Performance Evaluation of Various Solar Forecasting Models for Structural & Endogenous Datasets

Pardeep Singla[1,*], Manoj Duhan[1] and Sumit Saroha[2]

[1]*Deenbandhu Chhotu Ram University of Science & Technology, Sonepat, India*
[2]*Guru Jambheshwar University of Science and Technology, Hisar, India*
E-mail: pradeepsingla7@gmail.com
[*]*Corresponding Author*

Received 11 May 2022; Accepted 16 September 2022;
Publication 03 January 2023

Abstract

The forecasting of solar irradiation with high precision is critical for fulfilling electricity demand. The dataset used to train the learning-based models has a direct impact on the model's prediction accuracy. This work evaluates the impact of two types of datasets: structural and endogenous datasets over the prediction accuracy of different solar forecasting models (five variants of artificial neural network (ANN) based models, Support vector machine (SVM), Linear Regression, Bagged and Boosted Regression tree). The issue of variability estimation is also explored in the paper in order to choose the best model for a given dataset. The performance of the models is assessed using two essential error metrics: mean absolute percentage error (MAPE) and root mean square error (RMSE). The results shows that the MAPE and RMSE for structural data vary from 1.99% to 29.73% and 23.39 W/m^2 to 165.21 W/m^2, respectively, whereas these errors for endogenous dataset ranges from 1.98%

Distributed Generation & Alternative Energy Journal, Vol. 38_2, 467–490.
doi: 10.13052/dgaej2156-3306.3825

to 31.19% and 23.64 W/m^2 to 152.56 W/m^2. Moreover, these findings, together with the data variability findings, suggest that SVM is the best model for all forms of data variability, whereas CFNN may be employed for greater variability.

Keywords: Solar forecasting, learning rate, support vector machine, artificial neural network, moving window.

1 Introduction

In the field of renewable energy, the solar energy is the prominent source of energy used to generate the electricity. Various nations have already established rules for grid-connected solar and rooftop solar plants [1]. However, due to the intermittent behaviour of solar, the ability to anticipate accurate solar irradiance in advance is essential for the appropriate operation, scheduling, and balancing of grid-connected power systems. Furthermore, the quality of solar forecast is inextricably linked to meteorological and climatic circumstances, with shadow and panel location being secondary factors that influence accuracy [2]. Although, in many underdeveloped nations, the behavior of sun intensity at neighboring locations is used to observe the target sites. This indirect forecast for bad locations also necessitates accurate sun irradiation calculation. The literature is full of different models used for prediction of solar irradiation. Based on the literature, the prediction approach may be classified based on the type of input variables, temporal horizon, and the dataset type [3]. The input parameters are chosen from a list of strongly correlated meteorological parameters with the target parameter, and the temporal horizon is set in the future. The prior studies included a variety of time periods, including extremely short term, short term, midrange, and long-term forecasting. However, in case of real-time solar irradiance forecasting, very short-term forecasting is employed using the structural and endogenous dataset [4].

However, machine learning-based approaches especially, ANN is utilized by a huge number of forecasters in recent years to estimate the accurate solar irradiance. ANN methodology is a statistical method that uses a series of measurements gathered over time intervals for various meteorological and geographic characteristics [5]. To perform the experimental investigation on different datasets, this paper explores the performance of popular

learning-based models. The Feed-Forward Neural Network (M1), Non-linear Autoregressive Neural Network with exogenous inputs (M2), Elman Neural Network (M3), Cascade Forward Neural Network (M4), Generalized Regression Neural Network (M5), Support Vector Machine (M6), Linear Regression (M7), Boosted Regression Tree (M8) and Bagged regression tree (M9) methods are used to predict 24 hours ahead monthly solar irradiance for varied datasets. The prime objective of focusing on these models is that they are the most fundamental and widely utilized learning-based models in many research.

While taking into account prior works, Daniel et al. (2017) investigated the use of ANN-based approaches on solar data in order to increase prediction accuracy. To improve data accuracy, this work used the masking, tuning, and classification strategy as a feature selection method [6]. Similarly, Meenal et al. (2018) explored the SVM-based model for predicting the global solar radiation (GSR). The study utilized meteorological factors as input to forecast the GSR and assessed it using several error measures [7]. Zendehboudi et al. (2018) also looked at the SVM model for predicting solar and wind energy. Furthermore, this model was compared to other approaches, with SVM achieving high accuracy for the majority of the targeted regions [8]. Sharika et al. (2018) conducted a long-term horizon solar forecasting comparison research. For the chosen site, the Autoregressive integrated moving average (ARIMA), random forest (RF), SVM, and linear regression (LR) models were evaluated [9]. Exogenous data was utilized by Ozoegwu et al. (2019) to forecast daily and monthly GSR [10]. Furthermore, an ANN based forecasting model was implemented by Hameed et al. (2019) for solar GHI prediction of considered PV plants, whereas the PV power output was predicted by Junaat et al. (2018) in their study [11, 12]. These models were evaluated using RMSE and MAPE, which yielded different results depending on the location. Dahidi et al. (2019) discussed eleven ANN-based solar photovoltaic power production forecasting methods. The article used multiple error measures to evaluate a day ahead forecast [13]. Moreover, S. Rahman et al. (2021) predicted the solar radiation using the ANN based models. The study utilized the meteorological dataset consist of humidity, atmospheric pressure, temperature of air, time, wind speed, wind direction and solar radiation is used to train the model. To evaluate the performance of the developed models mean absolute error (MAE) and mean bias error (MBE) are used in the study [14]. In continuation, the boosted regression tree, extreme gradient boost and

regression model was used by L. Huang et al. (2021) to predict the solar radiation for extreme climates. The daily and monthly forecasting was performed in this study by the developed models for a highly variable climate [15]. A hybrid model with the least square SVM and artificial bee colony method was introduced by M. Guermoui et al. (2021) to estimate solar irradiance for the short-term time horizon. The model was trained and tested for the hourly forecast using the historical time series data with several combinations as an input [16].

However, in the literature, LR is also employed to create forecasting models. To characterize the input and output connection between the data, Mouatasim et al. (2018) used regression analysis [17]. This study used linear, multiple linear, and nonlinear regression approaches with meteorological factors as input to examine PV systems. According to aforementioned studies, the intensity of solar irradiation is directly impacted by several climatic factors such as relative humidity, cloud cover, temperature, sky index, dew point, wind direction, solar zenith angle, and wind speed. Furthermore, the choice of data set has an impact on predicting accuracy for various time horizons. In addition, hybrid models with different learning-based predictors are also developed utilizing wavelet transform (WT) [4], full wavelet packet transform [18], empirical mode decomposition, ensemble empirical mode decomposition, complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) [19].

