ISSN 2004-2965 Energy Proceedings, Vol. 18, 2021

International Conference on Applied Energy 2021
Nov. 29 - Dec. 2, 2021, in Bangkok, Thailand
Paper ID:69

Machine learning techniques for hourly global horizontal irradiance prediction: A case study for smart cities of India

Pratima Kumari[1*], Durga Toshniwal[2]

1* Indian Institute of Technology Roorkee, India (Corresponding Author)

2 Indian Institute of Technology Roorkee, India

ABSTRACT

The availability of solar irradiance is uncertain and time-dependent, which is influenced by several climatic factors. Therefore accurate solar irradiance prediction is required for planning, designing, and site selection to establish new solar power plants. This study utilizes eight machine learning techniques, including multivariate linear regression, ridge, lasso, elastic net, multilayer perceptron, k-nearest neighbors, decision tree, and random forest to develop hourly global horizontal irradiance prediction models. A feature selection procedure to select the most influential input features from different meteorological variables is also discussed in this study. To examine the accuracy of the developed models, this study employs the data of 21 locations of different climatic and geographic regions. The considered cities are categorized into three groups using k-means clustering in order to find out the favorable locations for solar power generation. The computational results suggest that the random forest and k-nearest neighbor are the most efficient prediction model, which outperformed other machine learning models with an average forecast skill of 37% and 35% over the smart persistence model. Overall, this study may be exercised for the selection of an efficient GHI prediction model and location for the installation, designing and planning of new solar power plants.

Keywords: Renewable energy, global horizontal irradiance, machine learning, forecasting, random forest, clustering

1. INTRODUCTION

The integration of renewable energy sources, specifically solar and wind, to the global energy supply is the most challenging part in energy supply system. The intermittent and uncertain nature of solar energy brings along several challenges, including voltage fluctuation, power quality, higher cost, power output instability, etc. [1]. Therefore, solar irradiance forecasting plays a vital role in the effective execution of power grid systems. It is essential for the management of storage reserves of backup energy sources, power distribution, scheduling of the plants, power trading and finally, optimizing the overall power production cost [2]. In recent years, accurate solar irradiance forecasting has drawn enormous attention from researchers [3-6].

Over the past two decades, several solar irradiance forecasting methods have been proposed, which are divided into three types: physical, statistical and machine learning methods. With the advancements in technology and development of artificial intelligence techniques, machine learning-based solar irradiance forecasting models provide more promising results as compared to physical and statistical models. The capability to extract complex non-linear features from the high dimensional solar irradiance data and meteorological parameters makes machine learning models excellent f or solar irradiance forecasting [7]. Artificial neural netw ork decision tree, support vector machine (SVM), random forest, etc. are a few extensively applied machine learning-based solar irradiance forecasting models [8,9].

2. RESEARCH CONTRIBUTION

This work provides the comparison of several most widely applied machine learning techniques for hourly solar irradiance prediction. In particular, eight popular techniques, including multivariate linear regression (MLR), ridge regression, lasso regression, elastic net

Selection and peer-review under responsibility of the scientific committee of the 13th Int. Conf. on Applied Energy (ICAE2021).

regression, multilayer perceptron (MLP), k-nearest neighbors (kNN), artificial neural network, decision tree regression and random forest regression are compared in the present study. An explicit feature selection procedure is also employed in this study to determine the most influencing input features for efficient solar irradiance prediction. A significant restriction of many previous studies is that they only demonstrate the model validation for a particular or limited locations. To overcome this issue, the effectiveness of considered models is evaluated on the data of 21 smart cities which belong to different climatic zones of India. It facilitates a richer comparison of prediction models across different locations and climate conditions. Further, based on the accuracy of prediction models for different locations, these locations are classified into different categories using k-means clustering. This classification may give an insight about the favorable sites for the establishment of new power plants.

3. METHODOLOGY

3.1 Data collection

In this work, 5 years (2010-2014) dataset from National Solar Radiation Database (NSRDB) with a temporal resolution of 1 hour and spatial resolution of 10 x 10 km is considered to affirm the efficiency of considered machine learning models [10]. The data is recorded using NSRDB SUNY model, developed through semi-empirical satellite model. The dataset of 21 cities, namely Agartala, Ahemdabad, Bhopal, Bhubneshwar, Chennai, Coimbatore, Davangere, Diu, Guwahati, Imphal, Indore, Jaipur, Kakinada, Kochi, Ludhiana, New Delhi, Pune, Solapur, Surat, Udaipur and Vishakhapatnam located in different climatic zones of India, selected under "Smart cities mission" is used for training and testing of the models. Table 1 shows the geographical coordinates and Fig. 1 shows the location of the selected cities on the Indian map.

