Renewable Energy 215 (2023) 118904

Contents lists available at ScienceDirect

Renewable Energy

journal homepage: www.elsevier.com/locate/renene

Using artificial intelligence for global solar radiation modeling from meteorological variables

Salma Zaim [a], Mohamed El Ibrahimi [a,b], Asmae Arbaoui [a,b], Abderrahim Samaouali [b], Mouhaydine Tlemcani [c], Abdelfettah Barhdadi [a,*]

[a] Physics of Semiconductors and Solar Energy Research Team (PSES), Energy Research Centre (CRE), Ecole Normale Supérieure, Mohammed V University in Rabat, Morocco
[b] Thermodynamic-Energy Team, Energy Research Centre (CRE), Faculty of Sciences, Mohammed V University in Rabat, Morocco
[c] Department of Mechatronics Engineering, Institute of Earth Sciences, School of Sciences and Technology, University of Evora, Portugal

ARTICLE INFO

Keywords:
Global solar radiation
Modeling
Artificial neural network
Levenberg marquardt algorithm
EXtreme gradient boosting
Morocco

ABSTRACT

Long-term quantification of solar energy variables at ground level is not easily achievable in many locations. In order to overcome this limitation, use of artificial intelligence such as the application of machine learning methods is commonly used for solar irradiance prediction.

In this context, this study proposes the implementation of artificial neural networks as deep learning and the XGBoost algorithm as a machine learning method for modeling the hourly global solar radiation for a humid climate such as the Rabat region. For this purpose, hourly meteorological data from the city of Rabat in Morocco are chosen in order of importance using the random forests method, for training and testing the models, namely date and time, sunshine duration, temperature, relative humidity, wind speed/direction and pressure. Subsequently, models are selected after the validation phase for testing, whose performance is evaluated using relevant statistical indicators. As a result, we retain 2 ANN and 1 XGBoost models which are eventually very close in terms of performance with a coefficient of determination value equal to 98% and 97% respectively. However, statistical indicators have proven to be effective in assessing the accuracy and fidelity of each model.

Ultimately, the intent of the modeling in terms of accuracy, simplicity or fidelity is a crucial factor in the selection of the model algorithm to adopt.

1. Introduction

In light of global climate change, the growing concern and interest in energy conservation and environmental protection is becoming an opportunity for countries and communities to develop their energetic infrastructure and accelerate their energy transition from near total dependence on fossil fuels to greater use of the alternatives low carbon renewable energy sources. Given its inexhaustibility, environmental sustainability, and ease of access at low cost in vast regions of the globe, solar energy is at the core of the consortium of energy generation technologies. This makes it the most abundant renewable energy resource in the world. Morocco is granting a particular interest for the clean energy production sector with an increasing installed capacity from renewable sources which is presently about 4 GW, including 750 MW from solar energy. This has allowed reaching a contribution of 37% of renewable energies in the total installed power during the year 2020 [1].

Morocco benefits from a huge solar energy potential as shown in Fig. 1 which illustrates the frequency distribution of the global horizontal radiation having an average of 5.54 kWh/m^2 [2].

It is commonly accepted that the solar energy source is permanent and abundant in nature and does not need to be replenished. But the distribution of solar irradiation intensity varies significantly in each area of the globe [3]. Therefore, the knowledge of the availability of solar radiation on horizontal and inclined planes as well as the consideration of the solar radiation mapping of an area remain indispensable, not only for the implementation of conversion systems, but also for the analysis of the solar potential that intrigues researchers in several fields. Practically, various types of tools or devices are usually used for solar radiation measurement depending on the requirement such as solar-meters, pyranometers and pyrheliometers. However, the availability of these tools remains limited due to their cost and the need for regular maintenance, whether corrective or preventive, not to mention the potential

* Corresponding author.
E-mail address: abdelfettah.barhdadi@ens.um5.ac.ma (A. Barhdadi).

https://doi.org/10.1016/j.renene.2023.118904
Received 4 December 2022; Received in revised form 12 May 2023; Accepted 9 June 2023
Available online 9 June 2023

Nomenclature

ANN	Artificial Neural Network
XGBoost	EXtreme Gradient Boosting
GSR	Global Solar Radiation (W.m^{-2})
FFNN	Feed Forward Neural Network
LM	Levenberg-Marquardt
BP	Back Propagation
M	Month
H	Hour
S	Sunshine duration
RH	Relative Humidity (%)
Wd	Wind direction (°)
Ws	Wind speed (m.s^{-1})
T	Air Temperature (°C)
ICE	Individual Conditional Expectation
PDP	Partial Dependence Plot
n	Number of observations
Xobs (i)	The i-th observed value of GSR
Xsim (i)	The i-th simulated value of GSR
RMSE	Root Mean Square Error
MSE	Mean Square Error
MAE	Mean Absolute Error
R^2	Coefficient of determination
NS	Nash-Sutcliffe criterion
RVE	Relative Volume Error criterion

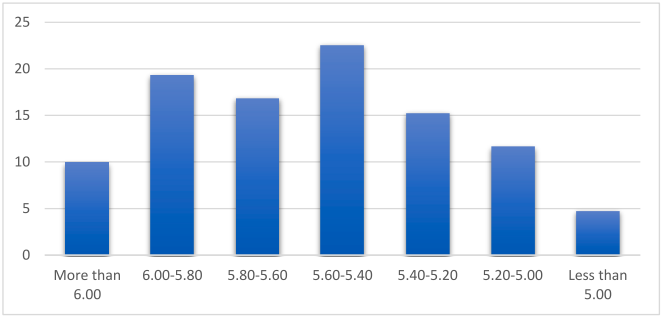

Fig. 1. Frequency distribution of GHI (kWh/m^2) on the Moroccan territory.

missing data and low accuracy of some of them.

To overcome such limitations, modeling remains among the best solutions. The types of modeling appropriate to this context can be classified into two sections. The first concerns physics-based models, including satellite image-based models, numerical weather prediction, numerical regression, and sky imagery. The second section includes statistical artificial intelligence methods, i.e., machine learning (ML) algorithms. In fact, ML-based artificial intelligence approaches are widely used and are recognized for their effectiveness in solving complex environmental and energy engineering problems [4–6]. Depending on the objectives, they can be used in different applications such as classification, clustering and regression [7]. Indeed, many ML techniques have been used for global solar radiation prediction. Among these models, ANN algorithms are the most frequently used [8]. For instance, Koca et al. have worked on solar radiation prediction for about seven cities in Turkey using an ANN. To this end, they used various activation functions in the hidden layer of the ANN model and then chosen the most appropriate models for the selected regions [9]. On the other hand, Geetha et al. revealed, in their study conducted in India, that the ANN model they developed by LM algorithm, can be used to efficiently estimate the hourly solar radiation in a shorter time and with minimum error [4].

