Research Article

A New Hybrid Model for Hourly Solar Radiation Forecasting Using Daily Classification Technique and Machine Learning Algorithms

Hamza Ali-Ou-Salah ¹, Benyounes Oukarfi,¹ Khalid Bahani,² and Mohammed Moujabbir³

¹Laboratory of Physic of Condensed Matter & Renewable Energy, Faculty of Sciences and Techniques, Hassan II Casablanca University, Mohammedia 28806, Morocco

²Department of Computer Science, Faculty of Sciences and Techniques, Hassan II Casablanca University, Mohammedia 28806, Morocco

³Laboratory of Computer Science, Polydisciplinary Faculty of Khouribga, Sultan Moulay Slimane University, Khouribga 2500, Morocco

Correspondence should be addressed to Hamza Ali-Ou-Salah; hamza.alioussalah-etu@etu.univh2c.ma

Received 24 December 2020; Revised 9 February 2021; Accepted 2 March 2021; Published 16 March 2021

Academic Editor: Shigeyuki Hamori

Copyright © 2021 Hamza Ali-Ou-Salah et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Photovoltaic power generation depends significantly on solar radiation, which is variable and unpredictable in nature. As a result, the production of electricity from photovoltaic power cannot be guaranteed permanently during the operational phase. Forecasting global solar radiation can play a key role in overcoming this drawback of intermittency. This paper proposes a new hybrid method based on machine learning (ML) algorithms and daily classification technique to forecast 1 h ahead of global solar radiation in the city of Évora. Firstly, several comparative studies have been done between random forest (RF), gradient boosting (GB), support vector machines (SVM), and artificial neural network (ANN). These comparisons were made using annual, seasonal, and daily testing sets in order to determine the best ML algorithm under different meteorological conditions. Subsequently, the daily classification technique has been applied to classify the original training set into sunny and cloudy training subsets in order to enhance the forecasting accuracy. The evaluation of the proposed ML algorithms was carried out using the normalized root mean square error (nRMSE) and the normalized absolute mean error (nMAE). The results of the seasonal comparison show that the RF model performs well for winter and autumn seasons with nRMSE equaling 22.53% and 23.42%, respectively. While the SVR model gives good results for winter and summer seasons with nRMSE equaling 24.31% and 8.41%, respectively. In addition, the daily comparison demonstrates that the RF model performs well for cloudy days with nRMSE = 41.40%, while the SVR model yields good results for sunny days with nRMSE = 8.88%. The results show that the daily classification technique enhances the forecasting accuracy of ML models. Furthermore, this study demonstrates that the forecasting accuracy of ML algorithms depends significantly on sky conditions.

1. Introduction

Solar radiation is the most important environmental parameter in solar energy applications [1]. It plays a vital role in the management of solar systems including photovoltaic and solar thermal power technology [2]. However, solar radiation depends largely on meteorological conditions that are variable and unpredictable. Consequently, solar energy sources cannot guarantee the continuous production of electricity over time. Furthermore, energy storage

technologies are not sufficiently developed for electricity storage when necessary. As a result, this intermittency introduces a big challenge for the grid operators to integrate solar energy sources into the electric grid [3]. In fact, the grid operators should ensure permanently the stability of the grid in such a way that supply matches demand. For this purpose, the grid operator plans in advance by mobilizing different sources such as conventional and renewable energy sources to match a peak demand and avoid a blackout. Nevertheless, the fluctuation of solar radiation caused mainly by clouds

decreases electricity production and complicates the grid operator task to maintain a balance between production and consumption [4]. Forecasting of PV power generation can help significantly to overcome this obstacle by facilitating the management of the electric grid by planning, scheduling, and maintenance [4]. For this purpose, different categories of methods have been developed by researchers. The first category is the mathematical approaches that are divided into two types of methods: the persistence model and the statistical approaches. The persistence model is considered as the benchmark to evaluate the forecasting accuracy of other models, while the statistical techniques are composed of autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), autoregressive moving average with exogenous inputs (ARIMAX), and bagged and boosted regression trees [5]. The ARMA and ARIMA models require stationary times series as input, while the ARIMAX receive exogenous inputs such as meteorological parameters. Besides, the bagging and boosting techniques enhance the performance of classical regression trees. In fact, the forecasting accuracy of these mathematical methods decreases with time horizon, which makes them pertinent for short-term forecasting. The second category is machine learning algorithms (ML) that can overcome the shortcomings of mathematical models. The main ML algorithms are artificial neural network (ANN), support vector machine (SVM), and extreme learning machine (ELM). These algorithms can handle efficiently nonstationary time series [5]. In fact, the ANN is a powerful tool for dealing with nonlinear problems due to their high capacity in the training phase. However, the ANN has many disadvantages such as overfitting, local minima, random initialization and complexity of its structure. In addition, the SVR model deals well with nonlinear systems, and it does not have a local minima problem contrary to the ANN. Nevertheless, the SVR is highly sensitive to its hyper-parameters as kernel parameters, epsilon margin, and cost parameter. Furthermore, the ELM model selects randomly the input weights and hidden nodes, which represents a challenge in the training process. The third category is hybrid models that consist of the combination of two or many methods to improve forecasting accuracy. In fact, the hybrid models take advantages of each individual method, which enhances the performance of the overall model. Many hybrid approaches have been proposed in the literature combining machine learning algorithms, mathematical methods, physical models, and optimization algorithms [5]. It was shown that the hybrid models outperform the individual methods in terms of forecasting accuracy. However, the hybrid models have a high computational complexity, which requires important resources for their running as time and space. Also, their performances depend significantly on historical inputs that should be selected carefully.

This paper aims to forecast global solar radiation in order to subsequently predict PV power generation. In the literature, several approaches have been proposed for short-term forecasting of solar radiation. We group them into three categories: The first category is the numerical weather prediction (NWP)-based models, which approximate the

solutions of physical equations that describe the motion and the direction of clouds. They give acceptable results for forecasting 6 h to several days ahead of solar radiation [6–8]. The second category is satellite- and sky-images-based models that detect the cloud motion structures by using two consecutive cloud index images [9, 10]. In fact, the sky-image-based models perform well for intra-hour forecasting, while the satellite-image-based models are commonly used for intraday forecasting [7, 8]. The third category is statistical and ML models that use a historical measurement of solar and meteorological data to forecast different time horizons (5 min to one day ahead). In fact, ML models are a data analytic technique that uses computational methods to learn from data in order to find patterns without requiring any explicit algorithms or equations.

