International Journal on Optimization and Applications

IJOA. Vol. 05, Issue No. 02, Year 2025, www.ijoa-journal.com

Predicting Solar Energy using Artificial Intelligence Models

GHIATE Saida

Faculty of Economics and Management Hassan First University, Laboratory of Mathematical Modeling and Economic Calculations (LM2CE), Morocco

saida.ghiate@gmail.com

jamal.zahi

Faculty of Economics and Management Hassan First University, Laboratory of Mathematical Modeling and Economic Calculations (LM2CE), Morocco

jamal.zahi@uhp.ac.ma

Abstract

Accurate forecasting of Global Horizontal Irradiance (GHI) is essential for estimating photovoltaic power generation, yet it is strongly affected by climatic variability. This study compares five machine learning methods for GHI forecasting in Morocco: Convolutional Neural Networks (CNN), Support Vector Regression (SVR), Random Forests (RF), Artificial Neural Networks (ANN), and Long Short-Term Memory networks (LSTM). The dataset was obtained from the NSRDB and covers the period from January to December 2022. Model performance was evaluated using the coefficient of determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). The results indicate that the ANN model achieves the highest accuracy, with an RMSE of 72.19 W/m², an MAE of 41.19 W/m², and an R^2 of 0.9142, demonstrating a strong agreement between predicted and observed values. This superior performance suggests that ANN models are particularly effective in capturing the nonlinear relationships between meteorological variables and GHI. The findings provide valuable insights for photovoltaic system optimization, energy planning, and predictive maintenance applications.

Keywords: Machine Learning, Global Horizontal Irradiation, Solar Forecasting, Renewable Energy, Artificial Intelligence.

XVII. INTRODUCTION

Solar energy is a virtually inexhaustible clean energy source and plays a key role in supporting the energy transition and reducing dependence on fossil fuels [8]. Today, thanks to photovoltaic systems, it provides electricity to homes, industries, and large-scale infrastructure, contributing to the sustainable development of the energy sector.

In Morocco, the energy sector plays a strategic role in economic and social development. Between 2010 and 2023, the installed capacity of renewable energy sources in Morocco has grown steadily, reflecting the country's commitment to sustainable development and energy transition. The country has committed to increasing the share of renewable energy to more than 52% of installed capacity by 2030. In 2023, national solar capacity reached approximately 831 MW, representing 7.3% of total capacity and nearly 18% of installed renewable power [2]. Annual solar production amounted to approximately 2,148 GWh, or 5.1% of national electricity production [2]. These results demonstrate the growing role of solar energy, with flagship projects such as the Noor Ouarzazate complex, one of the largest in the world.

Globally, 2023 set a record with 473 GW of new renewable capacity installed, nearly 75% of which came from solar photovoltaics [10]. This trend confirms the dominant role of this technology in the energy transition and highlights the importance of developing forecasting and optimization tools adapted to local specificities such as those in Morocco.

However, solar power generation is highly dependent on global horizontal irradiation (GHI), which varies directly with local weather conditions. Its intermittent and unpredictable nature poses a major challenge to power system stability and reliability [14].

To address this issue, accurate GHI forecasting is essential. This not only makes it possible to anticipate photovoltaic production, but also to improve energy management, operational planning and the integration of renewable sources into the grid [12]. In this context, the use of advanced machine learning techniques is a promising solution for modeling the complex relationships between climate variables and solar radiation.

This study provides a comparative analysis of several machine learning algorithms applied to WHI forecasting in Morocco using data from NSRDB for the period January–December 2022. Various approaches are used in this analysis, such as ANN, CNN, SVR, LSTM, and RF. These models were chosen for their ability to capture nonlinear relationships between climate variables. Their performance was measured using metrics such as RMSE, MAE, and R2. The objective is to identify the most effective algorithms for predicting GHI in Morocco. The results show that ANN provides the best prediction with an R2 of 0.91 and an RMSE of 72.19 W/m².

