Available online at www.sciencedirect.com

ScienceDirect

Solar Energy 115 (2015) 632–644

SOLAR ENERGY

www.elsevier.com/locate/solener

A support vector machine–firefly algorithm-based model for global solar radiation prediction

Lanre Olatomiwa [a,c], Saad Mekhilef [a,*], Shahaboddin Shamshirband [b,*], Kasra Mohammadi [d], Dalibor Petković [e], Ch Sudheer [f]

[a] Power Electronics and Renewable Energy Research Laboratory (PEARL), Department of Electrical Engineering, Faculty of Engineering, University of Malaya, 50603 Kuala Lumpur, Malaysia

[b] Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

[c] Department of Electrical & Electronic Engineering, Federal University of Technology, PMB 65, Minna, Nigeria

[d] Faculty of Mechanical Engineering, University of Kashan, Kashan, Iran

[e] University of Niš, Faculty of Mechanical Engineering, Department for Mechatronics and Control, Aleksandra Medvedeva 14, 18000 Niš, Serbia

[f] Department of Civil and Environmental Engineering, ITM University, Gurgaon, Haryana 122017, India

Received 18 September 2014; received in revised form 16 December 2014; accepted 9 March 2015
Available online 2 April 2015

Communicated by: Associate Editor Christian A. Gueymard

Abstract

In this paper, the accuracy of a hybrid machine learning technique for solar radiation prediction based on some meteorological data is examined. For this aim, a novel method named as SVM–FFA is developed by hybridizing the Support Vector Machines (SVMs) with Firefly Algorithm (FFA) to predict the monthly mean horizontal global solar radiation using three meteorological parameters of sunshine duration ($\bar{n}$), maximum temperature (T_{max}) and minimum temperature (T_{min}) as inputs. The predictions accuracy of the proposed SVM–FFA model is validated compared to those of Artificial Neural Networks (ANN) and Genetic Programming (GP) models. The root mean square (RMSE), coefficient of determination (R^2), correlation coefficient (r) and mean absolute percentage error (MAPE) are used as reliable indicators to assess the models' performance. The attained results show that the developed SVM–FFA model provides more precise predictions compared to ANN and GP models, with RMSE of 0.6988, R^2 of 0.8024, r of 0.8956 and MAPE of 6.1768 in training phase while, RMSE value of 1.8661, R^2 value of 0.7280, r value of 0.8532 and MAPE value of 11.5192 are obtained in the testing phase. The results specify that the developed SVM–FFA model can be adjudged as an efficient machine learning technique for accurate prediction of horizontal global solar radiation.

Keywords: Support vector machine; Firefly algorithm; Hybrid model; Global solar radiation prediction; Meteorological parameters

1. Introduction

The long-term knowledge of solar radiation at any particular locations is essential for variety of areas such as agricultural, hydrological, ecological as well as solar energy applications. It has been proved that the abundant potential of solar energy can play an important role to meet the ever-growing energy demand of the world (Ming et al., 2014; Akikur et al., 2013; Azoumah et al., 2011; Bajpai and Dash, 2012; Hasan et al., 2012). Among different types of renewable resources, solar energy has attracted enormous attention because not only it is

* Corresponding authors. Tel.: +60 146266763 (S. Shamshirband).
E-mail addresses: saad@um.edu.my (S. Mekhilef), shamshirband@um.edu.my (S. Shamshirband).

http://dx.doi.org/10.1016/j.solener.2015.03.015
0038-092X/

Nomenclature

ANFIS	adaptive neuro fuzzy inference system	$\bar{n}$	monthly mean sunshine duration hour (h)
ANN	Artificial Neural Network	RBF	radial basis function
FFA	Firefly Algorithm	RMSE	root-mean-square error
FG	fuzzy genetic	R^2	coefficient of determination
GP	Genetic Programming	SVM	Support Vector Machine
$\overline{H}$	monthly mean global solar radiation ($MJ/m^2/day$)	$\overline{T}_{max}$	monthly mean maximum temperature (°C)
		$\overline{T}_{min}$	monthly mean minimum temperature (°C)

sustainable, but also it is abundant and environmental friendly (Akikur et al., 2013). Solar energy exploitation is beneficial in abatement of prevalent global warming, since it does not emit CO_2 or hazardous greenhouse gases. In electricity production, solar radiation study is a prerequisite for design and prediction of energy output of solar conversion system. The best way to obtain solar radiation data is from measurements taken remotely at a particular location using designated measuring instruments; due to required high cost for calibration and maintenance of the instruments, solar radiation data are limited in many meteorological stations around the world (Hunt et al., 1998). The difficulties and uncertainty involve in the measurement of global solar radiation have resulted in development of so many models and algorithms for its estimation from some routinely measured meteorological variables consisting sunshine hour, maximum, minimum and average air temperature, relative humidity, cloud factor, etc. In Nigeria, numerous of the government owned meteorological stations have no record of solar radiation data, even where the record are available there are some missing days or month without record possibly due to improper calibration of measuring equipment employed. Over the past years, a vast number of methods including the empirical models (Angstrom, 1924; Hargreaves and Samani, 1982; Bristow and Campbell, 1984; Besharat et al., 2013; Halawa et al., 2014), satellite-derived model (Pinker et al., 1995; Viana et al., 2011) and stochastic algorithm model (Markov chain) (Hocaoğlu, 2011; Amato et al., 1986; Aguiar et al., 1988) have been developed for estimating the global solar radiation on a horizontal surface. Empirical models have been widely developed and used to correlate the global solar radiation with various routinely measured meteorological and geographical parameters. In many researches, the parameters such as sunshine duration, maximum and minimum temperatures have been recognized as the most proper elements for solar radiation prediction (Besharat et al., 2013; Trnka et al., 2005; Chen and Li, 2013; Wu et al., 2007). However, due to inaccessibility of sunshine duration data in some locations, some studies have proved that good estimations can be attained by using measured maximum and minimum temperature as inputs (Hargreaves and Samani, 1982; Bristow and Campbell, 1984; Liu et al., 2009). Although, application of satellite-based methods seems promising for estimation of solar radiation over a large region, their main drawbacks are the required cost and lack of sufficient historical data because it is relatively new. These methodologies have shown low performance when forecasting/modeling data on long term basis; they are also not suitable when there are some missing data in the database. However, one way to overcome these problems is utilization of artificial intelligence techniques.

In Nigeria, several works have been carried out on predictions of solar radiation using the conventional empirical models (Ezekwe and Ezeilo, 1981; Sambo, 1986; Akpabio and Etuk, 2003; Layi Fagbenle, 1993; Ajayi et al., 2014). Nevertheless, due to necessity of accurate and reliable solar radiation, artificial and computational intelligence techniques have been broadly applied to estimate solar radiation in many regions around the word. Al-Alawi and Al-Hinai (1998) predicted solar radiation for a location with no availability of measured data. They used monthly mean daily values of temperature, pressure, relative humidity, sunshine duration hours and wind speed as inputs for Artificial Neural Networks (ANN) technique to predict global solar radiation. They compared the results with empirical methods model and found more accuracy for ANN-based model. Mellit et al. (2006) employed the combination of neural and wavelet network to forecast daily solar radiation for photovoltaic (PV) sizing application. In their study, wavelets served as activation function. Their results of the forecast demonstrated the more favorable performance of the approach compared to other neural network models. In Jiang (2009), a ANN model was developed to estimate monthly mean daily solar radiation for eight typical cities in China. The achieved results were compared to those of conventional empirical models. The statistical analysis results indicated a good correlation between estimated values by the ANN model and the actual data with higher accuracy than other empirical models.

