Iranian Journal of Science and Technology, Transactions of Electrical Engineering (2024) 48:1249–1260
https://doi.org/10.1007/s40998-024-00716-y

RESEARCH PAPER

Binning Based Data Driven Machine Learning Models for Solar Radiation Forecasting in India

Anuradha Munshi[1] · R. M. Moharil[1]

Received: 26 April 2022 / Accepted: 19 February 2024 / Published online: 26 March 2024

Abstract
Energy is the primary driving force in improvement of the human life cycle. All the activities for the betterment of human life are dependent on some form of energy. Conventional energy sources rely on fossil fuels which have limited reserves and we are bound to exhaust them soon. On the other hand, non-conventional/renewable energy sources are produced on a regular basis and are clean without any polluting emissions. These sources include solar, wind, hydraulic, biomass/bio gas, geothermal, tidal, etc. Solar energy is one of the primary sources in countries like India, but it does have drawbacks like high initial cost, dependency on weather, expensive storage, space requirement, etc. It is therefore imperative to create accurate solar radiation forecasting models to identify and address these issues. Forecasting models are created based on daily or hourly data and are location specific. In this work, binning based machine learning models are proposed for accurately forecasting hourly solar radiation. These models are data driven clustering based models. The clusters are identified based on geographic locations. The proposed approach also helps reduce the number of required models without compromising the high accuracy. In this work, global and diffuse solar radiation data, gathered from five geographically distinct stations from India, is analyzed. Validation of these models demonstrate increased performance. The number models required are also significantly smaller compared to the daily or hourly models.

Keywords Binning · Climate variables · Machine learning · Solar radiation forecasting · Sunshine hour

1 Introduction

Solar energy is one of the cleanest forms of energy available among all renewable sources in India. A few advantages of solar energy are consistency of its availability, accessibility, low maintenance costs, and long-term reliability. Specially, in last few years solar energy has had a visible impact on India's energy sector. According to Ministry of New and Renewable Energy (MNRE) data, nearly 5000 trillion kWh energy is incident over India's land area per year with most areas receiving 4–7 kWh per sq. m per day. Although abundance of solar energy is one of its key advantages, high initial cost makes accurately forecasting generation capacity crucial. Forecasting solar radiation is a difficult problem to solve owing to the intermittent nature of solar energy, and its variability due to irradiance, temperature, humidity, pressure, clouds, etc. add to the difficulty.

Several models are available in the literature to predict solar radiation using various climatic parameters, like, sunshine duration, altitude, longitude, temperatures, humidity, wind speeds, clearness indexes, soil temperatures, pressure, etc. Numerous empirical equations have also been proposed to predict solar radiation; however, the variables involved are extremely uncertain, which make accuracy of these models unreliable. A lot of prediction models, based on various techniques are reported in the literature that try to improve the prediction accuracy. Two primary research areas are solar radiation forecasting and solar power generation forecasting. Forecasting the capacity even before generation is referred to as solar irradiance forecasting. In this area, special attention has been given to diffuse and global radiation forecasting. Metrological or climate data for one or more stations has

✉ Anuradha Munshi
aam.nyss@gmail.com

[1] Department of Electrical Engineering, Yeshwantrao Chavan College of Engineering (YCCE), Hingna Rd, Nagpur, Wanadongri, Maharashtra 441110, India

been analyzed through empirical equations modelling, machine learning models, and even through some hybrid combination.

Artificial neural network (ANN) is employed to estimate the monthly mean daily diffuse solar radiation in Jiang (2008). They gathered the diffuse solar radiation data from nine distinct locations with varying climatic conditions. Structure of such a network dictates the forecasting accuracy. The work in Kashyap et al. (2015) takes this research further by analyzing solar radiation forecasting with multi-parameter neural network. An overview of forecasting methods of solar irradiation using various machine learning approaches is available in Voyant et al. (2017). Sunshine based empirical models for daily global solar radiation estimation in China are evaluated in Fan et al. (2018). They analyzed meteorological solar radiation data between 1966 and 2015 from twenty stations in humid regions. They have used four statistical indicators and a global performance index. Support vector machine (SVM) based approaches for forecasting solar irradiance are available in Melzi et al. (2016); Belaid and Mellit (2016). Group method of data handling (GMDH), that comprises of models such as, multilayer feed-forward neural network (MLFFNN), adaptive neuro fuzzy inference system (ANFIS), particle swarm optimization (PSO), etc. The data from twelve sites across Iran's various climate zones has been analyzed (Khosravi et al. 2018). A novel model for the computation of global solar radiation on the horizontal surface in Muğla/Turkey is presented in Bayrakçı et al. (2018) along with a comparison against the empirical models from literature. There are several studies to improve the solar radiation prediction accuracy using empirical as well as ANN models (Jahani and Mohammadi 2019; Feng et al. 2019; Kim et al. 2019). A study on artificial intelligence (AI) applications with focus on machine learning (ML), deep learning (DL), and hybrid methods is presented in Mellit et al. (2020). Meteorological parameters have also been studied using hybrid models with SVM and empirical models to improve forecasting accuracy (Liu et al. 2020; Gürel et al. 2020). A feed forward back propagation three layered neural network has been employed for solar radiation forecasting for fourteen stations in Uttar Pradesh, India (Choudhary et al. 2020). A study in Álvarez-Alvarado et al. (2021) provides insight on finding the optimal parameters to reduce prediction error using the meteorological factors. They make use of several hybrid SVM models that use the search optimization algorithm (SOA). Comparison of some of the most widely used machine learning algorithms, namely, SVM, ANN, and extreme learning machine (ELM) to achieve best daily solar prediction is presented in de Freitas Viscondi and Alves-Souza (2021). A big dataset containing daily solar radiation data gathered from NASA's POWER project repository over a 36-year period (1983–2019) from two sites in India is used to develop DL based models to estimate daily solar irradiance (Brahma and Wadhvani 2020).