The document is structured as follows: Section 2 presents a literature review. Section 3 describes the methodology, including data collection and preprocessing. Section 4

presents the models and performance metrics. Section 5 presents the results and concludes the study. h trends in the optimization techniques and their applications in various fields.

XVIII. MOTIVATION & METHODOLOGY

A. Motivation

Time series have been extensively studied, and numerous studies have proven their effectiveness in estimating solar energy generation employing machine learning algorithms.

In study by El Maghraoui et al., the authors focused on the application of machine learning approaches to forecast energy consumption in open-pit mines [6]. Their study compared the performance of four models: ANN, SVR, DT, and RF. The results revealed that the RandomForest algorithm offered the best accuracy for this type of prediction.

Chu et al. proposes two approaches based on LSTM neural networks to develop a new solar irradiance prediction model, based on a dataset consisting of images [4]. The objective was to generate forecasts with horizons ranging from 5 to 60 minutes. The first approach uses as input variables the irradiance measured five minutes earlier, the current irradiance, and a central value extracted from the images. The second method improves these inputs by adding variance, a comparison between the red and blue channels of the image, and a three-step search technique. The experimental results showed that the second approach offered better predictive performance, highlighting the importance of leveraging more complex features derived from image processing.

In another study conducted by the same authors, the objective was to evaluate several machine learning techniques for forecasting the electricity consumption of hotel buildings employing a hotel in Shanghai as an example [5]. The algorithms studied included SVM, ANN, DT, and Random Forest (RF). The study compared their performance and identified Random Forest as the most accurate and robust, particularly for complex and nonlinear relationships between variables.

Chandola et al. built an LSTM model to predict solar irradiance 3, 6, and 24 hours in advance in dry regions of India [3]. They used a range of meteorological data such as temperature, humidity, wind, atmospheric pressure and different forms of irradiance (GHI, DHI, DNI). The model was tested on data from 2010 to 2014 collected in the Thar Desert. It performed well with a MAPE between 6.8% and 10.5%, which is quite accurate given the harsh climate.

Zhao et al. exploited a 3D-CNN model to anticipate direct normal irradiance with a 10-minute horizon [17]. The approach is based on the joint extraction of spatial and temporal features from cloud image sequences captured on the ground. The algorithm was trained using GBC images and DNI data collected between 2013 and 2014 via the NREL database. The model achieved a forecast accuracy of 17.06%.

B. Methodology

- *Data description*

The data used in this study to develop the CHI prediction model comes from the Marrakech site located at 31.6269°N latitude and 7.9881°W longitude at an altitude of approximately 466 m and characterized by a semi-arid climate with high solar exposure and low annual cloud cover. The variables considered for modeling include SHI, temperature, pressure, relative humidity, precipitable water, wind direction and wind speed covering the period from January to December 2022.

- *Data preprocessing*

Before modeling, the dataset was cleaned by processing missing values and detecting and correcting outliers. Next, a selection of explanatory variables was made in order to retain only those variables that made a significant contribution to the prediction. All variables were normalized to bring their values to the same scale. The data was then divided into training set (80%) and test set (20%). The Scikit-learn library was used to perform these operations and prepare the data for modeling.

- *Random Forest regressor*

Random Forest regressor is a statistical technique used to identify the most relevant input variables for a regression task. Each tree in the forest estimates the importance of each variable by measuring how much it reduces impurity at each split. The importance scores from all trees are then averaged to obtain a relative importance for each feature (Table 1). This approach is simple, computationally efficient, and effective in eliminating irrelevant or low-variance features, thereby improving model performance and focusing on the most informative variables. The model identified the five most predictive variables for GHI which are Relative Humidity, Wind Speed, Wind Direction, Temperature, and Precipitable Water.

