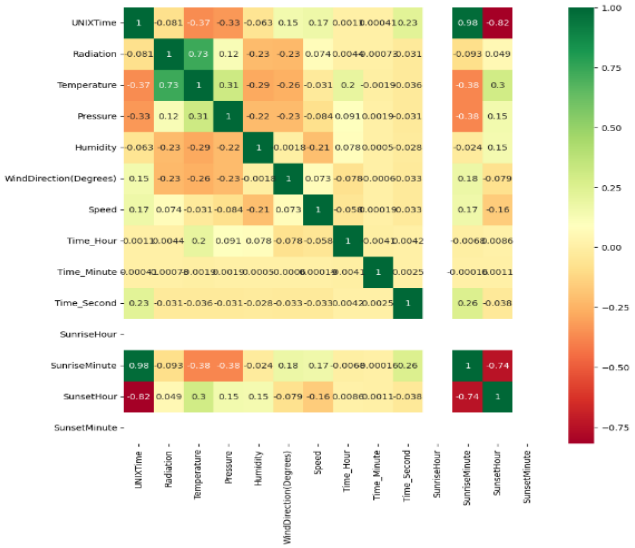
The methodology that is proposed starts with gathering a radiation prediction dataset, but the source of this data is Kaggle. The dataset is subjected to data preprocessing, which involves cleaning it to eliminate inconsistencies and missing values, extracting time-related features to provide insights into the temporal trends, and engineering features that are meaningful to provide insightful variables, and the selection of features which are the most significant to use in modeling. The dataset is divided into training (70%) and testing (30%) sets to assess the model performance in an effective way after preprocessing. Different machine learning algorithms, such as K-Nearest Neighbors (KNN), Extra Trees, Random Forest (RF), and XGBoost Regressor are used to estimate the radiation. Measures used to determine model performance include R2, RMSE, MSE and MAE which are used to measure the accuracy of the prediction. Lastly, the findings are discussed to identify the best model that could be used to predict radiation precisely. The fig. 1 presents the flow of the methodology.



Proposed Methodology for Solar Radiation Forecasting using AI Models

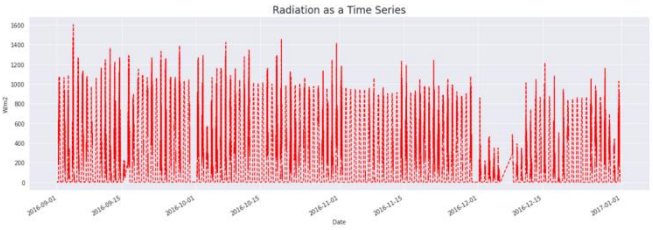
Dataset

The dataset utilized in this paper is the Solar Radiation Prediction Dataset1 from Kaggle. It also contains the meteorological parameters like temperature, humidity, atmospheric pressure, and timestamps that are applicable to forecast the patterns of the solar radiation. It includes time-series type of data of different meteorological values that are important to predict the solar energy, including solar radiation, temperature, humidity, pressure, wind speed and direction, and the time-related variables, like date, time, sunrise, and sunset. The data is covered between September 2016 and January 2017, and makes high frequency observations at fixed intervals, which is optimal in short and long-term trend analysis. The correlation between features is shown in the Fig. 2.



Correlation Heatmap

This heatmap shows Pearson correlation coefficients of Radiation Prediction Dataset, which can be summarized as the linearity of solar and meteorological variables in fig. 2. Radiation and Temperature have shown the highest positive relationship (0.73) and it is true that the higher the levels of the Sun the more the heat. On the other hand, Radiation is negatively correlated with Humidity (-0.23) and Wind Direction (-0.23). It is noteworthy that UNIXTime has a very high correlation with SunriseMinute (0.98), whereas it has a high negative correlation with SunsetHour (-0.82), which can probably be attributed to the seasonal variations in daylight. Some of the features like Time_Minute, and Time hour have near-zero values meaning that they have no or minimal linear relationships with the radiation levels.