A comparative study of several forecasting strategies and numerous successful uses of solar forecasting methods at the utility scale is presented in Inman et al. (2013) from the perspective of both the solar resource and the electricity output of solar plants. Two years of solar radiation data collected from Macau has been analyzed using data mining techniques like, ANN, SVM, k-Nearest Neighbour, and Multivariate Linear Regression (MLR) to estimate daily solar power output of up to 3 days (Long et al. 2014). Global ensemble forecast system (GEFS) is used as basis for deriving the Numerical weather prediction (NWP) models (Aler et al. 2015). They are tested on various nodes in a grid to study performance of several machine learning algorithms for forecasting. Renewable energy management centers (REMCs) are proposed in Mitra et al. (2016). The centres would be co-located with the load dispatch centres and will be responsible for a handful of tasks including forecasting. A partial functional linear regression model (PFLRM) which is the generalization of the classic multiple linear regression model with the nonlinearity structures is proposed in Wang et al. (2016). A comparative study of deterministic and stochastic models for day-ahead projections is presented in Ogliari et al. (2017). A novel hybrid approach for PV forecasting in Massucco et al. (2019) makes use of a decision tree. A selection is made among clear-sky models and an ensemble of artificial neural networks. Another hybrid approach involving the convolutional neural networks (CNN) and long-short term memory recurrent neural networks (LSTM) is available in Li et al. (2020). A multivariate strategy based on the LSTMs to anticipate short-term solar power generation is proposed in Ahmad and Kumar (2021).

A high-quality measured data for meteorological factors in Qassim, Saudi Arabia comprising of numerous climatic factors is analyzed using ensemble tree-based machine learning approach in Alaraj et al. (2021). Twenty two multivariate numerical models that incorporate solar radiation, temperature, cloud cover, sunshine, humidity, and wind speed are made available in Son and Jung (2021) to produce an effective energy management system (EMS). A modified version of LSTM technique is used to compare the models' performance. A new system based on mathematical probability density functions, climatic factors, and DL methods is proposed in Rodríguez et al. (2022). Rather than relying solely on massive data and ML algorithms, Luo et al. (2021) presents a physics-constrained LSTM (PC-LSTM) system to forecast PV generation on an hourly day-ahead basis. A comparative study of performance of several current ML methods for hourly prediction is available in Chahboun and Maaroufi (2021). Methods

compared include Bayesian regularised neural networks, random forest, k-nearest neighbours, gradient boosting, SVM, etc. Another study (Zulkifly et al. 2021) has investigated SVM, Decision Trees, Linear regression, Gaussian Process Regression (GPR), etc. based on high-quality measured data. A daily prediction model based on weather forecast information from Korea is proposed in Kim et al. (2017). This model is also integrated in a commercially available solar PV monitoring system.

All these studies try and find a suitable model for a given database. Furthermore, the models created are based on daily or hourly solar radiation data. In this work a novel approach is proposed which provides following key advantages.

1. Analysis of the data for trends based on time of the day and geographical locations. Exploratory data analysis (EDA) is performed to determine the dominant features present in the data.
2. The proposed models are built based on these dominant features instead of basing them on daily or hourly station data.
3. The proposed models decrease the number of models significantly and helps to generalize them for stations with similar dominant features.

We analyze global as well as the diffuse solar radiation data gathered from five geographically distinct stations in India. We validate the proposed models and demonstrate their increased performance compared to the daily as well as hourly data based models. Rest of this paper is organized as follows. Section 2 provides a brief information on the collected data, terminologies used, geographical significance of each station, and exploratory data analysis. Section 3 describes the proposed method in detail, followed by quantitative evaluation through results in Sect. 4. Finally, Sect. 5 concludes the paper with an analysis of the results and a brief discussion on the future work.

2 Data and Terminology

The data used in this work consists of solar irradiance hourly data from 2007 to 2019 for five different stations namely New Delhi, Ahmedabad, Kolkata, Goa-Panjim and Thiruvanthapuram. For each of the stations, following surface data parameters are available (please refer to Table 1).

