Environmental Science and Pollution Research (2021) 28:26571–26583
https://doi.org/10.1007/s11356-021-12435-6

RESEARCH ARTICLE

Solar radiation prediction using boosted decision tree regression model: A case study in Malaysia

Ellysia Jumin[1] · Faridah Bte Basaruddin[2] · Yuzainee Bte. Md Yusoff[1] · Sarmad Dashti Latif[1] · Ali Najah Ahmed[3]

Received: 12 April 2020 / Accepted: 7 January 2021 / Published online: 23 January 2021

Abstract

Reliable and accurate prediction model capturing the changes in solar radiation is essential in the power generation and renewable carbon-free energy industry. Malaysia has immense potential to develop such an industry due to its location in the equatorial zone and its climatic characteristics with high solar energy resources. However, solar energy accounts for only 2–4.6% of total energy utilization. Recently, in developed countries, various prediction models based on artificial intelligence (AI) techniques have been applied to predict solar radiation. In this study, one of the most recent AI algorithms, namely, boosted decision tree regression (BDTR) model, was applied to predict the changes in solar radiation based on collected data in Malaysia. The proposed model then compared with other conventional regression algorithms, such as linear regression and neural network. Two different normalization techniques (Gaussian normalizer binning normalizer), splitting size, and different input parameters were investigated to enhance the accuracy of the models. Sensitivity analysis and uncertainty analysis were introduced to validate the accuracy of the proposed model. The results revealed that BDTR outperformed other algorithms with a high level of accuracy. The funding of this study could be used as a reliable tool by engineers to improve the renewable energy sector in Malaysia and provide alternative sustainable energy resources.

Keywords Solar radiation prediction · Weather parameters · Machine learning algorithm · Boosted decision tree regression · Neural network and linear regression · Correlation coefficient

Introduction

A tremendous increase in the world population by almost five times in 2025, as predicted according to the United Nations, will result in a great reliance on an ample and uninterrupted supply of energy to live and work (Kitani et al. 1999). An alternative sustainable energy resource is essential to overcome global environmental problems and energy-related fossil resource exhaustion, which present significant challenges. Solar energy is a major type of renewable energy, and its estimation is important for decision-makers (Ghazvinian et al. 2019). Accurate global solar radiation data are fundamental information for the allocation and design of solar energy systems (Feng et al. 2019). Vast knowledge of

Responsible Editor: Philippe Garrigues

✉ Sarmad Dashti Latif
Sarmad.latif@uniten.edu.my

Ellysia Jumin
ellysiaelis@gmail.com

Faridah Bte Basaruddin
Faridah@uniten.edu.my

Yuzainee Bte. Md Yusoff
Yuzainee@uniten.edu.my

Ali Najah Ahmed
Mahfoodh@uniten.edu.my

[1] Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), 43000 Kajang, Selangor Darul Ehsan, Malaysia

[2] Department of Mechanical Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), 43000 Kajang, Selangor Darul Ehsan, Malaysia

[3] Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional (UNITEN), 43000 Kajang, Selangor, Malaysia

daily solar radiation that reaches the surface of the earth is essential as the radiation affects the energy balance of the earth's atmospheric system. Estimating future energy output will also entail predicting solar radiation (Wu et al. 2014; Olatomiwa et al. 2015a; Qazi et al. 2015; Alsina et al. 2016; Aybar-Ruiz et al. 2016; Wu and Wang 2016; Chia et al. 2020). Accurate estimation of solar radiation due to the lack of measured solar radiation has been a challenging task (Rabehi et al. 2020). Various models incorporating weather parameters have been developed and applied in predicting solar radiation because of the lack of the instrument of solar radiation measuring at most meteorological stations (Ghimire et al. 2019b). The instruments are also very costly and need calibration (Ghimire et al. 2019b).

As an alternative solution in the lack of measured solar radiation, Chen et al. (2013) proposed a potential support vector machine (SVM) using sunshine duration for estimating daily solar radiation. Seven models of SVM with different input attributes and five models of empirical sunshine are tested using climatological data at three stations in the province of Liaoning in China. All models of SVM outperformed the empirical models considerably. The SVM model utilizing sunshine ratio as an attribute that performs better in winter is preferred because its accuracy is greater and also due to its simple input attribute. However, a higher number of root mean square error (RMSE) and also relative root mean square error (RRMSE) were achieved in the summer season. The season-dependent SVM model in estimating the regular solar radiation in the winter is superior to the set one while fixing seasonal variation of the sets of data does not lead to improve the result in summer, spring, and autumn. Besides sunshine duration, weather parameters are routinely measured since many studies showed that these climatological variables, in addition to sunshine, can enhance the model's accuracy (Chen et al. 2013). In case if the data is unavailable, the daily solar radiation could well be estimated using the data from the nearby meteorological station that covers all areas of the province.

Ramedani et al. (2014) compared support vector regression (SVR) and fuzzy linear regression (FLR) models for universal solar radiation forecast in Iran. Two SVR models with polynomial functions and radial basis were investigated. The performance of SVR is better than FLR, and the result showed that SVR with radial basis function produced the best estimation of universal solar radiation with shorter computation time. Previous studies on solar radiation forecasting using artificial neural network (ANN) techniques and regression analysis are employed and have shown significant prediction results. Due to its self-learning and adaptive power, ANN has the ability to allow nonlinear neural architectures to achieve accurate simulation results, which reduce human interventions (Zou et al. 2017).

One study found that ANN techniques more reliably forecast solar radiation than traditional methods. However, the forecasting accuracy of ANN models depends on combinations of input parameters, training algorithms, and configurations of architecture (Yadav and Chandel 2014). A comparison study has been done between the fuzzy genetic approach, ANFIS, and ANN model's ability to estimate the solar radiation in Turkey (Kisi 2014). Olatomiwa et al. (2015b) applied an adaptive neuro-fuzzy inference system (ANFIS) model to simulate solar radiation in Nigeria efficiently.

The proposed ANFIS model incorporated monthly minimum and maximum monthly mean temperature and sunshine. ANFIS network composed of three input layer neurons and one output layer neuron was used to simulate the solar radiation. The prediction results of RMSE were 1.0854, and the coefficient of determination, R^2 was 0.8544 that obtained in the training phase, and RMSE was 1.7585, and R^2 was 0.6567 in the testing phase. The model's output is entirely location-based, so a general model calibration may be possible if the climate conditions around the area are identical. The ANFIS model could be combined with other soft computing techniques as well, and more meteorological input variables should be analyzed to enhance prediction accuracy (Olatomiwa et al. 2015b). Machine learning (ML) models were used to identify climate patterns contributed by meteorological variables such as sunshine, humidity, and temperature embedded in atmospheric data to simulate daily solar radiation (Falayi et al. 2008; Bilgili and Ozgoren 2011; Yacef et al. 2012).

Integrated supporting vector machine and discrete wavelet transformation algorithm in the development of short- and long-term global incident solar radiation forecasting model applied at several meteorological stations in Australia. Solar exposure has proved to be the most powerful predictor variable for the daily forecasting model for all the stations. However, the wavelet-couple model used all the inputs to generate the best forecast for the Brisbane City and Cairns Aero stations. As contrary to the above, for Townsville Aero, the incorporation of precipitation and wind speed time series appeared to deteriorate the performance. The geographic location of the weather station is playing a significant effect in forecasting accuracy (Deo et al. 2016).