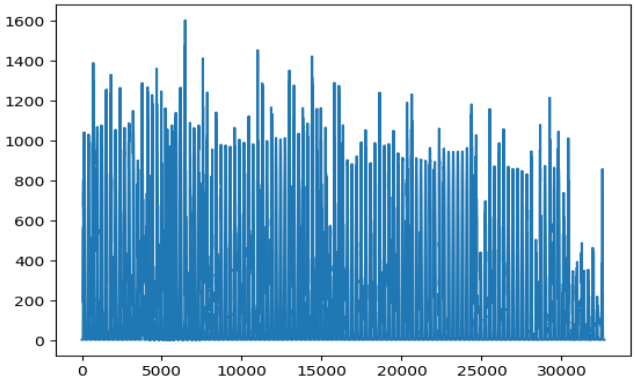


Time Series Plot of Radiation

Fig. 3 depicts the solar radiation levels (in W/m2) during the period of September 2016 and the beginning of January 2017.The picture shows a very oscillatory nature of solar radiation, which is a natural diurnal cycle with the maximum values of solar radiation in daytime and the minimum values equal to zero in the evening. Peak values are often over 1200 W/m2 with the highest recorded in mid-September at

[1] https://www.kaggle.com/datasets/dronio/SolarEnergy/data

around1600 W/m2. There is an observable negative tendency of increase in peak intensity throughout the year to December as is typical of the seasonal variation. It is interesting to note that there is a huge hole or time of no activity during the early months of December indicating that there was a lot of cloud cover or the lack of data.



The Plot for Radiation

Fig. 4 gives a time-series plot of the Radiation Prediction Dataset of the solar radiation intensity (in W/m2) in September 2016 and early January 2017, which indicates a very oscillatory diurnal pattern, with high radiation intensities during the day and zero radiation levels during the nights. Often the peaks are more than 1200 W/m2 and the maximum recorded has been about 1600 W/m2 in mid September. One can see a gradual decrease in the height of the peaks due to a change in the season to winter. It is notable that there is a major decline in data or shift at the very beginning of December, which may refer to extreme weather or the breakdown of the sensors.

Dataset Preprocessing

The pre-processing of data sets is an important step in the data analysis pipeline as it is necessary to convert raw data into a clean and organized format that can be utilized in the machine learning models. The process of preprocessing follows:

Data Cleaning

Data cleaning is applied to improve the quality of the data by fixing the errors and bringing uniformity in the data. It eliminates noise, unreliable records that can influence the learning of the model hence enhancing the robustness and forecasting accuracy of the solar radiation forecasting model.

- **Handling Missing Values:** Missing values are managed with the help of correct representation techniques, including the removal or statistical imputation to ensure the completeness of the data.
- **Dropping Irrelevant Columns:** Non informative or redundant attributes are dropped in order to diminish the dimension and enhancing the efficiency of the model.
- **Eliminating Outliers:** Outlier or extreme values are identified and eliminated so as to avoid bias and enhance the stability of the model.

Extracting Time-Based Features

Solar radiation time-based feature extraction is done to reflect time-related features in solar radiations. Some of the attributes like the hour, day, month, and season are gotten out of the timestamp data. These characteristics assist machine learning models to learn daily and seasonal changes in solar irradiance to a large extent enhance prediction on time sensitive solar radiation data.

Feature Engineering

The feature engineering boosts the predictive capability of the data by either transforming the existing variables or introducing new valuable features. Statistical transformations, interaction features and domain specific enhancers are used to enhance data representation. This step allows the machine learning models to improve the nonlinear correlation among the meteorological parameters and the output of solar radiations.

Feature Selection

The feature selection is a method that determines the most pertinent input variables that affect solar radiation. Correlation analysis together with ranking of importance is used to remove redundant and weak features. The dimensionality reduction aids in enhancing the calculation efficiency, overfitting and overall performance and generalization of the predictive models.

Data Splitting

A processed dataset is split into training and testing subsets to allow the assessment of the model without any bias. A common ratio is 70:30 in terms of training and testing respectively. This division is what enables the models to be taught by past trends and tested on hidden information in order to determine the accuracy of prediction.

