Comparison and Statistical analysis of various machine learning techniques for daily prediction of solar GHI representing India's overall solar radiation

Bharat Girdhani (bharatgirdhani2@gmail.com)
Maulana Azad National Institute of Technology Bhopal
Meena Agrawal
Maulana Azad National Institute of Technology Bhopal

Research Article

Keywords: Machine and Deep learning Methods, Solar Energy output, Solar Radiation (GHI) Prediction, Statistical Analysis, AI Model Comparison

Posted Date: April 19th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2801060/v1

Additional Declarations: No competing interests reported.

Comparison and Statistical analysis of various machine learning techniques for daily prediction of solar GHI representing India's overall solar radiation

Bharat Girdhani[1*], Meena Agrawal[1]

[1]Energy Centre, Maulana Azad National Institute of Technology Bhopal, Madhya Pradesh, India, 462003

[*]Corresponding author: Bharat Girdhani

***Corresponding author email - bharatgirdhani2@gmail.com**

Comparison and Statistical analysis of various machine learning techniques for daily prediction of solar GHI representing India's overall solar radiation

Abstract

Solar energy integration into the grid is a significant challenge because of its varying and unpredictable nature. Therefore, accurate solar energy prediction is vital in ensuring grid stability. To achieve this, the present study uses machine and deep learning methods to estimate the solar global horizontal irradiance. This study aims to predict daily solar GHI for four Indian states (Rajasthan, Madhya Pradesh, Assam, and Meghalaya) with different solar radiation distributions ranging from very high to very low. Four machine-learning techniques (linear regression, support vector machine, ANN and random forest) are used in the present study. Specific sites (Bhadla - Rajasthan, Rewa - Madhya Pradesh, Amguri-Assam, and Shillong-Meghalaya) were chosen in the respective states. The results of the sites represent the overall results for the entire state in this study. The dataset utilized for the study pertains to the selected sites and encompasses the period from January 2019 to November 2022. The study has focused on evaluating the success of machine learning techniques based on seven statistical metrics, including MBE, MAE, MSE, RMSE, Max. Error, R^2, and MAPE. The result analysis indicates that all ML techniques' R^2, MAPE, and MBE values lie between 0.6108 to 0.9152, 0.0432 to 0.2248, and -0.2271 to 0.63704 MJ/m2, respectively. The study concludes that all of the machine learning techniques can accurately predict daily solar GHI, with ANN being the best-performing model.

Keywords:

Machine and Deep learning Methods, Solar Energy output, Solar Radiation (GHI) Prediction, Statistical Analysis, AI Model Comparison.

Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
PV	Photo Voltaic
RF	Random Forest
SVM	Support Vector Machine
ANN	Artificial Neural Network
LR	Linear Regression
DT	Decision Tree
MLP	Multi-layer perceptron
GBT	Gradient boosting tree
MBE	Mean bias error
MAE	Mean absolute error
MSE	Mean squared error
RMSE	Root mean squared error
MAPE	Mean Absolute Percentage Error
GHI	Global Horizontal Irradiance
kNN	K Nearest Neighbor
PCA	Principal Component Analysis

1. Introduction

As a result of the extensive use of fossil fuels in recent times, the effects of the energy crisis and global warming have been felt on various fronts, such as climatic patterns, energy security concerns, and government economic policies. This has led to a push for clean and alternative(sustainable) energy sources to replace current energy production from conventional energy sources. Solar energy emanates from the sun in the form of heat and radiant light. It is a type of renewable energy that can be harnessed to generate solar power. Unlike traditional energy sources, such as fossil fuels, solar power is clean and sustainable (does not emit GHGs into the atmosphere). The globe is shifting towards using solar energy to generate power and reshape the energy mix into the grid. To develop an environmentally friendly and long-lasting electrical system, including a more significant renewable energy fraction in this mix is critical. However, solar power generation is impacted by the volatility (demand, price) caused by the meteorological and climate change conditions and can lead to the insecurity of solar power generation. Therefore, utilizing a suitable solar energy forecast model capable of effectively learning from the past meteorological dataset is crucial. This model can aid in ensuring optimal outcomes in the energy sector, making informed decisions, and optimizing operations. Implementing such a model would provide valuable advanced information on the availability of solar power, which can significantly contribute to grid stability and enable optimal load dispatch and unit commitment in the energy sector. In this respect, the effectiveness of the prediction model is determined by the importance of the characteristics (input variables) chosen for prediction and the forecasting horizon to the available data. It is thus critical to decide on the climatic factors that may aid in providing accurate estimates in the solar energy field. Using an integrated solution (a single prediction model) is also crucial to meet the many problems and goals related to the various horizons of solar prediction.

Based on the associated problems and goals, the photovoltaic predictions can be classified into the short-term or real-time horizon, medium horizon, and long-term horizon. For PV storage management and power trading, a real-time horizon is required. The short-term help makes decisions such as economic dispatch of load and commitment to the unit. At the same time, medium-horizon and long-term horizons are used in effective planning and scheduled maintenance of the PV plant.

The phrase "solar radiation" refers to "global solar radiation or solar GHI" throughout this work. SR measuring devices are often used to monitor solar radiation. Nevertheless, this equipment requires significant installation, maintenance, and calibration expenditure. Consequently, they are unavailable at the majority of stations worldwide. That is why it is necessary to forecast solar radiation statistics based on easily quantifiable climatic elements such as humidity, temperature, wind speed, cloud cover, etc. Several models have been presented to forecast solar radiation data from this perspective. Some of them are termed empirical models since mathematical formulae are their basis. Because of their superficial computational characteristics, these models are frequently recognized as a technique for estimating solar GHI. Even though these models are extensively used to estimate the monthly mean daily global solar GHI, they can-not anticipate short-term solar radiation data owing to rapid changes in climatic circumstances such as cloud cover, rainy days, and so on. Certain studies suggest that solar radiation may be heavily influenced in areas with high humidity and significant cloud coverage during rainfall, making it difficult for models to accurately capture the complexity and nonlinearity between dependent and independent variables (Fan et al. 2018). Previous research revealed that for global solar radiation (daily) data, these empirical models produced partly satisfactory prediction out-comes (Jiang 2008; Sharifi et al. 2016). Fig. 1 and Fig. 2 show the techniques for solar energy prediction and the classification of various forecasting models.

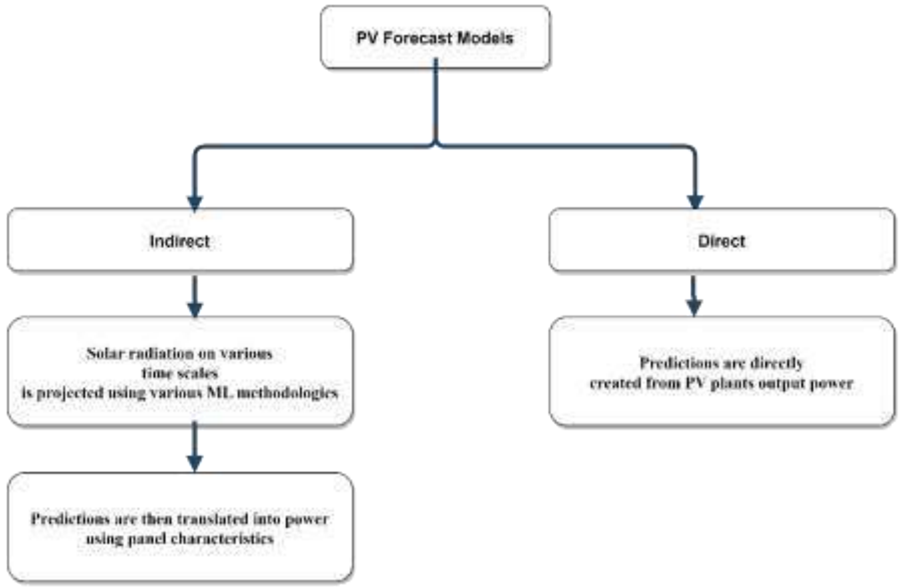


Figure 1 Techniques for solar energy prediction(Alkhayat and Mehmood 2021a)

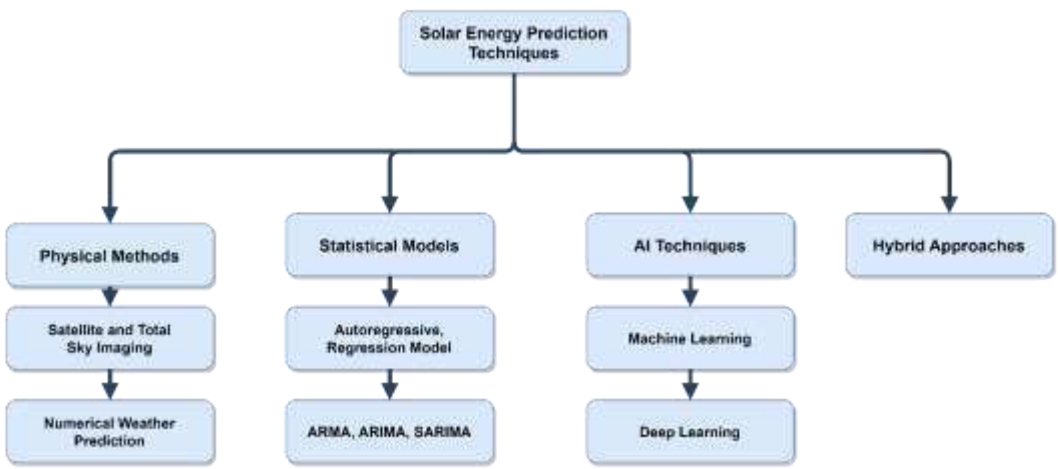


Figure 2 Classification of various forecasting models (Alkhayat and Mehmood 2021b)

AI and Machine Learning (ML) approaches (Benali et al. 2019) have found extensive applications in this field due to their exceptional regression and learning capabilities. AI approaches such as LR, SVM, RF, kNN, ANN, and others have been famous for predicting solar energy. Prior research has demonstrated that for solar radiation prediction, AI techniques offer more precise outcomes than empirical models. According to (Yang et al. 2020), an RF PV prediction technique based on PCA and the K-means technique has been developed. (Lai et al. 2020) provided an overview of various ML algorithms that can be used for the prediction of renewable energy. (Lahouar et al. 2017) proposed an RF-based predictor for estimating day-ahead power incorporating a Smart Persistence advanced ML approach. (Kim et al. 2019) used machine learning to demonstrate a two-stage technique for daily estimation of solar energy based on meteorological data. (Quej et al. 2017) projected solar GHI on a daily basis for Mexican city. ANN, ANFIS, and SVM were used as the ML models. The researchers trained the algorithms using meteorological data's rainfall, temperature, and solar radiation variables. (Ağbulut et al. 2021) predicted solar GHI on a daily basis using kNN, ANN, DL, and SVM ML techniques for four regions of Turkey. Based on the statistical analysis results, ANN and DL models outperformed the other models. (Mehdizadeh et al. 2016) calculated day-ahead solar GHI for Iran using three distinct models: GEP, ANN, and ANFIS. ANN (R^2 =0.935) produced the best outcomes. (Gürel et al. 2020) predicted monthly solar GHI utilizing feed-forward NN techniques, empirical Angstrom technique, RSM, and time series model for turkey. The obtained meteorological dataset contained variables such as humidity, surface pressure, ambient temperature, sunshine duration, and wind speed. The ANN model outperformed all the other models with $R^2 = 0.9911$. (Yadav et al. 2022) predicted solar power for Malaysian and Indian regions using the ANN model, and RMSE was considered the primary evaluation parameter. (Jebli et al. 2021) utilized the Pearson correlation coefficient for solar energy prediction using SVM, LR, ANN, and RF ML models. The projections were made for the Morocco region, and the results were compared with the Pirapora region of Brazil. The Artificial neural network and the Random-forest models provided effective prediction results for these regions. (Yıldırım et al. 2018) daily forecasted the solar GHI for Turkey using regression analysis and ANN techniques. Extraterrestrial solar radiation, sunshine duration, temperature, day of the year, and humidity were the variables considered in the dataset to train the selected algorithms. Based on the statistical analysis, ANN (RMSE = 0.14 and R2 = 0.961) was the outperforming model.

(Essam et al. 2022) investigated forecasting of solar PV output using extreme gradient boosting, long short-term memory and decision tree, RF, and ANN techniques. The dataset was collected from NREL and consisted of meteorological variables and monocrystalline PV plant output data for the cocoa region in Florida. ANN was the best algorithm with MAE = 0.4693, RMSE = 0.8816 W, and $R^2 = 0.9988$, respectively. (Makade et al. 2021) provided a detailed assessment of the existing solar GHI models and statistical analysis specifically for Indian regions. The authors provided a review and evaluated various models that have been developed to estimate solar GHI and assess their applicability and accuracy for Indian regions. (Khastagir et al. 2022) forecasted rainfall using ANN and multiple linear regression ML techniques for western Australia. The authors concluded that both ML techniques could predict accurately based on RMSE and MAE. Rahul et al utilized DT regression ML model for forecasting PV output generation. (Feng et al. 2020) mapped and quantified PV power and daily solar GHI using the PSO-ELM model, and the results were compared with SVM, ELM, autoencoder, RNN, and Decision tree ML techniques for the China region. As a result of the expensive costs of solar GHI measuring equipment, their calibration, and their maintenance cost, a large number of meteorological stations worldwide are incapable of gathering data related to solar radiation (Gürel et al. 2020). Thus, it's vital to forecast solar radiation data using more superficial characteristics to monitor in the appropriate location. Several ML methods have lately been applied for the daily prediction of solar GHI, as shown by earlier research publications. These algorithms are often utilized primarily for daily solar GHI because they may provide forecast outcomes more accurately compared to empirical models. However, past research has shown that the evaluated algorithms produce similar outcomes for the study locations. As a result, discussing the performance success of the algorithms using just a few indicators may need to be improved. With these perspectives, the following are the key contributions and novelty of this study:

- **To employ** the Pearson correlation coefficient to cross-correlations between various meteorological variables
- **To predict** the distribution of India's solar radiation nationwide, using four specific sites located in four states with varying levels of solar GHI potential - very low, low, medium, and high. These chosen states illustrate India's overall solar GHI (solar radiation).
- **A thorough examination of the findings**: Seven error metrics (MBE, MAE, MSE, RMSE, Max Error, R^2, MAPE) are used for the comparison and evaluation of the four ML models for daily prediction of solar GHI.