Table 1 Nomenclature

Variable	Description
SLP	Station level pressure (in hpa)
DBT	Dry bulb temperature (in °C)
DPT	Dew point temperature (in °C)
VP	Vapour pressure (in hpa)
FFF	Wind speed (in kmph)
VV	Visibility (in code)
Dl	Direction of low cloud (in 8 points of compass)
Dh	Direction of high cloud (in 8 points of compass)
Ht	Height of base individual cloud layer (in code)
MSLP	Mean sea-level pressure (in hpa)
WBT	Wet bulb temperature (in °C)
RH	Relative humidity (in %)
DD	Wind direction (in 16 points of compass)
AW	Average wind speed (in kmph)
h	Height of lowest cloud (in code)
Dm	Direction of medium cloud (in 8 points of compass)
TC	Total amount of clouds (in oktas)
RF	Rainfall (in mm) since previous observation
Cl	Form of low cloud
A	Amount of low cloud (in oktas)
Cm	Form of medium cloud (in codes)
A	Amount of medium cloud (in oktas)
Ch	Form of high cloud (in code)
A	Amount of high cloud (in oktas)
c	Form of individual layer of cloud (in code)
a	Amount of individual layer of cloud (in oktas)
EVP	Total evaporation (in mm)
DW	Direction of wind wave (in 16 points of compass)
P	Period of wave (in code)
H	Height of wind wave (in code)

2.1 Station Summary

The five stations under consideration are in very different geographical conditions. A summary of their location parameters is tabulated above (please refer to Table 2). A detailed description of the geographical significance of each station is mentioned below. The five stations under consideration are in very different geographical conditions. A summary of their location parameters is tabulated above (please refer to Table 2). A detailed description of the geographical significance of each station is tabulated in Table 3.

2.2 Data Cleaning

There are some erratic entries in the database as well as some missing values owing to machine and human errors. These extreme values as well as the missing values are first replaced by meaningful values using data imputation which makes use of the median of all available values.

Table 2 Geographical details of 5 stations under study

SR. No	Station Name	Latitude	Longitude	Elevation above sea level
1	New Delhi	28° 38′ 08″ N	77° 13′ 28″ E	212 m = 695 ft
2	Ahmedabad	23° 01′ 32″ N	72° 35′ 14″ E	56 m = 183 ft
3	Kolkata	22° 33′ 45″ N	88° 21′ 46″ E	11 m = 36 ft
4	Goa-Panjim	15° 29′ 44″ N	73° 49′ 34″ E	16 m = 52 ft
5	Thiruvanthapuram	08° 29′ 07″ N	76° 56′ 57″ E	18 m = 59 ft

Table 3 Geographical features and their significance on the five stations under study

	New Delhi	Ahmedabad	Kolkata	Goa-Panjim	Thiruvanthpuram
Location in India	North	West	East	Westcoast of the subcontinent	South
Major Geographical features	Dominated by river Yamuna, and the Aravalli range, and the plains in between	Located on the banks of the River Sabarmati	Spread linearly along the banks of Hooghly River	State is bordered in the west by the Arabian Sea. Panaji lies on the left bank of the River Mandovi	Built on seven hills by the seashore on the west coast, near the southern tip of mainland India
Average Rainfall	714 mm	932 mm	1582 mm	2813 mm	1500 mm
Weather conditions in Summer	Temperatures range is 40–45 °C	Temperature range is 23–43 °C	Max. temperatures can exceed 40 °C during May–June	Temperature range is 21–26 °C	It is hot tropical region February to May. temperature goes up to 35 °C
Weather conditions in Monsoon	Monsoon starts in late June and lasts until mid-September temperatures range is 25–32 °C	The south-west monsoon winds bring humid climate to Ahmedabad from mid-June to mid-September	Maximum rainfall occurs during the monsoon in August	Maximum rainfall occurs between the months of June and September	Average temperature varying between 23.2 and 29.8°C
Weather conditions in Winter	temperatures can fall to 4–5°C	November to February temperature 15–36 °C Cold northerly winds are responsible for a mild chill during January	Winter tends to last from December to early-February, with the min temperatures of 12–14 °C during December and January	Winter season, the temperatures do not fall to a great extent	December to February average temperature goes down to 20 °C
Peculiarity	Far from the sea and hemmed by mountains-creates an unusual dry continental climate in a largely humid, subtropical region	Two main lakes located in the city limits - the Kankaria lake, and the Vastapur lake	Often during early summer, spells of thunderstorm and heavy rains lashes the city	The winds from the sea influence the day and night temperatures of the city	The city is on the west coast of India and is bounded by the Laccadive Sea to its west and the Western Ghats to its east. relative humidity is generally high

2.3 Exploratory Data Analysis

To create meaningful machine learning models, it is imperative to gain a better understanding of the data. Identification of primary features form a long list of features (i.e., feature selection) is crucial. Therefore, we find correlation amongst features to estimate the relationships amongst features. Correlation heatmaps for global and diffuse radiations for a couple of stations, namely, New Delhi and Ahmedabad are illustrated in Fig. 1.