Proposed Models

- **KNN Model**

K-Nearest Neighbors (KNN) regressor is an instance based learning algorithm that is a non-parametric model that estimates the target value of a datum by comparing it to the nearest similar datum in the feature space[13]. KNN does not need to acquire an explicit model in the course of the training process: the full set of training data is stored, and the computation is only performed at prediction time. Given a particular test point, the algorithm calculates the distance of the test point to each training point, the Euclidean distance is usually used shown by eq 1:

$$d(x_i,x_j)=\sqrt{\sum_{ik=1}^{n}(x_{ik}-x_{ik})^2} \quad (1)$$

The model then selects the K closest neighbors and then estimates the predicted solar radiation value as the average of their target values given in eq 2:

$$\hat{y}=\frac{1}{K}\sum_{i=1}^{K}y_i \quad (2)$$

- **Proposed Extra Trees**

Extra Trees Regressor is also an ensemble learning algorithm, which relies on a set of decision trees, much like Random Forest, and with an increased level of randomness[14]. As compared to Random Forest, Extra Trees does not attempt to find the best split threshold on any given feature, but rather it picks random split thresholds on each feature which greatly decreases variance and costs involved in computing them.

The output of each of the decision trees is predicted as and the prediction of the final decision tree is found by averaging the predictions of all the trees presented in equation 3:

$$\hat{y}=\frac{1}{T}\sum_{t=1}^{T}\hat{y}_t(x) \quad (3)$$

Where T is the total numbers.

- **Proposed Random Forest**

Random Forest Regressor (RFR) is an ensemble prediction method that has been demonstrated to be effective in a number of classification and regression tasks since it uses multiple decision trees to come up with a final prediction. In addition to this, it enhances the overall performance of classifier by making a random choice of data nodes to create the decision tree[15]. The feature space is divided into L regions represented as RL by the decision tree. This feature space is used in predicting the final decision of a decision tree which can be formulated mathematically as (4) and (5) and the final outcome of the prediction is based on the majorities of all the trees.

$$f(x)=\sum_{i=1}^{L}\text{constant}_L * \Pi(x,R_L) \quad (4)$$

$$\Pi(x,R_L)=\begin{cases}1 & if\ x\epsilon R_L\\ 0 & otherwise\end{cases} \quad (5)$$

Random Forest minimizes overfitting by decorrelating single trees and enhances prediction accuracy thus being highly suitable in modeling.

- **XGBoost Regressor**

XGBoost is a robust gradient boosting model that constructs trees in a serial manner with each succeeding tree trying to address the errors that the last ensemble did[16]. It helps in maximizing a smoothed version of an objective function which allows the trade-off between the predictive quality and the complexity of the model.

The general objective functions is described in the equation 6:

$$L=\sum_{i=1}^{N} l(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k) \quad (6)$$

XGBoost proves to be the most suitable model when it comes to solar radiation prediction because it is able to capture complex patterns of nonlinearity, manage missing values internally, and offers superior predictive capability with small training efforts.

- **Performance Metrics**

The evaluation parameters are critical instruments of assessing the performance of regression models based on quantification of the differences between the predicted and actual values:

- **Mean absolute error (MAE):** MAE is a type of significant metrics to evaluate the regression models. If $\hat{y}_i$ is the predicted value of the ith sample and yi is the corresponding true value, then the MAE can be computed from the following Eq 7:

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \quad (7)$$

- **Mean Squared Error (MSE):** It is a commonly used evaluation metric in regression analysis that measures the average of the **squared** differences between actual values and predicted values[17]. MSE calculated as follows in equation 8:

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \quad (8)$$

- **Root mean squared error (RMSE):** It is applied to the differences between the values that a model forecasts and the values that are actually observed[18]. Stated differently, these individual differences are referred to as the residuals and the RMSE consolidates them into a single metric of predictive capability as presented in eq(9).

