Transactions on Energy Systems and Engineering Applications

Article

Geographically agnostic machine learning model for forecasting solar irradiance

Anupama R Itagi[1,*], Mrityunjaya Kappali[1], Rakhee Kallimani[2] and Krishna Pai[3]

[1] Department of Electrical and Electronics Engineering, KLE Technological University, Hubballi, Karnataka, India.
[2] Department of Electrical and Electronics Engineering, KLE Technological University Dr. M. S. Sheshgiri Campus, Belagavi, Karnataka, India.
[3] Independent Researcher, Bengaluru, Karnataka, India.
* Correspondence: anupama_itagi@kletech.ac.in

Received: 18 March 2025; Accepted: 07 April 2026; Published: 17 April 2026

Abstract: Solar Photovoltaic (PV) systems are essential in combating the energy crisis. The intermittent nature of solar isolation is a significant setback. Forecasting of solar irradiance offers a viable solution. The Artificial Intelligence (AI) models highlighted in the literature for solar irradiance forecasting are tailored to specific locations. Models designed for specific places cannot accurately predict solar irradiance in other locations. These models employ diverse techniques and are typically evaluated using at least 4 input parameters, thereby increasing model complexity. These limitations are addressed in this work. The authors present a geographically agnostic Machine Learning (ML) model that forecasts solar irradiance using a universal dataset that spans extreme climatic conditions. An ML-based model is developed with time, temperature, and dew factors as input parameters. Time series analysis is performed separately across distinct areas within all climatic zones to categorize solar irradiance data. Different ML algorithms are ensembled to formulate the model. Simulation results show acceptable values of the Root Mean Square Error (RMSE) and the coefficient of determination (R^2), thus validating the performance of the proposed model. The proposed geographically agnostic model distinguishes itself by its innovative approach and promising performance, with a moderate RMSE.

1. Introduction

The depletion of conventional energy sources has become a major concern due to their limited availability. As global energy demand increases, it is necessary to move towards sustainable and renewable energy solutions [1]. Distributed energy resources (DERs), including solar photovoltaic (PV) systems, wind turbines, and small-scale hydroelectric power, offer a reliable alternative to bridge the gap between energy generation and consumption. Among these, solar PV systems stand out as an Indigenous and

How to cite this article: Itagi, Anupama R; Kappali, Mrityunjaya; Kallimani, Rakhee; Pai, Krishna. Geographically agnostic machine learning model for forecasting solar irradiance. *Transactions on Energy Systems and Engineering Applications*, 7(1): 864, 2026. DOI:10.32397/tesea.vol7.n1.864

widely accessible choice due to the numerous benefits they offer [2]. However, one of the main challenges associated with integrating solar energy into the grid is its intermittent and unpredictable nature [3]. This leads to instability in energy supply, posing challenges for grid reliability and energy management systems. To mitigate these issues, accurate forecasting of solar irradiance becomes crucial [4]. By predicting the amount of solar irradiance available at any given time, energy operators can optimize power generation, storage, and distribution, ensuring a balanced and efficient energy supply [5]. One of the most effective approaches to improving solar irradiance forecasting is the use of Artificial Intelligence (AI) models [2]. AI techniques can analyze substantial amounts of real-time meteorological data to produce accurate, low-cost forecasts [6].

Several methods have been reported in the literature for solar irradiance forecasting. The authors in [1] introduce a novel short-term solar power forecasting method with a closed-loop configuration that combines point estimation and range classification. The proposed method uses 14 input features and is evaluated on two real-world datasets with forecasting horizons of 1 hour and 1 day. The study in [2] evaluates the effectiveness of six different ML algorithms in forecasting solar energy. The work focuses on the effectiveness of integrating XGBoost and Random Forest techniques to improve the accuracy of solar power forecasts. The work involves using at least 10 inputs. The work in [3] uses a deep neural network approach combined with geostationary satellite data to improve short-term, three-hour-ahead solar radiation forecasts over Northeast Asia. A key focus was improving the smoothness of the predicted image patterns, predominantly for long-term forecasts, by applying the HRNet model. The total solar insolation incident on a horizontal surface is forecasted using a clearness index model based on the ensemble Kalman Filter (KF). The model is developed by collecting data from 12 locations across India spanning a wide range of latitudes. A comparative study of recursive least squares (RLS) and KF methods is presented in [4]. To reduce the effect of seasonality and improve the accuracy of solar irradiance forecasts, the layering and stacking of weather data clusters are proposed in [5]. Data were collected from the NSRDB in the USA. Various High-dimensional weather data are added to the dataset during training.

A model based on a Long Short-Term Memory (LSTM) hybrid technique is formulated with four input parameters. Data are collected from 4 public places to develop autoregressive and neural network models to forecast solar energy. This article [6] investigates the use of tiny machine learning to forecast in real-time the cost-effective solar energy yield on resource-constrained edge IoT devices, such as microcontrollers, to improve residential and industrial energy management. The study uses a dataset collected from renewable energy facilities in China, which incorporates seven input variables. A novel deep learning model designed for multi-horizon probabilistic forecasting of Global Horizontal Irradiance (GHI) is presented in [7]. The study uses historical GHI data and natural irradiance impact factors to effectively capture correlations and temporal dependencies. The model is designed with two locations in the UK. In [8], a forecasting model is introduced to predict solar irradiance over noticeably short time intervals, specifically 10 minutes ahead. The study uses two ANN-based techniques: LSTM and CNN. Direct Normal Irradiance (DNI forecasting is proposed in [9] using the Variational Mode Decomposition-Whale Optimization Algorithm-Deep Extreme Learning Machine (VMD-WOA-DELM) to enhance plant control and operation. A solar thermal power plant in Qinghai Province serves as the case study for this research. A simple open-loop ANN model for predicting solar insolation is discussed in [10]. The authors used eight inputs to train the model. The multilayer neural network with nine inputs with increased generalization ability is discussed in [11].

The AI models recommended in the literature for solar irradiance forecasting are location-specific. The models developed for a single or a few specific locations cannot forecast solar irradiance with acceptable accuracy at other locations. The models are created using various techniques, and in most cases, performance is evaluated using at least 4 weather parameters as inputs. However, as input increases, the system's complexity and associated costs rise. Hence, there is a need for an approach that develops

a Geographically Agnostic (GA) model to forecast solar irradiance with satisfactory precision while minimizing the number of input variables.

The main contributions of the present work are as follows:

1. Formulation and use of a universal data set to forecast solar irradiance.
2. Development of a geographically agnostic ML model to forecast solar irradiance with acceptable accuracy.
3. Formulation of the above model with minimal inputs. Hence, the number of sensors needed is reduced, and the system's hardware implementation is simpler.
4. Assimilation of different ML algorithms to ensure system performance with satisfactory accuracy.
5. Formulation of a model that can be integrated into grid-level forecasting or IoT-based solar systems.

The paper is organized as follows. Section 2 explains the formation of the universal data set. Section 3 includes the choice of input parameters. Section 4 deals with the proposed method, and the evaluation of the model is detailed in Section 5. Section 6 explains existing challenges, and future work with the conclusion of the work is presented in Section 7.

2. Formation of universal data set

Data collection is one of the primary stages in developing an ML model [12]. It is not just the quality but also the quantity of data that decides the model's performance. The desired data are collected from NASA Power | Data Access Viewer [13]. Each sample of the information consists of observations from various National and State Laboratories worldwide. Weather data included timestamps, temperature at 2 meters, solar irradiance, relative humidity at 2 meters, Dew/Frost Point at 2 meters, and precipitation. Data from twenty-nine different climatic zones are classified into major climatic groups according to the Koppen-Geiger climatic classification [14]. 54,00,960 raw data samples are collected. Data cleaning reduced the data set to 53,29,898. This data is used to create a universal data set. The universal dataset contains solar irradiance and meteorological data from various locations worldwide. Its advantages are as follows: i) Enables the identification of patterns and trends that might be specific to certain regions or climatic conditions, including possible extreme values. It also aids in the appropriate selection of algorithms to build the model; ii) It encompasses data from remote places where the physical measurement of solar irradiance is challenging; iii) Facilitates a more holistic analysis of solar irradiance and its relationship with other weather parameters; iv) Improves the generalization capability of the model; v) Enhance the accuracy and robustness of the forecasting model; and vi) Beneficial for academic research work. The drive behind creating a universal dataset is to build a geographically agnostic ML model to forecast solar irradiance.

