Contents list available at IJRED website

Int. Journal of Renewable Energy Development (IJRED)

Journal homepage: https://ijred.undip.ac.id

Research Article

Machine Learning Models Based on Random Forest Feature Selection and Bayesian Optimization for Predicting Daily Global Solar Radiation

Mohamed Chaibi[a*], El Mahjoub Benghoulam[a], Lhoussaine Tarik[b], Mohamed Berrada[c], Abdellah El Hmaidi[d]

[a]*Moulay Ismail University, Faculty of Science, Department of Physics, Team of Renewable energy and energy efficiency, BP 11201, Zitoune, Meknes, Morocco*

[b]*Moulay Ismail University, Faculty of Science and Technique, Mining, Water and Environmental Engineering Laboratory, BP 509, Boutalamine, Errachidia, Morocco*

[c]*Moulay Ismail University, ENSAM, Laboratory of Mathematical and Computational Modeling, Marjane II, BP 15290, Al Mansour, 50000, Meknes, Morocco*

[d]*Moulay Ismail University, Faculty of Science, Department of Geology, Laboratory of Water Sciences and environmental engineering, BP 11201, Zitoune, Meknes, Morocco*

Abstract. Prediction of daily global solar radiation (H) with simple and highly accurate models would be beneficial for solar energy conversion systems. In this paper, we proposed a hybrid machine learning methodology integrating two feature selection methods and a Bayesian optimization algorithm to predict H in the city of Fez, Morocco. First, we identified the most significant predictors using two Random Forest methods of feature importance: Mean Decrease in Impurity (MDI) and Mean Decrease in Accuracy (MDA). Then, based on the feature selection results, ten models were developed and compared: (1) five standalone machine learning (ML) models including Classification and Regression Trees (CART), Random Forests (RF), Bagged Trees Regression (BTR), Support Vector Regression (SVR), and Multi-Layer Perceptron (MLP); and (2) the same models tuned by the Bayesian optimization (BO) algorithm: CART-BO, RF-BO, BTR-BO, SVR-BO, and MLP-BO. Both MDI and MDA techniques revealed that extraterrestrial solar radiation and sunshine duration fraction were the most influential features. The BO approach improved the predictive accuracy of MLP, CART, SVR, and BTR models and prevented the CART model from overfitting. The best improvements were obtained using the MLP model, where RMSE and MAE were reduced by 17.6% and 17.2%, respectively. Among the studied models, the SVR-BO algorithm provided the best trade-off between prediction accuracy (RMSE=0.4473kWh/m²/day, MAE=0.3381kWh/m²/day, and R^2=0.9465), stability (with a 0.0033kWh/m²/day increase in RMSE), and computational cost.

Keywords: Feature selection, Mean Decrease in Accuracy, Mean Decrease in Impurity, Bayesian optimization, Solar radiation

Article history: Received: 16th Sept 2021; Revised: 15th Nov 2021; Accepted: 28th Nov 2021; Available online: 6th Dec 2021
How to cite this article: Chaibi, M., Benghoulam, E.M., Tarik, M., Berrada, M., El-Hmaidi,A.(2022) Machine Learning Models Based on Random Forest Feature Selection and Bayesian Optimization for Predicting Daily Global Solar Radiation. *International Journal of Renewable Energy Development,* 11(1), 309-323
https://doi.org/10.14710/ijred.2022.41451

1 Introduction

Despite the global decline in energy demand due to the COVID-19 crisis, the growth of renewable energies has remained constant. Solar photovoltaic and onshore wind are now the most cost-effective options to install new electricity-generating plants in most countries (*World Energy Outlook 2020 – Analysis*, n.d.). Morocco is considered as one of the leading African countries in solar energy thanks to its sustainable policy that aims to produce 20% of its electricity from solar energy by 2030 (Merrouni *et al.*, 2018). Nowadays, there are two types of large solar energy systems: Concentrated Solar Power (CSP) and Photovoltaic (PV). The CSP systems use direct solar radiation while the PV systems use global solar radiation. Global solar radiation is measured by pyranometers commonly installed in weather stations. However, these measurements are not available for the majority of worldwide stations owing to the expensive cost of calibrating and maintaining these devices (Olatomiwa *et al.*, 2015). An alternative way to get information about H is by developing estimation models, including empirical models (Halawa *et al.*, 2014), reanalysis models (Dee *et al.*, 2011),(Gelaro *et al.*, 2017), satellite-based models (Bamehr & Sabetghadam, 2021), interpolation models (Alsamamra *et al.*, 2009), (Ruiz-Arias *et al.*, 2011), and machine learning models.

* Corresponding author: moha.chaibi@edu.umi.ac.ma

Because of their strong predictive capability and their ability to fit nonlinear data, ML models have been widely used in the literature for solar simulation (Tao *et al*., 2021). Kumar *et al*. (Kumar *et al*., 2015) compared artificial neural network (ANN) models to the regression models for estimating monthly global solar radiation. The results of this research revealed that the ANN models performed better than the regression methods. In ref (Piri *et al*., 2015), the authors compared the SVR algorithms to traditional empirical methods to predict *H* at two sites in Iran. The findings of this research indicated that SVR models were the most suitable for *H* estimation. To predict *H* in humid subtropical China, Fan *et al*. (Fan *et al*., 2018) compared SVR, Extreme Gradient Boosting (XGBoost), and four empirical models. SVR and XGBoost showed comparable prediction accuracy and outperformed the empirical models. Additionally, the XGBoost model was the most efficient in terms of stability and computational cost. Quej *et al*. (Quej *et al*., 2017a) assessed the potential of SVR, ANN, and Neuro-Fuzzy Inference System (ANFIS) models to estimate daily global solar radiation in the Yucatán Peninsula, México. They showed that the SVR model outperformed both the ANN and ANFIS techniques. In (Benali *et al*., 2019), the authors used smart persistence (SP), ANN, and RF models to forecast hourly components of solar radiation (global horizontal, beam normal, and diffuse horizontal) in Odeillo, France. They concluded that the RF method was the most robust. Benouna *et al*. (Bounoua *et al*., 2021) compared 22 empirical models, ANNs, and tree-based ensemble methods for estimating *H* in five locations in Morocco. Their results revealed the superiority of the RF model. Hassan *et al*. (Hassan *et al*., 2017) investigated the efficiency of gradient boosting, BTR, and RF models for predicting different solar radiation components over five stations in the MENA countries (Middle East and North Africa). The tree ensemble models yielded comparable prediction accuracy with ANN and SVR models while exhibiting less computational cost. Fan *et al*. (Fan *et al*., 2019a) compared the performances of 12 variants of Ångström–Prescott models and 12 machine learning algorithms for estimating daily solar radiation in different climatic zones of China. The authors showed that the ML models outperformed generally the empirical models and recommended the ANFIS model due to its high prediction capability and low computational cost.

Optimization techniques are frequently used to improve the predictive ability of ML algorithms because the performances of these models are highly sensitive to their hyperparameters. Ibrahim and Khatib (Ibrahim & Khatib, 2017) proposed a hybrid model incorporating the RF technique and the firefly algorithm (RF-FFA) for estimating hourly global solar irradiance. The novel model outperformed the conventional RF and MLP models as well as the optimized MLP-FFA model. Feng *et al*. (Feng *et al*., 2020) coupled particle swarm optimization (PSO) and extreme learning machine model (ELM) to estimate *H* in seven locations in China. The new model had better performances than ELM, SVR, Generalized Regression Neural Networks (GRNN), M5 model tree (M5tree), and autoencoder models. Lotfinejad *et al*. (Lotfinejad *et al*., 2018) combined the MLP model with BAT (BA) algorithm to estimate *H* at four sites in Iran. The proposed model performed better than ANFIS and GRNN models. Bayesian optimization (BO) is a powerful optimization technique that has gained much attention in many fields such as agriculture (Sameen *et al*., 2020), geosciences (Z. Zhang *et al*., 2021), medicine (Dhamala *et al*., 2020), and engineering (Y. Wang *et al*., 2021), (Wu *et al*., 2019), (Q. Zhang *et al*., 2020). However, there is no documented paper on solar radiation modeling using Bayesian optimization to the knowledge of the authors.