Table 1 Smart cities of India selected used to conduct the study

Agartala	Ahmedabad	Bhopal	Kochi
Chennai	Coimbatore	Davangere	Diu
Guwahati	Bhubaneshwar	Indore	Jaipur
Kakinada	Vishakhapatnam	Ludhiana	Solapur
Pune	New Delhi	Surat	Udaipur
Imphal			

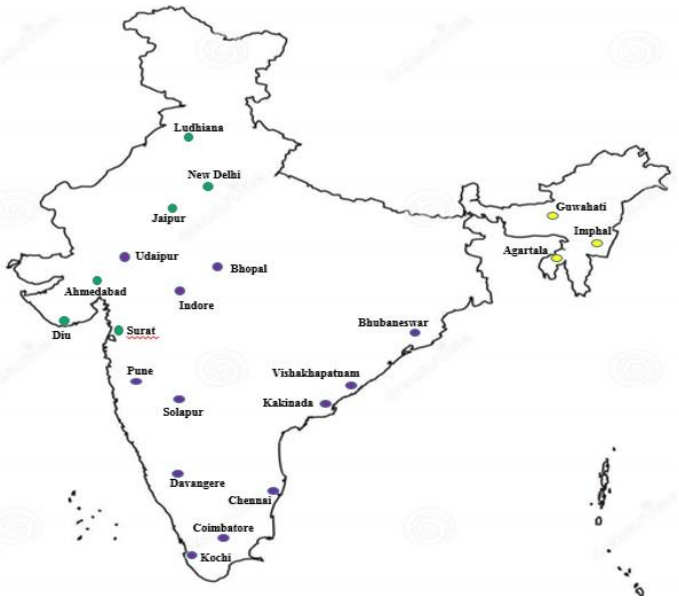

Figure 1. Physical map of India showing selected smart cities.

3.2 Feature selection

The list of variables, which consists of several GHI historical states and meteorological variables, is shown in Table 2.

Table 2 List of variables.

Input variables	Explanation
I_{t-1}	GHI measured at 1 hour ago on same day
I_{t-2}	GHI measured at 2 hours ago on same day
I_{t-3}	GHI measured at 3 hours ago on same day
Q_t	Temperature at time t
H_t	Relative humidity at time t
V_{dirt}	Wind direction at time t
V_t	Wind speed at time t
Q_{dewt}	Dew point at time t
R_t	Precipitable water at time t
P_t	Pressure at time t

The autocorrelation coefficient between the present GHI and historical GHI values is calculated to analyze the impact of historical GHI states on current GHI state as described below [11].

$$r_p = \frac{\sum_{j=1}^{n-p}(y_{(t)} - \bar{y})(y_{(t-p)} - \bar{y})}{\sum_{j=1}^{n}(y_{(t)} - \bar{y})^2}$$

where t denotes the sampling interval, r_p denotes the autocorrelation coefficient at p lag, $\bar{y}$ is the mean of time-series $y_{(t)}$ and n represents the total number of samples. Table 3 shows the value of autocorrelation coefficient between current and historical GHI, including t–T, t–2T, t–3T, t–24T, t–48T, and t–72T. The value in cell (t-2T, t-48T) of Table 3 is 0.7569, represents the

autocorrelation coefficient between the latest GHI value and GHI value obtained 2 days and 2 h (i.e.50 h) ago. GHI data at the same hour of previous 3 days (i.e., t-24T, t-48T, t-72T) is considered in the further parameter selection process.

Table 3 Autocorrelation coefficients of solar irradiance.

Hours	I_t	I_{t-24T}	I_{t-48T}	I_{t-72T}
It	1	0.9626	0.9575	0.9561
I_{t-T}	0.9297	0.9094	0.9039	0.9028
I_{t-2T}	0.7728	0.7616	0.7569	0.7554
I_{t-3T}	0.5558	0.5497	0.5456	0.5447
I_{t-4T}	0.3107	0.3075	0.3041	0.3037

RMSE and R-square value of hourly GHI prediction is calculated using historical GHI state and meteorologic -al parameters as an input to random forest model as shown in Table 4. Finally, GHI of past three days (i.e., I_t-24T ,I_t-48T and I_t-72T), temperature (Q_t), relative humidity (H_t) and pressure (P_t) are considered as the final input features for training the GHI prediction models.