3. Choice of input parameters

The different meteorological parameters of the data set are wind speed, temperature, wind direction, relative humidity, dew factor, time, and precipitation [15, 16]. Selecting the right features avoids overfitting, reduces model size, and improves forecast accuracy.

3.1. Analysis of the impact of the time stamp on solar irradiance data

Data collected for solar irradiance forecasting are time-dependent. Analyzing the impact of the timestamp on solar irradiance data is essential for determining whether time-series analysis is suitable for the data.

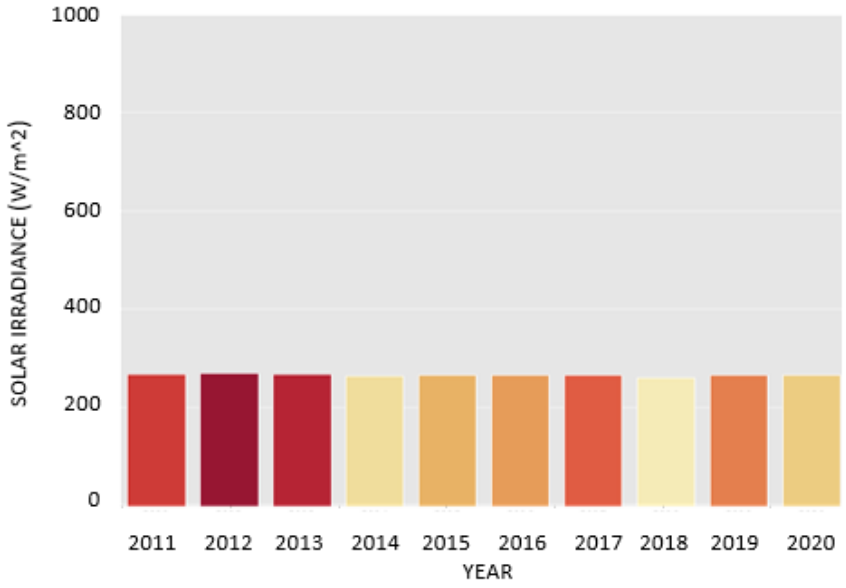

Figure 1. Mean solar irradiance versus year.

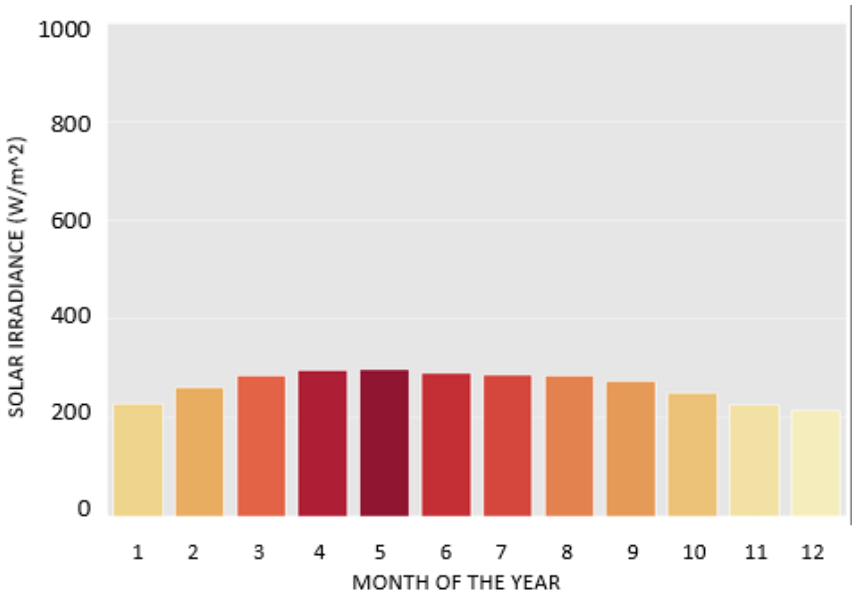

Figure 2. Mean solar irradiance versus months of the year.

The plot of mean solar irradiance versus year is shown in Figure 1. Since there is no notable change in the mean solar irradiance over the year, the seasonal year-to-year seasonality can be ignored. The plot of mean solar irradiance versus month is given in Figure 2. The graph of solar irradiance versus month follows a specific pattern when data from individual climatic zones are considered. Considering the monthly seasonality might deteriorate the model's performance when data from regions near the poles are included. However, examining the global data showed no appreciable pattern and can be ignored. The plot of mean solar irradiance versus the day of the month is given in Figure 3. The data do not follow daily or weekly seasonality.

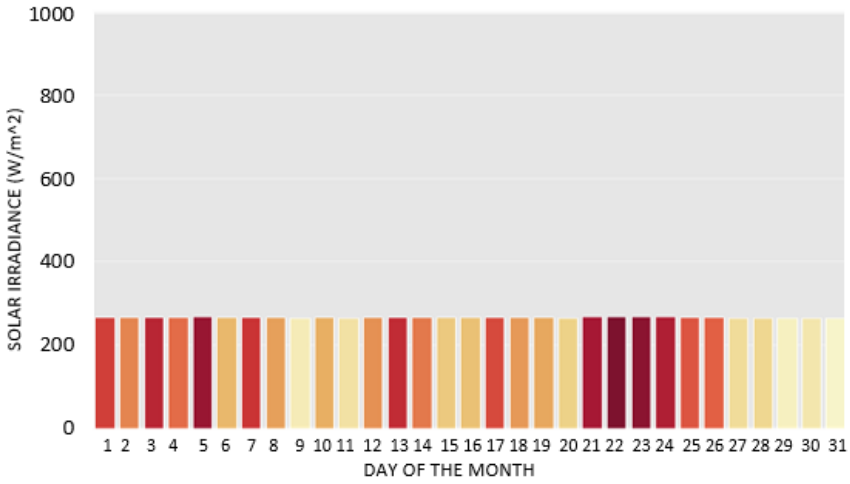

Figure 3. Mean solar irradiance versus day of the month.

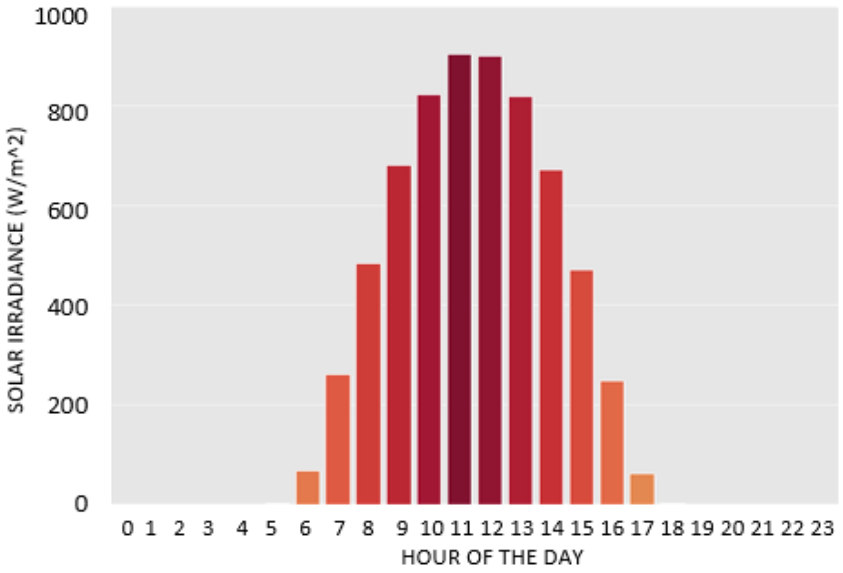

Figure 4. Mean Solar Irradiance versus Hour.

The variation in solar irradiance levels does not show any seasonal pattern. Thus, the day and week factor can be eliminated. The plot of mean solar irradiance versus hour of day is shown in Figure 4, and it is clear that solar irradiance exhibits hourly seasonality. Thus, the hour is the most prominent time stamp in the data set. The year, month, day, and week factors are less significant. Hence, the temporal pattern in the data is primarily driven by the hour of the day rather than longer-term trends or seasonal variations. In such cases, traditional time series algorithms that focus on capturing long-term patterns and seasonality are irrelevant.