Behrang et al. (2011) applied particle swarm optimization (PSO) technique to estimate monthly mean daily global solar radiation on a horizontal surface for 17 cities in different regions of Iran. Their results showed better performance of PSO-based models compared to the traditional empirical models. Mohandes (2012) employed PSO algorithm to train ANN in other to model the monthly mean daily global solar radiation values in Saudi Arabia. Different parameters such as month number, sunshine duration, latitude, longitude, and altitude of the location were considered as inputs. The developed hybrid PSO–ANN model showed a better performance compared to back-propagation trained neural network (BP-NN). Benghanem et al. (2009; Ornella and Tapia, 2010) developed six ANN-based models to estimate horizontal global solar radiation at Al-Madinah in Saudi Arabia. They utilized different combinations of input parameters consisting sunshine hours, ambient temperature, relative humidity and the day of year. Their results showed that the model with higher accuracy is dependent upon sunshine duration and air temperature. Ramedani et al. (2014; Jain et al., 2009) employed support vector regression (SVR) technique to develop a model for prediction of global solar radiation in Tehran, Iran. They used two SVRs models of radial basis function (SVR-rbf) and polynomial function (SVR-poly). They found more superiority for SVR-rbf technique. In another study, Ramedani et al. (2014; Bao et al., 2013) performed a comparative investigation between fuzzy linear regression (FLR) and support vector regression (SVR) techniques to predict global solar radiation in Tehran, Iran. They found that SVR-rbf approach enjoy superior performance compared to FLR. Also, in some studies, different techniques were combined to propose a hybrid approaches with more accuracy. Wu et al. (2014; Friedrichs and Igel, 2005) developed a genetic algorithm combing multi-model framework to predict solar radiation. Bhardwaj et al. (2013; Lorena and De Carvalho, 2008) proposed a hybrid approach which includes hidden Markov models and generalized fuzzy models to estimate solar irradiation in India. They assessed the influence of different meteorological parameters for estimation of solar radiation using the developed model. Wu et al. (2014; Hsu et al., 2003) combined the Autoregressive and Moving Average (ARMA) model with the controversial Time Delay Neural Network (TDNN) for prediction of hourly solar radiation. Salcedo-Sanz et al. (2014; Chung et al., 2003) assessed the capability of a novel Coral Reefs Optimization–Extreme Learning Machine (CRO–ELM) algorithm to predict the global solar radiation at Murcia (southern Spain) using different meteorological data. Hung et al. (2013; Chapelle et al., 2002) developed a hybrid Auto Regressive and Dynamical System (CARDS) model to forecast hourly global solar radiation in Mildura, Australia.

Generally, Support Vector Machines (SVMs) is a type of machine learning technique that has gained importance in environmental related applications (Ornella and Tapia, 2010; Jain et al., 2009). SVM is a learning algorithms employing high dimensional feature. The correctness of an SVM model is to a great extent relies on determination of its model parameters. Even though organized strategies for selecting parameters are important, model parameters alignment also need to be made. In the past, although some researchers have applied various conventional optimization algorithms to select these parameters, the achieved results have not been so effective due to the complex nature of the parameters (Bao et al., 2013; Friedrichs and Igel, 2005; Lorena and De Carvalho, 2008). Grid search algorithm (Hsu et al., 2003) and gradient decent algorithm (Chung et al., 2003; Chapelle et al., 2002) are among the algorithms that have been employed earlier. Computational complexity is a major drawback of grid search algorithm; thus, it only applicable to area involving fewer parameter selection. On the other hand, grid search algorithm is usually prone to local minima. In most optimization problems, multiple local solution do exist, but evolutionary algorithms seems to be the best approach due to the fact that they are capable of providing global solution to such optimization problems.

In this study, a hybrid approach by integrating Support Vector Machine (SVM) and Firefly Algorithm (FFA) has been developed to predict the global solar radiation. The Firefly Algorithm (FFA) is applied to determine optimal SVM parameters. The main objective of the study is to investigate the suitability of the proposed combined method (SVM–FFA) for prediction of monthly mean daily global solar radiation on a horizontal surface. To achieve this, three locations distributed in different regions of Nigeria have been considered to analyze the influence of weather conditions on the capability of the developed approach. Three widely available meteorological parameters of sunshine duration, maximum air temperature and minimum temperature are considered as inputs to predict the global solar radiation. These inputs are chosen because of their high availability in most areas and their strong correlations with the global solar radiation. The motivation behind this investigation is centered upon the significance of reliable solar radiation data in many applications including agricultural productions, hydrological and ecological studies as well as assessments and prediction of energy output of solar systems. Also, in most cases the solar radiation data are not readily available due to several issues. To validate the precision of developed SVM–FFA approach its capability is compared to Artificial Neural Network (ANN) and Genetic Programming (GP).

2. Materials and methods

2.1. Descriptions of study sites and data set

In this study, long-term monthly average daily global solar radiation on a horizontal surface ($\overline{H}$), sunshine

Table 1
The geographical information of the nominated locations.

Location	Zone	Latitude (°N)	Longitude (°E)	Altitude (m)
Iseyin	SW	7.96	3.60	330
Maiduguri	NE	11.83	13.15	353.8
Jos	NC	9.92	8.9	1217

SW (South-West); NE (North-East); NW (North-Central).

duration ($\bar{n}$), maximum air temperature ($\overline{T}_{max}$) and minimum air temperature ($\overline{T}_{min}$), for the period of 21 years from 1987 to 2007 for three sites of Iseyin, Maiduguri and Jos distributed in different regions of Nigeria were used. These data were measured at respective metrological station located in each sites courtesy of the Nigerian Meteorological Agency (NIMET) NIMET, 2014. The geographical information of the selected sites is presented in Table 1. Also, Fig. 1 shows the locations of the considered sites on the map of Nigeria. According to NIMET (NIMET, 2014), the measured solar radiation data were recorded using Gunn-Bellini radiometer. This instrument produce a time-oriented parameter of solar radiation falling on a black body by measuring volume of the liquid distilled in a calibrated tube (Ajayi et al., 2014; McCulloch and Wangati, 1967). To measure the sunshine duration, Campbell strokes sunshine recorder were used. Also, minimum and maximum dry bulb thermometers were used to measure both maximum and minimum air temperatures at the selected stations. The monthly mean daily data used for this research work were divided in two sets of training and testing. For the experiments, 80% (202 data set) for the period 1987–2003 were used for sample training and the remaining 20% (50 data set) in the period 2004–2007 are used for testing.

The variation of long-term averaged monthly mean daily horizontal global solar radiation, sunshine duration, maximum ambient temperature and minimum ambient temperature is shown in Fig. 2(a–d) for one of the sites considered in the study (Iseyin). From this figure, it can be seen that the variation of each parameters are closely related to solar radiation data. The annual mean solar radiation of this site is 16.34 MJ/m^2/day, while the annual mean bright day sunshine hour found to be 5.5 h, with highest value (7 h) observed in November and lowest (3.2 h) in August. The monthly mean daily maximum temperature ranges between 27.4 °C in August and 35.5 °C in February, while the minimum value ranged from 20.3 °C in January to 23.6 °C in March.