Meanwhile, to the best of our knowledge, no experiment has been performed in the literature, that discusses the implications of several datasets on model's performance. As a result, the impact of the structural and endogenous dataset on the forecasting accuracy of forecasting model is presented in this paper. The clear sky index (CSI) measure is used to forecast the global horizontal irradiance (GHI), and the assessment of variability in the dataset is used to choose the best model out of the nine available.

The paper is organized is as follows: Section 1 provides the introduction to solar forecasting and the brief of literature. Section 2 is intended for discussion on the subject area and motivation for the effort. The steps involved in the forecasting process and the history of the models used in the study are covered in Section 3. Further, the error metrics used to evaluate the models and the experimental findings for various datasets are presented in Sections 4 and 5. The data variability computation for model selection is discussed in Section 6. The paper's conclusion is presented in Section 7.

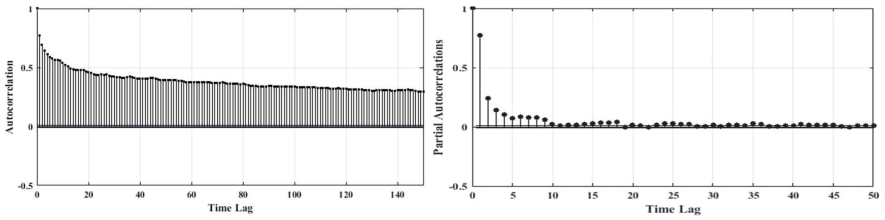
2 Study Area & Motivation

Any site or place can be utilized to test a model; however, this study incorporates meteorological and geographical historical data from Indian sites. The chosen places provide a wide range of weather conditions throughout the year, and multiple weather features may be seen at a single site. Another reason for choosing Indian site is its fastest-growing economy's commitment to the renewable energy sector [20]. The datasets utilized in this study are gathered for the city of Ahmedabad in Gujarat, India (23°.05′/72°.35′). Ahmedabad has a severe climate with four distinct seasons: summer, monsoon, autumn, and winter. Every season has its own unique and varying qualities [21]. The dataset is gathered for year 200 t0 2014 from the National Solar Radiation Database (NSRDB), where the data from year 2012 to 2014 is being utilized in the experiments. The structural dataset comprises of Temperature (T), Relative humidity (RH), precipitation (prec.), Dew Point (DP), pressure (P), wind speed (WS), solar zenith angle (SZA), and wind direction (WD) as a meteorological variable, whereas endogenous data is represented by an appropriate time lag time-series data shown in Figure 1.

3 Adopted Methodology

The following stages are used in this study to investigate the impact of the dataset on model accuracy, as well as model selection based on the assessment of variability and CSI. Figure 2 depicts the process flow of the forecasting approach employed in the study.

1. *Data Quality*: Data from solar data suppliers is always in raw format. Due to instrument problems, the data may contain incorrect entries and missing values. The removal of such fraudulent observations is part of the data quality assessment and is done at the beginning [2, 22].
2. *Pre-Processing*: Night hours, data immediately after sunrise, and data just before sunset are all removed throughout this procedure. The data just after and before the sunrise and sunset, is deemed erroneous readings due to the sensors' cosine error. In other words, data having a solar zenith angle (SZA) greater than 80° are excluded from this investigation [19]. After pre-processing, the dataset of nine hours of a day for the specified location remained. As a result, the one-year data has 3285 data points, with monthly data points serving as validation and testing.
3. *Clear sky Index Transformation*: The data received from any agency are often random and non-stationary in nature. The CSI for target vector i.e.,



(a) (b)

Figure 1 (a) and (b): ACF and PACF of the historical solar GHI.

GHI must be determine to change the target data into stationary form using the following mathematical calculations:

$$K_t = \frac{GHI}{GHI_{cs}} \tag{1}$$

Where GHI_{cs} = extraterrestrial solar GHI.

4. *Feature Selection*: After completing steps 1–3, the next step is to select appropriate meteorological data in the case of structural dataset. The study used Pearson's coefficient to determine the relevance of the input variable to the target variable. In comparison to the target GHI, a Pearson coefficient near to 1 is considered high and 0 is poor. To forecast GHI, structural variables RH, DP, T, prec., P, SZA, WS, and WD are found suitable for the selected dataset. The endogenous dataset, which is selected using autocorrelation function (ACF) and partial autocorrelation (PACF), uses the necessary time delays. Fig.1 shows the ACF and PACF to demonstrate the consecutive lags where the lags correlation rapidly decreases with time.
   As the all the time lags are having better ACF but lags 1 to 10 have better PACF, presenting confidence interval of 95%. Therefore, the time lag from 1 to 10 are used for preparing the endogenous dataset.
5. *Moving Window Mechanism*: The moving window method is used to automatically determine the training and testing window. The function of 't' denotes the value of a certain variable at time 't', whereas 't + 1' denotes one step ahead of future value. In structural data set, the input function 'f(*)' can be represented as:
   For 1-step ahead

$$I_{t+1} = f_1[((DP_1, DP_2, \ldots, DP_t), (RH_1, RH_2, \ldots, RH_t), (T_1, T_2, \ldots, T_t), \ldots)] \tag{2}$$

For n-step ahead

$$I_{t+n} = f_n \left[\begin{pmatrix} (DP_{1+n-1}, DP_{2+n-1}, \ldots, DP_{t+n-1}), \\ (RH_{1+n-1}, RH_{2+n-1}, \ldots, RH_{t+n-1}), \ldots, \\ (T_{1+n-1}, T_{2+n-1}, \ldots, T_{t+n-1}), \ldots \end{pmatrix} \right] \quad (3)$$

