JCEC
THE JOURNAL OF ENGINEERING AND EXACT SCIENCES

OPEN ACCESS

The Journal of Engineering and Exact Sciences – jCEC, Vol. 09 N. 04 (2023)
journal homepage: https://periodicos.ufv.br/jcec
eISSN: 2527-1075
ISSN: 2446-9416

Hybrid Models for Daily Global Solar Radiation Assessment

Modelos híbridos para la evaluación diaria de la radiación solar global

Article Info:
Article history: Received 2023-04-04 / Accepted 2023-06-06 / Available online 2023-06-06
doi: 10.18540/jcecvl9iss4pp15926-01e

Abdelkerim Souahlia
ORCID: https://orcid.org/0000-0002-3393-1608
Telecommunications and smart systems Laboratory, University of ZianeAchour, Djelfa 17000, Algeria.
E-mail: abdelkerim.souahlia@univ-djelfa.dz
Abdelhalim Rabehi
ORCID: https://orcid.org/ 0009-0006-2506-3966
Telecommunications and smart systems Laboratory, University of ZianeAchour, Djelfa 17000, Algeria.
Applied Automation and Industrial Diagnostics Laboratory (LAADI), Faculty of Science and Technology University of Ziane Achour of Djelfa, 17000, Algeria
E-mail: a.rabehi@univ-djelfa.dz
Abdelaziz Rabehi
ORCID: https://orcid.org/00000001-8684-4754
Telecommunications and smart systems Laboratory, University of ZianeAchour, Djelfa 17000, Algeria
E-mail: rab_ehi@hotmail.fr

Resumo
La previsión diaria de la radiación solar se ha vuelto fundamental recientemente en el desarrollo de la energía solar y su integración en los sistemas de red. A pesar de la gran cantidad de técnicas de pronóstico propuestas, una estimación precisa sigue siendo un desafío importante debido a la variación no estacionaria de los componentes de la radiación solar debido a las condiciones climáticas en constante cambio. Por lo general, se utilizan varios predictores de datos de entrada para el proceso de pronóstico, lo que puede causar redundancia y correlación entre las características de los datos. Este trabajo evalúa un conjunto de técnicas de selección de características para verificar su capacidad para seleccionar los predictores relevantes y reducir la información redundante e irrelevante. Se utiliza una red neuronal artificial para ajustar la radiación solar medida en función de las características seleccionadas. El modelo desarrollado se evalúa a través de varias métricas de evaluación objetiva utilizando datos históricos de tres años medidos en la región de Ghardaia en Argelia. Los resultados muestran la efectividad del método propuesto, donde se han encontrado valores de 0,0189, 0,0286, 5,4387 y 98,28% como MABE, RMSE, nRMSE y r, respectivamente.
Palabras clave: Radiación solar, energías renovables, selección de características, Forecasting, Redes Neuronales Artificiales.

Abstract
Daily solar radiation forecasting has recently become critical in developing solar energy and its integration into grid systems. Despite the huge number of proposed forecasting techniques, an accurate estimation remains a significant challenge because of the non-stationary variation of solar radiation components due to the continuously changing climatic conditions. Usually, several input data predictors are used for the forecasting process, which can cause redundancy and correlation between data features. This work assesses a set of feature selection techniques to check their ability to select the relevant predictors and reduce redundant and irrelevant information. An Artificial

Neural Network is used to fit the measured solar radiation based on the selected features. The developed model is evaluated through various objective evaluation metrics using historical data of three years measured at the Ghardaia region in Algeria. Results show the effectiveness of the proposed method, where values of 0,0189, 0.0286, 5.4387, and 98.28% have been found as MABE, RMSE, nRMSE and r, respectively.
Keywords: Solar radiation, renewable energy, features selection, Forecasting, Artificial Neural Networks.

Nomenclature

ANN: Artificial Neural Network

FS: Features Selection

MLP: Multi-layer perceptron

GSR: Global Solar Radiation

MABE: Mean Absolute Bias Error

Min CFS: Minimum Correlation Features Selection

NCA: Neighborhood Component Analysis

SR: Solar Radiation

RMSE: Root Mean Square Error

nRMSE: Normalized Root Mean Square Error

r: Correlation Coefficient

1. Introduction

Solar radiation (SR) and daylight are required for life on Earth to exist. The natural environment is influenced by the earth's meteorological systems, which are mainly controlled by solar radiation. Its presence on the earth's surface is critical for meeting the human race's energy demands. As a result, understanding the physics of solar radiation and daylight is critical, particularly determining the quantity of energy intercepted by the earth's surface.

At different locations, information about the measured solar radiation is required to assess the solar potential of a region. Generally, a pyranometer, solarimeter, pyrheliometer, with the data acquisition system, are used to measure the solar radiation components. On the other hand, by installing measuring instruments at every site it would be impossible to cover the high expenses resulting from the non-availability of measured solar radiation data for most sites worldwide. In addition, the lack of accuracy of the measuring equipment caused records in the data set to be misplaced. Likewise, solar radiation prediction, system design, and installation necessitate the utilization of the solar radiation measured at nearby meteorological stations (Yadav et al., 2014).

Numerous models for estimating global Solar Radiation (GSR) has been included in the bibliography. They can be divided into three groups:empirical models, physical models, and soft computing using machine learning techniques (Chen et al., 2013).The first group presents models that are acquired by statistical means from observed irradiance data. Meanwhile, the second category is based on first-principle equations. At last, the third category includes techniques that have been used widely in recent years; among these techniques, we can cite Artificial Neural Networks (ANN), support vector machine (SVM) (Chen et al., 2013), extreme learning machine (ELM) (Şahin et al., 2014),and Gaussian Process Regression (GPR).

Previously, measured from conventional meteorological stations and evaluated indirectly as a function of SR, parameters such as air temperature, humidity, sunshine duration, and cloud coverage were developed in harmony with Solar Radiation prediction models (Kiziltan & Şahin, 2016). In the literature, by using three categories as inputs, Khorasanizadeh and Mohammadi(Khorasanizadeh

& Mohammadi, 2013) tried to speculate the monthly means of global solar radiation through the function of sunshine duration only, the function of sunshine duration as well as relative humidity and ambient temperature, and the function of relative humidity with ambient temperature whether being maximum or minimum, of course, independent of sunshine duration.

Based on meteorological parameters, such as daily mean air temperature, relative humidity, sunshine hours, evaporation, and wind speed, two kinds of ANN were initiated by Behrang and Assareh(Behrang et al., 2010) for the estimation of DGSR. Meanwhile, Rahimikhoob(Rahimikhoob, 2010) employed ANN in a semi-arid environment as a function of air temperature data for the estimation of GSR. Mellit et al. (Mellit et al., 2011) developed six ANN models basing their deed on three meteorological parameters like air temperature, sunshine duration, and relative humidity. Behrang et al. (Behrang et al., 2010) have trained seven ANN models using daily values of measured sunshine duration, theoretical sunshine duration, maximum temperature, and the month number.