3.2. Identification of other Prominent Input Parameters

The Pearson Correlation coefficient, including night-time data, is given in Figure 5. This approach provides a more comprehensive assessment of the relationship between solar irradiance and other input variables, considering the impact of zero values on the correlation. The main climatic parameters that

influence solar irradiance are temperature, dew, and relative humidity. The dew factor and the relative humidity show the same correlation with the solar irradiance. Hence, anyone among these can be considered. Since visibility is one of the essential factors in calculating solar irradiance [17], the dew factor is chosen over relative humidity.

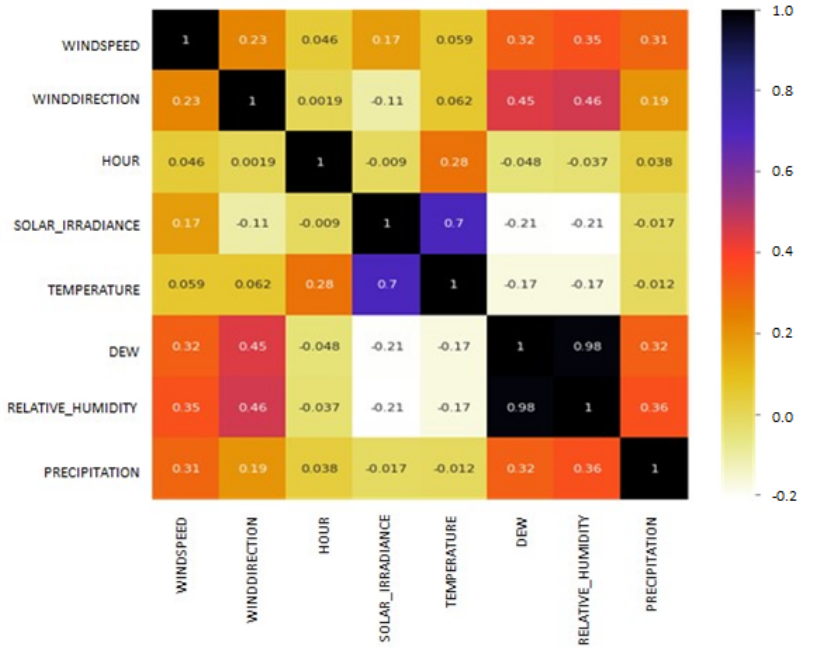

Figure 5. Pearson correlation including night-time data.

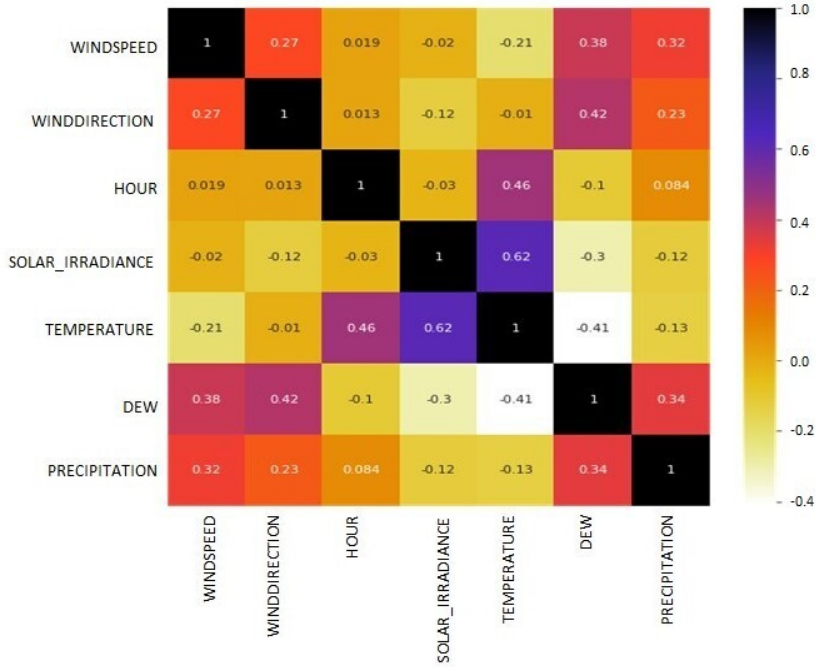

Figure 6. Pearson correlation dropping relative humidity and all night-time values.

The Pearson correlation shown in Figure 5 is obtained by dropping all zero solar irradiance values from the data set. This is because zero values in the data can skew the correlation analysis and yield misleading results. The adopted approach helps capture the relationship between meaningful data points and avoids any distortion caused by zeros. Therefore, it becomes possible to improve forecast accuracy further. In Figure 6, the prominent factors identified are temperature and dew factors. The temperature shows a high correlation of 0.28 with the hour of the day. The mean temperature versus hour graph is shown in Figure 7. During the sunlight hours, the temperature follows a pattern similar to that of solar irradiance. Since solar irradiance also shows consistent, predictable patterns by hour of day, a categorization-based prediction method is proposed, with the hour of day and temperature as the primary inputs.

The parameter selection process is performed meticulously so that augmenting the chosen input variables with a new meteorological variable will not substantially change the accuracy of the results. However, it increases the system's complexity.

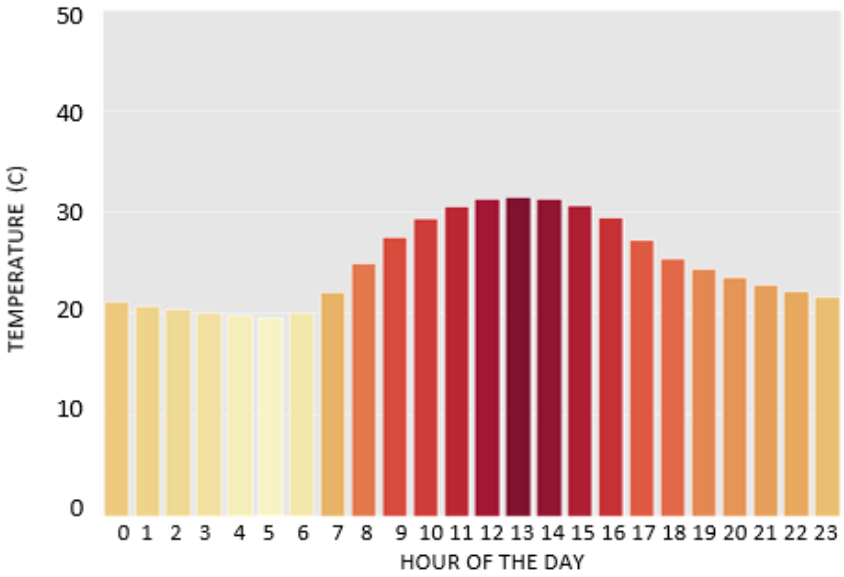

Figure 7. Mean temperature versus hour of the day.

4. Proposed methodology

Figure 8 details the systematic overview of the proposed methodology. The data collection process starts with the global collection of solar irradiance and meteorological data. Data preprocessing is followed by the identification of the main parameters. Then, a comprehensive universal dataset is constructed. The forecast data target is divided into five distinct categories. For each category, an ensemble machine learning model is developed by selecting appropriate sub-models. Thereafter, the sub-models are integrated using a weighted averaging ensemble technique to produce the final solar irradiance forecast. The detailed steps of the proposed approach are outlined below.

4.1. Categorization of solar irradiance data

The first step is to perform time series analysis in a specific area to forecast solar irradiance. The analysis revealed that the data can be grouped into five distinct categories, as shown in Figure 9. This pattern is consistently observed when conducting time-series analysis independently across regions representing different climatic zones. Consequently, the data are categorized into five classes. A novel method is

proposed to determine the appropriate number of categories. The classifications based on solar irradiance are Clear, Overcast, Mild, Moderate, and High [18], as listed in Table 1.