2.2. Support Vector Machine (SVM)

Given a set of data points represented by $= \{x_i, d_i\}_i^n$, where x_i is the input space vector of the data sample, d_i is the target value and n is the number of data points. Support Vector Machine (SVM) equations based on Vapnik's theory (Vapnik and Vapnik, 1998; Yang et al., 2009; Vapnik, 2000) approximates the function as:

$$f(x) = w\varphi(x) + b \tag{1}$$

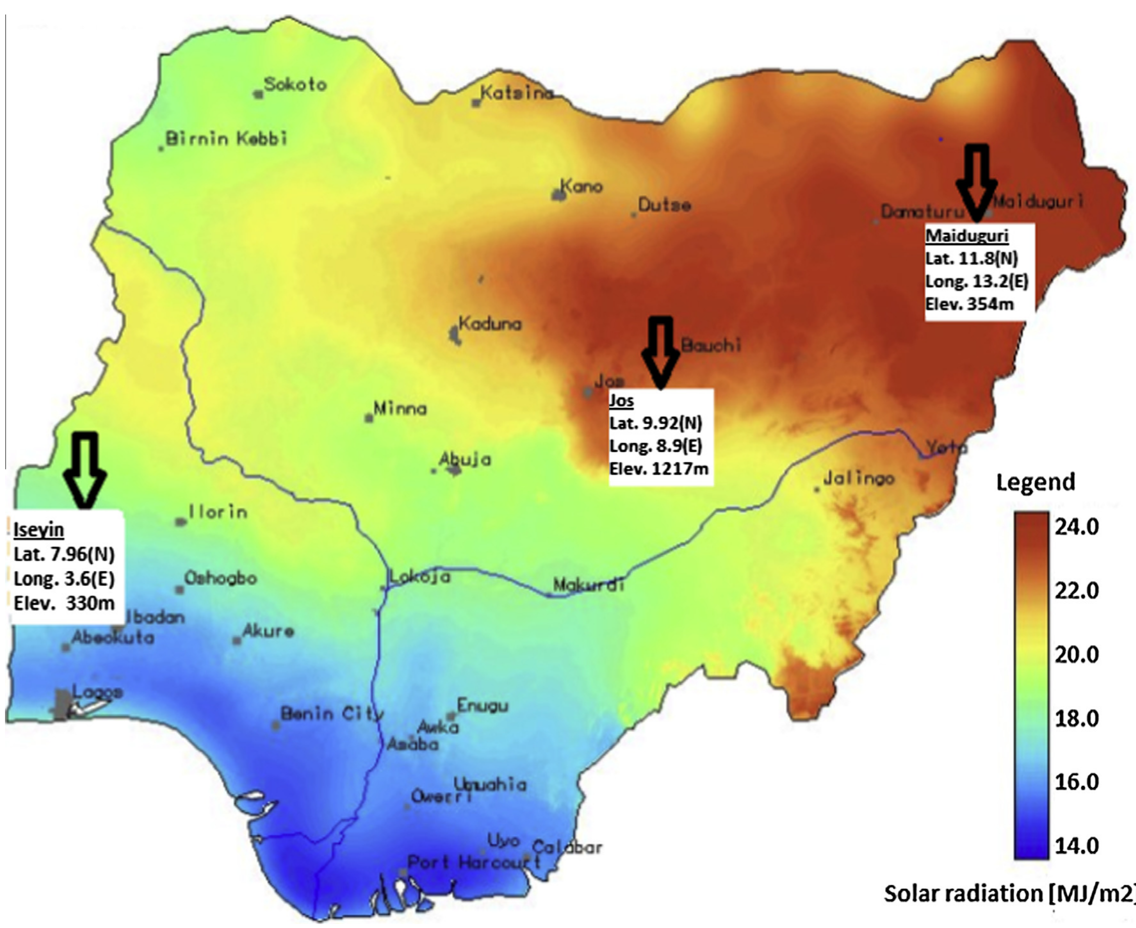


Fig. 1. Nigeria map showing the locations of the considered case studies.

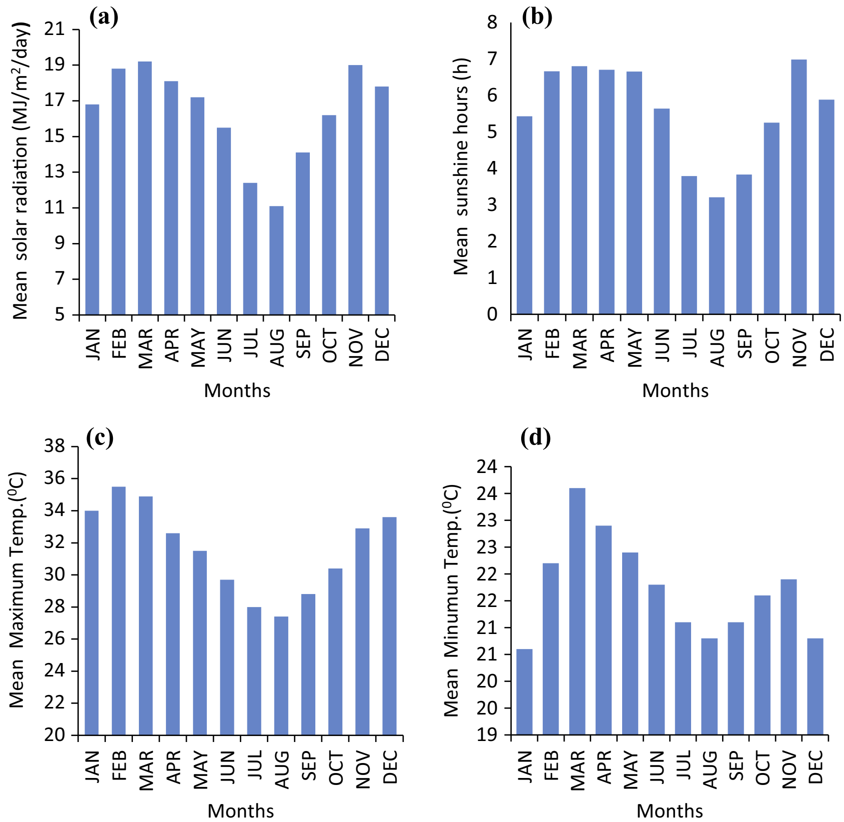


Fig. 2. Mean monthly distribution of (a) solar radiation; (b) sunshine hour; (c) minimum temperature; (d) maximum temperature.

$$R_{\mathrm{SVMs}}(C) = \frac{1}{2}\|w\|^2 + C\frac{1}{n}\sum_{i=1}^{n} L(x_i, d_i) \tag{2}$$

where $\varphi(x)$ represents high dimensional-space features that map the input space vector x, w is a normal vector, b is a scalar, and $C\frac{1}{n}\sum_{i=1}^{n} L(x_i, d_i)$ represents the empirical risk. The parameters w and b can be estimated by minimization of regularized risk function after introduction of positive slack variables ξ_i and ξ_i^* that represent upper and lower excess deviation (Vapnik and Vapnik, 1998).

$$\text{Minimize } R_{\mathrm{SVMs}}(w, \xi^*, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{3}$$

$$\text{Subject to } \begin{cases} d_i - w\varphi(x_i) + b_i \leqslant \varepsilon + \xi_i \\ w\varphi(x_i) + b_i - d_i \leqslant \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geqslant 0, \quad i = 1, \ldots, l \end{cases}$$

where $\frac{1}{2}\|w\|^2$ represent the regularization term, C is the error penalty factor used to control the trade-off between the regularization term and empirical risk, ε is the loss function, which equates to approximation accuracy of the training data point, and l is the number of elements in the training data set.