In Abha (Saudi Arabia), Rehman and Mohandes(Rehman & Mohandes, 2008) utilized ANN as a function of air temperature and relative humidity for the estimation of GSR. As a result, it was found that in locations where only temperature and relative humidity are available, ANNs are capable to estimate the GSR. In addition, forty Chinese cities, covering nine major thermal climatic zones and sub-zones, wherein Lam and Wan (Lam et al., 2008) employed measured sunshine duration along with the use of ANNs in the purpose of developing prediction models for the estimation of the daily GSR. Obtained results show that the coefficients of determination (R2) for all the 40 cities and nine climatic zones/sub-zones are 0.82 or higher, manifesting a reasonably strong relationship existing between daily GSR and the corresponding sunshine hours.

Furthermore, in Dezful city (Iran), Asl et al. (Asl et al., 2011) have anticipated the amount of daily GSR with an absolute percentage error of 6.08%; while Ramedani et al (Ramedani et al., 2013) have estimated the amount of global solar radiation of Tehran city using multilayer perceptron (MLP) neural network along with three layers with neuron number of 6-37-1 and some minimum input parameters: maximum daily temperature, relative humidity, sunshine duration and the amount of precipitation. They have attained the optimal model with a root-mean-square error of 3.09.

Meanwhile, a study was made by Mellit and (Mellit & Pavan, 2010) in which 14 months of data measured in Trieste, Italy, and MLP were used to predict the solar irradiation for 24 h from the daily average value of the global solar irradiation and the air temperature. Consequently, after several simulations, they have found that the optimal configuration is obtained with an input layer of 3 inputs, two hidden layers of 11and 17 neurons, and an output layer of 24 outputs. Endogenous and exogenous meteorological data were used as input by Voyant et al. (Voyant et al., 2011) who applied ANN to the forecast of the daily horizontal global irradiation. Moreover, by using only endogenous input data, they tried to compare this optimal network with an ANN structure. As a consequence, the relative RMSE was 0.5% and 1% in two Corsican stations. Monthly average values of daily global radiation were estimated on horizontal surfaces in most of the studies.

Ahmet Koca et al. (Koca et al., 2011) used ANN to estimate the solar radiation within the Mediterranean region of Anatolia in Turkey. The number of input variables increased from 2 to 6 (latitude, longitude, altitude, month, average cloudiness, and sunshine duration) and the effective number of input variables used for estimation is studied.

A random choice of which predictors should be used to result in a good solar radiation forecasting is a time and effort consuming. Also, the absence of a wise decision of selecting appropriate predictors can cause redundancy and correlation between data features which increase the input dimensionality and the complexity of the forecasting process in addition to decrease the forecasting performance. Therefore, the purpose of our work is the search of the optimal input features, which can reduce the complexity and increase the performance of the daily GSR forecasting process. To reach this objective, we propose the use of a set of features selection techniques to assess the relevance of several widely used predictors. Then, using an ANN to fit the measured daily global solar radiation using the found optimal predictors. To assess the performance of our proposed method, a real dataset collected in Ghardaïa area, Algeria, during the period of 2013

to 2015 is used, where a set of objective evaluation metrics is computed. As a result, a ranking score for each predictor is delivered to exprime their relevance. Consequently, the most relevant predictors are then used to train an MLP to fit the measured daily global solar radiation.

The remain of the paper is organized as follows: Section 2 presents theoretical background of features selection techniques as well as the MLP, in addition to site location and the data set collection. Section 4 demonstrates the setup of our experiments. The found results with thorough discussions are presented in section 4. At last, the final section stands for the conclusion of the work.

2. Theory

In this section, we will present the theoretical background of the different aspects treated in the present work, such as the features selection techniques and the used classifier, which is MLP, an idea about the dataset, and the evaluation metrics used to assess our found results.

2.1. Features selection

Usually, collected data used to predict solar radiation has multiple predictors, including but not limited to temperature (T), humidity (H), pressure (Pr), and sunshine duration (SS). Having a group of features in line and selecting the features to build the best prediction model, features selection techniques help to find the relevance weight of each input parameter, usually called predictor, and its likelihood to participate in predicting the response. Features selection techniques aid in reducing the dimension of the input data and deleting the redundant noisy data, which can reduce the prediction process's complexity and enhance the predicted response's performance.

As we have claimed early in this work, our objective is to search for the optimal combination of best-input parameters that can result in a good prediction of the daily GSR. To reach this goal, we have used many feature selection techniques:

Relief technique

In 1992 Kira and Rendell developed a filter-method technique called the Relief features selection algorithm (Kira, Kenji, n.d.)(Kira, K., & Rendell, 1992). Relief was initially conceived to solve binary classification problems with discrete features. The principle of work of the Relief algorithm consists of taking a dataset with *n* samples of *p* features, scaling each feature between zero and one, and then computing a weight or a score for each input predictor. Predictors' importance is ranked according to their scores. Afterwards, a set of best-scoring predictors is selected to predict the response further. Relief scores are computed using differences between feature values and nearest neighbour instance pairs. A neighbour of each instance is called either 'a hit' if it belongs to the same class or 'a miss' if it belongs to another class. The neighborhood relation is computed based on the difference between instances features. The feature score $W_i$is updated in each iteration of the algorithm according to the following equation:

$$W_i = W_i - (x_i - nearHIT_i)^2 + (x_i - nearMiss_i)^2 \quad (1)$$

Where W_i decreases if the neighbour is a hit and increases otherwise. After the maximum number of iterations (m) is reached, each element of W_i is divided by m. The final value of W_i presents the weight or the relevance of the feature i.

Several Relief-based feature selection algorithms have been developed, including the well-known ReliefF algorithm proposed by Kononenko et al. (Kononenko, 1994), which we have used in the present work. The reliefF algorithm uses the Manhattan (L1) norm instead of Euclidean (L2) norm to search the near-hit and near-miss neighbors. In addition, Kononenko et al. have found utile to take the absolute differences and to update the weight vector instead of the square value of the differences (Kononenko, 1994). ReliefF algorithm has several advantages like(1) being more

reliable in noisy situations (Kononenko, 1994), (2) extensible to multi-class problems (Kononenko, 1994), (3) having the ability to deal with regression problems (Robnik-Šikonja & Kononenko, 1997) and (4) the robustness against missing data (Kononenko, 1994). Nevertheless, they have some cons, such as the inability to distinguish between redundant features. Also, the small size of the training samples can result in poor performance.

Minimum correlation technique

Minimum Correlation Features Selection (minimum CFS) is a very directed and simple technique based on computing the correlations between the input features.

Assuming that we have n instances with p features. The minimum CFS computes a matrix of correlations *C* of *p* by *p* elements

$$C = \begin{bmatrix} c_{1,1} & c_{1,2} & c_{1,3} & & c_{1,p} \\ c_{2,1} & c_{2,2} & c_{2,3} & & c_{2,p} \\ c_{3,1} & c_{3,2} & c_{3,3} & & c_{3,p} \\ & & & & \\ & & & & \\ c_{p,1} & & & & c_{p,p} \end{bmatrix} \quad (2)$$

Where each element $c_{i,j}$ is a correlation between f_i and f_j features.