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \quad (9)$$

R2 score: R2 score is one of the most popular and standard measurements to test the regression models. Assuming that yi is the predicted value of the jth sample, y_i is the actual value, the following equations are obtained. The formulae (10) used to compute the R2 score value are as follows:

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2} \quad (10)$$

Here, y indicates the true value, $\hat{y}$ indicates predicted value, $\bar{y}$ indicates the average of all the true values.

The metrics will give a complete understanding of model performance to make objective comparisons and decisions made in predictive analysis.

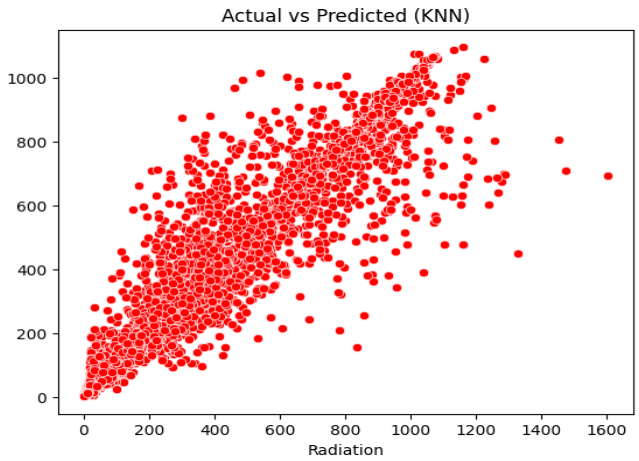
Results & Discussions

This section involves an overall assessment and comparison of various machine learning models in forecasting solar radiation during different meteorological conditions. The experiments were carried out on a system that has a minimum of an Intel Core i5/AMD Ryzen 5 processor, 8GB RAM (16GB preferable), sufficient storage, and optional NVIDIA graphics card. The virtual engine consisted of windows or Ubuntu, python 3.8 and above, and its libraries, scikit-learn, TensorFlow/Keras and visualization software. A few forecasting models have been tested on the basis of MAE, MSE, RMSE, and R2 score: Linear Regression (LR), Support Vector Regression (SVR), Multi-layer Perceptron (MLP), Random Forest (RF), Extra Trees, XGBoost and K-Nearest Neighbors (KNN). The outcome shows that nonlinear models, instance-based models, and ensemble models are always effective to explain the complicated patterns of

solar radiation data as compared to linear models. KNN showed the most excellent overall results with the smallest error values (MAE: 26.14, RMSE: 73.21) and the highest value of the R^2 of 0.946 and was closely followed by Extra Trees and RF, which showed good predictive power and robustness.

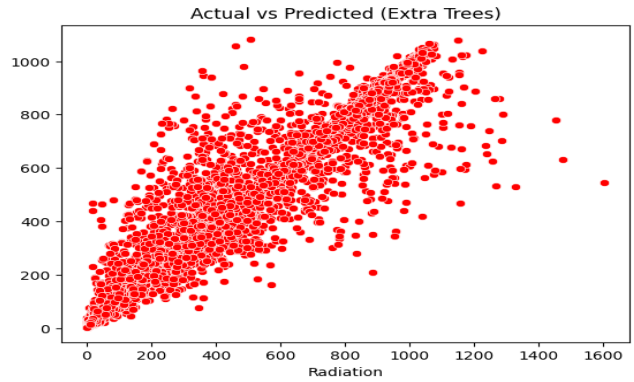
Performance of the Proposed Techniques for Solar Radiation Forecasting

Models	MAE	MSE	RMSE	R^2 score
KNN	26.1359	5359.95	73.2117	0.94619
Extra Trees	30.3336	6283.82	79.2705	0.93691
RFR	34.0253	6941.91	83.3181	0.93030
XGBoost	33.1510	7888.78	88.8188	0.92080