Selecting the most effective features for *H* prediction is a critical task in solar modeling. This improves the predictive efficiency of machine learning models, speeds up the training process, and eliminates redundant features (Almaraashi, 2018). Numerous studies have been conducted in recent years to identify the most significant features for global solar radiation estimation. Almaraashi (Almaraashi, 2018) integrated four feature selection methods and an ANN model to estimate daily solar radiation in eight sites in Saudi Arabia. The achieved results demonstrated that the models based on feature selection methods were better than those using all features. Alsina *et al*. (Alsina *et al*., 2016) employed the Automatic Relevance Determination method (ARD) to determine the most relevant inputs for an ANN model used to predict the monthly solar radiation in Italy. The best results were obtained with 7 inputs, namely, top of atmosphere radiation, day length, number of rainy days, rainfall, latitude, period, and altitude. Mgouchi *et al*. (El Mghouchi *et al*., 2019) identified clearness index, top of the atmosphere, and function of average ambient temperature as the optimal combination of attributes using ANNs for estimating *H* at 35 stations in Morocco and neighboring countries. In another study conducted in Morocco, Marzouq *et al*. (Marzouq *et al*., 2019) demonstrated via Evolutionary Artificial Neural Networks (EANN) that rainfall, wind direction, daily temperature gradient, and extraterrestrial solar radiation are the optimal combination of features. Using an ELM algorithm, Shamshirband *et al* (Shamshirband *et al*., (2015) demonstrated that the most influential single feature is relative sunshine duration, and that sunshine duration and the difference between maximum and minimum temperatures represent the optimal combination of two inputs. To identify the most influential input on *H* estimation, inputs. In Mashhad, Iran, Rohani *et al*. (Rohani *et al*., 2018) demonstrated that sunshine fraction duration, mean temperature, relative humidity, and extraterrestrial radiation are the most effective inputs for daily and monthly *H* prediction with the Gaussian process (GP) model. Zeng *et al*. (Zeng *et al*., 2020) showed that daily sunshine duration, daily maximum land surface temperature, and day of the year are the most important attributes for *H* modeling across China using the RF model. In ref (Sun *et al*., 2016), the authors employed the same approach and demonstrated that sunshine duration is the most effective feature in estimating solar radiation. Table 1 summarizes some of the feature selection approaches used in the literature.

Table 1
Summary of some of the features selection methods used in the literature

References	Location	Feature selection method	Comments
(Almaraashi, 2018)	8 sites, Saudi Arabia	ReliefF algorithm, Monte Carlo uninformative variable elimination algorithm,random-frog algorithm, and Laplacian score algorithm.	Models based on feature selection provided best performances than models with all features
(Alsina *et al*., 2016)	45 sites, Italy	ARD	Best results were obtained using: top of atmosphere radiation, day length, number of rainy days, rainfall, latitude, period time, and altitude
(Mgouchi *et al*., 2019)	27 stations in Morocco and 8 in neighboring countries	ANNs	The optimal combination was: clearness index, top of the atmosphere, and function of average ambient
(Marzouq *et al*., 2019)	1site, Morocco	EANN	The best combination of features was: rainfall, wind direction, daily temperature gradient, and extraterrestrial solar radiation
(Shamshirband *et al*., 2015)	1 site, Iran	ELM	Sunshine duration is the most important attribute while sunshine duration and the difference between the maximum and minimum temperatures represent is the best combination of two inputs
(Rohani *et al*., 2018)	1 site, Iran	GP	The most important features were: sunshine fraction duration, mean temperature, relative humidity, and extraterrestrial radiation
(Zeng *et al*., 2020)	130 sites, China	RF	Daily sunshine duration, daily maximum land surface temperature, and day of the year were the most impactful features
(Sun *et al*., 2016)	3 sites, China	RF	Daily sunshine duration is the most important input variable
(Yadav *et al*., 2014)	26 sites, India	Waikato environment for knowledge analysis	The most relevant input variables were: average temperature, maximum temperature, minimum temperature, altitude, and sunshine hours

As this brief review indicates, the BO technique has not been applied in global solar radiation modeling and the RF's model application as a feature importance technique is still limited. Furthermore, the two feature importance techniques, MDI and MDA, have not been compared yet. Hence, this study aims first to select the most important features for estimating H in the city of Fez, Morocco using MDA and MDI techniques. Second, to optimize 5 ML models including CART, RF, BTR, SVR, and MLP via the Bayesian optimization algorithm. Finally, to compare the best models in terms of predictive accuracy, stability, and computational cost. The methodology followed in this study is reported in Fig.1.

The rest of this paper is structured as follows: Section 2 describes different models and techniques used in this paper. It also presents the study area, data preprocessing, and evaluation criteria. The main results were presented and discussed in Section 3. Finally, Section 4 provides the conclusions with future work.

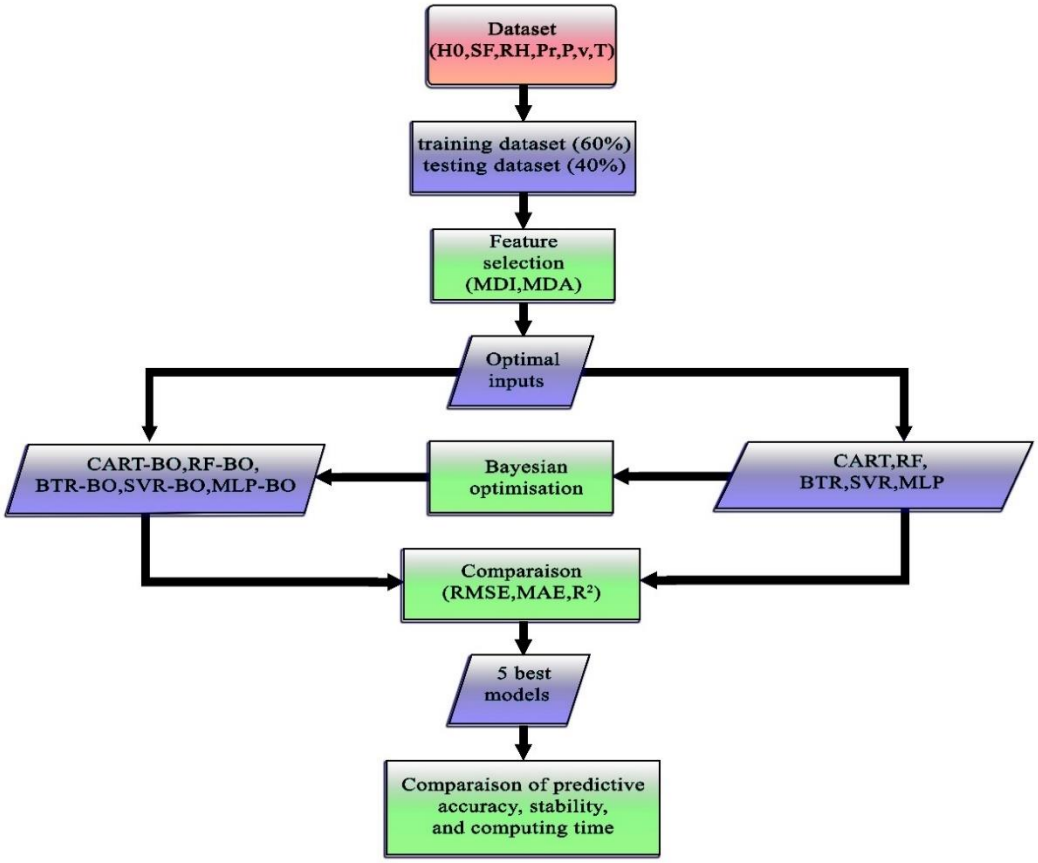

Fig. 1 Flowchart of the methodology used in this study

2 Materials and Methods

2.1 Predictive models

2.1.1 Classification and Regression Trees (CART)

The CART model is a kind of decision tree introduced by Breiman *et al.* in 1984 (Breiman, 2017). This model is constructed by partitioning the data space recursively and fitting a simple prediction model to each partition. Each decision tree includes a root node (top of the tree), multiple branch nodes (internal nodes), and several leaf nodes (terminal nodes) (Z. Wang *et al.*, 2018).

Let us denote a set of observed data S_n={X_i, Y_i} i = 1: n

where $Y \epsilon \mathbb{R}$ is the output variable, $X \epsilon \mathbb{R}^m$ is the input vector containing m features, and n is the number of observations.

The first step to construct a CART model is to split the root node p into two different children as:

$$\{X^j < d\} U \{X^j > d\} \quad (1)$$

where $j \in \{1, \dots, m\}$ and $d \epsilon \mathbb{R}$.

The best couple (j, d) is obtained by minimizing the child node variance function defined by:

$$V(p) = \sum_i (Y_i - \bar{Y}_p)^2 \quad (2)$$

where $\bar{Y}_p$ is the mean of the scalars Y_i present in the node p (Lahouar & Slama, 2015).