The score or the weight W_i of the feature f_i is the mean of values of the column i (i.e the mean of the correlation factor between the feature f_i and the remaining features):

$$W_i = score_i = \frac{1}{p}\sum_{j=1}^{p} C_{j,i} \quad (3)$$

The most interesting feature is the least correlated one with other features, which corresponds to :

$$Bestf_i = min\big(mean(C)\big) \quad (4)$$

According to this rule, we sort the mean of correlations in ascending order and take the features with minimum weights to be used as relevant predictors.

Chi-square test technique

The Chi-square test is a technique that helps to remedy the problem of feature selection by testing the relationship between each two features. It is a statistical independence test used to find the dependency of two variables (Alisha Sikri, 2023). Its principle of feature selection is based on the computation of the dependency of each feature of our input data with the response using chi-square statistics. Lower is the Chi-Square value, lower is the dependence of the feature to the response, and is more likely to be discarded from the predictors selected for the prediction process. Chi-square score, usually called the $\chi 2$ test, which is given by :

$$\chi 2 = \frac{(Observedfrequency - Expectedfrequency)^2}{Expectedfrequency} \quad (5)$$

Observed frequency is the number of class observations. Expected frequency is the number of expected observations of class if there was no relationship between the feature and the target.

F-test technique

In this technique, features' scores are computed based on the variance's ratio computation using the F-test parametric statistical test. In this work, we adopted one F-test version among the known versions of the F-test technique, namely the ANOVA F-test method, standing for Analysis of Variance. This latter divides the variance between groups by the variance within a group for a given feature. Here, the groups are the samples of data with the same response. ANOVA F-test checks the relevance of every feature individually based on the F-test. Each F-test checks the hypothesis that the response values grouped by predictor variable values come from data samples having the same mean against a second hypothesis that the data samples' means are not all the same. The weaker the f-score, the greater the importance of the corresponding predictor (Dhanya et al., 2020).

Neighborhood Component Analysis technique

Neighborhood Component Analysis (NCA) is a non-parametric method used to select features to maximize the prediction accuracy of regression algorithms. NCA aims at "learning" a distance metric by searching to linearly transform the input data in such a way that the average leave-one-out (LOO) classification performance is maximized in the transformed space (Jacob Goldberger et al., 2005).

NCA technique is driven by the functionality of the well-known K- Nearest Neighbors algorithm, which is an easy classification technique based on finding the label of a test point using the labels of its k nearest training points. KNN has two main drawbacks: (1) it is computationally expensive to store and compare the test point with the entire training set points. (2) The result performance is strongly related to the distance metric searching "nearest" points. NCA remedies these two problems by using a quadratic distance metric, which can be restricted to a low rank, and reduce the dimensionality; consequently, the storage and search times are shortened.

NCA features selection technique can be used efficiently for regression by making the response value continuous.

2.2. Multi-Layer Perceptron neural network (MLP)

A multilayer perceptron (MLP) is a static artificial neural network that contains more than one perceptron. They comprise numerous layers: a layer of input receiving the signal, an output layer that makes a decision or prediction from the input's information, and an adjustable number of hidden layers that serve as the MLP's actual processing engine. Figure 1 illustrates the architecture of a generic MLP model, which is often applied to supervised learning problems: They train on a set of input-output pairs and learn to model the correlation between those inputs and outputs. Training involves adjusting the weights and biases of the model to minimize error (equation 4). Backpropagation algorithms like gradient descent, conjugate gradient, Levenberg Marquardt (LM), and an activation function are used to adjust weight and bias relative to the error.

$$y_j = f\left(\left(\sum_i w_{ij} * X_{ij}\right) - b_j\right) \quad (6)$$

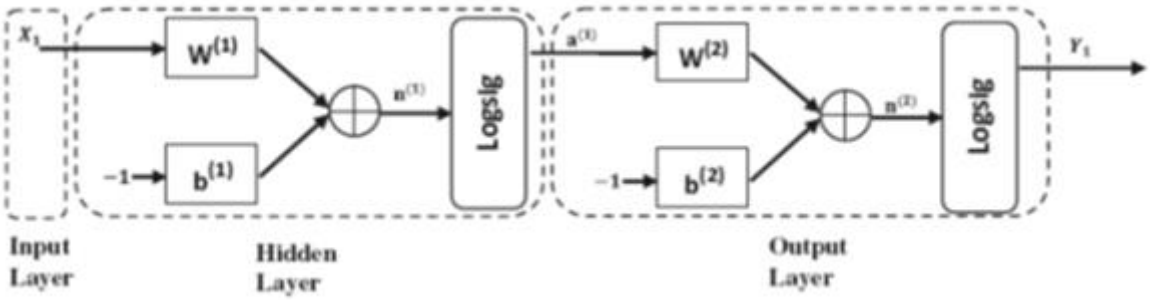

Figure 1 - Typical flowchart of Feed-forward Neural Network MLP.

2.3. Site and dataset

The experimental dataset used in this work is available by the Applied Research Unit for Renewable Energies (URAER) situated in the south of Algeria about 20 km from Ghardaïa city with (latitude = +32.37, longitude = +3.77 and altitude = 450 m). Figure 1 shows the site location. Its site is characterized by exceptional sunshine where the rate of insolation is significant; in a sunshine duration that is more than 3000 (hours/year) and on a horizontal plane, it was found that the mean annual global solar radiation exceeded 6000 (Wh/m2). On the contrary, winter in Ghardaïa is recognized as extreme cold due to the windblown snow from the highlands; Temperature is high in summer as it can exceed 45ºC while being relatively cool in winter; the jellies are exceptional and small (Guermoui, Abdelaziz, et al., 2022)(Guermoui & Rabehi, 2020).

Figure 2 - Geographical Location of the subject site: Ghardaia, Algeria(Gairaa et al., 2019).

The database contains the following parameters: Day's number (ND), DGSR, daily Air Temperature (Tmin, Tmax, Tmean), Relative Humidity (RHmin, RHmax, RHmean), Pressure (Pmin, Pmax, Pmean), Sushine Duration (SD). The data are recorded every 5 min with high precision by a radiometric station installed at the rooftop of the URAER, as shown in Figure 2. The daily data are obtained using a Trapezoidal integral on the collected data of the 5-min steps from sunrise to sunset. Figures 3 and 4 show examples of the evolution of two parameters from this dataset: daily global solar radiation (DGSR), and sunshine duration received along three years (2013, 2014 and 2015) on horizontal surface.