Fan et al. (2018) performed a comprehensive review of fourteen existing and the development of six new temperature-based empirical models for solar radiation estimation in humid regions. For the humid subtropical regions of China, the accuracy and suitability of the models were further evaluated as a case study using meteorological data from 20 radiation stations during 1966–2015 suggested that the accuracy of single temperature-based models was greatly improved when daily precipitation and relative humidity were included in the models. All the new models, whether single or complex temperature-based, have shown better results for the prediction of global solar radiation when applied to humid tropical or subtropical regions of China.

In view that solar radiation plays an important role in energy balance, energy applications, and climate change, an adaptive nonlinear empirical neuro-fuzzy inference system (ANFIS) with input parameters daily sunshine duration, precipitation, relative humidity, air pressure, and the daily temperature was proposed to predict daily solar irradiance in China. The results indicated that the model is superior to two other comparing models, the Bristow–Campbell and Improved Yang Hybrid with RMSE and mean absolute error (MAE) ranged from 0.59 to 1.60 MJm^{-2} day^{-1} and 0.42–1.21 MJm^{-2} day^{-1} respectively (Yadav and Chandel 2014).

Another case study using artificial neural network (ANN) and support vector machine (SVM) was proposed to forecast solar radiation of a tilted surface in Saudi Arabia (Ramli et al. 2015). The optimum solar radiation value was achieved with a tilt angle of 16° and 37.5°, respectively, for locations in Jeddah and Qasim. SVM outperformed ANN at both locations, with correlation coefficient (CC) between 0.918 and 0.967 for training and for the testing was in the range of 0.91981–0.97641 while for the training of ANN is in the range of 0.517–0.9692 and for the testing is 0.0361–0.0961 at Jeddah. The prediction result at Qassim gave a CC of 0.999 for training and 0.987 for testing. Results that were obtained while training and testing ANN at Qassim were poor.

A hybrid support-vector machine-wavelet transform approach for estimation of daily and monthly horizontal global solar radiation for an Iranian coastal city demonstrated good performance of coefficient of determination, $R^2 = 0.9086$ and 0.9742, respectively (Mohammadi et al. 2015).

Three separate sets of climatological parameters have been used as inputs for developing three models, and the results suggested the model utilizing relative sunshine period, variations between air temperatures, relative humidity, atmospheric temperature, and extraterrestrial solar radiation as inputs showed good output compared with other models (Olatomiwa et al. 2015a). The significance of extraterrestrial solar radiation to enhance the prediction accuracy could not be ignored.

Ji and Chee (2011) proposed an hourly solar radiation prediction model using time-series autoregressive moving average (ARMA) and time delay neural network (TDNN) model. The solar radiation series contain both linear and nonlinear components. ARMA was used to predict the linear component, and TDNN handled the nonlinear component. The result was quite good due to the stability and accuracy of the hybrid model. RMSE values ranging from 0.0231 to 0.0459 were obtained when the model was applied to a dataset detrended by four different models, Jain's, Baig's, S. Kaplanis', and Al-Sadah's models.

Sharafati et al. (2019) investigated the ability of four data-mining computer models to predict daily measured solar radiation at four locations in Burkina Faso, i.e., Bur Dedougou, Bobo-Dioulasso, Fada-Ngourma, and Ouahigouya, namely, random forest (RF), random tree, reduced error pruning trees, and a hybrid model of random committee with random tree reduction (RC). For the prediction of solar radiation, they used regular data from seven climatic variables, namely, maximum and minimum air temperature, maximum and minimum relative humidity, wind speed, evaporation, and vapor pressure deficit, for the 1998–2012 season. According to the correlation coefficient between the predictors and the predictand, various combinations of input variables were used, and the best input combination was chosen based on the sensitivity of the model output calculated in terms of the statistical indices. For all meteorological stations, the findings of their research were found to be reliable. When all the climate variables are used as data, the highest accuracy in prediction has been found. The minimum absolute error in prediction was shown by the RC and RF at all the stations. In the range of 0.03–0.05 and 0.77–0.91 for RC and 0.03–0.05 and 0.78–0.92 for RF at various stations, the RMSE and NSE are found. The results show that the data mining models proposed can predict solar radiation over Burkina Faso reliably. A hybrid model using firefly and random forests were proposed to predict hourly global solar radiation (Ibrahim and Khatib 2017). However, hybridization has some limitations, such as high computational time complexity and slow convergence speed (Wang and Liu 2019).

Ghimire et al. (2019a) proposed a study to review, build, and evaluate a suite of artificial neural network (ANN)-based machine learning (ML) models versus several other types of data-driven models such as support vector regression (SVR), Gaussian process machine learning (GPML), and genetic programming (GP) models generated by the European Centre for Medium Ran Ranking for the prediction of daily I_{rad}. In their research, to train these models for 5 solar-rich metropolitan sites (i.e., Brisbane, Gold Coast, Sunshine Coast, Ipswich, and Toowoomba, Australia), 87 different predictor variables from the ERA-Interim reanalysis dataset (01 January 1979 to 31 December 2015) were extracted. According to their results, the performance of ANN was significantly better than the other models (SVR, GPML, GP).

Even though during the last decade, ANN models contribute significantly to the estimation of solar radiation, the time series model is still popular and applied on its own or coupled with ANN. Huang et al. (2013) used a combined auto-regressive and dynamic system (CARDS) model to forecast solar radiation on an hourly time scale. The model could predict solar radiation an hour ahead of when climatic conditions change significantly with clouds covering the sun. At present, the use of solar radiation values and the one-as a correction to a forecast value increased the predictive accuracy by 30% relative to models without this adjustment. The CARDS model gave normalized root mean square error (NRMSE) of 16.5% for all days and is favorably compared with a similar model from the literature that had 16–17% and 32% for mostly clear and cloudy days, respectively.

A benchmarking of machine learning techniques composed of NN, Gaussian processes (GP), and SVM for intraday solar forecasting was proposed against simple models such as AR and scale persistence, reference model (Lauret et al. 2015).

The performance of the model was assessed on the historical Global Horizontal Solar Irradiance (GHI) data measured at three French islands. The machine learning techniques outperformed the comparing models for forecasting horizons greater than 1 h. For an hour ahead of solar forecasting, the sky conditions play a significant role whereby the nonlinear methods slightly improve the scale persistence for stable, clear sky conditions.

Under unstable sky conditions, the discrepancy between machine learning methods and a simple model is more prevalent, with a 2% relative root mean square error (rRMSE) difference on average (Lauret et al. 2015).

The above literature studies indicated both conventional and artificial neural network models have different abilities in the prediction of solar radiation and very much dependent on the input parameters and the quality of data (Huang et al. 2013; Lauret et al. 2015).

In this study, three models, boosted decision tree regression (BDTR), neural network regression with three different normalizers, and linear regression, have been proposed and investigated using historical solar radiation with various months of years as the input parameters. The prediction accuracy of the models was evaluated and tabled out based on the following: correlation coefficient (R), coefficient of determination (R^2), root mean square error (RMSE), relative absolute error (RAE), and relative square error (RSE). The study area and hydrological data with the proposed models are briefly listed in “Methodology.” “Results and discussion” displays the results of the proposed models and their comparisons. The conclusion of the study is explained in “Conclusion.”