Eq. (1) can be solved with the introduction of Lagrange multiplier and optimality constraints, hence obtaining a generic function given by

$$f(x) = \sum_{i=1}^{l}(\beta_i - \beta_i^*)K(x_i, x_j) + b \tag{4}$$

where $K(x_i, x_j) = \varphi(x_i)\varphi(x_j)$ and the term $K(x_i, x_j)$ is called the kernel function, which is an inner product of the two vector x_i and x_j in the feature space $\varphi(x_i)$ and $\varphi(x_j)$, respectively. This inner product space is a vector space with an additional structure called an inner product. This additional structure associates each pair of vectors in the space with a scalar quantity known as the inner product of the vectors. Inner products allow the rigorous introduction of intuitive geometrical notions such as the length of a vector or the angle between two vectors. The main purpose of SVMs is to carry out data correlation via non-linear mapping. Kernel methods enables to operate in a high-dimensional, implicit feature space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data in the feature space. This operation is often computationally cheaper than the explicit computation of the coordinates. This approach is known as a direct computation method of a kernel function, denoted by K. The results obtained in the higher-dimensional feature space correspond to the results of the original, lower-dimensional input space.

There are four basic kernel functions provided by SVM, namely, lineal, sigmoid, polynomial, and radial basis functions. But over the years, radial basis function (RBF) has been proved to be the best kernel function due to its computationally efficiency, simplicity, reliability, ease of adaptation to optimization and other adaptive techniques as well as its adaptability in handling parameters that are very complex (Yang et al., 2009; Rajasekaran et al., 2008; Wu and Wang, 2009). RBF kernel function only need the solution of a set of linear equations instead of the lengthy and computationally demanding quadratic programming problem for its training (Shamshirband et al., 2014). The non-linear radial basis kernel function is defined as:

$$K(x_i, x_j) = e^{\gamma \|x_i - x_j\|^2} \tag{5}$$

where x_i and x_j are vectors in the input space, i.e. vectors of features computed from training or test samples. Parameter γ is represented as $\gamma = -\frac{1}{2\sigma^2}$, where σ denotes Gaussian noise level of standard deviation.

The three parameters associated to RBF Kernels are γ, ε and C. The accuracy of SVM model is principally dependent on model parameter selection. In our scheme, a default value of $e = 0.1$ seemed to perform well. To select user-defined parameters (i.e. γ, ε and C), a large number of trials were carried out with different combinations of C and γ for the radial basis function kernel.

2.3. SVM parameters selection using firefly optimization algorithm

Firefly Algorithm (FFA) is a metaheuristic search algorithm, which is based on the social dashing behavior of fireflies in nature (Łukasik and Żak, 2009; Yang, 2010a,b). The two main issues in FFA are the variation of light intensity and formulation of attractiveness. In the case of optimal design considering maximization of objective function, the objective function is proportional to the brightness or light intensity emitted by a firefly. The Gaussian form of the light intensity I with varying distance can be written as

$$I = I_0 e^{-\gamma r^2} \tag{6}$$

where I is the light intensity at distance r from a firefly, I_0 represent initial light intensity, i.e. when $r = 0$ and γ is the light absorption coefficient which value varies between 0.1 and 10, As a firefly's attractiveness is proportional to the light intensity observed by adjacent fireflies, the attractiveness ω at a distance r from the firefly is given as:

$$\omega(r) = \omega_0 e^{-\gamma r^2}, \tag{7}$$

where ω_0 is the attractiveness at $r = 0$. Cartesian distance between any two fireflies i and j at x_i and x_j, respectively, is represented as:

$$r_{ij} = \|x_i - x_j\| = \sqrt{\sum_{k=1}^{n} (x_{i,k} - x_{j,k})}, \tag{8}$$

where n denotes the dimensionality of the problem, $x_{i,k}$ is the kth component of the spatial coordinate x_i of the ith firefly and $x_{j,k}$ is the kth component of the spatial coordinate x_j of the jth firefly. The movement of firefly I, attracted to another brighter firefly j can be represented as:

$$x_i = x_i + \omega_0 e^{-\gamma r^2}(x_j - x_i) + \alpha\varepsilon_i \tag{9}$$

where the first term in Eq. (9) is due to the attraction, the second term represents the randomization with α as randomization coefficient whose value is lies between 0 and 1. And ε_i the random number vector derived from a Gaussian distribution. Fig. 3 depicts the flow chart for obtaining the optimal SVM parameters.

2.4. Input parameters

The capability of the SVM to make good estimations is dependent on input parameters selection. In this study, the monthly mean daily values of $\bar{n}$, $\overline{T}_{\min}$ and $\overline{T}_{\max}$ for the period of 1987 to 2007 were used as inputs to generate the SVM model. The criteria for choosing these meteorological parameters as inputs include their high availability in most areas and their strong correlations with the horizontal global solar radiation. Thus, it is anticipated that the developed models using these inputs provide favorable precision to predict the global solar radiation. Fig. 4 presents schematic diagram of the proposed SVM–FFA global solar radiation model based upon the considered input parameters.

In order to obtain reliable evaluation and comparison, SVM model are tested with data set that have not been used during the training process. The statistical parameters (minimum value, maximum value, mean, standard deviation and variation coefficient) for data sets are calculated and given in Tables 2. The standard deviation in the table

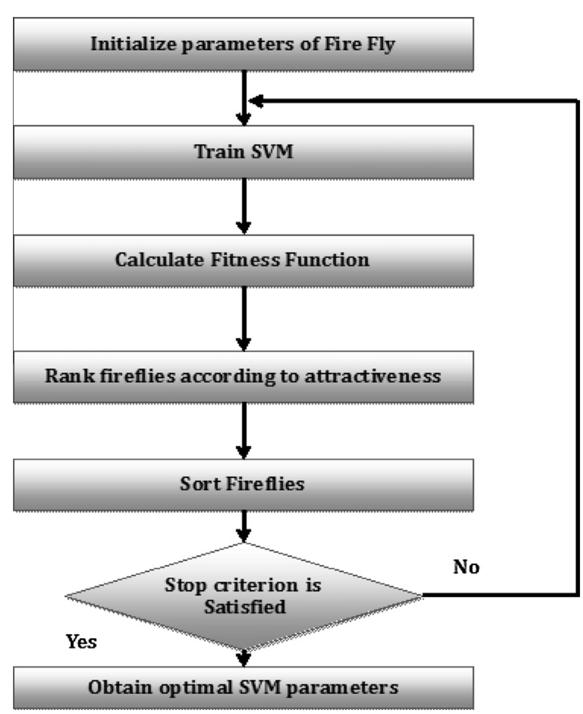


Fig. 3. Flow chart of the proposed FFA-based parameter determination approach for the SVM classifier.

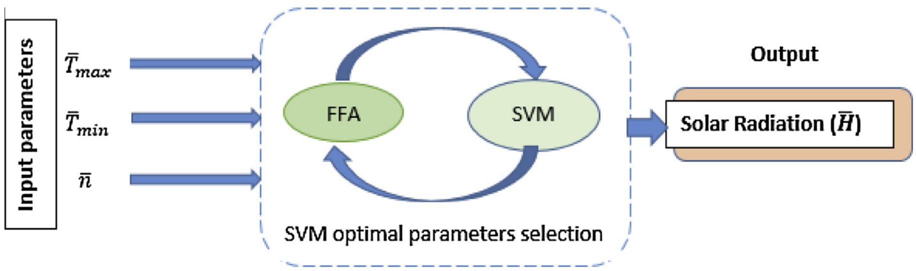


Fig. 4. Schematic diagram of the proposed SVM–FFA model for global solar radiation prediction.

Table 2
Statistical parameters for data sets.