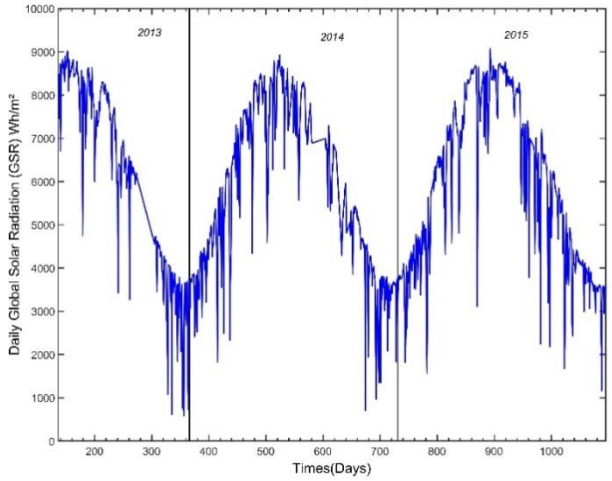

Figure 3 - The daily Global Solar Radiation.

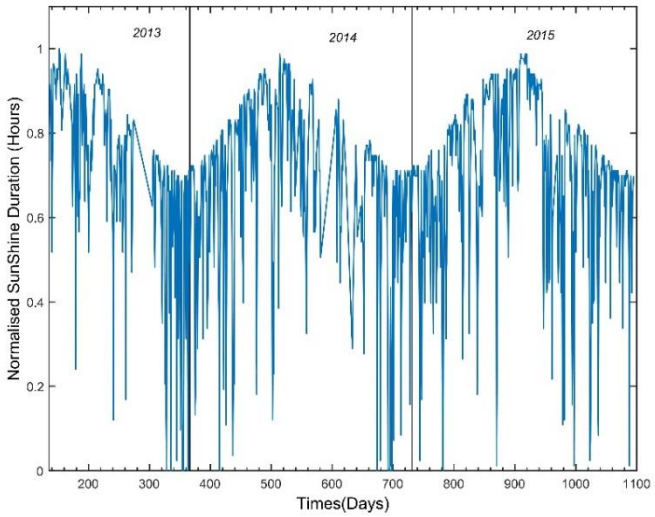

Figure 4 - An example of the used dataset predictors: "The Sunshine Duration".

3. Experimental Setup

3.1. Proposed method steps

The architecture of our proposed method to predict daily GSR consists of the following steps :

(1) Preparation of the data to be further used in the training and the test of our MLP.

(2) Selection of the relevant features of the input data using the five different feature selection techniques mentioned in section 4.4.

(3) Creation of an MLP.

(4) Training and testing the proposed model using train and test data, respectively.

(5) Validation of the model using an objective evaluation based on Mean Absolute Bias Error (MABE), Root Mean Square Error (RMSE), Normalised Root Mean Square Error (MMSE) and Determination Coefficient (r) detailed in section 4.4.

4.2. Data pre processing

The collected data are scaled linearly to the range [0, 1] using the following equation:

$$Xnor_i = \frac{x_i - \min(X)}{\max(X) - \min(X)} \quad (7)$$

Where X is an input data vector, $Xnor_i$ is the normalized value of the input x_i, min and max are the minimum and the maximum values.

After the scaling, the data is then divided into two major sets training and test.The former is used for model hyper parameters tuning (700 values), while the latter (142 values) is used for examining the behavior of the studied model. In addition, 100 samples from the training sat are used for the validation to prevent the overfitting phenomena.

4.3. MLP Setup

An MLP is used to realize the function of daily GSR prediction. MLPS are strong predictors which can learn non-linear mappings between input and output data. The MLP used in our work contains two hidden layers of 10 sigmoidal neurons each and an output layer of one linear neuron. We have chosen this MLP configuration through the trial and error principle after testing several architectures with different numbers of layers and different neurons in each layer. Furthermore, the Levenberg-Marquardt backpropagation algorithm adapts connection weights during the training process. We have chosen the latter because of its high efficiency and speed, particularly with small training data, as in our case in this work.

4.4. Model Validation Metrics

The estimated values and the measured values, when compared, in this work, using different statistical indexes, results in the assessment of the performance of models: Root Mean Square Error (RMSE), Relative Square Error (rRMSE), Determination Coefficient (R2), and Correlation Coefficient(r) (Gairaa et al., 2019)(Rabehi et al., 2020).

Hence, the difference between the anticipated values considered by the model and the measured values is displayed in RMSE. The RMSE identifies the model's accuracy calculated by:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\grave{H} - H\right)^2} \quad (8)$$

where $\grave{H}$ stands for the estimated values, and H denotes the measured values.

The nRMSE can be calculated by splitting the RMSE into the average of measured data as (Guermoui, Benkaciali, et al., 2022)(Guermoui, Bouchouicha, et al., 2022):

$$\mathrm{nRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\grave{H} - H\right)^2}}{\frac{1}{N}\sum_{i=1}^{n} H} \times 100 \quad (9)$$

The judgment of the assessed model can be built upon nRMSE as follows(Rabehi et al., 2021):

The model is considered:

Excellent if: nRMSE< 10%

Good if: 10% <nRMSE< 20%

Fair if: 20% <nRMSE< 30%

And Poor if: nRMSE> 30%

r can be calculated according to the following formula, in which they indicate the strength of the linear relationship existing between the measured and predicted values:

$$r = \frac{\sum_{i=1}^{n}(H_P - \bar{H}_P).(H_M - \bar{H}_M)}{\sqrt{\sum_{i=1}^{n}(H_P - \bar{H}_P).\sum_{i=1}^{n}(H_M - \bar{H}_M).}} \quad (10)$$

The MABE, give the mean absolute value of bias error. Its expression is given by:

$$MABE = \frac{1}{n}\sum_{i=1}^{n}(H_P - H_M) \quad (11)$$

4. Results and discussion

The main goal of this work is to study the influence of the selected input predictors on the prediction of the daily GSR. As mentioned in section 2.3, our dataset has several predictors, such as the minima, the maxima and the means of temperature, humidity, and pressure, as well as the sunshine duration and the number of days. Using the whole pole of these features can increase the dimension of the input data and result in redundant noisy data, which increases the prediction process's complexity and breaks down the predicted response's performance. Therefore, our objective is to assess a set of feature selection techniques to extract relevant features and consequently increase the daily GSR prediction.

Figures (5-9) present the scores ranking of the 11 predictors of the training input data for ReliefF, Minumum CFS, Chi-square test, F_test, and NCA features selection techniques, respectively.