The remaining work structure is as follows: Section 2 discusses the study sites, forecasting, and overall methodology, and finally, the introduction to the various ML techniques and evaluation parameters under the study's consideration. **Section 3** contains the time series prediction results and the obtained statistical results. **Section 4** finally summarizes the conclusion.

2. Material and Methodology

There are four subsections in this section. The first section gives information regarding the research sites. The second section provides data gathering and pre-processing information. The third section includes a summary of the machine learning techniques. Finally, in the last subsection, statistical metrics are presented.

2.1 Description of the study sites

India is situated to the north of the equator, with latitudes ranging between 8°4' north (on the mainland) and 37°6' north and longitudes ranging between 68°7' east to 97°25' east. It is the world's seventh-largest nation with a total land area of 3,287,263 km^2 (Makade et al. 2021) (1,269,219 sq mi). India experiences an average of 250-300 clear and sunny days per year, with approximately 2300-3200 hours of sunlight. India has enormous potential for solar energy. According to (Kirmani et al. 2015), around 5,000 trillion kWh of energy is incident on India's geographical surface annually, with most regions receiving 4-7 kWh per square meter per day and 1200-2300 kWh/m2 per year. This case makes India a desirable place for solar energy investments. In this article, significant sites in four different Indian states that represent India's overall solar radiation were selected for daily prediction of solar GHI. Table 1 and Fig. 3 provide some essential geographical facts and a perspective of the states on the Indian map. Fig. 3 shows India's solar radiation distribution. It demonstrates that the solar GHI levels of the four study states vary significantly in magnitude. For example, Meghalaya state (particularly Shillong) has extremely low solar radiation, and Rajasthan has the highest global solar radiation. Compared to the other two states, Assam has intense solar radiation, while high solar GHI is characteristic of the MP state. Though there are other states, only these four states are selected with a viewpoint to represent India's overall solar radiation. The significance of the selected regions in the particular state is shown in Table 1.

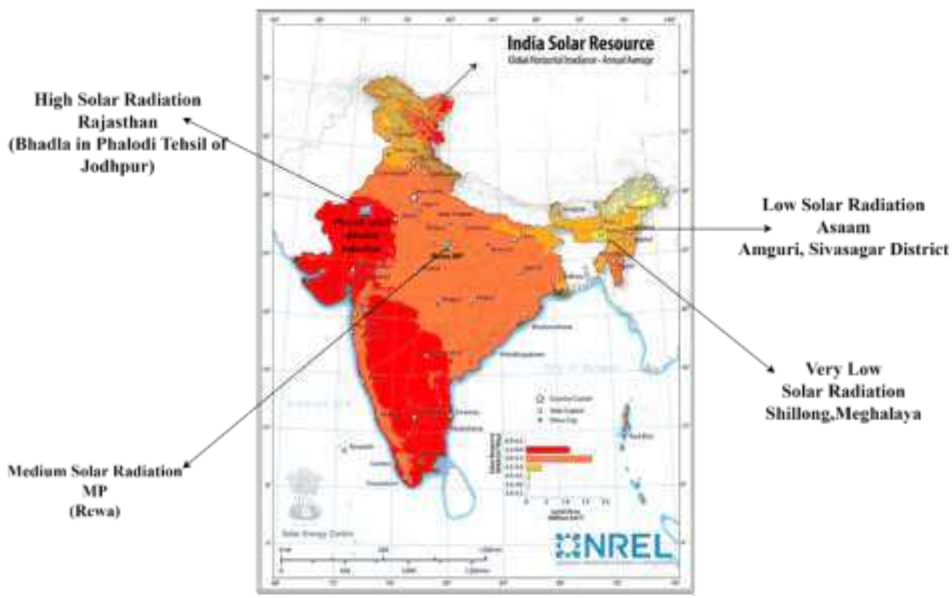


Figure 3 India's solar radiation distribution (Elchinger 2016)

Table 1 Specific features of the selected sites

S. No.	Site	Representation State	GHI Status	Site Feature
1.	Bhadla in Phalodi Tehsil of Jodhpur	Rajasthan	Very high (5.5-6.0 kWh/m^2/day)	World's largest solar park (capacity=2245 MW) is located here
2.	Rewa	Madhya Pradesh	Medium (5.0-5.5 kWh/m^2/day)	MP's largest (Capacity = 750 MW) solar power plant
3.	Amguri, Sivasagar	Assam	Low (4.5-5.0 kWh/m^2/day)	Biggest solar plant (Capacity = 70 MW) in North-East India is to be installed here
4.	Shillong	Meghalaya	Very Low (4.0-4.5 kWh/m^2/day)	Site with very low solar radiation

2.2. Forecasting Methodology

To assess the model's performance, the entire forecasting methodology contain various divisions-

- Collection of data
- Data preprocessing
- Feature identification and extraction
- ML model training
- ML model testing

The overall forecasting methodology is shown in Figure 4.

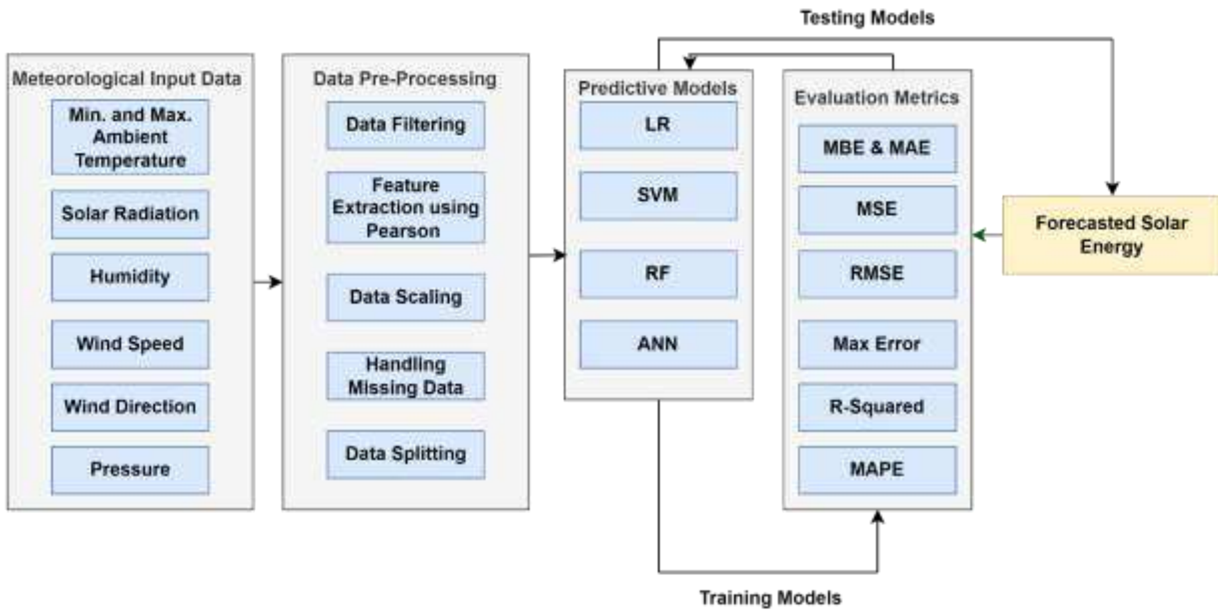


Figure 4 Overall Methodology

2.2.1 Meteorological collection of data

The primary objective of this paper is to predict the daily mean solar GHI (global solar radiation) for four specific sites to represent India's overall solar radiation. Raw data (01 Jan 2019 to 31 Dec 2022) for Bhadla (Rajasthan), Rewa (MP), Amguri Sivasagar (Assam), and Shillong (Meghalaya) was collected from Nasa's solar radiation database.

2.2.2 Data preprocessing and feature extraction using pearson correlation coefficient

Before applying ML techniques, input data cleaning and normalization is essential. Due to unpredictability and changing weather conditions, the dataset could have several non-stationary components and peaks, resulting in significant prediction errors caused by inadequate training of the model. Additionally, the entire meteorological parameters in the dataset may not be helpful [ ]. Data preparation is thus a crucial step in getting the data ready for an ML model. One of the most critical processes in solar GHI forecasting is feature extraction and identification, which comes after data preparation. The most significant and relevant features strongly correlated with the solar GHI were determined using the r_{xy} coefficient, also known as the Pearson Correlation coefficient. rxy is defined as follows-

$$r_{XY} = \frac{\sum(X_i - \bar{X})\sum(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}} \tag{1}$$

where,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{N} X_i \tag{2}$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{N} Y_i \tag{3}$$

$\bar{X} - Mean\ value\ of\ X$ and $\bar{Y} - Mean\ value\ of\ y$

$-1 \leq r_{xy} \leq 1$ is the range of Eq. 1

Table 2 Remarks on Correlation status

r_{XY}	**Correlation status between X and Y**
1	Perfectly positive
0	No linear relationship
-1	Perfectly negative

Table 2 represents the remarks on correlation status. After data preprocessing and feature identification, we observed that temperature holds a strong positive correlation with solar GHI, which means that as the temperature increases, the solar GHI will also increase. However, cloud cover is negatively correlated with the solar GHI.

2.3 ML Methodology

AI&ML is becoming more appealing and popular as new application domains emerge. ML gives systems the capacity to grasp and estimate unknown outputs on their own. A Machine Learning system's performance heavily depends on selecting relevant attributes and the success of the training process. Four distinct ML algorithms were employed in this study:

- Linear regression (LR),
- Support vector machine (SVM),
- Random forest (RF), and
- ANN

Python and its associated libraries, including Pandas, SciPy, NumPy, Matplotlib, and Scikit-learn, were used to implement all of the algorithms in Google Collab. During the training of models, the training and testing split approach (80% for training:20% for testing) was adopted to minimize bias in learning.

2.3.1 Linear Regression (LR)

LR is a popular ML algorithm used in predictive modeling. It is a supervised learning algorithm that models the linearity between a dependent variable and one or more independent variables. The objective of LR is to find the best-fit line, which can accurately predict the value of the dependent variable for any new data points. The basic idea behind linear regression is to fit a straight line that minimizes the distance between the predicted and actual values of the dependent variable. Linear regression is a simple and effective algorithm. It is easy to interpret, and the results can be used to make informed decisions. However, it assumes that the relationship between the independent and dependent variables is linear and that the errors are normally distributed. If these assumptions are not met, the results may be inaccurate(Hastie et al. 2021).

$$z = c_0 + c_1 * y \tag{4}$$

Where, z - dependent variable, y - independent variable, c_0 - intercept, and c_1 - slope.

2.3.2 SVM Model

It is based on locating a hyperplane with the greatest margin of separation between data points of various classes. The goal of SVM is to identify a decision boundary that minimizes the margin between classes. The margin measures the distance between the hyperplane and the nearest data points in each class. A collection of support vectors, which are the data points closest to the hyperplane, establish the decision boundary. By projecting the data into a higher-dimensional feature space, where the classes become linearly separable, SVM can deal with non-linearly separable data. The inner product of the data points in the feature space is calculated using a kernel function to achieve this goal. It's a powerful algorithm that can handle complex datasets with high accuracy. However, SVM is sensitive to the choice of hyperparameters and can be computationally expensive for large datasets(Corinna Cortes & Vladimir Vapnik 1995).

$$\hat{z} = \frac{1}{N}\sum_{i=1}^{N} z_i \tag{5}$$

Where, $\hat{z}$- predicted value, z_i - prediction from the ith decision tree, and N - number of trees in the forest.

For our analysis, we have considered: Default parameters (Gamma = 'scale' and C =1) AND kernel type – Radial Bias function.

2.3.3 Random Forest (RF)

It works by building many decision trees at training time and outputting the class that is the individual trees' mean prediction (regression). Each decision tree constructed by RF is based on a unique combination of input characteristics and data points drawn randomly from the whole training set. This helps to prevent overfitting and makes the algorithm more robust to noise in the data. RF is known for its high accuracy and ability to handle high-dimensional datasets with complex interactions between the input features (Breiman 2001; Hastie et al. 2021).

