Estimation of Daily Direct Normal Solar Irradiation Using Machine-Learning Methods

Zineb Bounoua and Abdellah Mechaqrane

Abstract The sizing and simulation of all solar systems require the availability of reliable measurements of solar radiation at different time steps. Unfortunately, solar radiation measurements are not readily available for most worldwide locations. For this reason, it is desirable to develop accurate prediction models by developing relationships between available meteorological data and solar irradiation. Artificial Neural Networks (ANN) have been widely used for the estimation of different solar irradiation components. Recently, some machine learning methods have been reported and appear to be very promising. In this paper, we are interested in comparing the performance of ANN and three ensemble methods (Bagging, Boosting and Random Forests) in estimating the daily direct normal (DNI) solar irradiation from some commonly measured meteorological variables. Our study is performed using measurements data from five Moroccan cities: Oujda, Missour, Erfoud, Zagora, and Tan-Tan. The achieved results show that all developed models give good performances on training and validation datasets with a normalized Root Mean Squared Error (nRMSE) < 20%.

Keywords Direct normal irradiation · MLP · Boosting · Bagging · Random forest

1 Introduction

After the energy crisis of 1973, researchers devoted an important part of their works to the development of alternative solutions to increase the independence of fossil fuels, while taking into account the environment protection. Of course, the use of solar energies is an obvious solution, which represents enormous advantages. After several difficulties, solar energy has gained a lot of interest in recent years. Indeed, several countries, including Morocco, have opted to diversify their energy mix by installing large photovoltaic or concentrated solar power plants, or both. However, and as it well known, solar energy is a fluctuating and intermittent renewable energy source.

Z. Bounoua (✉) · A. Mechaqrane
SIGER, Intelligent Systems, Georesources and Renewable Energies Laboratory, Faculty of Sciences and Techniques, Sidi Mohamed Ben Abdellah University, PO. Box 2202 Fez, Morocco
e-mail: zineb.bounoua@usmba.ac.ma

S. Bennani et al. (eds.), *WITS 2020*, Lecture Notes in Electrical Engineering 745,
https://doi.org/10.1007/978-981-33-6893-4_58

Indeed, its availability and its quantity depend on weather conditions and therefore exhibit a high degree of random variability over time. Therefore, the injection of energy produced from solar installations into an electrical grid creates problems in its management. To remedy to these problems, it is important to have, in advance, an idea about the amount of energy which will be produced in a given time horizon step. Depending on the solar conversion system, it is necessary to have a good knowledge of the global solar irradiation for photovoltaic and water heating systems, or of the direct normal solar irradiation for Concentrated Solar systems (CSP) or Concentrated Photovoltaic systems (CPV).

Solar component measurements (Global Horizontal solar Irradiation (GHI), Diffuse Horizontal solar Irradiation (DHI) and Direct Normal solar Irradiation (DNI)) are not always available in most regions of the world due to high equipment costs and maintenance problems. This strong demand for local and regional solar radiation and meteorological information databases prompted several researchers to develop different techniques for predicting solar radiation in a location using commonly measured meteorological variables.

According to the literature, Artificial Neural Networks (ANN) techniques have been widely employed for solar radiation prediction [1]. Their high performance and ability to solve complex non-linear problems by learning the relationship between inputs and outputs of systems were the main reasons for their wide use in different fields [2]. Mubiru [3] developed a prediction model using ANN to estimate the daily average monthly direct solar radiation for sites in Uganda. The results showed good agreement between the estimated and measured values of direct solar irradiation with a mean square error of 0.197 MJ/m^2.

Nowadays, several new machine-learning methods are increasingly used for solar radiation prediction. Hassan et al. [4] have explored the potential of tree-based ensemble methods (gradient boosting, bagging, and random forest (RF)) in estimating global, diffuse, and normal solar radiation components in daily and hourly time-scales. These models were tested on five sites in the MENA region and compared to multi-layer perceptron (MLP), support vector regression (SVR), and decision tree (DT) models, mentioning that the overcast-sky daily observations are eliminated before applying the machine-learning methods. Achieved results showed that ensemble models are more stable than decision trees and MLP and have close performances to the optimized SVR models, with significantly less computational costs.