Recent studies in the field of solar radiation forecasting became interested in ML models thanks to their high forecasting accuracy and their ease of use [7, 8]. Khosravi et al. [11] developed several ML algorithms such as multi-layer feed-forward neural network (MLFFNN), radial basis function neural network (RBFNN), support vector regression (SVR), fuzzy inference system (FIS), and adaptive neuro-fuzzy inference system (ANFIS) to predict hourly solar irradiance. It was found that the SVR model can forecast accurately the hourly solar irradiance using historical endogenous values. Lotfi et al. [12] used multilayer perceptron neural networks (MLP) and neural autoregressive (NARX) to predict the components of hourly solar radiation. The NARX model gives the best performance using available and cheap meteorological data. Anwar and Khatib [13] proposed a novel hybrid model based on random forests (RFs) and firefly algorithm (FFA) for predicting hourly global solar radiation. The results obtained indicated the superiority of the proposed hybrid model as compared with conventional RFs and ANN models. Belaid et al. [14] developed a new approach based on the SVM model and time series principle for forecasting 1h ahead of global solar radiation. The results showed a high accuracy of the proposed method using previous solar radiation values. Qing and Niu [15] proposed a novel prediction method using long short-term memory networks (LSTM) and weather forecasts. Their experimental results indicated the superiority of the LSTM algorithm compared to the MLFFN model. Benmouiza and Cheknane [16] developed several hybrid models based on the ANFIS model with different clustering algorithms as fuzzy c-means (FCM), subtractive clustering, and grid partitioning for forecasting an hour-ahead solar radiation in Algeria. The performance comparison showed that the ANFIS model with the FCM clustering algorithm provides the best forecasting results. Furthermore, Benali et al. [17] found that the RF algorithm gives accurate hourly forecasts for the different components of solar radiation compared to the ANN and smart persistence (SP) models. Urraca et al. [18] used the SVR model for forecasting 1h ahead of global solar irradiation in Southeast Spain. The results showed the high performance of the SVR model compared to the RF and classical linear models. In addition, Voyant et al. [7] showed that the SVM, RF, and gradient boosting (GB) models give very good prediction results, but

still more studies should be carried out using these techniques. All these recent works showed that RF, GB, SVM, and ANN models are very promising methods. However, the performance comparison between these abovementioned studies seems difficult due to several reasons, namely, the nature of the dataset used for training, the location of the study, the forecasting horizon, and the statistical indicators used for evaluation [7]. Furthermore, a model that performs well for a location does not give good results for other locations. Consequently, there is no universal model that performs well for all locations, but just a set of promising models.

In this paper, we propose a new hybrid method based on ML algorithms and the daily classification technique for forecasting 1h-ahead global solar radiation in the city of Évora located south of Portugal. Firstly, several comparative studies have been carried out between ML algorithms such as RF, GB, SVM, and ANN models. These ML algorithms represent some of the most promising forecasting models [7]. The recent literature showed their effectiveness in different fields such as civil engineering, renewable energy, and environmental engineering. Among their advantages are their high capabilities to deal with nonlinear multidimensional functions, their capacity to identify patterns and trends, continuous improvements using new training data, and their wide range of application. However, the ML algorithms have many disadvantages such as the requirement of massive data for training, their high computational complexity in particular for the RF and GB models, their dependence on the quality of the dataset, their sensitivity to the hyper-parameter values, specifically for the SVR model, and the random initialization and local minima plaguing, especially the ANN model.

Furthermore, the comparison studies have been performed using annual, seasonal, and daily testing sets. This allows us to understand how these ML algorithms behave toward the different climate conditions of Évora city. In fact, the results of daily comparison have shown that the performances of the proposed ML models depend significantly on daily sky conditions. For this reason, the original training set has been classified according to sky conditions using the daily classification technique in order to enhance the forecasting accuracy. In fact, the daily classification technique enables to form more homogenous subsets that were used for the training of the best selected ML models. This work takes into account the sky conditions for evaluation and training of ML models, which is not considered by the majority of existing studies. Moreover, the proposed method was applied on a dataset of Évora city, which is characterized by specific climate conditions that lead to identifying the best ML models for such climate and sky conditions.

The rest of this paper is organized as follows: Section 2 outlines the proposed methodology. Section 3 discusses the experimental results, and Section 4 concludes the paper and indicates some perspectives for future work.

2. Materials and Methods

2.1. Data. Five years of hourly global horizontal solar radiation data (from 2012 to 2016) have been used in this work.

These data have been collected using an Eppley pyranometer from Évora city's meteorological station located south of Portugal (38°34'N, 07°54'W). In fact, only data between sunrise and sunset are considered, which constitute the sunshine duration for Évora city. Our objective is to forecast 1h ahead of global solar radiation based on historical endogenous data. For this reason, the auto mutual information function has been used to define the most significant lagged values of observed solar radiation data [19].

In fact, the mutual information function is a function of entropy that estimates the linear and nonlinear relationship between two variables, X and Y . It is based on Shannon's information theory in order to measure the amount of information obtained on one random variable Y by observing the other random variable X . The auto mutual information function is estimated between two measurements of the same time series $x(t)$. It calculates the extent to which series $x(t)$ can be forecasted from delayed series $x(t - \tau)$. The following is the mathematical expression of the mutual information between discrete random variables $X_{t-\tau}$ and X_t :

$$I(X_{t-\tau}, X_t) = \sum_{x_{t-\tau} \in X} \sum_{x_t \in X} P_{xx}(x_{t-\tau}, x_t) \cdot \log \left(\frac{P_{xx}(x_{t-\tau}, x_t)}{P_x(x_{t-\tau})P_x(x_t)} \right), \quad (1)$$

where P_{xx} is the joint probability density function of $X_{t-\tau}$, and X_t , $P_x(x_{t-\tau})$ and $P_x(x_t)$ are, respectively, the marginal probability mass functions of $X_{t-\tau}$ and X_t .

In our case, several auto mutual information values have been calculated between $x(t)$ and its lagged values $x(t - \tau)$ with $1 \leq \tau \leq 75$. The results of applying the auto mutual information function can be expressed as follows:

$$G_d(t) = f(G_d(t), G_d(t-1), G_{d-1}(t), G_{d-2}(t), G_{d-3}(t), G_{d-4}(t), G_{d-5}(t)), \quad (2)$$

where $G_d(t)$ is the 1h-ahead solar radiation of day d , $G_d(t-i)$ is i -th lagged value of day d , and $G_{d-j}(t)$ is lagged value of j -th previous day of the same time of forecasted values. As results, the most significant lagged values are two lagged values of the same day of the forecasted solar radiation and five lagged values of the five previous days of the same time of predicted value.