The child's nodes are then divided in the same manner. This splitting process will be repeated recursively until a predefined stopping criterion is met, for instance, the minimum number of samples for the node split or the minimum number of samples in a leaf. Finally, a CART model $\hat{h}(X, S_n)$ is built over S_n. Fig.2 illustrates an example of a CART model with its partitioning space.

The CART model contains several hyperparameters. In this study, we focus on the maximum depth of the tree (max_depth) and the number of features to consider when looking for the best split (max_features).

Although CART models are fast and interpretable, they are unstable; a minor change made in the input data can lead to a significant influence on the output value [12]. To overcome this drawback and to get high prediction accuracy, ensemble methods such as Bagged Trees, Boosted Trees, and Random Forests have been proposed.

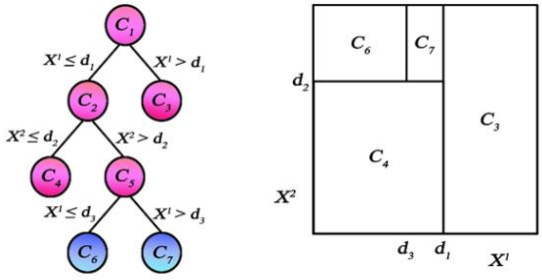

Fig. 2 An example of a CART model with its partitioning space

The BTR algorithm proposed by Breiman in 1996 (Breiman, 1996) uses the bagging (or bootstrap aggregation) process to create a set of decision trees (Li *et al.*, 2018). The bagging algorithm generates several bootstrap samples ($S_n^{\theta_1}$,….$S_n^{\theta_q}$) from the original training data S_n by randomly selecting n observations with replacement from S_n. These samples are then used as new datasets to train q independent trees $\hat{h} = \left(X, S_n^{\theta_1}\right) \ldots, \hat{h} = \left(X, S_n^{\theta_q}\right)$ (Li *et al.*, 2018). Finally, the outputs of all decision trees are averaged (Lahouar & Slama, 2015):

$$\hat{Y} = \frac{1}{q}\sum_{l=1}^{q} \hat{h}(X, S_n^{\theta_i}) \quad (3)$$

For the BTR algorithm, the main tuning hyperparameters include (1) the number of trees in the model (n_estimators) and (2) the maximum depth of the tree (max_depth). The structure of a BTR model is illustrated in Fig.3.

2.1.2 Random Forest (RF)

The RF algorithm proposed by Breiman in 2001 (Breiman, 2001) is a variant of the BTR model. Unlike BTR, RF selects a subset of features to create non-correlated trees that reach the averaging stage (Zhou, 2012). The main hyperparameters for this model are (1) the number of trees in the forest (n_estimators), (2) the maximum depth of the tree (max_depth), and (3) the number of features to consider when looking for the best split (max_features).

Besides its good prediction accuracy, RF also offers two methods for feature selection: Mean Decrease in Impurity (MDI) and Mean decrease in Accuracy (MDA).

2.1.3 Mean Decrease in Impurity (MDI)

In the case of a CART model, the MDI score for a variable X^j is calculated by summing decreases in node impurities (variances) during data partitioning using the feature X^j. The decrease in impurity is defined as (Scornet, 2020):

$$\Delta V(t) = V(t) - p_L V(t_L) - p_R V(t_R) \quad (4)$$

where t_L and t_R represent two child nodes generated when partitioning the data at node t, $p_L = N_{t_L}/N_t$ and $p_R = N_{t_R}/N_t$ are the proportion of data reaching the children nodes t_L and t_R, respectively.

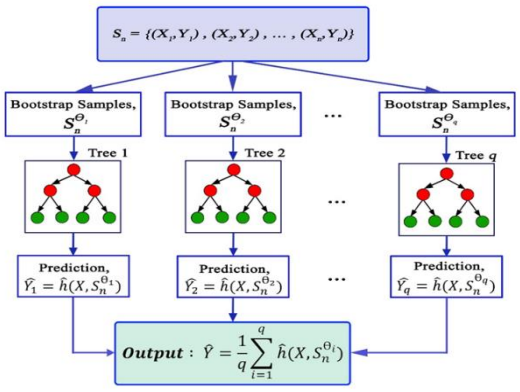

Fig. 3 Structure of BTR model

In the case of the RF model, the MDI score is obtained by averaging the scores of all q trees in the forest

$$MDI = \frac{1}{q}\sum_q \sum_{t \in q} p(t)\ \Delta V(t) \quad (5)$$

where $p(t) = N_t/N$ represents the proportion of samples reaching node t.

2.1.4 Mean Decrease in Accuracy (MDA)

MDA is a feature importance technique introduced first by Breiman for Random Forests and generalized by Fisher *et al.* in 2018 (Fisher *et al.*, 2018) for all kinds of machine learning models. It measures the decrease of accuracy when permuting the values of a variable X^j. The MDA score is calculated as (Molnar, 2020):

$$MDA = err^{perm} - err^{orig} \quad (6)$$

where err^{orig} is the original model error and $error^{perm}$ is the error obtained after permuting the values of X^j.

The error utilized in this study is the mean absolute error (MAE) calculated by equation 15.

2.1.5 Support Vector Regression (SVR)

SVRs are powerful ML derived from statistical learning theory and the structural risk minimization principle (Chen *et al.*, 2013). The SVRs transform the non-linear relationship between features and the outcome in the original space into a linear regression in a new higher dimensional feature space, implicitly using the kernel trick (Quej *et al.*, 2017b). A detailed description of the SVR model is given in (Vapnik, 2013). The hyperparameters of the SVR model are (1) the regularization parameter (C), (2) the width of the tube around the estimated function (epsilon), and (3) the kernel function (kernel).

2.1.6 Multilayer Perceptron (MLP)

MLP models are a kind of feedforward ANNs that are inspired by the functioning of biological neurons. It is composed of an input layer, an output layer, and one or several hidden layers. Fig.4 illustrates the structure of an MLP network with one hidden layer. In the first stage, the model propagates forward the input data from the input layer through the hidden layers to the output layer. In the second stage, the error is propagated back to the input layer. A learning algorithm is used to adjust the network's weights and bias until the error is minimized (R. Wang *et al.*, 2019). More details about this model can be found in ref (Hastie *et al.*, 2009). The tuned hyperparameters for an MLP model with one hidden layer are: (1) the activation function for the hidden layer (activation), (2) the learning algorithm (solver), and (3) the number of neurons in the hidden layer (hidden_layers_sizes).

2.1.7 Bayesian Optimization (BO)

BO is an effective strategy for optimizing unknown black box objectives functions (Snoek *et al.*, 2012). The key components in the BO algorithm are a Gaussian Process (GP) model of the unknown objective function $f(x)$, a Bayesian procedure for updating the GP model at each new evaluation of $f(x)$, and an acquisition function $a(x)$ that determines the next point to evaluate.

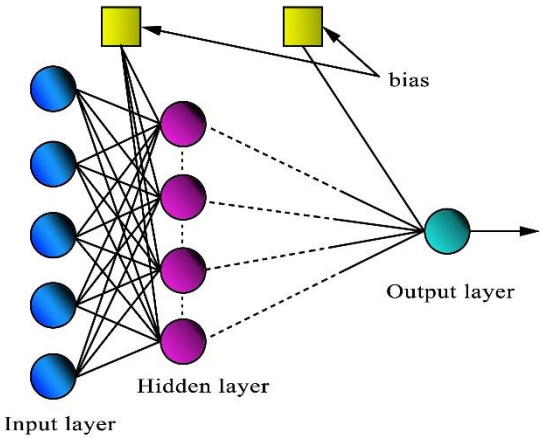

Fig.4 Structure of an MLP model with one hidden layer

GP generalizes the multivariate normal distribution over functions; it is defined by its mean function $m(x)$ and covariance matrix $k(x,x')$

$$GP \sim N(m(x), k(x,x')) \quad (7)$$

where N denotes the standard normal distribution (Shi & Choi, 2011).

Let D=$\{x_i, y_i\}$ $i = 1{:}\,t$ denote a set of data

A GP model for this data set can be specified as follows $f(x_i) = y_i + \varepsilon_i$ with $\varepsilon_i = N(0, \sigma^2)$ is Gaussian noise. We assume a zero-mean function in the GP prior and we choose ARD Matern 5/2 Kernel as a covariance function

$$k(x_i, x_j) = {\sigma_f}^2 \left(1 + \frac{\sqrt{5}r^2}{\sigma_l} + \frac{5r^2}{3{\sigma_l}^2}\right) \exp\left(-\frac{\sqrt{5}r}{\sigma_l}\right) \quad (8)$$

where $r = \sqrt{(x_i - x_j)^T (x_i - x_j)}$ is the Euclidean distance between x_i and x_j, σ_l is the characteristic length scale, and σ_f is the signal standard deviation (Snoek *et al*., 2012).