Furthermore, among the ML methods used in practice, ensemble methods such as bagging, boosting, and random forest are known to be highly effective, especially for tabular data such as weather data. Gradient tree boosting is a technique that is proving successful in many applications. Tree boosting has been shown to give top results on many Benchmarks. In its advanced version, EXtreme Gradient Boosting (XGBoost) is a scalable machine learning system for tree boosting. It has been widely recognized in a variety of data exploration challenges [8, 10]. Although it is widely used in many other fields, the application of the model remains limited in solar radiation prediction compared to other learning methods, especially for studies conducted in Morocco.

The objective of this study is to investigate the feasibility and applicability of using ANNs as deep learning along with the XGBoost algorithm to model the nonlinear relationship between solar radiation and other meteorological parameters.

2. Material and methods

2.1. Methodology

In this first section, we explore the various data measurement devices as well as the characteristics of the study site. After that, the program inputs will be selected in order of importance using the random forest method for the selection of each input weather parameter based on its relevance level. This method plays the same role as the genetic algorithm (GA) or the ant colony optimization (ACO) employed in Ref. [5]. In order to evaluate the consistency of the ML models with the target quantity, two ML models: ANN and XGBoost have been trained and validated. As a result, the best-performing models will be selected and then submitted to the testing phase with random data of the year 2021. The main steps of the methodology adopted are illustrated in the flowchart of Fig. 2.

2.2. Study area and measurement station

The meteorological parameters used in this study are measured in solar energy platform of our laboratory in the city of Rabat, capital of Morocco. This location is set at 33° 979106 latitude, - 06°827483 longitude, and 89 m altitude. It is also characterized by a Mediterranean climate, which refers to a warm temperate climate with dry summer according to the Köppen Geiger classification [11]. The input data for our model come from a very good meteorological station allowing real-time measurements of several meteorological parameters thanks to its high quality and accurate sensors. The data collection refers to the period from January 1, 2020 to December 31, 2020 with a time step of 30 min.

The main equipment and their technical characteristics allowing the measurement of the various climate parameters are listed in Table 1.

2.3. Variables dependency and data selection

In many cases, especially those of high dimension, the choice of the number and nature of predictors is of crucial importance for predictive modeling. Since the issue consists of linking the predictable output to a set of inputs, the use of inappropriate or inadequate number of inputs may lead to weak modeling and results. However, in most cases, the inputs that should be selected and used in the modeling are not so obvious. There is often some uncertainty for which inputs should be used. This is the reason why, in the framework of this study, we started by evaluating first the effect of each potential input on the predictable

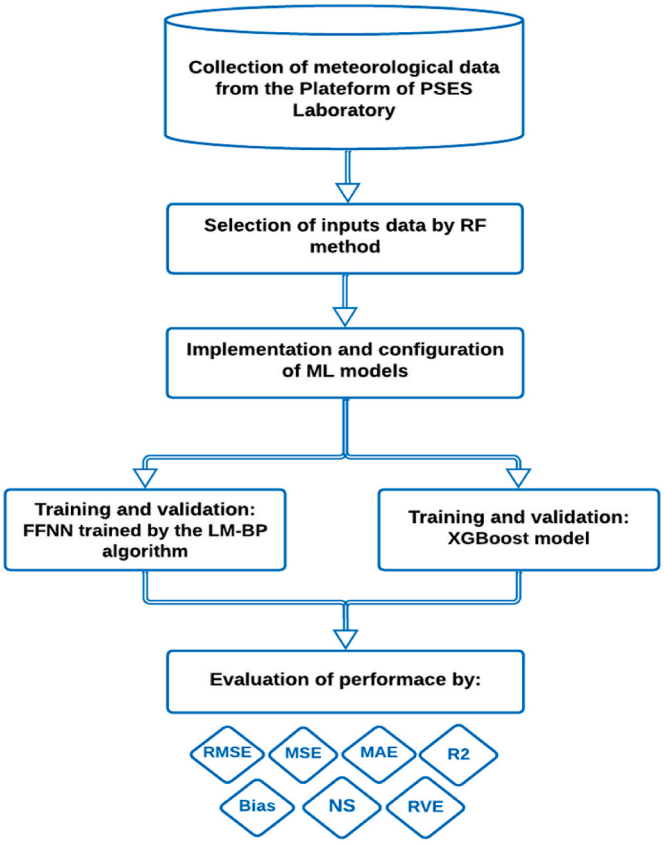

Fig. 2. Flowchart of the adopted methodology.

output by using the Random-Forest method. This algorithm evaluates the relevance of each predictor by building a hierarchy of all potential inputs with respect to the predictable output [12]. This index allows us to rank the variables from the most important to the least important. The evaluation was carried out on 9 predictor parameters. According to the percentage of importance obtained by this method, 8 predictor variables were retained as shown in the histogram (Fig. 3), while precipitation has been excluded as having a rather negligible percentage.

Furthermore, we evaluated the interaction and influence of the predictor's inputs on our target response which is the GSR following an approach based on the Individual Conditional Expectation (ICE) and Partial Dependence Plots (PDP). In fact, PDP are used to implement the marginal contribution of different characteristics on the output. They are closely related to the ICE graphs as well. The difference is that ICE graphs show changes in the prediction for each instance of the data, resulting in one line per instance for ICEs, as opposed to an overall line in PDP [13]. To simplify, a PDP can be expressed as the average of the lines in an ICE graph. The main purpose of using these graphs is to show how a change in a feature can concretely affect a given output. Fig. 4 gathers the ICE and PDP for the first 3 parameters ranked as most important by the random forest method, namely sunshine duration, hour and relative humidity.

Fig. 4 a shows that GSR, represented by its ICE graph, increases from 100 to 270 W/m^2 with increasing insolation duration from 0 to 90 min. On the contrary, GSR in Fig. 4 c, shows a slight decrease which does not exceed 10 W/m^2, when the relative humidity goes from 12% to 100%. Fig. 4 b shows that GSR increases from 6 a.m. reaching its maximum value at 12 p.m., then decreases and reaches its minimum value around 8 p.m. The time of day has an important effect on the value of the radiation, and it varies from 20 to 450 W/m^2 in the ICE graphs.

Based on the previous arguments, the ultimate selection of input variables is composed of relative humidity, sunshine duration, temperature, pressure, wind speed/direction, month and time. Table 2 groups the different statistical factors (Minimum, Maximum and Average) related to the input and output parameters.

Table 1
Technical characteristics of the instruments used to measure the various meteorological parameters.

