Computer Modeling in Engineering & Sciences CMES, vol.123, no.1, pp.49-73, 2020

Simulation of Daily Diffuse Solar Radiation Based on Three Machine Learning Models

Jianhua Dong[1], Lifeng Wu[2], Xiaogang Liu[1,*], Cheng Fan[1], Menghui Leng[3] and Qiliang Yang[1]

Abstract: Solar radiation is an important parameter in the fields of computer modeling, engineering technology and energy development. This paper evaluated the ability of three machine learning models, i.e., Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM) and Multivariate Adaptive Regression Splines (MARS), to estimate the daily diffuse solar radiation (R_d). The regular meteorological data of 1966-2015 at five stations in China were taken as the input parameters (including mean average temperature (T_a), theoretical sunshine duration (N), actual sunshine duration (n), daily average air relative humidity (RH), and extra-terrestrial solar radiation (R_a)). And their estimation accuracies were subjected to comparative analysis. The three models were first trained using meteorological data from 1966 to 2000. Then, the 2001-2015 data was used to test the trained machine learning model. The results show that the XGBoost had better accuracy than the other two models in coefficient of determination (R^2), root mean square error (RMSE), mean bias error (MBE) and normalized root mean square error (NRMSE). The MARS performed better in the training phase than the testing phase, but became less accurate in the testing phase, with the R^2 value falling by 2.7-16.9% on average. By contrast, the R^2 values of SVM and XGBoost increased by 2.9-12.2% and 1.9-14.3%, respectively. Despite trailing slightly behind the SVM at the Beijing station, the XGBoost showed good performance at the rest of the stations in the two phases. In the training phase, the accuracy growth is small but observable. In addition, the XGBoost had a slightly lower RMSE than the SVM, a signal of its edge in stability. Therefore, the three machine learning models can estimate the daily R_d based on local inputs and the XGBoost stands out for its excellent performance and stability.

Keywords: Diffuse solar radiation, extreme gradient boosting, multivariate adaptive regression splines, statistical indices, support vector machine.

[1] Faculty of Agriculture and Food, Kunming University of Science and Technology, Kunming, 650500, China.

[2] School of Hydraulic and Ecological Engineering, Nanchang Institute of Technology, Nanchang, 330099, China.

[3] Jiangxi Key Laboratory of Hydrology-Water Resources and Water Environment, Nanchang Institute of Technology, Nanchang, 330099, China.

[*] Corresponding Author: Xiaogang Liu; Email: liuxiaogangjy@126.com.

Received: 02 November 2019; Accepted: 19 December 2019

Nomenclature

Variables		Abbreviations	
T_a	Average air temperature (°C)	R^2	Coefficient of determination
N	Theoretical sunshine duration (h)	MBE	Mean bias error ($MJ \cdot m^{-2} \cdot d^{-1}$)
n	Actual sunshine duration (h)	RMSE	Root mean square error ($MJ \cdot m^{-2} \cdot d^{-1}$)
R_a	Extra-terrestrial solar radiation ($MJ \cdot m^{-2} \cdot d^{-1}$)	NRMSE	Normalized root mean square error
		R	Correlation coefficient
		SD	Standard deviation ($MJ \cdot m^{-2} \cdot d^{-1}$)
R_d	Diffuse solar radiation ($MJ \cdot m^{-2} \cdot d^{-1}$)	SVM	Support Vector Machine
RH	Daily average air relative humidity (%)	XGBoost	Extreme Gradient Boosting
		MARS	Multivariate Adaptive Regression Splines

1 Introduction

Solar radiation plays a fundamental role in engineering [Rehman and Uzair (2017); Yousuf, Siddiqui and Rehman (2018)], computer modeling [Carballo, Bonilla, Berenguel et al. (2019); Charvát, Klimeš, Pech et al. (2019)] and solar energy utilization [Daus, Yudaev, Taranov et al. (2019); Qu, Hong and Jin (2019); Sarnavi, Nikbakht, Hasanpour et al. (2019)]. Solar energy is a promising renewable source of energy, and considered to be one of the most promising sources of energy. If were used rationally, it could replace traditional energy sources on the earth. With the development of technology, solar energy has been increasingly used in many fields [Velmurugan and Kalaivanan (2015); Wenceslas and Ghislain (2018)]. To make full use of solar energy, it is necessary to collect and process the data on solar radiation in an accurate manner. The total solar radiation reaching the ground can be divided into direct solar radiation and diffuse solar radiation [Boland, Huang and Ridley (2013)]. Computer monitoring and modeling of solar radiation and the application of solar energy in energy-saving buildings [Chen, Ding and Liu (2019); He, Lei, Bi et al. (2016); Li and Ratti (2019)]. Accurate solar radiation can promote the development of intelligent society in the future. Compared with direct radiation, diffuse solar radiation is backed up with fewer data across the globe. Thus, it is of great scientific significance to compute the diffuse solar radiation (R_d).

The R_d is more difficult to measure than regular meteorological data like temperature, rainfall and relative humidity. There is no universal or highly reliable method for R_d measurement. The high cost and sophisticated technology add to the difficulty in R_d

measurement [Jahani, Dinpashoh and Nafchi (2017)], especially in developing countries. All these have led to a severe lacking of R_d data around the world. To solve the problem, many approaches have been developed to estimate the R_d, ranging from empirical model [Fan, Wu, Zhang et al. (2019a); Jamil and Bellos (2019)], regression model [Liu, Zhou, Chen et al. (2020); Song, Ren, Deng et al. (2020)] to artificial neural network (ANN) [Shakouri and Banihashemi (2019)]. Simultaneously, Boland et al. [Boland, Scott and Luther (2001)) employed the regression model and the fuzzy logic model to estimate the hourly R_d in some regions of Australia, revealing that the fuzzy logic model, with more meteorological parameters, outperformed the regression model in most cases. In 1982, Hargreaves et al. [Hargreaves and Samani (1982)] proposed an empirical coefficient model that estimates the daily R_d with empirical coefficient based on the highest and lowest temperatures, abbreviated as H-S. Elminir et al. [Elminir, Azzam and Younes (2007)] put forward an ANN model to estimate the R_d in some regions of Egypt, created some empirical models for the target stations, and proved that the ANN model is more suitable for R_d estimation in Egypt. In addition, the R_d has also been explored by correlation models between scatter ratio, scattering rate and meteorological factors like the clearness index, sunshine duration ratio (the ratio of actual sunshine duration to the theoretical maximum sunshine duration) and cloud cover. For instance, Liu et al. [Liu and Jordan (1960)] were the first to set up the linear relationship between the scatter ratio and the clearness index, and use the relationship to estimate the daily radiant exposure of the horizontal R_d. This model has been applied to different regions, yielding the empirical coefficient of each region. The relationship between scattering rate and sunshine duration ratio or daily cloud cover could be explained by linear, quadratic and cubic polynomial, exponential functions. Furthermore, some scholars combined meteorological factors (e.g., daily average air temperature and daily average air relative humidity) with sunshine duration ratio, and inputted the integrated parameter to estimate the daily radiant exposure of R_d [Fan, Wu, Zhang et al. (2019b)].

Though more and more methods are widely used to estimate the daily R_d now, the problem of the influence of complex and multi-parameter variables on accuracy has not been solved. In recent years, artificial intelligence methods have received increasing attention. These methods utilize flexible combinations of input parameters, and boast better accuracy than empirical models. The typical artificial intelligence approaches include SVM [Fan, Wu, Zhang et al. (2018)], MARS [Fan, Wu, Ma et al. (2020)], XGBoost [Fan, Wang, Wu et al. (2018)], gene expression programming (GEP) and random forest (RF) [Dong, Wu, Liu et al. (2020)]. With the aid of the SVM model and the ANN model, Ramli et al. [Ramli, Twaha, Al-Turki (2015)] estimated the inclined surface solar radiations in two spots of Saudi Arabia based on the direct radiational data and the global horizontal surface R_d, and compared the performances of the two models. The comparison shows that the SVM is more accurate, stable and efficient than the ANN. Taking the sunshine durations of three stations in China as the inputs, Chen et al. [Chen, Li and Wu (2013)] estimated the R_d with seven SVM models and five sunlight-based empirical models, and demonstrated that all SVM models outshined the empirical models. Shamshirband et al. [Shamshirband, Mohammadi, Khorasanizadeh et al. (2016)] integrated the SVM and the wavelet transform (WT) into a coupled model (Cluster-Based Approach (SVM-WT)) of solar radiation and horizontal diffusion, verified the

effectiveness of the SVM-WT using the set of daily R_d data measured in Kerman, Iran, and proved that this model is way more accurate than other models. Torabi et al. [Torabi, Mosavi, Ozturk et al. (2019)] combined the SVM and the ANN into a cluster-based approach (CBA) to estimate the horizontal R_d, and manifested that the CBA is better than ANN and SVM through experiments. Proposed by Jerome Friedman [Fisher (2015)] in 1991, the MARS can process a huge amount of high-dimensional data in a rapid and accurate manner, and has been applied in various fields. For example, Zhang et al. [Zhang and Goh (2013)] used the MARS and the neural network to solve geotechnical problems lacking accurate analytical theories, and discovered the strong generalization ability and accuracy of the MARS. Leathwick et al. [Leathwick, Elith and Hastie (2006)] analyzed the relationship between the distributions of 15 freshwater fishes and the environment using the generalized additive model (GAM) and the MARS, concluding that the MARS is more suitable to analyze large datasets than the GAM. The MARS has also been adopted for estimation of daily R_d. Using Kringing, MARS, M5Tree and Reynolds stress equation model (RSM), Keshtegara et al. [Keshtegar, Mert and Kisi (2018)] explored the influence of periodic data input, accurately estimated the R_d of the target stations, and compared the performances of the four models. The results show that the Kriging model achieved better performance than the other three models.