The function values $y_{1:t}$ jointly follow a multivariate Gaussian distribution as $y_{1:t} \sim N(0, K)$, where covariance matrix K is given as:

$$K = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_t) \\ \vdots & \ddots & \vdots \\ k(x_t, x_1) & \cdots & k(x_t, x_t) \end{bmatrix} \quad (9)$$

The predictive mean and the variance for or a new data point x_{t+1} are (Joy *et al*., 2019):

$$m(x_{t+1}) = k^T [K + \sigma^2 I]^{-1} y_{1:t} \quad (10)$$

$$\sigma^2(x_{t+1}) = k(x_{t+1}, x_{t+1}) - k^T [K + \sigma^2 I]^{-1} k \quad (11)$$

The next step in the BO algorithm is to choose the next x^* to evaluate by maximizing the acquisition function $a(x)$. This function balances exploration against exploitation. Exploration attempts to improve the model in the less explored regions of the search space, while exploitation favors parts that the model predicts as promising (Calandra *et al*., 2016). In this paper, we used the Expected Improvement acquisition function given by (Cheng *et al*., 2019):

$$EI(x) = \begin{cases} \big(m(x) - f(x^*)\big)\emptyset(z) + \sigma(x)\varphi(z), if\ \ \sigma(x) > 0 \\ 0,\ if\ \sigma(x) = 0 \end{cases} \quad (12)$$

where $z = \frac{m(x) - f(x^*)}{\sigma(x)}$, $\emptyset(.)$, and $\varphi(.)$ are the cumulative density function (CDF) and the probability density function (PDF) of standard normal distribution, respectively.

2.2 Study area and data processing

2.2.1 Case study and data collection

The data used in this study were measured from the 1st of January 2016 to the 31st of December 2017 at a meteorological station in Fez (latitude 33°55'58" N, longitude 4°58'30" W, altitude 571.3m). Seven parameters were included, namely global solar radiation (H), sunshine duration (N), average temperature (T), atmospheric pressure (P), relative humidity (RH), precipitations (P_r), and wind speed (v). The database was supported by the daily calculated extraterrestrial solar radiation (H_0) and the daily sunshine duration fraction (SF). The detailed equations of these two quantities can be found in (Kalogirou, 2013). Fig.5 presents the boxplots of the daily variables used in this study. The dataset contains some incorrect and missing values that must be removed. For this aim, we excluded all daily clearness index ($K_t = H/H_0$) and SF values that were outside of the ranges $0.015 < K_t < 1$ and $0 \leq \mathrm{SF} \leq 1$, respectively (Quej *et al*., 2017a), (Hassan *et al*., 2017). We found five days with missing H values and four days with SF incorrect values among the 731 daily data used in this study.

2.2.2 Data preprocessing and evaluation criteria

The dataset was randomly divided into two groups (60% for training and 40% for testing) for constructing the predictive ML models. The models: CART, RF, and BTR, do not necessitate data normalization. However, preprocessing of data is required for SVR and ANN models. The normalized value X_{norm} of an instance of the dataset is calculated as (R. Wang *et al*., 2019):

$$X_{norm} = \frac{X_i - X_{i,min}}{X_{i,max} - X_{i,min}} \quad (13)$$

where X_i, $X_{i,min}$, $X_{i,max}$ denote the real, the minimum, and maximum values, respectively.

To tune the hyperparameters of ML algorithms with BO, we used the 5-fold cross-validation technique, which divides the training dataset into five groups. Four groups were used to train the ML methods and the remaining group was used to validate them. This process is repeated 5 times.

All the simulations were written with Python 3.8 language on a computer with a 2.53 GHz processor and 16 GB RAM. The ML models were developed with scikit-learn library (Pedregosa *et al*., 2011), and the BO algorithm with scikit-optimize library.

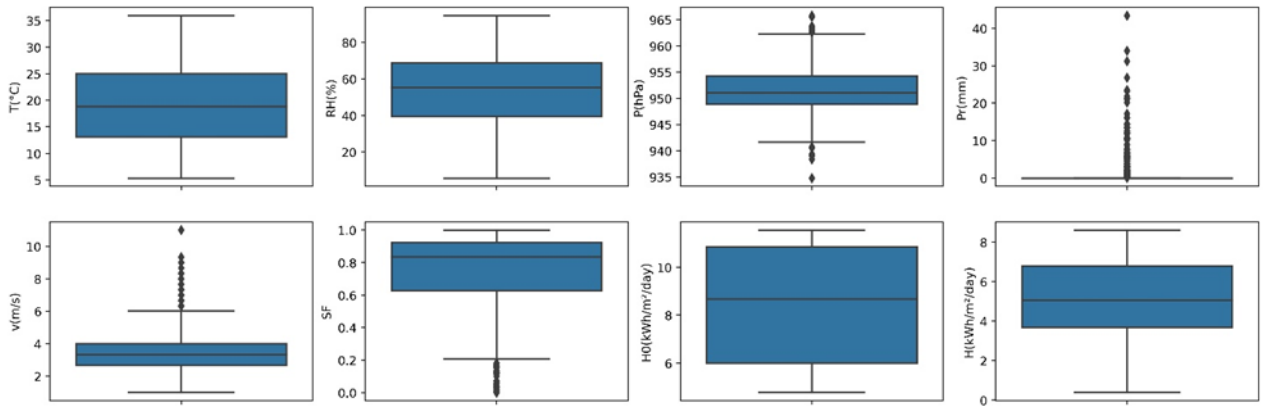

Fig. 5 Boxplots of the daily parameters used in this study

We used three statistical indicators to assess the predictive accuracy of the ML models: root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) (Ahmad *et al.*, 2018):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(H_{i,c}-H_{i,m})^2}{n}} \quad (14)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|H_{i,c} - H_{i,m}\right)| \quad (15)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(H_{i,c}-H_{i,m})^2}{\sum_{i=1}^{n}(H_{i,m}-H_{m,avg})^2} \quad (16)$$

where n is the number of observations, $H_{i,c}$ is the calculated solar radiation, $H_{i,m}$ denotes the measured solar radiation, and $H_{m,avg}$ is the mean of the measured values.

A model is more accurate when RMSE and MAE are close to 0 and R^2 is close to 1.

3 Results and Discussion

3.1. Feature selection and performance analysis of RF models

In this section, we used and compared MDA and MDI techniques to measure the importance of the input variables and to identify the most influential of them. Fig.6(a) and Fig.6(b) represent the MDA and MDI scores of the seven investigated predictors. Both methods produced the same order of the top five features. Among all variables, H_0 and SF were found to be the most important, using the two techniques. H_0 is 1.58 and 1.41 more important than SF, using the MDI and MDA methods, respectively. These two variables are highly correlated to H and are widely used in solar modeling because the solar radiation reaching the ground is the fraction of H_0 that passes through the atmosphere and SF is an indirect index of the cloudiness of a site (Paulescu *et al.*, 2016). T and RH are the next two most relevant features; they showed a small impact in H predicting compared to H_0 and SF. The MDI score of H_0 is 67.4 and 88.2 times the MDI score of T and RH, respectively. While the MDA score of H_0 is 56 and 88.5 times that of T and RH, respectively. The remaining features had insignificant scores, basically when using the MDA technique.

To evaluate the efficacy of feature importance techniques and to consider the interaction of features, we compared the RF model with all inputs to several RF models with different combinations. Table 2 summarizes the obtained results. According to this table, the performances of the RF models varied significantly under various input combinations.