It should be stressed that the majority of research works concern the GHI prediction. For the DNI component, the most existing models have been developed based on Weather Research and Forecasting methods [5], satellite methods [6], and image processing techniques [7].

The main objective of this work is to evaluate the ability of ensemble methods based on regression trees (Bagging, Boosting, and Random Forests) in estimating the daily direct normal solar irradiation (DNI). A comprehensive study aiming at assessing and comparing the performance of these new methods will play a crucial role in choosing the best method for estimating solar irradiation. First, each model is trained and tested using data from each city individually, as well, using another database, which is constructed by aggregating data from the five cities (All cities). Such a comparison performed for different climates will lead to a more realistic and general conclusion about the models' suitability. To evaluate the effectiveness of these models, they were compared to the Multi-Layers Perceptron ANN (MLP-ANN) model. The accuracy of these models was assessed based on three major statistical indicators (R, nRMSE and nMAE). This work will help the researchers in evaluating the daily solar energy resource, which is essential to determine the performance of different solar energy systems and study their feasibility at any location with similar climatic condition.

2 Methodology

2.1 Multi-layer Perceptron Artificial Neural Network (MLP-ANN)

Inspired by biological neural networks, Artificial Neural Networks (ANN) represent the simplest models that have a great ability to solve complex nonlinear problems by deducing the relationship between inputs and outputs in several areas of prediction, optimization, and classification. In this study, we used a Multilayer Perceptron Neural Network model (MLP) applying the feed-forward backpropagation, and the Levenberg-Marquardt algorithm as an approximation function to reduce the error of nonlinear functions. It consists of M inputs, one hidden layer with N tangent hyperbolic neurons, and a linear output neuron. Figure 1 shows the MLP-ANN architecture adopted in this study.

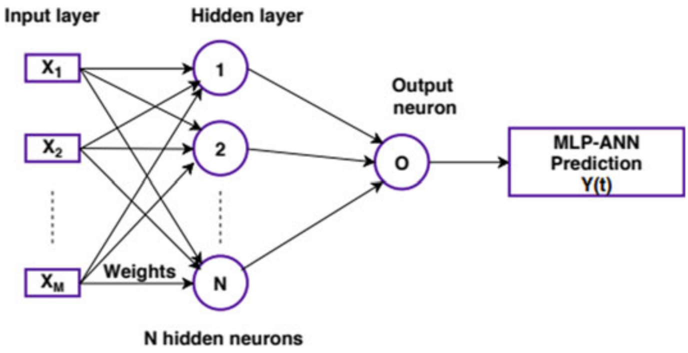

Fig. 1 MLP-ANN architecture

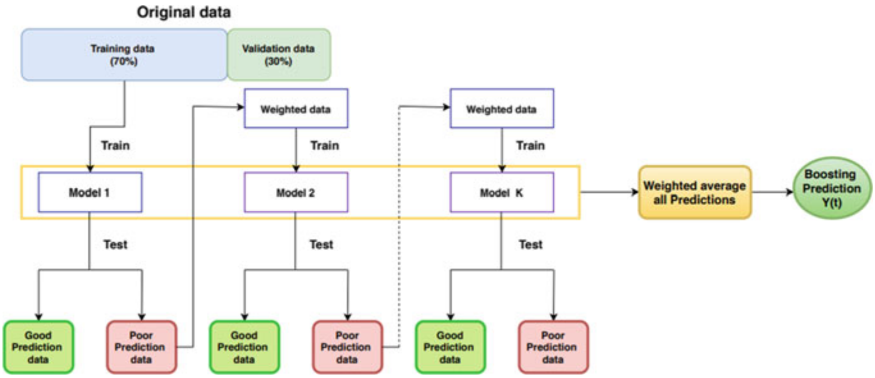

Fig. 2 Gradient boosting process diagram

2.2 Gradient Boosting

The most popular algorithm in the ensemble methods family is called Boosting and was introduced in 1996 by FREUND and SCHAPIRE [8]. The main characteristic of this algorithm is its ability to convert weak learners into strong learners by: (a) sequentially associating the outputs of individual weak learners during the training phase, and (b) by assigning, at each iteration, higher weights to poorly predicted samples.