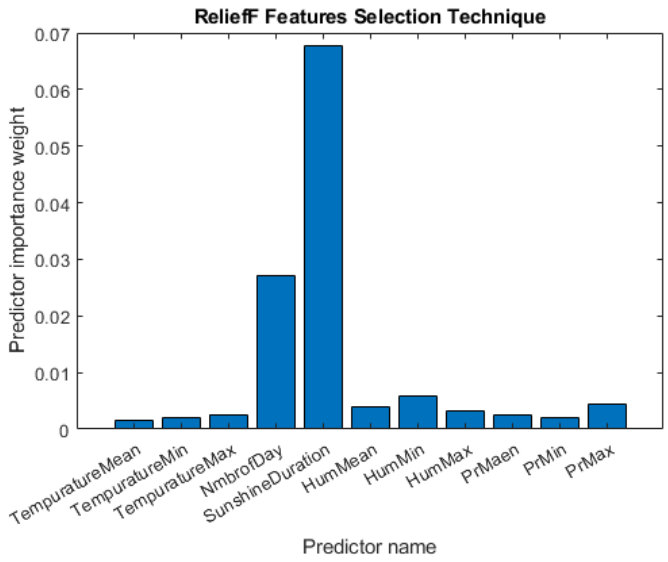

Figure 5 - Predictors' importance Scores of the 11 predictors of the training input data for ReliefF features selection technique.

According to the ReliefF features selection technique, the most interesting feature is the Sunshine duration, with a score of 0.0677, followed by the Number of the Day feature, with a score of 0.0271. The remaining feature scores vary from 0.0058 to 0.0015. The Minimum CFS technique puts out "Sunshine duration", "Pressure Max" and "number of the day" as the most relevant features with scores of 0.2028, 0.2028 and 0.2025, respectively.

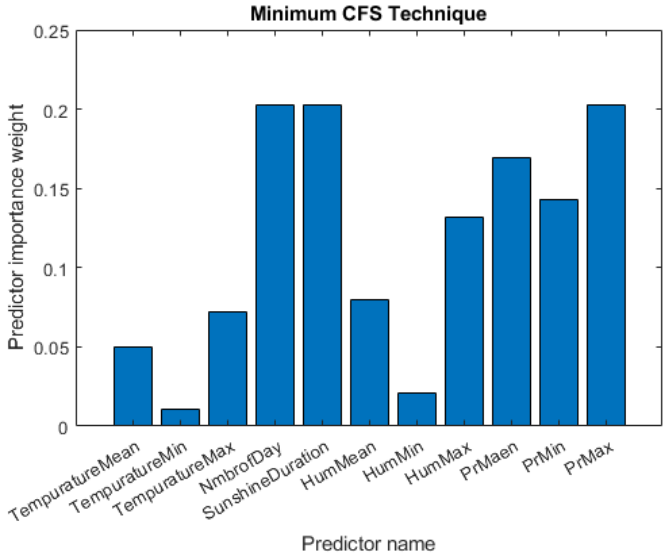

Figure 6 - Predictors' importance Scores of the 11 predictors of the training input data for minimum CFS features selection technique.

For the Chi-square test technique, "Temperature Max" and "Sunshine Duration" are selected as the most interesting predictors, with scores of 0.8872 and 0.8773, respectively. Another remark that can be seen from Figure 7 is that the predictors' relevance is very similar for this feature selection technique. They come in a tight range varied from 0.7265 to 0.8872.

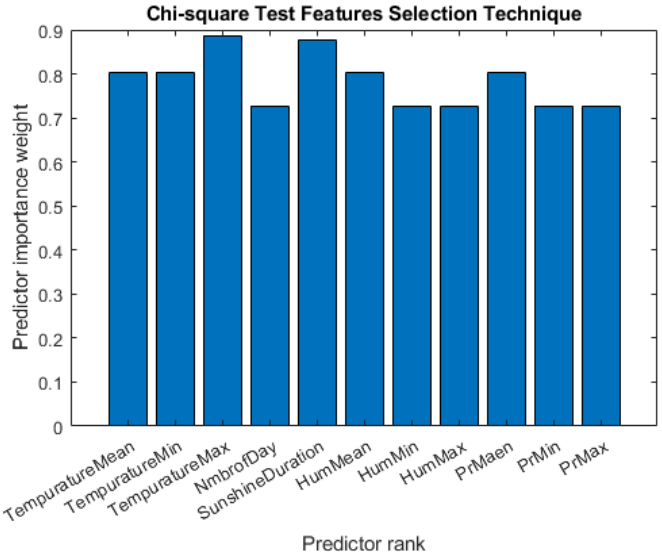

Figure 7 - Predictors' importance Scores of the 11 predictors of the training input data for Chi-Square Test features selection technique.

The f-Test features selection technique selected the "Number of the Day" feature as the more interesting feature with a score of 564.2552, followed by the "Sunshine Duration" with a score of 395.9209. The least interesting feature according to this technique is "Pressure Min" with a score

of 56.7554. Note that the scores scale differs from one feature selection technique to another due to the different methods of score computing.

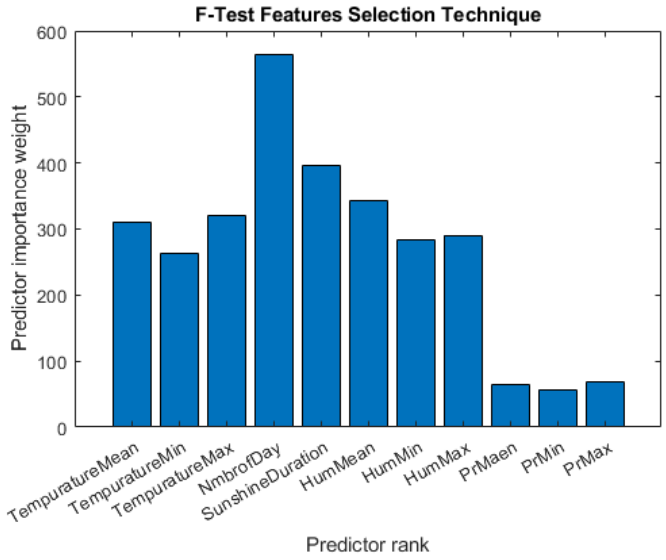

Figure 8 - Predictors' importance Scores of the 11 predictors of the training input data for F-Test features selection technique.

The last evaluated feature selection technique is the NCA. It selected "Sunshine Duration" as the more suited feature to predict the daily GSR with a score of 3.1566, followed by the "Number of the Day" feature with a score of 0.6985. NCA gives almost a score of Zero for the all-remaining features. The scores of the remaining features vary from 9.9533e-05 to 1.6970e-06.

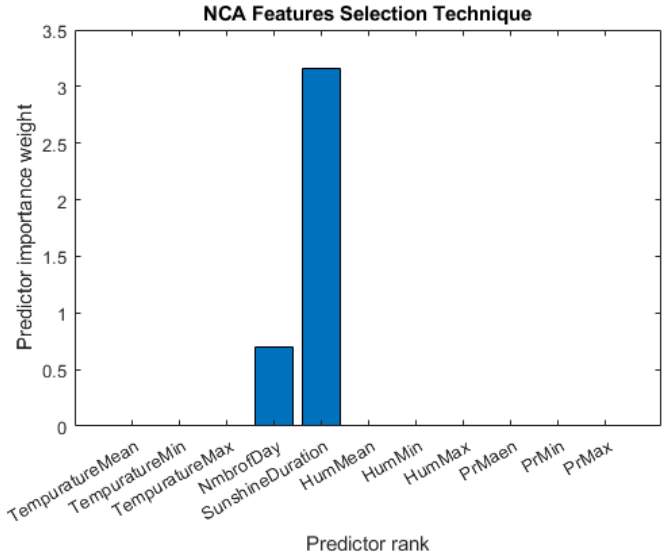

Figure 9 - Predictors' importance Scores of the 11 predictors of the training input data for minimum NCA features selection technique.