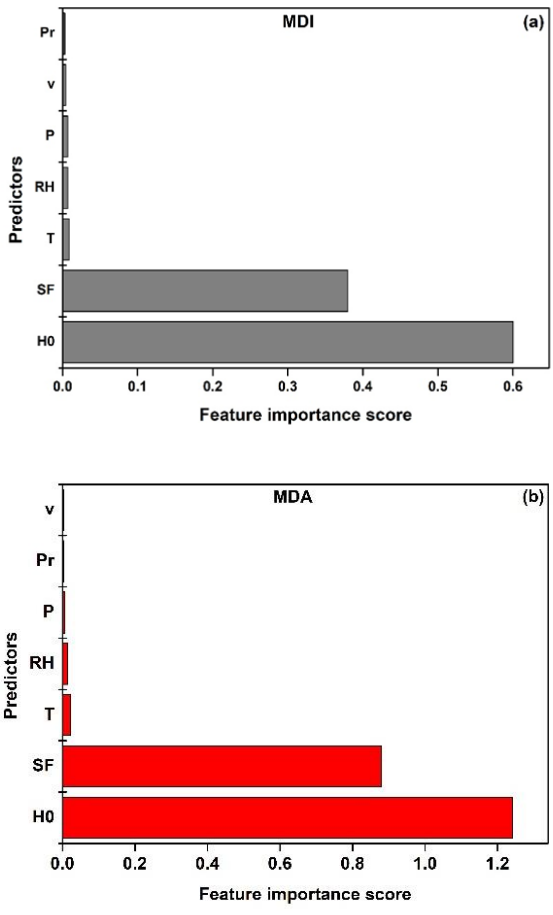

Fig. 6 Variable importance input predictors(a) MDI method, (b) MDA method

The RF model with the complete features did not achieve the best performances because some of them were redundant. All models that used the combination of H_0 and *SF* performed better than the models without it. For instance, the model with only H_0 and *SF* outperformed all models without this association. These results confirmed that these two features are the most important for *H* estimating. Among the sixteen developed soft computing techniques, the model incorporating H_0, *SF,* and *T* showed the best accuracy during the testing phase with R^2=0.9462, RMSE=0.4488kWh/m²/day, and MAE=0.3448 kWh/m²/day. It is followed by the model based on the combination *H*, *SF,* and *RH*. These results showed the benefits of using feature selection to obtain simple models without compromising prediction quality.

3.1 Bayesian Optimization results

Bayesian optimization (BO) aims at improving the performances of five standalone ML models: CART, RF, BTR, SVR, and MLP using the three optimal inputs H_0, *SF* and *T*. The hyperparameters of each model were optimized over the 5-fold-cross-validation technique. Table 3 depicts the range of hyperparameters, and their final values obtained by the BO algorithm.

Table 2
Statistical results obtained for the RF models with different features (bold represents the best result)

Input variables	R2		RMSE (kWh/m²/day)		MAE (kWh/m²/day)	
	Training	Testing	Training	Testing	Training	Testing
All	0.9919	0.9406	0.1769	0.4715	0.1299	0.3586
H_0, SF, T, RH	0.9921	0.9446	0.1750	0.4554	0.1281	0.3471
H_0, T, RH, P	0.9745	0.7573	0.3138	0.9529	0.2239	0.6998
SF, T, RH, P	0.9599	0.6839	0.3932	1.0874	0.3092	0.8398
T, RH, P	0.9412	0.5794	0.4765	1.2544	0.3658	1.0094
H_0, SF, RH	0.9920	0.9459	0.1757	0.4501	0.1303	0.3465
H_0, SF, T	0.9928	**0.9462**	0.1667	**0.4488**	0.1271	**0.3448**
H_0, T, RH	0.9674	0.7149	0.3547	1.0327	0.2613	0.7777
SF, T, RH	0.9614	0.6664	0.3862	1.1172	0.3015	0.8543
T, RH, P	0.9271	0.5329	0.5303	1.3220	0.3980	1.0277
SF, T	0.9543	0.6160	0.4200	1.1986	0.3287	0.9398
SF, RH	0.9357	0.5547	0.4982	1.2907	0.3911	1.0074
H_0, T	0.9421	0.5467	0.4728	1.3023	0.3406	0.9427
H_0, RH	0.9617	0.6887	0.3843	1.0792	0.2869	0.8296
H_0, SF	0.9909	0.9381	0.1877	0.4810	0.1393	0.3668
T, RH	0.9229	0.4176	0.5454	1.4761	0.4122	1.1508

Table 3
Range of hyperparameters and their final values obtained by the BO algorithm

Models	Range of hyperparameters	Final values
CART-BO	max_depth=2-14	6
	max_features =1-3	3
RF-BO	n_estimators=10-500	480
	max_depth=2-14	10
	max_features =1-3	2
BTR-BO	n_estimators=10-500	376
	max_depth=2-14	8
SVR-BO	C=1-100	11.30
	epsilon=10^{-3}-0.3	0.1166
	kernel=linear- poly-rbf	rbf
MLP-BO	activation=tanh-sigmoid-relu	relu
	solver=bgfs-Adam	bgfs
	hidden_layers_sizes=2-100	16

To demonstrate the robustness of the BO approach, we compared the five models with default hyperparameters to their corresponding tuned models. Table 3 illustrates the obtained results. The BO approach improved the predictive accuracy of MLP, CART, SVR, and BTR models. Besides, the BO algorithm prevented the CART model from overfitting. On the contrary, the RF model was less responsive to the hyperparameters changes. This was generally consistent with the findings of Wang et. al (Y. Wang *et al*., 2021). The best improvements were obtained using the MLP model, where RMSE and MAE were reduced by 17.6% and 17.2%, respectively. The BO algorithm decreased the RMSE-MAE of the CART, BTR, and SVR models by 3-1.42%, 2.80-4.38%, and 5.4-5.9%, respectively.

Table 4
Statistical indicators for models with defaults and tuned hyperparameters (bold represents the best result)

Models	R^2		RMSE (kWh/m²/day)		MAE (kWh/m²/day)	
	Training	Testing	Training	Testing	Training	Testing
CART	1.0000	0.8942	0.0000	0.6293	0.0000	0.4566
CART-BO	0.9326	0.9006	0.4199	0.6099	0.3190	0.4501
RF	0.9928	0.9462	0.1667	0.4488	0.1271	0.3448
RF-BO	0.9923	0.9463	0.1725	0.4484	0.1294	0.3446
BTR	0.9896	0.9423	0.2008	0.4647	0.1380	0.3651
BTR-BO	0.9925	0.9455	0.1704	0.4517	0.1280	0.3491
SVR	0.9526	0.9406	0.4276	0.4715	0.3135	0.3581
SVR-BO	0.9512	**0.9465**	0.4340	**0.4473**	0.3045	**0.3381**
MLP	0.9257	0.9191	0.5356	0.5503	0.4116	0.4244
MLP-BO	0.9638	0.9451	0.3738	0.4534	0.2785	0.3511

3.2 Comparison of the optimized models

3.2.1 Comparison of prediction accuracy and stability

We can see from table 4 that all optimized models provided low values of RMSE and MAE and high values of R^2. We can also see that the RMSE and MAE values increased relatively in the testing phase while R^2 decreased. Fig.7 depicts the scatter plots between the measured and estimated global solar radiation values of the five tuned models for both training and testing phases. As shown in Fig. 7 (a), the two ensemble methods BTR-BO (R^2=0.9925) and RF-BO (0.9923) demonstrated the best correlations during the training process, while the CART-BO model showed the worst correlation (R^2=0.9326). In the testing phase (Fig.7 (b)), the CART-BO model produced again more scattered estimates than the other models with a value of R^2=0.9006. The four techniques, RF-BO, BTR-BO, SVR-BO, and MLP-BO showed close correlations with a slight superiority of the SVR-BO algorithm (R^2=0.9465).

Fig.8 shows the RMSE and MAE metrics in the testing phase for the five studied models. We can observe from this figure that the CART-BO model provided the worst performances with RMSE=0.6099kWh/m²/day and MAE= 0.4501kWh/m²/day. On the other hand, the SVR-BO algorithm offered the best results with RMSE=0.4473kWh/m²/day and MAE=0.3381 kWh/m²/day. This can be seen graphically in Fig.9, where the SVR-BO followed the daily irregular variation of the measured solar radiation in both training and testing stages