The type of boosting algorithm chosen to solve our regression problem is Gradient Boosting, It is an algorithm, which creates the model by steps. Each model is obtained by optimizing the absolute differentiable loss function. For each weak learner added, a new model is created which gives a more accurate estimate of the output. The gradient boosting process diagram is presented in Fig. 2.

2.3 Bagging

Leo BREIMAN developed the bagging algorithm in 1996 [9]. This algorithm is based on two main steps; bootstrap and aggregation. Bagging adopts the bootstrap distribution to generate different base learners by applying bootstrap sampling. A sample that is taken from the original dataset with replacement is called a bootstrap sample, which provides the data subsets for training the base learner. When this process is repeated several times, several training samples can be obtained and from each sample, a base learner is trained using the base learning algorithm. For the regression problems, Bagging adopts a popular strategy to aggregate the base learner's outputs by calculating the average of the obtained estimations from each sample. This process is explained as a diagram in Fig. 3.

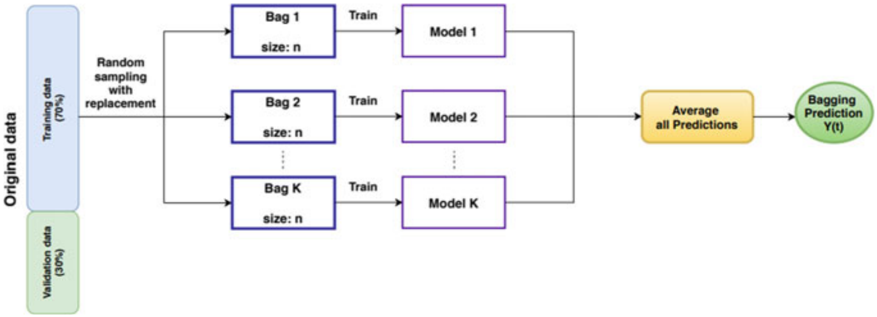

Fig. 3 Bagging process diagram

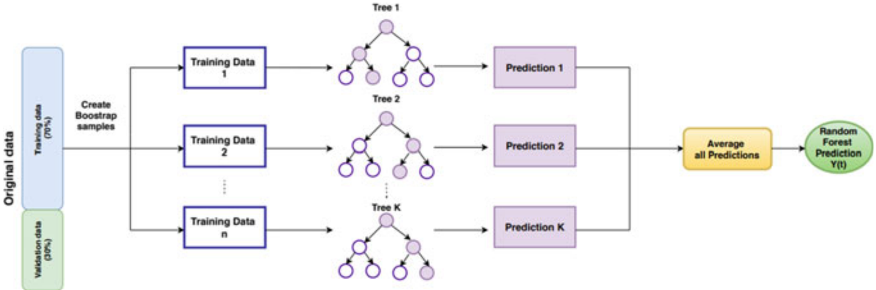

Fig. 4 Random forest process diagram

2.4 *Random Forest*

Random forests are part of the machine learning techniques proposed by Breiman, which combines several random trees constructed using the Bagging method [10]. This method represents the association of a number of predictors. A regression tree is developed for each sample with a random selection of input variables and an arbitrary selection of a subset of predictors. The generalization error tends towards a limit only if the number of trees initially selected in the forest is high. The final estimate of the set is obtained by calculating the average estimate of all developed trees. The process diagram of Random Forest is shown in Fig. 4.