Solar Radiation Actual Predicted Value of KNN Model

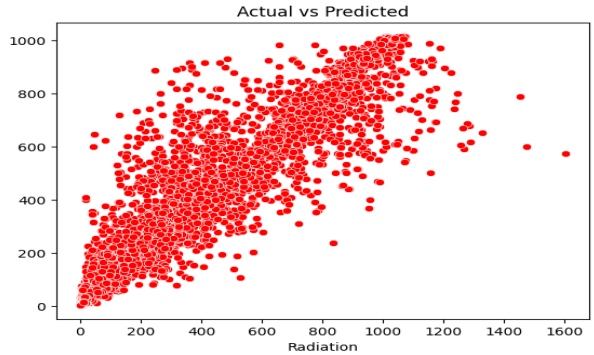
Fig. 5 is an assessment of K-Nearest Neighbors regression model performance in predicting solar radiation. The x-axis is the ground-truth values of the radiation and the y-axis displays the predictions of the model. The data points are tightly clustered in a vertical, diagonal form meaning that there is high positive correlation and a moderately high level of accuracy. The dispersion however increases with high radiation levels implying that the model suffers a lot of variance in forecasting peak solar intensity. Although the KNN model shows the general trend, there are few outliers that can be noticed where the actual radiation is greater than 1400 W/m2 but the prediction is much lower. In general, this visualization supports the idea that ML algorithms can be effectively used to predict the relationship between meteorological characteristics and the amount of radiation produced, but additional adjustments might be necessary to consider severe atmospheric conditions.



Solar Radiation Actual Predicted Value of Extra Tree Model

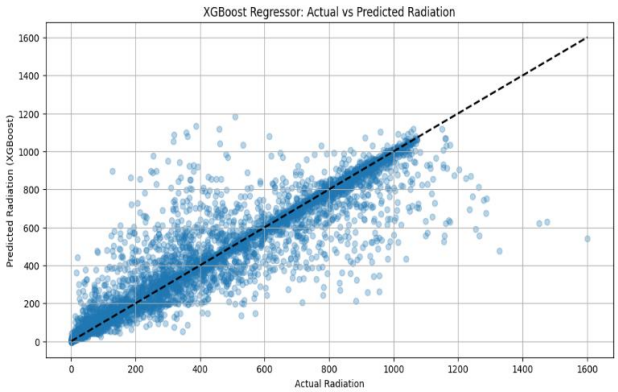
Fig. 6 shows the performance of an Extra Trees Regressor model in the scenario of solar radiation prediction. The x-axis is used to indicate the actual radiation distribution, which was the ground-

truth (Actual), and the y-axis indicates the result of the model. There is a very high positive correlation between the data points that are closely clustered about an upward trajectory which gives the indication that the model is effective in capturing the major patterns of the sun. Nevertheless, like the KNN model, the variance is high at higher levels of radiation, and specifically above 800 W/m2, the model will underestimate peak solar intensity. Although there are outliers at the very high side of the scale, the close clustering of the lower and middle values display that the Extra Trees ensemble technique is a very efficient machine learning strategy to model a complex, non-linear correlation of meteorological factors and solar energy production.



Solar Radiation Actual Predicted Value of Random Forest Model

Fig. 7 displays the accuracy of a random forest regressor (RFR) model to predict solar radiation. The horizontal axis shows the values of the ground-truth of the radiation and the vertical axis shows the predictions of the model. The visualization indicates that the data points are heavily linearly focused indicating that the ensemble-based RFR model is able to represent the latent trends of solar intensity. Although low to mid-range predictions are very precise, the scattering is evident towards the 1000 W/m2 radiation levels. At these intensities, the model tends to give values lower than the real ones, which is usually the case with regression tasks in extreme atmospheric conditions. Altogether, the RFR model is very reliable and offers a strong frame of reference to map the sophisticated meteorological conditions to the solar energy output.



Solar Radiation Actual Predicted Value of XGBoost Model

Fig. 8 shows the results of an XGBoost machine learning model in solar radiation prediction. The x- axis is the ground-truth radiation values (W/m2) and the y-axis displays the model prediction. The data points are largely clustered around the identity line dashed, which represents a strong positive relationship as well as a high predictive power of most of the data. The model is very effective to simulate the mid-range of the sun intensity, but can see that it has a visible dispersion in the extreme high end of the scale especially when the value is above 1200 W/m2. Although there are a few outliers of these and minor underestimates of the peak radiation, the general consistency supports the argument that gradient boosting is an effective method used to address the non-linear complexities of solar data.