Methodology

Data

The area of investigation is in Kuala Terengganu, Malaysia. This study will be using historical solar data only to predict solar radiation at any desired time possible. Raw solar data were obtained from the Department of Meteorology Malaysia (MMD). The data used in this study measured hourly from 7 a.m. to 6 p.m. Therefore, the selected months of data used composed of March and April 2008; January, February, March, and April 2009; and April 2010 (Table 1).

Data pre-processing

The data will undergo a pre-processing stage whereby clean missing data, normalization, and filter-based feature selection module were applied. There were large missing data; hence, probabilistic PCA cleaning mode was applied. Cleaned missing data of each feature have distinctive value ranges; thus, normalization is essential to alter the values of numeric columns within the dataset to a common scale without disfiguring contrast within the ranges of values. Min-max transformation method is used whereby the min-max normalizer linearly rescales every feature to the [0, 1] interval. The values in each column are transformed by using the equation as follows:

$$z = \frac{x - \min(x)}{[\max(x) - \min(x)]} \tag{1}$$

x is the original number, and min and max value of x in order to compute min-max transformation.

The final step of data pre-processing is the filter-based feature selection module. This step is imperative in performing a machine learning algorithm model as it helps to identify the columns in the input dataset that have the greatest predictive power toward solar radiation. This study is using the filter selection metric of Pearson’s correlation. Pearson’s

Table 1 Statistics of raw solar data.

Solar statistics	Month and Year						
	Mar 08	Apr 08	Jan 09	Feb 09	Mar 09	Apr 09	Apr 10
Data count	60	338	233	277	242	174	345
Maximum (MJ/m^2)	1226	1323	1173	1234	1260	1273	1291
Minimum (MJ/m^2)	108	99	98	98	98	98	101
Mean	584	656	507	613	602	593	673
Mean deviation	278	333	303	304	340	299	325
Median	607	651	399	575	553	557	689
Sample variance	107,821	141,755	117,939	120,621	145,163	118,389	133,284
Standard deviation	328	377	343	347	381	344	365

correlation between independent variables and dependent variables was done by using Eq. 2:

$$r_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}} \quad (2)$$

r_{xy} is the correlation function, n is the sample size of data, x_i and y_i are the sample points, and $\overline{x}$ and $\overline{y}$ are the sample mean, respectively.

Figure 1 illustrates the prediction methodology of solar radiation using machine learning algorithms.

Machine learning algorithms modeling

Two data partitions were attempted in this study at 80% and 75% training to compare which data partition relative to the machine learning algorithm performs better. Eighty percent of randomly selected independent variables data will go through intensive training using a machine learning algorithm. By

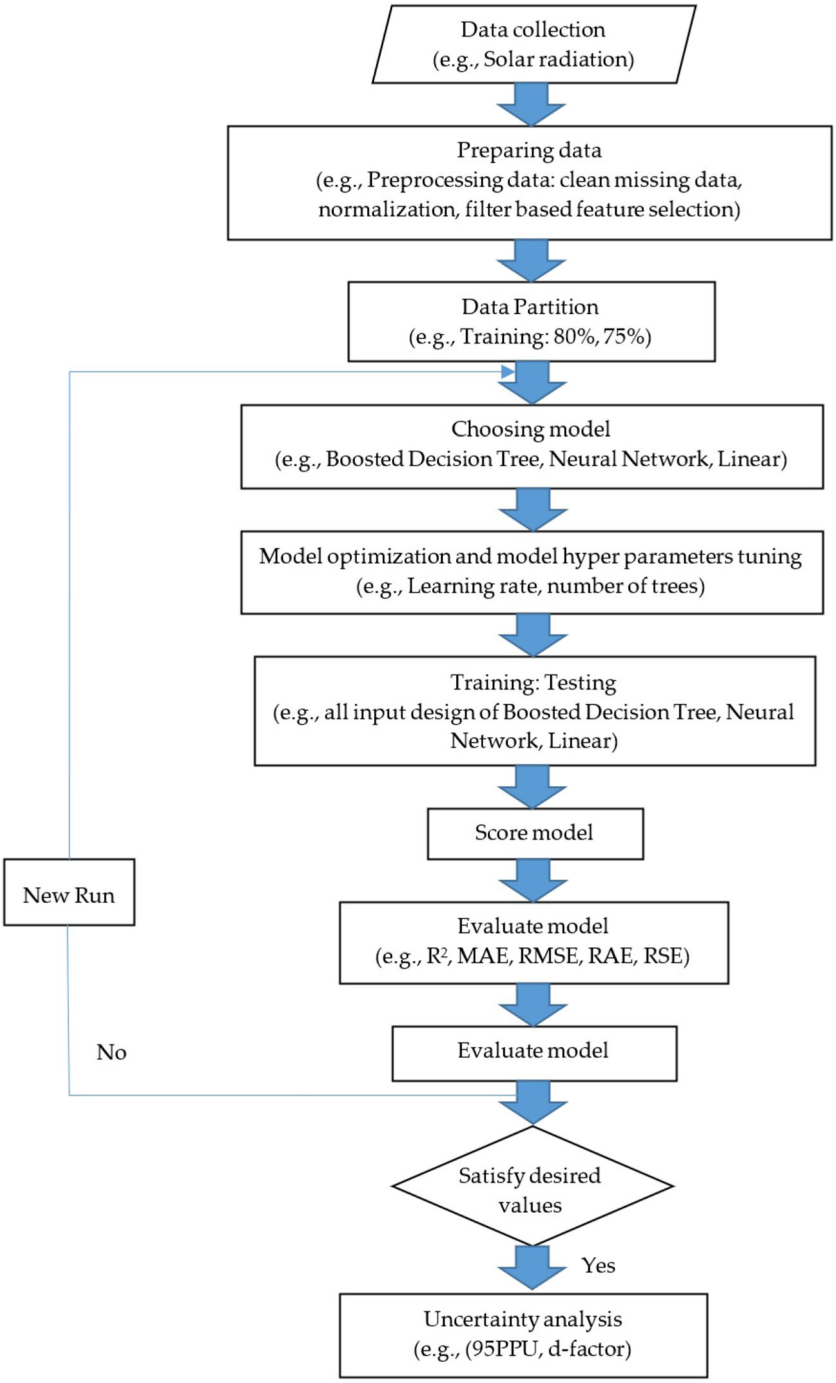

Data collection
(e.g., Solar radiation)
Preparing data
(e.g., Preprocessing data: clean missing data, normalization, filter based feature selection)
Data Partition
(e.g., Training: 80%, 75%)
Choosing model
(e.g., Boosted Decision Tree, Neural Network, Linear)
Model optimization and model hyper parameters tuning
(e.g., Learning rate, number of trees)
Training: Testing
(e.g., all input design of Boosted Decision Tree, Neural Network, Linear)
Score model
Evaluate model
(e.g., R2, MAE, RMSE, RAE, RSE)
Evaluate model
New Run
No
Satisfy desired values
Yes
Uncertainty analysis
(e.g., (95PPU, d-factor)

Fig. 1 Prediction methodology of solar radiation using machine learning algorithms

using a trained dataset, then the remaining untrained data will be used to test the model performance. The same process is applied to the second data partition. The machine learning algorithms used in this study are boosted decision tree regression, neural network regression, and linear regression.