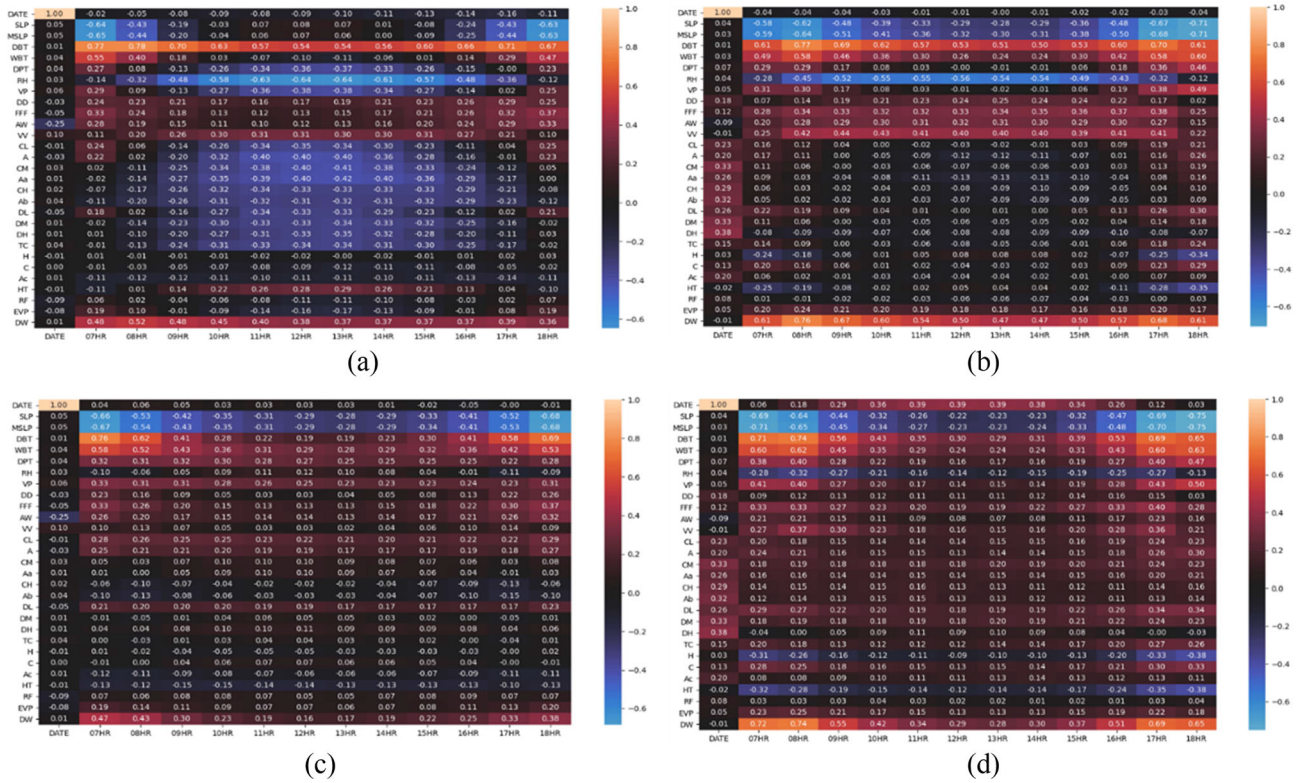

Fig. 1 Global radiation correlation heatmap with different features for **a** Ahmedabad and **b** New Delhi, and Diffuse radiation correlation heatmap with different features for **c** Ahmedabad and **d** New Delhi

3 Proposed Method

In this section, we discuss the steps involved in the proposed method. Identification of important features and feature space reduction is crucial for creating meaningful and accurate forecasting models. It helps reduce the training time and helps remove the less important data.

3.1 Feature Selection

While developing forecasting models for the solar radiation data, we carried out the feature selection process. We employ the SelectKBest algorithm from Scikit-learn API to identify important features. This algorithm selects the best features using K highest scores. We set the value of K to number of features available to us to get the scores for all the features. The resulting graphs for Ahmedabad station are illustrated in Fig. 2.

3.2 Binning Approach

In this work, we propose binning approach-based machine learning models. Binning refers to dividing or partitioning the input space into subspaces and developing a model for each of these subspaces. Collectively, these models cover the entire input space. In Mendhurwar et al. (2008), authors use a similar technique to model a complex input space with a highly nonlinear input parameter. They remove this parameter from the model and then use it to partition the input space in uniform grids along its axis. This parameter is referred to as the binning parameter.

For instance, if vector $\boldsymbol{x} = [x_1, x_2, x_3, \ldots, x_n]$ represents the input space (please refer to Table 1 for input parameters), one of the parameters x_i is identified as the binning parameter. This parameter is then removed from the input space and a separate model is developed for various values/ranges of this parameter with $n-1$ inputs. Selection of x_i is based on the data and the sensitivity of the output to the input parameters. Typically, input parameter, to which the output is most sensitive to, is selected as the binning parameter. If a model is given by equation $y = f(\boldsymbol{x})$, and if x_3 is the chosen binning parameter, then the binned models y_i for all the i values/ranges of are represented by equation 1. The concept is also illustrated in Fig 4.

$$y_i = f(\hat{x}) | \forall i \in x_3 \tag{1}$$

where,

$$\hat{x} = [x_1, x_2, x_4, \ldots, x_n] \tag{2}$$

Fig. 2 Feature selection results on Ahmedabad station data using K-Best method

All the features listed in Table 1 were considered for the choice of the binning parameter. We needed a parameter with a well defined range as well as intervals. We noticed that that the weather parameters had a wider range and they also had the possibility of outliers due to measurement/human error. All the measurements; however, had the hourly timestamps and their ranges were similar in certain time intervals (see Figs. 2 and 3). We used the feature selection results to validate the choice of hour of the day as the binning parameter. It is evident from Fig. 2 that the graphs plotted on hourly basis can be grouped together based on the dominant features. For instance, the plots between 9am to 4pm (see Fig. 2c–j) all have same dominant features. Similarly, plots for 5pm, 6pm and 7pm (see Fig. 2k, l, and a) also have same dominant features. To confirm our hypothesis, we applied two additional feature selection methods, namely, Pearson and Spearman. These methods also provided us with the same set of features. Feature selection results obtained using all the feature selection methods on New Delhi station dataset at 2.00 PM is illustrated in Fig. 3. As such, we use hour of the day as