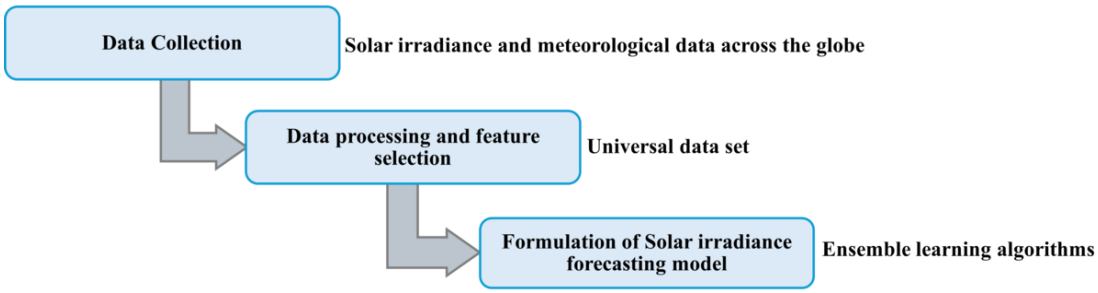

Figure 8. Overview of the proposed method.

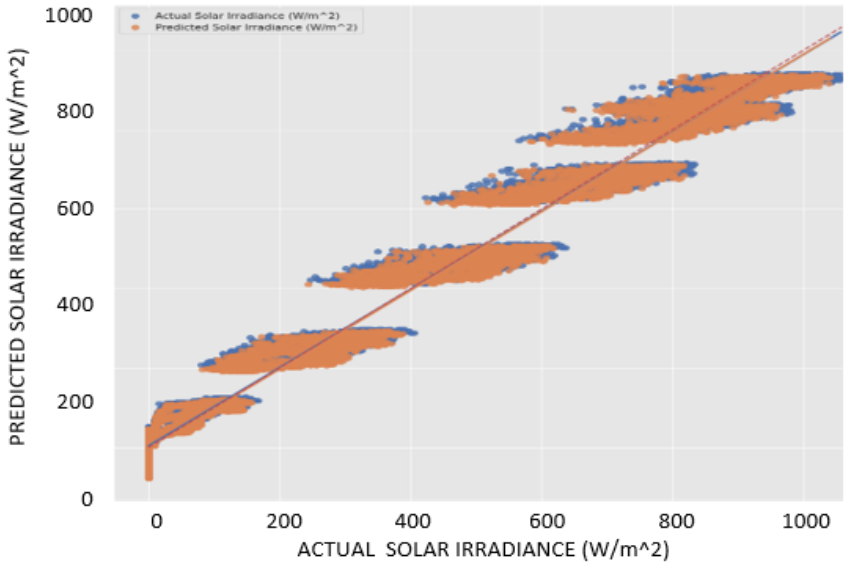

Figure 9. Predicted versus Actual Solar Irradiance.

Table 1. Five class categorisations of Solar Irradiance.

Class number	Class	Range of Solar Irradiance (W/m^2)
0	Clear	0 – 20
1	Overcast	21 – 180
2	Mild	181 – 435
3	Moderate	436 – 795
4	High	796 – 1000

The classification is determined based on the time of day and temperature from the given dataset. Figure 10 displays the confusion matrix for the test dataset, indicating that in certain boundary cases, data can be assigned to adjacent classes. To address this issue, the authors suggest employing an ensemble machine learning algorithm for solar irradiance prediction. The proposed model incorporates time, temperature, and dew point as input features. A total of 4,263,918 data points are used for training, while 1,065,979 are used for model validation. The model is tested exclusively on data from 2021 and 2022.

Figure 10. Confusion matrix indicating various classes.

4.2. Ensembled models

Developing Machine learning models using a single algorithm tends to be simple; however, it has several limitations: (i) inadequate performance while handling large datasets; (ii) a tendency to over-fit in the presence of noisy data; (iii) difficulty in accurately modeling complex relationships among variables; and (iv) limited generalization ability. These challenges can be effectively addressed through ensemble models. Ensemble learning involves combining multiple distinct models (sub-models) to improve overall prediction accuracy. When appropriately integrated, these sub-models produce more robust and dependable outcomes. Ensemble techniques exploit the individual strengths of each sub-model while minimizing their weaknesses [19]. The final prediction is obtained by aggregating the outputs of the sub-models.

4.3. Developing ensemble model

The selection of appropriate ensemble models depends on the specific application and the dataset's inherent attributes [20]. The considered dataset is complex, exhibiting a non-linear relationship between input features and the target variable. Key factors considered for the selection of sub-models for the ensemble: i) The chosen sub-models exhibit unique characteristics and produce diverse predictions [21]; ii) Each model is adept at addressing particular data patterns or challenges; iii) The models are expected to perform reliably across various datasets, even noise presence; iv) The balance between model complexity and performance is maintained [21]; v) The ensemble includes conceptually distinct models to ensure methodological diversity. The algorithm developed for the prediction model is shown below.

Algorithm 1: Modeling pipeline for solar irradiance forecasting

Input: Scaled features X_{scaled}, target y
Output: Ensemble performance metrics (RMSE, R^2)
Step 1: Define Models
Initialize SVM, XGBoost, AdaBoost, GaussianProcess, DecisionTree
Step 2: Train-Test Split
Split X_{scaled}, y into training and testing sets
Step 3: Train Ensemble
for each model **do**
    Fit model on training data
end for
Combine models into VotingRegressor ensemble
Fit ensemble on training data
Step 4: Evaluate
Predict on test data
Compute RMSE, R^2

4.4. Architecture

The architecture of the proposed ensemble modeling framework spanning classes 0 through 4 is depicted in Figures 11 to 15. A Support Vector Machine (SVM) classifier, using the hour of the day and temperature as input features, is used to determine the class label. The SVM outputs a value between 0 and 4, corresponding to classes 0, 1, 2, 3, and 4. For class 0, the predicted output is zero. For classes 1 through 4, the model incorporates the dew point as an additional input to estimate solar irradiance. When the SVM predicts class 1, the model uses an XGBoost Regressor and an Ada regressor as submodels. If the predicted value lies within the predefined range specified in Table 1, it is accepted as the final output. If it exceeds the lower bound, it is reassigned to class 0. Conversely, if it exceeds the upper threshold, the ensemble method for the next higher class (class 2) is invoked. The final prediction is then computed as the average of the outputs from the relevant sub-models.

This methodology is applied consistently across all classes. For class 3, the ensemble includes the AdaRegressor, the Gaussian Process Regressor, and the Decision Tree Regressor. Class 4 utilizes the Gaussian Process Regressor and Decision Tree Regressor. If the predicted value exceeds the maximum allowed limit of 1000 W/m^2, it is capped at that value and returned as the final result.

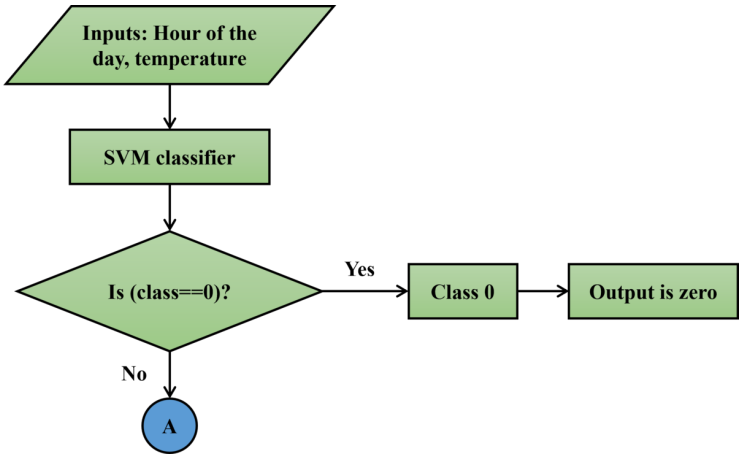

Figure 11. Architecture for the proposed methodology: Class 0.