. Table 1. The features importance scores

Features	Importance
Relative Humidity	0.429126
Wind Speed	0.130911
Wind Direction	0.121440
Temperature	0.094753
Precipitable Water	0.088843
Dew Point	0.056122
Pressure	0.041142
Surface Albedo	0.027324
Fill Flag	0.010339

- *the min–max scaling method*

Since the variables in the dataset have different scales (Table 2), which could affect model performance, all features were normalized using the min–max scaling method. This transformation brings all variables into the [0, 1] range using the formula:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where:

x= original value, x_{min} = minimum value of the variable, x_{max} = maximum value of the variable, x' = normalized value between 0 and 1.

Table 2. List of meteorological data in used dataset.

Meteorological variable	unit
GHI	w/m2
Temperature	°c

Pressure	mbar
Relative Humidity	%
Precipitable Water	cm
Wind Direction	Degrees
Wind speed	m/s

- *RandomizedSearchCV*

RandomizedSearchCV is a module of the Scikit-learn library used for hyperparameter optimization. It samples a predefined number of parameter combinations from specified distributions. This approach helps identify the optimal configuration by selecting the set of parameters that provides the best performance [1]. In this study, RandomizedSearchCV was applied to explore the hyperparameter space while reducing computation time.

C. Forecast models

- *The support vector Regression*

Support Vector Regression is a supervised machine learning approach widely applied for regression tasks particularly in energy forecasting applications [11,15]. SVR aims to construct a function in an N-dimensional feature space that approximates the target values within a defined margin of tolerance (ε), while simultaneously minimizing model complexity. This approach allows the algorithm to capture both linear and nonlinear relationships between input variables and the target.

- *Artificial Neural Network*

An Artificial Neural Network (ANN) is inspired by the functioning of biological neurons in the human brain. It is composed of interconnected layers of nodes, each performing weighted computations followed by activation functions to capture non-linear relationships. An ANN typically includes an input layer, one or more hidden layers, and an output layer. Each node (artificial neuron) is associated with a weight and a bias (threshold). When the weighted sum of the inputs exceeds the threshold, the neuron is activated and passes its output to the next layer [13].

- *Convolutional Neural Network*

The convolutional neural network (CNN) has become one of the most widely used models in deep learning because of its strength in identifying complex, non-linear patterns within data. A standard CNN is usually built from three main types of layers: convolutional, pooling, and fully connected layers [16]. Among these, the convolutional layer plays the central role. It applies multiple kernels to the input, producing feature maps that allow the model to capture increasingly abstract and meaningful representations.

- *Random Forest*

Random Forest (RF) is a robust and widely used supervised machine learning algorithm that builds an ensemble of decision trees and aggregates their outputs to improve prediction accuracy. It is applicable to both regression and classification tasks [7]. Each tree is trained on a random subset of the data and features, which introduces diversity and reduces correlation among trees. By averaging predictions in regression or using majority voting in classification, Random Forest mitigates overfitting and delivers more stable, accurate and generalizable results in comparison with single decision tree.

- *Long Short-Term Memory*

Long short-term memory (LSTM) neural networks are an advanced variant of recurrent neural networks (RNN), designed to mitigate the problem of gradient vanishing when processing long time series. Using a system of gates (input, forget, and output), LSTM regulates the flow of information: it decides which data should be retained, which irrelevant data should be forgotten, and which information should be transmitted to the output [9].

- *Performance Indicators*

The purpose of verification is to evaluate the quality of forecasts. Several metrics can be used to assess the performance of machine learning models, and their suitability depends on the specific use case. In this study, the forecasting methods are evaluated in terms of accuracy and efficiency using the following statistical indicators: RMSE, R^2, and MAE.

Root Mean Squared Error: RMSE is a fundamental evaluation measure based on residual squaring, it takes the square root of the output of MSE. Its problem is that it penalizes larger errors.