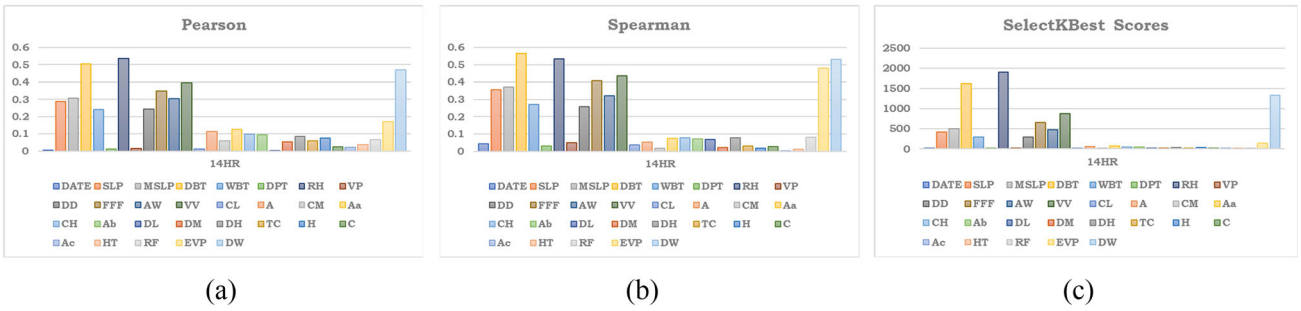

Fig. 3 Comparison of selected features obtained from **a** Pearson, **b** Spearman, and **c** SelectKBest Scores methods on New Delhi station dataset at 2pm

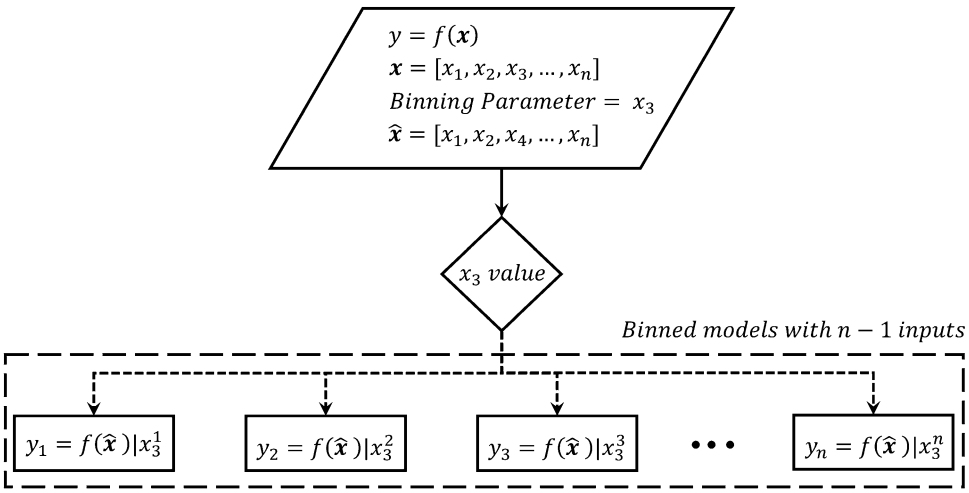

Fig. 4 Binned models developed with x_3 as the binning parameter

binning parameter and create bins as mentioned above. We compared the dominant features across the five stations at a given hour (see Fig. 5). It is evident that plots from same station can be grouped; however, features from any two different stations at the same hour are not the same. As such, the models cannot be grouped across stations. Distinct geographical locations of all five stations also contribute towards different parameters being dominant features. It is possible however, that stations in proximity of each other might have similar dominant features. In that case, this approach can then be used to group two or more stations based on a common binning parameter.

3.3 Evaluation Metrics

It is critical to evaluate the model performance and as such model evaluation is vital. There exist several evaluation metrics that help people understand the performance of a machine learning model. However, depending on the model type, namely, classification model or regression model, only some of them are useful. In this work we are building a regression model and as such corresponding evaluation metrics are being used. Unlike, the classification model where an item is classified correctly or incorrectly, regression model accuracy is closeness of the predicted value to the actual value (Pedregosa et al. 2019). Three primary evaluation metrics used for regression models are

1. R^2: R square also referred to as the coefficient of determination and it predicts variation in output based on all the inputs combined. This value typically ranges between 0 and 1 with higher value indicating a better model. This value can also be negative in case the model fits worse than a horizontal line. It is given by

$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y_i})^2}{\sum_i^n (y_i - \bar{y})^2} \tag{3}$$

2. RMSE: Root Mean Square Error helps us estimate the standard deviation of error. It provides us an absolute value as opposed to R^2 that provides us a relative value. It is calculated as some form of a normalized distance between the recorded and the predicted values. It is given by

Fig. 5 Comparison of selected features for all the five stations at hours 07, 12, 15 and 18. Hourly plots **a–d** represents data for Ahmedabad station, **e–h** for New Delhi station, **i–l** Kolkata station, **b–p** for Goa-Panjim station and **q–t** for Thiruvanthapuram

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2} \tag{4}$$

3. MAE: Mean Absolute Error is essentially the same as Mean Squared error, but it sums up the absolute value of error instead of the squared value.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y_i} - y_i| \tag{5}$$

In all the above equations n is the sample size of test data, y_i and $\hat{y_i}$ represent measured data and predicted data, respectively, and $\bar{y}$ represents the mean value.