2.2. Machine Learning Algorithms. Machine learning (ML) is an artificial intelligence method that learns patterns from data without requiring explicit programs or algorithms [7]. Two main techniques are used in ML models: supervised and unsupervised learning. In supervised learning, the algorithm uses a set of known inputs and their desired outputs to learn the rule that maps inputs to outputs [20]. Subsequently, the model will be used to predict new data. The unsupervised learning algorithms are used to find hidden structures in data inputs without any knowledge of desired outputs. In this work, we have used four supervised ML algorithms such as ANN, SVM, GB, and RF models.

2.2.1. Artificial Neural Network. Artificial neural network (ANN) is the most used ML algorithm in the field of solar radiation forecasting [7]. Among the used ANN variants, the multilayer perceptron neural network (MLFFNN) is commonly employed by researchers for solar radiation forecasting [11, 21, 22]. In this work, the MLFFNN has been developed to forecast accurately 1h ahead of global solar radiation. The MLFFNN approximates nonlinear functions efficiently using three interconnected layers. The first layer is the input layer, the second layer is composed of one or several hidden layers, and the third layer is the output layer [11]. Each layer is composed of processing units called neurons. Also, each neuron of hidden and output layers contains an activation function that defines its output. The following equation describes the output of the artificial neuron:

$$y_j = f\left(\sum_{i=1}^n w_{ji}x_i\right), \quad (3)$$

where y_j is the neuron output of the j -th hidden layer, w_{ji} is the synaptic weight between the j -th layer and the i -th layer, x_i is the input of the j -th layer, and $f(x)$ is the activation function. The expressions of activation functions are defined as follows:

$$\begin{aligned} \text{tansig}(n) &= \frac{2}{(1 + e^{-2n})^{-1}}, \\ \text{purelin}(n) &= n. \end{aligned} \quad (4)$$

The information is transferred from layer to layer using synaptic connections called weights and biases [21]. In fact, the learning of the network is carried out by adjusting weights and biases using a backpropagation (BP) training algorithm. Among the activation functions, we find linear, sigmoid, hyperbolic, and rectified linear unit activation functions. In fact, the linear function allows a wide range of outputs, unlike the binary step function that gives two outputs, one or zero. However, its derivative function, which is constant, prevents the model from using the backpropagation in the training process, which is considered as the main training algorithm for neural networks. The linear function reduces the complexity of the model because its last layer is just a linear function of its first layer, which makes the model with only one layer whatever the number of layers. This decreases the ability of the model to deal with nonlinear data. In addition, the sigmoid and hyperbolic activation functions overcome the disadvantages of linear function thanks to their derivative functions and the use of several layers. Furthermore, they have almost the same advantages such as the smooth gradient, which avoids the jumping of output values that are normalized between two bounds. Also, these activation functions allow clear predictions by bringing the outputs to the limit of the curves. In fact, the hyperbolic function is distinguished from the sigmoid function by its zero-centered propriety, which allows it to handle strongly positive or negative values. Among their disadvantages is the disappearance of the gradient for very

high or very low input values, which causes slow convergence or learning saturation. Besides, the two activation functions are expensive in terms of computational complexity and require more resources and time for complex modelling. Contrary to sigmoid and hyperbolic activation functions, the rectified linear unit has a low computational complexity, which allows a fast convergence of the model. It is a nonlinear function capable of handling complex data using the backpropagation algorithm. Nevertheless, its gradient becomes null when the inputs tend to zero or to negative values which stops the learning process. In this work, we have used the Levenberg Marquardt (LM) as a training algorithm, and the hyperbolic tangent sigmoid and the linear activation functions, respectively, in the hidden and output layers. This configuration of the MLPNN model gives the best forecast results according to the literature [11, 21, 22]. Finally, the performance of the MLFFNN model depends significantly on the number of hidden layers (*Hidden_layer*) and hidden neurons (*Hidden_Neurons*). Karsolliya [23] shows that one hidden layer, with a sufficient number of hidden neurons, is enough to map any nonlinear complex function. This parameter should be optimized by exploring different configurations of the MLFFNN model [11].

2.2.2. Support Vector Machine. Support vector machine (SVM) was initially proposed by Vapnik and Cortes in 1995 [24]. It is a supervised learning algorithm used for classification and regression analysis. In fact, the SVM is fast, reliable, and capable of dealing with complex nonlinear problems; it was successfully applied in the field of solar energy forecasting [3, 25]. The principle of the SVM is to map vector inputs using a nonlinear transformation into a higher dimensional feature space, in which it carries out a linear separation. The SVR minimizes the following equation:

$$\begin{aligned} &\min \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \right) \\ &\text{subject to} \begin{cases} y_i - (\langle w, x_i \rangle + b) \leq \varepsilon + \xi_i \\ (\langle w, x_i \rangle + b) - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0, \text{ clear,} \end{cases} \end{aligned} \quad (5)$$

where x_i and y_i are input and target sets. w is the weight vector, and b is the bias. ξ_i, ξ_i^* are the slack variables, ε is the epsilon margin, and C is the cost parameter [3].

Furthermore, the nonlinear transformation is performed using kernel functions. There exist different types of kernel functions such as the linear, polynomial, Gaussian, and sigmoid functions. In fact, the linear function is the simplest kernel that offers a high degree of interpretability, which allows for determining the feature importance. However, the linear kernel function performs well just for linear problems. Unlike the linear function, the polynomial kernel function can solve nonlinear functions and provide accurate

predictions by defining the pertinent polynomial degree. But, the polynomial function requires more parameters and is unsuitable when the selected polynomial degree is too high. Besides, the Gaussian kernel function can map data into infinite dimensions with various decision boundaries. It has a simple tuning with only one parameter to select. On the other hand, the Gaussian kernel function is slow in terms of calculation speed and has poor interpretability. Its performance is very sensitive to the value of its selected parameter. Also, the sigmoid kernel function finds its origin from the neural network field. When the SVM model uses it as kernel function, the developed model becomes a multilayer perceptron neural network, which is very powerful for non-linear problems. This multilayer perceptron takes advantages of the SVM model in such a way that it does not have local minima and has good generalization capability.