Statistical parameter	Iseyin			Maiduguri			Jos		
	$\bar{n}$	$\bar{T}_{min}$	$\bar{T}_{max}$	$\bar{n}$	$\bar{T}_{min}$	$\bar{T}_{max}$	$\bar{n}$	$\bar{T}_{min}$	$\bar{T}_{max}$
Min	1.30	18.0	22.8	4.40	9.20	28.0	3.10	7.20	22.6
Max	8.40	33.7	37.1	10.8	28.1	42.6	10.7	24.8	33.3
Mean	5.50	21.7	31.6	8.31	20.3	35.2	7.33	15.8	27.8
Std. deviation	1.44	1.31	2.84	1.28	4.72	3.43	1.86	2.98	2.33
Variation coefficient	2.08	1.72	8.067	1.63	11.8	22.3	3.45	8.88	5.41

indicates the distribution of the data around the mean, indicating the degree of consistency of the data.

2.5. Artificial Neural Networks (ANN)

Artificial Neural Network (ANN) is a mathematical model that performs a computational simulation of the behavior of neuron in the human brain by replicating the brain's pattern to produce results based on the learning of set of training data (Izgi et al., 2012). The multi layer feed-forward network with a back-propagation learning algorithm is one of the most popular neural network architectures. It has been deeply studied and widely used in many fields (Chen and Jain, 1994; Gardner and Dorling, 1998; Behrang et al., 2010). Typically, a neural network consists of three layers: (1) an input layer; (2) an output layer; and (3) an intermediate or hidden layer (Schalkoff, 1997). The input vectors are $\epsilon\ R^n$ and $D = (X_1, X_2, \ldots, X_n)^T$; the outputs of q neurons in the hidden layer are $Z = (Z_1, Z_2, \ldots, Z_n)^T$; and the outputs of the output layer are $Y \epsilon\ R^m$, $Y = (Y_1, Y_2, \ldots, Y_n)^T$. Assuming that the weight and the threshold between the input layer and the hidden layer are w_{ij} and y_j, respectively, and that the weight and the threshold between the hidden layer and output layer are w_{jk} and y_k respectively, the outputs of each neuron in a hidden layer and output layer are;

$$Z_j = f\left(\sum_{i=1}^{n} w_{ij} X_i - \theta_j\right) \tag{10}$$

$$Y_k = f\left(\sum_{j=1}^{q} w_{kj} Z_j - \theta_k\right) \tag{11}$$

where $f()$ is a transfer function, which is the rule for mapping the neuron's summed input to its output, and by a suitable choice it is a means of introducing a non-linearity into the network design. One of the most commonly used functions is the sigmoid function, which is monotonic increasing and ranges from 0 to 1.

For the validation of the performance of the proposed model, a typical feed forward neural network consisting of three (3) input, one (1) hidden layer with seven (7) neuron and one (1) output layer were used. The structure of the neural network is shown in Fig. 5, while Table 3 summarizes the parameters used in the ANN model.

2.6. Genetic Programming (GP)

Genetic Programming (GP) is a systematic and domain-independent method based on Darwinian theories of natural selection and survival to approximate the equation in symbolic form (Koza, 1992). The algorithm considers an initial population of randomly generated programs (equations), derived from the random combination

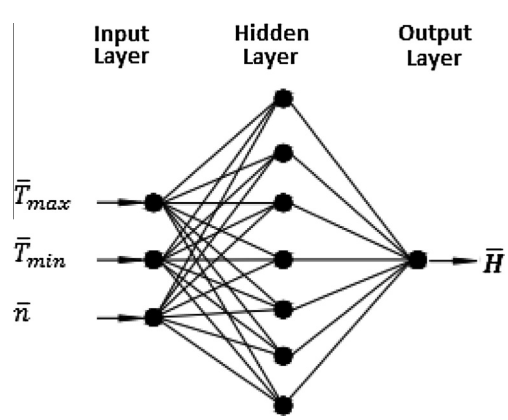


Fig. 5. ANN model use for the validation.

Table 3
User-defined parameters for ANN.

ANN parameters				Activation function
Learning rate	Momentum	Hidden node	Number of iteration	
0.2	0.1	3, 6, 10	1000	Continuous log-sigmoid function

of input variables, random numbers and functions, which include arithmetic operators (+, −, ×, ÷), mathematical functions (sin, cos, exp, log), logical/comparison functions, etc., which have to be appropriately chosen based on some understanding of the process. This population of potential solutions is then subjected to an evolutionary process and the 'fitness' of the evolved programs is evaluated. Individual programs that best fit the data are then selected from the initial population. The programs that are the best fit are then selected to exchange part of the information between them to produce better programs through 'crossover' and 'mutation', which mimics the natural world's reproduction process. Exchanging the parts of the best programs with each other is called crossover, and randomly changing programs to create new programs is called mutation. The programs that fitted the data less well are discarded. This evolution process is repeated over successive generations and is driven toward finding symbolic expressions describing the data, which can be scientifically interpreted to derive knowledge about the process. The parameters used per run of GP are summarized in Table 4.

2.7. Model performance evaluation

To assess the success of the SVM models and other selected techniques, some statistical indicators were examined as follows:

(1) Root-mean-square error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}}, \tag{12}$$

(2) Coefficient of determination (R^2)

$$R^2 = \frac{\left[\sum_{i=1}^{n}\left(O_i - \overline{O_i}\right)\cdot\left(P_i - \overline{P_i}\right)\right]^2}{\sum_{i=1}^{n}\left(O_i - \overline{O_i}\right)\cdot\sum_{i=1}^{n}\left(P_i - \overline{P_i}\right)} \tag{13}$$

(3) Correlation coefficient (r)

$$r = \frac{\sum_{i=1}^{n}\left(O_i - \overline{Q_i}\right)\cdot\left(P_i - \overline{P_i}\right)}{\sqrt{\sum_{i=1}^{n}\left(O_i - \overline{Q_i}\right)\cdot\sum_{i=1}^{n}\left(P_i - \overline{P_i}\right)}} \tag{14}$$

(4) Mean absolute percentage error (MAPE)

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{x}\left|\frac{O_i - P_i}{P_i}\right| \times 100, \tag{15}$$

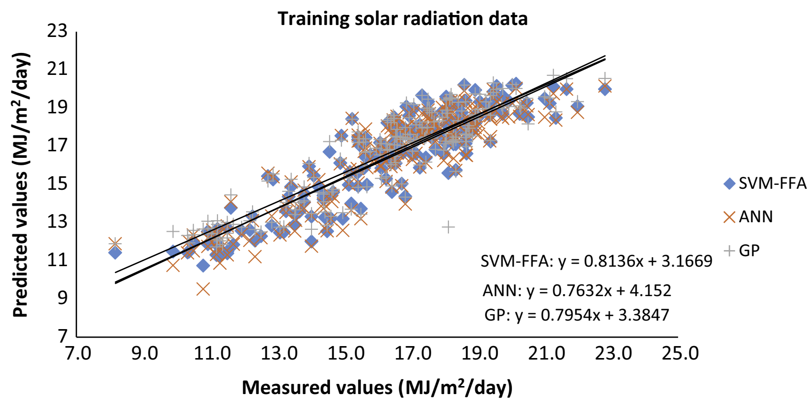
where P_i and O_i are known as the experimental and forecast values, respectively, while $\overline{P_i}$ and $\overline{Q_i}$ are the mean value of P_i and O_i respectively and n is the total number of test data. The RMSE value provides information on the short term performance of the correlation by comparing the extent of deviation of the predicted value from the actual measured value, while R^2 and r is a measure that allows one to determine the certainty of the predictions from the actual value. The smaller the value of RMSE and MAPE the the better the performance model and vice versa in the case of R^2 and r.