2.5 *Performance Evaluation*

Trained models are evaluated using three statistical indicators, the coefficient of correlation (R), the normalised mean absolute error (nMAE), and the normalised root mean square error (nRMSE). Normalized MAE and RMSE are used to avoid the scale dependency. A low value of nRMSE indicates a very good prediction model.

$$R = \frac{\sum_{i=1}^{n}\left(y_{ei} - \overline{y}_e\right)\left(y_{mi} - \overline{y}_m\right)}{\sqrt{\sum_{i=1}^{n}\left(y_{ei} - \overline{y}_e\right)^2 \sum_{i=1}^{n}\left(y_{mi} - \overline{y}_m\right)^2}} \tag{1}$$

$$nMAE = \frac{\frac{1}{n}\sum_{i=1}^{n}|(y_{ei} - y_{mi})|}{\overline{y}_m} \tag{2}$$

$$nRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{ei} - y_{mi})^2}}{\overline{y}_m} \tag{3}$$

With y_{mi} and y_{ei} are the ith measured value and the ith estimated value, respectively. $\overline{y}_m$ and $\overline{y}_e$ are the mean value of n measurements of y_{mi} $\left(\overline{y}_m = \frac{1}{n}\sum_{i=1}^{n} y_{mi}\right)$ and the mean value of n estimations of y_{ei} $(\overline{y}_e = \frac{1}{n}\sum_{i=1}^{n} y_{ei})$, respectively.

3 Locations and Databases

In this work, daily solar and meteorological data are measured at five locations in Morocco as part of the enerMENA project. This project was initiated in 2010, by the Institute of Solar Research at the German Aerospace Center (DLR). The "thermal sensor" stations installed in Oujda, Missour and Tan-Tan include a CHP1 pyrheliometer to measure DNI and two Kipp & Zonen CMP21 pyranometers to measure DHI and GHI. Whereas, the Erfoud and Zagora stations are equipped with a "Rotating Shadowband Irradiometer" (RSI), it is the least expensive technique for measuring the two solar components (GHI and DHI). These stations also contain temperature, relative humidity and pressure sensors, as well as an anemometer to measure the wind speed and direction at 10 m height. The measurement period is between 2013 and 2015 for Erfoud, Zagora and TanTan, between 2014 and 2015 for Missour, and between 2011 and 2015 for Oujda. Table 1 shows the geographical coordinates of each station and the used instruments for measuring solar irradiation.

4 Results and Discussions

In this study, 70% of historic data is used for training phase and the network is adjusted according to its error. The remaining 30% of the data provides independent validation of network.

Building an efficient estimation model requires a good choice of relevant input variables with the most significant impact on the output variable fluctuations. For this, some possible combinations between the input parameters have been tested using an ANN-MLP model with a variable architecture, the structure of each trained model is

Table 1 Moroccan stations of the enerMENA network

Site	Latitude (°N)	Longitude (°W)	Altitude (m)	Measurement period	Station type	Measured solar components	Type of instruments	Model
OUJDA	34.650	1.900	617 m	18/08/2011 to 30/09/2015	TS	GHI DHI DNI	Pyranometer Pyranometer Pyrheliometer	Kipp & Zonen CMP21 Kipp & Zonen CMP21 Kipp & Zonen CHP1
MISSOUR	32.860	4.107	1107 m	01/01/2014 to 30/09/2015	TS	GHI DHI DNI	Pyranometer Pyranometer Pyrheliometer	Kipp & Zonen CMP21 Kipp & Zonen CMP21 Kipp & Zonen CHP1
ERFOUD	31.491	4.218	859 m	02/06/2/013 to 30/09//2015	RSI	GHI DHI	Pyranometer Pyranometer	RSI PY67982
ZAGORA	30.272	5.852	783 m	03/06/2013 to 30/09/2015	RSI	GHI DHI	Pyranometer Pyranometer	RSI PY75722
TAN-TAN	28.498	11.322	75 m	01/06/2013 to 30/09/2015	TS	GHI DHI	Pyranometer Pyranometer Pyrheliometer	Kipp & Zonen CMP21 Kipp & Zonen CMP21 Kipp & Zonen CHP1

optimized using a different number of hidden neurons (from 1 to 20) and evaluated on five runs. The performed tests reveal that the best combination to estimate daily DNI is the same for all the five stations, and it is composed of the day of the year, daily normal top of atmosphere solar irradiation, daily mean air temperature and the daily clearness index (J_{an}, $\text{G}_{\text{toa\_n}}$, T_{mean}, K_t), with 10 neurons in the hidden layer.