Table 4 The list of input features with different combination of historical states of GHI and meteorological variables.

Parameters	R-square	RMSE
I_t-24T ,I_t-48T, I_t-72T, Q_t	0.9555	57.59
I_t-24T ,I_t-48T, I_t-72T, Q_t, P_t	0.9513	59.23
I_t-24T ,I_t-48T, I_t-72T, Q_t, H_t	0.9673	56.07
I_t-24T ,I_t-48T, I_t-72T, Q_t, R_t	0.9568	58.51
I_t-24T ,I_t-48T, I_t-72T, Q_t, V_t	0.9671	57.20
I_t-24T ,I_t-48T, I_t-72T, Q_t, H_t, P_t	0.9828	54.62
I_t-24T ,I_t-48T, I_t-72T, Q_t, H_t, V_t	0.9673	56.09
I_t-24T ,I_t-48T, I_t-72T, Q_t, H_t, R_t	0.9677	55.70
I_t-24T ,I_t-48T, I_t-72T, Q_t, H_t, P_t, R_t	0.9676	55.69

3.3 GHI Prediction models

3.3.1 Multivariate linear regression (MLR)

Multivariate Linear Regression Multivariate linear regression (MLR) is the simplest data-driven model that works on the concept of deriving a linear relationship between dependent and independent [12].

3.3.2 Ridge regression

Ridge regression deals with the problem of multicollinearity and over-fitting in data by introducing a regularization term [13].

3.3.3 Lasso regression

Least absolute shrinkage and selection operator is a regression technique that differs from standard linear regression as it has an embedded quality of feature selection [14].

3.3.4 Elastic net regression

Elastic net utilizes the idea of both lasso and ridge regression by linearly combining the penalty terms i.e., L1 and L2 regularization [15].

3.3.5 Multilayer Perceptron (MLP)

MLP is comprised of fully connected multiple layers stacked between input and output layers [16]. The first layer depicts the input layer where number of nodes is equal to the number of input parameters. The second layer is the hidden layer and the third layer is the output layer, where the number of nodes is decided on the basis of the intended output.

3.3.6 K-Nearest Neighbor (kNN)

kNN is generally used as a clustering algorithm. However, it may also be utilized to solve regression problems [17]. Distance of each new test point to all training data point is calculated using several distance metrics such as euclidean, manhattans, etc.

3.3.7 Random forest

Random forest, proposed by Breiman [18], is a flexible machine learning algorithm that is an extension of decision tree and does not demand much time in the tuning of hyper-parameters. It is an ensemble of decision trees and has comparatively less chances of over-fitting. Random forest selects the subsets of features and constructs smaller decision trees, results in good performance due to its default hyperparameters.

3.3.8 Decision tree

Decision tree is a widely used rule-based machine learning technique that can be used for regression and classification [19]. It consists of nodes and leaves, arranged in a hierarchical tree-like structure. Each test data point starts from the root node and follows the path which is tested to be true for that data point. The

procedure is repeated until a leaf node arrives, then the value of the leaf node is allocated as the predicted value to the test point.

4. RESULTS AND DISCUSSION

This section presents and compares the results of considered models developed for hourly GHI prediction across different locations in India. The performance of the developed models is evaluated for multiple locations, which belong to different climatic regions of India. Fig. 2 and Fig. 3 show results in terms of MAE and R-square across different locations. Considered machine learning models can be divided into two groups based on their performance. The first group is comprised of MLR, Lasso, Ridge and Elastic net, which shows considerably low performance. While the other group is consists of kNN, MLP, decision tree and random forest, which shows higher prediction performance. The maximum error is observed for MLR model with an average MAE value of 50 W/m^2 followed by lasso, ridge and elastic net model. The lasso and ridge model has shown comparable performance having average MAE values of 47 W/m^2 and 46 W/m^2, respectively. The improvement in the performance of lasso and ridge regression model is observed due to the incorporation of L1 and L2 regularization term in these models. The regularization term in the model overcomes the problem of overfitting and hence enhances the accuracy of prediction models. To further enhance the performance over lasso and ridge regression models, the elastic net model is used. The incorporation of both L1 and L2 regularization term in elastic net has slightly enhanced the prediction accuracy. The minimum error is observed for random forest model followed by kNN, MLP and decision tree. Random forest has delivered the most promising results among all discussed algorithms with an average MAE of 32 W/m^2. Furthermore, decision tree and MLP have shown good accuracy in prediction with an average MAE values 36 W/m^2 and 39 W/m^2, respectively. Whereas kNN outperformed both MLP and decision tree with an average MAE of 34 W/m^2, attributed to the principle of localization. The statistical analysis shows that kNN and random forest have performed well and provide comparable results, which indicates their potential for GHI forecasting. The similar trend of performance is observed in terms of R-square, as shown in Fig. 3. Overall, the considered models can be arranged in descending order of their performance as follows: random forest > kNN >MLP > decision tree > elastic net > ridge > lasso > MLR.