3. Results and discussions

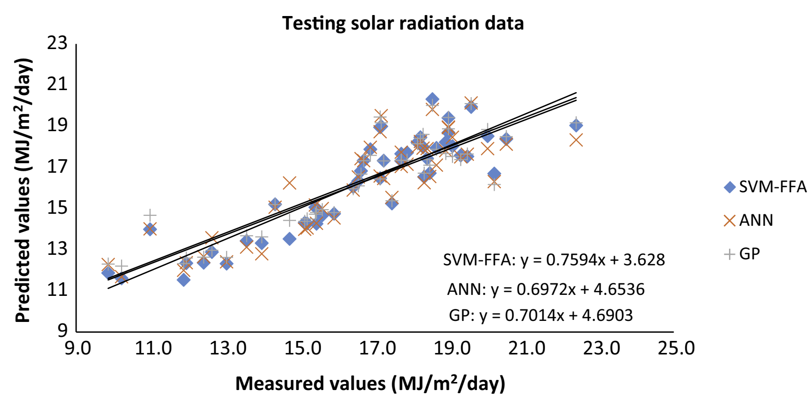
In this study, a hybrid approach by integrating the Support Vector Machine (SVM) with Firefly Algorithm (FFA) has been proposed to predict the monthly mean

Table 4
Parameters used in GP modeling.

Population size	512
Function set	$+, -, *, /, \sqrt{}, x^2, \ln(x), e^x, a^x$
Chromosomes	20–30
Head size	5–9
Number of genes	2–3
Linking functions	Addition, subtraction, arithmetic, trigonometric, multiplication
Fitness function error type	RMSE
Mutation rate	91.46
Inversion rate	108.53
Crossover rate	30.56
Homologs crossover rate	98.46
One-point recombination rate	0.2
Two-point recombination rate	0.2
Gene recombination rate	0.1
Gene transposition rate	0.1



(a)



(b)

Fig. 6. (a) Scatter plots of training data and predicted values using three machine learning models and (b) scatter plots of tested data and predicted values using three machine learning models.

global solar radiation on a horizontal surface in three locations distributed in different parts of Nigeria. Three widely available meteorological parameters consisting sunshine duration, maximum and minimum ambient temperatures have been considered as input elements to simulate the solar radiation. The suitability level of new hybrid approach named SVM–FFA is compared to Artificial Neural Network (ANN) and Genetic Programming (GP).

3.1. SVM model analysis

At the beginning, the SVM network was trained with measured data by above presented experimental procedure. After training process the SVM network were tested to determine the solar radiation. Based on the experiments, the input parameters (monthly mean value of minimum temperature, maximum temperature and sunshine duration) and output (solar radiation) are collected and defined for the learning techniques. For the experiments, 80% (202 data set) for the period 1987–2003 were used for sample training and the remaining 20% (50 data set) in the period 2004–2007 are used for testing. We analyzed the SVM model for solar radiation estimation based on the three inputs, monthly mean minimum temperature, monthly mean maximum temperature and monthly mean sunshine duration hours.

The estimated solar radiation is represented in Fig. 6 in the form of a scatterplot by three methodologies, SVM–FFA, ANN and GP. The training data of solar radiation and predicted values are shown in Fig. 6(a), while Fig. 6(b) presents testing data of solar radiation and predicted values by using the three machine learning models.

Finally, Fig. 7(a) and (b) shows comparative forecasting of solar radiation by SVM–FFA technique with ANN and GP results. It can be observed that SVM–FFA has better forecasting abilities for global solar radiation prediction than ANN and GP methods.

3.2. Performance analysis

In order to evaluate the performance of the proposed model, experimental work was carried out to determine the importance of each independent input variable on the output. Root-mean-square error (RMSE), coefficient of determination (R^2), correlation coefficient (r) and mean absolute percentage error (MAPE) served to evaluate the differences between the predicted and actual values for both SVMs models. Tables 5–7 compares the single SVM–FFA model with Artificial Neural Network (ANN)

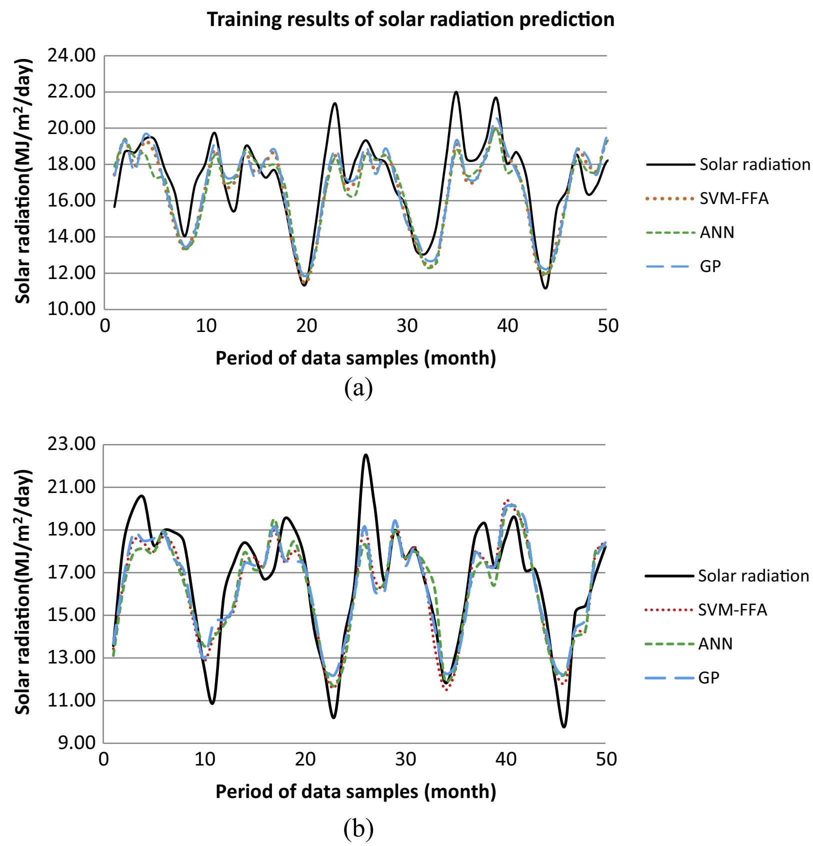


Fig. 7. Forecasting of solar radiation by SVM–FFA, ANN and GP in (a) training and (b) testing phase.

Table 5
Performance statistics of the SVM–FFA model compares to other methodologies for Iseyin.

		RMSE	R^2	r	MAPE
SVM–FFA	Training	0.4662	0.8183	0.9046	6.1754
	Testing	0.4935	0.7953	0.8918	6.2253
ANN	Training	0.4801	0.7895	0.8937	6.6195
	Testing	0.5502	0.7457	0.8635	6.9862
GP	Training	0.5301	0.7987	0.8885	7.0605
	Testing	0.5202	0.7678	0.8762	6.4681

Table 6
Performance statistics of the SVM–FFA model compares to other methodologies for Maiduguri.

		RMSE	R^2	r	MAPE
SVM–FFA	Training	0.5357	0.8253	0.9084	6.2616
	Testing	2.4934	0.2095	0.4577	13.985
ANN	Training	0.6317	0.7782	0.8821	6.6821
	Testing	2.6083	0.1334	0.3652	16.075
GP	Training	0.5522	0.7872	0.8872	6.6035
	Testing	2.5498	0.1632	0.4039	15.685

and Genetic Programming (GP) models for Iseyin, Maiduguri and Jos site respectively, while Table 8 summarizes the comparison results of the predictions in the three considered case studies in Nigeria. The results in Tables 5–8 indicate that the SVM model has the best capabilities of estimating the solar radiation.