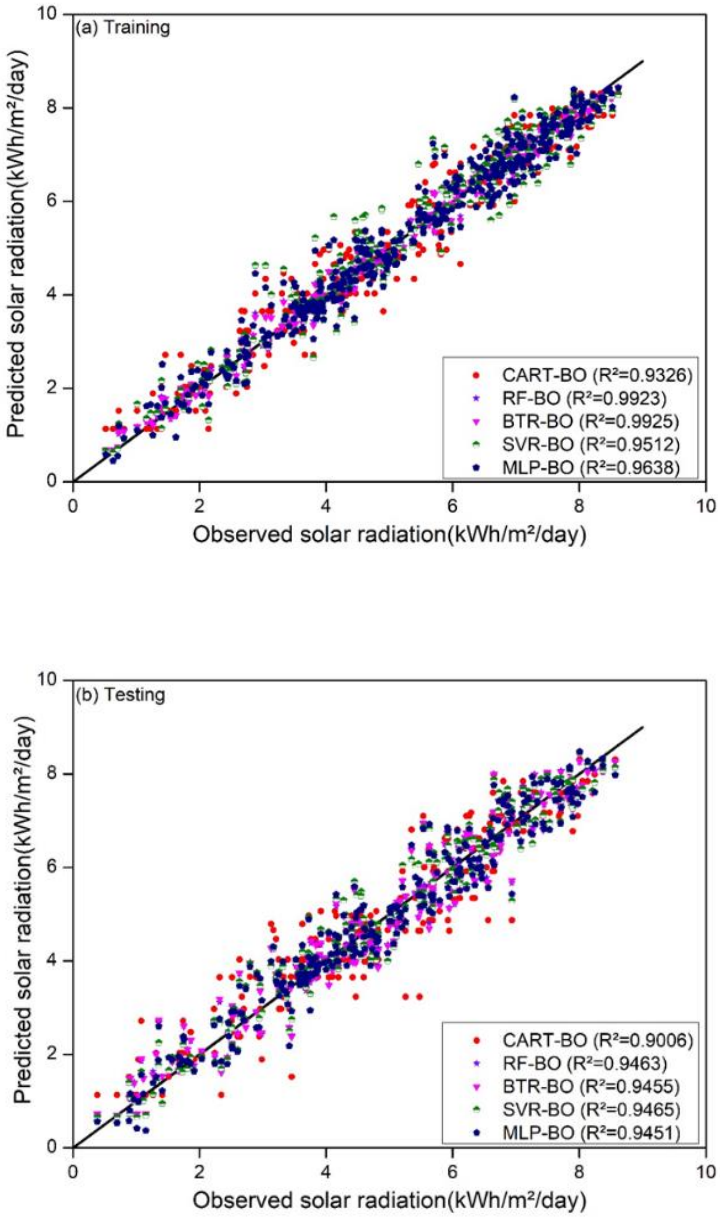

Fig. 7 Scatter plot for the 5 models between predicted and observed solar irradiation for (a) training dataset, (b) testing dataset

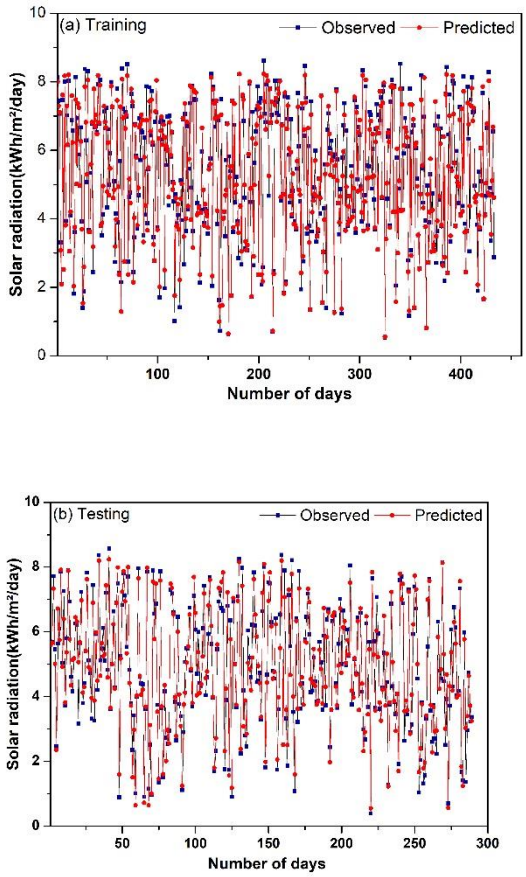

Fig. 9 Day by day comparison between the observed data and the predicted by the SVR-BO model irradiation for (a) training dataset, (b) testing dataset

The three techniques, MLP-BO, RF-BO, and BTR-BO gave close prediction performances compared to the SVR-BO model. Both the RF-BO and BTR-BO models outperformed their base learner CART; this demonstrates the benefit of using the ensemble strategy. Moreover, because of incorporating randomized feature selection, the RF-BO model was slightly better than the BTR-BO model.

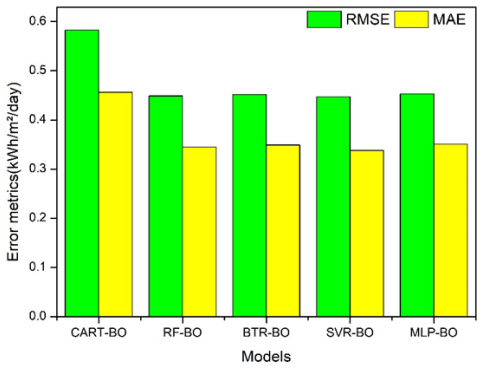

Fig.8 Metric errors (RMSE and MAE) obtained by the five tuned models.

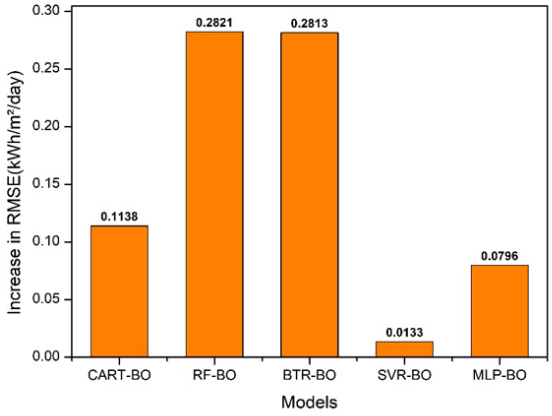

Fig.10 Increase in RMSE for the five studied models

Table 5 compares the statistical results of the current study to some previously conducted in the literature for daily global solar radiation estimation based on the same metrics. As can be seen from this table, the SVR-BO achieved good prediction accuracy compared to those developed in the literature.

When comparing machine learning models, the stability of these models is also an important issue to consider (Hassan *et al.*, 2017). Stability is measured by the increased RMSE between the training and the testing stages. According to Fig.10, the SVR-BO model was the most stable, where the RMSE increased only by 0.0133 kWh/m²/day. This is in agreement with the findings of Hassan *et al.* (Hassan *et al.*, 2017) and Fan *et al.* (Fan *et al.*, 2018). MLP-BO and CART-BO were the next two stable models. In contrast, the least stable models were the RF-BO and BTR-BO; these two methods exhibited the largest increase in RMSE with values of 0.2813 kWh/m²/day and 0.2821 kWh/m²/day, respectively.

3.2.2 Comparison of computational time

Prediction accuracy and stability are the critical factors to consider when using machine learning models. Nonetheless, the computational time is also a factor to take into account, particularly when a large amount of data is available (Fan *et al.*, 2019b). Fig.11 presents the average computational time of the five models in the training and testing phases. The results showed that the testing time was always less than the training time. The results also revealed that the average training and testing time consumed by the RF-BO and BTR-BO was much longer than those of the other algorithms because of the large number of trees employed in these models. The RF-BO algorithm was the slowest during the training phase, followed by the BTR-BO algorithm. In contrast, the CART-BO algorithm was the fastest. The RF-BO cost was approximately 1.64, 4.22, 44.36, and 173 much higher than the computational cost of the BTR-BO, MLP-BO, SVR-BO, and CART-BO models, respectively.

In the testing phase, the RF-BO model again exhibited the highest computational time. The cost of RF-BO is 1.4 times that of the BTR-BO. On the other hand, the CART-BO, SVR-BO, and MLP-BO models had insignificant computational time.

Table5
Metric comparison of present paper with the literature studies in the prediction of daily global solar radiation

Reference	Location	Predictive models	Best model	Evaluation metrics		
				R^2	RMSE (kWh/m²/day)	MAE (kWh/m²/day)
(Moreno *et al.*, 2011)	40 sites in Spain	ANN, Empirical, Kernel Ridge Regression	ANN	0.8600	0.8800	0.6500
(L. Wang *et al.*, 2017)	21 sites in China	ANFIS, M5 tree, Empirical	ANFIS	0.9100	0.5700	
(Antonopoulos *et al.*, 2019)	2 sites in Greece	ANN, Empirical, Multi-Linear Regression	ANN	0.8840	0.8810	
(Hassan *et al.*, 2017)	5 sites in MENA countries	CART, BTR, RF, ANN, Boosted Trees	SVR	0.9860	0.2200	
(Ağbulut *et al.*, 2021)	4 sites in Turkey	SVR, ANN, Kernel and Nearest-Neighbor, Deep Learning	ANN	0.9320	0.6000	
(Fan *et al.*, 2018)	3 sites in China	XGBoost, SVR, Empirical	SVR	0.7760	1.002	0.7291
(Piri *et al.*, 2015)	2 sites in Iran	SVR, Empirical	SVR	0.9330	0.4515	
Present paper	1 site in Morocco	SVR, SVR-BO, CART, CART-BO, MLP, MLP-BO, BTR, BTR-BO, RF, RF-BO	SVR-BO	0.9465	0.4473	0.3381

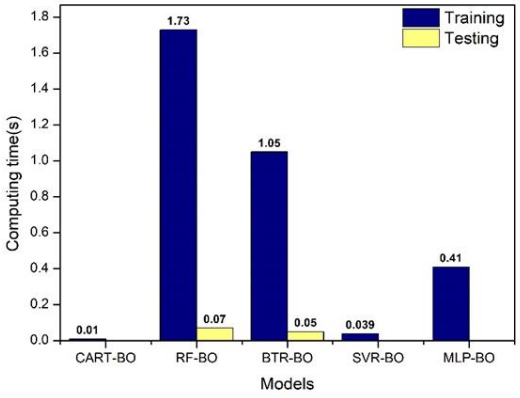

Fig.11 The computational time of the five models for both training and testing phases.