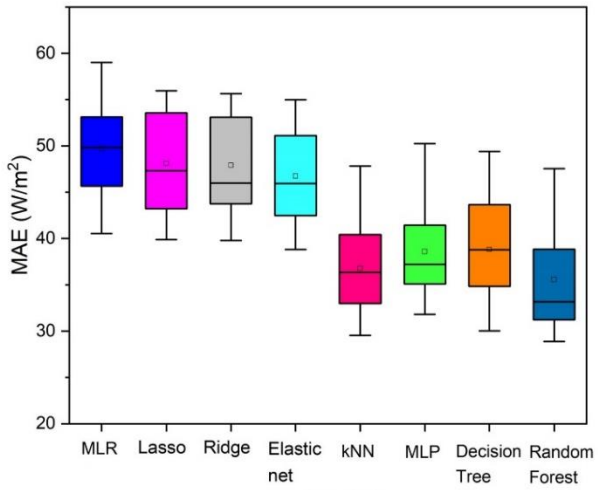

Figure 2 Box plot representation of prediction accuracy of developed models in terms of MAE

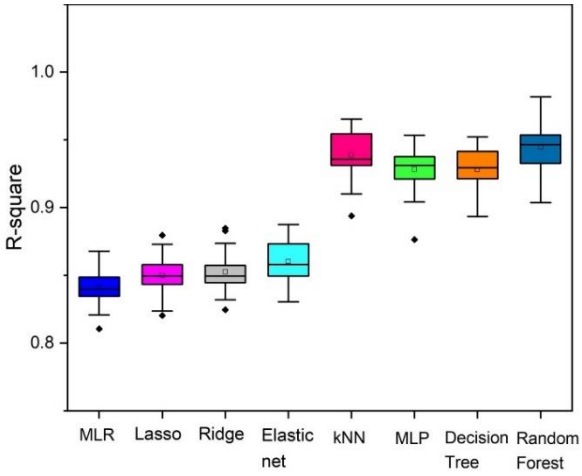

Figure 3 Box plot representation of prediction accuracy of developed models in terms of R-square.

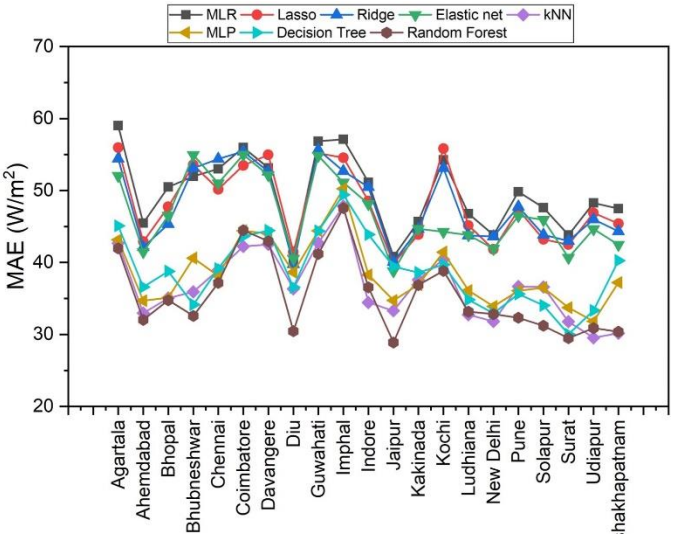

Figure 2 Comparison of prediction models across all selected locations in terms of MAE