4 Results

Four popularly used machine learning models, namely, linear regression, polynomial regression, decision trees and random forests, have been employed. Results for the New Delhi station are illustrated in Fig. 6. Random forest

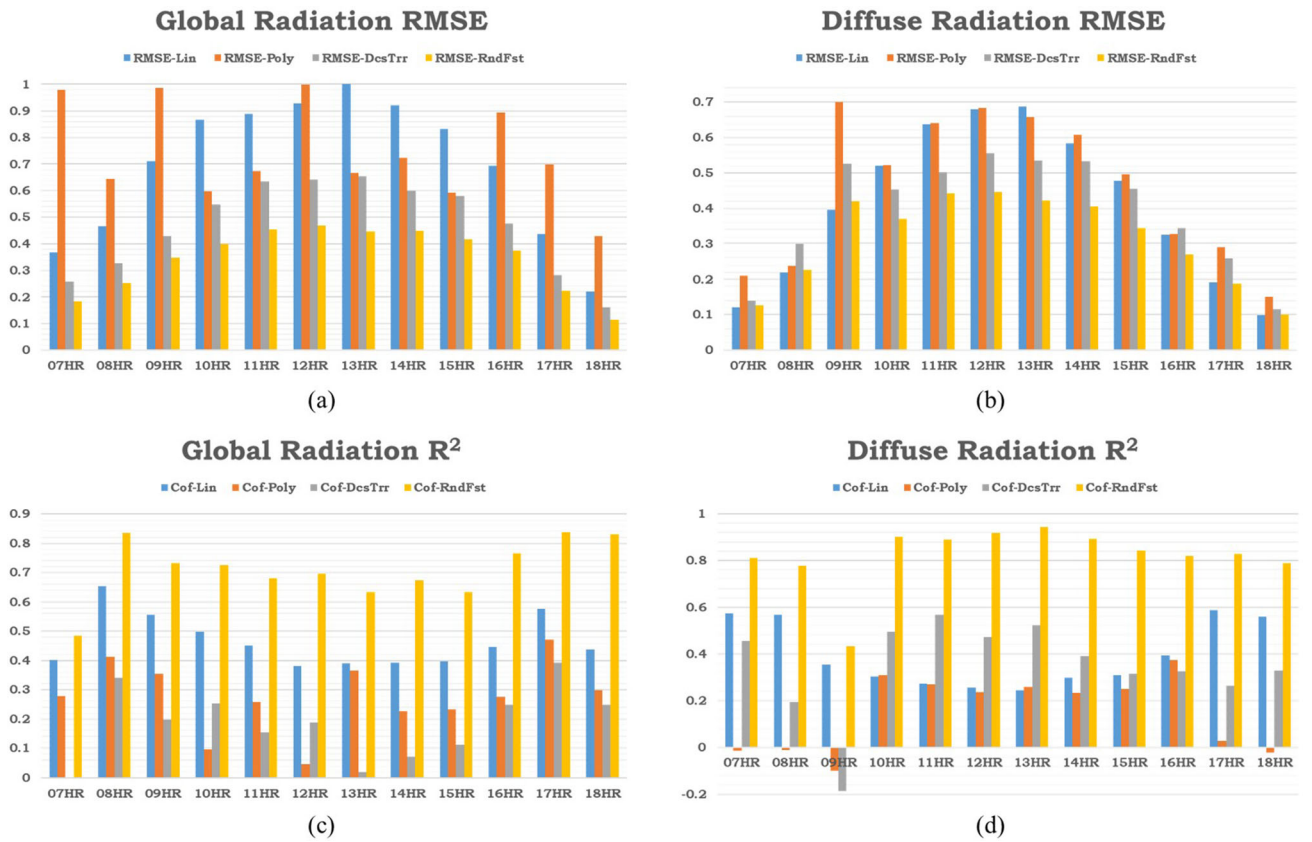

Fig. 6 Quantitative comparison of all the employed machine learning methods, namely, **a** Global radiation RMSE, **b** Diffuse radiation RMSE, **c** Global radiation R2, and **d** Diffuse radiation R2 for New Delhi Station

Table 4 Quantitative results of the proposed machine learning models on all the stations

Station Name	Models	R^2		RMSE		MAE	
		Global	Diffuse	Global	Diffuse	Global	Diffuse
New Delhi	Daily	0.728	0.817	0.313	0.339	0.246	0.215
	Hourly	0.711	0.607	0.114	0.153	0.198	0.169
	Binned	0.879	0.880	0.135	0.171	0.219	0.176
Ahmedabad	Daily	0.741	0.832	0.342	0.311	0.239	0.193
	Hourly	0.905	0.851	0.257	0.243	0.167	0.158
	Binned	0.920	0.868	0.272	0.260	0.178	0.166
Kolkata	Daily	0.651	0.752	0.449	0.332	0.333	0.247
	Hourly	0.842	0.810	0.342	0.266	0.234	0.185
	Binned	0.831	0.795	0.363	0.288	0.247	0.199
Goa-Panjim	Daily	0.706	0.689	0.429	0.354	0.298	0.221
	Hourly	0.857	0.724	0.370	0.306	0.216	0.189
	Binned	0.872	0.750	0.361	0.318	0.232	0.207
Thiruvanthapuram	Daily	0.729	0.812	0.354	0.288	0.337	0.199
	Hourly	0.846	0.823	0.266	0.243	0.225	0.189
	Binned	0.859	0.827	0.288	0.270	0.246	0.193

method clearly outperforms the other methods and similar observations can be made for all the five stations. We create three types of models, namely, Daily, Hourly, and Binned, for each of the dataset. As the name suggests, Daily model makes use of the data from a given day to create the model and this model considers the same