In fact, the radial basis function (RBF) has been widely used by researchers in the field of solar radiation forecasting [3, 11, 14, 18]. For this reason, the RBF has been utilized as a kernel function whose expression is formulated as follows:

$$\varphi(x_i, x_j) = e^{-\|x_i - x_j\|^2 / 2\sigma^2}, \quad (6)$$

where σ is a parameter that defines the width of the Gaussian function.

The performance of the SVM depends on hyper-parameters as C , ε , and σ . The optimal combination of these hyper-parameters should be determined using different predefined values [11].

2.2.3. Gradient Boosting. Gradient boosting (GB) algorithm is an ensemble learning technique that combines a sequence of weak learners to form a strong model. In fact, the GB algorithm tries to improve the prediction accuracy of weak learners by adding iteratively new learners in such a manner that the added learner enhances the accuracy of its predecessor [26]. In addition, the GB algorithm is viewed as a gradient descent algorithm for regression problems in some function space. It evaluates the shortcoming of weak learners by using the gradients in the loss function [27]. In fact, the most common learner is the regression tree, which is weak as an individual learner. However, the combination of several weak learners in an ensemble significantly improves the accuracy [28]. The main objective is to find the approximation function $\hat{f}(X)$ that minimizes the loss function $L(f(X))$ [26].

$$\hat{f}(X) = \arg \min_f L(f(X)), \quad (7)$$

where X is the input vector.

We consider that $f_0(X)$ and $f_{t-1}(X)$ are the initial and current approximations, respectively, of $\hat{f}(X)$; the algorithm updates the approximation function $f_t(X)$ using the gradient $g_t(X)$ and the step size ρ_t ($\rho_t > 0$) as follows [26]:

$$f_t(X) = f_{t-1}(X) - \rho_t \cdot g_t(X). \quad (8)$$

The process of updating continues until the number of iterations is reached. The negative gradient $-g_t(X)$ is

approximated using the parametric function $h(X; \theta_t)$, which is described for the regression terminal node J as

$$h(X; \theta_t) = \sum_{j=1}^J \bar{Y}_j I_{[X \in R_j]}. \quad (9)$$

where θ_t is a parameter vector, $\bar{Y}_j$ and R_j are, respectively, the mean and separated input space of the J -th terminal node of the regression tree [27].

The accuracy of the GB model depends on hyper-parameter values such as the number of trees m , the learning rate η , and the maximum number of splits (*MaxNumSplits*). These hyper-parameters are tuned by exploring predefined values to find the optimal architecture of the GB model [29].

2.2.4. Random Forest. Tree-based models are increasingly used in the field of solar radiation forecasting due to their high performance [8]. Among these models, we find the random forest (RF) algorithm that performs well in terms of forecasting accuracy [13, 17]. In fact, the RF algorithm is an ensemble learning technique that combines several decision trees to improve the performance of individual tree as well as to overcome its shortcomings [17, 30]. The RF predicts the output value by averaging the predicted values of all regression trees, which leads to more accurate predictions [31]. The RF algorithm uses two random concepts to build the model. The first one is the bagging technique (or bootstrap aggregating), which consists of dividing randomly the original dataset with replacement in such a way that each decision tree is trained with different subsets [13]. The second concept is the random feature selection, which is used to grow each regression tree. In fact, at each node split, random sampling of features is performed and only some of them are selected for splitting [17].

The regression trees that form the RF model are trained with different subsets given by bootstrap sampling. One-third of these bootstrap samples are left out to validate the RF model, which are called out-of-bag (OOB) data, while the remaining samples are called in-of-bag data [13]. Furthermore, at each node, iterative binary splits using randomly selected input variables are carried out to grow each decision tree until reaching the best architecture. Finally, the predicted value of the RF model is the average value of all outputs of the trees [32].

The performance of the RF model depends on several hyper-parameter values such as the number of trees in the forest (m), the number of selected variables at each node split (*NumVarSplits*), the minimum samples per leaf (*MinLeafSize*), and the maximum number of splits (*NumMaxSplits*). These hyper-parameters should be optimized to avoid overfitting or underfitting. For this reason, the grid search technique has been used to find the optimal combination of hyper-parameters of the RF model [29]. It involves searching through manually predefined values of hyper-parameters by building several models for each possible combination. The model that gives accurate predictions is selected as the best model with its correspondent combination of hyper-parameters.

2.3. Architectures of ML Models. Firstly, the original dataset has been divided into 4 years of the training set and 1 year of the testing set. Then, the grid search hyper-parameter tuning has been used to determine the best architecture of each ML model. In fact, the architecture of each model is a combination of hyper-parameters that should be tuned.

For this purpose, the grid search technique explores several predefined combinations of hyper-parameters to find the ML model that gives the best prediction results [29]. Each combination of these hyper-parameters represents a point of the grid and corresponds to the architecture of the candidate model.

In fact, the 5-fold cross-validation technique has been used to evaluate the performance of ML candidate models during the grid search process [26]. This technique involves dividing randomly the dataset into 5 folds of data. Each fold is used as a testing set and the rest of the folds are used as a training set. The validation error is the mean of testing sets errors. Figure 1 shows a flowchart of the grid search technique.

In fact, a grid search program has been developed using the functions of Statistics and Machine learning Toolbox™ and Deep learning Toolbox™ available in MATLAB® Software (Ver. 9.2, R2017a). These functions propose a set of hyper-parameters that should be optimized for each ML model. The grid search technique has been carried out using the following ranges of hyper-parameter values:

Concerning the ANN model, we have used the “fitnet” function that receives as hyper-parameters the number of *Hidden_Neurons*. The range of this hyper-parameter is

$$\text{Hidden_Neurons} = [10, 20, 30, 40, 45, 50, 55, 60, 65, 70, 75, 80, 90, 100]$$

$$\text{Number of possible combinations} = 19$$

For the SVR model, the “fitrsvm” function was utilized with three hyper-parameters: *C*, ϵ , and σ . The used ranges are

$$C = [100, 200, 300, 400, 500, 600, 700]$$

$$\epsilon = [10, 20, 30, 40, 50, 60, 70]$$

$$\sigma = [0.5, 1.23, 1.5, 2, 3]$$

$$\text{Number of possible combinations} = 245$$

To build The GB model, the “fitrensemble” function was used by tuning the following hyper-parameters: *m*, η , and *MaxNumSplit*. The ranges of these hyper-parameters are

$$m = [25, 50, 75, 100, 150, 200]$$

$$\eta = [0.1, 0.25, 0.5, 0.75, 1]$$

$$\text{MaxNumSplits} = [1, 5, 10, 20, 30, 40, 50, 60, 80, 100, 200, 300, 400, 500]$$

$$\text{Number of possible combinations} = 450$$

The RF model was designed using the “fitrensemble” function that receives as hyper-parameters: *m*, *NumVarSplits*, *MinLeafSize*, and *NumMaxSplits*. The ranges of these hyper-parameters are

$$m = [100, 150, 200, 250, 750, 1000]$$

$$\text{NumVarSplits} = [4, 5, 6]$$

FIGURE 1: The grid search technique.