Proposed by Chen et al. [Chen, Li, Xiao et al. (2015)], XGBoost is an improved gradient boosting (GB) method based on the decision tree. This method enjoys a high computing efficiency and handles over-fitting problems excellently. In recent years, the XGBoost has become a popular tool in many areas. For example, Fan et al. [Fan, Wang, Wu et al. (2018)] verified that XGBoost excels over the SVM in estimating the daily global R_d of humid subtropical climate based on temperature and rainfall. Urraca et al. [Urraca, Antonanzas and Antonanzas-Torres (2017)] relied on XGBoost to estimate the daily global horizontal radiation at places with no temperature records, proving that the model is highly universal and better than the previous models in the Spanish literature. The XGBoost has also been utilized to estimate the daily reference evapotranspiration (ET_0), and the results are more accurate and stable than those of other models [Fan, Yue, Wu et al. (2018)]. Aler et al. [Aler, Galvan, Ruiz-Arias et al. (2017)] improved the separation of direct and diffuse solar radiation radiations with GB machine learning. Considering multiple influencing factors, some scholars confirmed the high accuracy and reliability of XGBoost in predicting regional power consumption [Gumus and Kiran (2017); Zheng, Yuan and Chen (2017)]. Son et al. [Son, Jung, Park et al. (2016)] proposed an online tracking algorithm based on Gradient Boosting Decision Tree (GBDT), which is more accurate than the advanced segmentation-based tracking method. Inspired by the GBDT, Wang et al. [Wang, Li, Wang et al. (2017)] designed a target recognition model based on high-resolution range profile (HRRP), examined the parameter selection for the model, and experimentally verified the advantage of the GBDT over the SVM in recognition and efficiency. Comparing XGBoost, RF and neural network on 30 internal datasets, Sheridan et al. [Sheridan, Wang, Liaw et al. (2016)] discovered that the XGBoost can run on a single CPU, costs less than one-third of the time of any other method, and has an obvious advantage in computing speed. Babajide et al. [Babajide and Saeed (2016)] predicted the bioactivity of compounds by the XGBoost. Their experimental results show that the XGBoost outperformed machine learning models like RF, SVM and Naïve Bayes (NB)

in predicting bioactivity, and exceled in the estimation based on diverse datasets. Wang et al. [Wang, Dong and Tian (2017)] used the improved XGBoost to estimate the loss of distribution feeder, verified the effectiveness of the algorithm by an example of 762 distribution feeders in the Shanghai Distribution Network, and validated the better accuracy of the XGBoost than the neural network. Thanks to its simplicity, high-speed, good effect and big data processing ability, the XGBoost has been extensively applied to various fields. However, there is less report on the application of XGBoost in R_d estimation, not to mention the coupling between XGBoost and machine learning models in R_d estimation. This also provides motivation for the research of this paper. Therefore, the purpose of this study is evaluated the ability of three machine learning models, i.e., XGBoost, SVM and MARS, to estimate the daily R_d. The regular meteorological data of 1966-2015 at five stations in humid subtropical China were taken as the input parameters (including mean average air temperature (T_a), theoretical sunshine duration (N), actual sunshine duration (n), daily average air relative humidity (RH), and extra-terrestrial solar radiation (R_a)). The accuracies of the three models were compared in the training phase and the testing phase, the effects of the input parameters on the daily R_d estimation were analyzed, and the optimal estimation algorithm for daily R_d was identified.

2 Materials and methods

2.1 Research location and data

In this study, data came from five stations in China, from 1966 to 2015.Three models were trained using meteorological data from 1966 to 2000. Then, use the 2001-2015 data to test the trained machine learning model. Each of the five stations can represent the meteorological conditions of the local region. The information and positions of the stations are respectively presented in Tab. 1 and Fig. 1. The technical roadmap of this paper is shown in Fig. 2. Besides, extra-terrestrial solar radiation (R_a) and theoretical sunshine duration (N) was calculated by geographic, season and solar information [Quej, Almorox, Arnaldo et al. (2017)]. The meteorological data were provided and checked by the National Meteorological Information Center (NMIC) of China Meteorological Administration (CMA). Some data entries were removed from the original data, because they were incomplete or had a greater-than-one ratio between measured R_d and theoretical R_d (the scattering ratio). During the quality check, the partially incorrect data entries were assigned the quality control code of zero and removed. In general, the original data were processed by the following principles:

(1) If one or more entries on a day are lost, all data of that day will be deleted;

(2) If the actual sunshine duration on a day is greater than the theoretical maximum sunshine duration on that day, all data of that day will be deleted.

The MARS programs were written in the MATLAB software (version 2011b, The MathWorks Inc.), while the XGBoost and SVM programs were written in the R software (version 3.2.3; R Project for Statistical Computing). All the simulations were performed in a computer with a single Intel Core i7-6700 @ 3.4-4.0 GHz and 16 GB of random-access memory (RAM).

Table 1: Basic information of selected sites in this study

Number	Station	Latitude (°N)	Longitude (°E)	Elevation (m)	H (MJ·m^{-2}·d^{-1})	T (°C)	n (h)	RH (%)	R_d (MJ·m^{-2}·d^{-1})	Data omission (%)
1	Beijing	39.80	116.47	31.3	14.1	12.7	7.1	55.9	6.64	0.1
2	Kunming	25.01	102.41	1897.0	15.0	16.0	6.1	71.2	6.89	5.0
3	Zhengzhou	34.72	113.65	110.4	13.2	15.1	5.7	64.2	7.25	0.1
4	Wuhan	30.38	114.04	27.0	12.1	17.3	5.2	76.6	6.77	0.6
5	Guangzhou	23.17	113.33	41.0	11.6	22.8	4.5	76.7	6.98	0.3

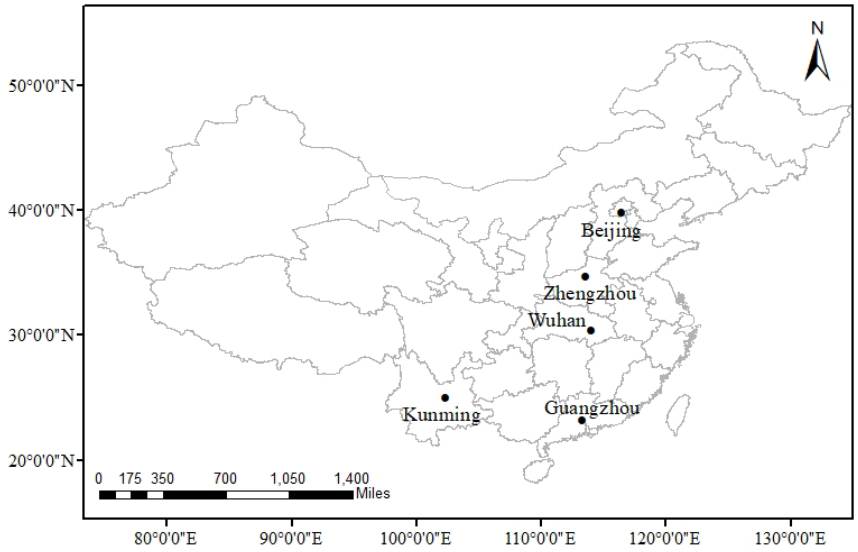


Figure 1: The geographical locations of the five diffuse global solar radiation stations in China