Measurement	Equipment	Technical features
Global, direct and diffuse solar radiation	Solar tracker *Kipp & Zonen SOLYS2:* - 1 pyrheliometer *CHP1* for direct radiation. - 2 pyranometers *CMP10*, the first one for global radiation and the second one with shading ball for the diffuse radiation.	- Range of values up to 4000 W/m^2 - Operational temperature range: from - 40 °C to 80 °C
Air temperature	*CS215 Campbell* Sensor	- Measurement range: from - 40 °C to 70 °C - Accuracy: ±0.3 °C (at 25 °C) and ±0.4 °C (from 5 °C to 40 °C)
Relative humidity	*CS215 Campbell* Sensor	- Measurement range of 0–100% (from - 20° to + 60 °C). - Accuracy of ±2% (between 10% and 90%) at 25 °C and ±4% (from 0% to 10% and from 90% to 100%) at 25 °C
Atmospheric pressure	Numerical barometer Vaisala *PTB330*	Accuracy of ±0.10 hPa at 20 °C and above
Sunshine duration	Sensor type *CSD3 Kipp & Zonen*	- Global spectral range: 400 nm–1100 nm - Accuracy: more than 90%
Wind speed and direction	2D ultrasonic anemometer *WindSonic4*	- Speed: measurement range from 0 to 60 m/s with ±2% accuracy - Direction: measurement range from 0° to 359° (no dead band) with ±2° accuracy
Precipitation	Rain gauge *Lambrecht 15188*	Gauge with tilt system and heating

3. Machine learning models for GSR modeling

3.1. ANN theory

ANN derived from artificial intelligence concepts, are commonly used to solve complex problems that are difficult to model in analytic ways. It is indeed a concept inspired analogously from the efficient behavior of the human brain. As in the brain, a set of identical artificial neurons are connected in series to each other to form the whole network. Networks are distinguished according to different criteria, either by their architecture with the number of layers used, or by their complexity including the number of neurons, but also by the objective aim for optimization, supervised learning, etc [14].

In a multilayer ANN, the neurons are distributed in different layers: an input layer, one or more hidden layers and an output layer. The first input layer receives the collected data and transfers the input signal to the next layer thanks to its ability to communicate with the other neurons called neuron weight wi,j. Also, with each neuron not belonging to the input layer, is associated a constant b called bias [14]. It is worth mentioning that the function which receives the input signal and generates the output one, taking into account a certain type of threshold, is called activation function.

Most of applications associated with the GSR modeling context use Feed-Forward Neural Networks (FFNN) which are usually trained with the Back-Propagation (BP) training algorithm. This is indeed a

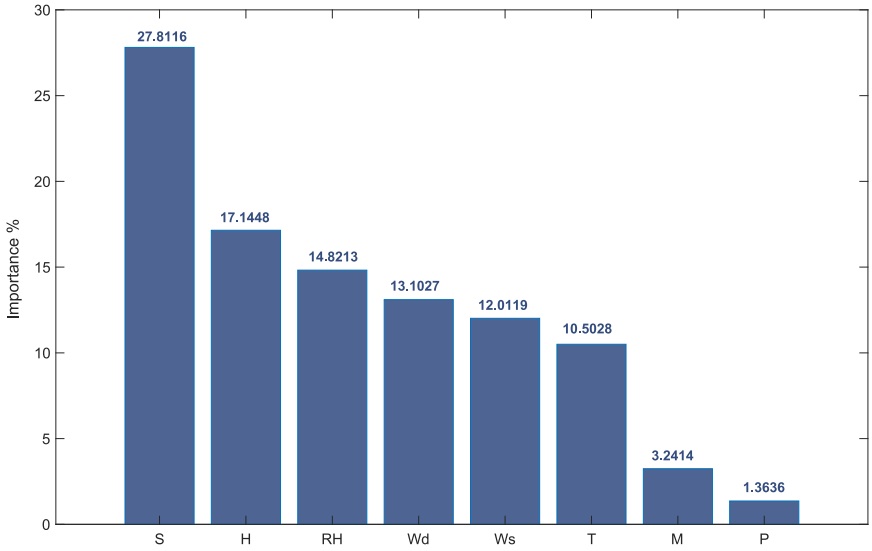

Fig. 3. Importance of the 8 predictors estimated by RF method.

supervised iterative learning process based on finding the global minimum of the error surface, which is the difference between the model output and the target, based on the weights and biases of the ANN [15]. The BP process in turn contains different algorithms; we mention for example the Gradient Descent (GD), Levenberg-Marquardt (LM), Resilient Propagation (RP) and Graded Conjugate Gradient (GCS) [16].

3.2. ANN configuration and implementation

In first instance, the 2020 data set is randomly divided into two samples with different percentages: 77% for training, 23% for validation. In order to use this random sampling method, the reproducibility of the results is ensured by using a random number generation command "rng". Two different samples rng (0) and rng (1) were used for each simulated architecture. This is indeed a legal strategy that brings an offload in terms of time and memory.

The model adopted in this work is a FFNN trained by the LM-BP algorithm, developed in MATLAB. The activation function of the hidden layer is a sigmoid hyperbolic tangent function and the one of the output is an identity function. Table 3 provides a brief description of each of the mentioned functions. Also, Fig. 5 shows concurrently the flowchart of the adopted algorithm and the optimal structuring of a generalized neural network.

As all optimization methods are iterative algorithms, they require stopping criteria. The MATLAB toolbox proposes several criteria such as the number of iterations, time, performance, premature stop, etc. Two stopping criteria were considered as the most critical. These are the performance measured on the basis of the mean square error as well as the premature stop, which allows to avoid the over-adjustment. Indeed, the performance criterion has been set to zero to converge to the lowest possible error, while the premature stopping criterion has been chosen in order to stop the learning before 40 successive epochs with overfitting of the resulting models. It should be noted that the over-adjustment leads to a deviation of the validation curve from the learning curve. This means that the model is very accurate with the learning inputs. However, this process generates large errors in the test and validation data. Regarding the criteria of time and number of iterations, these can be important when comparing different networks or optimization algorithms. Therefore, the time criterion was set to infinity and a large value was assigned to the maximum number of epochs (20 000 epochs).

Tuning ML models is a type of optimization problem, and as mentioned earlier the objective function considered here is the MSE. Fig. 6 shows the evolution of the MSE for the training and validation samples during the learning process. The represented case is the model of one hidden layer with 15 neurons.

3.3. XGBoost theory

As mentioned earlier, XGBoost is one of the most popular boosting tree algorithms for the gradient boosting machine (GBM). It has been widely used due to its high problem solving performance and minimal feature requirements [18]. Practically, it can be used for regression and classification problems. Its operating principle is based on generating a weak learner at each step and then accumulating it in the total model. It was conceived largely to boost the performance of ML models and computational speed. With this algorithm, trees are built in a parallel way, instead of being built sequentially. It follows a level-based strategy, scanning the gradient values and using these subsets to evaluate the quality of the splits at each potential split in the training set.

Compared to deep learning algorithms, XGBoost is known to be easier to use for small datasets running on the CPU. On the other hand, comparing it with the random forest method, the main difference between them is that in RF, the trees are built independently of each other, while GBM adds a new tree to complete the already built trees.

Specifically, the XGBoost algorithm is a highly accurate and evolutionary implementation of gradient augmentation which extends the limits of computational power for augmented tree algorithms.