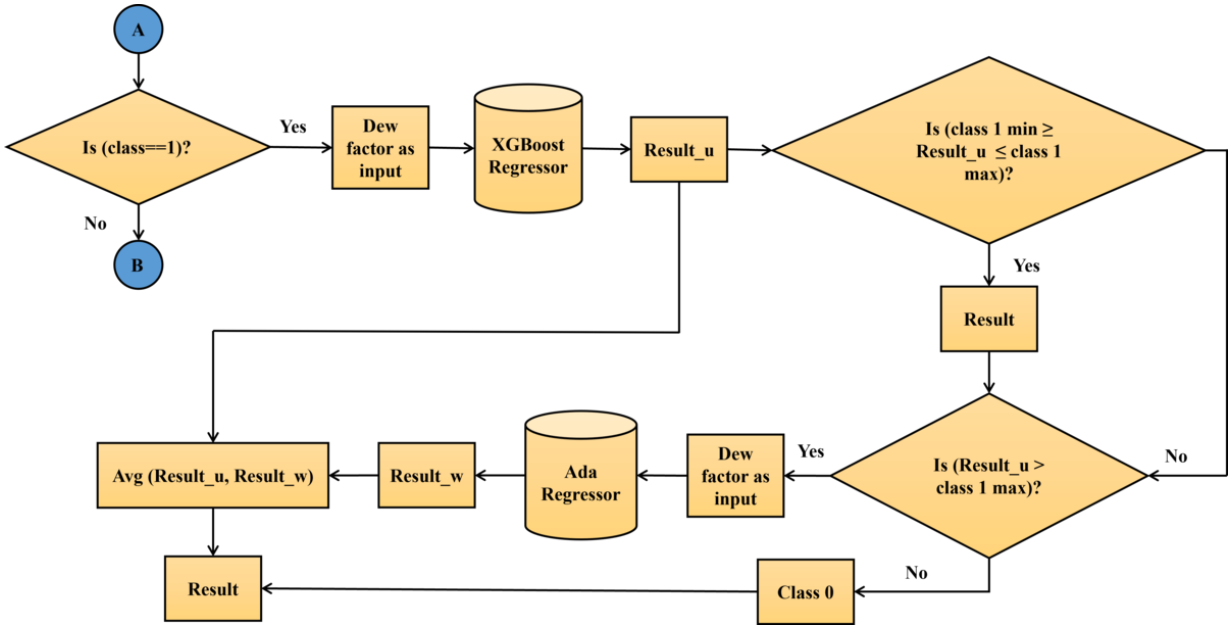

Figure 12. Architecture for the proposed methodology: Class 1.

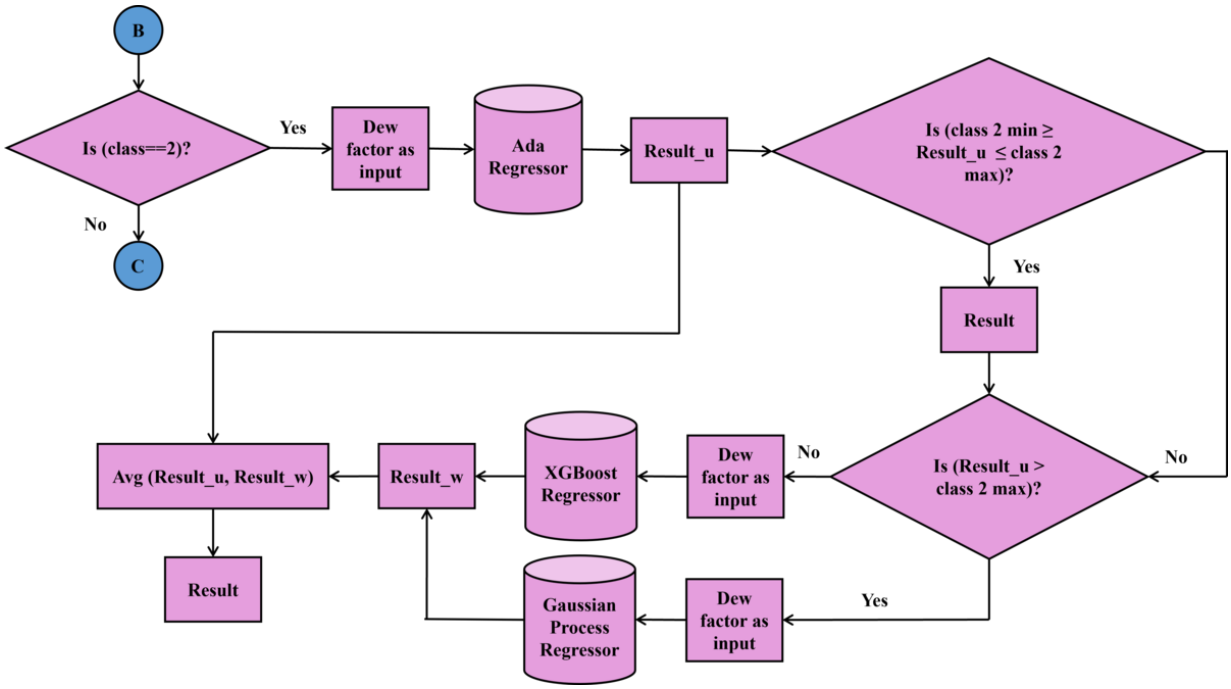

Figure 13. Architecture for the proposed methodology: Class 2.

4.5. Description of sub-models of the proposed ensemble model

The primary task of the model is to manage large-scale, non-linear datasets and classify them into distinct categories as outlined in Table 1. A Support Vector Machine (SVM) classifier is employed due to its ability to handle both linear and non-linear data patterns effectively [22]. For predicting solar irradiance in class 1, the model combines an XGBoost Regressor and an AdaBoost Regressor. XGBoost is particularly

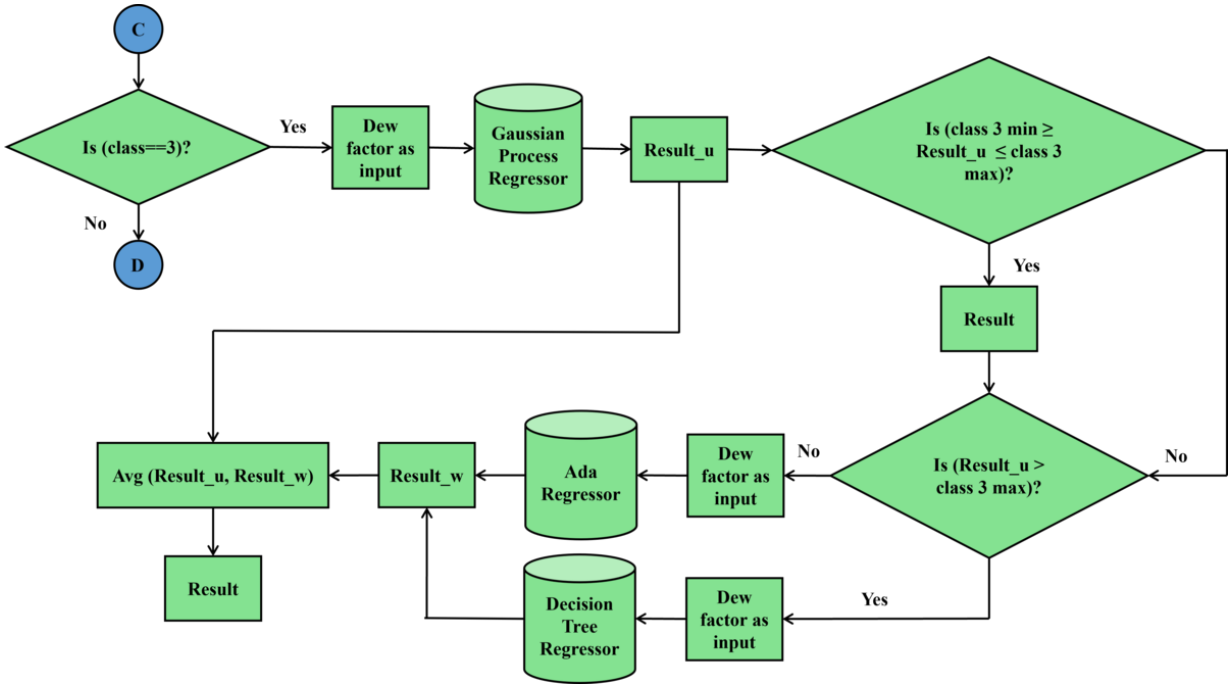

Figure 14. Architecture for the proposed methodology: Class 3.