$$\text{MinLeafSize} = [5, 7, 9, 11, 13]$$

$$\text{NumMaxSplits} = [500, 1000, 2000, 3000, 4000, 5000, 6000]$$

$$\text{Number of possible combinations} = 525$$

The best ML model corresponds to the combination that gives a low 5-fold cross validation error. Then, each ML model was trained with the selected combination of hyper-parameters. In addition, several comparative studies have been performed using annual, seasonal, and daily testing sets in order to determine the best ML models under different climate and sky conditions. To do this, 1 year of the testing set has been divided into seasonal and daily testing subsets.

The accuracy of the proposed ML models was evaluated using the root mean square error (RMSE) and the mean absolute error (MAE). The RMSE is very sensitive to high errors, while the MAE is used to measure the direct magnitude of the model deviation [8, 29]. Lower values of the RMSE and MAE mean that the forecasts are accurate.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}, \quad (10)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|,$$

where $\hat{y}$ is the forecasted value, y is the measured data, and n is the number of observations. In addition, the errors are normalized using the mean value of the solar radiation time series.

$$n\text{RMSE}(\%) = \frac{\sqrt{(1/n) \sum_{i=1}^n (\hat{y}_i - y_i)^2}}{(1/n) \sum_{i=1}^n y_i} \times 100, \quad (11)$$

$$n\text{MAE}(\%) = \frac{(1/n) \sum_{i=1}^n |\hat{y}_i - y_i|}{(1/n) \sum_{i=1}^n y_i} \times 100.$$

The Akaike Information Criterion (AIC) has been used to evaluate the developed ML models [33].

$$AIC = n \times \log \left(\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \right) + 2k. \quad (12)$$

where k is the number of optimized parameters by the model.

2.4. Daily Classification Technique. The daily classification technique was used to classify sky conditions using daily clearness index (k_t), which is widely used for solar radiation characterization [34]. In fact, the daily clearness index is the attenuation factor of the atmosphere, which is defined as the ratio of the received global horizontal solar radiation (G) to the corresponding extraterrestrial global horizontal solar radiation (G_0).

$$k_t = \frac{G}{G_0}. \quad (13)$$

The sky conditions depend significantly on daily clearness index values. For this reason, different kt intervals available in the literature [34] have been used to classify the sky conditions into two groups of days: sunny or very sunny days ($kt > 0.6$) and cloudy or partly cloudy days ($kt \leq 0.6$). The daily clearness index data have been obtained from power project datasets supported by NASA [35]. Figure 2 shows Five years of daily clearness index data of Évora city.

Firstly, the daily classification technique was used to divide the testing set into daily testing subsets in order to evaluate the ML models under different sky conditions. Then, the best ML models were trained with specific data, with which each model performs well. For this purpose, the original training set was divided into sunny and cloudy training subsets using the daily classification technique to form more homogenous training subsets in order to improve the forecasting accuracy. Figure 3 illustrates the overall research methodology.

3. Results and Discussion

In this work, the auto mutual information function has been applied to solar radiation time series to identify the most significant historical values for forecasting 1h ahead. In fact, the results show that seven lagged values can accurately forecast future solar radiation. Furthermore, the grid search hyper-parameter tuning has been applied to determine the architectures of the ML models. Indeed, several combinations of hyper-parameters have been tested to find the optimal architecture of each model using the 5-fold cross-validation technique for evaluation. Table 1 summarizes the best hyper-parameter values of the ML models.

3.1. Annual Comparison. A comparative study between the ML models has been carried out using one year of the testing set. The following Table 2 illustrates the results of the annual comparison.

As shown in Table 2, the RF model is slightly better than the other ML models in terms of RMSE. Besides, The ANN and SVR models have almost the same errors, while the GB

FIGURE 2: Daily clearness index data of Évora city.

FIGURE 3: The overall research methodology.

lacks in performance compared to other ML models. However, The GB is the best model in terms of accuracy and complexity according to the AIC. In terms of nMAE, the SVR and RF models are better than the ANN and GB models.

In fact, the comparison between these ML models seems somewhat difficult, because some models such as the RF model is a slightly better than the other ML models according to RMSE, while the SVR model is slightly better in terms of MAE. For this reason, an in-depth comparative study between the ML models has been performed using seasonal testing subsets in order to investigate the performances of each ML model according to seasons.

TABLE 1: The best hyperparameters of the ML models.

Models	Configurations	Validation RMSE (W/m ²)	AIC
ANN	<i>Hidden_layer = 1; Hidden_Neurons = 60</i>	63.82	5131
SVM	<i>C = 600; $\epsilon = 10$; $\sigma = 2$</i>	62.35	26245
GB	<i>m = 50; MaxNumSplits = 40; $\eta = 0.1$</i>	67.13	4577
RF	<i>m = 750; NumVarSplits = 6; MinLeafSize = 11; MaxNumSplits = 5000;</i>	66.82	20472

TABLE 2: The annual comparison between the ML models.

Error metrics	ANN	SVR	GB	RF
RMSE (W/m ²)	62.73	62.80	63.40	62.44
nRMSE (%)	18.70	18.73	18.91	18.62
MAE (W/m ²)	33.94	32.52	34.57	32.75
nMAE (%)	10.12	9.69	10.31	9.76
AIC	6136	27273	5546	21428

3.2. Seasonal Comparison. One year of the testing set is split into four seasonal testing subsets; every one of them represents a season of the year. Table 3 highlights the results of the performance comparison between the ML models for each season.