3.4. XGBoost configuration

By simplifying the objective functions that combine the predictive and regularization terms from an optimal computational speed, XGBoost aims to avoid overfitting while optimizing computational resources. As shown in Fig. 7, the additive training process in XGBoost begins with fitting the first learner to the entire input data set, and a second model is then fitted to this residual data to address the drawbacks of a weak learner. This fitting process is repeated several times until the stopping criterion is reached. The final model prediction is obtained by summing the predictions of each learner. The general function for the prediction at step t is shown in eq. (1) [10]:

$$f_i^{(t)} = \sum_{k=1}^{t} f_k(xi) = f_i^{(t-1)} + f_t(xi) \qquad \text{(eq.1)}$$

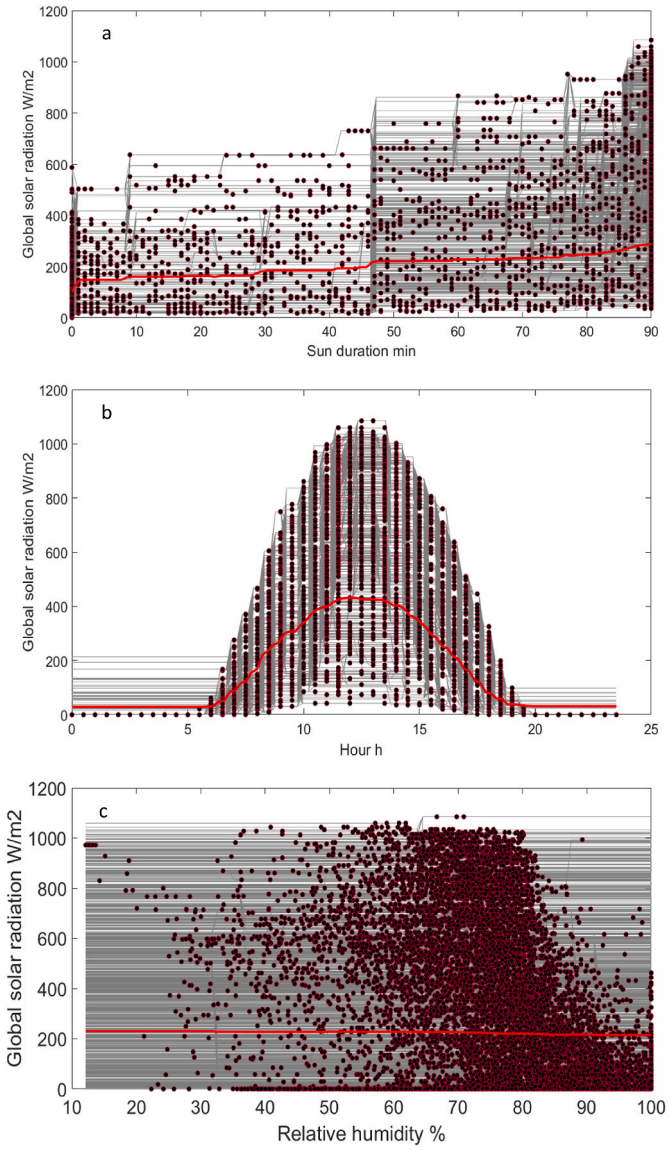

Fig. 4. ICE and PDP plots for the variables a. Sunshine duration b. time and c. relative humidity.

Table 2
Statistical parameters for the input and output variables.

Variable		Unit	Minimum	Maximum	Average
Input	RH	%	12.1	100	79.9
	S	–	0	91	33.6
	T	°C	7.5	40.6	18.7
	P	hPa	993.9	1022	1007
	Ws	m/s	0.1	8.5	1.9
	Wd	°	0	360	208.9
	M	–	1	12	–
	H	–	0	23.5	–
Output	GSR	W/m^2	0	1130.6	223.3

Where ft(xi) is the learner at step t; xi is the input variable; fi(t) and fi(t - 1) are the predictions at t and t – 1.

3.5. Hyper-parameter tuning for model refinement

In ML, a hyper-parameter is a parameter whose value is used to control the learning process. Hyper-parameters cannot be inferred when fitting the ML to the training set because they refer to the model selection task, or as algorithm hyper-parameters, which in principle have no influence on the performance of the model but affect the speed and quality of the learning process.

Table 3
Description of the activation functions used [17].

Function	Definition	Description	Range variation of F (x)
Identity	F(x) = x	Linear or identity activation function is the most basic one, it copies the input to the output. For neural networks the activation of the neuron is transmitted directly to the output.	[-∞, +∞]
Hyperbolic tangent	$F(x) = \frac{2}{1+e^{-2x}} - 1$	It is a scaled sigmoid function characterized by an "S" curve, and generally gives good results because of its symmetry. It is indeed adapted to multilayer perceptrons.	[-1, +1]

3.5.1. ANN hyper-parameters

An example of a model hyper-parameter for ANN is the topology and size of a neural network. The primary goal is to find the right combination of the values of these hyper-parameters to determine the global minimum of the implemented function. In this sense, the number of hidden layers, the number of neurons and the learning rate are the most crucial hyper-parameters. Therefore, on the one hand, we performed the execution of two programs, the first one with a single hidden layer with a number of neurons ranging from 9 to 30 and the second one with two hidden layers and a number of neurons ranging from 1 to 16 for each layer separately. On the other hand, the learning rate is a hyper-parameter that controls how much the weight of the ANN is adjusted with respect to the lost gradient. Choosing a too-low value for the learning rate may result in a long learning process that could get stuck, while a too high value may result in learning a suboptimal set of weights too quickly or an unstable learning process. To avoid this problem, two adaptation coefficients were used. The variation in terms of samples and tests of different combinations of hyper-parameters led to 3920 FFNN models. Fig. 8 shows the final architecture and configuration of the neural networks adopted for 1 and 2 hidden layers.

3.5.2. XGBoost hyper-parameters

As mentioned before, the main objective of this step for any ML model is to find the right combination of the values of these hyper-parameters to determine the global minimum of the implemented function. In this sense, based on the studies previously conducted with the XGBoost algorithm [10,19] and taking into consideration the relevance of each parameter, the hyper-parameters selected for this study are grouped in the Table 4.

3.6. Statistical performance metrics

The evaluation in terms of performance of the different models was carried out on the basis of validation sample using several statistical indicators, namely Bias, RMSE-val, MSE-val, MAE-val, and R^2-val which can be defined as follows:

Bias: It is generally recognized as a criterion of fidelity, it represents the difference between observations and measurements, indicating whether the model systematically overestimates or underestimates the predicted values. It can be calculated using eq. (2).