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(Yi - \hat{Y}i)^2} \quad (2)$$

Mean Absolute Error: The MAE calculates the absolute difference between the actual values and the predicted values. The lower the value of the MAE, the better the model, which indicates a better agreement between predicted and observed values. the formula to calculate the MAE:

$$\frac{1}{N}\sum_{i=1}^{N}|Yi - \hat{Y}i| \quad (3)$$

The coefficient of determination: The R^2 is the coefficient of determination. It is an indicator for judging the quality of a regression. It explains the strength of the relationship between an independent variable and a dependent variable in the regression model. The R^2 is calculated as follows:

$$1 - \frac{\sum_{i=1}^{N}(Yi-\hat{Y}i)^2}{\sum_{i=1}^{N}(Yi-\bar{Y}i)^2} \quad (4)$$

D. Results and Conclusion

This study aims to identify the most effective models among SVR, ANN, CNN, Random Forest, and LSTM for HSI prediction, using data from the NSRDB database over a one-year period.

Among the five models tested, ANN achieved the best performance, with the most favorable values for all evaluation metrics. The results of the models, expressed in terms of RMSE, MAE, and R^2 are presented in Table 3 and Figure 1. ANN ranks first, followed by LSTM.

The ANN model stands out for its superior accuracy, with an RMSE of 72.19 W/m², an MAE of 41.19 W/m², and an R^2 of 0.9142, indicating a strong fit between predicted and actual values. The LSTM (RMSE 75.63 W/m², R^2 0.9059) and Random Forest (RMSE 76.36 W/m², R^2 0.9040) also perform well, confirming that models capable of capturing nonlinear and temporal relationships are particularly effective for this

type of prediction. In contrast, SVR and CNN show poorer performance, suggesting that they are less effective at capturing SHI variations in this context.

The use of RandomizedSearchCV played a key role in optimizing the models. Table 3 summarizes the optimal hyperparameters for each algorithm. The ANN model appears to be the most suitable for solar power plants in Morocco, with the following optimal hyperparameters identified: learning_rate = 0.001, hidden_units = 64, hidden_layers = 3, dropout_rate = 0.1, activation = relu, epochs = 50, and batch_size = 32.

The results confirm that the ANN outperforms the SVR, RF, LSTM, and CNN models for GHI prediction. The study is based on data from the NSRDB database, and the five most relevant variables were selected using a Random Forest Regressor. These results have direct implications for energy producers and decision-makers. The use of ANNs makes it possible to optimize resource allocation, better plan energy demand, and reduce waste and dependence on non-renewable sources.

In terms of future prospects, one initial approach is to develop hybrid models combining deep learning and statistical methods. Another direction is the integration of additional data such as satellite observations and ground camera images to enhance the accuracy of forecasts.

In conclusion, this study highlights the importance of accurate SHI forecasts for improving energy management. The application of the ANN algorithm makes it possible to optimize the use of solar energy, reduce dependence on fossil fuels, and contribute to the transition to renewable sources.

Table 3. The prediction accuracy of studied models

Model	Model Accuracy		
	RMSE W/m²	MAE W/m²	R2
CNN	98.45	77.70	0.8405
LSTM	75.63	48.40	0.9059
ANN	72.19	41.19	0.9142
SVR	91.34	71.01	0.8627
RF	76.36	43.95	0.9040

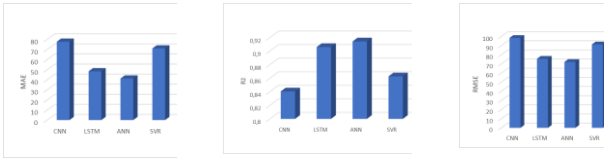

Fig.1. Ranking of the five algorithms

REFERENCES

[67] T. Agrawal (2021). *Hyperparameter optimization using scikit-learn*. In Hyperparameter Optimization in Machine Learning. https://doi.org/10.1007/978-3-030-81859-9_7

[68] ANRE (2023). *Rapport d'activité ANRE*. https://anre.ma/wp-content/uploads/2024/12/Rapport-dactivite-ANRE-2023_FR-1

[69] D. Chandola, H. Gupta, V. A. Tikkiwal, & M. K. Bohra (2020). Multi-step ahead forecasting of global solar radiation for arid zones using deep learning. *Procedia Computer Science*, 167, 626–635. https://doi.org/10.1016/j.procs.2020.03.326

[70] T.-P. Chu, J.-H. Jhou, & Y.-G. Leu (2020). Image-based solar irradiance forecasting using recurrent neural networks. In *Proceedings of the International Conference on System Science and Engineering (ICSSE)* (pp. 1–4), Kagawa, Japan.