In same way, if *'GHI'* is a time series, then the same function for one step ahead forecast can be represented as for endogenous datasets:

$$I_{t+1} = f_1(I_1, I_2, \ldots, I_t) \quad (4)$$

Likewise, for the case of n- step ahead

$$I_{t+1} = f_n(I_{1+n-1}, I_{2+n-1}, \ldots, I_{t+n-1}) \quad (5)$$

6. *Training & Testing*: The data is divided into three primary sequences: training data, validation data, and testing data in this stage. However, choosing training data for a model is one of the most important project since it directly affects the model's predicting ability. A detailed experiment is undertaken in this research to determine the appropriate training window for the proposed models. Table 1(a) & 1(b) show the outcomes of the various experiments for various training window sizes. To construct the forecast, the authors in the literature, sometimes choose the training and testing data ranges for a specific day, month, or year. However, the vast majority of studies employed this data in a ratio of 70%:30% for training and testing [18]. This data division makes the model reliant on a certain event that occurs only in that data range. The k-fold cross-validation approach, which splits the dataset into samples equal to the value assigned to k (this study used k = 10), is commonly used to overcome this problem [23, 24]. The k-fold approach produces samples that are utilized at least once for training and testing. The forecasting is carried out for each month of the year 2014.
7. *Forecasting*: Once the dataset is divided into training, validation and testing phases, different predictors i.e., learning based models are employed to the dataset. This study applied total nine models to analyze the effect of different dataset. The brief explanation of the theoretical background of all models are as follows:

Feed-Forward Neural Network (M1):

This network is made up of three layers: input, hidden, and output. If the connection weight between input and hidden hidden layer is 'ωij',

hidden layer bias is 'boj', hidden layer to output layer connection weight is 'ωjt', and bias at output is 'Ot', then the net of the model may be described as [25].

$$O_{net,j} = \sum_{i=1}^{N} I_i \omega_{ij} + b_{oj} \tag{6}$$

$$I_j = g(O_{net,j}) \tag{7}$$

where, $g(O_{net,j}) = \frac{1}{1-exp^{-O_{net,j}}}$ is sigmoid function used as activation.

Generalized Regression Neural Network (M5):

The GRNN structure is divided into four layers: input layer, hidden layer, summation layer, and division layer, with the hidden layer employing a Gaussian activation function [26, 27]. The network can be represented as:

$$I(x) = \frac{\sum_{k=1}^{N} I_k k(x, x_k)}{\sum_{k=1}^{N} k(x, x_k)} \tag{8}$$

where $I(x)$ represents the predicted value of x.
'I_k' represents activation weight at neuron *'k'* and $(x, x_k) = e^{-d_k/2\sigma^2}$, $d_k = (x - x_k)'(x - x_k)$ which is Gaussian kernel and *'d_k'* is squared distance between training & input samples.

Nonlinear Autoregressive (M2):

This model's feedback layer encloses the network's multiple levels and may be employed in both open and closed loop configurations. This network may be represented mathematically as [28]:

$$I(t+1) = F \begin{bmatrix} I(t)I(t-1), \ldots, & I(t-n_{GHI}), x(t+1) \\ x(t), x(t-1) \ldots & x(t-n_x) \end{bmatrix} \tag{9}$$

$$I(t+1) = F \begin{bmatrix} I_p(t)I_p(t-1), \ldots, & I_p(t-n_{GHI}), x(t+1) \\ x(t), x(t-1) \ldots & x(t-n_x) \end{bmatrix} \tag{10}$$

where $I(t+1)$ = forecasted output at time *'t'*
$I_p(t),\ I_p(t-1) \ldots$ = Past time output values
$I(t),\ I(t-1) \ldots$ = Present time output value
$x(t), x(t-1) \ldots$ = input values
n_x & n_y = input & output delays

Elman Neural Network (M3):

Elman introduced this network in 1990 as a sort of recurrent neural network. A buffer layer coupled the feedback from the output layer to itself. A recurrent layer is another name for this buffer layer [29]. If $I(k)$ is the network output & *'h(k)'* is hidden layer output then

$$I(k) = g(\omega_o)h(k) \tag{11}$$

where *'ω_o'* is the weight between hidden layer and output layer.
Assume 'm' is the input layer, 'n' is the output layer and 'j' is the hidden neuron. If $x(k-1)$ is the input to the network then hidden layer output '$h(k)$' & recurrent layer output $h_c(k)$ can be expressed as:

$$h(k) = f(\omega_r h_c(k) + \omega_i x(k-1))$$

Then $h_c(k) = h(k-1)$.
Where *'ω_i'* is the weight of input layer & hidden layer & *'ω_r'* is the weight of recurrent layer and hidden layer. *'f'* is sigmoid function used as activation function.

Cascade Forward Neural Network (M4):

Perceptron and multilayer networks are combined in this network. This network has a direct link between the input and output layers, as well as a link from the preceding layer to the output layer [30].

Support Vector Machine (M6):

This model belongs to the kernel-based technique category and is stated as follows for time sequence [31]:

$$I(t+h) = \sum_{i=1}^{N} \alpha_i k(I_t, I_i) + b \tag{12}$$

where α_i represents Lagrange multiplier, $k(I_t,\ I_i)$ is the kernel function and $k(I_t,\ I_i) = exp^{(\frac{-\|I_t - I_i\|^2}{\sigma^2})}$.
Where σ is RBF kernel function.

Linear Regression (M7):

This is one of the most straightforward models for estimating the output value by matching the input and output variables together.

The mathematical expression for this model is expressed as [32]:

$$I(t+h) = v_o + v_1 x_i + \varepsilon_i \tag{13}$$

where $x_i =$ input variable, $I =$ output target value

$$v_o = \overline{I} - v_1 \overline{x}$$

and

$$v_1 = \frac{\sum_{t=1}^{n} I_i x_i - \frac{\left(\sum_{i=1}^{n} I_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

and

$$\overline{I} = \frac{1}{n}\sum_{i=1}^{n} I_i \quad \& \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Bagged & Boosted Regression Tree (M8 & M9):

These models are widely used for classifications that follow an if-then logic. However, this approach was also used to forecast solar GHI [33]. This model functions as a decision tree, with each node working as a regression model and continually predicting future values. The boosted regression tree approach considers the average of different regression tree classifiers' predictions. The data is anticipated by consecutive trees, with poorly forecasted data giving the following tree a greater weight. Both strategies are used to increase forecast accuracy and are stated mathematically as [24]:
For Bagged Regression Tree

$$I(t+h) = \alpha v_p \phi_p(I(t+h)) \tag{14}$$

where 'ϕ_p' & 'αv_p' are the before aggregation predictor & average of predictors respectively.
For Boosted Regression Tree

$$I(t+h) = \sum_{n} \beta_n L(I(t+h), \gamma_n \tag{15}$$

L is used for representation of each tree with split variable γ_n and weight of each node β_n.