Comparative Analysis & Discussion

The performance of the different machine learning and deep learning models to predict solar radiation using the R2 metric has been compared in the table. GRNN and MLP neural networks models have moderate predictive power (R2 0.83 and 0.78 respectively) which implies that they are not capable of establishing complicated nonlinear correlations. Model tree-based representation shows better results with a score of R2 of 0.9348 of the Decision Tree. AE-BiGRU model also gives a better forecasting power which has an R2 of 0.901 which indicates the superiority of hybrid deep learning. The KNN model has the highest R2 of 0.94619 which shows that it is more accurate than all the other models assessed. Extra Trees, Random Forest Regressor, and XGBoost are also ensemble-based methods, which show a competitive performance, which proves the efficiency of the ensemble and instance-based learning approaches in predicting solar radiation.

Performance Comparison of Different Models for Solar Radiation Forecasting

Models	R2
GRNN[19]	0.83
MLP[20]	0.78
Decision Tree[21]	0.9348
AE-BiGRU[22]	0.901
KNN	0.94619
Extra Trees	0.93691
RFR	0.93030
XGBoosts	0.92080

The experiment suggests clearly that machine learning models can be extremely useful in forecasting solar radiation, especially models that are able to model nonlinear and local trends. KNN model demonstrated the best prediction accuracy, and this indicates that instance based learning could be effective in localizing variation in meteorological conditions. Extra Trees and random forest are also ensemble techniques, and they both showed high and stable performance, with lower variance, and higher generalization. Nevertheless, there was a general trend of all models to predict with errors higher at extreme radiation levels, which showed difficulty in operating in unusual atmospheric conditions. On the whole, the findings do support the fact that nonlinear and ensemble learning methods are better than the traditional model predictors of solar radiation.

Conclusion & Future Work

Artificial intelligence-based machine learning algorithms used to predict solar radiation has been conducted. The findings indicate that learning models which can learn nonlinear relationships especially instance-based and ensemble methods provide significantly better accuracy as compared to classical regression models. K-Nearest Neighbors model turned out to be the best predictor and Extra Trees and Random Forest regressors have a good and consistent predictive power. These results underscore the success of data-driven AI based on the complexity of meteorological interactions of data with solar irradiance and provide viable returns of improving photovoltaic efficiency, grid stability, and energy management plans. However, there are various issues that have not been addressed. The use of model performance is likely to degrade under conditions of extreme radiation values prediction and the predictor (training) data are limited in both time span and geographic variation. Further research in this area should focus on the application of large, multi-location studies to enhance the generalization and strength. The implementation of more advanced deep learning models (e.g. LSTM, GRU, transformer-based) may help to reinforce the learning of temporal features. Furthermore, explainable AI models would enhance interpretability and probabilistic prediction methods may give uncertainty-sensitive forecasts. Another promising direction of the further improvement of the accuracy of solar radiation forecasting is the development of hybrid systems that would combine physical modeling and machine learning.

References

1. Y. Kashyap, A. Bansal, and A. K. Sao, “Solar radiation forecasting with multiple parameters neural networks,” *Renew. Sustain. Energy Rev.*, vol. 49, pp. 825–835, 2015.
2. E. Chodakowska, J. Nazarko, Ł. Nazarko, H. S. Rabayah, R. M. Abendeh, and R. Alawneh, “ARIMA models in solar radiation forecasting in different geographic locations,” *Energies*, vol. 16, no. 13, p. 5029, 2023.