These training datasets undergo two different approaches. The first approach is a conventional way whereby the model is optimized by manually adjusting the learning rate or a number of trees of the algorithms. The second approach is by introducing the tune model hyperparameter module to the model. Tune model hyperparameter determines the optimum hyperparameter for a given machine learning algorithm through different combinations and settings of multiple models and compares the metric to get the best combination of settings. The tune model hyperparameter is used to aid the model performances.

Boosted regression tree (BRT) models are a combination of two techniques, which are decision tree algorithms and boosting methods. BRTs repeatedly fit many decision trees to improve the accuracy of the model. While boosted decision tree regression is an algorithm used to train the model by implementing the MART gradient boosting algorithm. Boosting builds a series of trees in a stage-wise fashion, and each tree is dependent on prior trees. Therefore, each error on the prior tree is measured by using a predefined loss function and correct it in the next tree. This infers that the prediction is an ensemble of a group of weaker prediction models and formed a robust prediction model. The boosted decision tree regression algorithm is as follows:

$$\hat{y}(x) = \sum_t w_t h_t(x) \tag{3}$$

$$O(x) = \sum_i l\left(\hat{y}_i, y_i + \sum_t \Omega(f_t)\right. \tag{4}$$

where $h(x)$ is the tree's output, w is the weight; $l(\hat{y}_i\ , y_i)$ is the loss function, distance between the truth, and the prediction in ith sample; and $\Omega(f_t)$ is the regularization function. Figure 2 shows the structure of the Boosted Regression Tree model.

Neural network regression is used in classification and regression problems (Ehteram et al. 2020). Generally, it consists of three arranged layers; input layer, hidden layer(s), and output layer (Dashti Latif et al. 2020). The hidden layer will transform feed input data from the input layer into high dimensional space, and each neuron in the hidden layer applies radial function. All hidden neurons are connected to the output neurons by regulating output weights at the last layer of the output layer.

Linear regression will be the last machine learning algorithm performed. It shows a linear relationship between one or more independent variables and a dependent variable outcome. The algorithm works in a way as follows:

$$y = \alpha + \beta x \tag{5}$$

β denotes the slope of the line, and α is the y-intercept of the linear relationship between regression γ dependent variable and x independent variable.

Two different normalizers have been adopted in this study for standardizing the dataset, namely, Gaussian normalizer binning normalizer.

Gaussian normalization technique is proposed to normalize the data to have a mean equal to 0 and variance equal to 1. While binning normalization is used to scale the observed data to a range between 0 and 1 by grouping the data into classes (bins) with equal size and then normalize each value by dividing the index value for the class by the total number of classes. Therefore, in this study, both techniques will be investigated to compare the effect on the accuracy of machine learning algorithms.

Data scoring is relative to each of the machine learning algorithms that will be compared, and the best model is chosen according to the performance indices used in this studies that are coefficient of determination (R^2), mean absolute error (MAE), root mean square error (RMSE), relative absolute error (RAE) and relative square error (RSE). The performance indices used to evaluate the scored model on how much close the computed solar radiation to the real values is as follows:

1. Coefficient of determination (R^2):

$$R^2 = \frac{\sum_{i=1}^{n}\left(y_i - \overline{Y}_i\right)^2 - \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}_i\right)^2} \tag{6}$$

The higher the R^2 value indicates good model performance.

2. Mean absolute error (MAE):

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{7}$$

MAE measures the accuracy of continuous variables.

3. Root mean square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{8}$$

y_i and $\hat{y}_i$ are the observation and prediction in the ith step. RMSE gives big errors a fairly high weight.

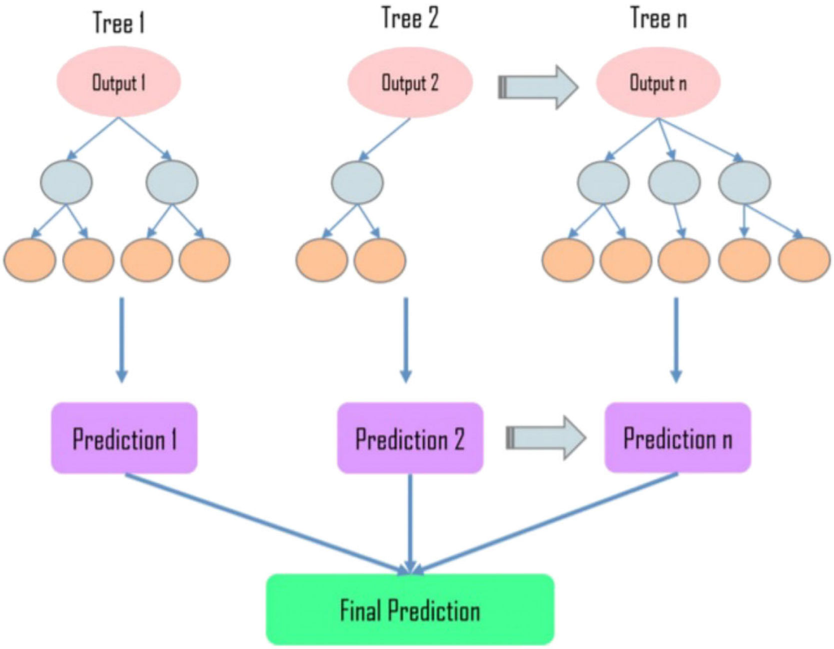


Fig. 2 The structure of typical boosted tree regression (Lai et al. 2019)

Both RMSE and MAE were used to measure the residual error and indicate the unit error of output. Both metrics can range from 0 to ∞, and the lower the values are better.

4. Relative absolute error (RAE):

$$\mathrm{RAE} = \frac{\sum_{i=1}^{n}\left|\widehat{y}_i - y_i\right|}{\sum_{i=1}^{n}\left[\overline{y} - y_i\right]^2} \tag{9}$$

RAE is a normalized value by dividing the total absolute error by a simple predictor total absolute error. A good forecast model will yield a ratio close to 0 whilst a weak model will yield a ratio greater than 1.

5. Relative square error (RSE):

$$\mathrm{RSE} = \frac{\sum_{i=1}^{n}\left(\widehat{y}_i - y_i\right)^2}{\sum_{i=1}^{n}\left(\overline{y} - y_i\right)^2} \tag{10}$$

RSE is a normalized value by dividing the total square error by the simple predictor total square error.

Finally, once the best model is chosen, it will be evaluated for uncertainties test to determine whether the model can be used at different location environments; hence, 95PPU and d-factor are used. The test of uncertainty aims to estimate the variation in output due to the input variability. It is done to identify the range of possible results based on input uncertainty and to analyze the effect of the lack of information or errors of the model (Noori et al. 2010). The model is reliable to use at any location if the values bracketed by 95PPU fall into the band range 95PPu (between 80 and 100% of observed data), and the d-factor value is lesser than 1 as the best d-factor is 0.