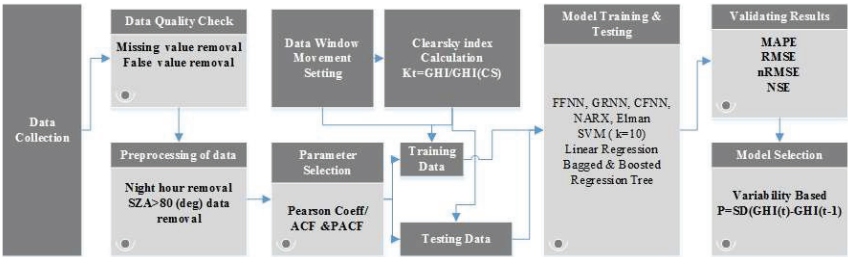


Figure 2 Forecasting mechanism used in the study.

8. *Error Calculations*: After training the model, various errors: MAPE, RMSE are computed. Section 4 provides into further information about these errors.
9. *Model selection*: To choose a model, the user must first take the variability of a certain month into account as a statistical parameter. For the estimate of variability, the study used the Equation (16). The variable 'P' offers information on the data spread in a testing month, which may be used to choose a specific model for any uncertainty. Table 4 shows the variability in the testing dataset.

$$P = Std.Dev.(k_t - k_{t-1}) \tag{16}$$

where k_t & k_t = Clear sky index at time 't' and '$t-1$'.

4 Performance Evaluation

The error indices used to evaluate the models are as follows [34]:

Mean absolute percentage error: MAPE used to find forecasting accuracy of any model in percentage form. This is simply the representation of uniform error in percentage form.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{I_{f,i} - I_{r,i}}{I_{r,i}}\right| \times 100 \tag{17}$$

where $I_{f,i}$ is forecasted GHI and $I_{r,i}$ is real/actual GHI.

Root Mean Square Error: It is one of the popular error measures used to evaluate the performance of any forecasting model. This measure identifies

the largest error in the forecasted sequence.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(I_{f,i} - I_{r,i}\right)^2} \tag{18}$$

5 Experimental Results

This section outlines the experimental setup that is used to evaluate the models' performance. This study is being carried out to compare the performance of several models using structural and endogenous data. The models generated for the various datasets are initially configured using the training and validation dataset. The study employed a manual search strategy to find the best parameters for the constructed models. The structural dataset uses a total of 8 input neurons: dew point, temperature, pressure, relative humidity, sun zenith angle, precipitation, wind direction, and wind speed. Whereas, f or endogenous datasets, the input neurons are 10 (t-1 to t-10 historical solar GHI). The time delays are chosen based on the ACF and PACF graphical patterns presented in Section 2. M1 and M4 model are configured with 1 hidden layer, LM algorithm for learning, 'tansig', 'purelin' as transfer function, learning rate 0.001 and momentum constant of 0.06 for both datasets. The M6 model's Kernel function is set to a 'Gaussian function' with a kernel scale of 2.8 with k-fold value of 10. The M2 and M3 model uses a input delay of 1:2. In M7, M8 and M9 models, the k-fold value used same as of M6 model. The default settings for all other parameters in the models stay unaltered. The models are created using the MATLAB 2020b environment. To train the models, the dataset is separated into training and testing data. The comprehensive experiment with training window of 1 year, 2 years, 3 years and 4 years are conducted to select the best training size using RMSE as a loss function. Table 1(a) and 1(b) show the real report of all model's MAPE and RMSE for each month for 1 year, 2 years, 3 years, and 4 years.

The performance of the various models varies depending on the training size of the data. However, it has been observed that all discussed models attained better performance with a training data size of three years in case of structural data. Despite this, the M1 and M2 performed better for two years, M3, M5, M8, and M9 performed better for three years, while M4, M6, and M7 showed improvement for four years of training data size in the endogenous dataset. All of the models were run for a one-day GHI forecast for each month, with data flowing according to the moving window

Table 1(a) Observed statistical results for structural dataset with different training size

	1 yr		2 yr		3 yr		4 yr	
Models	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
M1	5.11	39.96	4.99	39.48	4.97	40.87	5.06	39.64
M2	5.09	39.25	5.06	39.33	4.96	39.71	5.12	38.95
M3	5.45	41.45	5.17	40.89	4.99	40.67	4.99	40.65
M4	5.88	41.88	5.19	40.56	5.18	40.41	5.20	41.91
M5	5.13	39.85	5.00	39.47	4.88	40.20	5.09	39.90
M6	4.95	41.45	4.85	41.29	4.84	41.28	4.84	41.13
M7	5.93	41.54	5.42	41.23	5.29	40.87	5.42	40.45
M8	7.77	45.37	7.78	45.34	7.52	44.48	7.56	44.61
M9	5.44	41.22	5.04	40.32	4.91	40.55	4.97	40.48

Table 1(b) Observed statistical results for endogenous dataset with different training size

	1 yr		2 yr		3 yr		4 yr	
Models	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
M1	6.49	46.52	6.49	46.01	6.95	48.92	7.05	49.53
M2	6.66	46.27	6.52	46.36	6.61	47.38	6.80	45.86
M3	8.12	46.95	7.91	46.15	7.73	45.49	7.74	45.50
M4	6.62	47.28	6.28	45.56	5.90	45.01	6.64	45.30
M5	6.97	48.91	6.74	49.49	7.04	48.80	6.74	46.82
M6	5.95	46.36	5.96	45.86	6.07	47.00	5.89	44.82
M7	7.65	52.46	7.51	51.16	7.51	51.60	7.48	50.83
M8	10.42	65.60	10.07	62.01	9.90	61.72	10.01	61.16
M9	8.03	58.57	7.74	55.02	7.60	54.89	7.71	54.86

technique. Table 2(a) & 2(b) for MAPE & RMSE show the comparison findings for different models on both datasets.