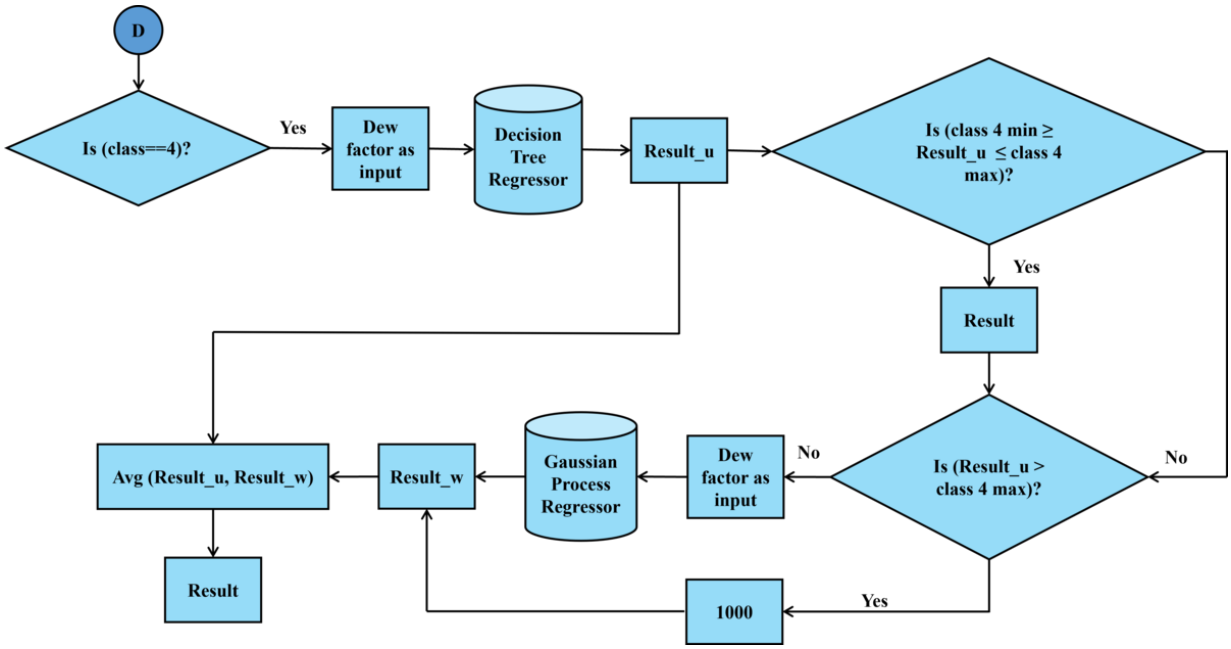

Figure 15. Architecture for the proposed methodology: Class 4.

suited to capturing complex relationships between input and output variables and to efficiently processing large datasets [20]. AdaBoost is used to correct misclassifications, especially those that involve neighboring categories [23]. Given the time-dependent nature of the data, this combination of algorithms improves the model's accuracy, reliability, and generalization. Additionally, Gaussian Process Regression (GPR) is incorporated to construct the predictive model for class 2.

Gaussian Process Regression (GPR) uses prior knowledge to make predictions and can effectively quantify the uncertainty in its outputs [24]. Unlike models based on predefined mathematical functions, GPR identifies relationships between data points based on their similarities, allowing it to adapt to a wide range of data patterns [25]. It is particularly well-suited for handling non-linear datasets and tends to produce predictions that closely align with actual values, as evidenced by its lower Root Mean Square Error (RMSE) [26]. Additionally, GPR helps mitigate the risk of over-fitting, especially when dealing with large and complex datasets [27]. The Decision Tree Regressor is another model incorporated in classes 3 and 4. It excels at capturing non-linear relationships between variables and is highly effective for multivariate analysis. Decision trees accommodate categorical and continuous input and output variables and require minimal data preprocessing compared to other forecasting methods [28].

5. Results and evaluation of the proposed model

The evaluation of the proposed model is crucial to determine its performance and reliability. It provides insights into the model's accuracy and generalization capability to unseen data. It also helps to understand the model's strengths and weaknesses.

5.1. Quality measure

The R^2 measures the model's predictive ability. The value of R^2 can range from 0 to 1. The bigger the number, the better the model performs. For the proposed model, the achieved R^2 value is 0.9455. The RMSE shows how concentrated the data are around the best-fit line. The 5-Fold Cross-Validation Metrics plot is shown in Figure 16. The value of RMSE is 1.71. The lower the RMSE, the better the model's performance.

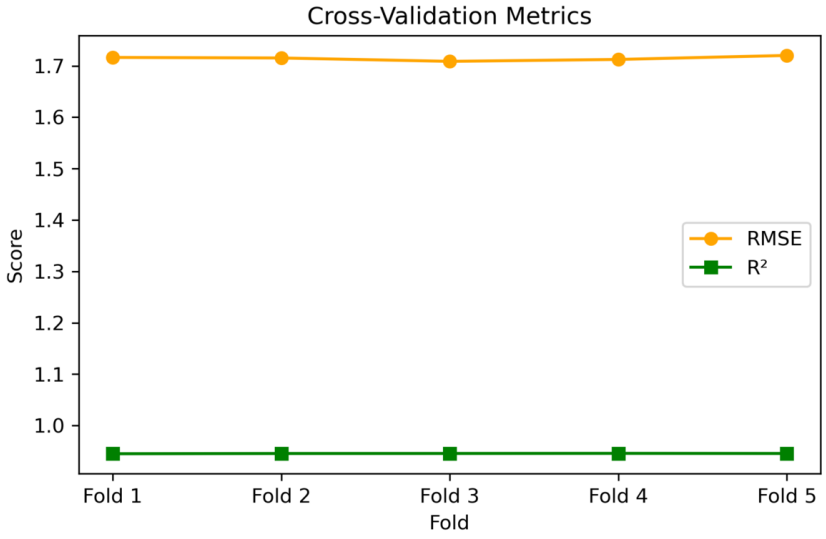

Figure 16. 5-Fold Cross-Validation metrics.

5.2. Statistical and sensitivity analysis

To strengthen the robustness and interpretability of the model, we used both statistical error assessment and sensitivity analysis. The core performance indicators, Mean Bias Error (MBE), Mean Absolute Percentage Error (MAPE), and the standard deviation of the residuals, were used to evaluate the consistency and reliability of the predictions. Beyond statistical evaluation, we conducted exclusion and

perturbation tests on key variables, including temperature and dew point. Sensitivity analysis, performed by systematically omitting influential input features, revealed that excluding the dew factor significantly degraded model performance, with RMSE increasing to 1.95 and R^2 dropping to 0.9288. This underscores the critical role of the dew factor in predictive precision, as illustrated in Figure 17. These insights enabled the identification of dominant features, informed feature selection, and guided model optimization.

Furthermore, performance metrics were compared across multiple geographic regions to assess spatial variability in the model's effectiveness. Table 2 summarizes the regional performance indicators, highlighting the disparities in error rates and model responsiveness. These insights are critical for tailoring forecasting strategies to specific environments.

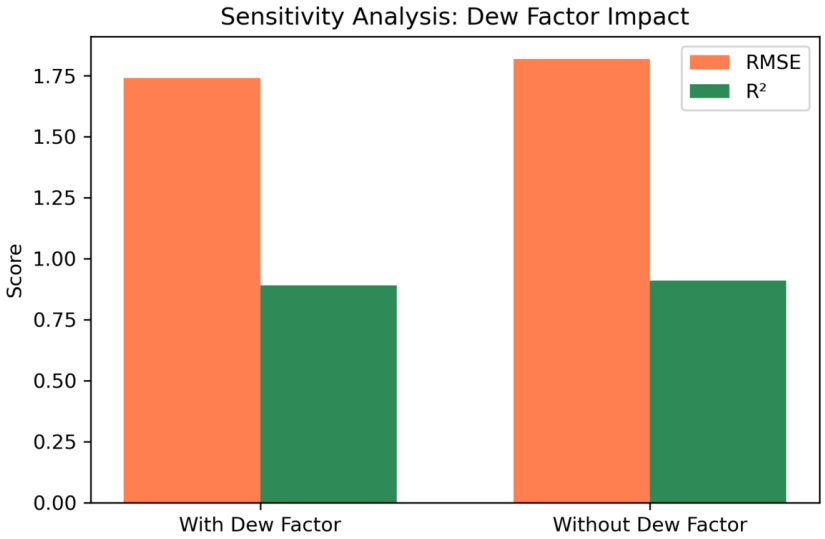

Figure 17. Sensitivity Analysis: Dew Factor Impact

Table 2. Performance metrics across different regions.