6. Bracketed by 95PPU:

$$\text{Bracketed by 95PPU} = \tfrac{1}{k}\text{count}(K|X_L \le K \le X_U) \times 100 \tag{11}$$

K is the number of observed data at the testing stages. By referring to Eq. (11), the value of bracketed by 95PPU is optimum or 100% if all the measured data are placed between X_L and X_U. Percentage of measured data obtained by 2.5% of X_L and 97.5% of X_U.

7. d-factor

$$d\text{−}factor = \frac{\overline{d}_x}{\sigma_x} \tag{12}$$

d-factor measured the average width of the confidence interval band. σ_x is the standard deviation of observed data x, and $\overline{d}_x$ is the average distance between upper and lower bands computed as follows:

8. $\overline{d}_x$:

$$\overline{d}_x = \frac{1}{k}\sum_{i=1}^{k}(X_U - X_L) \tag{13}$$

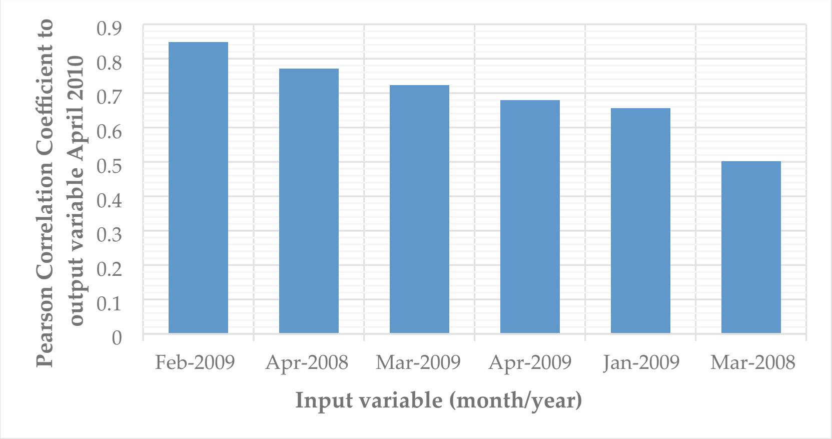


Fig. 3 Bar chart of Pearson's correlation coefficient of input variables relative to the output variable

Results and discussion

The capability of different machine learning algorithm models was explored for average hourly solar radiation prediction using only historical solar radiation data. Two different data partitions were applied for better evaluation of the methods.

Correlation coefficient

For the purpose of these studies, the prediction of average hourly solar radiation in April 2010 was focused. Figure 3 shows the correlation between the input variables and the output variable. It can be seen that February 2009 has the highest correlation of 0.85 coefficient, and March 2008 is the lowest correlated with 0.50 correlation, in relation to the output variable of April 2010. Due to the limitation in the data availability, these five parameters have been selected to predict the solar radiation changes during the month of April 2010.

Model performance

The first approach of machine learning modeling was by using a conventional way without the module of tune model hyperparameter. Table 2 (a) and (b) show the performance of the model after the dataset was performed for 80–20% and 75–25% data splitting, respectively. Based on Table 2 (a), the R^2 for each model from highest to lowest values are BDTR 0.89125, LR 0.82789, NNBN 0.76327, and NNGN 070640. Though, LR is overfitted, as R^2 of the test (20%) dataset is 0.82789 higher than the train (80%) dataset, which is 0.81683. For 75–25% data splitting in Table 2 (b), the R^2 values in descending order are BDTR 0.90183, LR 0.84529, NNGN 0.80527 and NNBN 0.79877. However, LR and NNGN are overfitted since the R^2 values in the test (25%) dataset are higher than the train (75%) dataset. By comparing both data splitting, BDTR outperformed the other models, and with 75–25% data split has higher R^2 compared with 80–20%. In addition, the BTDR model also has MAE 0.06625, RSME 0.08551, RAE 0.27746, and RSE 0.09817, which are

Table 2 Performance indices for testing dataset without tune model hyperparameter (a) 80–20% and (b) 75–25% data splitting

Dataset	Train (80%)					Test (20%)				
Algorithm	R^2	MAE	RMSE	RAE	RSE	R^2	MAE	RMSE	RAE	RSE
(a) 80–20% data splitting										
Neural network regression: Gaussian normalizer	0.76603	0.10807	0.14265	0.40623	0.23397	0.70640	0.11217	0.14089	0.50263	0.29360
Neural network regression: binning normalizer	0.81802	0.09395	0.12581	0.35314	0.18198	0.76327	0.09771	0.12652	0.43781	0.23673
Boosted decision tree regression	0.99956	0.00340	0.00616	0.01277	0.00044	0.89125	0.06691	0.08575	0.29980	0.10875
Linear regression	0.81683	0.09452	0.12622	0.35529	0.18318	0.82789	0.08570	0.10787	0.38401	0.17211
(b) 75%-25% data splitting										
Neural network regression: Gaussian normalizer	0.79294	0.09995	0.13305	0.37991	0.20706	0.80527	0.09227	0.12043	0.38645	0.19473
Neural network regression: binning normalizer	0.81217	0.09454	0.12672	0.35935	0.18784	0.79877	0.09305	0.12242	0.38969	0.20123
Boosted decision tree regression	0.99953	0.00320	0.00634	0.01216	0.00047	0.90183	0.06625	0.08551	0.27746	0.09817
Linear regression	0.80977	0.09575	0.12753	0.36395	0.19023	0.84529	0.08405	0.10734	0.35201	0.15471

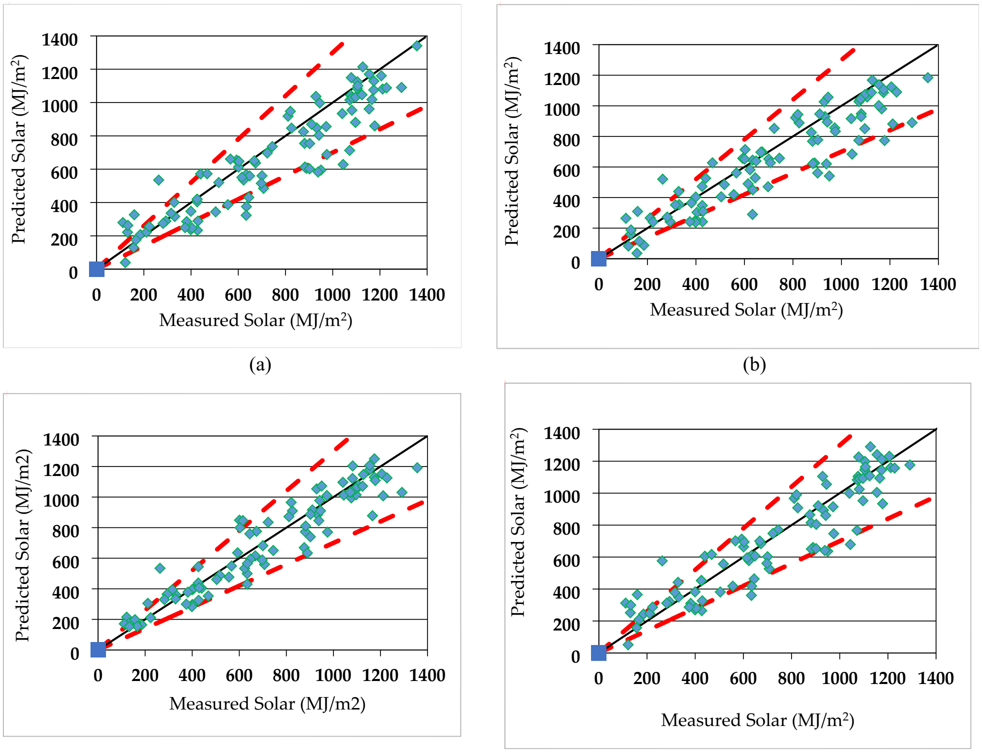
Table 3 Performance indices for testing dataset with tune model hyperparameter (a) 80–20% and (b) 75–25% data splitting