$$BIAS = \frac{1}{n}\sum_{i=1}^{n}[Xobs(i) - Xsim(i)] \quad \text{(eq. 2)}$$

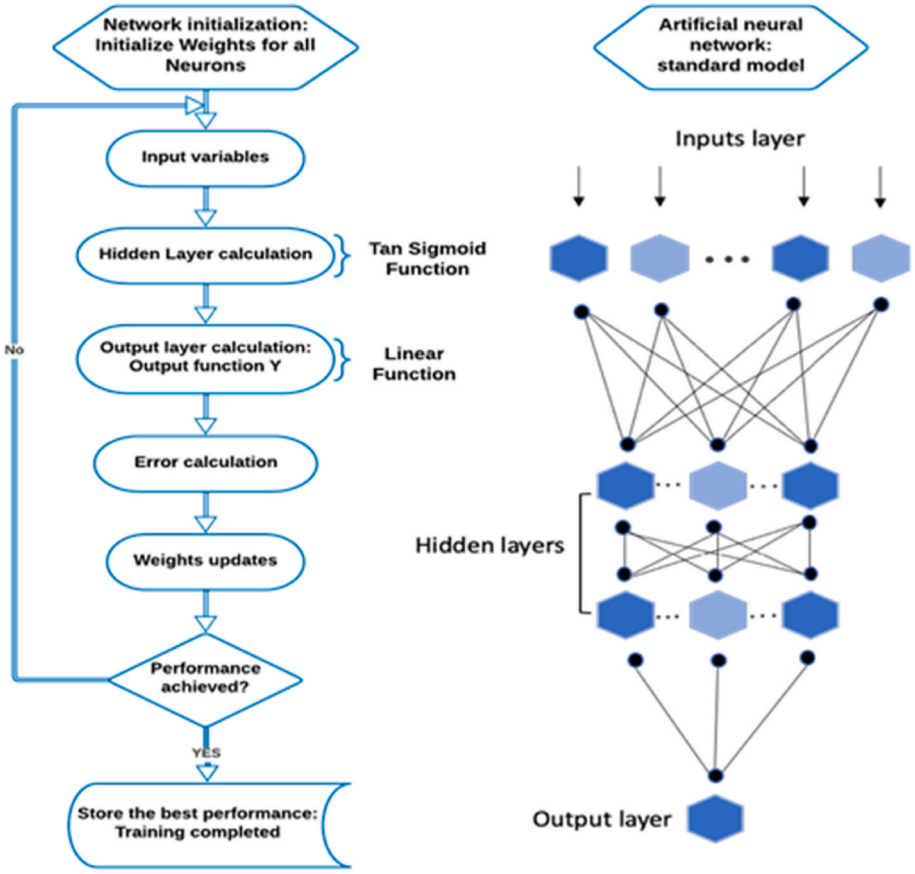

Fig. 5. Flowchart of the training process: BP algorithm and optimal ANN structure.

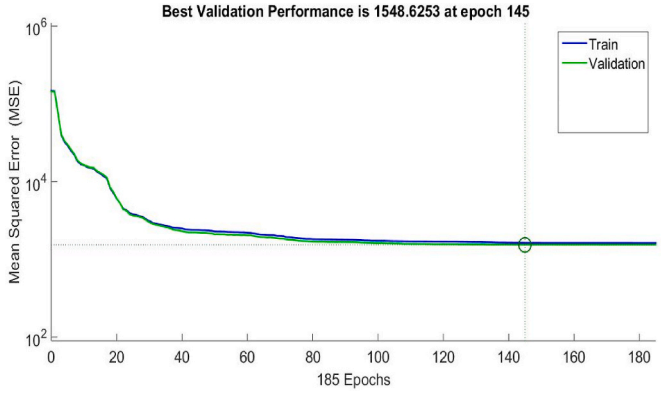

Fig. 6. Evolution of the MSE as a performance function of the training and validation data.

RMSE: The Root Mean Square Error is an accuracy criteria measuring the variation of the predicted values compared to the measured ones which allows to characterize the size of the gaps. It is calculated from eq. (3).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[\mathrm{Xobs}(i) - \mathrm{Xsim}(i)]^2} \qquad \text{(eq.3)}$$

MSE: The Mean Square Error is the RMSE without the square root as indicated in eq. (4).

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}[\mathrm{Xobs}(i) - \mathrm{Xsim}(i)]^2 \qquad \text{(eq. 4)}$$

MAE: The Mean Absolute Error is the average of the absolute differences between the real-time observation and the prediction on the test sample where all differences have the same weight. In other words, it measures the average magnitude of the errors regardless of their direction. It is calculated from eq. (5).

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\mathrm{Xobs}(i) - \mathrm{Xsim}(i)| \qquad \text{(eq. 5)}$$

R^2: The coefficient of determination represents the proportion of variability measured quantitatively as the sum of squared deviations in the data set. This variability is represented by eq. (6).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}[\mathrm{Xobs}(i) - \mathrm{Xsim}(i)]^2}{\sum_{i=1}^{n}[\mathrm{Xsim}(i)]^2} \qquad \text{(eq. 6)}$$

Nonetheless, judging a model on the basis of the indicators mentioned above remains difficult, since each indicator is particularly dependent on the data used. This is where standardized indicators should come in. Being dimensionless parameters, they allow establishing a relative performance value for each indicator, so that we can standardize the evaluation of the model in question and also compare the models to each other afterwards [20]. It is noteworthy to mention that for the MSE/RMSE cases, this reference performance value is defined, as shown in eq. (7), as the variance of the measured values noted σ_x^2. It refers to a representation of the MSE or RMSE committed by a model where we simulate the output X as the average of the observations denoted $\overline{X}$ obs.

$$\sigma_x^2 = \mathrm{MSE} - \mathrm{BIAS}^2 \qquad \text{(eq. 7)}$$

The Nash-Sutcliffe criterion (NS) is a performance indicator that

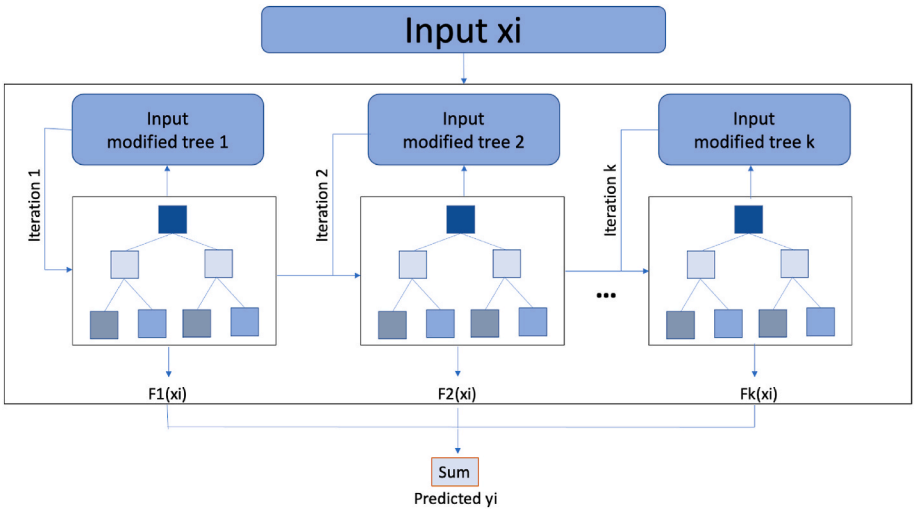

Fig. 7. Diagram of the XGBoost regression tree model.