The results of this study established the advantages of combining RF feature selection methods and the BO algorithm to obtain accurate, stable, and fast models. Particularly, the SVR-BO that offered the best combination of the three criteria: accuracy, stability, and computational cost. Nevertheless, our study suffers from the fact that data are restricted to one geographical site. As a result, additional research should be undertaken using data collected from a variety of locations.

4 Conclusion

In this paper, we proposed an ML methodology based on two RF feature selection methods and Bayesian optimization for H estimating in Fez city, Morocco. The results indicated that the proposed approach is highly recommended to obtain simple and accurate models for solar radiation prediction. H_0 and SF had the highest impact in solar radiation modeling and the combination H_0, SF and T yielded the best prediction accuracy. The results confirmed that the optimized ML models outperformed their corresponding standalone models. Among the proposed optimized models, the SVR-BO model provided the best trade-off between prediction accuracy, stability, and computational cost. The two ensemble models, RF-BO and BTR-BO outperformed the CART-BO and MLP-BO models. However, these models were unstable and necessitated high computational costs in both the training and testing stages. Additional assessment is required to determine the efficiency of this methodology in different regions and periods (hours-months) using other machine learning models.

Acknowledgments

The authors would like to thank the Moroccan Department of National Meteorology for providing the data used in this study.

Author Contributions: M.C., E.M.B., and L.T: Conceptualization, methodology, formal analysis, writing—original draft, E.M.B and A.E.H; supervision, resources, project administration, M.C.and M.B. writing—review and editing, project administration, validation. All authors have read and agreed to the published version of the manuscript.

Funding: The author(s) received no financial support for the research, authorship, and/or publication of this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

Ağbulut, Ü., Gürel, A. E., & Biçen, Y. (2021). Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. Renewable and Sustainable Energy Reviews, 135, 110114; doi.org/10.1016/j.rser.2020.110114

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2018). Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression. Energy, 164, 465–474; doi:org/10.1016/j.energy.2018.08.207

Almaraashi, M. (2018). Investigating the impact of feature selection on the prediction of solar radiation in different locations in Saudi Arabia. Applied Soft Computing, 66, 250–263; doi:org/10.1016/j.asoc.2018.02.029

Alsamamra, H., Ruiz-Arias, J. A., Pozo-Vázquez, D., & Tovar-Pescador, J. (2009). A comparative study of ordinary and residual kriging techniques for mapping global solar radiation over southern Spain. Agricultural and Forest Meteorology, 149(8), 1343–1357; doi:org/10.1016/j.agrformet.2009.03.005

Alsina, E. F., Bortolini, M., Gamberi, M., & Regattieri, A. (2016). Artificial neural network optimisation for monthly average daily global solar radiation prediction. Energy Conversion and Management, 120, 320–329; doi:org/10.1016/j.enconman.2016.04.101

Antonopoulos, V. Z., Papamichail, D. M., Aschonitis, V. G., & Antonopoulos, A. V. (2019). Solar radiation estimation methods using ANN and empirical models. Computers and Electronics in Agriculture, 160, 160–167; doi.org/10.1016/j.compag.2019.03.022

Bamehr, S., & Sabetghadam, S. (2021). Estimation of global solar radiation data based on satellite-derived atmospheric parameters over the urban area of Mashhad, Iran. Environmental Science and Pollution Research, 28(6), 7167–7179; doi:org/10.1007/s11356-020-11003-8

Benali, L., Notton, G., Fouilloy, A., Voyant, C., & Dizene, R. (2019). Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components. Renewable Energy, 132, 871–884; doi:org/10.1016/j.renene.2018.08.044

Bounoua, Z., Chahidi, L. O., & Mechaqrane, A. (2021). Estimation of daily global solar radiation using empirical and machine-learning methods: A case study of five Moroccan locations. Sustainable Materials and Technologies, 28, e00261; doi.org/10.1016/j.susmat.2021.e00261

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140; doi:org/10.1007/BF00058655

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32; doi:org/10.1023/A:1010933404324

Breiman, L. (2017). Classification and regression trees. Routledge

Calandra, R., Seyfarth, A., Peters, J., & Deisenroth, M. P. (2016). Bayesian optimization for learning gaits under uncertainty. Annals of Mathematics and Artificial Intelligence, 76(1–2), 5–23; doi:org/10.1007/s10472-015-9463-9

Chen, J.-L., Li, G.-S., & Wu, S.-J. (2013). Assessing the potential of support vector machine for estimating daily solar radiation using sunshine duration. Energy Conversion and Management, 75, 311–318; doi:org/10.1016/j.enconman.2013.06.034

Cheng, H., Ding, X., Zhou, W., & Ding, R. (2019). A hybrid electricity price forecasting model with Bayesian optimization for German energy exchange. International Journal of Electrical Power & Energy Systems, 110, 653–666; doi:org/10.1016/j.ijepes.2019.03.056

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., & Bauer, d P. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. Quarterly Journal of the Royal Meteorological Society, 137(656), 553–597; doi:org/10.1002/qj.828

Dhamala, J., Bajracharya, P., Arevalo, H. J., Sapp, J. L., Horácek, B. M., Wu, K. C., Trayanova, N. A., & Wang, L. (2020). Embedding high-dimensional Bayesian optimization via generative modeling: Parameter personalization of cardiac electrophysiological models. Medical Image Analysis, 62, 101670; doi:org/10.1016/j.media.2020.101670

El Mghouchi, Y., Chham, E., Zemmouri, E. M., & El Bouardi, A. (2019). Assessment of different combinations of meteorological parameters for predicting daily global solar radiation using artificial neural networks. Building and Environment, 149, 607–622; org/10.1016/j.buildenv.2018.12.055

Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X., Lu, X., & Xiang, Y. (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, 164, 102–111; doi:org/10.1016/j.enconman.2018.02.087

Fan, J., Wu, L., Zhang, F., Cai, H., Zeng, W., Wang, X., & Zou, H. (2019a). 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, 100, 186–212; doi:org/10.1016/j.rser.2018.10.018

Feng, Y., Hao, W., Li, H., Cui, N., Gong, D., & Gao, L. (2020). Machine learning models to quantify and map daily global solar radiation and photovoltaic power. Renewable and Sustainable Energy Reviews, 118, 109393; doi:org/10.1016/j.rser.2019.109393

Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. Journal of Machine Learning Research, 20(177), 1–81

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., & Reichle, R. (2017). The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). Journal of Climate, 30(14), 5419–5454; doi:org/10.1175/JCLI-D-16-0758.1

Halawa, E., GhaffarianHoseini, A., & Li, D. H. W. (2014). Empirical correlations as a means for estimating monthly average daily global radiation: A critical overview. Renewable Energy, 72, 149–153; doi:org/10.1016/j.renene.2014.07.004

Hassan, M. A., Khalil, A., Kaseb, S., & Kassem, M. A. (2017). Exploring the potential of tree-based ensemble methods in solar radiation modeling. Applied Energy, 203, 897–916; doi.org/10.1016/j.apenergy.2017.06.104

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer Science & Business Media;

Ibrahim, I. A., & Khatib, T. (2017). A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm. Energy Conversion and Management, 138, 413–425; doi:org/10.1016/j.enconman.2017.02.006

Joy, T. T., Rana, S., Gupta, S., & Venkatesh, S. (2019). A flexible transfer learning framework for Bayesian optimization with convergence guarantee. Expert Systems with Applications, 115, 656–672; doi:org/10.1016/j.eswa.2018.08.023

Kalogirou, S. A. (2013). Solar energy engineering: Processes and systems. Academic Press.