Table 5 Quantitative results of Binning based models created using various machine learning algorithms

Algorithms	Models	RMSE	
		Global	Diffuse
Linear rgression	Daily	0.696	0.770
	Binned	0.222	0.252
Polynomial regression	Daily	0.741	0.801
	Binned	0.427	0.441
Decision trees	Daily	0.465	0.479
	Binned	0.161	0.203
Random forest	Daily	0.313	0.339
	Binned	0.135	0.171

features throughout the day. Hourly model may have different features per hour as a separate model is created for each hour of the day. Binned model groups the hourly data based on the similar dominant features and creates a few models for a day. Hour parameter is treated as the binning parameter and is used to select the appropriate binned model. For every station, data is split randomly for training and testing with 80% data used for training and 20% for testing. All the models are thus tested against the data not used during training. K-Fold validation is performed with value of K set to 6. It repeats training and testing K times and mean of all the metrics are then recorded for evaluation purposes. A summary of results for all the stations is tabulated in Table 4.

4.1 Analysis

As can be seen from Table 4 performance of the binning approach based models is far superior than the daily models. This is owing to a variety of fluctuations in several parameter values throughout the day. Binning based models also perform almost in par with the hourly models, while requiring only 20–30% of the number of models. Creating hourly models for several stations is cumbersome and won't help us generalize them for nearby stations. Binning based models on the other hand help cluster the input space using a binning parameter. This can help us generalize these models for stations with similar binned input space. If the binning parameter is time, it can also help in situations where the timestamps (time resolution) for measurements at different stations are not same.

We have also carried out experiments to demonstrate the effectiveness of binning on other machine learning algorithms. Table 5 validates the fact that binning approach increases the accuracy for any machine learning algorithm and is in fact independent of the algorithm used. In certain cases, a simple algorithm like linear regression can also perform very similar to a superior algorithm like random forest. This is due to the fact that a complex non-linear problem can also be piece-wise linear. This can help us create a bank of simpler models in place of a single or a few complex models.

5 Conclusion

Solar radiation prediction is imperative for producing optimal solar power and plays a key role in reducing power station expenses. Machine learning models provide efficient ways of accurately forecasting the solar radiation. We have gathered the solar radiation data from five geographically distinct solar power stations. We have processed the data to remove outliers due to machine and human error as well as to synthesize missing data. We have performed EDA to identify the dominant features for building the machine learning models. We have proposed binning approach-based machine learning models to improve the performance of these machine learning models. We have evaluated these models quantitatively using commonly used evaluation metrics. Our binning-based models have shown improved performance. These models not only yield better results compared to the daily model, but also yield results in par with the hourly models while using much smaller number of models. Unique geographical locations and variations in climatic conditions throughout the day does not allow a single model to perform equally well throughout the day and creating several models per station is not an optimal solution. The proposed approach reduces number of models significantly (from 12 to a maximum of 3 per station). In addition, this being a data driven approach, similar models can be employed for a power station in the nearby vicinity. Moreover, simpler models such as linear regression models can also be used to model non-linearity as this approach makes use of piecewise linear input space. As a future work, we will record data from more stations and will try to form clusters of these stations using the binning approach. We will be able to demonstrate that binning can help us build a single model for multiple stations for a specific time of the day. This approach will help us create data dependent models rather than station dependent models. This will greatly reduce the number of models if we were to do this for several stations across India.

Acknowledgements Authors would like to acknowledge India Meteorological Department, Climate Research and Services, Pune, Ministry of Earth Sciences, Government of India for providing the necessary solar radiation and surface data. An appendix contains supplementary information that is not an essential part of the text itself but which may be helpful in providing a more comprehensive

understanding of the research problem or it is information that is too cumbersome to be included in the body of the paper.

References

Jiang Y (2008) Prediction of monthly mean daily diffuse solar radiation using artificial neural networks and comparison with other empirical models. Energy Policy 36(10):3833–3837

Kashyap Y, Bansal A, Sao AK (2015) Solar radiation forecasting with multiple parameters neural networks. Renew Sustain Energy Rev 49:825–835

Voyant C, Notton G, Kalogirou S, Nivet ML, Paoli C et al (2017) Machine learning methods for solar radiation forecasting: a review. Renewable Energy 105:569–582

Fan J, Wang X, Wu L, Zhang F, Bai H 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 Manage 156:618–625

Melzi FN, Touati T, Same A, Oukhellou L (2016) Hourly solar irradiance forecasting based on machine learning models. In: 15th IEEE international conference on machine learning and applications (ICMLA), pp 441-446

Belaid S, Mellit A (2016) Prediction of daily and mean monthly global solar radiation using support vector machine in an arid climate. Energy Convers Manage 118:105–118

Khosravi A, Nunes RO, Assad MEH, Machado L (2018) Comparison of artificial intelligence methods in estimation of daily global solar radiation. J Clean Prod 194:342–358