Table 2 presents the performance of daily DNI estimation models obtained by applying the three ensemble methods and the artificial neural network for each used database. The results corresponding to the best estimation model are shaded in gray.

Despite differences in climate and geographical positions between stations, it can be observed that the performances obtained by grouping the five stations data (Erfoud, Missour, Oujda, Tan-Tan and Zagora) for all developed methods, vary between 94.79 and 95.44% for R-values and between 11.87 and 12.65% for nRMSE values. The results obtained at Erfoud progress from 94.96 (Boosting) to 95.30% (RF) for R-values and from 11.69 (Boosting) to 11.35% (RF) for nRMSE values. The four machine learning methods applied to Missour gave results ranging between 95.97 and 96.25% for R-values and between 10.86 and 11.17% for nRMSE values. In Oujda, R-values vary between 96.46 and 96.76% and nRMSE values between 10.40 and 10.80%. The performance results of Zagora show a variation from 94.68 to 95.30% for the R-values, and from 10.60 to 11.25% for the nRMSE values. For Tan-Tan, R-values vary between 94.75 and 95.10% and nRMSE values between 12.64 and 13.03%.

According to [11], TanTan is characterized by frequent clouds and fogs, and a high relative humidity of 83% averaged all over the year. These conditions cause the formation of small droplets of water on the measuring instruments, resulting in the most of the time of errors in solar radiation measurements as long as the instruments are not cleaned. However, this station only experienced 84 cleaning events for the period between July 26, 2014 and June 1, 2015 [11]. Hence, the slightly weak performances obtained in TanTan compared to the other stations.

In general, the Random Forest model outperforms all the other three methods (ANN, Boosting and, Bagging) in the estimation of daily DNI for all used datasets. Figures 5 and 6 represent the training and validation scatter plots of the four developed machine-learning methods for each database. A good agreement between the estimated and the measured values for all databases can be noticed.

The obtained results prove that the tree ensemble methods (Boosting, Bagging and Random Forest) can compete the MLP-ANN to estimate daily direct normal solar irradiation. The performances of the database containing all cities with different climates are encouraging and to a certain extent general. From Fig. 7, which shows the validation results of randomly 100 days of DNI estimation for all cities using the four machine learning models, it can be noted that all developed models give good performances on validation datasets with a slight better performance of Random forest estimation that closely follows the measurement values against other models.

According to this study, it is possible to have an idea about the solar potential of new cities with the same climatic conditions with no DNI solar data measurements in order to design and analyse the performance of any concentrated solar system.

Table 2 Machine learning methods performance for daily DNI estimation in train and validation phases for each station and for all grouped stations

Stations	Model	Structure	Training			Validation		
			R (%)	nMAE (%)	nRMSE (%)	R (%)	nMAE (%)	nRMSE (%)
Erfoud	ANN	10 neurons	95.347	8.573	11.071	95.081	8.832	11.602
	Boosting	238 learners	95.638	8.454	10.729	94.962	9.233	11.695
	Bagging	71 learners	97.617	6.317	8.175	94.979	9.108	11.634
	RF	139 learners	**97.560**	**9.874**	**8.290**	**95.307**	**8.785**	**11.357**
Missour	ANN	10 neurons	96.828	7.755	10.377	96.168	8.473	10.909
	Boosting	254 learners	97.214	7.317	9.778	95.977	8.671	11.177
	Bagging	48 learners	98.417	5.559	7.575	96.112	8.467	11.018
	RF	137 learners	**98.441**	**5.560**	**7.499**	**96.257**	**8.228**	**10.864**
Oujda	ANN	10 neurons	97.386	6.656	9.091	96.655	7.672	10.476
	Boosting	257 learners	97.266	6.950	9.304	96.461	8.100	10.801
	Bagging	53 learners	98.619	4.895	6.764	96.692	7.870	10.499
	RF	165 learners	**98.638**	**4.908**	**6.745**	**96.764**	**7.818**	**10.407**
Tan-Tan	ANN	10 neurons	96.629	9.018	11.437	95.024	10.214	12.733
	Boosting	275 learners	96.797	8.744	11.210	94.751	10.272	13.034
	Bagging	35 learners	97.037	8.372	10.742	94.945	10.033	12.822
	RF	185 learners	**97.452**	**7.738**	**9.954**	**95.109**	**9.828**	**12.648**
Zagora	ANN	10 neurons	95.677	7.786	10.089	95.129	8.784	10.794
	Boosting	246 learners	96.708	6.792	8.821	94.682	8.989	11.254
	Bagging	42 learners	97.793	5.614	7.412	95.103	8.650	10.928