In winter, the SVR model gives better results than the GB, ANN, and RF models. Moreover, the RF model provides better results than the other ML models during spring. In summer, the ANN and SVR models give better results than the GB and RF models according to RMSE, while the SVR model is better than the other ML models in terms of MAE. Also, the RF model is slightly better than the other ML models during autumn. Besides, the GB is the better than the other ML models during all seasons according to AIC. In summary, the RF model performs well for spring and autumn, which are very difficult to forecast due to the strong variation of global solar radiation. On the other hand, the SVR model achieves good results for winter and summer, which are characterized by low variability of global solar radiation.

In fact, each season represents a mix of daily sky conditions (sunny or cloudy skies), which does not allow us to determine which model is the best performer for a particular sky condition. Consequently, a comparison study based on daily testing subsets was performed to study the performance of each model according to sky conditions.

3.3. Daily Comparison. A daily comparison between the ML models has been carried out using the daily testing subsets obtained by the classification of sky conditions. In fact, the daily classification technique has been used to classify one year of the testing set into daily testing subsets using the daily clearness index (*kt*) of Évora city. Table 4 shows the results of the comparison between the ML models using daily testing subsets, including 219 sunny days and 147 cloudy days.

For sunny days, the SVR model significantly outperforms the other ML models according to nRMSE and nMAE. For cloudy days, the ANN and RF models yield better results than the GB and SVR models. In fact, the ANN and RF models have marginal variation in terms of nRMSE.

However, the RF model is better than the ANN model in terms of nMAE. Therefore, the RF model is the best performer for cloudy days. For cloudy and sunny days, the GB is the best model in terms of the AIC. In conclusion, the results show that there is not just one best model, but the two models are complementary in such a way that the SVR model performs well for sunny days while the RF model gives good results for cloudy days. Contrary to the annual and seasonal comparisons, the daily comparison better highlights the performances of each model, which allows us to make definitive conclusions about the performances of the ML models.

3.4. The Proposed Hybrid Model. The high accuracy of the SVR and RF models for forecasting sunny and cloudy days, respectively, led us to train the two ML models with specific data with which each model performs well in order to improve the forecasting accuracy. For this reason, the original training set has been classified into sunny and cloudy training subsets using the daily classification technique. The first training subset represents sunny days which is used to train the SVR model, while the RF model is trained with a cloudy days training subset. Table 5 shows the accuracy of the SVR and RF models using the new training subsets obtained by applying the daily classification technique on the original training set.

As shown in Table 5, the training of the SVR and RF models with new homogenous training subsets enhances the forecasting accuracy. In fact, the accuracy of the SVR model is improved according to nRMSE and nMAE values, while the RF model is superior in terms of nRMSE with marginal variation according to nMAE. The AIC shows that the SVR and RF models have been enhanced in terms of accuracy and complexity. Therefore, a new hybrid method is developed using the ML models, such as the SVR and RF models, and daily classification of the original training set to forecast 1h-ahead of global solar radiation. Figure 4 shows the structure of the proposed hybrid model.

Figure 5 shows the regression plot of the sunny testing subset using the SVR model. We notice that the majority of

TABLE 3: The seasonal comparison between the ML models.

Error metrics	Winter				Spring				Summer				Autumn			
	ANN	SVR	GB	RF	ANN	SVR	GB	RF	ANN	SVR	GB	RF	ANN	SVR	GB	RF
RMSE (W/m^2)	50.06	48.76	50.29	48.94	93.89	94.43	94.27	93.17	41.97	42.01	44.71	43.30	51.42	51.97	51.44	51.22
nRMSE (%)	24.96	24.31	25.07	24.40	22.71	22.84	22.80	22.53	8.40	8.41	8.95	8.67	23.51	23.76	23.52	23.42
MAE (W/m^2)	29.12	27.84	29.45	28.13	59.04	57.09	59.14	56.76	20.24	17.88	22.37	19.34	27.42	27.36	27.35	26.83
nMAE (%)	14.52	13.88	14.68	14.02	14.28	13.81	14.30	13.73	4.05	3.58	4.48	3.87	12.54	12.51	12.51	12.26
AIC	2336	23464	1735	17627	2563	23700	1962	17858	2346	23482	1763	17653	2369	23547	1767	17666

TABLE 4: The daily comparison between the ML models.

Error metrics	Sunny days ($kt > 0.6$)				Cloudy days ($kt \leq 0.6$)			
	ANN	SVR	GB	RF	ANN	SVR	GB	RF
RMSE (W/m^2)	40.77	38.80	40	39	85.56	87.05	86.91	85.77
nRMSE (%)	9.61	9.15	9.57	9.36	42.09	42.82	42.75	42.19
MAE (W/m^2)	22.47	19.72	23.18	21.03	51.02	51.58	51.55	50.19
nMAE (%)	5.30	4.65	5.46	4.96	25.1	25.37	25.36	24.69
AIC	3846	24947	3231	19113	3332	20790	2737	18631

TABLE 5: Performance of the ML models using two types of training sets.

Error metrics	Original training set		New training subsets	
	SVR	RF	SVR	RF
RMSE (W/m^2)	38.80	85.77	37.64	84.17
nRMSE (%)	9.15	42.19	8.88	41.40
MAE (W/m^2)	19.72	50.19	19.17	50.25
nMAE (%)	4.65	24.69	4.52	24.72
AIC	24974	18631	13664	18622

FIGURE 4: The proposed hybrid model.

data points are falling on the regression line, which means that the SVR model gives very accurate forecasts for sunny days. Figure 6 shows 1h-ahead forecasting of global solar radiation using the SVR model for 240 h of sunny days. As can be seen, the graphs of forecasted and measured values are almost identical which indicates the high accuracy of forecasts.

In addition, Figure 7 shows the regression plot of the cloudy testing subset using the RF model. The regression plot of cloudy days indicates that the predictions are less accurate than the forecasts of sunny days due to the high variability of global solar radiation during cloudy days. However, the RF model fits well with the cloudy data considering their complexity and their high nonlinearity. Figure 8 shows 1h-ahead forecasting of global solar radiation using the RF model for 240 h of cloudy days.

In the literature, several studies have been developed using hybrid models. VanDeventer et al. [36] proposed a new hybrid model using genetic algorithm (GA) and SVM model for solar PV power forecasting. Firstly, the SVM classifies historical meteorological parameters in order to form an ensemble of classifiers for each range of PV power. Besides, the architecture of the SVM was optimized by the GA, which was used again for the optimization of the weight matrix. The results show the superiority of the hybrid model compared to the conventional SVM. In fact, our findings are consistent with this study in terms of the use of the SVM model and ensemble technique to enhance forecasting accuracy. However, in our study, we used grid search technique for the optimization of the architectures of ML algorithms, while VanDeventer et al. [36] used the GA optimization technique to enhance the performance of the SVM model. In the literature, other optimization

FIGURE 5: The regression plot of the SVR model using the sunny testing subset.