Let X be the input vector containing k features with X = {x_1, x_2, …, x_k}

Y – the output scalar and

S_n – the training set having n observations can be represented as

$$S_n = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}, X \in R_k, Y \in R \tag{6}$$

The prediction $\hat{p}$ is the average of the results of different regression and can be determined by,

$$\hat{p} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{7}$$

where, B – number of trees and n – node size

In our analysis, we have considered B = 100 and n_{min} = 1

2.3.4 ANN (Artificial Neural Network)

ANN is an excellent method to model non-linear systems. Therefore, it's unnecessary to characterize such systems using intricate mathematical formulas. ANN is a structural network of neurons or parts having processing power. The generated structure of the neural network requires training to determine the link between the output and the input variables. It closely resembles how the human brain processes information (Behrang et al. 2010). The trained NN structure can handle noisy datasets with missing variables and quickly provide accurate answers. For modeling with ANN, the data needed for training may be more than what can be accomplished using straightforward techniques if the systems are sophisticated (Azlah et al. 2019; Liu and Lang 2019). In ANN, hidden layer numbers may increase based on the network's size. The network may begin to remember the data set as a result, and even noise (Ramil et al. 2018). The feed-forward neural network (ANN) was trained across all study sites using the multi-layer perceptron back-propagation technique. The expression for the output neuron is given as,

$$A_i = g\left(\sum_{k=0}^{n} W_{jk} * a_j\right) \tag{8}$$

Where,

A_i - network output, W_{ij} - weight assigned to the connection from the jth neuron to the ith layer neuron, and

a_j - the neuron input

We analyzed the study sites using the ReLU activation function and ADAM optimizer.

2.4 Statistical Metrics

Assessing the success of prediction algorithms largely depends on the accuracy factor, considered the most crucial aspect. Evaluation of the prediction model output and their comparison has been made using the error metrics The comparison of the performance success of the prediction models used in this paper has been made using error metrics such as Mean bias error (MBE), Max Error, Mean Absolute Error (MAE), Mean squared error (MSE), Root mean squared error (RMSE), Coefficient of determination (R^2), and Mean absolute percentage error (MAPE). **Table 3** lists provides error metrics description.

Table 3 Error Metrics Description

Equation	Description	Equation No.
MBE = $\frac{1}{n}\sum_{j=1}^{n}\langle y_j - x_j\rangle$	MBE is a crucial statistic for predicting long-term model performance. A low MBE value suggests that the prediction model performs well. Furthermore, zero represents the ideal scenario (Fan et al. 2018).	9
MAE = $\frac{1}{n}\sum_{j=1}^{n}\lvert y_j - x_j\rvert$	The Mean Absolute Bias Error (MABE) is a metric that measures the correlation's quality in a prediction model. It is calculated as the absolute value of the difference between the predicted and actual values, averaged over all predictions. A value close to zero indicates a better correlation between the predicted and actual values. MABE gives information on prediction models' long-term performance(Rehman 1998).	10
MSE = $\frac{1}{n}\sum_{j=1}^{n}\left(y_j - x_j\right)^2$	MSE is the average of the squared errors (Error is the difference between the real values y and what is estimated x). A low MSE value indicates that the prediction model performs well (Wang et al. 2012).	11
RMSE = $\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - x_j\right)^2}$	The short-term performance of a prediction model is analyzed by RMSE. It should be close to zero and its always positive(Zang et al. 2020).	12
ME = $\max_{1\le J\le n}\lvert y_j - x_j\rvert$	The worst error between the predicted and the real value is represented by the maximum residual error (ME) (Monjoly et al. 2017).	13
$R^2 = 1 - \frac{\Sigma(x_j - y_j)^2}{\Sigma(x_j - \bar{x}_j)^2}$	This method explains how effectively a model predicts a measurable data's set. An R^2 value approaching 1 indicates a better performance of the model. A higher R^2 value indicates that the model is a better fit for the given data set(Gouda et al. 2019).	14
MAPE = $\frac{1}{n}\sum_{j=1}^{n}\left\lvert\frac{x_j - y_j}{x_j}\right\rvert * 100$	MAPE is the mean absolute value of prediction errors divided by the mean absolute value of actual data. A low MAPE value indicates that the prediction model performs well. The prediction model's success based on MAPE is as follows: (Emang et al. 2010)	15

MAPE Value (Val)	Prediction Accuracy
Val ≤ 10% (or 0.1)	High
10% < Val ≤ 20%	Good
20% < Val ≤ 50%	Reasonable
Val > 50% (or 0.5)	Inaccurate

Where,

n – no. of observations, y_j – predicted solar GHI (global solar radiation) on a daily basis,

x_j – Actual solar GHI (global solar radiation) on a daily basis, and

$\overline{x_j}$ - Average value of the solar GHI (global solar radiation) on a daily basis

3. Discussion on Results

The focus of the current paper is to examine the predictability of daily solar radiation on selected sites located in four different states of India. This study utilizes four various ML techniques to predict the solar GHI (amount of solar radiation falling on a horizontal surface). The study discusses seven regularly used statistical measures often utilized in literature for performance analysis of these algorithms. These error metrics are used to evaluate the performance of the ML algorithms in accurately predicting daily solar radiation (solar GHI). Table 4 represents the calculations of metrics computed for the selected sites using various ML techniques.

Table 4 Performance comparison of study sites for each ML technique

Study Sites	Statistical Metric	LR	SVM	RF	ANN
Bhadla, Rajasthan	MBE(MJ/m^2)	-0.02578	-0.033	0.1344	0.2489
	MAE(MJ/m^2)	2.033	1.5165	1.156	1.169
	MSE(MJ/m^2)	6.51	3.9797	2.676	2.283
	RMSE(MJ/m^2)	2.5515	1.9949	1.636	1.511
	Max Error (MJ/m^2)	8.9524	7.146	7.8069	6.355
	R^2	0.7373	0.8394	0.892	0.905
	MAPE	0.1253	0.08866	0.0649	0.0615
Rewa, Madhya Pradesh	MBE(MJ/m^2)	-0.198	0.5004	-0.0789	0.2795
	MAE(MJ/m^2)	2.1732	1.544	1.0387	1.122
	MSE(MJ/m^2)	8.146	4.5979	2.521	2.4703
	RMSE(MJ/m^2)	2.854	2.144	1.5877	1.5717
	Max Error (MJ/m^2)	9.546	18.0123	12.248	8.3065
	R^2	0.6769	0.8132	0.8983	0.9152
	MAPE	0.1551	0.1055	0.0754	0.06956
Amguri, Sivasagar, Assam	MBE(MJ/m^2)	-0.1513	0.6338	0.3936	0.3667
	MAE(MJ/m^2)	2.2933	1.38	1.49025	1.4153
	MSE(MJ/m^2)	8.6564	3.7093	3.895	3.93
	RMSE(MJ/m^2)	2.9421	1.926	1.974	1.983
	Max Error (MJ/m^2)	11.1641	7.87	10.4428	6.58
	R^2	0.6108	0.8174	0.7807	0.857
	MAPE	0.2248	0.1137	0.1048	0.0623
Shillong, Meghalaya	MBE(MJ/m^2)	-0.2271	0.63704	0.5062	0.3231
	MAE(MJ/m^2)	1.5518	1.6923	1.0906	1.07466
	MSE(MJ/m^2)	4.4706	6.1153	2.4725	2.433
	RMSE(MJ/m^2)	2.1143	2.473	1.5724	1.5601
	Max Error (MJ/m^2)	10.44	11.1644	9.146	8.37
	R^2	0.7959	0.7092	0.87744	0.9027
	MAPE	0.09864	0.1082	0.06434	0.0432

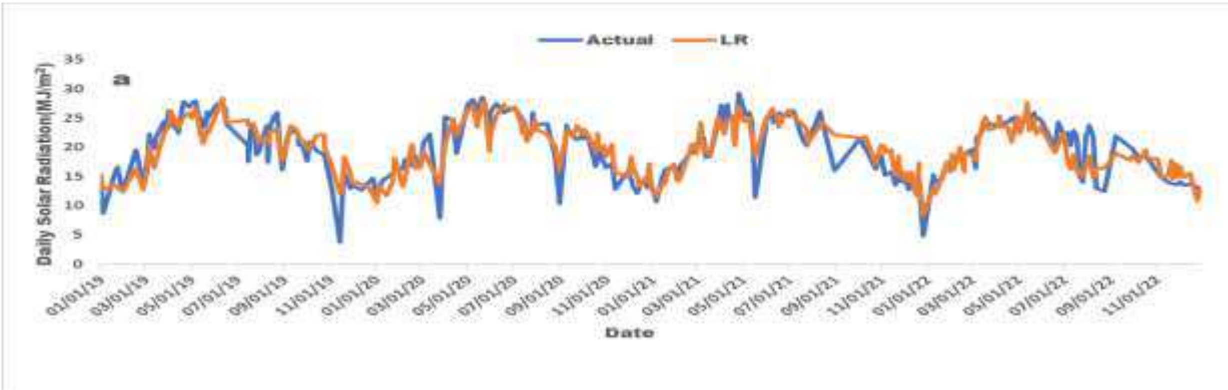
Table 4 shows that based on the specific site and the machine learning technique employed, Metric R^2 ranges from 0.6108 to 0.9152. This section will compare and discuss all study sites and ML techniques using Table 4 as the reference.

Bhadla site, Rajasthan

The site located in Bhadla, Rajasthan, exhibits a very high potential for solar GHI compared to all other sites studied in this work**.** The prediction results using various ML algorithms for Bhadla site are given below-

LR algorithm

The prediction results are shown in Fig. 5. In predicting solar radiation data for the Bhadla site, the LR method has been the most unsuccessful regarding statistical metrics. Compared to the other algorithms for the Bhadla site, the R^2 value (=0.7373) is the lowest, and the MAPE (=12.53%) is the greatest, which is undesirable.



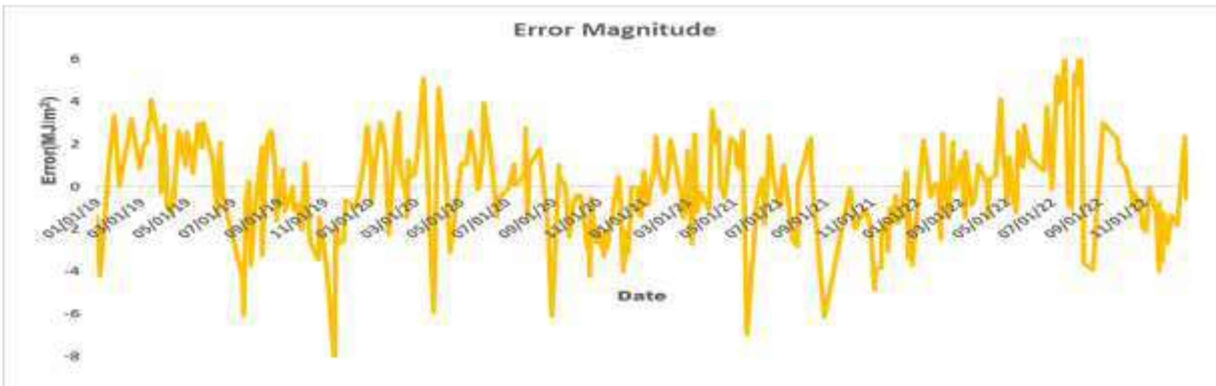


Figure 5 Actual vs predicted values and error metrics for the LR model

SVM algorithm

The prediction results are shown in Fig. 6. The R^2 and MAPE values are 0.8393 and 0.08866, respectively. The SVM algorithm's performance is better than the LR algorithm for the Bhadla site, but still, the results are unsatisfactory. Among all the algorithms used to predict solar radiation data for the Bhadla site in Rajasthan, the Support Vector Machine (SVM) method has demonstrated the third highest level of success after ANN (Artificial Neural Network) and RF (Random Forest).

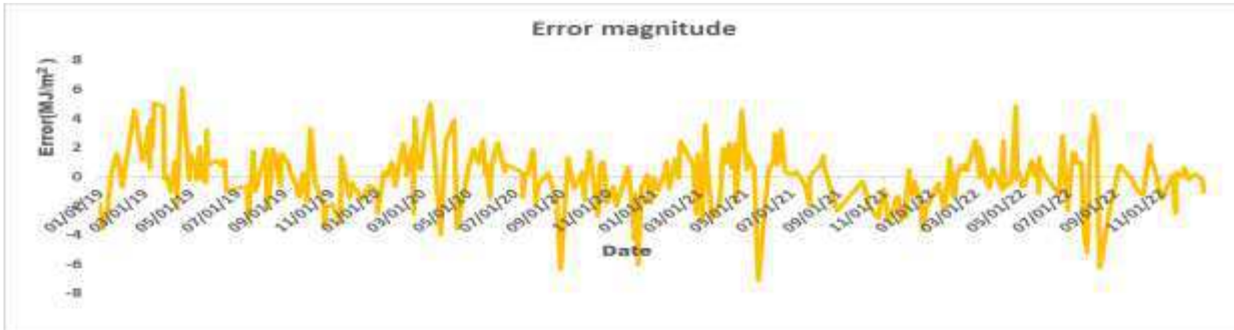


Figure 6 Actual vs predicted values and error metrics for the SVM model

RF algorithm

The prediction results are shown in Fig. 7. The R^2 and MAPE values are 0.892 and 0.0649, respectively. Table 4 indicates that the RF (Random Forest) algorithm outperforms the LR (Linear Regression) and SVM (Support Vector Machine) techniques in terms of both MAPE and R^2 metrics. The error metrics calculated for the RF algorithm are lower than those for the LR and SVM algorithms, indicating the better predictive performance of the RF algorithm for solar radiation data.