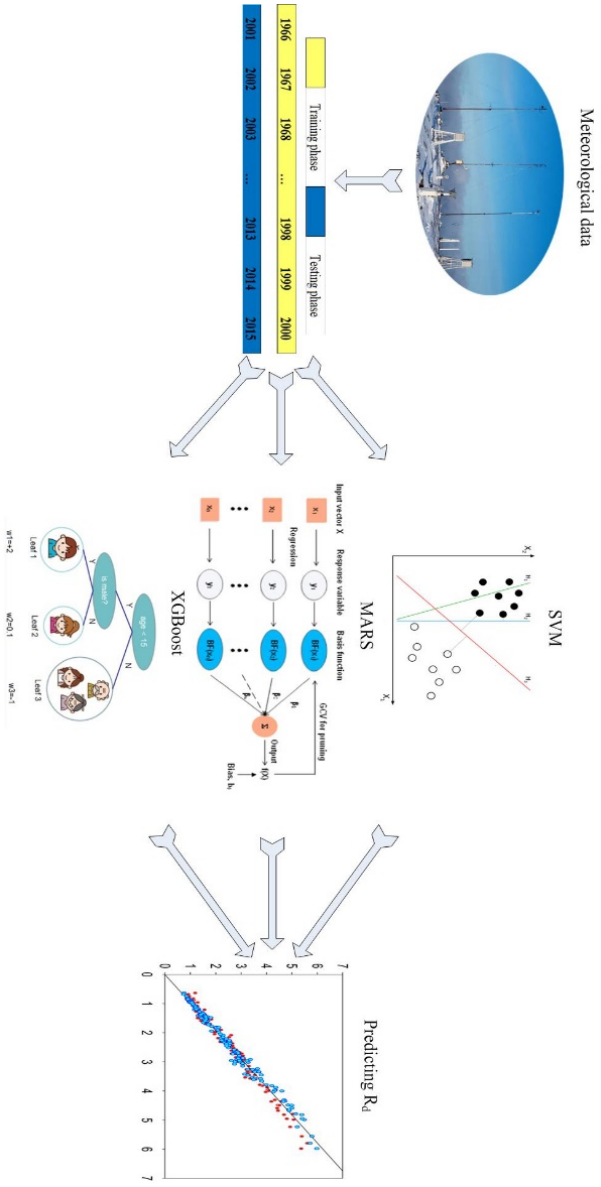


Figure 2: Technical roadmap for this article

2.2 Machine learning models for estimating diffuse solar radiation

2.2.1 Support vector machine (SVM)

The Support Vector Machine (SVM) method is a novel and effective method for dealing with non-linear classification and regression that has become popular internationally in recent years. It is based on the statistical learning theory proposed by Vapnik et al. [Cortes and Vapnik (1995)], with the help of the Mercer kernel expansion theorem and the results of the modern optimization method. The sample space is mapped to a higher-dimensional feature space, and the problem of seeking the optimal regression hyperplane is attributed to a quadratic convex programming problem under convex constraints in the feature space, and the optimal solution is obtained. Compared with the commonly used neural network model, the SVM model gives an unique solution due to the convexity of the optimality problem [Chen, Li and Wu (2013)].

The approximated function in the SVM algorithm is presented as follows:

$$f\left(x\right)=\omega\varphi\left(x\right)+b \tag{1}$$

where $\varphi\left(x\right)$ is a high-dimensional hyperplane function, ω is the weight and b is the bias, ω and b can be determined by minimizing the risk function ($R\left(\omega\right)$):

$$R\left(\omega\right)=C\frac{1}{p}\sum_{i=1}^{p}L(d_i,f(x_i))+\frac{1}{2}\left\|\omega\right\|^2 \tag{2}$$

where C is the penalty parameter of the error, d_i is the desired value, $\frac{1}{2}\left\|\omega\right\|^2$ is a confidence risk item and p is the number of observations. $C\frac{1}{p}\sum_{i=1}^{p}L(d_i,f(x_i))$ is an empirical risk function, it is the arithmetic mean of the target sample and the estimated value of the error, and the function L_ε can be determined below:

$$L_\varepsilon(d_i,f(x_i))=\begin{cases}\left|d-f(x)\right|-\varepsilon & \left|d-f(x)\right|\geq\varepsilon\\ 0 & \text{otherwise}\end{cases} \tag{3}$$

The relationship exists as the larger the C, the greater the effect of the empirical risk term, and the smaller the effect of the confidence risk term. ε is the error within the acceptable range of the training sample.

By introducing Lagrange multipliers and exploiting the optimality constraints, the approximate function in Eq. (2) can be expressed as:

$$f\left(x,a_i,a_i^*\right)=\sum_{i=1}^{p}\left(a_i-a_i^*\right)K\left(x\cdot x_i\right)+b \tag{4}$$

where α_i is a Lagrangian multiplier. This function satisfies the kernel function of the Mercer condition, and $K\left(x\cdot x_i\right)$ is called the kernel function. This paper used the radial

basis function as the basis function (RBF) [Chen and Li (2014); Fan, Wang, Wu et al. (2018)].

$$K\left(x \cdot x_i\right) = e^{-\gamma\|x-x_i\|^2}, \gamma > 0 \tag{5}$$

where x and x_i are the vectors in the input space. γ is the parameter of the kernel function. More information on the SVM model can be found in Vapnik's article [Cortes and Vapnik (1995)].

2.2.2 Multivariate adaptive regression splines (MARS)

MARS is a non-linear model proposed by Friedman et al. [Friedman and Stuetzle (1981)]. The MARS method has obvious advantages in dealing with large quantity sets and it has certain expansion capabilities. The MARS model has the feature of dividing the computation space into different regions, each with its own basis function to define the relationship between input and output parameters. The basis function is the basic unit of MARS and its form is as follows:

$$Y = \max(0, |D - x|) \tag{6}$$

Among them, Y - the basic function, x - input parameter, D - the threshold of the input parameter. The overall form of the MARS model can be expressed as

$$f(x) = b + \sum_{m=1}^{M} \beta_m h_m(x) \tag{7}$$

In the aquation, $f(x)$ -output result, b -the bias, M -the number of basic functions, $h_m(x)$ -the m basic function, β_m -the coefficient of the corresponding basic function. Developing a MARS model can be divided into two steps: the first step, be ready for the basic process. In this step, an over-fitting problem may occur; In the second step, the basis function that is not important to the result will be clipped. The process follows the generalized cross-validation (GCV) principle, which has the following form:

$$G_{cv} = \frac{\frac{1}{p}\sum_{i=1}^{p}(y_i - f(x))^2}{(1 - (\frac{C(B)}{p}))^2} \tag{8}$$

$$C(B) = B + 1 + dM \tag{9}$$

$C(B)$ -penalty function, B -the number of significant coefficients in the model, usually the same as M , d -penalty factor. The model with the smallest G_{cv} value in each model is the final MARS model.

2.2.3 Extreme gradient boosting (XGBoost)

XGBoost was first proposed by Chen et al. [Chen and Guestrin (2016)], and is a new model of Gradient Boost Machines (GBMs). XGBoost is highly efficient and accurate.

XGBoost is optimized for decision tree algorithms, it improves the processing of the database, and solves over-fitting problems through regularizeation and built-in cross-validation, improves accuracy and achieves optimal computational speed. In addition, during the training phase, the functions in XGBoost will automatically run and calculate. Therefore, it is widely used in the research of dimensionality reduction and feature extraction [Guo, Yang, Bie et al. (2019)], classification [Dong, Xu, Wang et al. (2018)] and behavior prediction [Ho, Wong, Yau et al. (2019)].

In the addition learning process in XGBoost, the algorithm combines all the predictions of a group of "weak" learners based on the idea of "enhancement", and cultivates "strong" learners through the addition training strategy [Fan, Wang, Wu et al. (2018)]. The function of step t is as follows:

$$f_i^{(t)} = \sum_k^t f_k\left(x_i\right) = f_i^{(t-1)} + f_t\left(x_i\right) \tag{10}$$

where $f_t\left(x_i\right)$ is the learner of step t, $f_i^{(t)}$ and $f_i^{(t-1)}$ are steps t and t-1, and x_i is the input variable.

To prevent over-fitting without affecting the model's computational speed, the XGBoost model can derive the following formula:

$$Obj^{(t)} = \sum_{k=1}^{p} l\left(\overline{y}_i, y_i\right) + \sum_{k=1}^{p} \Omega\left(f_i\right) \tag{11}$$

l is the loss function, p is the sum of the number of observations and Ω is a regularization term, the formula is:

$$\Omega\left(f\right) = \rho T + \frac{1}{2}\lambda\left\|\omega\right\|^2 \tag{12}$$

where ρ is the leaf node where the lowest loss requires further partitioning, λ is the regularization parameter, and ω is the weight. The details and calculation steps of the XGBoost algorithm can be found in the studies of Chen et al. [Chen and Guestrin (2016)].

2.3 Model comparison and statistical error analysis

Four common statistic indices were selected to estimate the daily R_d and compare the accuracies and performances of different estimation models for the daily R_d. These indices, namely, R^2 (coefficient of determination), RMSE (root mean square error), NRMSE (normalized root mean square error) and MBE (mean bias error). The mathematical equations of the statistical indicators are described below:

$$R^2 = \frac{\sum_{i=1}^{p}\left(O_{i,m} - O_{i,e}\right)^2}{\sum_{i=1}^{p}\left(O_{i,m} - \overline{O}_{i,m}\right)^2} \tag{13}$$

$$RMSE=\sqrt{\frac{1}{p}\sum_{i=1}^{p}(O_{i,m}-O_{i,e})^2} \tag{14}$$

$$MBE=\frac{1}{p}\sum_{i=1}^{p}(O_{i,m}-O_{i,e}) \tag{15}$$

$$NRMSE=\frac{\sqrt{\frac{1}{p}\sum_{i=1}^{n}(O_{i,m}-O_{i,e})^2}}{\overline{O}_{i,m}} \tag{16}$$

where $O_{i,e}$, $O_{i,m}$, $\overline{O}_{i,m}$ and p are the measured R_d, estimated R_d, mean measured R_d and the number of measurements, respectively. The greater the R^2 (i.e., the closer it gets to 1), the better the regression curve fits the data, and the better the performance of the algorithm. Conversely, the algorithm performance is negatively correlated with the values of RMSE, NRMSE and absolute MBE.

According to the requirements of the machine learning, the original meteorological data were normalized into the range of 0-1 by the following equation:

$$x_P=\frac{x_i-x_{\min}}{x_{\max}-x_{\min}} \tag{17}$$

where x_P and x_i represent the moralized and raw training and testing data; $x_{\max}$ and $x_{\min}$ are the maximum and minimum of the training and testing data.