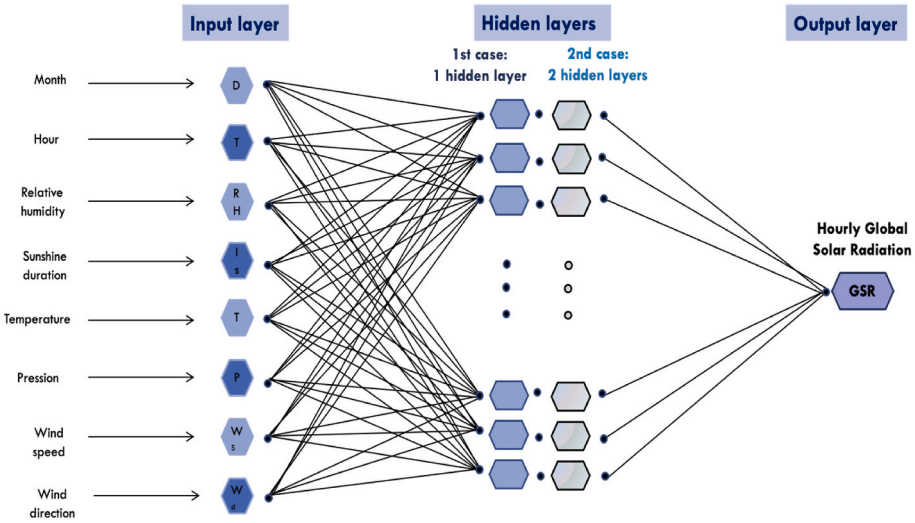

Fig. 8. Architecture of the neural network adopted.

estimates the ability of a model to reproduce an observed behavior. It is in fact constructed from the normalization of the MSE. The closer the value obtained for this criterion is to 1, the better the adequacy of the model to the observed values [20]. It can be calculated from eq. (8).

$$NS = 1 - \frac{MSE}{\sigma_x^2} \quad \text{(eq. 8)}$$

In addition, we also quote the Relative Volume Error Criterion (RVE), which is defined as the error on the modeled volume relative to the total observed volume. As with the NS indicator, this time we normalize the bias parameter presented in the previous section by a simulation with $\overline{X}$ obs equal to zero. However, we would like to precise that the volume invoked for this indicator is used in the sense of an overall quantity of the entity in question and not as the ordinary mathematical volume. RVE is then the sum of the errors related to the sum of the observed values, expressed as a relative value or as a percentage, as shown in eq. (9).

$$RVE = \frac{BIAS}{\sum_{i=1}^{n} X_{obs}(i)} \quad \text{(eq. 9)}$$

The overall performance of models is affected by σ_x^2 and bias, as shown in eq. (7). It should be mentioned that the calculation of the bias (or its normalization RVE) measure the fidelity while the calculation of σ_x^2, which is the variance of the Bias, measure the precision.

The results of these indicators were used to implement a selection of models that are expected to be the best performing. This preliminary selection was based on the best validation values of Bias, RMSE-val, MSE-val, MAE-val, R^2-val, NS-val and RVE-val. The selected models are then candidates for the test phase.

4. Results and discussion

4.1. Models performance

Once the models configuration has been established, the question that arises is regarding its reliability and relevance. The quality of a model is normally judged from the similarity of the measured data and those simulated by the model. However, this assessment has to be made from data that the model has never seen or used, which refers to data measured over a period other than the one used in the training phase. This being said, in order to concretely evaluate the best models selected during the validation, we used a database from January 1, 2021 to August 31, 2021 collected from the same weather station.

Table 4
Hyper-parameters used for the optimization of XGboost models.

Hyper-parameters	Significance	Range
N_trees	The number of trees in an XGBoost model is specified in the n_estimators argument.	[50 75 100 125 150]
Max_depth	Maximum depth of a tree. The increase in the value of this quantity is susceptible to the over-fitting of the model.	[3 4 5 6 7 8 9 10]
Learning rate	The step size at each iteration by moving towards minimization of a loss function.	[0.01 0.05 0.1 0.15 0.2 0.25 0.3]
Subsample	The subsample parameters in XGBoost control the percentage of rows used to build the tree.	[0.3 0.4 0.5 0.6 0.7 0.8 0.9 1]
Gamma	Gamma specifies the minimum loss reduction required to perform a cut and makes the algorithm conservative.	[0 0.2 0.4 0.6 0.8 1 1.2 1.5]
Early stopping rounds	Early stopping is used to control the patience of the number of iterations we will wait for the next decrease in the loss value.	[5 20 40]
Min_child_weight	It is used to control overfitting being defined as the minimum sum of the weights of all observations required in a child model.	[1 2 3 4 5 6 7 8 9 10 11 12 13 14 15]

4.1.1. ANN models performance

The results of the different indicators for the models selected in the validation phase are presented in Table 5. The choice of the best model is based on the best results of the indicators (optimal results are distinguished in bold). At first glance, the values of R^2 greater than 0.97 show that there is good agreement between the measured and predicted values. Based on the R^2 and MAE criteria, the two best ANN models are {15} and {8; 7} for the present study with R^2 values of 0.9805 and 0.9793, respectively, and MAE values of 25,7 W/m^2, and 23.6 W/m^2, respectively, which are indeed very close values.

The evaluation of the NS criterion linked to the normalization of the MSE confirms the choice of the models {15} and {8; 7}. It is indeed a criterion widely used and quoted in the literature for the evaluation and the global performance of modeling.

According to the results, the model with 2 hidden layers {8; 7} with RVE = 1.1 E−5 is more faithful than the model with 1 hidden layer {15} with RVE = 2.4 E−5. Also, the model {15} with $\sigma_x^2 = 2063.6$ W^2/m^4 is more accurate than the model {8; 7} with $\sigma_x^2 = 2203.4$ W^2/m^4. As mentioned earlier by the positive sign of the Bias (RVE) criterion, both models tend to underestimate the global values of solar radiation.

4.1.2. XGBoost performance

The XGBoost algorithm was run in the Python environment. By testing several combinations of the hyper-parameters, various models were obtained. As it was previously mentioned, the choice was based on the optimal statistical indicators by taking the 3 best combinations for each indicator. With the repetition of some combinations, we eventually select 8 models with the best performance results as shown in Table 6.

Based on the statistical performance metrics reported, model N°7 proved to be the optimal model with an R^2 value reaching 0.97, a minimum MAE equal to 27.84 W/m^2, an optimal RMSE equal to 51.23 and a variance of 2576.93 W^2/m^4. Table 7 groups the hyper-parameters and the values of the associated indicators. We could also choose model 8, if we wanted to gain a little in terms of fidelity based on accuracy. Table 7 represents the combination of hyperparameters involved in model N°7.

4.1.3. Comparison of ML-models

The best models obtained by ANN and XGBoost algorithm were applied to predict the evolution of global radiation on 8 random days of the year 2021 (Figs. 9 and 10).