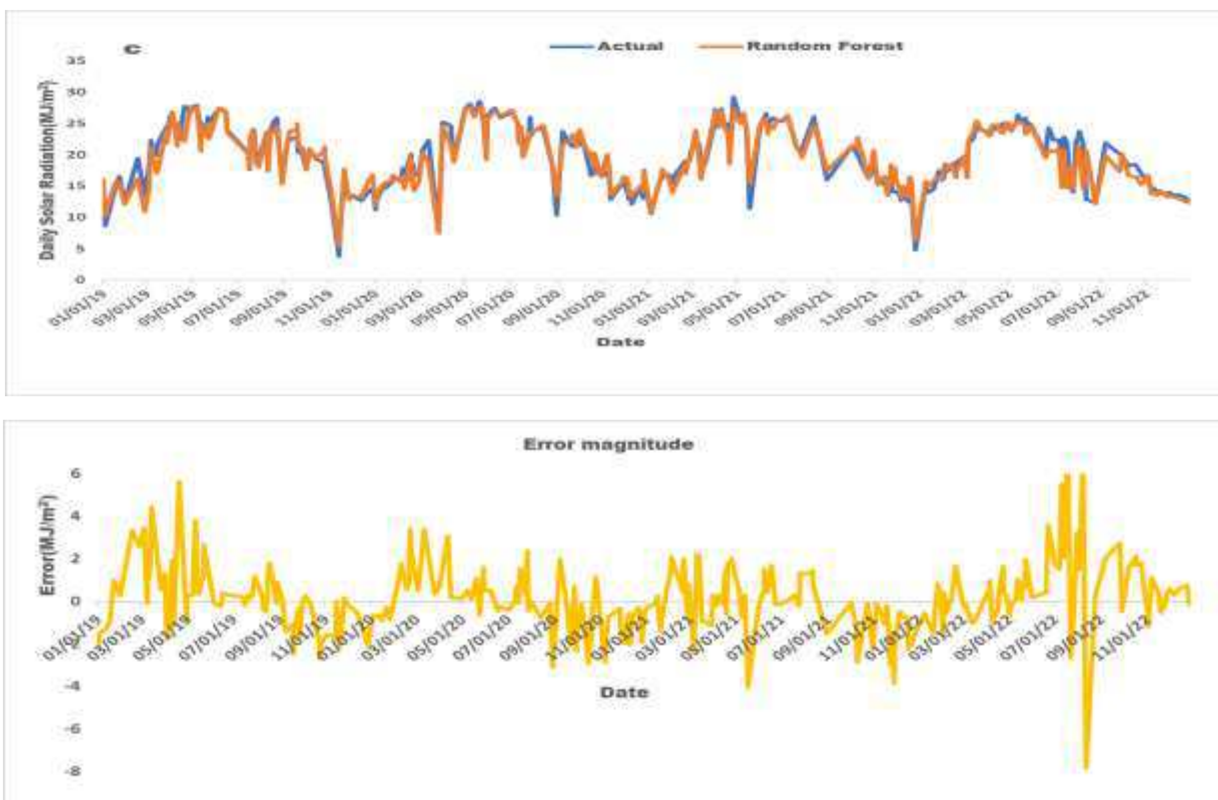
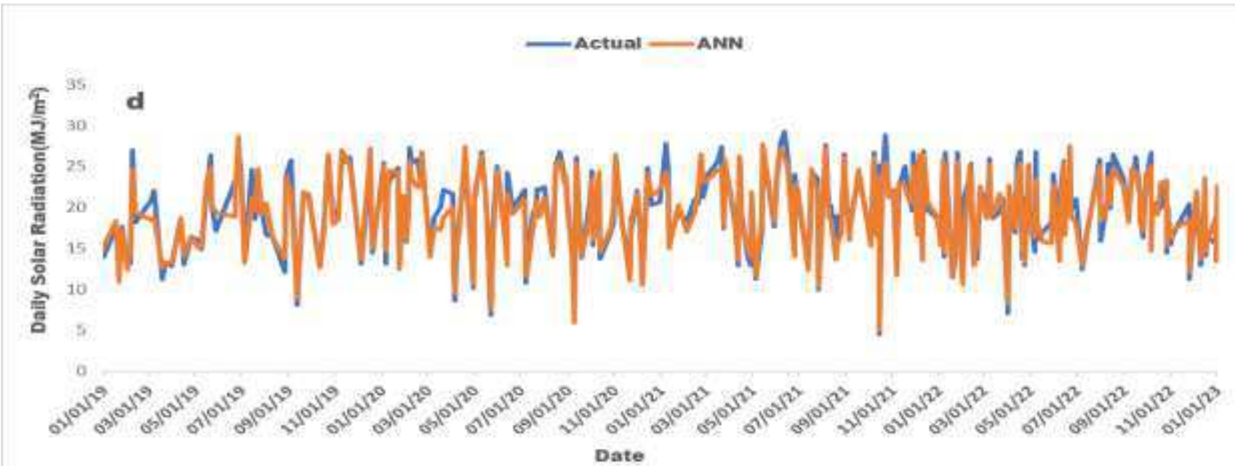


Figure 7 Actual vs predicted values and error metrics for RF model

ANN algorithm

The prediction results are shown in Fig. 8. The ANN algorithm has proven to be the most effective among all the algorithms used for the Bhadla site, as indicated by its high R^2 and MAPE values. Compared to other algorithms, the ANN algorithm demonstrates significantly lower error magnitudes, as observed through the analysis of its performance. The R^2(=0.905) value is very high compared to all other algorithms for this site, and correspondingly the MAPE (=0.0615) value is significantly less. To give a general assessment of the Bhadla site, the prediction success of ANN and RF techniques is very similar, especially when considering all error metrics. Additionally, the error magnitudes for both algorithms are closely aligned.



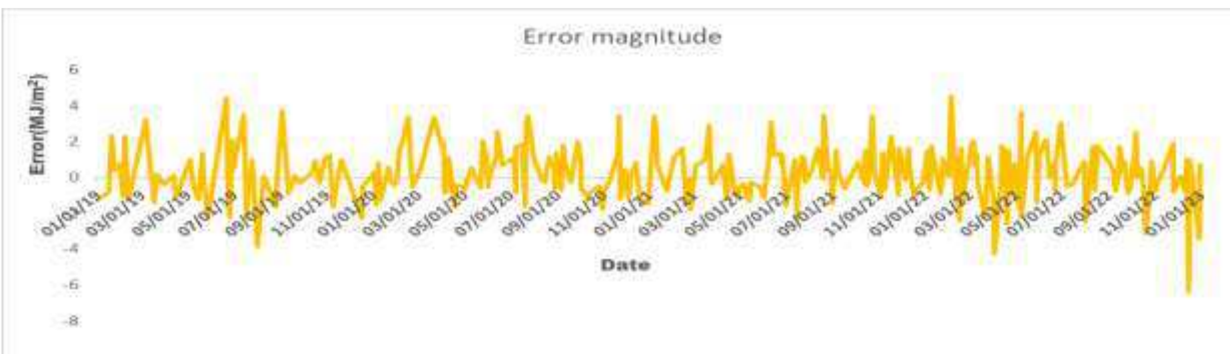


Figure 8 Actual vs predicted values and error metrics for the ANN model

The combined effect of all the algorithms

The combined performance of various algorithms for the Bhadla site is shown in Fig. 9. The spread of data for LR, representing the predicted values, is very high, i.e., the deviation from the actual values is high compared to all other algorithms for this site. It indicates the worst performance of this algorithm**.**

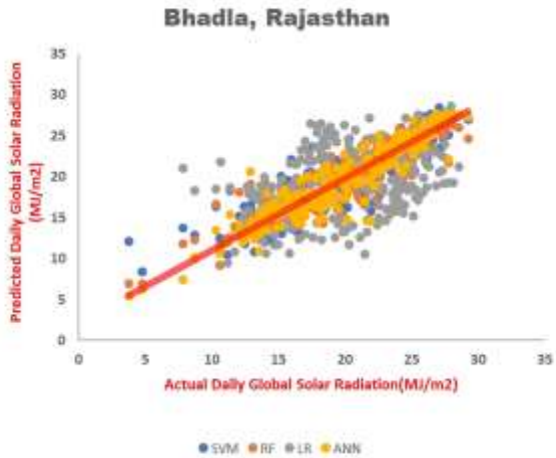


Figure 9 Combined effect of all the algorithms for Bhadla site

Rewa site, Madhya Pradesh

LR algorithm

The prediction results are shown in Fig. 10. It has the worst prediction performance for this site. Table 4 indicates that the R^2 value is 0.6769, and, correspondingly, MAPE is 0.155. Also, the MBE (= -0.198 MJ/m2) has a high negative value—however, the max. Error is significantly less compared to the other models for this site.

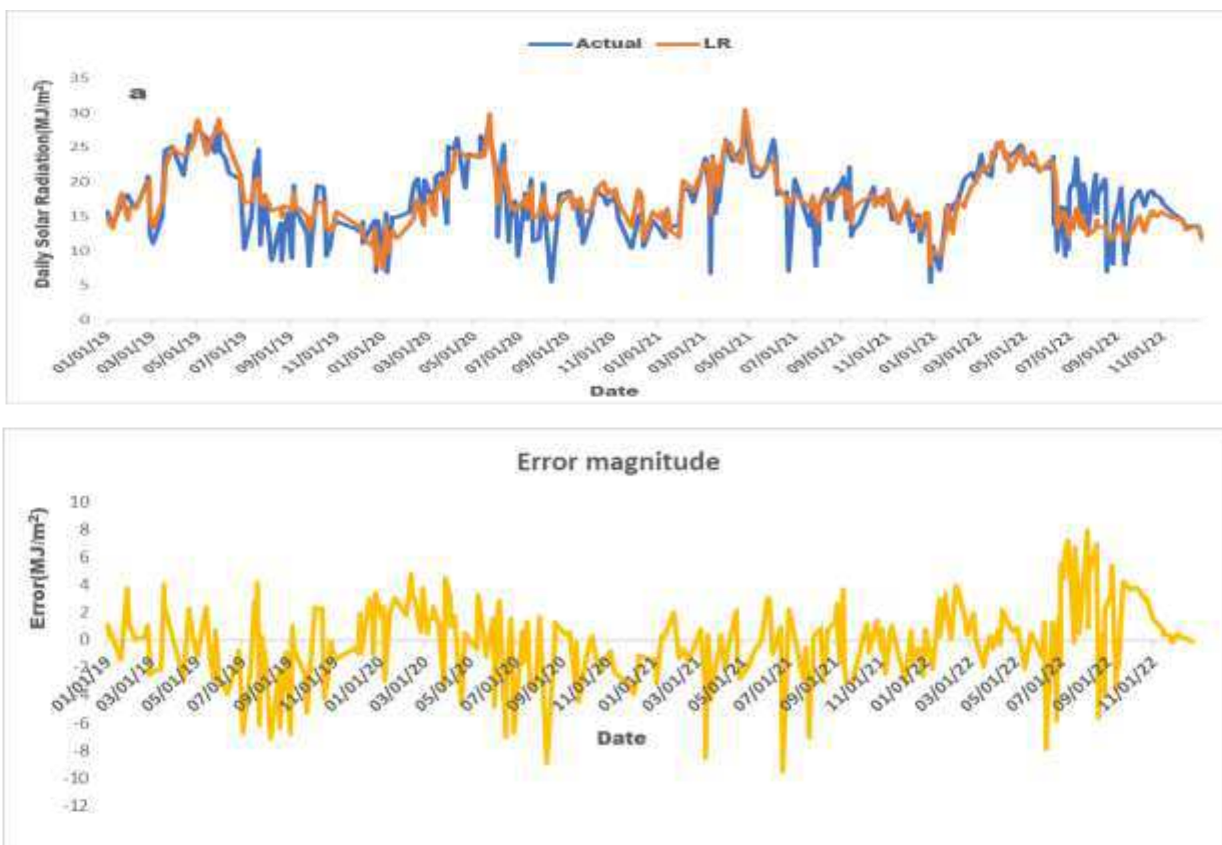


Figure 10 Actual vs predicted values and error metrics for the LR model- Rewa site

SVM algorithm

Fig. 11 shows the prediction results. SVM method results could be more satisfactory in this regard. Even the R^2(=0.8132) value and MAPE (= 0.1055) predicted by the SVM method for this site are not satisfactory compared to the predictions made for the Bhadla site. Also, it has the highest value of Max error (=18.0123 MJ/m2).