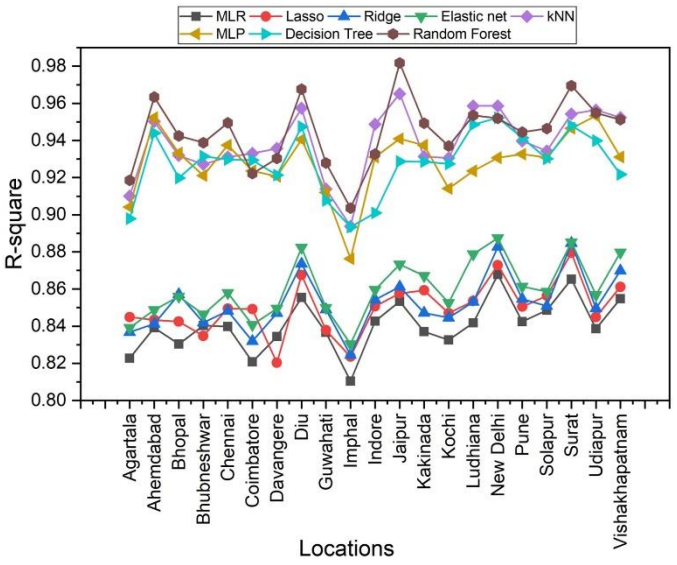

Figure 3 Comparison of prediction models across all selected locations in terms of R-square.

Further, the performance of developed machine learning models is observed at each considered locations as shown in Fig. 4 and Fig. 5. It can be observed from Fig. 4 that prediction models show very high error at few locations such as Agartala, Davangere, Guwahati, Imphal and Kochi. However, the prediction models show highest performance at Ahmedabad, Diu, Jaipur and Surat. Random forest has shown the highest accuracy at most of the locations, while MLR has shown the lowest performance at most of the locations.

4.1 Clustering of cities

The forecasting machine learning models are developed for considered 21 locations, and evaluated using three different performance metrics, i.e., MAE and R-Square [20]. Further, these cities are clustered based on the performance metrics for each algorithm. The clustering has been performed using an unsupervised clustering algorithm, i.e. k-means clustering [19]. To determine the value of k in the k-means algorithm, elbow method is used. The optimized value of k has come out to be 3. After applying k-means clustering, the considered locations are categorized into three clusters as shown in Fig. 6. Cluster I is shown in violet color, which consists of 7 cities, including Ahmedabad, Diu, Jaipur, Kakinada, Ludhiana, New Delhi, and Surat. These locations have shown high GHI prediction accuracy, which might be attributed to the highly stable weather conditions throughout the year. Cluster II is shown in green color, which is comprised of 8 cities, Bhopal, Bhubaneshwar, Chennai, Indore, Pune, Solapur, Udaipur and

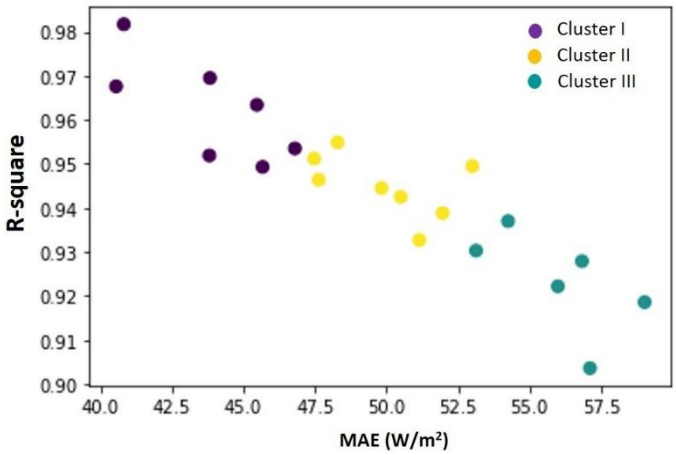

Figure 4 Scatter plot of smart cities clustered using k-means.

Vishakhapatnam. Cluster III shown in violet color is comprised of 6 cities, Agartala, Coimbatore, Davangere, Guwahati, Imphal and Kochi. For these locations, proposed models have shown poor prediction accuracy. In these locations, meteorological variables are not contributing significantly due to fluctuating weather conditions and unpredictable high rainfalls throughout the year.

5. CONCLUSIONS

Global horizontal irradiance (GHI) is the most primary element for the designing and installation of operating solar power plants. However, the availability of GHI is influenced by many climatic factors. Therefore, the accurate prediction of GHI is necessary for the successful planning of solar power plants. In this work, several machine learning models, including multivariate linear regression, ridge, lasso, elastic net, multilayer perceptron, k-nearest neighbors, decision tree, and random forest are developed for hourly GHI prediction. A feature selection procedure is also employed to determine the most influential input features. The performance of the developed models is compared to each other and to a benchmark model, namely smart persistence model. The performance of the developed models is evaluated using meteorological data of 21 different locations. Finally, k-means clustering is applied to categorize the considered locations into three different categories. Based on the obtained simulation results achieved in the present study, the following conclusions can be drawn:

- The historical GHI data of the previous three days at the time of prediction (i.e., t-24T, t-48T, t-72T), temperature, relative humidity and pressure are used as input features to train the GHI prediction models.