3 Results and discussion

In this research, three machine learning models, i.e., MARS, SVM and XGBoost, were adopted to estimate the R_d based on the meteorological data on such five parameters as T_a, N, n, R_a and RH. The data were collected from five stations in the humid subtropical China, including Beijing, Kunming, Wuhan, Zhengzhou and Guangzhou. The five parameters were grouped into different combinations and inputted to each of the three models: (1) R_a, n, N, T_a and RH (2) R_a, n, N and T_a (3) R_a, n and N.

3.1 Prediction of machine learning models

Each of the three models showed different accuracies under different parameter combinations. Note that MARS 1-3 respectively denote the MARS algorithm inputted with parameter combinations 1-3. The notations of the other two models were similarly defined. The values of the four commonly used statistical indicators for the Beijing, Kunming, Wuhan, Zhengzhou and Guangzhou stations in the training and testing phases are respectively recorded in Tabs. 2-6. As shown in Tab. 2, the combination of R_a, n, N, T_a and RH led to the best accuracy in the testing phase at the Beijing station. While in the training phase, the XGBoost model achieved the best performance in this phase. In the testing phase, the SVM (on average R^2=0.777, RMSE=1.822 MJ·m^{-2}·d^{-1}, MBE=0.119 MJ·m^{-2}·d^{-1}, NRMSE=0.272) model performed better than two other models. But the

SVM model performance is very close to the XGBoost model. Thus, all three models underwent an increase in accuracy from the training phase to the testing phase. Moreover, the *RH* seems to be the best parameter for modelling at the Beijing station. Tab. 2 shows that the addition of T_a and *RH* could enhance the estimation accuracy of R_d. Judging by the MBE, the MARS1 outputted a small negative value in the testing phase at the Beijing station, while the other models all overestimated the R_d. However, the overestimation is not serious, for the MBE values were smaller than 0.15 $MJ \cdot m^{-2} \cdot d^{-1}$.

In the Kunming station, as shown in Tab. 3, from the overall performance of the three models in the training period, the MARS model estimated the accuracy of R_d better, followed by the XGBoost model. Nonetheless, the MARS did not have any advantage in RMSE, MBE and NRMSE in this phase. For example, the R^2 values of MARS1, SVM1 and XGBoost1 were 0.796, 0.741 and 0.781 in the training phase, which changed to 0.728 (-8.5%), 0.796 (+7.4%) and 0.796 (+1.9%) in the testing phase. In the training phase, the XGBoost1 model (R^2=0.731) realized better accuracy than XGBoost2 and XGBoost3 models. After entering the testing phase, the MARS model saw a decline in R^2 value, averaging at -9.08%. Meanwhile, the SVM and XGBoost models became more accurate, with mean R^2 values growing by 7.4% and 3.7%, respectively. However, in all models, although the XGBoost2 (R^2=0.717) model performs best, the performance and the XGBoost1(R^2=0.714) model are still similar. In the Kunming Station, the performance of the model is different from that of Beijing station. At the Beijing station, the XGBoost was the optimal algorithm in the training phase, while the SVM was the best in the testing phase. Tab. 3 show MBE of MARS far exceeded 0.15 $MJ \cdot m^{-2} \cdot d^{-1}$ in the training phase, but that of SVM or XGBoost model was below that figure. This means all three models overestimate the daily R_d in the training phase at the Kunming station, and the biggest overestimation go to MARS model. The results were completely the opposite in the testing phase. In that phase, SVM and XGBoost seriously overestimated the daily R_d, while the MARS model remained relatively stable. Only MARS1 model had a negative MBE in the testing phase, which did not greatly affect the estimation of daily R_d. As for the Beijing station, the three models were all below 0.15 $MJ \cdot m^{-2} \cdot d^{-1}$ in terms of the MBE, revealing their stability in estimating daily R_d at this station. According to the statistic indices at Wuhan, Zhengzhou and Guangzhou, the performance curves of the three models obeyed the same trends as those at the Kunming station in both the training and testing phases. Feng et al. [Feng, Lin, Wang et al. (2018)] also used the model to estimate the R_d at station such as Zhengzhou, and achieved good estimation accuracy. As can be seen from Tabs. 4-6, the SVM and XGBoost models had basically the same index values in the testing phase, showing no obvious performance gap. Hence, the two models are both applicable to the R_d estimation, despite a slight advantage of the XGBoost. Judging by the testing phase RMSEs, the XGBoost model outputted smaller RMSEs than the SVM model at all stations except Beijing. As a result, the XGBoost model is more stable than the SVM model in R_d estimation.

Table 2: Comparison of MARS, SVM and XGBoost models for predicting R_d at Beijing station

	Beijing	Training				Testing			
		R^2	RMSE ($MJ \cdot m^{-2} \cdot d^{-1}$)	MBE ($MJ \cdot m^{-2} \cdot d^{-1}$)	NRMSE	R^2	RMSE ($MJ \cdot m^{-2} \cdot d^{-1}$)	MBE ($MJ \cdot m^{-2} \cdot d^{-1}$)	NRMSE
MARS1	R_a *n N* T_a *RH*	0.776	1.835	0.064	0.274	0.748	1.720	-0.001	0.260
MARS2	R_a *n N* T_a	0.762	1.885	0.143	0.281	0.741	1.742	0.000	0.263
MARS3	R_a *n N*	0.736	1.977	0.129	0.295	0.734	1.766	0.000	0.267
SVM1	R_a *n N* T_a *RH*	0.767	1.656	0.031	0.250	0.808	1.715	0.140	0.256
SVM2	R_a *n N* T_a	0.752	1.709	0.007	0.258	0.783	1.799	0.094	0.269
SVM3	R_a *n N*	0.733	1.772	0.006	0.268	0.741	1.952	0.122	0.291
XGBoost1	R_a *n N* T_a *RH*	0.806	1.512	0.002	0.228	0.792	1.783	0.126	0.266
XGBoost2	R_a *n N* T_a	0.771	1.641	0.001	0.248	0.767	1.862	0.113	0.278
XGBoost3	R_a *n N*	0.752	1.705	0.000	0.258	0.734	1.983	0.124	0.296

Table 3: Comparison of MARS, SVM and XGBoost models for predicting R_d at Kunming station

	Kunming	Training				Testing			
		R^2	RMSE ($MJ \cdot m^{-2} \cdot d^{-1}$)	MBE ($MJ \cdot m^{-2} \cdot d^{-1}$)	NRMSE	R^2	RMSE ($MJ \cdot m^{-2} \cdot d^{-1}$)	MBE ($MJ \cdot m^{-2} \cdot d^{-1}$)	NRMSE
MARS1	R_a *n N* T_a *RH*	0.709	1.752	0.178	0.244	0.664	1.914	-0.001	0.282
MARS2	R_a *n N* T_a	0.714	1.762	0.344	0.245	0.645	1.972	0.000	0.291
MARS3	R_a *n N*	0.702	1.822	0.441	0.253	0.623	2.032	0.000	0.300
SVM1	R_a *n N* T_a *RH*	0.682	1.863	0.101	0.275	0.702	1.786	0.273	0.248
SVM2	R_a *n N* T_a	0.654	1.944	0.104	0.287	0.709	1.789	0.405	0.249
SVM3	R_a *n N*	0.618	2.048	0.113	0.302	0.694	1.871	0.539	0.260
XGBoost1	R_a *n N* T_a *RH*	0.731	1.719	0.000	0.253	0.714	1.739	0.149	0.242
XGBoost2	R_a *n N* T_a	0.677	1.879	-0.002	0.277	0.717	1.747	0.302	0.243
XGBoost3	R_a *n N*	0.646	1.966	0.000	0.290	0.699	1.831	0.445	0.255

Table 4: Comparison of MARS, SVM and XGBoost models for predicting R_d at Zhengzhou station

	Zhengzhou	Training				Testing			
		R^2	RMSE ($MJ \cdot m^{-2} \cdot d^{-1}$)	MBE ($MJ \cdot m^{-2} \cdot d^{-1}$)	NRMSE	R^2	RMSE ($MJ \cdot m^{-2} \cdot d^{-1}$)	MBE ($MJ \cdot m^{-2} \cdot d^{-1}$)	NRMSE
MARS1	R_a *n N* T_a *RH*	0.796	1.971	0.523	0.249	0.728	1.818	-0.003	0.259
MARS2	R_a *n N* T_a	0.745	2.199	0.789	0.278	0.692	1.935	0.000	0.276
MARS3	R_a *n N*	0.752	2.150	0.727	0.272	0.682	1.966	0.000	0.281

SVM1	R_a *n N* T_a *RH*	0.741	1.769	0.002	0.252	0.796	1.942	0.566	0.245
SVM2	R_a *n N* T_a	0.714	1.864	-0.009	0.266	0.778	2.043	0.738	0.258
SVM3	R_a *n N*	0.682	1.969	0.030	0.281	0.750	2.148	0.761	0.271
XGBoost1	R_a *n N* T_a *RH*	0.781	1.628	0.001	0.233	0.796	1.942	0.524	0.245
XGBoost2	R_a *n N* T_a	0.734	1.796	0.001	0.257	0.773	2.091	0.769	0.264
XGBoost3	R_a *n N*	0.704	1.896	0.000	0.271	0.759	2.123	0.730	0.268