Table 1 summarizes the ranking of each of the 11 predictors for the five used features selection techniques. The seventh column of the table namely "Mean rank" presents the mean of the ranking of each predictor by all the used features selection in this study. One can note easily from the Table that the most interesting predictor for all experiments is the "Sunshine Duration" with a mean rank of 1.4. It is ranked first by three technique and second by two techniques. The second

interesting predictor is the “Number of the Day” feature with a mean rank of 3. The worst predictor is the “Pressure Min” with a rank mean of 8.2.

Table 1 - The ranking of each of the 11 predictors according to all FS techniques

Predictor	Ranking according to					Mean rank	Ranking according to Mean rank
	ReliefF	Min CFS	Chi-squaretest	F-test	NCA		
‘TempuratureMean'	11	9	5	5	10	8	9
‘TempuratureMin'	9	11	3	8	5	7,2	7
‘TempuratureMax'	7	8	1	4	3	4,6	3
‘NmbrofDay'	2	3	7	1	2	3	2
‘SunshineDuration'	1	1	2	2	1	1,4	1
‘HumMean'	5	7	4	3	6	5	4
‘HumMin'	3	10	10	7	4	6,8	6
‘HumMax'	6	6	11	6	11	8	9
‘PrMaen'	8	4	6	10	8	7,2	7
'PrMin'	10	5	8	11	7	8,2	11
'PrMax'	4	2	9	9	9	6,6	5

Detailed numerical forecasting findings are displayed in Table 2. It presents the obtained values of MABE, RMSE, nRMSE, and r between the measured and predicted daily GSR in function of the number of selected predictors from the input data using ReliefF, Min CFS, Chi-square Test, F-Test, and NCA features selection techniques. It is worth noting that selected predictors for each feature selection technique start from the most to the worst relevant predictors according to this technique and as presented in Table 1. For instance, when we say that we have taken only one predictor for the ReliefF technique, this predictor is the ”Sunshine Duration”, as it is the most relevant predictor for this technique. By the same way, if we take two predictors for the min CFS for example, then the taken predictors are the two best predictors for this FS technique, which are ‘SunshineDuration' and 'PrMax' as reported in Table 1, and so on.

As can be seen from Table 2, all models show reasonable forecasting accuracy when speaking about statistical indexes. For instance, the best r for all the FS techniques varies from 98.02 (corresponds to Min CFS using three predictors) to 98.44 (corresponds to NCA using three predictors too). Another remark is that Min CFS, F-test and NCA achieves their best statistical indexes values for a number of selected predictors of three only, which present a good dimensionality reduction of the input data. Chi-square technique achieved its best for a number of selected predictors of seven; however, RelifF is the worst technique according to this point of comparison as it is necessary for it to use all the 11 predictors to reach its best.

Overall experiments show that the NCA technique outperforms others as it records the best scores for the four used statistical indicators with values of 0.0176, 0.0266 and 5.0575 and 98.44 for MABE, RMSE, nRMSE and r, respectively. This best result corresponds to a number of only three selected predictors, which are “Sunshine duration”, “Number of the Day” and “Temperature Max”.

Table 2 - The performance of daily GSR forecasting in function of the number of selected predictors using different features selection techniques.

	Number of Predictors	MABE (MJ/m2)	RMSE (MJ/m2)	nRMSE	r (%)
	1	0,0960	0,1126	21,3892	0,6676
	2	0,0204	0,0323	6,1394	0,9769
	3	0,0201	0,0306	5,8095	0,9793
	4	0,0198	0,0312	5,9163	0,9786
	5	0,0208	0,0296	5,6178	0,9807
ReliefF	6	**0,0196**	0,0301	5,7188	0,9800
	7	0,0199	0,0295	5,6069	0,9808
	8	0,0205	0,0310	5,8910	0,9788
	9	0,0214	0,0315	5,9732	0,9781
	10	0,0242	0,0333	6,3300	0,9754
	11	0,0202	**0,0291**	**5,5322**	**0,9813**
	1	0,0933	0,1121	21,2915	0,6714
	2	0,0564	0,0764	14,5054	0,8632
	3	**0,0191**	**0,0299**	5,6831	**0,9802**
	4	0,0205	0,0316	5,9960	0,9780
	5	0,0199	0,0312	5,9225	0,9785
Min CFS	6	0,0203	0,0306	5,8184	0,9793
	7	0,0216	0,0310	5,8904	0,9788
	8	0,0218	0,0322	6,1145	0,9771
	9	0,0222	0,0311	5,9034	0,9787
	10	0,0215	0,0302	5,7447	0,9798
	11	0,0212	0,0299	**5,6823**	0,9802

Table 2 –(continued)

	Number of Predictors	MABE (MJ/m2)	RMSE (MJ/m2)	nRMSE	r (%)
Chi-square Test	1	0,1436	0,1787	33,9306	0,6284
	2	0,0693	0,0829	15,7368	0,8366
	3	0,0693	0,0832	15,8098	0,8350
	4	0,0545	0,0688	13,0660	0,8906
	5	0,0619	0,0726	13,7907	0,8773
	6	0,0507	0,0699	13,2737	0,8869
	7	**0,0200**	**0,0295**	**5,5993**	0,9808
	8	0,0203	0,0296	5,6181	0,9807
	9	0,0255	0,0343	6,5069	0,9740
	10	0,0210	0,0279	5,3054	**0,9828**
	11	0,0202	0,0313	5,9449	0,9784
F_test	1	0,0688	0,1044	19,8337	0,7235
	2	0,0197	0,0310	5,8859	0,9788
	3	0,0201	**0,0294**	**5,5748**	**0,9810**
	4	0,0215	0,0296	5,6294	0,9806
	5	**0,0196**	0,0295	5,5985	0,9808
	6	0,0209	0,0309	5,8725	0,9789
	7	0,0202	0,0294	5,5926	0,9809
	8	0,0220	0,0328	6,2210	0,9763
	9	0,0221	0,0334	6,3384	0,9754
	10	0,0213	0,0311	5,9046	0,9787
	11	0,0239	0,0324	6,1565	0,9768

Table 2 –(continued)

	Number of Predictors	MABE (MJ/m2)	RMSE (MJ/m2)	nRMSE	r (%)
	1	0,0873	0,1057	20,0676	0,7156
	2	0,0198	0,0309	5,8750	0,9789
	3	**0,0176**	**0,0266**	**5,0575**	**0,9844**
	4	0,0202	0,0303	5,7555	0,9797
	5	0,0191	0,0306	5,8034	0,9794
NCA	6	0,0194	0,0279	5,3072	0,9828
	7	0,0200	0,0289	5,4822	0,9816
	8	0,0192	0,0280	5,3103	0,9828
	9	0,0206	0,0300	5,6989	0,9801
	10	0,0195	0,0289	5,4852	0,9816
	11	0,0207	0,0304	5,7745	0,9796

A sample from forecasting errors reported in Table 3, which is MABE is vividly visualized in Figure 10 in order to deliver a graphical visualization of the obtained results.