Region	RMSE	R^2	MBE	MAPE	Std Dev of Residuals
Region_01	3.01	0.5426	0.62	8.84	2.95
Region_02	2.01	0.4270	0.67	6.36	1.14
Region_03	1.99	0.8176	0.41	7.88	1.95

5.3. Cross-regional generalization

To explicitly demonstrate geographic transferability, we implemented a leave-one-climate-out cross-validation (LORO-CV) scheme. In this design, the ensemble was trained on all but one climatic zone and evaluated exclusively on the held-out zone. This procedure enforces spatial independence between training and testing sets, thereby simulating deployment in previously unseen geographies. Unlike random k-fold splits, which assume i.i.d. samples, LORO-CV accounts for spatial autocorrelation inherent in climate data.

The results revealed substantial variability between regions. For example, RMSE ranged from 1.99 (Region_03) to 3.01 (Region_01), while R^2 varied between 0.43 and 0.82. These differences highlight that model precision is climate-dependent: tropical regions exhibit stronger predictability, while temperate

regions show greater residual variance. Residual plots confirmed that errors remained centered around zero, with no systematic bias across irradiance ranges.

Leave-one-climate-out validation revealed that ensemble performance is not uniform across geographies. RMSE increased by approximately 50% when temperate climates were excluded from the training data, underscoring the importance of climatic diversity in training data. This experiment demonstrates that geographic generalization is feasible but sensitive to climatic heterogeneity.

5.4. Ensemble architecture transparency

The ensemble was implemented using `VotingRegressor` with uniform averaging across constituent models (SVR, XGBoost, AdaBoost, Gaussian Process, Decision Tree). Equal weighting was chosen for three technical reasons: (i) variance control, since different learners exhibit distinct bias–variance trade-offs; (ii) interpretability, as equal weights provide a transparent baseline that is easy to reproduce; and (iii) benchmarking, where uniform averaging serves as a neutral starting point before introducing adaptive schemes.

Formally, the ensemble prediction is given by:

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} \hat{y}_m, \tag{1}$$

where M is the number of regressors. This ensures balanced contributions. Although classification boundaries are not directly relevant (this is a regression task), instability manifests itself as a prediction variance near climatic transition zones. Residual analysis was performed by plotting $y - \hat{y}$ against irradiance for each region. The errors remained centered around zero, with no systematic drift at transition boundaries, indicating robustness across heterogeneous regimes.

Preliminary experiments with alternative weighting strategies showed marginal improvements ($<$ 2% reduction in RMSE) but introduced instability when one model's validation score dominated. For transparency and reproducibility, equal weighting was retained. Future work may incorporate Bayesian model averaging or stacked generalization to adaptively balance regressors.

5.5. Sample results

The actual and predicted solar irradiance values for data from distinct locations are tabulated in Table 3. The acceptable difference validates the performance of the developed model.

Table 3. Sample results (all values in W/m^2).

Sr. No.	Location details	Actual value	Predicted value	Difference value
1	Latitude: 3.1390°, Longitude: 101.6869°	131	128	03
2	Latitude: 15.8402°, Longitude: 70.0219°	108	114	-06

5.6. Strengths of the proposed model

Table 4 compares the performance of the proposed ensemble model with various models from the literature in terms of the approach followed, the type of model, the number of input variables, and the RMSE values. It helps evaluate the strengths and weaknesses of the proposed model compared to existing models. The following are observations.

Table 4. Performance comparison of the proposed model with existing models.

Year	Approach	References	Model type	No. of inputs	RMSE W/m^2	Quantitative Benchmarking
2024	GRU-Attention	[29]	LS	8	20.84	+18.56%
2024	Temporal convolutional network (TCN)	[29]	LS	8	20.75	+18.91%
2024	TCN-GRU attention-MLP (TGAM)	[29]	LS	8	15.7	+38.65%
2024	LSTM	[6]	LS	07	0.955	+96.27%
2024	HRNet	[3]	LS	-	98.570	-285.32%
2024	VMD-WOA-DELM	[9]	LS	-	43.15	-68.61%
2023	Point estimation range classification	[1]	LS	14	0.494	+98.07%
2023	LSTM	[30]	LS	7	12.0	+53.10%
2023	GRU	[30]	LS	7	11.9	+53.47%
2023	MLP	[30]	LS	7	11.8	+53.84%
2022	CNN	[8]	LS	04	52.580	-105.52%
2022	Open loop ANN	[11]	LS	09	8.9185	+65.13%
2022	ANN	[24]	LS	08	1.2886	+94.96%
2022	SCFT	[5]	LS	04	13.38	+47.71%
2020	LSTM	[31]	LS	05	62.540	-144.47%
2020	LSTM	[32]	LS	06	30.210	-18.06%
-	Proposed model	-	GA	03	1.71	+93.36%

1. Approach: A unique approach is discussed in the present work compared to a diverse range of techniques reported in the literature.
2. Model type: The existing models are developed considering the relevant data of some specific places. Hence, the models are location-specific (LS). If the prediction is required in another location, the entire process is repeated to make it relevant to that location. However, the proposed model is formulated and tested globally using the desired data, making it geographically agnostic.
3. Number of Input Variables: Existing models utilize different numbers of input variables, ranging from 4 to 14. The proposed ensemble model focuses on three key input variables, demonstrating a targeted approach compared to other models.
4. RMSE Values: These values reflect the prediction accuracy of the models, with lower values indicating better performance. The proposed ensemble model has an RMSE of 1.71, indicating its performance falls within the moderate range compared to existing models.
5. Quantitative Benchmarking: To ensure a fair comparison between diverse model architectures and input configurations, we recalculated performance improvements using the average RMSE of all benchmarked models (25.59 (W/m^2)) as a baseline.
6. The Worldwide analysis is carried out, and the model's performance is evaluated using data from major climatic groups, including tropical, dry, temperate, continental, and polar. The model's performance is satisfactory.

Thus, the proposed model differs in its unique approach, global applicability, and promising results, with a moderate RMSE. Thus, the proposed model differs in its unique approach, global applicability, and promising results, with a moderate RMSE.

6. Future work

In the future, there are several opportunities to strengthen and extend this work. An important direction is to test the model's robustness to sensor noise, since real-world measurements often suffer from calibration drift or environmental interference. Another is to adapt the framework for regions with sparse or incomplete datasets, where creative solutions such as satellite integration or transfer learning could make the model more practical. Similarly, atypical irradiance behaviors, such as prolonged low-light conditions during high-latitude winters, deserve closer attention, as they may challenge the model's generalisability. Finally, practical implementation will require careful consideration of operational challenges, including sensor maintenance, infrastructure reliability, computational efficiency, and integration into forecasting or energy management systems. Addressing these aspects will help ensure that the model is not only theoretically sound but also resilient and useful in diverse real-world settings.

7. Conclusion

Inconsistent energy generation of solar PV systems disturbs energy management in a solar PV-based system. This issue can be resolved by accurate forecasting of solar irradiance. The AI models emphasized in the solar irradiance forecasting literature are tailored to specific locations, limiting their accuracy when applied to other areas. These models require at least 4 weather parameters as input, thereby increasing the cost. The work presented here addresses the constraints of creating location-specific ML models for solar irradiance forecasting by formulating a geographically agnostic ML model. Creating a universal dataset that includes solar irradiance and global climatic parameters is proposed. The universal dataset is analyzed to uncover hidden patterns and trends. Ensembled ML algorithms are deployed in the process. Only three inputs are used in the formulation of the model: time, temperature, and the dew factor. Hence, the number of sensors to be incorporated is reduced. Thus, the model is simple and exhibits reduced system complexity. The performance of this model is validated by the values of RMSE and R^2, which are 1.71 and 0.9455, respectively.

Acknowledgements

The authors express their sincere gratitude to the authorities of KLE Technological University, Hubballi, Karnataka, India. The authors also acknowledge the active association of Sheshank Shyam Kindalkar during his studentship in the Department of Electrical and Electronics Engineering at KLE Technological University, Hubballi, Karnataka, India.

Funding: This research received no external funding.