FIGURE 7: The regression plot of the RF model using the cloudy testing subset.

FIGURE 6: 1 h-ahead forecasting of global solar radiation using the SVR model for 240 h of sunny days.

techniques have been proposed such as particle swarm optimization (PSO), ant colony optimization (ACO), fruit fly optimization algorithm (FOA), firefly algorithm (FF), and chaotic artificial bee colony algorithm. Furthermore, several researchers developed hybrid models based on daily weather classification [37]. Yang et al. [38] used self-organizing map (SOM) and learning vector quantization (LVQ) for the classification of PV power into six daily weather types. Then, many submodels have been used for the forecasting stage based on the SVR algorithm. Similarly, Shi et al. [39] used weather classification technique and SVM model to forecast PV power output. The weather conditions are classified into four types of days (clear,

FIGURE 8: 1 h-ahead forecasting of global solar radiation using the SVR model for 240 h of cloudy days.

cloudy, foggy, and rainy days). In fact, these studies are in good agreement with our findings in the use of the classification of weather conditions to improve forecasting accuracy. Nevertheless, they used only one ML algorithm for all types of days which is different from our methodology that involves the use of a specific ML model for each daily weather condition.

The significance of our study lies in the assessment of the ML models using the daily testing subsets, which allows us to determine the best ML model according to each sky

condition. In fact, our findings indicate that the daily comparison represents the best way to evaluate the forecasting accuracy of the ML models. Furthermore, the daily classification technique was used to divide the original training set into daily training subsets. This enables to train the best selected ML models with new training subsets with which each model gives good prediction results. In fact, the daily classification technique enhances the forecasting accuracy of the ML models by means of forming small and homogenous training subsets according to the sky conditions of Évora city.

Among the limitations of this work, the grid search technique is computationally expensive. In fact, the grid search explores several combinations of hyper-parameters, which means a large number of experiences. The use of a 5-fold cross-validation technique increases the number of experiences. Furthermore, the number of training data is high and the computation power is limited. The combination of all these factors such as the use of 5-fold cross-validation technique, large number of training samples, and limited computational power renders the grid search technique a time-consuming process and does not allow exploring a wide range of hyper-parameter values. The use of optimization techniques and parallel computing can overcome this drawback. In addition, the lack of relevant inputs such as hourly clearness index and a longer length of solar data prevents us from achieving more accurate predictions results. Also, the generalization of this method to other sites in the world is not possible due to the absence of reliable hourly solar radiation time series of other locations.

4. Conclusions

In this work, a new hybrid model is proposed using the daily classification technique and ML models for forecasting 1h ahead of global solar radiation in the city of Évora. Firstly, several comparative studies have been carried out between ML models such as RF, GB, SVM, and ANN using annual, seasonal, and daily testing sets. In fact, the results of the seasonal comparison show that the spring and autumn seasons are very difficult to forecast due to the strong variability of global solar radiation. Furthermore, the results of comparative studies show that the daily comparison is the best way to evaluate the performances of ML models. Moreover, the daily comparison indicates that there is not just one best model, but the ML models are complementary in such a way that the SVR model performs well for sunny days, while the RF model gives good results for cloudy days. In addition, the daily classification technique has been applied to divide the original training set into small and homogenous training subsets. Then, the best selected ML models such as the SVR and RF models were trained with new subsets with which each model performs well. The results demonstrate that the daily classification technique improves the forecasting accuracy of the ML models. Finally, the proposed hybrid model was compared to other methods in the recent literature. The result of the comparison demonstrates the significant improvement in forecasting

accuracy through the use of the proposed hybrid model. Further work will focus on more efficient methods that can improve the forecasting accuracy during cloudy sky conditions that are characterized by a strong fluctuation of solar radiation time series.

Data Availability

The solar radiation data used to support the findings of this study may be released upon application to the Institute of Earth Sciences that can be contacted at <http://www.ict.uevora.pt/g1/index.php/data-request-form/>. Daily clearness index data used in this article were obtained from the NASA.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors thank Professor Mouhaydine Tlemcani from the Institute of Earth Sciences for providing the solar radiation data of Évora city. Daily clearness index data used in this article were obtained from the NASA Langley Research Center (LaRC) POWER Project funded through the NASA Earth Science/Applied Science Program. This work was supported by the National Centre for Scientific and Technical Research under Grant number 4UH2C2017, 2017.

References

- [1] A. Mellit and A. M. Pavan, "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*, vol. 84, no. 5, pp. 807–821, 2010.
- [2] A. Ahmad, T. N. Anderson, and T. T. Lie, "Hourly global solar irradiation forecasting for New Zealand," *Solar Energy*, vol. 122, pp. 1398–1408, 2015.
- [3] K. Y. Bae, H. S. Jang, and D. K. Sung, "Hourly solar irradiance prediction based on support vector machine and its error analysis," *IEEE Transactions on Power Systems*, vol. 32, pp. 935–945, 2017.
- [4] T. T. Teo, T. Logenthiran, and W. L. Woo, "Forecasting of photovoltaic power using extreme learning machine," in *Proceedings of the 2015 IEEE Power & Energy Society's Innovative Smart Grid Technologies (ISGT) Asia*, vol. 1–6, Bangkok Thailand, 2015.
- [5] M. N. Akhter, S. Mekhilef, H. Mokhlis, and N. Mohamed Shah, "Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques," *IET Renewable Power Generation*, vol. 13, no. 7, pp. 1009–1023, 2019.
- [6] P. Mathiesen and J. Kleissl, "Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States," *Solar Energy*, vol. 85, no. 5, pp. 967–977, 2011.
- [7] C. Voyant, G. Notton, S. Kalogirou et al., "Machine learning methods for solar radiation forecasting: a review," *Renewable Energy*, vol. 105, pp. 569–582, 2017.