Even that all the used FS techniques in this work are starting with a poor performance of prediction using only one or two predictors. However, the performance increases and reaches excellent values beyond the use of three predictors or more for all the features techniques, excluding the Chi-square test technique. For these latter features selection technique, statistics remain weak for a number of selected predictors varied from 1 to 6. For a number of predictors equal to or greater than 7, the Chi-square technique performance also increases and achieves comparable performance to the remaining techniques.

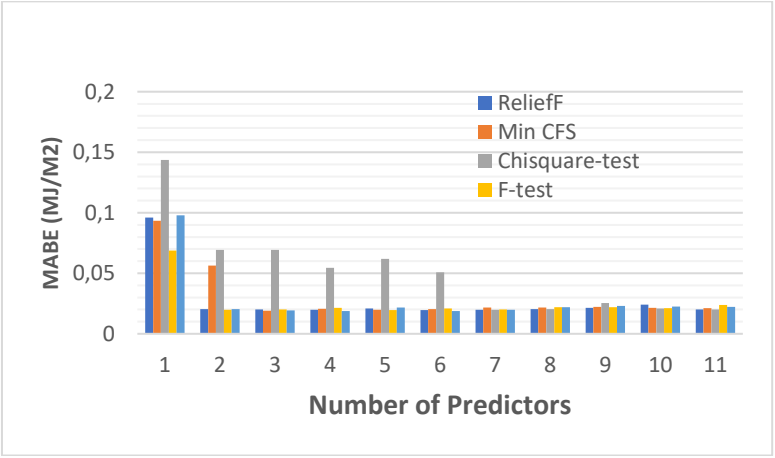

Figure 10 - The MABE in function of the number of selected predictors for the five different used features selection techniques

Furthermore, it can be seen clearly from figure 10 that all the features' techniques have achieved an MABE error close to the value 0.02, which is significantly low and proves the effectiveness of the proposed method.

An example of the obtained regression between the measured and the predicted daily GSR using our proposed method is presented in Figure 11. It refers to the best-obtained result in overall our experiments, which corresponds to three selected predictors using the NCA technique. It can be seen from the graph that the dispersion between the measured and predicted samples of data is very low, which proves the high performance of the proposed model.

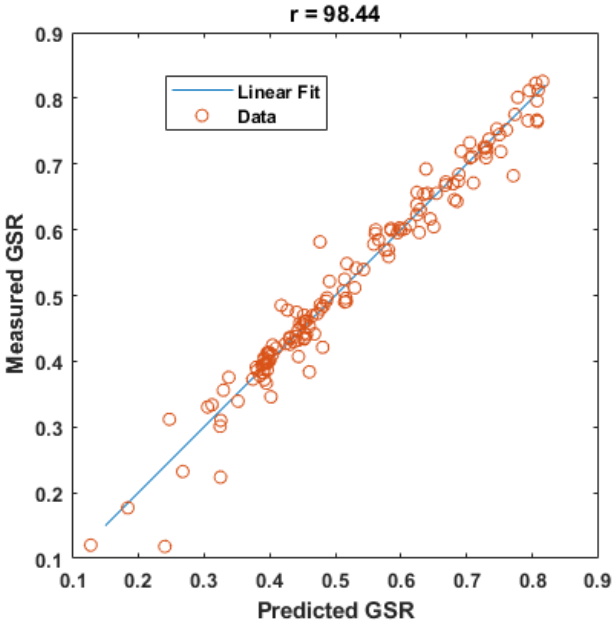

Figure 11 - Correlation between measured and predicted Global Solar Radiation using our proposed method

Furthermore, prediction outputs based on the proposed method against the measured data are shown in Figure 12 to graphically analyze the best-found model's performance. Our proposed model achieves good agreement between the measured and predicted values. The two graphs look very close, which again proves our proposed method's excellent performance.8,5.

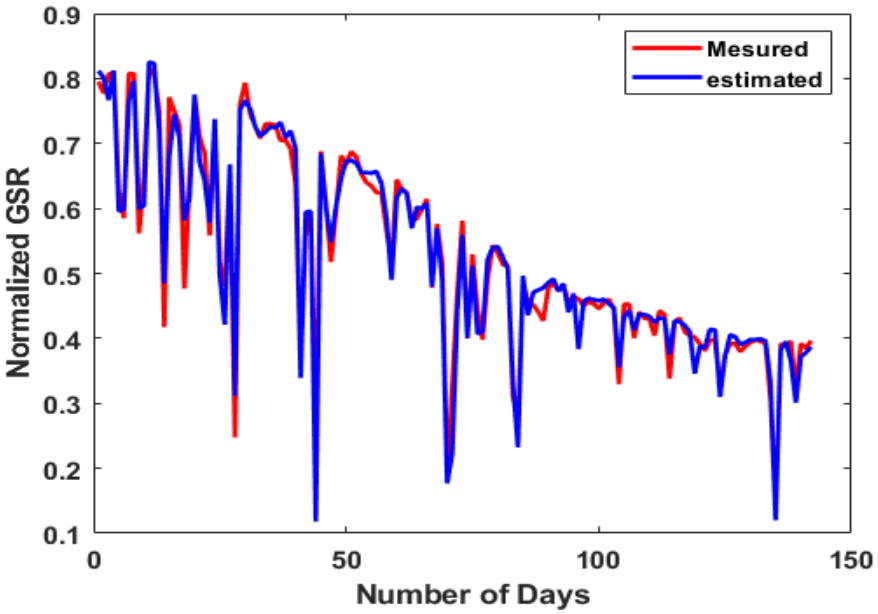

Figure 12 - Comparison between the normalized values of measured and predicted Global Solar Radiation using our proposed method

5. Conclusion

This paper compares several feature selection techniques proposed to select the best combinations of input data predictors for estimating the daily global solar radiation. Five different feature selection techniques are investigated using a daily database of three years of measurements in a semi-arid climate. It has been found that Neighborhood Component Analyses slightly outperform other techniques. Furthermore, we have found that the "Sunshine duration" is the most relevant feature, followed by "the number of the Day". However, using only one or two predictors cannot result in good prediction performance. Using the three best predictors for most of the feature selection techniques is enough to provide a good trade-off between acceptable performance of daily GSR prediction and a reasonable dimensionality reduction.

As a perspective, one can elaborate on this study by considering additional cases studies with different climate conditions. Furthermore, we suggest testing other techniques to search the best predictors combinations based on optimization methods like Genetic Algorithms.