Kumar, R., Aggarwal, R. K., & Sharma, J. D. (2015). Comparison of regression and artificial neural network models for estimation of global solar radiations. Renewable and Sustainable Energy Reviews, 52, 1294–1299; doi:org/10.1016/j.rser.2015.08.021

Lahouar, A., & Slama, J. B. H. (2015). Day-ahead load forecast using random forest and expert input selection. Energy Conversion and Management, 103, 1040–1051; doi:org/10.1016/j.enconman.2015.07.041

Li, Y., Zou, C., Berecibar, M., Nanini-Maury, E., Chan, J. C.-W., van den Bossche, P., Van Mierlo, J., & Omar, N. (2018). Random forest regression for online capacity estimation of lithium-ion batteries. Applied Energy, 232, 197–210; doi:org/10.1016/j.apenergy.2018.09.182

Lotfinejad, M. M., Hafezi, R., Khanali, M., Hosseini, S. S., Mehrpooya, M., & Shamshirband, S. (2018). A comparative assessment of predicting daily solar radiation using bat neural network (BNN), generalized regression neural network (GRNN), and neuro-fuzzy (NF) system: A case study. Energies, 11(5), 1188; doi:org/10.3390/en11051188

Marzouq, M., Bounoua, Z., El Fadili, H., Mechaqrane, A., Zenkouar, K., & Lakhliai, Z. (2019). New daily global solar irradiation estimation model based on automatic selection of input parameters using evolutionary artificial neural networks. Journal of Cleaner Production, 209, 1105–1118; org/10.1016/j.jclepro.2018.10.254

Merrouni, A. A., Elalaoui, F. E., Mezrhab, A., Mezrhab, A., & Ghennioui, A. (2018). Large scale PV sites selection by combining GIS and Analytical Hierarchy Process. Case study: Eastern Morocco. Renewable Energy, 119, 863–873; doi:org/10.1016/j.renene.2017.10.044

Molnar, C. (2020). Interpretable Machine Learning. Lulu. com.

Moreno, A., Gilabert, M. A., & Martínez, B. (2011). Mapping daily global solar irradiation over Spain: A comparative study of selected approaches. Solar Energy, 85(9), 2072–2084; org/10.1016/j.solener.2011.05.017

Olatomiwa, L., Mekhilef, S., Shamshirband, S., & Petković, D. (2015). Adaptive neuro-fuzzy approach for solar radiation prediction in Nigeria. Renewable and Sustainable Energy Reviews, 51, 1784–1791; doi:org/10.1016/j.rser.2015.05.068

Paulescu, M., Stefu, N., Calinoiu, D., Paulescu, E., Pop, N., Boata, R., & Mares, O. (2016). \AAngström–Prescott equation: Physical basis, empirical models and sensitivity analysis. Renewable and Sustainable Energy Reviews, 62, 495–506; doi:org/10.1016/j.rser.2016.04.012

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830

Piri, J., Shamshirband, S., Petković, D., Tong, C. W., & ur Rehman, M. H. (2015). Prediction of the solar radiation on the earth using support vector regression technique. Infrared Physics & Technology, 68, 179–185; doi:org/10.1016/j.infrared.2014.12.006

Quej, V. H., Almorox, J., Arnaldo, J. A., & Saito, L. (2017a). 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, 155, 62–70; doi:org/10.1016/j.jastp.2017.02.002

Rohani, A., Taki, M., & Abdollahpour, M. (2018). A novel soft computing model (Gaussian process regression with K-fold cross validation) for daily and monthly solar radiation forecasting (Part: I). Renewable Energy, 115, 411–422; doi:org/10.1016/j.renene.2017.08.061

Ruiz-Arias, J. A., Pozo-Vázquez, D., Santos-Alamillos, F. J., Lara-Fanego, V., & Tovar-Pescador, J. (2011). A topographic geostatistical approach for mapping monthly mean values of daily global solar radiation: A case study in southern Spain. Agricultural and Forest Meteorology, 151(12), 1812–1822; doi:org/10.1016/j.agrformet.2011.07.021

Sameen, M. I., Pradhan, B., & Lee, S. (2020). Application of convolutional neural networks featuring Bayesian optimization for landslide susceptibility assessment. Catena, 186, 104249; doi:org/10.1016/j.catena.2019.104249

Scornet, E. (2020). Trees, forests, and impurity-based variable importance. ArXiv Preprint ArXiv:2001.04295.

Shamshirband, S., Mohammadi, K., Yee, L., Petković, D., & Mostafaeipour, A. (2015). A comparative evaluation for identifying the suitability of extreme learning machine to predict horizontal global solar radiation. Renewable and Sustainable Energy Reviews, 52, 1031–1042; doi:org/10.1016/j.rser.2015.07.173

Shi, J. Q., & Choi, T. (2011). Gaussian process regression analysis for functional data. Chapman and Hall/CRC.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. Advances in Neural Information Processing Systems, 2951–2959.

Sun, H., Gui, D., Yan, B., Liu, Y., Liao, W., Zhu, Y., Lu, C., & Zhao, N. (2016). Assessing the potential of random forest method for estimating solar radiation using air pollution index. Energy Conversion and Management, 119, 121–129; doi:org/10.1016/j.enconman.2016.04.051

Tao, H., Ewees, A. A., Al-Sulttani, A. O., Beyaztas, U., Hameed, M. M., Salih, S. Q., Armanuos, A. M., Al-Ansari, N., Voyant, C., & Shahid, S. (2021). Global solar radiation prediction over North Dakota using air temperature: Development of novel hybrid intelligence model. Energy Reports, 7, 136–157; doi:org/10.1016/j.egyr.2020.11.033

Vapnik, V. (2013). The nature of statistical learning theory. Springer science & business media.

Wang, L., Kisi, O., Zounemat-Kermani, M., Zhu, Z., Gong, W., Niu, Z., Liu, H., & Liu, Z. (2017). Prediction of solar radiation in China using different adaptive neuro-fuzzy

methods and M5 model tree. International Journal of Climatology, 37(3), 1141–1155; doi.org/10.1002/joc.4762

Wang, R., Lu, S., & Li, Q. (2019). Multi-criteria comprehensive study on predictive algorithm of hourly heating energy consumption for residential buildings. Sustainable Cities and Society, 49, 101623; doi:org/10.1016/j.scs.2019.101623

Wang, Y., Kandeal, A. W., Swidan, A., Sharshir, S. W., Abdelaziz, G. B., Halim, M. A., Kabeel, A. E., & Yang, N. (2021). Prediction of tubular solar still performance by machine learning integrated with Bayesian optimization algorithm. Applied Thermal Engineering, 184, 116233; doi:org/10.1016/j.applthermaleng.2020.116233

Wang, Z., Wang, Y., & Srinivasan, R. S. (2018). A novel ensemble learning approach to support building energy use prediction. Energy and Buildings, 159, 109–122; doi:org/10.1016/j.enbuild.2017.10.085

World Energy Outlook 2020 – Analysis. (n.d.). IEA. Retrieved 14 January 2021, from https://www.iea.org/reports/world-energy-outlook-2020

Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., & Deng, S.-H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. Journal of Electronic Science and Technology, 17(1), 26–40; doi:org/10.11989/JEST.1674-862X.80904120

Yadav, A. K., Malik, H., & Chandel, S. S. (2014). Selection of most relevant input parameters using WEKA for artificial neural network based solar radiation prediction models. Renewable and Sustainable Energy Reviews, 31, 509–519; doi.org/10.1016/j.rser.2013.12.008

Zeng, Z., Wang, Z., Gui, K., Yan, X., Gao, M., Luo, M., Geng, H., Liao, T., Li, X., & An, J. (2020). Daily Global Solar Radiation in China Estimated From High-Density Meteorological Observations: A Random Forest Model Framework. Earth and Space Science, 7(2), e2019EA001058; doi:org/10.1029/2019EA001058

Zhang, Q., Hu, W., Liu, Z., & Tan, J. (2020). TBM performance prediction with Bayesian optimization and automated machine learning. Tunnelling and Underground Space Technology, 103, 103493; doi:org/10.1016/j.energy.2018.08.207

Zhang, Z., Wang, G., Liu, C., Cheng, L., & Sha, D. (2021). Bagging-based positive-unlabeled learning algorithm with Bayesian hyperparameter optimization for three-dimensional mineral potential mapping. Computers & Geosciences, 104817; doi:org/10.1016/j.cageo.2021.104817

Zhou, Z.-H. (2012). Ensemble methods: Foundations and algorithms. Chapman and Hall/CRC.