- The random forest model has shown the maximum prediction accuracy and outperformed the other machine learning models at 14 out of 21 locations.
- kNN came out to be the second most accurate GHI prediction model. kNN has outperformed other machine learning models at 7 out of 21 locations.
- The analysis of error metrics and clearness index demonstrates that the locations in cluster I are best suited for establishment of solar power plants while those are in cluster III are not much suitable for solar power generation.

To sum up and based on the outcomes of present study, government and energy industries may make effective policies to advance their economic expediency, profitable market penetration and finally lead the way for cracking the biggest source of energy. Alternative non-conventional energy resource projects such as wind energy, tidal energy should be preferred at sites where harnessing solar energy is not profitable.

ACKNOWLEDGEMENT

Authors would like to thank the Dean of Resources & Alumni Affairs (DORA), IIT Roorkee for providing the financial support.

REFERENCES

[1] Diagne, M., David, M., Lauret, P., Boland, J., & Schmutz, N. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. Renewable and Sustainable Energy Reviews 2013; 27: 65-76.

[2] Kumari, P., & Wadhvani, R. Wind power prediction using klms algorithm. IEEE International Conference on Inventive Research in Computing Applications (ICIRCA) 2018; 154-161.

[3] Chen C, Duan S, Cai T, Liu B. Online 24-h solar power forecasting based on weather type classification using artificial neural network. Sol Energy 2011;85:2856–70.

[4] Kumari, P., & Toshniwal, D. (2019). Hourly solar irradiance prediction from satellite data using lstm.ICAE 2019.

[5] Kumari, P., & Toshniwal, D. Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance. Journal of Cleaner Production, 2021;279:123285.

[6] Qing X, Niu Y. Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. Energy 2018;148:461–8.

[7] Kumari, P., & Toshniwal, D. Analysis of ANN-based daily global horizontal irradiance prediction models with different meteorological parameters: a case study of mountainous region of India. International Journal of Green Energy 2021;1-20.

[8] Kumari, P., & Toshniwal, D. Long short term memory–convolutional neural network based deep hybrid approach for solar irradiance forecasting. Applied Energy 2021;295:117061.

[9] Kumari, P., & Toshniwal, D. Deep learning models for solar irradiance forecasting: A comprehensive review. Journal of Cleaner Production 2021;128566.

[10] Sengupta M, Xie Y, Lopez A, Habte A, Maclaurin G, Shelby J. The National Solar Radiation Data Base (NSRDB). Renew Sustain Energy Rev 2018.

[11] Box GE and Jenkins GM. Time series analysis: forecasting and control, revised ed. Holden-Day 1976.

[12] Şahin, M., Kaya, Y., & Uyar, M. Comparison of ANN and MLR models for estimating solar radiation in Turkey using NOAA/AVHRR data. Advances in Space Research 2013;51:891-904.

[13] Hoerl AE and Kennard RW Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 1970;12:55–67.

[14] Tibshirani R Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2016;73:273–282.

[15] Salcedo-Sanz, S., Casanova-Mateo, C., Muñoz-Marí, J., & Camps-Valls, G. Prediction of daily global solar irradiation using temporal Gaussian processes. IEEE Geoscience and Remote Sensing Letters 2014;11:1936-1940.

[16] Mashaly AF and Alazba A Mlp and mlr models for instantaneous thermal efficiency prediction of solar still under hyper-arid environment. Computers and Electronics in Agriculture 2016;122:146–155.

[17] Pedro HT and Coimbra CF Nearest-neighbor methodology for prediction of intra-hour global horizontal and direct normal irradiances. Renewable energy 2015;80:770–782.

[18] Breiman L Random forests. Machine learning 2018;45:5–32.

[19] Kumari, P., & Toshniwal, D. (2020, November). Real-time estimation of COVID-19 cases using machine learning and mathematical models-The case of India. IEEE 15th International Conference on Industrial and Information Systems (ICIIS) 2020; 369-374.

[20] Kumari, P., & Toshniwal, D. (2021). Advanced machine learning techniques for short-term solar irradiance forecasting. First International Conference on AI-ML Systems