Dataset	Train (80%)					Test (20%)				
Algorithm	R^2	MAE	RMSE	RAE	RSE	R^2	MAE	RMSE	RAE	RSE
(a) 80–20% data splitting										
Neural network regression: Gaussian normalizer	0.74337	0.11502	0.14940	0.43235	0.25663	0.66410	0.12169	0.15070	0.54528	0.33590
Neural network regression: binning normalizer	0.80619	0.09828	0.12983	0.36943	0.19381	0.73774	0.10607	0.13316	0.47526	0.26227
Boosted decision tree regression	0.98992	0.02377	0.02961	0.08934	0.01008	0.86691	0.06927	0.09486	0.31037	0.13309
Linear regression	0.81683	0.09452	0.12622	0.35529	0.18318	0.82789	0.08570	0.10787	0.38401	0.17211
(b) 75–25% data splitting										
Neural network regression: Gaussian normalizer	0.73392	0.11728	0.15082	0.44577	0.26608	0.70940	0.11834	0.14712	0.49564	0.29061
Neural network regression: binning normalizer	0.78790	0.10389	0.13466	0.39489	0.21210	0.75445	0.10643	0.13523	0.44575	0.24555
Boosted decision tree regression	0.99056	0.02298	0.02841	0.08735	0.00944	0.88277	0.06724	0.09344	0.28162	0.11723
Linear regression	0.80977	0.09575	0.12753	0.36395	0.19023	0.84529	0.08405	0.10734	0.35201	0.15471

relatively low compared with other models and close to 0 that indicate a better model.

The second approach was by implementing the tune model hyperparameter module to the models. Table 3 (a) and (b) show the performance of the model after the dataset was performed for 80–20% and 75–25% data splitting, respectively. Based on Table 3 (a), the R^2 for each model in descending order is BDTR 0.86691, LR 0.82789, NNBN 0.73774, and NNGN 0.66410. As shown in Tables 3 and 2, LR shows no changes in R^2 value regardless of variations in data splitting applied and is overfitted. This implies that LR is not a suitable model used to predict solar radiation. For Table 3 (b), the descending order of R^2 values is BDTR 0.88277, LR 0.84529, NNBN 0.75445, and NNGN 0.70940. Again, by implementing the tuned model hyperparameter module to the models, BDTR outperformed the other models, and BDTR with 75–25% data split has higher R^2 compared with 80–20% data split. In Table 3, although the R^2 of all models was slightly lower than in Table 2, most of the models were performed well and stable without overfitting except for LR. This infers that the tune model hyperparameter module significantly helps in stabilizing the models' performance by giving aids in optimizing the models largely.

Fig. 4 Scatter plot of predicted versus measured solar for test (25%) dataset without tune model hyperparameter **a** NNGN, **b** NNBN, **c** BDTR, and **d** LR



(a) (b) (c) (d)

It can be concluded that in both scenarios, without implementing the tuning technique and with it, the most suitable model that can be used to predict the solar radiation is BDTR with train (75%) and test (25%) data splitting. However, BDTR without tuning outperformed the proposed model with tuning. The tuning technique used in this study is a random search method, which depends on choosing values for the hyperparameters without checking the previous training results, which can lead to miss the optimal values of the hyperparameters.

For more visual comparison, Fig. 4 shows the scatter plot of predicted versus measured solar (a) NNGN, (b) NNBN, (c) BDTR, and (d) LR, respectively, for Table 2 (b). It can be seen clearly that the proposed BDTR algorithm outperformed all algorithms in mimicking the changes in solar radiation. In addition to that, it approximated the actual observation with an acceptable level of accuracy.

Taylor diagram

Taylor's diagram represents a brief statistical summary of how fit patterns match their correlation and standard deviation. Taylor diagram formula is as follow:

$$R = \frac{\frac{1}{N}\sum_{n=1}^{N}\left(f_n-\overline{f}\right)\left(r_n-\overline{r}\right)}{\sigma_f\sigma_r} \qquad (13)$$

Table 4 Uncertainty analysis of 75–25% data splitting models for the testing dataset

Model	Statistic	
	U95PPU	d-factor
NNBN	87.4749	0.03579
BDTR	97.2539	0.03731

R is a correlation, N is the number of discrete points, f_n and r_n are two variables, σ_f and σ_r are the standard deviation of f and r, and $\overline{f}$and $\overline{r}$ are the mean values of σ_f and σ_r.

Figure 5 illustrates the relationship between standard deviation and correlation of predicted solar and measured solar for all models from Table 2 (b). BDTR prediction is highly correlated with the actual value, and the standard deviation is closest to the actual standard deviation compared with other models. This proved that with an R^2 of 0.90183, the BDTR is the most reliable model for solar prediction, among other models. The standard deviation of LR is closer to the actual value; however, it has a lower correlation with the actual value compared with NNBN. Meanwhile, neural network Gaussian normalizer is the least correlated and farther from the actual standard deviation. However, LR and NNGN are overfitted, as seen in Table 2 (b).

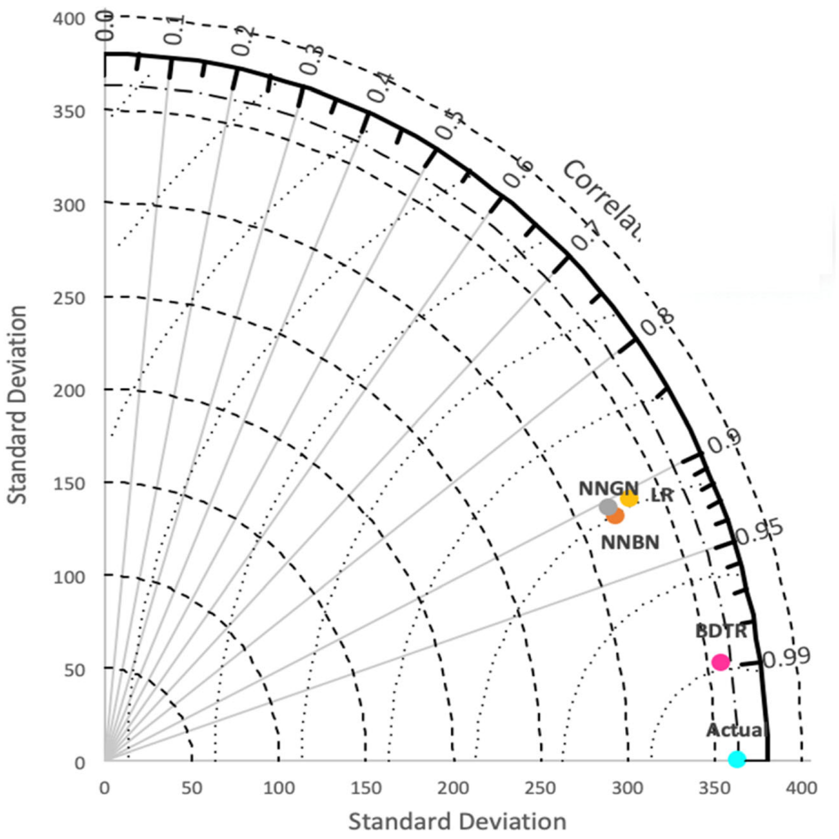


Fig. 5 Taylor diagram of correlation and standard deviation of 75–25% data splitting without tune model hyperparameter