(continued)

Table 2 (continued)

Stations	Model	Structure	Training			Validation		
			R (%)	nMAE (%)	nRMSE (%)	R (%)	nMAE (%)	nRMSE (%)
	RF	138 learners	**97.799**	**5.645**	**7.436**	**95.307**	**8.492**	**10.604**
All cities	ANN	10 neurons	95.494	9.132	11.817	95.063	9.384	12.330
	Boosting	186 learners	95.614	9.130	11.662	94.793	9.670	12.654
	Bagging	63 learners	97.927	6.203	8.135	94.867	9.602	12.567
	RF	145 learners	**98.025**	**6.092**	**7.960**	**95.440**	**9.046**	**11.877**

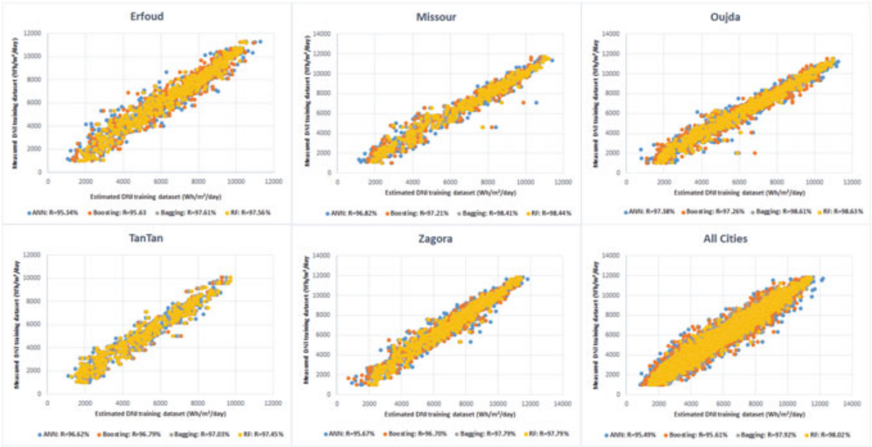

Fig. 5 Scatter plots of daily measured and estimated DNI using the four machine-learning models and the six training datasets

5 Conclusion

The objective of this work was to evaluate the potential of four machine-learning methods (Multilayer Perceptron Artificial Neural Network (MLP-ANN), Boosting, Bagging and Random Forest) in estimating the daily direct normal solar irradiation using measurement databases of five Moroccan stations (Erfoud, Missour, Oujda, Zagora and Tan-Tan). The achieved results show that all developed models give good performances on training and validation datasets (nRMSE < 20%).

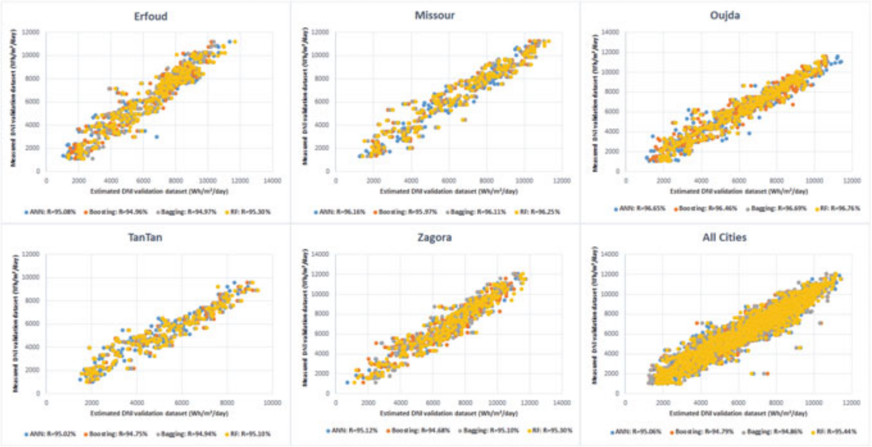

Fig. 6 Scatter plots of daily measured and estimated DNI using the four machine-learning models and the six validation datasets