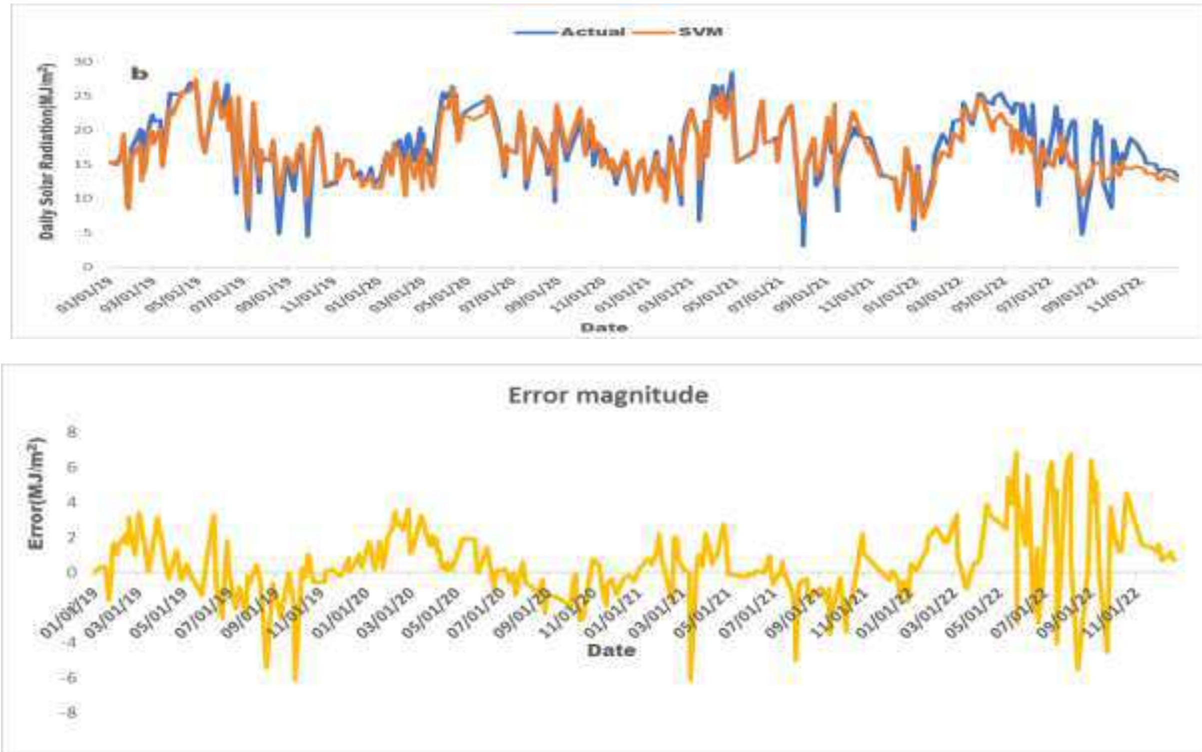


Figure 11 Actual vs predicted values and error metrics for the SVM model- Rewa site

RF algorithm

The prediction results are shown in Fig. 12. Rewa was identified as the site where both ANN and RF algorithms demonstrated the highest level of performance based on statistical parameters. The RF method, in particular, delivered highly satisfactory results for this site, with the greatest value of R^2 (=0.8983) and the lowest MBE (-0.0789 MJ/m^2).

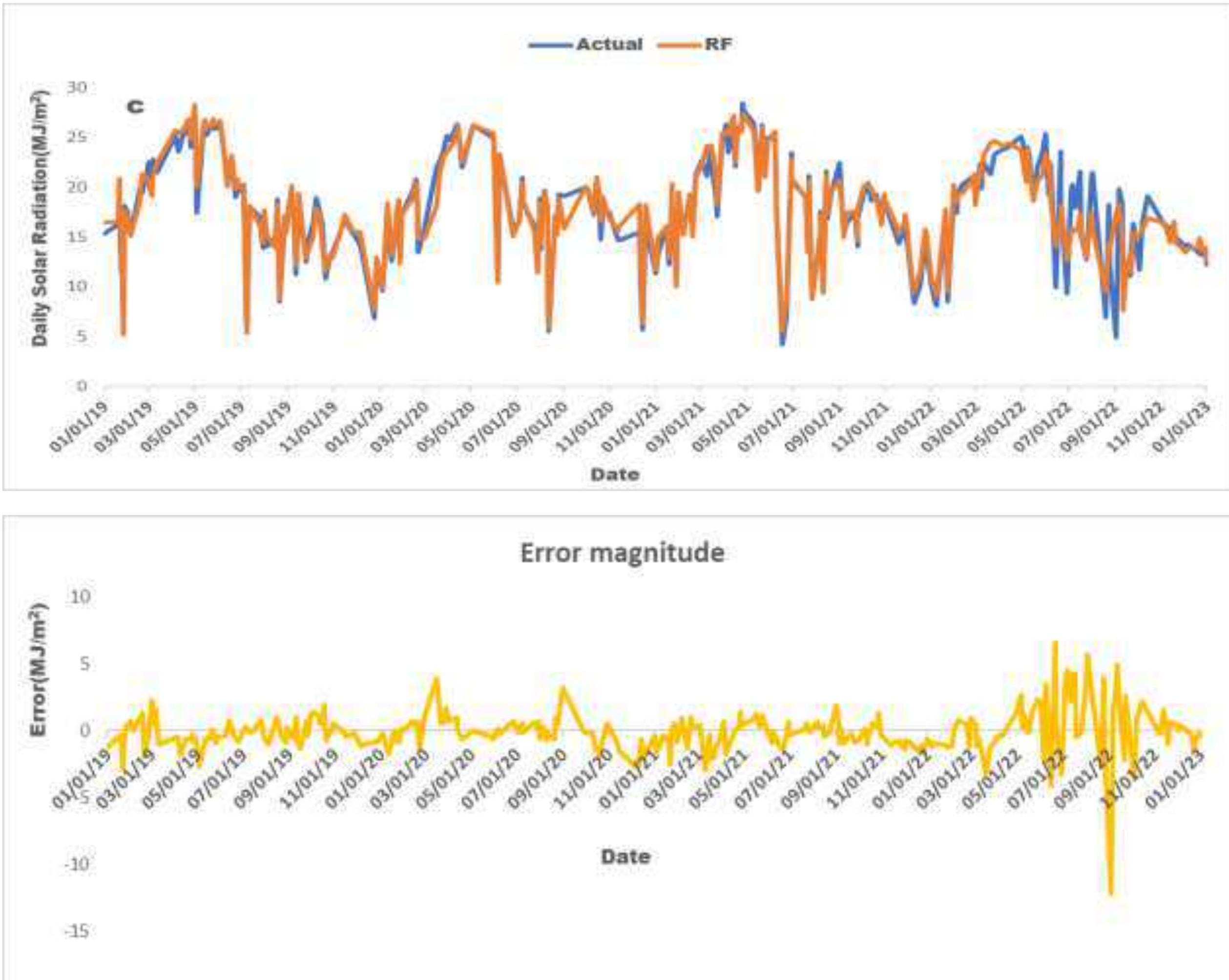
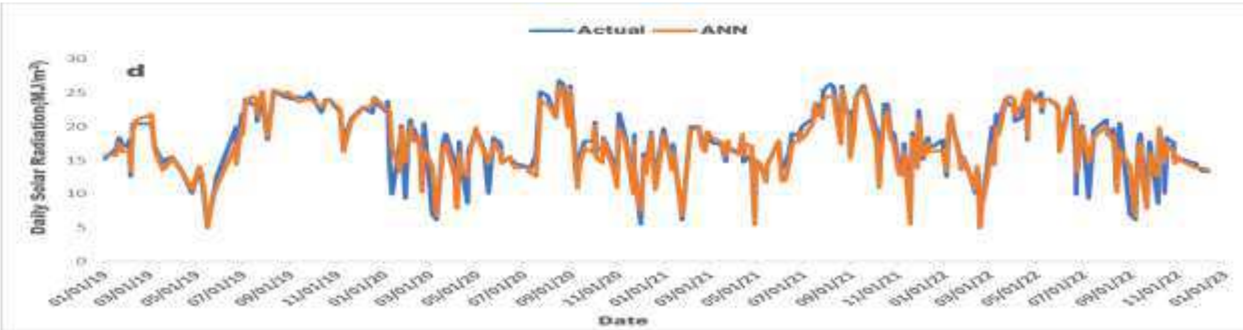


Figure 12 Actual vs predicted values and error metrics for the RF model- Rewa site

ANN algorithm

The prediction results are shown in Fig. 13. With R^2 (=0.9152) and MAPE (=0.06956), it is the best-performing algorithm for this site. However, it has a slightly higher value of MBE (=0.2792 MJ/m2) than the SVM algorithm. To give a general assessment of the Rewa site, ANN is the best-performing algorithm when all the statistical metrics are considered. However, the MBE value for RF is the least compared to all the other algorithms, but it's the second-performing algorithm considering R^2 and MAPE.



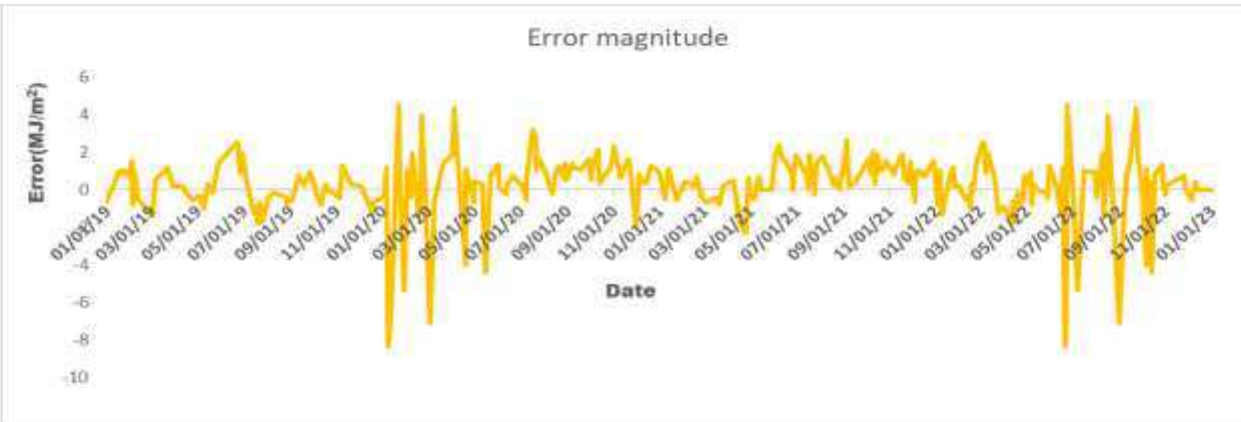


Figure 13 Actual vs predicted values and error metrics for the ANN model- Rewa site

The combined effect of all the algorithms

The combined performance of various algorithms for the Rewa site is shown in Fig. 14. For this site, the spread of data for LR is very high, i.e., the deviation from the actual values is high compared to all other algorithms for this site. It again indicates the worst performance of this algorithm. Also, the spread of ANN is significantly less, making it the best-performing algorithm.

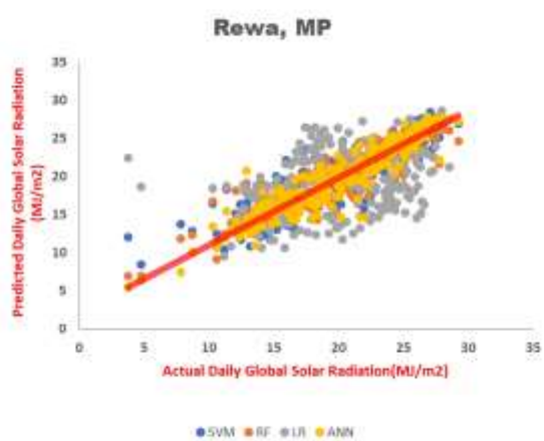
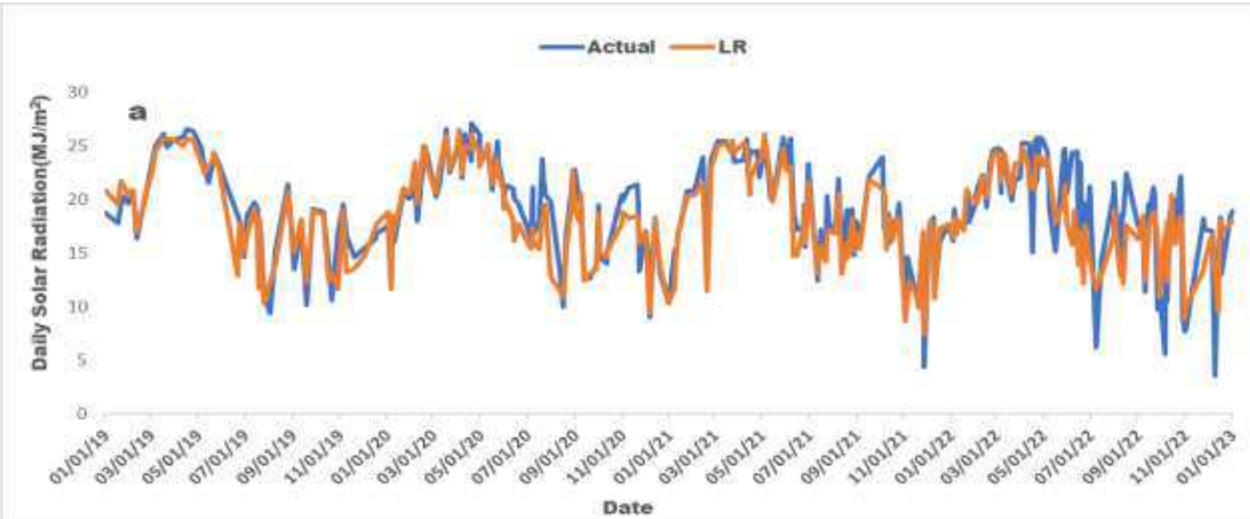


Figure 14 Combined effect of all the algorithms for Rewa site

Amguri, Sivasagar site, Assam

LR algorithm

The prediction results are shown in Fig. 15. The LR algorithm is again the worst-performing algorithm for this site. It has a very low R2 (=0.6108) value and a very high MAPE (=0.2248). The algorithm demonstrated the lowest MBE value of -0.1513 and also the Max. Error (=11.164) value is the highest of all the algorithms.