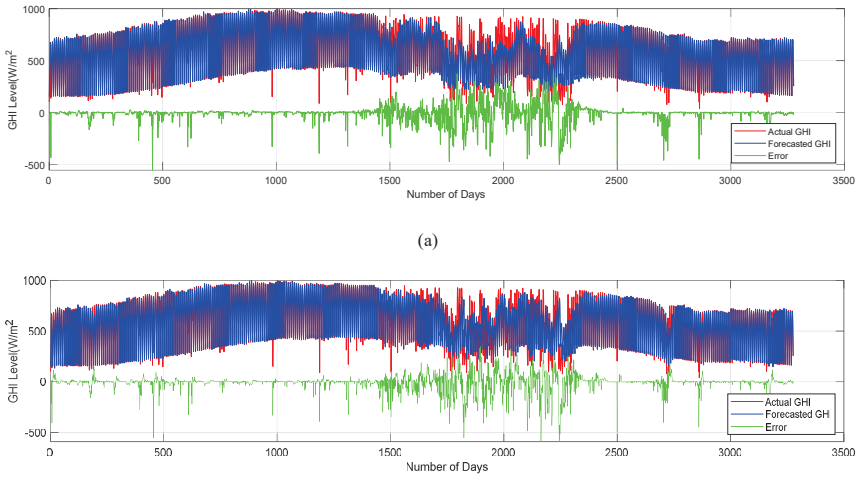
Table 2(a) and 2(b) compare the MAPE (%) and RMSE (W/m^2) of all models for both datasets. For all models, the structural datasets produced better results than the endogenous datasets. For the structural data set, the M6 model outperformed the other models in terms of MAPE (4.84%, 5.22%, 2.30%, 1.99%, 4.30%, 26.14%, 7.63%, 5.47%, and 3.91% for months 1, 2, 3, 4, 5, 7, 10, 11, and 12) and RMSE (41.28 W/m^2, 53.54 W/m^2, 32.90 W/m^2, 23.39 W/m^2, 42.39 W/m^2, 138.98 W/m^2, 67.15 W/m^2, 36.06 W/m^2, and 38.48 W/m^2 for months 1, 2, 3, 4, 5, 7, 10, 11, and 12) Similarly, for endogenous datasets, this model performed better in months 1, 2, 3, 4, 5, 6, 11, and 12 with MAPE of 5.89%, 6.03%, 2.44%, 1.98%, 4.60%, 9.27%, 6.10%, and 4.35%; RMSE of 44.83 W/m^2, 55.87 W/m^2, 32.53 W/m^2, 23.64 W/m^2, 43.02 W/m^2, 72.38 W/m^2, 39.41 W/m^2. However, with an

Table 2(a) MAPE Results for Structural dataset & Endogenous dataset (TS)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
M1	4.97	5.33	3.01	2.31	4.56	9.16	24.46	26.02	30.96	7.88	5.60	3.98
M1-TS	6.49	7.28	3.10	2.27	5.52	9.56	28.97	27.89	31.56	6.36	6.41	4.63
M2	4.96	5.18	2.63	2.26	4.77	9.30	26.50	26.18	29.73	7.92	5.52	3.96
M2-TS	6.52	6.28	3.01	2.23	5.27	9.39	29.64	28.97	32.40	6.49	7.72	4.34
M3	5.18	5.31	2.70	2.36	5.11	8.98	24.83	26.91	32.16	8.53	5.76	4.07
M3-TS	5.90	5.98	3.03	2.64	5.59	9.42	29.42	28.24	31.71	6.54	6.85	4.90
M4	4.88	5.25	2.40	2.04	4.41	8.98	25.01	26.87	30.73	7.77	5.29	4.08
M4-TS	6.74	7.41	2.93	2.43	5.66	9.54	28.77	27.66	31.19	6.38	6.68	4.64
M5	4.99	5.50	3.17	2.72	4.66	9.18	28.49	29.44	33.60	7.79	5.51	4.20
M5-TS	7.73	7.72	6.19	6.20	8.38	9.91	39.23	32.71	40.29	10.19	9.02	6.34
M6	4.84	5.22	2.30	1.99	4.30	9.28	26.14	28.36	34.01	7.63	5.47	3.91
M6-TS	5.89	5.43	2.44	1.98	4.60	9.27	32.62	29.74	33.18	6.81	6.10	4.35
M7	5.29	5.55	3.18	2.43	4.52	9.46	25.58	27.37	32.06	8.49	5.99	4.29
M7-TS	7.48	8.02	3.88	3.10	6.31	9.82	29.65	27.94	32.86	6.86	7.64	5.20
M8	7.52	7.51	6.23	6.09	8.16	11.65	25.36	28.24	31.04	10.56	9.18	6.43
M8-TS	9.90	10.19	7.23	6.47	9.40	12.25	29.17	28.03	33.15	10.09	10.41	7.79
M9	4.91	5.41	3.30	2.48	4.59	9.23	25.51	27.19	32.70	8.06	5.58	3.92
M9-TS	7.60	8.16	3.75	2.96	5.91	10.05	29.48	27.78	32.07	6.97	7.41	5.03

Table 2(b) RMSE (W/m^2) results for Structural dataset & Endogenous dataset (TS)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
M1	40.87	51.97	34.05	24.15	41.49	71.39	130.95	148.79	135.44	67.44	34.48	38.13
M1-TS	46.01	61.81	35.82	25.70	47.11	70.31	148.99	155.35	142.43	59.15	39.80	38.19
M2	39.71	52.78	33.48	24.29	42.62	69.89	136.74	145.92	137.32	64.16	35.71	37.34
M2-TS	46.36	54.53	33.33	26.67	45.90	70.55	153.85	158.01	151.77	56.94	45.72	35.96
M3	40.41	52.32	32.87	24.03	41.59	66.92	133.93	150.06	139.51	66.53	34.30	37.20
M3-TS	45.01	54.15	33.86	26.92	46.95	69.22	150.00	158.15	139.82	57.69	40.38	37.51
M4	40.20	51.47	33.17	25.37	43.76	70.94	132.73	154.53	133.58	64.47	34.51	37.48
M4-TS	46.82	61.71	35.04	28.26	47.18	71.07	148.91	152.56	139.75	58.25	42.19	36.63
M5	40.67	52.62	33.48	26.68	43.75	70.20	144.13	156.92	151.86	67.32	39.45	38.61
M5-TS	45.49	56.13	48.01	48.04	57.98	72.22	187.43	175.27	172.85	65.31	44.52	41.08
M6	41.28	53.54	32.90	23.39	42.39	72.47	138.98	164.07	155.22	67.15	36.06	38.48
M6-TS	44.83	54.87	32.53	23.64	43.02	72.38	169.02	167.22	153.92	61.98	39.41	40.26
M7	40.87	50.75	33.92	24.01	40.87	72.16	134.53	154.90	139.44	69.15	35.77	37.52
M7-TS	50.83	64.28	38.33	31.40	49.64	73.41	151.71	154.33	146.89	59.53	48.76	39.36
M8	44.48	54.89	47.91	46.34	55.90	89.13	137.41	165.21	134.57	66.81	44.85	41.54
M8-TS	61.72	73.57	56.46	50.53	65.32	91.65	151.86	155.15	145.85	65.97	53.87	47.03
M9	40.55	52.59	34.58	24.24	41.28	70.25	138.75	161.39	144.41	66.00	34.27	37.06
M9-TS	54.89	67.97	39.48	29.32	48.69	78.70	149.20	152.73	144.21	58.47	47.81	39.48