Table 5: Comparison of MARS, SVM and XGBoost models for predicting R_d at Wuhan station

	Wuhan	Training				Testing			
		R^2	RMSE ($MJ \cdot m^{-2} \cdot d^{-1}$)	MBE ($MJ \cdot m^{-2} \cdot d^{-1}$)	NRMSE	R^2	RMSE ($MJ \cdot m^{-2} \cdot d^{-1}$)	MBE ($MJ \cdot m^{-2} \cdot d^{-1}$)	NRMSE
MARS1	R_a *n N* T_a *RH*	0.748	2.153	0.350	0.298	0.684	1.994	-0.002	0.301
MARS2	R_a *n N* T_a	0.731	2.226	0.480	0.308	0.663	2.058	0.000	0.312
MARS3	R_a *n N*	0.719	2.249	0.457	0.311	0.645	2.113	0.000	0.320
SVM1	R_a *n N* T_a *RH*	0.697	1.951	0.072	0.295	0.750	2.133	0.402	0.295
SVM2	R_a *n N* T_a	0.672	2.030	0.059	0.307	0.745	2.201	0.561	0.305
SVM3	R_a *n N*	0.642	2.127	0.119	0.322	0.714	2.279	0.556	0.316
XGBoost1	R_a *n N* T_a *RH*	0.741	1.805	0.002	0.273	0.759	2.101	0.349	0.291
XGBoost2	R_a *n N* T_a	0.694	1.964	-0.002	0.297	0.741	2.190	0.487	0.303
XGBoost3	R_a *n N*	0.666	2.048	0.001	0.310	0.719	2.247	0.463	0.311

Table 6: Comparison of MARS, SVM and XGBoost models for predicting R_d at Guangzhou station

	Guangzhou	Training				Testing			
		R^2	RMSE ($MJ \cdot m^{-2} \cdot d^{-1}$)	MBE ($MJ \cdot m^{-2} \cdot d^{-1}$)	NRMSE	R^2	RMSE ($MJ \cdot m^{-2} \cdot d^{-1}$)	MBE ($MJ \cdot m^{-2} \cdot d^{-1}$)	NRMSE
MARS1	R_a *n N* T_a *RH*	0.709	1.651	0.297	0.221	0.627	1.929	-0.002	0.283
MARS2	R_a *n N* T_a	0.694	1.746	0.518	0.234	0.585	2.042	0.000	0.300
MARS3	R_a *n N*	0.677	1.809	0.578	0.242	0.562	2.098	0.000	0.309
SVM1	R_a *n N* T_a *RH*	0.661	1.850	0.061	0.272	0.721	1.645	0.426	0.220
SVM2	R_a *n N* T_a	0.615	1.972	0.062	0.290	0.716	1.727	0.621	0.231
SVM3	R_a *n N*	0.555	2.120	0.114	0.312	0.658	1.899	0.708	0.254
XGBoost1	R_a *n N* T_a *RH*	0.707	1.718	0.001	0.253	0.726	1.637	0.415	0.219
XGBoost2	R_a *n N* T_a	0.637	1.912	-0.002	0.281	0.712	1.718	0.568	0.230
XGBoost3	R_a *n N*	0.587	2.040	0.000	0.300	0.671	1.824	0.584	0.244

3.2 Comparison of model accuracy under various input combinations

Three artificial intelligence models and three meteorological data parameters were combined to estimate the daily R_d value and the measured values of the Guangzhou Meteorological Station in China. The scatter plots are plotted in Figs. 3 and 4 (the training and testing period of MARS1, SVM1 and XGBoost1 models, respectively) Comparing the results of MARS1, SVM1 and XGBoost1 models in the training phase and the testing phase, it is obvious that the scatter points of SVM1 and XGBoost1 models were closer to the fitting line and more evenly distributed than those of MARS1 model. Besides, all three models witnessed an increase of accuracy from the training phase to the testing phase. As shown in Tab. 6, the models outputted relatively small R^2 values under the combination of R_a, n and N, and the values did not obey any obvious law. Thus, this combination suppressed the estimation accuracy. This is because Guangzhou, as a subtropical humid region, has a high perennial rainfall, which affects the R_d estimation. This conclusion is consistent with the argument of Fan et al. [Fan, Wang, Wu et al. (2018)] that rainfall makes the estimation inaccurate. The low accuracy is also attributable to the limited number of parameters. Clearly, the same algorithm yielded more accurate R_d under the combination of R_a, n, N, T_a and RH than that under the other combinations. The RMSE values in the table indicate that XGBoost is slightly more stable than SVM model, which is consistent with the conclusions obtained in Figs. 3 and 4.

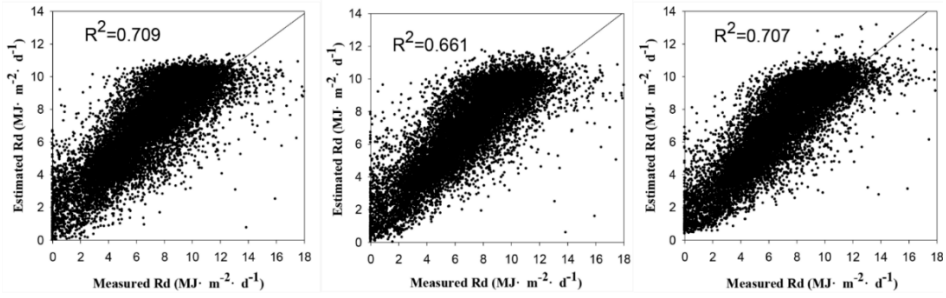


Figure 3: Scatterplot of global diffuse solar radiation values estimated from the three selected models versus corresponding values observed at Guangzhou station for the training phase (note: fine line is the best fitted line)

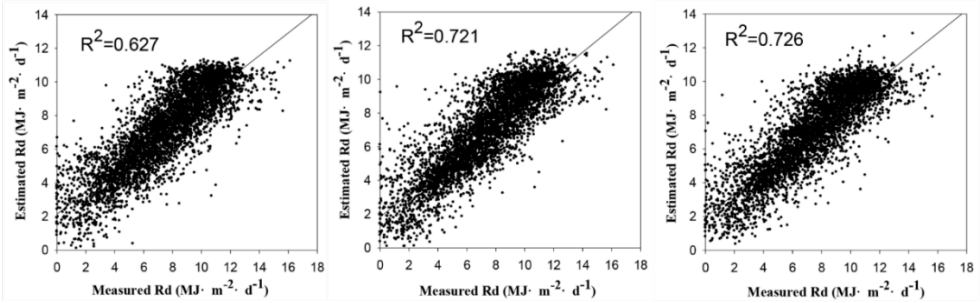
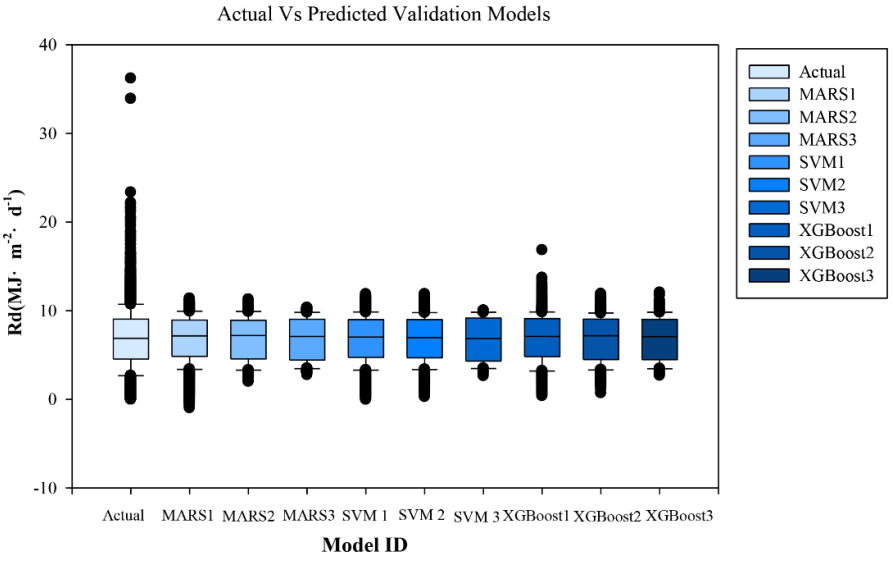


Figure 4: Scatterplot of global diffuse solar radiation values estimated from the three selected models versus corresponding values observed at Guangzhou station for the testing phase (note: fine line is the best fitted line)

3.3 Comparison of stability of various machine learning models

Fig. 5 are the box plots that compare the measured R_d values with the R_d values estimated by the three models, respectively inputted with the three parameter combinations, for the Guangzhou station. In the training phase, the SVM faced deviation between the median values of measured and estimated R_d values, and the largest deviation belonged to SVM3. Meanwhile, the other models all performed stably. The models faced similar deviations between the quartiles of the measured and estimated R_d values. On the extreme R_d values, MARS1 model failed to simulate the minimum R_d; the best simulation was realized by XGBoost1 model, followed by SVM1 and SVM2 models. In the testing phase, all models saw deviation between the median values of measured and estimated R_d values. MARS1, SVM1 and XGBoost1 model shad relatively small deviations, while SVM3 had the greatest deviation. The models all had deviations between the quartiles of the measured and estimated R_d values, with MARS2, SVM1 and XGBoost1 performing relatively well. On the extreme R_d values, MARS1 outputted negative value, failing to simulate the standard minimum value; XGBoost1 achieved the best performance, followed by SVM1 and XGBoost2. To sum up, XGBoost was the best performer of R_d estimation at Guangzhou, and the second-best performer was SVM. It is also found that, with the increase in the number of parameters, the models were improved in R_d estimation ability and accuracy.