[71] A. El Maghraoui, F.-E. Hammouch, Y. Ledmaoui, & A. Chebak (2022). Smart energy management system: A comparative study of energy consumption prediction algorithms for a hotel building. In *Proceedings of the 4th Global Power, Energy and Communication Conference (GPECOM)* (pp. 529–534). https://doi.org/10.1109/GPECOM55404.2022.9815807

[72] A. El Maghraoui, Y. Ledmaoui, O. Laayati, H. El Hadraoui, & A. Chebak (2022). Smart energy management: A comparative study of energy consumption forecasting algorithms for an experimental open-pit mine. *Energies*, 15(13), 4569. https://doi.org/10.3390/en15134569

[73] G.-F. Fan, L.-Z. Zhang, M. Yu, W.-C. Hong, & S.-Q. Dong (2022). Applications of random forest in multivariable response surface for short-term load forecasting. *International Journal of Electrical Power & Energy Systems*, 139, 108073. https://doi.org/10.1016/j.ijepes.2022.108073

[74] S. K. Jha, J. Bilalovic, A. Jha, N. Patel, & H. Zhang (2017). Renewable energy: Present research and future scope of artificial intelligence. *Renewable and Sustainable Energy Reviews*, 77, 297–317. https://doi.org/10.1016/j.rser.2017.04.018

[75] C.-H. Liu, J.-C. Gu, & M.-T. Yang (2021). A simplified LSTM neural networks for one day-ahead solar power forecasting. *IEEE Access*, 9, 17174–17195. https://doi.org/10.1109/ACCESS.2021.3052959

[76] REN21 (2021). *Renewables 2024 Global Status Report: Global overview*. https://www.ren21.net/gsr-2024/modules/global_overview/

[77] M. Sharifzadeh, A. Sikinioti-Lock, & N. Shah (2019). Machine-learning methods for integrated renewable power generation: A comparative study of artificial neural networks, support vector regression, and Gaussian process regression. *Renewable and Sustainable Energy Reviews*, 108, 513–538. https://doi.org/10.1016/j.rser.2019.03.040

[78] J. J. Song, Y. S. Jeong, & S. H. Lee (2014). Analysis of prediction model for solar power generation. *Journal of Digital Convergence*, 12(3), 243–248. https://doi.org/10.14400/JDC.2014.12.3.243

[79] R. Uhrig (1995). Introduction to artificial neural networks. In *Proceedings of IECON'95 – 21st Annual Conference on IEEE Industrial Electronics* (Vol. 1, pp. 33–37). https://doi.org/10.1109/IECON.1995.483329

[80] K. Wang, X. Qi, & H. Liu (2019). A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network. *Applied Energy*, 251, 113315. https://doi.org/10.1016/j.apenergy.2019.113315

[81] Q. Wang (2022). Support vector machine algorithm in machine learning. In *Proceedings of the IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)* (pp. 750–756). https://doi.org/10.1109/ICAICA54878.2022.9844516

[82] R. Yamashita, M. Nishio, R. K. G. Do, & K. Togashi (2018). Convolutional neural networks: An overview and application in radiology. *Insights Into Imaging*, 9, 611–629. https://doi.org/10.1007/s13244-018-0639-9

[83] X. Zhao, H. Wei, H. Wang, T. Zhu, & K. Zhang (2019). 3D-CNN-based feature extraction of ground-based cloud images for direct normal irradiance prediction. *Solar Energy*, 181, 510–518. https://doi.org/10.1016/j.solener.2019.02.057