References

Alisha Sikri, N. P. S. (2023). Chi-Square Method of Feature Selection: Impact of Pre-Processing of Data. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(3s), 241–248. https://ijisae.org/index.php/IJISAE/article/view/2680

Asl, S. F. Z., Karami, A., Ashari, G., Behrang, A., Assareh, A., & Hedayat, N. (2011). Daily global solar radiation modeling using multi-layer perceptron (MLP) neural networks. *World Academy of Science, Engineering and Technology*, *79*, 740–742. https://doi.org/10.1016/j.energy.2011.02.048.K.

Behrang, M. A., Assareh, E., Ghanbarzadeh, A., & Noghrehabadi, A. R. (2010). The potential of different artificial neural network (ANN) techniques in daily global solar radiation modeling based on meteorological data. *Solar Energy*, *84*(8), 1468–1480. https://doi.org/10.1016/j.solener.2010.05.009

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. https://doi.org/10.1016/j.enconman.2013.06.034

Dhanya, R., Paul, I. R., Akula, S. S., Sivakumar, M., & Nair, J. J. (2020). F-test feature selection in Stacking ensemble model for breast cancer prediction. *Procedia Computer Science*, *171*(2019), 1561–1570. https://doi.org/10.1016/j.procs.2020.04.167

Gairaa, K., Benkaciali, S., & Guermoui, M. (2019). Clear-sky models evaluation of two sites over Algeria for PV forecasting purpose. *European Physical Journal Plus*, *134*(10), 1–17. https://doi.org/10.1140/epjp/i2019-12917-2

Guermoui, M., Abdelaziz, R., Gairaa, K., Djemoui, L., & Benkaciali, S. (2022). New temperature-based predicting model for global solar radiation using support vector regression. *International Journal of Ambient Energy*, *43*(1), 1397–1407. https://doi.org/10.1080/01430750.2019.1708792

Guermoui, M., Benkaciali, S., Gairaa, K., Bouchouicha, K., Boulmaiz, T., & Boland, J. W. (2022). A novel ensemble learning approach for hourly global solar radiation forecasting. *Neural Computing and Applications*, *34*(4), 2983–3005. https://doi.org/10.1007/s00521-021-06421-9

Guermoui, M., Bouchouicha, K., Benkaciali, S., Gairaa, K., & Bailek, N. (2022). New soft computing model for multi-hours forecasting of global solar radiation. *European Physical Journal Plus*, *137*(1), 1–28. https://doi.org/10.1140/epjp/s13360-021-02263-5

Guermoui, M., & Rabehi, A. (2020). Soft computing for solar radiation potential assessment in Algeria. *International Journal of Ambient Energy*, *41*(13), 1524–1533. https://doi.org/10.1080/01430750.2018.1517686

Jacob Goldberger, Sam Roweis, Geoff Hinton, & Ruslan Salakhutdinov. (2005). Neighbourhood components analysis. *Advances in Neural Information Processing Systems*, *17*, 513–520.

Khorasanizadeh, H., & Mohammadi, K. (2013). Introducing the best model for predicting the monthly mean global solar radiation over six major cities of Iran. *Energy*, *51*, 257–266.

https://doi.org/10.1016/j.energy.2012.11.007
Kira, K., & Rendell, L. A. (1992). A Practical Approach to Feature Selection. *In Machine Learning Proceedings*, 249–256.
Kira, Kenji, and L. A. R. (n.d.). The feature selection problem: Traditional methods and a new algorithm. *In Aaai*, *2*(1992a), 129–134.
Kiziltan, & Şahin, M. (2016). Calculation of Solar Radiation by Using Regression Methods. *Journal of Physics: Conference Series*, *707*(1), 012049. https://doi.org/10.1088/1742-6596/707/1/012049
Koca, A., Oztop, H. F., Varol, Y., & Koca, G. O. (2011). Estimation of solar radiation using artificial neural networks with different input parameters for Mediterranean region of Anatolia in Turkey. *Expert Systems with Applications*, *38*(7), 8756–8762. https://doi.org/10.1016/j.eswa.2011.01.085
Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *784 LNCS*, 171–182. https://doi.org/10.1007/3-540-57868-4_57
Lam, J. C., Wan, K. K. W., & Yang, L. (2008). Solar radiation modelling using ANNs for different climates in China. *Energy Conversion and Management*, *49*(5), 1080–1090. https://doi.org/10.1016/j.enconman.2007.09.021
Mellit, A., Mekki, H., Messai, A., & Kalogirou, S. A. (2011). FPGA-based implementation of intelligent predictor for global solar irradiation, Part I: Theory and simulation. *Expert Systems with Applications*, *38*(3), 2668–2685. https://doi.org/10.1016/j.eswa.2010.08.057
Mellit, A., & Pavan, A. M. (2010). A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste, Italy. *Solar Energy*, *84*(5), 807–821. https://doi.org/10.1016/j.solener.2010.02.006
Rabehi, A., Guermoui, M., Khelifi, R., & Mekhalfi, M. L. (2020). Decomposing global solar radiation into its diffuse and direct normal radiation. *International Journal of Ambient Energy*, *41*(7), 738–743. https://doi.org/10.1080/01430750.2018.1492445
Rabehi, A., Rabehi, A., & Guermoui, M. (2021). Evaluation of Different Models for Global Solar Radiation Components Assessment. *Applied Solar Energy (English Translation of Geliotekhnika)*, *57*(1), 81–92. https://doi.org/10.3103/S0003701X21010060
Rahimikhoob, A. (2010). Estimating global solar radiation using artificial neural network and air temperature data in a semi-arid environment. *Renewable Energy*, *35*(9), 2131–2135. https://doi.org/10.1016/j.renene.2010.01.029
Ramedani, Z., Omid, M., & Keyhani, A. (2013). Modeling solar energy potential in a tehran province using artificial neural networks. *International Journal of Green Energy*, *10*(4), 427–441. https://doi.org/10.1080/15435075.2011.647172
Rehman, S., & Mohandes, M. (2008). Artificial neural network estimation of global solar radiation using air temperature and relative humidity. *Energy Policy*, *36*(2), 571–576. https://doi.org/10.1016/j.enpol.2007.09.033
Robnik-Šikonja, M., & Kononenko, I. (1997). An adaptation of Relief for attribute estimation in regression. *Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97)*, *5*, 296–304.
Şahin, M., Kaya, Y., Uyar, M., & Yıldırım, S. (2014). Application of extreme learning machine for estimating solar radiation from satellite data. *International Journal Of Energy Research*, *38*(4), 205–212. https://doi.org/10.1002/er.3030
Voyant, C., Muselli, M., Paoli, C., & Nivet, M. L. (2011). Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation. *Energy*, *36*(1), 348–359. https://doi.org/10.1016/j.energy.2010.10.032
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. https://doi.org/10.1016/j.rser.2013.12.008