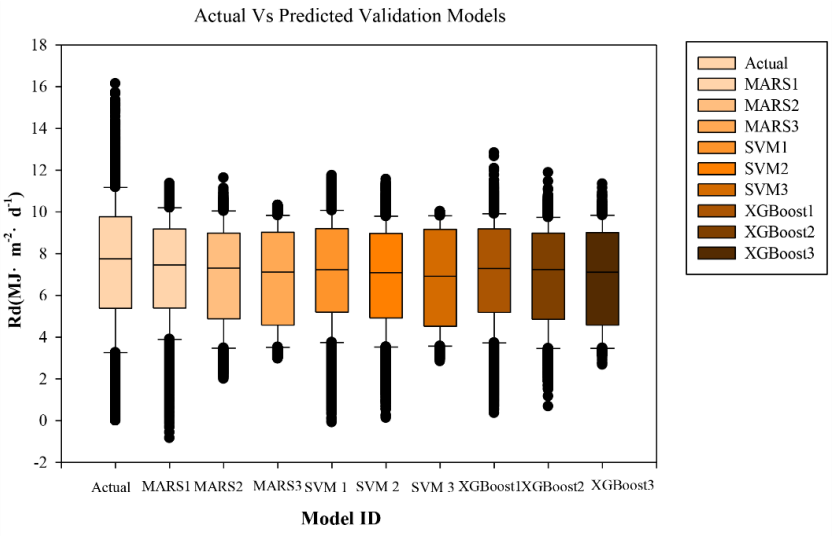
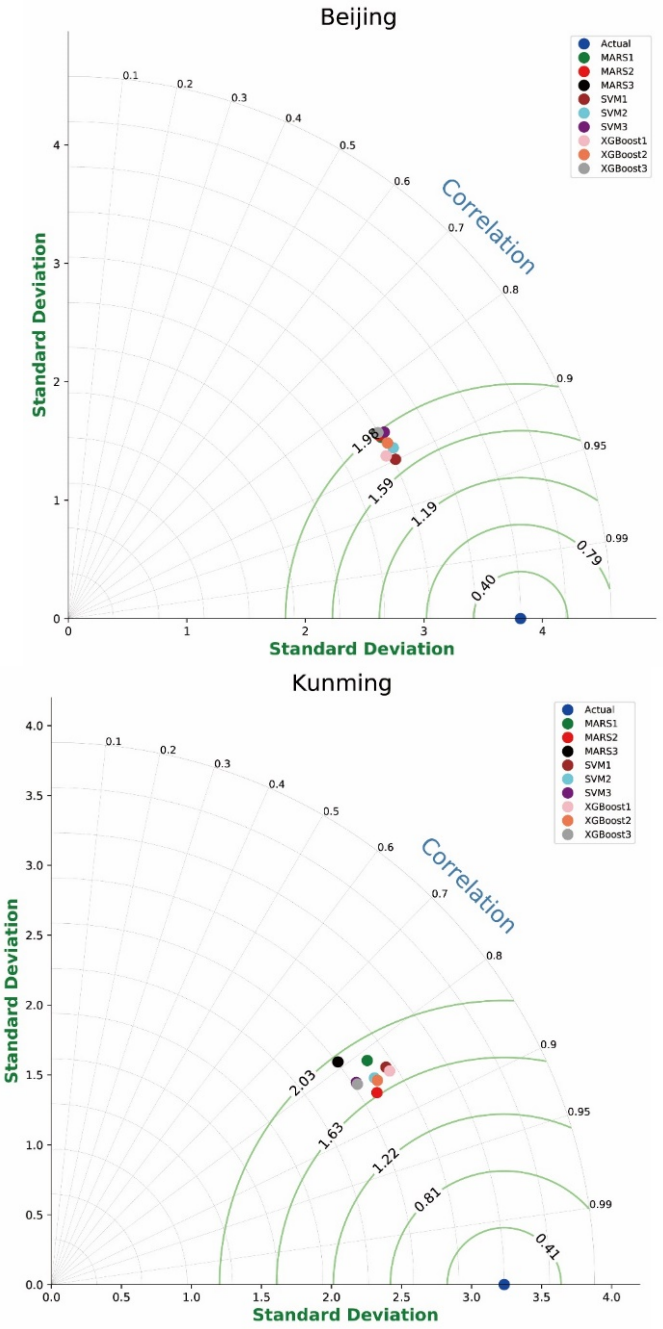


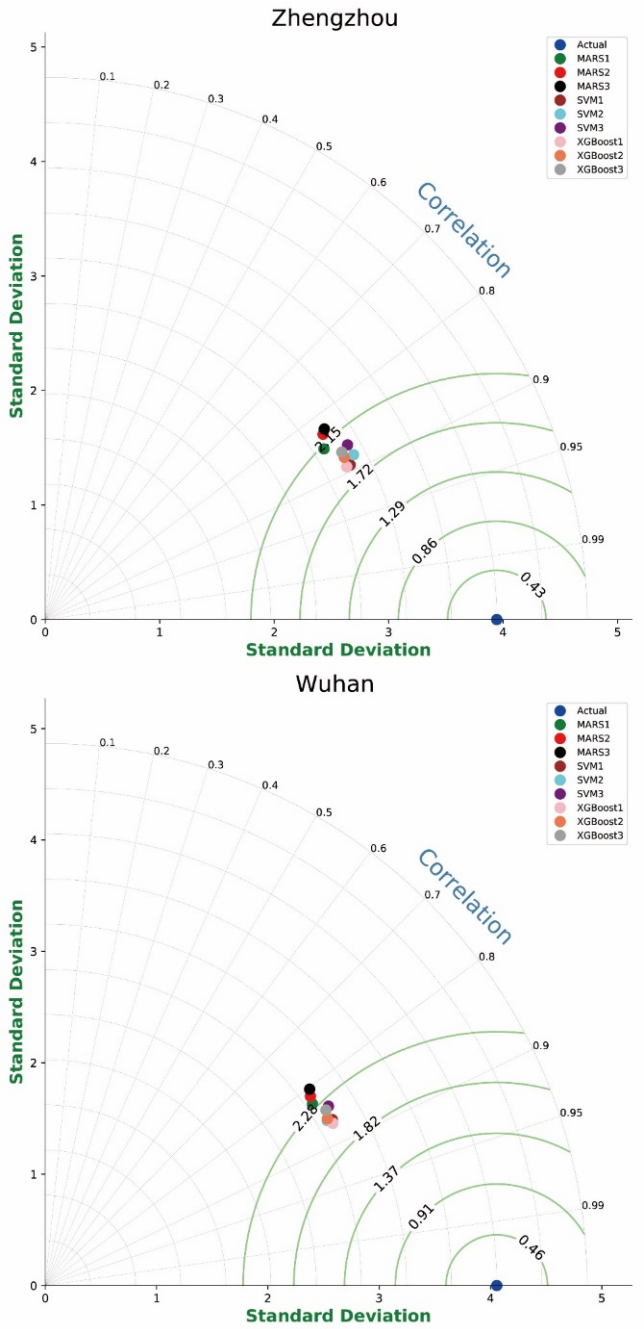
Figure 5: Estimated R_d for actual and nine selected models for training(up) and testing(down) phase of the Guangdong station

3.4 Comparison of comprehensive performance of various machine learning models

Fig. 6 provides the Taylor diagrams of the algorithm results at the five stations. The Taylor diagram has two major advantages. First, the RMSE, standard deviation (SD) and correlation coefficient (R) values are geometrically presented, which facilitate data comparison [Simon-Martin, Alonso-Tristan and Diez-Mediavilla (2017)]. Second, the particular advantage of presenting statistical results using Taylor diagrams is that the models are apparently clustered according to their performance [Despotovic, Nedic, Despotovic et al. (2016)]. At the Beijing station, SVM1 and XGBoost1 models were the best and second-best performers. MARS1, MARS2, MARS3, SVM3 and XGBoost3 had similar R, RMSE and SD values and relatively concentrated positions. However, these model positions were far from the reference point, indicating that these models performed poorly. At the Kunming station, MARS2 model (R=0.861, RMSE=1.972 $MJ{\cdot}m^{-2}{\cdot}d^{-1}$ and SD=2.700 $MJ{\cdot}m^{-2}{\cdot}d^{-1}$) achieved the best performance. After MARS2, the following pairs of models had similar performances, such as XGBoost2 and SVM2, XGBoost1 and SVM1, as well as XGBoost3 and SVM3. It can be inferred that XGBoost and SVM models had similarly good performances, when the R_d at Kunming was estimated under the same parameter combination. Specifically, the MARS3 model had the worst performance, because its position was the farthest one from the reference point. At the Zhengzhou Station, XGBoost1 and SVM1 were the closest to the reference point, making them the top performers. Overall, XGBoost and SVM both had concentrated patterns. The worst performing algorithm at this station was MARS3. At the Wuhan Station, XGBoost1 (R=0.871, RMSE=2.101 $MJ{\cdot}m^{-2}{\cdot}d^{-1}$ and SD=2.966 $MJ{\cdot}m^{-2}{\cdot}d^{-1}$) boasted the best performance, followed by SVM1, SVM2, XGBoost1and XGBoost2; MARS3 (R=0.803, RMSE=2.113 $MJ{\cdot}m^{-2}{\cdot}d^{-1}$ and SD=2.958 $MJ{\cdot}m^{-2}{\cdot}d^{-1}$) still had the poorest performance. At the Guangdong Station, there were huge differences in performance among the models. XGBoost1 and SVM1 had the best performance, on average R=0.851,

RMSE=1.641 MJ·m^{-2}·d^{-1} and SD=2.440 MJ·m^{-2}·d^{-1}, followed by XGBoost2 and SVM2, while MARS3 remained as the worst performer, on average R=0.845, RMSE=1.723 MJ·m^{-2}·d^{-1} and SD=2.378 MJ·m^{-2}·d^{-1}.