(a)

(b)

Figure 3 (a) and (b): Yearly forecasted solar GHI visualization for structural & endogenous datasets resp.

endogenous dataset, M4 produces better results for months 7, 8, 9, and 10, but with a structural data set, it produces better results for month 6. For a year Figure 3(a) & (b) present graphical depictions and tracing of the real GHI curve by model's anticipated values for both datasets. The largest error production in the anticipated solar GHI occurs when the irradiation is uncertain, as shown in the Figure 3. The wet and foggy qualities of the day cause this fluctuation in the solar GHI, which may be identified using the variability estimation or clear sky index as discussed in Section 3.

6 Discussion

It's tough to extrapolate generic conclusions for every site from this little study. However, this part attempts to offer broad assertions concerning dataset and model selection in order to achieve excellent accuracy for one-day ahead monthly forecasting.

***Selection of datasets*:** The results of several forecasting models based on structural datasets and endogenous datasets are shown in Table 2(a) and 2(b). These findings show that the meteorological dataset, when compared to the endogenous dataset, gives better accuracy for day ahead forecasting by all

models. In the case of endogenous dataset, the findings of certain months are better, but the difference is minor and may be neglected. For instance, using endogenous data, for April month, the MAPE of M6 model is 1.98%; whereas, it is 1.99% using structural dataset. In other words, the MAPE for both data by M6 model differs by 0.04%. Despite this, there is a discernible difference in several months where endogenous data-based models outperformed structural data-based models. For a day ahead monthly forecast, it is obvious that meteorological data-based models give better accuracy than time series-based models.

***Selection of Models*:** For the same months, various models produce different outcomes for different regions. For both datasets, the MAPE derived by the M6 model is lower in January, February, March, April, May, November, and December. However, with M4 and M6 model for structural data and endogenous dataset, June has a lower MAPE. This pattern is also visible in July, August, September, and October. As a result, various models have lower MAPE in different months. These variances in the outcome, however, are related to the varying qualities of the days in a month, as well as the month's data unpredictability. Thus, what model should be used for a given sort of month or unique type of data variability? The answer to this question may be found by estimating the dataset's variability or by utilising 'kt'.

The clear sky index of a day and month may also be used to assess the qualities of a day. The value of 'kt' may be categorized into three categories, according to Sayago et al., which correspond to the distinct features of days [35]. The cloudy day can be considered with range of $0.03 \leq k_t < 0.30$, partially cloudy day can be considered as range of $0.3 \leq k_t < 0.7$ and clear day can be considered as $0.7 \leq k_t < 0.9$.

The developed models are practically synchronized in their responses to the same sort of variability. The value of factor 'P' must be found using Equation (7) to quantify the variability in the dataset. A large 'P' value is expected to have a high MAPE, whereas a smaller 'P' value is expected to have a lower MAPE. The largest MAPE represents wet or overcast months with undetermined types of fluctuations in the dataset, whereas the lowest MAPE represents bright days with low solar GHI variations. The study's dataset has three forms of variability: moderate, medium, and strong variability. The different variability values obtained from Equation (7) are shown in Table 3.

Table 3 shows that the months of January, February, March, April, May, and December has the least uncertainty. June, September, October, and

Table 3 Observed variability of different months

Variability of the Months												
Month	1	2	3	4	5	6	7	8	9	10	11	12
Variability	0.08	0.09	0.06	0.06	0.09	0.12	0.17	0.18	0.15	0.12	0.10	0.08

Table 4 Selection of model based on dataset variability

Weaker Variability	Medium Variability	Stronger Variability
M6	M4, M6	M2, M4, M6

November have a medium degree of fluctuation, whilst July and August have a higher level of variability, ranging from 0.12–0.17. The lower variability shows the smaller data spread in a given month's datasheet. The medium variability shows a respectable degree of data dispersion, whereas the higher variability represents massive fluctuations in the GHI in the past and present. The more the variability, the more doubtful the data is. For both data sets, the M6 model performed best for all months with lower variability. In both datasets, the M4 and M6 models show greater accuracy because to the medium variability of the data. Furthermore, M6 and M4 can tolerate higher fluctuation, such as in the months of July and August. M2, on the other hand, can manage medium variability of structural data well, as in September, but M4 can also handle both with a decent margin in structural data. Table 4 depicts model selection based on dataset variability.

Therefore, in the aforementioned study, the M6 model produces superior results for data with lesser variability, while the M6 and M4 are appropriate possibilities for data with medium and greater variability.

***Training Size*:** The impact of data training amount on every model may be observed immediately in Table 1. However, for the structural kind of dataset, a training size of three years is acceptable, but the endogenous dataset requires two to four years of data. Because the variability of the dataset varies for different types of datasets, a repeated experiment is required to determine the best endogenous data training size. However, three years of data can yield the best results from any model with any sort of datasheet in general.

Training Parameters: One of the most significant factors in obtaining correct results is the parameters used in model learning. The more relevant parameters in meteorological parameters, as in structural datasets, resulting in an accurate forecast. Air pollution and other air-related characteristics, in addition to some climatic elements, might alter the model's performance. However, air quality measures such as PM2.5, AQI, and others are not

included in this study. However, the study might be redone in the future with these factors included. Similarly, in the case of endogenous datasets, the correct selection of temporal delays leads to an accurate forecast.