Bayrakçı HC, Demircan C, Keçebaş A (2018) The development of empirical models for estimating global solar radiation on horizontal surface: A case study. Renew Sustain Energy Rev 81:2771–2782

Jahani B, Mohammadi B (2019) A comparison between the application of empirical and ANN methods for estimation of daily global solar radiation in Iran. Theoret Appl Climatol 137(1–2):1257–1269

Feng Y, Gong D, Zhang Q, Jiang S, Zhao L, Cui N (2019) Evaluation of temperature-based machine learning and empirical models for predicting daily global solar radiation. Energy Convers Manage 198:111780

Kim SG, Jung JY, Sim MK (2019) A two-step approach to solar power generation prediction based on weather data using machine learning. Sustainability 11(5):1501

Mellit A, Massi Pavan A, Ogliari E, Leva S, Lughi V (2020) Advanced methods for photovoltaic output power forecasting: a review. Appl Sci 10(2):487

Liu Y, Zhou Y, Chen Y, Wang D, Wang Y et al (2020) Comparison of support vector machine and copula-based nonlinear quantile regression for estimating the daily diffuse solar radiation: a case study in China. Renew Energy 146:1101–1112

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 277:122353

Choudhary A, Pandey D, Bhardwaj S (2020) Artificial neural networks based solar radiation estimation using backpropagation algorithm. Int J Renew Energy Res 10(4):1566–1575

Álvarez-Alvarado JM, Ríos-Moreno JG, Obregón-Biosca SA, Ronquillo-Lomelí G, Ventura-Ramos E et al (2021) Hybrid techniques to predict solar radiation using support vector machine and search optimization algorithms: a review. Appl Sci 11(3):1044

de Freitas Viscondi G, Alves-Souza SN (2021) Solar irradiance prediction with machine learning algorithms: A Brazilian case study on photovoltaic electricity generation. Energies 14(18):5657

Brahma B, Wadhvani R (2020) Solar irradiance forecasting based on deep learning methodologies and multi-site data. Symmetry 12(11):1830

Inman RH, Pedro HT, Coimbra CF (2013) Solar forecasting methods for renewable energy integration. Prog Energy Combust Sci 39(6):535–576

Long H, Zhang Z, Su Y (2014) Analysis of daily solar power prediction with data-driven approaches. Appl Energy 126:29–37

Aler R, Martín R, Valls JM, Galván IM (2015) A study of machine learning techniques for daily solar energy forecasting using numerical weather models. In: Camacho D, Braubach L, Venticinque S, Badica C (eds) Intelligent distributed computing VIII. Springer, Cham, pp 269–278

Mitra I, Sharma S, Kaur M, Ramanan A, Wypior M, Heinemann D (2016) Evolution of solar forecasting in India: The introduction of REMCs. In: EuroSun conference proceedings, international solar energy society, vol 1, pp 1–10

Wang G, Su Y, Shu L (2016) One-day-ahead daily power forecasting of photovoltaic systems based on partial functional linear regression models. Renew Energy 96:469–478

Ogliari E, Dolara A, Manzolini G, Leva S (2017) Physical and hybrid methods comparison for the day ahead PV output power forecast. Renew Energy 113:11–21

Massucco S, Mosaico G, Saviozzi M, Silvestro F (2019) A hybrid technique for day-ahead PV generation forecasting using clear-sky models or ensemble of artificial neural networks according to a decision tree approach. Energies 12(7):1298

Li G, Xie S, Wang B, Xin J, Li Y, Du S (2020) Photovoltaic power forecasting with a hybrid deep learning approach. IEEE Access 8:175871–175880

Ahmad R, Kumar R (2021) "Very short-term photovoltaic (PV) power forecasting using deep learning (LSTMs). In: IEEE international conference on intelligent technologies (CONIT), pp 1-6

Alaraj M, Kumar A, Alsaidan I, Rizwan M, Jamil M (2021) Energy production forecasting from solar photovoltaic plants based on meteorological parameters for Qassim region, Saudi Arabia. IEEE Access 9:83241–83251

Son N, Jung M (2021) Analysis of meteorological factor multivariate models for medium-and long-term photovoltaic solar power forecasting using long short-term memory. Appl Sci 11(1):316

Rodríguez F, Galarza A, Vasquez JC, Guerrero JM (2022) Using deep learning and meteorological parameters to forecast the photovoltaic generators intra-hour output power interval for smart grid control. Energy 239:122116

Luo X, Zhang D, Zhu X (2021) Deep learning-based forecasting of photovoltaic power generation by incorporating domain knowledge. Energy 225:120240

Chahboun S, Maaroufi M (2021) Novel comparison of machine learning techniques for predicting photovoltaic output power. Int J Renew Energy Res 11(3):1205–1214

Zulkifly Z, Baharin KA, Gan CK (2021) Improved machine learning model selection technique for solar energy forecasting applications. Int J Renew Energy Res 11(1):308–319

Kim JG, Kim DH, Yoo WS, Lee JY, Kim YB (2017) Daily prediction of solar power generation based on weather forecast information in Korea. IET Renew Power Gener 11(10):1268–1273

Mendhurwar KA, Devabhaktuni VK, Raut R (2008) Binning algorithm for accurate computer aided device modeling. In: IEEE international symposium on circuits and systems, pp 2773-2776

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al (2019) Scikit-learn: machine learning in Python. Theoret Appl Climatol J Mach Learn. Res 137(1):1257–1269