Zhengzhou
Actual
MARS1
MARS2
MARS3
SVM1
SVM2
SVM3
XGBoost1
XGBoost2
XGBoost3
Correlation
Standard Deviation
Standard Deviation
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.95
0.99
2.15
1.72
1.29
0.86
0.43
Wuhan
Actual
MARS1
MARS2
MARS3
SVM1
SVM2
SVM3
XGBoost1
XGBoost2
XGBoost3
Correlation
Standard Deviation
Standard Deviation
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.95
0.99
2.28
1.82
1.37
0.91
0.46

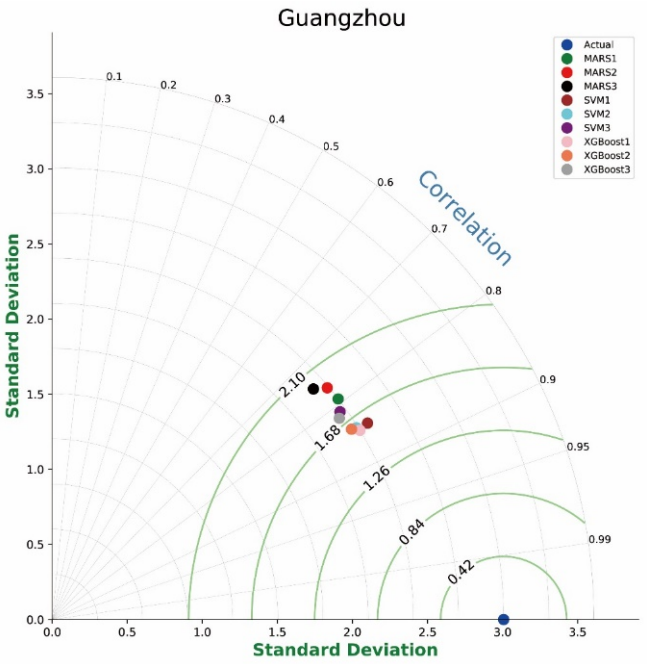


Figure 6: Taylor diagrams of the models investigated in five stations

Summing up the scatterplot, box plot and Taylor diagram comparisons, it can be concluded that XGBoost and SVM are more accurate and stable than MARS in R_d estimation accuracy and stability at the five station, and XGBoost is more stable over SVM. As a result, the performance level of the model is: XGBoost > SVM > MARS.

4 Conclusions

Solar radiation is the major source of energy on the surface of the Earth. It has a direct bearing on the survival of animals and plants, as well as our production and life. This paper evaluated the ability of three machine learning models, i.e., XGBoost, SVM and MARS, to estimate the daily R_d. The regular meteorological data of 1966-2015 at five stations in China were taken as the input parameters (including mean average air temperature (T_a), theoretical sunshine duration (N), actual sunshine duration (n), daily average air relative humidity (RH), and extra-terrestrial solar radiation (R_a)). The estimation results of the models were compared under each parameter combination. The comparison shows that the XGBoost and SVM models has better performance than the MARS model in estimating daily R_d. Overall, the tested models can be ranked as XGBoost>SVM>MARS in descending order of R_d estimation performance. The RH and T_a parameters can improve the R_d estimation accuracy. Moreover, XGBoost had a slight better performance than SVM, as well as a stronger stability. Considering accuracy and stability, XGBoost model is the most suitable algorithm for daily R_d estimation in China based on regular meteorological data. In future research, more parameters will be introduced to estimate daily R_d (such as precipitation, etc.). Of course, it should be noted that regional differences may affect the applicability of the algorithm and the integrity of the data. Therefore, it is recommended to

use the similarity of the development model in similar climates for other countries (such as Japan, Korea, etc.) for further research. Therefore, daily R_d estimates in other regions should be based on local conditions and using appropriate methods and parameters.

Acknowledgement: Thanks to the National Meteorological Information Center of China Meteorological Administration for offering the meteorological data.

Funding Statement: This study was jointly supported by National Natural Science Foundation of China (51769010, 51979133, 51469010 and 51109102).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Aler, R.; Galvan, I. M.; Ruiz-Arias, J. A.; Gueymard, C. A. (2017): Improving the separation of direct and diffuse solar radiation components using machine learning by gradient boosting. *Solar Energy*, vol. 150, pp. 558-569.

Babajide, M. I.; Saeed, F. (2016): Bioactive molecule prediction using extreme gradient boosting. *Molecules*, vol. 21, no. 8, pp. 983-994.

Boland, J.; Huang, J.; Ridley, B. (2013): Decomposing global solar radiation into its direct and diffuse components. *Renewable and Sustainable Energy Reviews*, vol. 28, pp. 749-756.

Boland, J.; Scott, L.; Luther, M. (2001): Modelling the diffuse fraction of global solar radiation on a horizontal surface. *Environmetrics*, vol. 12, no. 2, pp. 103-116.

Carballo, J. A.; Bonilla, J.; Berenguel, M.; Fernández-Reche, J.; García, G. (2019): New approach for solar tracking systems based on computer vision, low cost hardware and deep learning. *Renewable Energy*, vol. 133, pp. 1158-1166.

Charvát, P.; Klimeš, L.; Pech, O.; Hejčík, J. (2019): Solar air collector with the solar absorber plate containing a PCM-Environmental chamber experiments and computer simulations. *Renewable Energy*, vol. 143, pp. 731-740.

Chen, J.; Li, G. (2014): Evaluation of support vector machine for estimation of solar radiation from measured meteorological variables. *Theoretical and Applied Climatology*, vol. 115, no. 3-4, pp. 627-638.

Chen, J.; Li, G.; Wu, S. (2013): Assessing the potential of support vector machine for estimating daily solar radiation using sunshine duration. *Energy Conversion and Management*, vol. 75, pp. 311-318.

Chen, J.; Li, G.; Xiao, B.; Wen, Z.; Lv, M. et al. (2015): Assessing the transferability of support vector machine model for estimation of global solar radiation from air temperature. *Energy Conversion and Management*, vol. 89, pp. 318-329.

Chen, Q.; Ding, Q.; Liu, X. (2019): Establishment and validation of a solar radiation model for a living wall system. *Energy and Buildings*, vol. 195, pp. 105-115.

Chen, T.; Guestrin, C. (2016): XGBoost: a scalable tree boosting system. *Acm Sigkdd International Conference on Knowledge Discovery & Data Mining*, pp. 785-794.

Cortes, C.; Vapnik, V. (1995): Support-vector networks. *Machine Learning*, vol. 20, no. 3, pp. 273-297.

Daus, Y. V.; Yudaev, I. V.; Taranov, M. A.; Voronin, S. M.; Gazalov, V. S. (2019): Reducing the costs for consumed electricity through the solar energy utilization. *International Journal of Energy Economics and Policy*, vol. 9, no. 2, pp. 19-23.

Despotovic, M.; Nedic, V.; Despotovic, D.; Cvetanovic, S. (2016): Evaluation of empirical models for predicting monthly mean horizontal diffuse solar radiation. *Renewable & Sustainable Energy Reviews*, vol. 56, pp. 246-260.

Dong, H.; Xu, X.; Wang, L.; Pu, F. (2018): Gaofen-3 PolSAR image classification via xgboost and polarimetric spatial information. *Sensors*, vol. 18, no. 2, pp. 611-631.

Dong, J.; Wu, L.; Liu, X.; Li, Z.; Gao, Y. et al. (2020): Estimation of daily dew point temperature by using bat algorithm optimization based extreme learning machine. *Applied Thermal Engineering*, vol. 165, no. 114569, pp. 1-15.

Elminir, H. K.; Azzam, Y. A.; Younes, F. I. (2007): Prediction of hourly and daily diffuse fraction using neural network, as compared to linear regression models. *Energy*, vol. 32, no. 8, pp. 1513-1523.