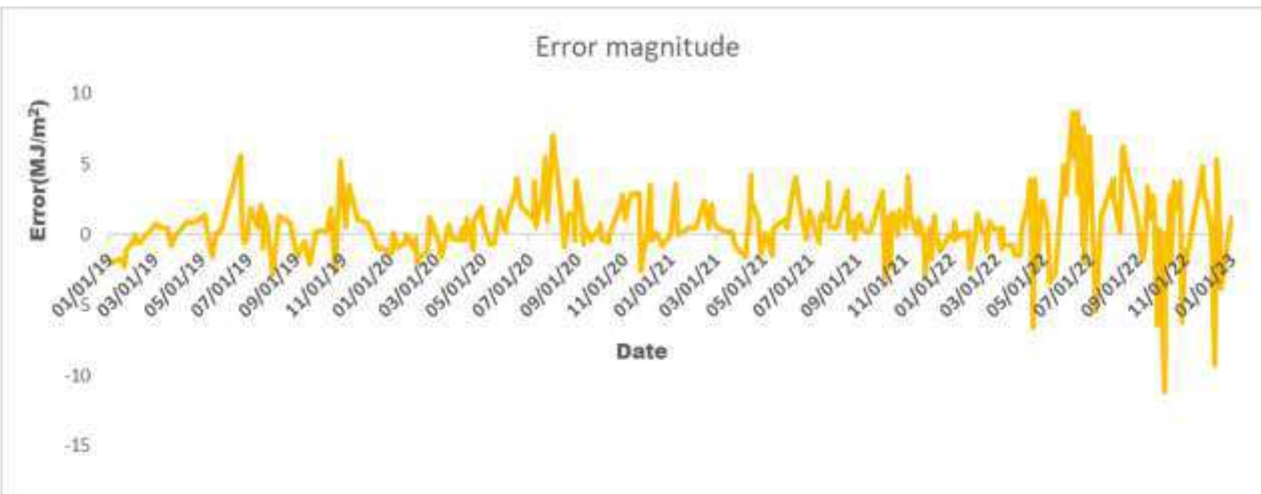


Figure 15 Actual vs predicted values and error metrics for the LR model- Amguri site

SVM algorithm

The prediction results are shown in Fig. 16. For this site, SVM is the second most successful algorithm after ANN regarding R^2. The R^2 value is relatively high, but the MAPE value is also high (=0.1137), which makes it lie under the category of good prediction. The value of MBE (=-0.6638) is also the highest for this algorithm.

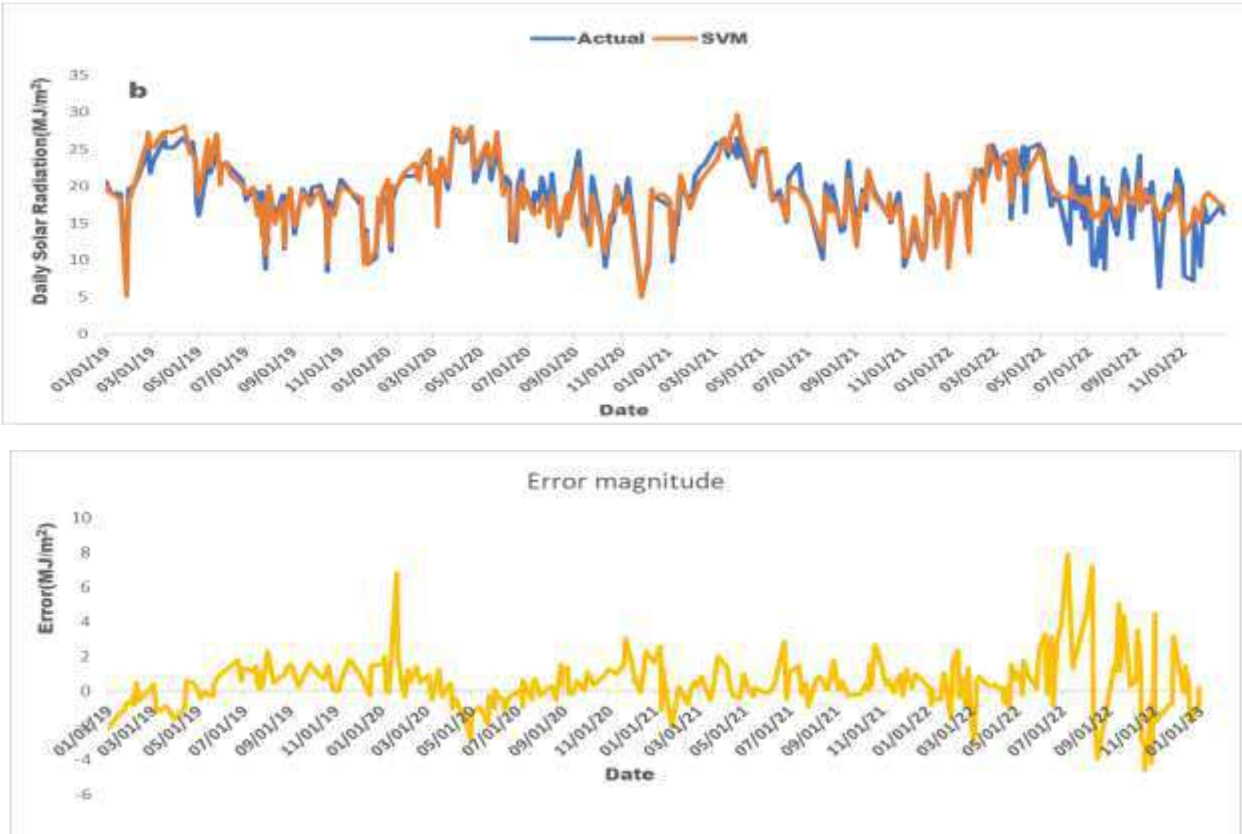
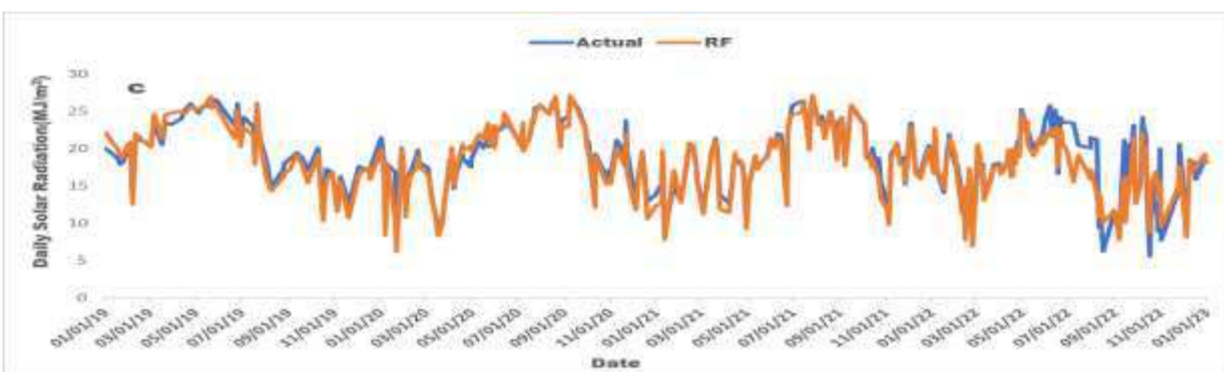


Figure 16 Actual vs predicted values and error metrics for the SVM model- Amguri site

RF algorithm

The prediction results are shown in Fig. 17. While the RF algorithm ranks third regarding R^2 metrics behind SVM and ANN, it has been the second most effective algorithm after ANN for both the Bhadla and Rewa sites based on performance. However, the model is not outperforming the SVM model for this site. The R^2 value is low (=0.7807), and it does not provide high prediction accuracy as per MAPE. However, it has the lowest MSE value, but this algorithm cannot be considered for this site.



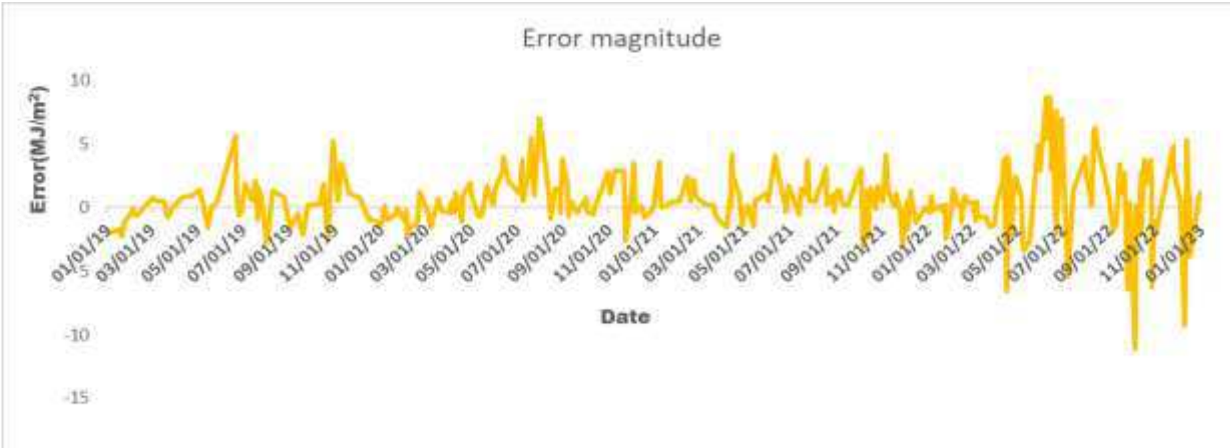


Figure 17 Actual vs predicted values and error metrics for the RF model- Amguri site

ANN algorithm

The prediction results are shown in Fig. 18. ANN is again the outperforming model for this site with a very high R^2 (=0.857) value and very low MAPE value (= 0.0623), which makes it fall under the category of a high prediction accuracy model. To give a general assessment of this site, ANN is the best-performing algorithm, but still, the R^2 value is low compared to the Bhadla and Rewa sites. All other algorithms have a very high value of MAPE for this site except ANN. RF is performing well in terms of MAPE but lags behind the SVM in terms of R^2.

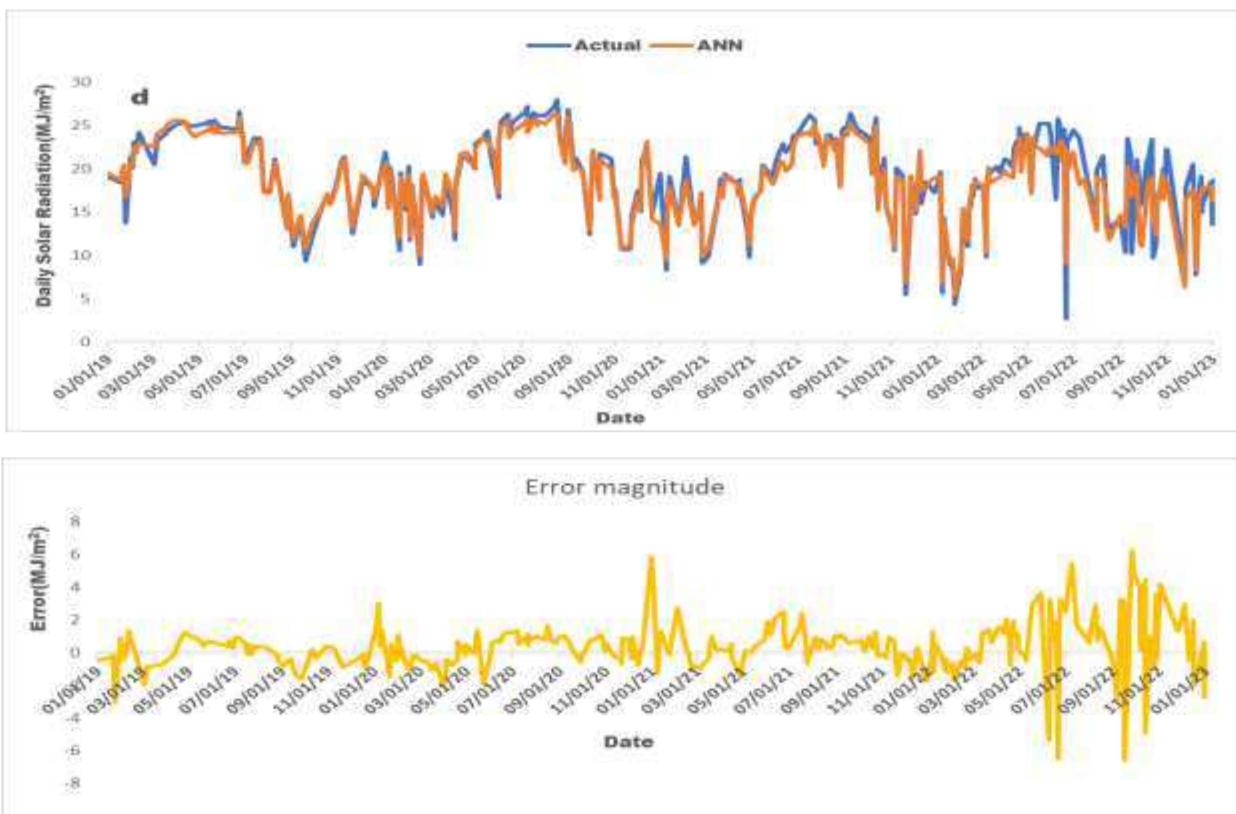


Figure 18 Actual vs predicted values and error metrics for the ANN model- Amguri site

The combined effect of all the algorithms

The combined performance of various algorithms for the Assam site is shown in Fig. 19. For this site, the spread of data for SVM is less as compared to the RF. It clearly indicates the out-performance of the SVM model after ANN. Again, the deviation for the LR model is very high compared to all the other models.