- [8] G. M. Yagli, D. Yang, and D. Srinivasan, "Automatic hourly solar forecasting using machine learning models," *Renewable and Sustainable Energy Reviews*, vol. 105, pp. 487–498, 2019.
- [9] R. Perez, S. Kivalov, J. Schlemmer, K. Hemker, D. Renné, and T. E. Hoff, "Validation of short and medium term operational solar radiation forecasts in the US," *Solar Energy*, vol. 84, no. 12, pp. 2161–2172, 2010.
- [10] H. Yang, B. Kurtz, D. Nguyen et al., "Solar irradiance forecasting using a ground-based sky imager developed at UC San Diego," *Solar Energy*, vol. 103, pp. 502–524, 2014.
- [11] A. Khosravi, R. N. N. Koury, L. Machado, and J. J. G. Pabon, "Prediction of hourly solar radiation in Abu Musa Island using machine learning algorithms," *Journal of Cleaner Production*, vol. 176, pp. 63–75, 2018.
- [12] H. Loutfi, A. Bernatchou, and R. Tadili, "Generation of horizontal hourly global solar radiation from exogenous variables using an artificial neural network in fes (Morocco)," *International Journal of Renewable Energy Research*, vol. 7, pp. 1097–1107, 2017.
- [13] I. Anwar and T. Khatib, "A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm," *Energy Conversion Management*, vol. 138, pp. 413–425, 2017.
- [14] S. Belaid, A. Mellit, H. Boualit, and M. Zaiani, "Hourly global solar forecasting models based on a supervised machine learning algorithm and time series principle," *International Journal of Ambient Energy*, vol. 41, pp. 1–12, 2020.
- [15] X. Qing and Y. Niu, "Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM," *Energy*, vol. 148, pp. 461–468, 2018.
- [16] K. Benmouiza and A. Cheknane, "Clustered ANFIS network using fuzzy c-means, subtractive clustering, and grid partitioning for hourly solar radiation forecasting," *Theoretical and Applied Climatology*, vol. 137, no. 1–2, pp. 31–43, 2019.
- [17] L. Benali, G. Notton, A. Fouilloy, C. Voyant, and R. Dizene, "Solar radiation forecasting using artificial neural network and random forest methods: application to normal beam, horizontal diffuse and global components," *Renewable Energy*, vol. 132, pp. 871–884, 2019.
- [18] R. Urraca, J. Antonanzas, M. Alia-Martinez, F. J. Martinez-De-Pison, and F. Antonanzas-Torres, "Smart baseline models for solar irradiation forecasting," *Energy Conversion and Management*, vol. 108, pp. 539–548, 2016.
- [19] B. Julitta, M. Vallverdú, U. S. P. Melia et al., "Auto-mutual information function of the EEG as a measure of depth of anesthesia," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2574–2577, Boston, MA, USA, 2011.
- [20] R. H. Inman, H. T. C. Pedro, and C. F. M. Coimbra, "Solar forecasting methods for renewable energy integration," *Progress in Energy and Combustion Science*, vol. 39, no. 6, pp. 535–576, 2013.
- [21] M. Alizamir, S. Kim, O. Kisi, and M. Zounemat-Kermani, "A comparative study of several machine learning based non-linear regression methods in estimating solar radiation: case studies of the USA and Turkey regions," *Energy*, vol. 197, Article ID 117239, 2020.
- [22] A. K. Yadav and S. S. Chandel, "Solar radiation prediction using Artificial Neural Network techniques: a review," *Renewable and Sustainable Energy Reviews*, vol. 33, pp. 772–781, 2014.
- [23] S. Karsoliya, "Approximating number of hidden layer neurons in multiple hidden layer BPNN architecture," *International Journal of Engineering Trends and Technology*, vol. 3, pp. 714–717, 2012.
- [24] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [25] A. Zendehboudi, M. A. Baseer, and R. Saidur, "Application of support vector machine models for forecasting solar and wind energy resources: a review," *Journal of Cleaner Production*, vol. 199, pp. 272–285, 2018.
- [26] M. A. Hassan, A. Khalil, S. Kaseb, and M. A. Kassem, "Exploring the potential of tree-based ensemble methods in solar radiation modeling," *Applied Energy*, vol. 203, pp. 897–916, 2017.
- [27] J. Lee, W. Wang, F. Harrou, and Y. Sun, "Reliable solar irradiance prediction using ensemble learning-based models: a comparative study," *Energy Conversion and Management*, vol. 208, Article ID 112582, 2020.
- [28] R. Aler, I. M. Galván, J. A. Ruiz-Arias, and C. A. Gueymard, "Improving the separation of direct and diffuse solar radiation components using machine learning by gradient boosting," *Solar Energy*, vol. 150, pp. 558–569, 2017.
- [29] A. Torres-Barrán, Á. Alonso, and J. R. Dorronsoro, "Regression tree ensembles for wind energy and solar radiation prediction," *Neurocomputing*, vol. 326–327, pp. 151–160, 2019.
- [30] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [31] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [32] M. W. Ahmad, M. Mourshed, and Y. Rezgui, "Trees vs Neurons: comparison between random forest and ANN for high-resolution prediction of building energy consumption," *Energy and Buildings*, vol. 147, pp. 77–89, 2017.
- [33] H. Bonakdari, H. Moeeni, I. Ebtehaj, M. Zeynoddin, A. Mahammadian, and B. Gharabaghi, "New insights into soil temperature time series modeling: linear or nonlinear?" *Theoretical and Applied Climatology*, vol. 135, no. 3–4, pp. 1157–1177, 2019.
- [34] C. Yousif, G. O. Quecedo, and J. B. Santos, "Comparison of solar radiation in marsaxlokk, Malta and valladolid, Spain," *Renewable Energy*, vol. 49, pp. 203–206, 2013.
- [35] <https://power.larc.nasa.gov>.
- [36] W. VanDeventer, E. Jamei, G. S. Thirunavukkarasu et al., "Short-term PV power forecasting using hybrid GASVM technique," *Renewable Energy*, vol. 140, pp. 367–379, 2019.
- [37] U. K. Das, K. S. Tey, M. Seyedmahmoudian et al., "Forecasting of photovoltaic power generation and model optimization: a review," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 912–928, 2018.
- [38] H. Yang, C. Huang, Y. Huang, and Y.-S. Pai, "A weather-based hybrid method for 1-day ahead hourly forecasting of PV power output," *IEEE Transactions on Sustainable Energy*, vol. 5, no. 3, pp. 917–926, 2014.
- [39] J. Shi, W.-J. Lee, Y. Liu, Y. Yang, and P. Wang, "Forecasting power output of photovoltaic systems based on weather classification and support vector machines," *IEEE Transactions on Industry Applications*, vol. 48, no. 3, pp. 1064–1069, 2012.