Fan, J.; Wang, X.; Wu, L.; Zhou, H.; Zhang, F. et al. (2018): Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: a case study in China. *Energy Conversion and Management*, vol. 164, pp. 102-111.

Fan, J.; Wu, L.; Ma, X.; Zhou, H.; Zhang, F. (2020): Hybrid support vector machines with heuristic algorithms for prediction of daily diffuse solar radiation in air-polluted regions. *Renewable Energy*, vol. 145, pp. 2034-2045.

Fan, J.; Wu, L.; Zhang, F.; Cai, H.; Ma, X. et al. (2019a): Evaluation and development of empirical models for estimating daily and monthly mean daily diffuse horizontal solar radiation for different climatic regions of China. *Renewable and Sustainable Energy Reviews*, vol. 105, pp. 168-186.

Fan, J.; Wu, L.; Zhang, F.; Cai, H.; Wang, X. et al. (2018): Evaluating the effect of air pollution on global and diffuse solar radiation prediction using support vector machine modeling based on sunshine duration and air temperature. *Renewable and Sustainable Energy Reviews*, vol. 94, pp. 732-747.

Fan, J.; Wu, L.; Zhang, F.; Cai, H.; Zeng, W. et al. (2019b): Empirical and machine learning models for predicting daily global solar radiation from sunshine duration: a review and case study in China. *Renewable and Sustainable Energy Reviews*, vol. 100, pp. 186-212.

Fan, J.; Yue, W.; Wu, L.; Zhang, F.; Cai, H. et al. (2018): Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China. *Agricultural and Forest Meteorology*, vol. 263, pp. 225-241.

Feng, L.; Lin, A.; Wang, L.; Qin, W.; Gong, W. (2018): Evaluation of sunshine-based

models for predicting diffuse solar radiation in China. *Renewable and Sustainable Energy Reviews*, vol. 94, pp. 168-182.

Fisher, N. I. (2015): A conversation with Jerry Friedman. *Statistical Science*, vol. 30, no. 2, pp. 268-295.

Friedman, J. H.; Stuetzle, W. (1981): Projection pursuit regression. *Publications of the American Statistical Association*, vol. 76, no. 376, pp. 817-823.

Gumus, M.; Kiran, M. S. (2017): Crude oil price forecasting using XGBoost. *International Conference on Computer Science and Engineering*, pp. 1100-1103.

Guo, J.; Yang, L.; Bie, R.; Yu, J.; Gao, Y. et al. (2019): An XGBoost-based physical fitness evaluation model using advanced feature selection and Bayesian hyper-parameter optimization for wearable running monitoring. *Computer Networks*, vol. 151, pp. 166-180.

Hargreaves, G. H.; Samani, Z. A. (1982): Estimating potential evapotranspiration. *Journal of the Irrigation and Drainage Division*, vol. 108, no. 3, pp. 225-230.

He, L.; Lei, B.; Bi, H.; Yu, T. (2016): Simplified building thermal model used for optimal control of radiant cooling system. *Mathematical Problems in Engineering*, vol. 2016, no. 2976731, pp. 1-15.

Ho, S. C.; Wong, K. C.; Yau, Y. K.; Yip, C. K. (2019): A machine learning approach for predicting bank customer behavior in the banking industry. *Machine Learning and Cognitive Science Applications in Cyber Security*, *IGI Global*, pp. 57-83.

Jahani, B.; Dinpashoh, Y.; Nafchi, A. R. (2017): Evaluation and development of empirical models for estimating daily solar radiation. *Renewable & Sustainable Energy Reviews*, vol. 73, pp. 878-891.

Jamil, B.; Bellos, E. (2019): Development of empirical models for estimation of global solar radiation exergy in India. *Journal of Cleaner Production*, vol. 207, pp. 1-16.

Keshtegar, B.; Mert, C.; Kisi, O. (2018): Comparison of four heuristic regression techniques in solar radiation modeling: Kriging method vs RSM, MARS and M5 model tree. *Renewable & Sustainable Energy Reviews*, vol. 81, no. 1, pp. 330-341.

Leathwick, J. R.; Elith, J.; Hastie, T. (2006): Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling*, vol. 199, no. 2, pp. 188-196.

Li, X.; Ratti, C. (2019): Mapping the spatio-temporal distribution of solar radiation within street canyons of Boston using Google Street View panoramas and building height model. *Landscape and Urban Planning*, vol. 191, no. 103387, pp. 1-12.

Liu, B. Y. H.; Jordan, R. C. (1960): The interrelationship and characteristic distribution of direct, diffuse and total solar radiation. *Solar Energy*, vol. 4, no. 3, pp. 1-19.

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

Qu, W.; Hong, H.; Jin, H. (2019): A spectral splitting solar concentrator for cascading

solar energy utilization by integrating photovoltaics and solar thermal fuel. *Applied Energy*, vol. 248, pp. 162-173.

Quej, V. H.; Almorox, J.; Arnaldo, J. A.; Saito, L. (2017): ANFIS, SVM and ANN soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment. *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 155, pp. 62-70.

Ramli, M. A. M.; Twaha, S.; Al-Turki, Y. A. (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 Conversion and Management*, vol. 105, pp. 442-452.

Rehman, N. U.; Uzair, M. (2017): The proper interpretation of analytical sky view factors for isotropic diffuse solar irradiance on tilted planes. *Journal of Renewable and Sustainable Energy*, vol. 9, no. 5, pp. 1-10.

Sarnavi, H. J.; Nikbakht, A. M.; Hasanpour, A.; Shahbazi, F.; Aste, N. et al. (2019): A novel stochastic energy analysis of a solar air heater: case study in solar radiation uncertainty. *Energy Systems*, vol. 10, no. 1, pp. 141-161.

Shakouri, M.; Banihashemi, S. (2019): Developing an empirical predictive energy-rating model for windows by using Artificial Neural Network. *International Journal of Green Energy*, vol. 16, no. 13, pp. 961-970.

Shamshirband, S.; Mohammadi, K.; Khorasanizadeh, H.; Yee, P. L.; Lee, M. et al. (2016): Estimating the diffuse solar radiation using a coupled support vector machine-wavelet transform model. *Renewable & Sustainable Energy Reviews*, vol. 56, pp. 428-435.

Sheridan, R. P.; Wang, W. M.; Liaw, A.; Ma, J.; Gifford, E. M. (2016): Extreme gradient boosting as a method for quantitative structure-activity relationships. *Journal of Chemical Information & Modeling*, vol. 56, no. 12, pp. 2356-2360.

Simon-Martin, M.; Alonso-Tristan, C.; Diez-Mediavilla, M. (2017): Diffuse solar irradiance estimation on building's facades: Review, classification and benchmarking of 30 models under all sky conditions. *Renewable & Sustainable Energy Reviews*, vol. 77, pp. 783-802.

Son, J.; Jung, I.; Park, K.; Han, B. (2016): Tracking-by-segmentation with online gradient boosting decision tree. *IEEE International Conference on Computer Vision*, pp. 3056-3064.

Song, Z.; Ren, Z.; Deng, Q.; Kang, X.; Zhou, M. et al. (2020): General models for estimating daily and monthly mean daily diffuse solar radiation in China's subtropical monsoon climatic zone. *Renewable Energy*, vol. 145, pp. 318-332.

Torabi, M.; Mosavi, A.; Ozturk, P.; Varkonyi-Koczy, A.; Istvan, V. (2019): A hybrid machine learning approach for daily prediction of solar radiation. *International Conference on Global Research and Education*, pp. 266-274.

Urraca, R.; Antonanzas, J.; Antonanzas-Torres, F. (2017): Estimation of daily global horizontal irradiation using extreme gradient boosting machines. *Advances in Intelligent Systems and Computing*, pp. 105-113.

Velmurugan, P.; Kalaivanan, R. (2015): Energy and exergy analysis of solar air heaters

with varied geometries. *Arabian Journal for Science & Engineering*, vol. 40, no. 4, pp. 1173-1186.

Wang, S.; Dong, P.; Tian, Y. (2017): A novel method of statistical line loss estimation for distribution feeders based on feeder cluster and modified XGBoost. *Energies*, vol. 10, no. 12, pp. 2067-2084.

Wang, S.; Li, J.; Wang, Y.; Yang, L. (2017): Radar HRRP target recognition based on gradient boosting decision tree. *International Congress on Image & Signal Processing*, pp. 1013-1017.

Wenceslas, K. Y.; Ghislain, T. (2018): Experimental validation of exergy optimization of a flat-plate solar collector in a thermosyphon solar water heater. *Arabian Journal for Science & Engineering*, pp. 1-15.

Yousuf, M. U.; Siddiqui, M.; Rehman, N. U. (2018): Solar energy potential estimation by calculating sun illumination hours and sky view factor on building rooftops using digital elevation model. *Journal of Renewable and Sustainable Energy*, vol. 10, no. 1, pp. 13703.

Zhang, W. G.; Goh, A. T. C. (2013): Multivariate adaptive regression splines for analysis of geotechnical engineering systems. *Computers and Geotechnics*, vol. 48, pp. 82-95.

Zheng, H.; Yuan, J.; Chen, L. (2017): Short-term load forecasting using emd-lstm neural networks with a xgboost algorithm for feature importance Evaluation. *Energies*, vol. 10, no. 11688, pp. 1-20.