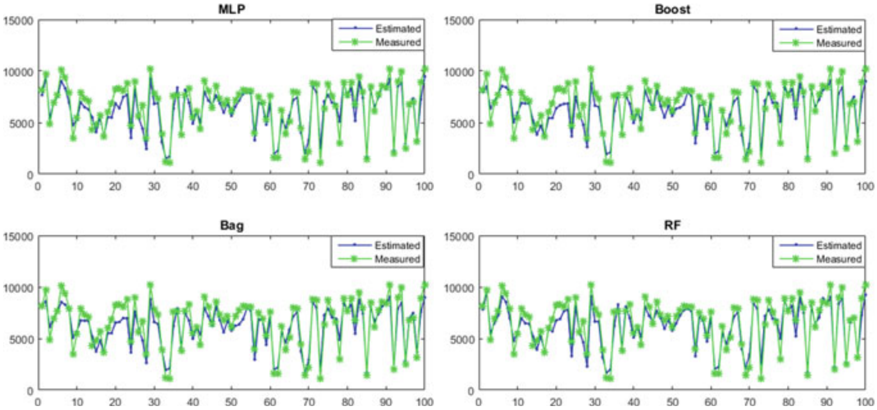

Fig. 7 Times series plot of measured and estimated daily DNI of a random selection of 100 days for all cities using the four machine learning models in the validation phase

The ensemble methods based on decisions trees (Bagging, Boosting, Random Forest) yielded a good performance and a high degree of stability. They are sufficiently efficient and give quite similar performance compared to MLP-ANN models. In addition, it was found that the Random Forest (RF) is the most efficient model for estimating daily DNI.

Finally, we can say that the ensemble methods based on decision trees can compete the Artificial Neural Networks models, which have been widely used in the domain of solar estimation and prediction. Therefore, it can be concluded that Bagging, Boosting and Random Forest are the future methods.

This work allowed to generate some reflections and to build solid ideas on prediction models. This will open up new horizons of research, and we anticipate to predict direct normal solar irradiation in different time horizons (hour, ½ hour, …, 10 min) by using the actually tested and eventually other machine learning methods.

Acknowledgements The authors would like to express their special gratitude to the German Aerospace Center (DLR) for their scientific support by making available to us the measurements of solar components and meteorological parameters used in this study.

References

1. Mohandes M, Rehman S, Halawani TO (1998) Estimation of global solar radiation using artificial neural networks. Renew Energy 14(1–4):179–184
2. Benghanem M, Mellit A, Alamri SN (2009) ANN-based modelling and estimation of daily global solar radiation data: a case study. Energy Convers Manage 50(7):1644–1655
3. Mubiru J (2011) Using artificial neural networks to predict direct solar irradiation. Adv Artif Neural Syst (2011)
4. Hassan MA et al (2017) Exploring the potential of tree-based ensemble methods in solar radiation modeling. Appl Energy 203:897–916
5. Lara-Fanego V et al (2012) Evaluation of DNI forecast based on the WRF mesoscale atmospheric model for CPV applications. In: AIP conference proceedings, vol 1477, No. 1. American Institute of Physics
6. Polo J et al (2011) Solar radiation estimations over India using Meteosat satellite images. Sol Energy 85(9):2395–2406
7. Marquez R, Coimbra CFM (2013) Intra-hour DNI forecasting based on cloud tracking image analysis. Solar Energy 91:327–336
8. Marsland S (2015) Machine learning: an algorithmic perspective. CRC Press
9. Zhou Z-H (2012) Ensemble methods: foundations and algorithms. CRC press
10. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
11. Schüler D et al (2016) The enerMENA meteorological network–solar radiation measurements in the MENA region. In: AIP conference proceedings, vol 1734, No. 1. AIP Publishing LLC