Author contributions: Conceptualisation, A.I., R.K., and M.K.; Methodology, A.I., M.K., R.K., and K.P.; Model Design, A.I., R.K., and K.P.; Validation, A.I., M.K., R.K., and K.P.; Formal Analysis, A.I., R.K., and K.P.; Data Curation, A.I., M.K., and R.K.; Writing — Original Draft Preparation, A.I., R.K., M.K., and K.P.; Writing — Review & Editing, A.I., M.K., and R.K.; Visualisation, A.I., and K.P.

Disclosure statement: The authors declare no conflict of interest.

References

[1] Somaieh Kharazi, Nima Amjady, Maryam Nejati, and Hamidreza Zareipour. A new closed-loop solar power forecasting method with sample selection. *IEEE Transactions on Sustainable Energy*, 15(1):687–698, 2023.

[2] Montaser Abdelsattar, Mohamed A Ismeil, Mohamed A Azim, Ahmed AbdelMoety, and Ahmed Emad-Eldeen. Assessing machine learning approaches for photovoltaic energy prediction in sustainable energy systems. *IEEE Access*, 2024.

[3] Hyojung Ahn, Jeongmin Yu, Jonghan Ko, and Jong-Min Yeom. Enhanced short-term prediction of solar radiation using hrnet model with geostationary satellite data. *IEEE Geoscience and Remote Sensing Letters*, 2024.

[4] Pravat Kumar Ray, Bidyadhar Subudhi, Ghanim Putrus, Mousa Marzband, and Zunaib Ali. Forecasting global solar insolation using the ensemble kalman filter based clearness index model. *CSEE Journal of Power and Energy Systems*, 8(4):1087–1096, 2022.

[5] Najiya Omar, Hamed Aly, and Timothy Little. Seasonal clustering forecasting technique for intelligent hourly solar irradiance systems. *IEEE Transactions on Industrial Informatics*, 19(3):2520–2529, 2022.

[6] Ali M Hayajneh, Feras Alasali, Abdelaziz Salama, and William Holderbaum. Intelligent solar forecasts: Modern machine learning models and tinyml role for improved solar energy yield predictions. *IEEE Access*, 12:10846–10864, 2024.

[7] Jonathan Fjord Jønler, Frederik Brunø Lottrup, Bogi Berg, Dalin Zhang, and Kaixuan Chen. Probabilistic forecasts of global horizontal irradiance for solar systems. *IEEE Sensors Letters*, 7(1):1–4, 2022.

[8] Garazi Etxegarai, Asier López, Naiara Aginako, and Fermín Rodríguez. An analysis of different deep learning neural networks for intra-hour solar irradiation forecasting to compute solar photovoltaic generators' energy production. *Energy for Sustainable Development*, 68:1–17, 2022.

[9] Siyuan Zhang, Dongsheng Niu, Zhi Zhou, Yanglong Duan, Jian Chen, and Genben Yang. Prediction method of direct normal irradiance for solar thermal power plants based on vmd-woa-delm. *IEEE Transactions on Applied Superconductivity*, 2024.

[10] Anupama R Itagi, Mrityunjaya Kappali, and Shrikant Karajgi. An open loop time series ann model for forecasting solar insolation for standalone pv applications. In *2022 2nd International Conference on Intelligent Technologies (CONIT)*, pages 1–4. IEEE, 2022.

[11] Anupama R Itagi, Mrityunjaya Kappali, Shrikant Karajgi, and Padmaja Kallimani. Prediction of solar insolation in a pv based dc micro grid to meet the ceaseless energy demand of critical loads. In *2022 International Conference for Advancement in Technology (ICONAT)*, pages 1–3. IEEE, 2022.

[12] Munir Husein and Il-Yop Chung. Day-ahead solar irradiance forecasting for microgrids using a long short-term memory recurrent neural network: A deep learning approach. *Energies*, 12(10):1856, 2019.

[13] Power Nasa. Nasa prediction of worldwide energy resources. In *Data Access Viewer*, 2022.

[14] Hylke E Beck, Niklaus E Zimmermann, Tim R McVicar, Noemi Vergopolan, Alexis Berg, and Eric F Wood. Present and future köppen-geiger climate classification maps at 1-km resolution. *Scientific data*, 5(1):1–12, 2018.

[15] Yunjun Yu, Junfei Cao, and Jianyong Zhu. An lstm short-term solar irradiance forecasting under complicated weather conditions. *IEEE Access*, 7:145651–145666, 2019.

[16] Jessica Wojtkiewicz, Matin Hosseini, Raju Gottumukkala, and Terrence Lynn Chambers. Hour-ahead solar irradiance forecasting using multivariate gated recurrent units. *Energies*, 12(21):4055, 2019.

[17] Liexing Huang, Junfeng Kang, Mengxue Wan, Lei Fang, Chunyan Zhang, and Zhaoliang Zeng. Solar radiation prediction using different machine learning algorithms and implications for extreme climate events. *Frontiers in Earth Science*, 9:596860, 2021.

[18] Bálint Hartmann. Comparing various solar irradiance categorization methods–a critique on robustness. *Renewable Energy*, 154:661–671, 2020.

[19] Theddeus T Akano and Chinemerem C James. An assessment of ensemble learning approaches and single-based machine learning algorithms for the characterization of undersaturated oil viscosity. *Beni-Suef University Journal of Basic and Applied Sciences*, 11(1):149, 2022.

[20] Iqbal H Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3):160, 2021.

[21] Nicholas Pudjihartono, Tayaza Fadason, Andreas W Kempa-Liehr, and Justin M O'Sullivan. A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*, 2:927312, 2022.

[22] Deepak Mehta and Chaman Verma. Prediction of cancer diagnosis patients from fine-needle aspirates using machine learning. In *International Conference on Intelligent Computing and Smart Communication 2019: Proceedings of ICSC 2019*, pages 337–348. Springer, 2019.

[23] Pritika Bahad and Preeti Saxena. Study of adaboost and gradient boosting algorithms for predictive analytics. In *International Conference on Intelligent Computing and Smart Communication 2019: Proceedings of ICSC 2019*, pages 235–244. Springer, 2020.

[24] Jie Wang. An intuitive tutorial to gaussian process regression. *Computing in Science & Engineering*, 25(4):4–11, 2023.

[25] Jinkyoo Park, David Lechevalier, Ronay Ak, Max Ferguson, Kincho H Law, Y-TT Lee, and Sudarsan Rachuri. Gaussian process regression (gpr) representation in predictive model markup language (pmml). *Smart and sustainable manufacturing systems*, 1(1):121–141, 2017.

[26] Teemu Mutanen, Laura Sirro, and Yrjö Rauste. Tree height estimates in boreal forest using gaussian process regression. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 1757–1760. IEEE, 2016.

[27] Rekar O Mohammed and Gavin C Cawley. Over-fitting in model selection with gaussian process regression. In *Machine Learning and Data Mining in Pattern Recognition: 13th International Conference, MLDM 2017, New York, NY, USA, July 15-20, 2017, Proceedings 13*, pages 192–205. Springer, 2017.

[28] Isha Arora, Jaimala Gambhir, and Tarlochan Kaur. Solar irradiance forecasting using decision tree and ensemble models. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 675–681. IEEE, 2020.

[29] Zhi Rao, Zaimin Yang, Xiongping Yang, Jiaming Li, Wenchuan Meng, and Zhichu Wei. Tcn-gru based on attention mechanism for solar irradiance prediction. *Energies*, 17(22):5767, 2024.

[30] Salwan Tajjour, Shyam Singh Chandel, Majed A Alotaibi, Hasmat Malik, Fausto Pedro García Márquez, and Asyraf Afthanorhan. Short-term solar irradiance forecasting using deep learning techniques: a comprehensive case study. *IEEE Access*, 11:119851–119861, 2023.

[31] Hui He, Nanyan Lu, Yongjun Jie, Bo Chen, and Runhai Jiao. Probabilistic solar irradiance forecasting via a deep learning-based hybrid approach. *IEEJ Transactions on Electrical and Electronic Engineering*, 15(11):1604–1612, 2020.

[32] Byung-ki Jeon and Eui-Jong Kim. Next-day prediction of hourly solar irradiance using local weather forecasts and lstm trained with non-local data. *Energies*, 13(20):5258, 2020.