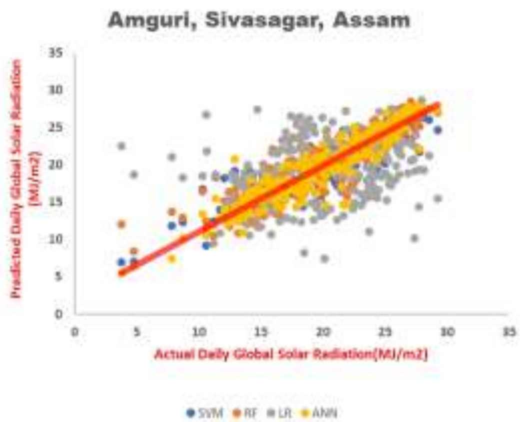


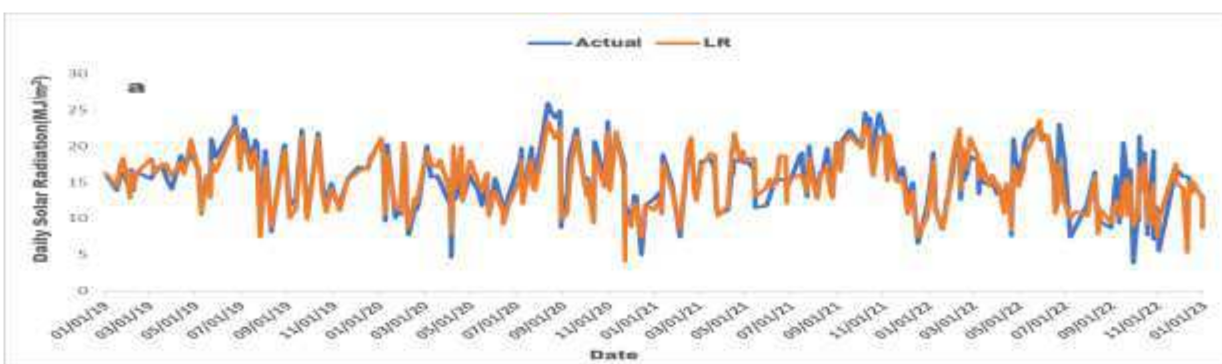
Figure 19 Combined effect of all the algorithms for Amguri site

Shillong site, Meghalaya

Shillong is the site with the lowest value of solar GHI.

LR algorithm

The results of prediction are shown in Fig. 20. In this case; the LR algorithm has outperformed the SVM algorithm regarding all the error metrics. The R^2 (=0.7959) value is high, and the MAPE (=0.09864) value is low compared to the SVM model. This is the main reason that the LR model was considered in this study. The superiority of a model and its ability to perform well for different sites can be distinct, as we have seen that the LR was the worst-performing algorithm for the above three sites. But, in this case, it has outperformed the Support Vector Machine technique, the second most effective technique for the Amguri Assam site.



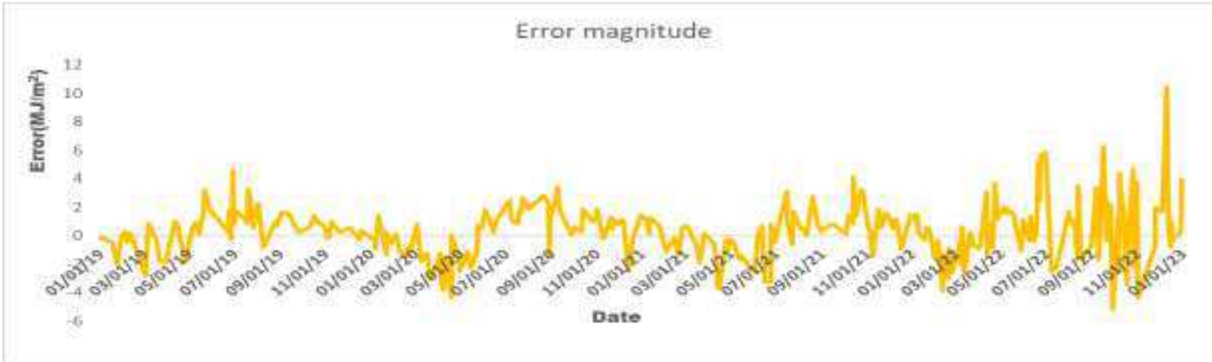


Figure 20 Actual vs predicted values and error metrics for the LR model- Shillong

SVM algorithm

The prediction results are shown in Fig. 21. The SVM algorithm is the worst-performing algorithm regarding this site's error metrics. The R^2 (=0.7092) value is the lowest, and its highest MAPE (=0.1082) value makes it fall outside the range of high prediction accuracy.

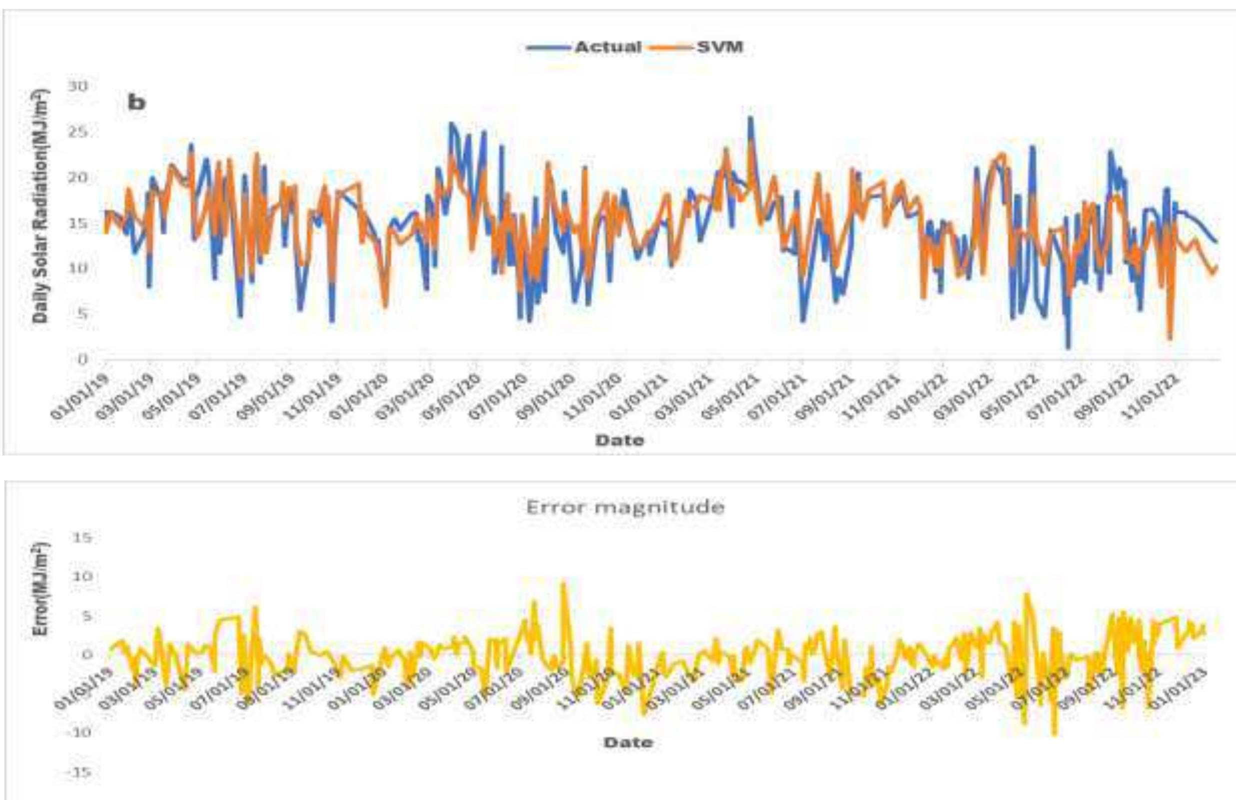


Figure 21 Actual vs predicted values and error metrics for the SVM model- Shillong

RF algorithm

The prediction results are shown in Fig. 22. RF is the second most successful algorithm for this site with a R^2 (=0.8774) and MAPE (=0.06434).

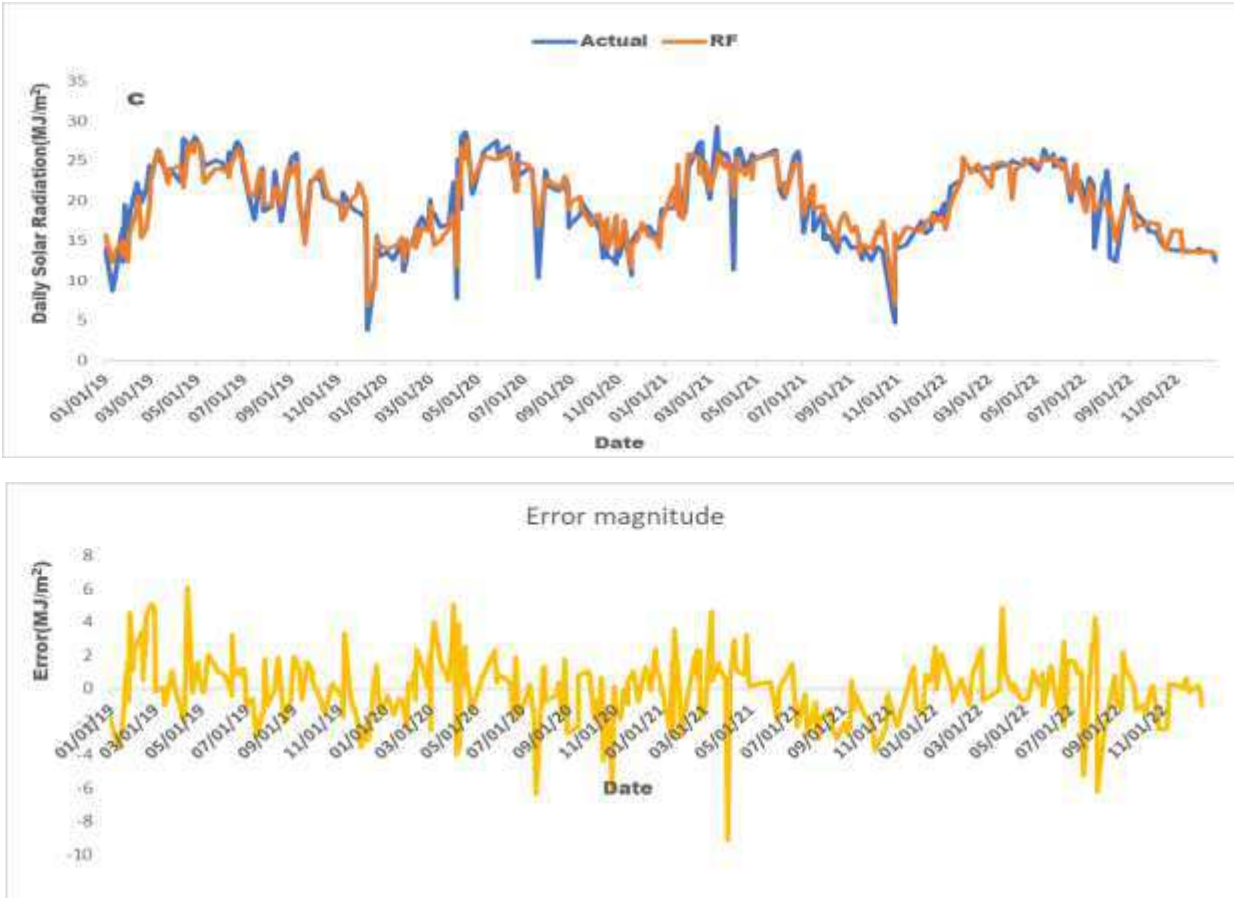
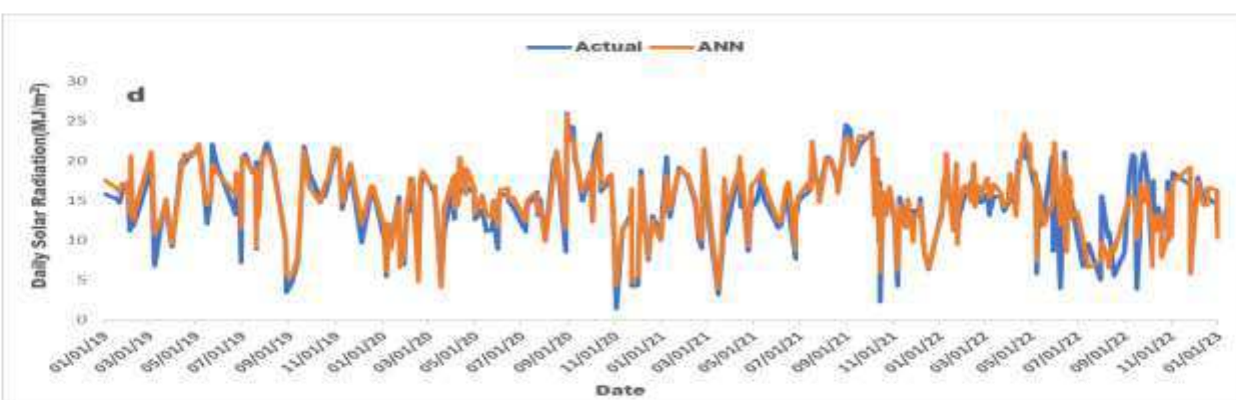


Figure 22 Actual vs predicted values and error metrics for the RF model- Shillong

ANN algorithm

The prediction results are shown in Fig. 23. ANN again out-wins in this case. It has an R^2 value of 0.9027, and its MAPE (=0.0432) is the lowest of all the instances considering all the sites. To give a general assessment of this site, ANN is again the best-performing algorithm. However, LR has been this site's third most successful algorithm after ANN and RF.