3. S. Shaw and M. Prakash, “Solar radiation forecasting using support vector regression,” in *2019 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, 2019, pp. 1–4.
4. D. Cannizzaro, A. Aliberti, L. Bottaccioli, E. Macii, A. Acquaviva, and E. Patti, “Solar radiation forecasting based on convolutional neural network and ensemble learning,” *Expert Syst. Appl.*, vol. 181, p. 115167, 2021.
5. C. N. Obiora, A. N. Hasan, A. Ali, and N. Alajarmeh, “Forecasting Hourly Solar Radiation Using Artificial Intelligence Techniques,” *IEEE Can. J. Electr. Comput. Eng.*, vol. 44, no. 4, pp. 497–508, 2021, doi: 10.1109/ICJECE.2021.3093369.
6. A. Mellit, “Artificial Intelligence technique for modelling and forecasting of solar radiation data: a review,” *Int. J. Artif. Intell. Soft Comput.*, 2008, doi: 10.1504/ijaisc.2008.021264.
7. Y. A. Kaplan and A. G. Kaplan, “Development of new solar radiation prediction models using different machine learning techniques,” *Electr. Power Syst. Res.*, vol. 251, p. 112343, 2026, doi: https://doi.org/10.1016/j.epsr.2025.112343.
8. A. R. Vijay Babu, N. Bharath Kumar, R. Patnaik Narasipuram, S. Periyannan, A. Hosseinpour, and A. Flah, “Solar Energy Forecasting Using Machine Learning Techniques for Enhanced Grid Stability,” *IEEE Access*, vol. 13, pp. 93735–93754, 2025, doi: 10.1109/ACCESS.2025.3574093.
9. I. K. Tanoli *et al.*, “Machine learning for high-performance solar radiation prediction,” *Energy Reports*, vol. 12, pp. 4794–4804, 2024, doi: https://doi.org/10.1016/j.egyr.2024.10.033.
10. M. S. Sevas, N. Sharmin, C. F. T. Santona, and S. R. Sagor, “Advanced ensemble machine-learning and explainable ai with hybridized clustering for solar irradiation prediction in Bangladesh.,” *Theor. \& Appl. Climatol.*, vol. 155, no. 7, 2024.
11. D. P. Mishra, S. Jena, R. Senapati, A. Panigrahi, and S. R. Salkuti, “Global solar radiation forecast using an ensemble learning approach,” *Int. J. Power Electron. Drive Syst.*, 2023, doi: 10.11591/ijpeds.v14.i1.pp496-505.
12. Z. Liu and Z. Zhang, “Solar forecasting by K-Nearest Neighbors method with weather classification and physical model,” in *2016 north american power symposium (NAPS)*, 2016, pp. 1–6.
13. M. W. Ahmad, J. Reynolds, and Y. Rezgui, “Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees,” *J. Clean. Prod.*, vol. 203, pp. 810–821, 2018.
14. A. Munshi and R. M. Moharil, “Solar radiation forecasting using random forest,” in *AIP Conference Proceedings*, 2022, p. 50003.
15. X. Li *et al.*, “Probabilistic solar irradiance forecasting based on XGBoost,” *Energy Reports*, vol. 8, pp. 1087–1095, 2022.
16. F. Rehman, Y. Kamal, and S. U. Amin, “The relationship between idiosyncratic, stock market volatility and excess stock returns,” *Public Financ. Quarterly= Pénzügyi Szle.*, vol. 62, no. 3, pp. 311–325, 2017.
17. V. Demir, “Evaluation of Solar Radiation Prediction Models Using AI: A Performance Comparison in the High-Potential Region of Konya, Türkiye,” *Atmosphere (Basel).*, vol. 16, no. 4, p. 398, Mar. 2025, doi: 10.3390/atmos16040398.
18. M. S. Naveed, I. Iqbal, M. F. Hanif, J. Xiao, X. Liu, and J. Mi, “Enhanced accuracy in solar irradiance forecasting through machine learning stack-based ensemble approach,” *Int. J. Green Energy*, pp. 1–24, 2025.
19. Y. Mariappan, K. Ramasamy, and D. Velusamy, “An optimized deep learning based hybrid model for prediction of daily average global solar irradiance using CNN SLSTM architecture,” *Sci. Rep.*, vol. 15, no. 1, p. 10761, 2025.
20. M. Chiranjeevi, S. Karlamangal, T. Moger, and D. Jena, “Solar irradiation prediction hybrid framework using regularized convolutional BiLSTM-based autoencoder approach,” *IEEE Access*, vol. 11, pp. 131362–131375, 2023.

❑❑❑