According to the results summarized in Table 8, the ANN model {15} is better than the XGBOOST one. However, the latter has proven to be a powerful learner machine method. More explicitly, on the basis of the variance σ_x^2, the results show that the ANN method gains in terms of accuracy with an interval of [2063.6–2374.8] (W^2/m^4). On the other hand, the dimensionless bias criterion (RVE) shows that the XGBoost method wins in fidelity with an optimal interval of [(- 5.07 E−6) – (-1.12 E−5)].

5. Conclusions

Given the importance of the magnitude of global solar radiation arriving at the Earth's surface for the optimal design and use of solar energy conversion systems as well as for other environmental applications that require knowledge of its values, this paper presents an application of two AI models (i.e. ANN and XGBoost) to accurately estimate hourly global solar radiation from meteorological data for a humid Mediterranean environment as in the Rabat region. In addition, a relevance analysis was conducted by the random-forest method to determine the most efficient input variables for the required modelling. Thus, 8 temporal and meteorological variables were adopted namely sunshine duration, relative humidity, pressure, temperature, wind speed and direction, month and time.

In first instance, the ANN network of multilayer perceptron (MLP) type with 1 and 2 hidden layers was developed with the adoption of Back-Propagation (BP) algorithm as the adjusting method. Secondly, among the ensemble ML methods, the XGBoost model was used with the variation of various hyper-parameters such as the number of trees, learning rate, early stopping and many others in order to refine the elaboration of the models as well as to avoid the problem of over-fitting.

The performance evaluation of the different models obtained was established by various statistical indicators including RMSE, MSE, MAE, R^2 and also the normalized indicators such as NS and RVE, which reflect the relevance of each model in question. The analysis of all these indicators allowed us to retain 2 ANN and 1 XGBOOST models, the first ANN model with 1 hidden layer model {15} and the second ANN model with 2 hidden layers model {8, 7}. The two ANN models were significantly close in terms of performance and accuracy with a R^2 of 98%. The XGboost model has also proved to have good results with a R^2 of 97%. The performances of the latter have revealed that this method, which is part of machine learning, is interesting and can be compared to the

Table 5
Optimum ANN model selection.

	n	MSE(W^2/m^4)	RMSE (W/m^2)	MAE (W/m^2)	BIAS(W/m^2)	R^2	NS	RVE	σ_x^2 (W^2/m^4)
1 hidden layer	**15**	**2068.2**	**45.5**	**25.7**	2.1	**0.9805**	**0.9804**	2.4 E−5	**2063.6**
	17	2336.0	48.3	27.5	1.8	0.9781	0.9779	2.0 E−5	2332.7
	27	2263.7	47.6	27.1	3.7	0.9787	0.9785	4.1 E−5	2263.7
	28	2156.1	46.4	26.1	**1.5**	0.9797	0.9796	**1.7 E-5**	2153.8
2 hidden layers	**6;8**	2334.1	48.3	26.4	1.6	0.9779	0.9779	1.8 E−5	2331.5
	4;10	2268.1	47.6	25.5	3.1	0.9787	0.9785	3.4 E−5	2258.8
	8;5	2376.6	48.8	24.3	1.4	0.9775	0.9775	1.5 E−5	2374.8
	13;4	2362.6	48.6	24.3	1.7	0.9777	0.9776	1.9 E−5	2359.7
	8;7	**2204.4**	**47.0**	**23.6**	**1.0**	**0.9793**	**0.9791**	**1.1 E-5**	**2203.4**

Table 6
Optimum XGBoost model selection.

Models	MSE(W^2/m^4)	RMSE (W/m^2)	MAE (W/m^2)	BIAS (W/m^2)	R^2	NS	RVE	σ_x^2 (W^2/m^4)
1	2638.38	51.36	28.88	- 7.14	0.9751	0.9751	- 5.07 E−6	2587.35
2	2633.80	51.32	28.83	- 7.09	0.9751	0.9751	- 5.04 E−6	2583.43
3	2626.55	51.25	28.00	- 6.85	0.9752	0.9751	- 4.86 E−6	2579.59
4	3939.05	62.76	35.73	- 15.78	0.9629	0.9628	- 1.12 E−5	3689.94
5	3939.05	62.76	35.73	- 15.78	0.9628	0.9629	- 1.12 E−5	3689.94
6	3939.05	62.76	35.7313	- 15.7832	0.9628	0.9629	- 1.12 E−5	3689.94
7	**2624.82**	**51.23**	**27.84**	- 6.92	**0.9753**	**0.9752**	- 4.91 E−6	**2576.93**
8	2630.75	51.29	27.88	**- 6.51**	0.9751	0.9751	- 4.62 E−6	2588.29

Table 7
Hyper-parameters values of the best XGBoost model obtained.

	Hyper-parameters						
	N tree	Max depth	Learning rate	Subsample	Gamma	Early stop	Min childW
Model N°7	100	20	0.1	1.0	0.0	40	15

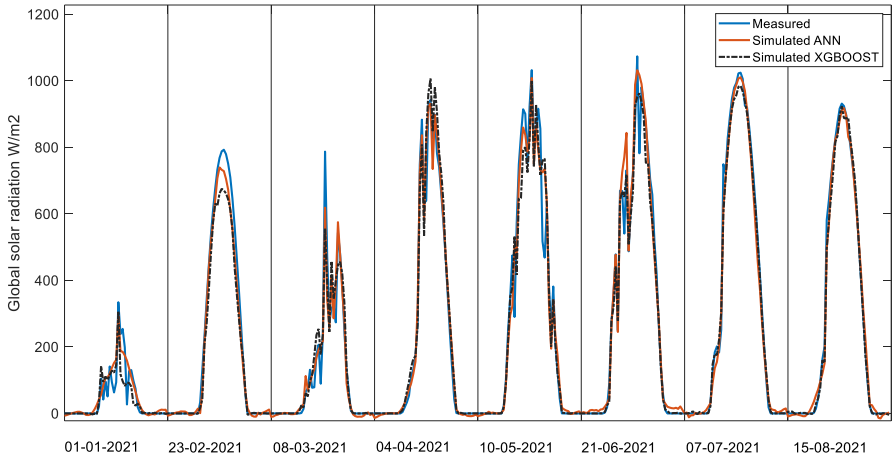

Fig. 9. Comparison between measured and simulated data by the 1 hidden layer {15} and XGBoost N°7 models.