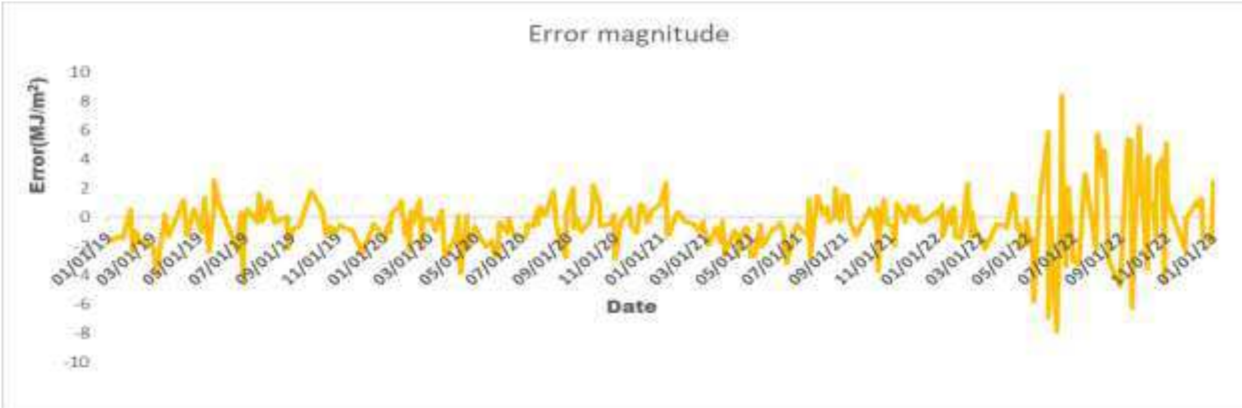


Figure 23 Actual vs predicted values and error metrics for the ANN model- Shillong

The combined effect of all the algorithms

The combined performance of various algorithms for Meghalaya is shown in Fig. 24. The LR model is more successful than the SVM model for this site. The spread of the SVM model data is large compared to all the other models for this site.

Figure 24 Combined effect of all the algorithms for Shillong

The net discussion regarding various model's performance is given in the table below-

Table 5 Various model's performance

Site	I	II	III	IV
Bhadla, Rajasthan	ANN	RF	SVM	LR
Rewa, Madhya Pradesh	ANN	RF	SVM	LR
Amguri, Sivasagar, Assam	ANN	SVM	RF	LR
Shillong, Meghalaya	ANN	RF	LR	SVM

I represent the best-performing model, and IV represents the worst-performing model.

4. Conclusion and future perspectives

Solar energy integration into the grid is challenging because of its uncertainty. Thus, AI techniques play a vital role in achieving it and ensuring grid parity. This paper's study uses four machine learning techniques (LR, SVM, RF, and ANN) for four different Indian states to represent India's overall solar GHI. Each of the sites chosen holds significance and illustrates a particular state. The detailed analysis was based on seven statistical metrics (MBE, MAE, MSE, RMSE, Max. Error, R^2, and MAPE). Based on the study's results, the following conclusions may be drawn: -

- The R^2 value for all the ML techniques ranges between 0.6108 to 0.9152. The worst value of R^2 (=0.6108) is for the LR model (Assam), and the highest value (=0.9152) is for the ANN model (Madhya Pradesh)
- The MAPE value for all the ML techniques ranges between 0.0432 to 0.2248. The highest MAPE value is 0.2248 for the LR model for the Rewa site, making it lie under the category of the worst prediction model. The ANN model gives the minimum MAPE value for the Shillong site. However, all the other models have also performed well regarding MAPE.
- The Max. Error value ranges from 6.355 MJ/m^2 for the ANN model-Bhadla site to 18.0123 MJ/m^2 for SVM model-Rewa site.
- Upon a general evaluation of MBE, it is observed that MBE ranges from -0.2271 to 0.63704 for all models and states. The worst value of MBE on the negative side is for the LR model (Shillong), and the positive side is for the SVM model (Shillong).
- The results of prediction by the Random Forest model are very close to the ANN model for Bhadla and Rewa sites regarding R^2 and MAPE. Therefore, other parameters need to be considered to evaluate and determine the success of these models.
- The Bhadla site outperforms the other three sites regarding overall statistical metrics.
- Considering all the statistical metrics, the ANN model has outperformed all the sites' models. The LR has been the worst-performing model for Bhadla, Rewa, and Amguri sites. However, it has exceeded the SVM model for Shillong.
- The study found that when considering the magnitudes of statistical errors, the ANN model exhibits lower error magnitudes than other models, particularly in observations with lower solar radiation.
- Based on the study's results, the ANN is the best-performing model for India. However, as a future perspective, while generalizing the ANN technique for prediction, the studies must be carried out for other sites in the selected states or other states representing India's overall solar radiation distribution.

Acknowledgements

The researchers express gratitude towards NASA for granting access to the data required for conducting their research and to the faculty members of NIT Bhopal for their support throughout the study. Additionally, the researchers extend their thanks to all the authors and co-authors cited in the references.

Authors' contributions

Bharat Girdhani contributed to the project's conceptualization, conducted the investigation and formal analysis, provided resources, and was responsible for drafting, editing, and revising the original manuscript. Meena Agrawal provided review and supervision for the project.

Funding: Not applicable.

Availability of data and materials

The National Aeronautics and Space Administration (NASA) provided the raw data for this study. The processed and derived data supporting this research's results and conclusions can be obtained from the corresponding author upon reasonable request.

Declarations

Conflict of interest

The authors affirm that they do not have any known financial interests or personal relationships that could have potentially influenced the findings presented in this paper.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable

References

Ağbulut Ü, Gürel AE, Biçen Y (2021) Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. Renew Sustain Energy Rev 135:. https://doi.org/10.1016/j.rser.2020.110114

Alkhayat G, Mehmood R (2021a) A review and taxonomy of wind and solar energy forecasting methods based on deep learning. Energy AI 4:100060. https://doi.org/10.1016/j.egyai.2021.100060

Alkhayat G, Mehmood R (2021b) Energy and AI A review and taxonomy of wind and solar energy forecasting methods based on deep learning. Energy AI 4:100060. https://doi.org/10.1016/j.egyai.2021.100060

Azlah MAF, Chua LS, Rahmad FR, et al (2019) Review on techniques for plant leaf classification and recognition. Computers 8:. https://doi.org/10.3390/computers8040077

Behrang MA, Assareh E, Ghanbarzadeh A, Noghrehabadi AR (2010) The potential of different artificial neural network (ANN) techniques in daily global solar radiation modeling based on meteorological data. Sol Energy 84:1468–1480. https://doi.org/10.1016/j.solener.2010.05.009

Benali L, Notton G, Fouilloy A, et al (2019) Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components. Renew Energy 132:871–884. https://doi.org/10.1016/j.renene.2018.08.044

Breiman L (2001) Random Forests. https://doi.org/https://doi.org/10.1023/A:1010933404324

Corinna Cortes & Vladimir Vapnik (1995) Support-vector networks. Springer. https://doi.org/https://doi.org/10.1007/BF00994018

Elchinger M (2016) India Solar Resource Data: Enhanced Data for Accelerated Deployment (Fact Sheet), NREL (National Renewable Energy Laboratory). 80401

Emang D, Shitan M, Abd. Ghani AN, Noor KM (2010) Forecasting with Univariate Time Series Models: A Case of Export Demand for Peninsular Malaysia's Moulding and Chipboard. J Sustain Dev 3:157–161. https://doi.org/10.5539/jsd.v3n3p157

Essam Y, Ahmed AN, Ramli R, et al (2022) Investigating photovoltaic solar power output forecasting using machine learning algorithms. Eng Appl Comput Fluid Mech 16:2002–2034. https://doi.org/10.1080/19942060.2022.2126528

Fan J, Wang X, Wu L, et al (2018) New combined models for estimating daily global solar radiation based on sunshine duration in humid regions: A case study in South China. Energy Convers Manag 156:618–625. https://doi.org/10.1016/j.enconman.2017.11.085

Feng Y, Hao W, Li H, et al (2020) Machine learning models to quantify and map daily global solar radiation and photovoltaic power. Renew Sustain Energy Rev 118:109393. https://doi.org/10.1016/j.rser.2019.109393

Gouda SG, Hussein Z, Luo S, Yuan Q (2019) Model selection for accurate daily global solar radiation prediction in China. J Clean Prod 221:132–144. https://doi.org/10.1016/j.jclepro.2019.02.211

Gürel AE, Ağbulut Ü, Biçen Y (2020) Assessment of machine learning, time series, response surface methodology and empirical models in prediction of global solar radiation. J Clean Prod 277:. https://doi.org/10.1016/j.jclepro.2020.122353

Hastie T, Tibshirani R, James G, Witten D (2021) An introduction to statistical learning (2nd ed.). Springer texts 102:618

Jebli I, Belouadha FZ, Kabbaj MI, Tilioua A (2021) Prediction of solar energy guided by pearson correlation using machine learning. Energy 224:120109. https://doi.org/10.1016/j.energy.2021.120109

Jiang Y (2008) Prediction of monthly mean daily diffuse solar radiation using artificial neural networks and comparison with other empirical models. Energy Policy 36:3833–3837. https://doi.org/10.1016/j.enpol.2008.06.030

Khastagir A, Hossain I, Anwar AHMF (2022) Efficacy of linear multiple regression and artificial neural network for long-term rainfall forecasting in Western Australia. Meteorol Atmos Phys 134:1–11. https://doi.org/10.1007/s00703-022-00907-4

Kim SG, Jung JY, Sim MK (2019) A two-step approach to solar power generation prediction based on weather data using machine learning. Sustain 11:. https://doi.org/10.3390/SU11051501

Kirmani S, Jamil M, Rizwan M (2015) Empirical correlation of estimating global solar radiation using meteorological parameters. Int J Sustain Energy 34:327–339. https://doi.org/10.1080/14786451.2013.826222

Lahouar A, Mejri A, Ben Hadj Slama J (2017) Importance based selection method for day-ahead photovoltaic power forecast using random forests. Int Conf Green Energy Convers Syst GECS 2017. https://doi.org/10.1109/GECS.2017.8066171

Lai J, Chang Y, Chen C, Pai P (2020) applied sciences A Survey of Machine Learning Models in Renewable Energy Predictions

Liu H, Lang B (2019) Machine learning and deep learning methods for intrusion detection systems: A survey. Appl Sci 9:. https://doi.org/10.3390/app9204396

Makade RG, Chakrabarti S, Jamil B (2021) Development of global solar radiation models: A comprehensive review and statistical analysis for Indian regions. J Clean Prod 293:126208. https://doi.org/10.1016/j.jclepro.2021.126208

Mehdizadeh S, Behmanesh J, Khalili K (2016) Comparison of artificial intelligence methods and empirical equations to estimate daily solar radiation. J Atmos Solar-Terrestrial Phys 146:215–227. https://doi.org/10.1016/j.jastp.2016.06.006

Monjoly S, André M, Calif R, Soubdhan T (2017) Hourly forecasting of global solar radiation based on multiscale decomposition methods: A hybrid approach. Energy 119:288–298. https://doi.org/10.1016/j.energy.2016.11.061

Quej VH, Almorox J, Arnaldo JA, Saito L (2017) ANFIS, SVM and ANN soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment. J Atmos Solar-Terrestrial Phys 155:62–70. https://doi.org/10.1016/j.jastp.2017.02.002

Ramil A, López AJ, Pozo-Antonio JS, Rivas T (2018) A computer vision system for identification of granite-forming minerals based on RGB data and artificial neural networks. Meas J Int Meas Confed 117:90–95. https://doi.org/10.1016/j.measurement.2017.12.006

Rehman S (1998) Solar radiation over Saudi Arabia and comparisons with empirical models. Energy 23:1077–1082. https://doi.org/10.1016/S0360-5442(98)00057-7

Sharifi SS, Rezaverdinejad V, Nourani V (2016) Estimation of daily global solar radiation using wavelet regression, ANN, GEP and empirical models: A comparative study of selected temperature-based approaches. J Atmos Solar-Terrestrial Phys 149:131–145. https://doi.org/10.1016/j.jastp.2016.10.008

Wang F, Mi Z, Su S, Zhao H (2012) Short-term solar irradiance forecasting model based on artificial neural network using statistical feature parameters. Energies 5:1355–1370. https://doi.org/10.3390/en5051355

Yadav O, Kannan R, Meraj ST, Masaoud A (2022) Machine Learning Based Prediction of Output PV Power in India and Malaysia with the Use of Statistical Regression. Math Probl Eng 2022:. https://doi.org/10.1155/2022/5680635

Yang L, Cao Q, Yu Y, Liu Y (2020) Comparison of daily diffuse radiation models in regions of China without solar radiation measurement. Energy 191:116571. https://doi.org/10.1016/j.energy.2019.116571

Yıldırım HB, Çelik Ö, Teke A, Barutçu B (2018) Estimating daily Global solar radiation with graphical user interface in Eastern Mediterranean region of Turkey. Renew Sustain Energy Rev 82:1528–1537. https://doi.org/10.1016/j.rser.2017.06.030

Zang H, Cheng L, Ding T, et al (2020) Application of functional deep belief network for estimating daily global solar radiation: A case study in China. Energy 191:116502. https://doi.org/10.1016/j.energy.2019.116502

s