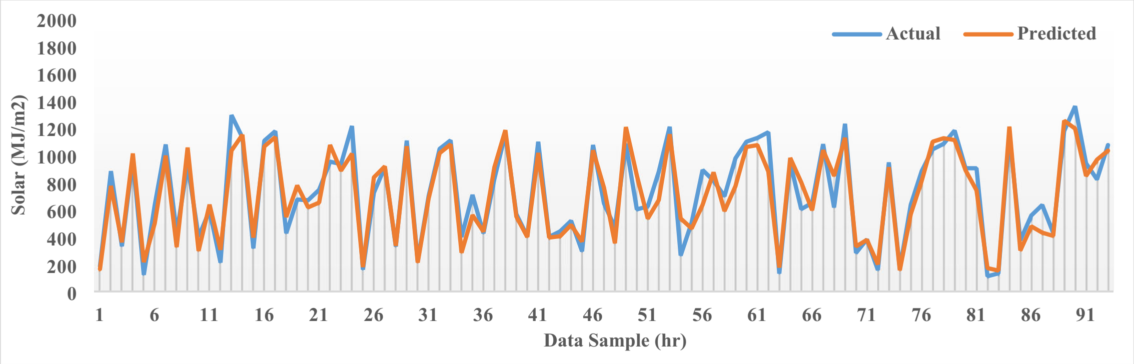


Fig. 6 Actual vs. predicted solar radiation using BDTR during the testing phase

Uncertainty analysis

Finally, the uncertainty analysis was calculated for the 75–25% data splitting without the tune model hyperparameter module of chosen models, which were NNBN and BDTR. Two different criteria were used for this purpose, namely, bracketed by 95PPU and d-factor. Uncertainty analysis is usually used to check the performance of the proposed model when there is a new dataset of the input that can be introduced. The best outcome of bracketed by 95PPU should fall in the band range of (80–100%) whilst the d-factor is 0. Table 4 shows the uncertainty analysis results for the testing dataset.

The values of bracketed by 95PPU are 87.5% and 97.3% of data for each model NNBN and BDTR relate to 75–25% data splitting. Based on these obtained values for both models, it can be concluded that all the observed data fall into the 95PPU band range of between 80 and 100% observed data. In addition, the d-factor values of 0.03579 and 0.03731 for NNBN and BDTR, respectively, were less than one, which is desirable since the best value for the d-factor is 0. Finally, Fig. 6 depicts the performance of the proposed model in predicting the actual observation of the solar radiation during the testing phase.

Conclusion

The capability of various models to predict solar radiation was assessed based on the available historical data of solar radiation itself as the input variables. Four prediction models were studied, composed of neural network Gaussian normalizer, neural network binning normalizer, boosted decision tree regression, and linear regression. By using two different data splitting, which was 80–20% and 75–25% data split, boosted decision tree regression outperformed all the other models with R^2 of 0.89125 and 0.90183, respectively, without implementing tune model hyperparameter module. Whilst by implementing the tune model hyperparameter module, the performance of boosted decision tree regression somehow decreased to 0.86691 and 0.88277 of R^2 for each 80–20% and 75–25% data splitting. This infers that data splitting of 75–25% gives better performance toward boosted decision tree regression by omitting the implementation of the tune model hyperparameter module. Detailed observation paid to this matter; only boosted decision tree regression and neural network binning normalizer models can be used, as the rest of the models were overfitted. The reliability of both models was calculated by uncertainty analysis known as 95PPU and d-factor. Based on the values of 95PPU and d-factor, it is concluded that both of these models have an acceptable low degree of uncertainty. In this study, only historical solar data composed of different months and years were used to predict solar radiation in April 2010 and are parsimonious enough to produce a good prediction model. The performance of the proposed model can be improved if more data incorporated, such as recent solar radiation and weather data at various meteorological stations, which were not available during this study. In addition to that, a high level of accuracy could be achieved if the proposed model augments with optimizers. On the other hand, the proposed model may be applied in other areas for solar radiation prediction.

Acknowledgments The author would like to thank the Department of Meteorology Malaysia (MMD) for providing us with the data.

Author contributions Data curation: Faridah Bte Basaruddin; Formal analysis: Yuzainee Bte. Md Yusoff; Methodology: Ali Najah Ahmed; Writing—original draft: Ellysia Jumin; Writing—review and editing: Sarmad Dashti Latif.

Funding This research was funded by Innovation & Research Management Center (iRMC), Universiti Tenaga Nasional, Malaysia, YCU grant coded:201901001YCU/23.

Data availability Not applicable

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval Not applicable

Consent to participate Not applicable

Consent to publish Not applicable

References

Alsina EF, Bortolini M, Gamberi M, Regattieri A (2016) Artificial neural network optimisation for monthly average daily global solar radiation prediction. Energy Convers Manag 120:320–329. https://doi.org/10.1016/j.enconman.2016.04.101

Aybar-Ruiz A, Jiménez-Fernández S, Cornejo-Bueno L, Casanova-Mateo C, Sanz-Justo J, Salvador-González P, Salcedo-Sanz S (2016) A novel grouping genetic algorithm-extreme learning machine approach for global solar radiation prediction from numerical weather models inputs. Sol Energy 132:129–142. https://doi.org/10.1016/j.solener.2016.03.015

Bilgili M, Ozgoren M (2011) Daily total global solar radiation modeling from several meteorological data. Meteorol Atmos Phys 112:125–138. https://doi.org/10.1007/s00703-011-0137-9

Chen JL, Li GS, Wu SJ (2013) Assessing the potential of support vector machine for estimating daily solar radiation using sunshine duration. Energy Convers Manag 75:311–318. https://doi.org/10.1016/j.enconman.2013.06.034

Chia MY, Huang YF, Koo CH, Fung KF (2020) Recent advances in evapotranspiration estimation using artificial intelligence approaches with a focus on hybridization techniques—a review. Agronomy

Dashti Latif S, Najah Ahmed A, Sherif M, Sefelnasr A, el-Shafie A (2020) Reservoir water balance simulation model utilizing machine learning algorithm. Alexandria Eng J. 60:1365–1378. https://doi.org/10.1016/j.aej.2020.10.057

Deo RC, Wen X, Qi F (2016) A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. Appl Energy 168:568–593. https://doi.org/10.1016/j.apenergy.2016.01.130