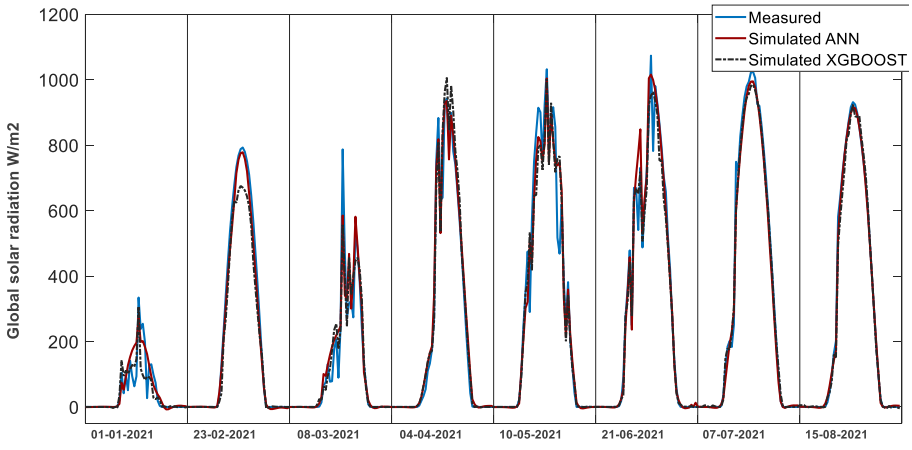

Fig. 10. Comparison between measured and simulated data by the 2 hidden layers {8; 7} and XGBoost N°7 models.

neural network method, which is part of the deep learning method.

By way of conclusion, the purpose of the modeling in terms of accuracy, simplicity or fidelity remains a decisive factor in the selection of the algorithm of the model to be adopted.

The results of this work thus contribute to the mutual intention of testing the applicability of a common set of methods and approaches to estimate solar radiation in similar geographical regions with wet aspect.

Table 8
Performance metrics gathered of the optimum 3 models.

MODEL	PERFORMANCE METRICS				
	RMSE (W/m^2)	MAE (W/m^2)	R^2	RVE	σ_x^2 (W^2/m^4)
ANN {15}	45.5	25.7	0.9805	2.4E-5	2063.6
ANN {8; 7}	47.0	23.6	0.9793	1.1E-5	2203.4
XGBOOST N°7	51.23	27.84	0.9753	−4.91E-6	2576.93

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the Moroccan Ministry for Higher Education, Scientific Research and Innovation, in the framework of Priority Research Project PPR1 Nr. 14/2016. The authors are thanking Mrs. LAARABI Bouchra, member of the PSES laboratory, for her help and her pertinent comments.

References

[1]] Ministry of Energy Transition and Sustainable Development, Key indicators (2021). https://www.mem.gov.ma/Pages/secteur.aspx?e=2, 2021. (Accessed 23 December 2021).
[2] The World Bank Group, GLOBAL PHOTOVOLTAIC POWER POTENTIAL | Country Factsheet, 2021 ([Online]. Available: globalsolaratlas.info/global-pv-potential-study/).
[3] M.R. Rashel, R. Melicio, M. Tlemcani, T. Goncalves, Modeling and simulation of PV panel under different internal and environmental conditions with non-constant load, in: IFIP Advances in Information and Communication Technology, Springer New York LLC, 2019, pp. 376–392, https://doi.org/10.1007/978-3-030-17771-3_33.
[4] A. Geetha, et al., Prediction of hourly solar radiation in Tamil Nadu using ANN model with different learning algorithms, Energy Rep. 8 (Apr. 2022) 664–671, https://doi.org/10.1016/j.egyr.2021.11.190.
[5] T. Beltramo, M. Klocke, B. Hitzmann, Prediction of the biogas production using GA and ACO input features selection method for ANN model, Information Processing in Agriculture 6 (3) (Sep. 2019) 349–356, https://doi.org/10.1016/j.inpa.2019.01.002.
[6] S.I. Kampezidou, A.T. Ray, S. Duncan, M.G. Balchanos, D.N. Mavris, Real-time occupancy detection with physics-informed pattern-recognition machines based on limited CO2 and temperature sensors, Energy Build. 242 (Jul. 2021), https://doi.org/10.1016/j.enbuild.2021.110863.
[7] H. Ali-Ou-Salah, B. Oukarfi, T. Mouhaydine, Short-term solar radiation forecasting using a new seasonal clustering technique and artificial neural network, Int. J. Green Energy 19 (4) (2022) 424–434, https://doi.org/10.1080/15435075.2021.1946819.
[8] J. Fan, et al., 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 Convers. Manag. 164 (May 2018) 102–111, https://doi.org/10.1016/j.enconman.2018.02.087.
[9] A. Koca, H.F. Oztop, Y. Varol, G.O. Koca, Estimation of solar radiation using artificial neural networks with different input parameters for Mediterranean region of Anatolia in Turkey, Expert Syst. Appl. 38 (7) (Jul. 2011) 8756–8762, https://doi.org/10.1016/j.eswa.2011.01.085.
[10] J. Yu, W. Zheng, L. Xu, L. Zhangzhong, G. Zhang, F. Shan, A pso-xgboost model for estimating daily reference evapotranspiration in the solar greenhouse, Intelligent Automation and Soft Computing 26 (5) (2020) 989–1003, https://doi.org/10.32604/iasc.2020.010130.
[11] M.C. Peel, B.L. Finlayson, T.A. Mcmahon, Updated world map of the Köppen-Geiger climate classification, Hydrol. Earth Syst. Sci. 11 (2007) 1633–1644 [Online]. Available: www.hydrol-earth-syst-sci.net/11/1633/2007/.
[12] R. Genuer, J.-M. Poggi, Arbres CART et Forêts aléatoires, Importance et sélection de variables Arbres CART et Forêts aléatoires Importance et sélection de variables [Online]. Available: https://hal.archives-ouvertes.fr/hal-01387654v2, 2017.
[13] Christoph Molnar, Interpretable Machine Learning, A Guide for Making Black Box Models Explainable, vol. 447, 2019 [Online]. Available: https://christophm.github.io/interpretable-ml-book/.
[14] S.A. Kalogirou, Artificial neural networks in renewable energy systems applications: a review, Renew. Sustain. Energy Rev. 5 (2001) 373–401 [Online]. Available: www.elsevier.com/locate/rser.
[15] Ian Goodfellow, Yoshua Bengio, Aaron Courville, Neural Networks and Deep Learning, Determination Press, 2015. Accessed: Jan. 03, 2022. [Online]. Available: http://neuralnetworksanddeeplearning.com.
[16] N. Premalatha, A. Valan Arasu, Prediction of solar radiation for solar systems by using ANN models with different back propagation algorithms, J. Appl. Res. Technol. 14 (3) (Jun. 2016) 206–214, https://doi.org/10.1016/j.jart.2016.05.001.
[17] B. Karlik, A.V. Olgac, Performance analysis of various activation functions in generalized MLP architectures of neural networks, Int. J. Artif. Intell. Expet. Syst. 1 (4) (2011).
[18] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, Aug. 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.
[19] V. Sansine, P. Ortega, D. Hissel, M. Hopuare, Solar irradiance probabilistic forecasting using machine learning, metaheuristic models and numerical weather predictions, Sustainability 14 (22) (Nov. 2022), https://doi.org/10.3390/su142215260.
[20] GRAIE-GT Autosurveillance - Sous-groupe Modélisation, Critère & indicateurs d'auto-évaluation des modèles. http://www.graie.org/graie/graiedoc/reseaux/autosurv/GRAIE-Criteres-INDICATEURS-AUTOEVALUTIONdesMODELES-AUTOSURVEILLANCE-WEB18-v1.pdf, 2018.