***Dataset selection*:** The results show that taking into account the structural dataset, such as meteorological characteristics, can result in high accuracy. The presence of mistakes in the meteorological datasets given by local weather forecasting organizations, on the other hand, has an impact on the model's performance. Nowadays, accurate measurement of meteorological data is also a cause of worry, posing a significant threat of prediction mistakes. More metrological parameters may result in more effective model learning, but they can result in a bigger investment in the instruments. Due to geographical and other factors, it is often impossible to monitor meteorological data for a specific site. In such instances, the time-series technique, i.e., endogenous datasets, emerges as a viable option for forecasting solar GHI. In such circumstances, the prediction of the following step GHI requires simply a historical GHI. A single device called a pyranometer must be positioned on the target site for this technique to work. The historical GHI of surrounding locations might also aid if the place is difficult to reach. As a result, in most prior investigations, time series forecasting methods were preferred to anticipate the solar GHI.

7 Conclusion

Using a structural and endogenous data set, nine machine learning based models are constructed and evaluated. These models are created to anticipate the monthly day ahead GHI using the clear-sky index for Indian location. From the study, it is observed that the structural dataset outperformed the endogenous dataset. For meteorological datasets, MAPE and RMSE vary from 1.99% to 29.73% and 23.39 W/m^{2}to 165.21 W/m^2 respectively; for endogenous datasets, these errors range from 1.98% to 31.19% and 23.64 W/m^2 to 152.56 W/m^2, respectively. Moreover, the study demonstrates that the data has various levels of variability. The lesser variability is discovered in March, whereas the larger variability is observed in August. The SVM model provided the best fit for all types of variability with fair output among contrast models. CFNN, on the other hand, may be utilized for medium and stronger variability. Furthermore, given the data used, the GRNN, Elman, NARX, LR, Boosted RT, and Bagged RT approaches did not perform better. Further, the training size of three years for structural dataset is suitable to achieve best results which are 2–4 years for endogenous dataset.

References

[1] P. Singla, M. Duhan, and S. Saroha, "A comprehensive review and analysis of solar forecasting techniques," *Front. Energy*, pp. 1–37, Mar. 2021, doi: 10.1007/s11708-021-0722-7.

[2] P. Singla, M. Duhan, and S. Saroha, "Solar Irradiation Forecasting by Long-Short Term Memory Using Different Training Algorithms," pp. 81–89, 2022, doi: 10.1007/978-981-16-4663-8_7.

[3] S. Sobri, S. Koohi-Kamali, and N. A. Rahim, "Solar photovoltaic generation forecasting methods: A review," *Energy Conversion and Management*, vol. 156. Elsevier Ltd, pp. 459–497, Jan. 15, 2018. doi: 10.1016/j.enconman.2017.11.019.

[4] P. Singla, M. Duhan, and S. Saroha, "An ensemble method to forecast 24-h ahead solar irradiance using wavelet decomposition and BiLSTM deep learning network," *Earth Sci. Informatics*, vol. 15, no. 1, pp. 291–306, Nov. 2021, doi: 10.1007/S12145-021-00723-1/TABLES/7.

[5] G. Notton, C. Voyant, A. Fouilloy, J. L. Duchaud, and M. L. Nivet, "Some applications of ANN to solar radiation estimation and forecasting for energy applications," *Appl. Sci.*, vol. 9, no. 1, pp. 1–21, 2019, doi: 10.3390/app9010209.

[6] D. O'Leary and J. Kubby, "Feature Selection and ANN Solar Power Prediction," *J. Renew. Energy*, vol. 2017, pp. 1–7, Nov. 2017, doi: 10.1155/2017/2437387.

[7] R. Meenal, A. Immanuel Selvakumar, S. Berclin JeyaPrabha, and E. Rajasekaran, "Solar Mapping of India using Support Vector Machine," *J. Phys. Conf. Ser.*, vol. 1142, p. 012010, Nov. 2018, doi: 10.1088/1742-6596/1142/1/012010.

[8] A. Zendehboudi, M. A. Baseer, and R. Saidur, "Application of support vector machine models for forecasting solar and wind energy resources: A review," *Journal of Cleaner Production*, vol. 199. Elsevier Ltd, pp. 272–285, Oct. 20, 2018. doi: 10.1016/j.jclepro.2018.07.164.

[9] W. Sharika, L. Fernando, A. Kanagasundaram, R. Valluvan, and A. Kaneswaran, "Long-term Solar Irradiance Forecasting Approaches – A Comparative Study," Dec. 2018. doi: 10.1109/ICIAFS.2018.8913381.

[10] C. G. Ozoegwu, "Artificial neural network forecast of monthly mean daily global solar radiation of selected locations based on time series and month number," *J. Clean. Prod.*, vol. 216, pp. 1–13, Apr. 2019, doi: 10.1016/J.JCLEPRO.2019.01.096.

[11] W. I. Hameed et al., "Prediction of Solar Irradiance Based on Artificial Neural Networks," *Inventions*, vol. 4, no. 3, p. 45, Aug. 2019, doi: 10.3390/inventions4030045.

[12] S. Amely Jumaat, F. Crocker, M. Helmy Abd Wahab, N. Hanis Mohammad Radzi, and M. Fakri Othman, "Prediction of Photovoltaic (PV) Output Using Artificial Neutral Network (ANN) Based on Ambient Factors," *J. Phys. Conf. Ser.*, vol. 1049, p. 012088, Jul. 2018, doi: 10.1088/1742-6596/1049/1/012088.

[13] S. Al-Dahidi, O. Ayadi, J. Adeeb, and M. Louzazni, "Assessment of Artificial Neural Networks Learning Algorithms and Training Datasets for Solar Photovoltaic Power Production Prediction," *Front. Energy Res.*, vol. 7, p. 130, Nov. 2019, doi: 10.3389/fenrg.2019.00130.

[14] S. Rahman, S. Rahman, and A. K. M. Bahalul Haque, "Prediction of Solar Radiation Using Artificial Neural Network," *J. Phys. Conf. Ser.*, vol. 1767, no. 1, p. 012041, Feb. 2021, doi: 10.1088/1742-6596/1767/1/012041.

[15] L. Huang, J. Kang, M. Wan, L. Fang, C. Zhang, and Z. Zeng, "Solar Radiation Prediction Using Different Machine Learning Algorithms and Implications for Extreme Climate Events," *Front. Earth Sci.*, vol. 9, p. 202, Apr. 2021, doi: 10.3389/FEART.2021.596860/BIBTEX.

[16] M. Guermoui, K. Gairaa, J. Boland, and T. Arrif, "A Novel Hybrid Model for Solar Radiation Forecasting Using Support Vector Machine and Bee Colony Optimization Algorithm: Review and Case Study," *J. Sol. Energy Eng.*, vol. 143, no. 2, Apr. 2021, doi: 10.1115/1.4047852.