As we can see from Tables 5–7, the performance capability of our SVM–FFA approach is different between

Table 7
Performance statistics of the SVM–FFA model compares to other methodologies for Jos.

		RMSE	R^2	r	MAPE
SVM–FFA	Training	1.094	0.7637	0.8739	6.0934
	Testing	2.611	0.5852	0.7650	14.347
ANN	Training	1.190	0.7112	0.8433	7.0421
	Testing	2.979	0.5187	0.7202	17.230
GP	Training	1.170	0.6999	0.8366	7.2141
	Testing	2.790	0.6233	0.7895	17.474

Table 8
Summary of performance statistics of the SVM–FFA model compares to other methodologies for Nigeria.

		RMSE	R^2	r	MAPE
SVM–FFA	Training	0.6988	0.8024	0.8956	6.1768
	Testing	1.8661	0.5300	0.7280	11.5192
ANN	Training	0.7673	0.7596	0.8730	6.7813
	Testing	2.0458	0.4659	0.6496	13.4305
GP	Training	0.7507	0.7619	0.8708	6.9594
	Testing	1.9532	0.5181	0.6899	13.2089

Table 9
Comparison between SVM–FFA models with other models.

Reference	Model type	Inputs parameters	Country of study	Coefficient of determination (R^2)
Yohanna et al. (2011)	Empirical	3	Nigeria	0.608
Ramedani et al. (2014)	ANN	7	Iran	0.799
Ramedani et al. (2014)	ANFIS	7	Iran	0.801
Ramedani et al. (2014)	SVR-rbf	7	Iran	0.790
Present study	SVM–FFA	3	Nigeria	0.802

the three considered sites. In one site the performance is higher than others. However, the main point is the fact that the performance of SVM–FFA model was compared to the ANN and GP models and the achieved results revealed that SVM–FFA is the superior approach. The accuracy of the proposed model is also compared with some conventional solar radiation prediction model and several artificial intelligent (AI) based model as presented in Table 9. According to the table, the proposed model is seen to give accurate results than the conventional methods and some of the previously proposed AI models in terms of coefficient of determination (R^2).

4. Conclusion

In this paper, a new hybrid machine learning approach for prediction of horizontal global solar radiation is proposed. To achieve this, we combined Support Vector Machine (SVM) with Firefly Algorithms (FFA) to enhance the predictions accuracy. The simulation studies using long-term measured data obtained from Nigerian meteorological Agency (NIMET) for three sites in different geopolitical zone of the country have yielded several conclusions. The main idea of the study centers on investigation of the feasibility of the proposed hybrid techniques to model the relationship between solar radiation and some other meteorological parameters. In the proposed model, temperature measurements (minimum and maximum) as well as sunshine duration serves as the inputs, and the choice of these input parameters is not far fetch from the obvious reasons of their high availability in most areas, their strong correlations with the global solar radiation, as well as the simplicity and cheapness of the equipment required for their measurements.

To validate the precision of developed SVM–FFA approach its capability is compared to Artificial Neural Network (ANN) and Genetic Programming (GP). It could be seen from the analysis, the performance of developed model including the ANN and GP models vary from one station to another, this is because the model is highly dependent upon the solar radiation characteristics and weather conditions of the locations. Basically, solar radiation estimation is totally location dependent; therefore calibrating a general model to estimate the solar radiation for an entire region including several stations would only be possible option if the climate conditions of the region are similar. Otherwise, the amount of errors obtained may be high for some stations with different weather conditions.

The statistical indicator used for performance evaluation of the proposed model indicates lower values of RMSE and MAPE and higher values of R^2 and r when compared to ANN and GP model for all the station considered. The obtained results by SVM–FFA model were: RMSE of 0.6988, R^2 of 0.8024, r of 0.8956 and MAPE of 6.1768 in training phase while, RMSE value of 1.8661, R^2 value of 0.7280, r value of 0.8532 and MAPE value of 11.5192 are obtained in the testing phase. The achieved results demonstrated that the proposed hybrid SVM–FFA approach would be an appealing option to predict global solar radiation since the results were favorable for all considered case studies despite different climate conditions.

Based on these, the proposed SVM–FFA model can therefore be adjudged an efficient machine learning approach for accurate prediction of horizontal global solar radiation. However, the model is open for further improvements, as several other combinations of meteorological data such as; air pressure, humidity, sunshine duration, cloud index and many more can be incorporated into the model and further analysis of this can be considered as future study.

Acknowledgments

The authors would like to thank the Ministry of Higher Education, Malaysia and the Bright Spark Unit of University of Malaya, Malaysia for providing the enabling environment and financial support under the grant No. UM.C/HIR/MOHE/ENG/16001-00-D000024. The authors also want to appreciate the effort of Nigerian Meteorological Agency (NIMET) for providing the required data for this research.