Ehteram M, Ahmed AN, Latif SD, Huang YF, Alizamir M, Kisi O, Mert C, el-Shafie A (2020) Design of a hybrid ANN multi-objective whale algorithm for suspended sediment load prediction. Environ Sci Pollut Res. 28:1596–1611. https://doi.org/10.1007/s11356-020-10421-y

Falayi EO, Adepitan JO, Rabiu AB (2008) Empirical models for the correlation of global solar radiation with meteorological data for Iseyin, Nigeria. Int J Phys Sci 3:210–216

Fan J, Chen B, Wu L, Zhang F, Lu X, Xiang Y (2018) Evaluation and development of temperature-based empirical models for estimating daily global solar radiation in humid regions. Energy 144:903–914. https://doi.org/10.1016/j.energy.2017.12.091

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 Manag. 198:111780. https://doi.org/10.1016/j.enconman.2019.111780

Ghazvinian H, Mousavi SF, Karami H, Farzin S, Ehteram M, Hossain MS, Fai CM, Hashim HB, Singh VP, Ros FC, Ahmed AN, Afan HA, Lai SH, el-Shafie A (2019) Integrated support vector regression and an improved particle swarm optimization-based model for solar radiation prediction. PLoS One. 14:e0217634. https://doi.org/10.1371/journal.pone.0217634

Ghimire S, Deo RC, Downs NJ, Raj N (2019a) Global solar radiation prediction by ANN integrated with European Centre for medium range weather forecast fields in solar rich cities of Queensland Australia. J Clean Prod. 216:288–310. https://doi.org/10.1016/j.jclepro.2019.01.158

Ghimire S, Deo RC, Raj N, Mi J (2019b) Deep learning neural networks trained with MODIS satellite-derived predictors for long-term global solar radiation prediction. Energies. 12. https://doi.org/10.3390/en12122407

Huang J, Korolkiewicz M, Agrawal M, Boland J (2013) Forecasting solar radiation on an hourly time scale using a Coupled AutoRegressive and Dynamical System (CARDS) model. Sol Energy 87:136–149. https://doi.org/10.1016/j.solener.2012.10.012

Ibrahim IA, Khatib T (2017) A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm. Energy Convers Manag 138:413–425

Ji W, Chee KC (2011) Prediction of hourly solar radiation using a novel hybrid model of ARMA and TDNN. Sol Energy 85:808–817. https://doi.org/10.1016/j.solener.2011.01.013

Kisi O (2014) Modeling solar radiation of Mediterranean region in Turkey by using fuzzy genetic approach. Energy 64:429–436. https://doi.org/10.1016/j.energy.2013.10.009

Kitani, O., Jungbluth, T., Peart, R. M., Ramdani, A (1999) CIGR Handbook of Agricultural Engineering (Energy and Biomass Engineering)

Lai V, Ahmed AN, Malek MA, Abdulmohsin Afan H, Ibrahim RK, el-Shafie A, el-Shafie A (2019) Modeling the nonlinearity of sea level oscillations in the Malaysian coastal areas using machine learning algorithms. Sustain. 11. https://doi.org/10.3390/su11174643

Lauret P, Voyant C, Soubdhan T, David M, Poggi P (2015) A benchmarking of machine learning techniques for solar radiation forecasting in an insular context. Sol Energy 112:446–457. https://doi.org/10.1016/j.solener.2014.12.014

Mohammadi K, Shamshirband S, Tong CW, Arif M, Petković D, Ch S (2015) A new hybrid support vector machine-wavelet transform approach for estimation of horizontal global solar radiation. Energy Convers Manag 92:162–171. https://doi.org/10.1016/j.enconman.2014.12.050

Noori R, Hoshyaripour G, Ashrafi K, Araabi BN (2010) Uncertainty analysis of developed ANN and ANFIS models in prediction of carbon monoxide daily concentration. Atmos Environ 44:476–482. https://doi.org/10.1016/j.atmosenv.2009.11.005

Olatomiwa L, Mekhilef S, Shamshirband S, Mohammadi K, Petković D, Sudheer C (2015a) A support vector machine-firefly algorithm-based model for global solar radiation prediction. Sol Energy 115:632–644. https://doi.org/10.1016/j.solener.2015.03.015

Olatomiwa L, Mekhilef S, Shamshirband S, Petković D (2015b) Adaptive neuro-fuzzy approach for solar radiation prediction in Nigeria. Renew Sustain Energy Rev 51:1784–1791. https://doi.org/10.1016/j.rser.2015.05.068

Qazi A, Fayaz H, Wadi A, Raj RG, Rahim NA, Khan WA (2015) The artificial neural network for solar radiation prediction and designing solar systems: a systematic literature review. J Clean Prod 104:1–12. https://doi.org/10.1016/j.jclepro.2015.04.041

Rabehi A, Guermoui M, Lalmi D (2020) Hybrid models for global solar radiation prediction: a case study. Int J Ambient Energy. 41:31–40. https://doi.org/10.1080/01430750.2018.1443498

Ramedani Z, Omid M, Keyhani A, Khoshnevisan B, Saboohi H (2014) A comparative study between fuzzy linear regression and support vector regression for global solar radiation prediction in Iran. Sol Energy 109:135–143. https://doi.org/10.1016/j.solener.2014.08.023

Ramli MAM, Twaha S, Al-Turki YA (2015) Investigating the performance of support vector machine and artificial neural networks in predicting solar radiation on a tilted surface: Saudi Arabia case study. Energy Convers Manag 105:442–452. https://doi.org/10.1016/j.enconman.2015.07.083

Sharafati A, Khosravi K, Khosravinia P, Ahmed K, Salman SA, Yaseen ZM, Shahid S (2019) The potential of novel data mining models for global solar radiation prediction. Int J Environ Sci Technol. 16: 7147–7164. https://doi.org/10.1007/s13762-019-02344-0

Wang C, Liu K (2019) A randomly guided firefly algorithm based on elitist strategy and its applications. IEEE Access. https://doi.org/10.1109/ACCESS.2019.2940582

Wu Y, Wang J (2016) A novel hybrid model based on artificial neural networks for solar radiation prediction. Renew Energy 89:268–284. https://doi.org/10.1016/j.renene.2015.11.070

Wu J, Chan CK, Zhang Y, Xiong BY, Zhang QH (2014) Prediction of solar radiation with genetic approach combing multi-model framework. Renew Energy 66:132–139. https://doi.org/10.1016/j.renene.2013.11.064

Yacef R, Benghanem M, Mellit A (2012) Prediction of daily global solar irradiation data using Bayesian neural network: a comparative study. Renew Energy 48:146–154. https://doi.org/10.1016/j.renene.2012.04.036

Yadav AK, Chandel SS (2014) Solar radiation prediction using Artificial Neural Network techniques: a review. Renew Sustain Energy Rev 33:772–781. https://doi.org/10.1016/j.rser.2013.08.055

Zou L, Wang L, Xia L, Lin A, Hu B, Zhu H (2017) Prediction and comparison of solar radiation using improved empirical models and adaptive neuro-fuzzy inference systems. Renew Energy 106: 343–353. https://doi.org/10.1016/j.renene.2017.01.042

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.