[17] A. El Mouatasim and Y. Darmane, "Regression analysis of a photovoltaic (PV) system in FPO," in *AIP Conference Proceedings*, Dec. 2018, vol. 2056, no. 1, p. 020008. doi: 10.1063/1.5084981.

[18] P. Singla, M. Duhan, and S. Saroha, "A Hybrid Solar Irradiance Forecasting Using Full Wavelet Packet Decomposition and Bi-Directional Long Short-Term Memory (BiLSTM)," *Arab. J. Sci. Eng.*, pp. 1–27, Mar. 2022, doi: 10.1007/S13369-022-06655-2/TABLES/18.

[19] P. Singla, M. Duhan, and S. Saroha, "A dual decomposition with error correction strategy based improved hybrid deep learning model to forecast solar irradiance," https://doi.org/10.1080/15567036.2022.2056267, vol. 44, no. 1, pp. 1583–1607, Mar. 2022, doi: 10.1080/15567036.2022.2056267.

[20] "Photovoltaic Power Systems Technology Collaboration Program Snapshot of Global PV Markets: Report IEA PVPS T1-35:2019," 2019.

[21] "Ahmedabad climate: Average Temperature, weather by month, Ahmedabad weather averages – Climate-Data.org." https://en.climate-data.org/asia/india/gujarat/ahmedabad-2828/ (accessed Apr. 24, 2020).

[22] C. Yousif, G. O. Quecedo, and J. B. Santos, "Comparison of solar radiation in Marsaxlokk, Malta and Valladolid, Spain," *Renew. Energy*, vol. 49, pp. 203–206, 2013, doi: 10.1016/j.renene.2012.01.031.

[23] J. Zeng and W. Qiao, "Short-term solar power prediction using a support vector machine," *Renew. Energy*, vol. 52, pp. 118–127, Apr. 2013, doi: 10.1016/j.renene.2012.10.009.

[24] C. Huang, L. Wang, and L. L. Lai, "Data-Driven Short-Term Solar Irradiance Forecasting Based on Information of Neighboring Sites," *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9918–9927, Dec. 2019, doi: 10.1109/TIE.2018.2856199.

[25] S. Y. Heng et al., "Artificial neural network model with different backpropagation algorithms and meteorological data for solar radiation prediction," *Sci. Reports 2022 121*, vol. 12, no. 1, pp. 1–18, Jun. 2022, doi: 10.1038/s41598-022-13532-3.

[26] M. Sridharan, "Application of Generalized Regression Neural Network in Predicting the Performance of Natural Convection Solar Dryer," *J. Sol. Energy Eng. Trans. ASME*, vol. 142, no. 3, Jun. 2020, doi: 10.1115/1.4045384/1066328.

[27] G. Dudek, "Generalized regression neural network for forecasting time series with multiple seasonal cycles," *Adv. Intell. Syst. Comput.*, vol. 323, pp. 839–846, 2015, doi: 10.1007/978-3-319-11310-4_73.

[28] Z. Boussaada, O. Curea, A. Remaci, H. Camblong, and N. M. Bellaaj, "A nonlinear autoregressive exogenous (NARX) neural network model for the prediction of the daily direct solar radiation," *Energies*, vol. 11, no. 3, 2018, doi: 10.3390/en11030620.

[29] I. Majumder, R. Bisoi, N. Nayak, and N. Hannoon, "Solar power forecasting using robust kernel extreme learning machine and decomposition methods," *Int. J. Power Energy Convers.*, vol. 11, no. 3, pp. 260–290, 2020, doi: 10.1504/IJPEC.2020.107958.

[30] "Cascade-forward neural network - MATLAB cascadeforwardnet." https://www.mathworks.com/help/deeplearning/ref/cascadeforwardnet.html;jsessionid=b7bd9600b0fddad3c28a11df0cd7 (accessed Apr. 24, 2020).

[31] S. Shamshirband et al., "Estimating the diffuse solar radiation using a coupled support vector machine-wavelet transform model," *Renewable*

and Sustainable Energy Reviews, vol. 56. Elsevier Ltd, pp. 428–435, Apr. 01, 2016. doi: 10.1016/j.rser.2015.11.055.

[32] S. Ibrahim, I. Daut, Y. M. Irwan, M. Irwanto, N. Gomesh, and Z. Farhana, "Linear regression model in estimating solar radiation in perlis," *Energy Procedia*, vol. 18, no. September 2015, pp. 1402–1412, 2012, doi: 10.1016/j.egypro.2012.05.156.

[33] R. Ahmed, V. Sreeram, Y. Mishra, and M. D. Arif, "A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization," *Renew. Sustain. Energy Rev.*, vol. 124, no. June 2019, p. 109792, 2020, doi: 10.1016/j.rser.2020.109792.

[34] P. Singla, M. Duhan, and S. Saroha, "Review of Different Error Metrics: A Case of Solar Forecasting," *AIUB J. Sci. Eng.*, vol. 20, no. 4, pp. 158–165, Dec. 2021, doi: 10.53799/AJSE.V20I4.212.

[35] S. Sayago, G. Ovando, J. Almorox, and M. Bocco, "Daily solar radiation from NASA-POWER product: assessing its accuracy considering atmospheric transparency," *Int. J. Remote Sens.*, vol. 41, no. 3, pp. 897–910, Feb. 2020, doi: 10.1080/01431161.2019.1650986.

Biographies

Pardeep Singla currently a research scholar at Department of Electronics and Communication Engineering, Deenbandhu Chhotu Ram University of Science and Technology, Sonepat, Haryana, India. He has published several research papers in reputed international SCI indexed journals. His area of research interest is solar forecasting, wind forecasting using machine and deep learning.

Manoj Duhan currently working as professor at the Department of Electronics and Communication Engineering, Deenbandhu Chhotu Ram University of Science and Technology, Murthal. Sonipat Haryana, India. He has published several research papers in reputed international SCI indexed journals. He does research in Electronic Engineering, solar forecasting, reliability Engineering and biomedical signal processing.

Sumit Saroha currently working as assistant professor at the Department of Electrical Engineering, Guru Jambheshwar University of Science and Technology, Hisar, Haryana, India. He has published several research papers in reputed international SCI indexed journals and earned various patents. He does research in solar forecasting and wind forecasting.