References

Aguiar, R., Collares-Pereira, M., Conde, J., 1988. Simple procedure for generating sequences of daily radiation values using a library of Markov transition matrices. Sol. Energy 40, 269–279.
Ajayi, O., Ohijeagbon, O., Nwadialo, C., Olasope, O., 2014. New model to estimate daily global solar radiation over Nigeria. Sustain. Energy Technol. Assess. 5, 28–36.
Akikur, R., Saidur, R., Ping, H., Ullah, K., 2013. Comparative study of stand-alone and hybrid solar energy systems suitable for off-grid rural electrification: a review. Renew. Sustain. Energy Rev. 27, 738–752.
Akpabio, L.E., Etuk, S.E., 2003. Relationship between global solar radiation and sunshine duration for Onne, Nigeria. Turk. J. Phys. 27, 161–167.
Al-Alawi, S., Al-Hinai, H., 1998. An ANN-based approach for predicting global radiation in locations with no direct measurement instrumentation. Renewable Energy 14, 199–204.
Amato, U., Andretta, A., Bartoli, B., Coluzzi, B., Cuomo, V., Fontana, F., Serio, C., 1986. Markov processes and Fourier analysis as a tool to describe and simulate daily solar irradiance. Sol. Energy 37, 179–194.
Angstrom, A., 1924. Solar and terrestrial radiation. Report to the international commission for solar research on actinometric investigations of solar and atmospheric radiation. Quart. J. Roy. Meteorol. Soc. 50, 121–126.
Azoumah, Y., Yamegueu, D., Ginies, P., Coulibaly, Y., Girard, P., 2011. Sustainable electricity generation for rural and peri-urban populations of sub-Saharan Africa: the "flexy-energy" concept. Energy Policy 39, 131–141.
Bajpai, P., Dash, V., 2012. Hybrid renewable energy systems for power generation in stand-alone applications: a review. Renew. Sustain. Energy Rev. 16, 2926–2939.
Bao, Y., Hu, Z., Xiong, T., 2013. A PSO and pattern search based memetic algorithm for SVMs parameters optimization. Neurocomputing 117, 98–106.
Behrang, M., Assareh, E., Ghanbarzadeh, A., Noghrehabadi, A., 2010. The potential of different artificial neural network (ANN) techniques in daily global solar radiation modeling based on meteorological data. Sol. Energy 84, 1468–1480.
Behrang, M., Assareh, E., Noghrehabadi, A., Ghanbarzadeh, A., 2011. New sunshine-based models for predicting global solar radiation using PSO (particle swarm optimization) technique. Energy 36, 3036–3049.
Benghanem, M., Mellit, A., Alamri, S., 2009. ANN-based modelling and estimation of daily global solar radiation data: A case study. Energ. Convers. Manage. 50 (7), 1644–1655.
Besharat, F., Dehghan, A.A., Faghih, A.R., 2013. Empirical models for estimating global solar radiation: a review and case study. Renew. Sustain. Energy Rev. 21, 798–821.
Bhardwaj, S. et al., 2013. Estimation of solar radiation using a combination of Hidden Markov Model and generalized Fuzzy model. Sol. Energy 93 (1), 43–54.
Bristow, K.L., Campbell, G.S., 1984. On the relationship between incoming solar radiation and daily maximum and minimum temperature. Agric. For. Meteorol. 31, 159–166.
Chapelle, O., Vapnik, V., Bousquet, O., Mukherjee, S., 2002. Choosing multiple parameters for support vector machines. Mach. Learn. 46, 131–159.
Chen, D.S., Jain, R.C., 1994. A robust backpropagation learning algorithm for function approximation. Neural Netw. IEEE Trans. 5, 467–479.
Chen, J.L., Li, G.S., 2013. Estimation of monthly average daily solar radiation from measured meteorological data in Yangtze River Basin in China. Int. J. Climatol. 33, 487–498.
Chung, K.-M., Kao, W.-C., Sun, C.-L., Wang, L.-L., Lin, C.-J., 2003. Radius margin bounds for support vector machines with the RBF kernel. Neural Comput. 15, 2643–2681.
Ezekwe, C., Ezeilo, C.C., 1981. Measured solar radiation in a Nigerian environment compared with predicted data. Sol. Energy 26, 181–186.
Friedrichs, F., Igel, C., 2005. Evolutionary tuning of multiple SVM parameters. Neurocomputing 64, 107–117.
Gardner, M., Dorling, S., 1998. Artificial neural networks (the multilayer perceptron) – a review of applications in the atmospheric sciences. Atmos. Environ. 32, 2627–2636.
Halawa, E., GhaffarianHoseini, A., Hin Wa Li, D., 2014. Empirical correlations as a means for estimating monthly average daily global radiation: a critical overview. Renewable Energy 72, 149–153.
Hargreaves, G.H., Samani, Z.A., 1982. Estimating potential evapotranspiration. J. Irrig. Drain. Div. 108, 225–230.
Hasan, M., Mahlia, T., Nur, H., 2012. A review on energy scenario and sustainable energy in Indonesia. Renew. Sustain. Energy Rev. 16, 2316–2328.
Hocaoğlu, F.O., 2011. Stochastic approach for daily solar radiation modeling. Sol. Energy 85, 278–287.
Hsu, C.-W., Chang, C.-C., Lin, C.-J., 2003. A Practical Guide to Support Vector Classification.
Huang, J. et al., 2013. Forecasting solar radiation on an hourly time scale using a Coupled AutoRegressive and Dynamical System (CARDS) model. Sol. Energy 87, 136–149.
Hunt, L., Kuchar, L., Swanton, C., 1998. Estimation of solar radiation for use in crop modelling. Agric. For. Meteorol. 91, 293–300.
Izgi, E., Öztopal, A., Yerli, B., Kaymak, M.K., Şahin, A.D., 2012. Short–mid-term solar power prediction by using artificial neural networks. Sol. Energy 86, 725–733.

Jain, P., Garibaldi, J.M., Hirst, J.D., 2009. Supervised machine learning algorithms for protein structure classification. Comput. Biol. Chem. 33, 216–223.
Jiang, Y., 2009. Computation of monthly mean daily global solar radiation in China using artificial neural networks and comparison with other empirical models. Energy 34, 1276–1283.
Koza, J.R., 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT press.
Layi Fagbenle, R., 1993. Total solar radiation estimates in Nigeria using a maximum-likelihood quadratic fit. Renew. Energy 3, 813–817.
Liu, X., Mei, X., Li, Y., Wang, Q., Jensen, J.R., Zhang, Y., Porter, J.R., 2009. Evaluation of temperature-based global solar radiation models in China. Agric. For. Meteorol. 149, 1433–1446.
Lorena, A.C., De Carvalho, A.C., 2008. Evolutionary tuning of SVM parameter values in multiclass problems. Neurocomputing 71, 3326–3334.
Łukasik, S., Żak, S., 2009. Firefly algorithm for continuous constrained optimization tasks. Computational Collective Intelligence. In: Semantic Web, Social Networks and Multiagent Systems. Springer, pp. 97–106.
McCulloch, J., Wangati, F., 1967. Notes on the use of the Gunn Bellani radiometer. Agric. Meteorol. 4, 63–70.
Mellit, A., Benghanem, M., Kalogirou, S., 2006. An adaptive wavelet-network model for forecasting daily total solar-radiation. Appl. Energy 83, 705–722.
Mohandes, M.A., 2012. Modeling global solar radiation using Particle Swarm Optimization (PSO). Sol. Energy 86, 3137–3145.
NIMET, 2014. Nigerian Meteorological Agency, Oshodi, Lagos State, Nigeria.
Ming, T., de_Richter, R., Liu, W., Caillol, S., 2014. Fighting global warming by climate engineering: is the Earth radiation management and the solar radiation management any option for fighting climate change? Renew. Sustain. Energy Rev. 31, 792–834.
Ornella, L., Tapia, E., 2010. Supervised machine learning and heterotic classification of maize (*Zea mays* L.) using molecular marker data. Comput. Electron. Agric. 74, 250–257.
Pinker, R., Frouin, R., Li, Z., 1995. A review of satellite methods to derive surface shortwave irradiance. Remote Sens. Environ. 51, 108–124.
Rajasekaran, S., Gayathri, S., Lee, T.-L., 2008. Support vector regression methodology for storm surge predictions. Ocean Eng. 35, 1578–1587.
Ramedani, Z., Omid, M., Keyhani, A., Shamshirband, S., Khoshnevisan, B., 2014. Potential of radial basis function based support vector regression for global solar radiation prediction. Renew. Sustain. Energy Rev. 39 (1), 1005–1011.
Salcedo-Sanz, S. et al., 2014. Daily global solar radiation prediction based on a hybrid Coral Reefs Optimization–Extreme Learning Machine approach. Sol. Energy 105, 91–98.
Sambo, A., 1986. Empirical models for the correlation of global solar radiation with meteorological data for northern Nigeria. Solar Wind Technol. 3, 89–93.
Schalkoff, R.J., 1997. Artificial Neural Networks. McGraw-Hill Higher Education.
Shamshirband, S., Petković, D., Saboohi, H., Anuar, N.B., Inayat, I., Akib, S., Ćojbašić, Ž., Nikolić, V., Mat Kiah, M.L., Gani, A., 2014. Wind turbine power coefficient estimation by soft computing methodologies: comparative study. Energy Convers. Manage. 81, 520–526.
Trnka, M., Žalud, Z., Eitzinger, J., Dubrovský, M., 2005. Global solar radiation in Central European lowlands estimated by various empirical formulae. Agric. For. Meteorol. 131, 54–76.
Vapnik, V., 2000. The Nature of Statistical Learning Theory. Springer.
Vapnik, V.N., Vapnik, V.N., 1998. Statistical Learning Theory. Wiley, New York.
Viana, T., Rüther, R., Martins, F., Pereira, E., 2011. Assessing the potential of concentrating solar photovoltaic generation in Brazil with satellite-derived direct normal irradiation. Sol. Energy 85, 486–495.
Wu, K.-P., Wang, S.-D., 2009. Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space. Pattern Recogn. 42, 710–717.
Wu, G., Liu, Y., Wang, T., 2007. Methods and strategy for modeling daily global solar radiation with measured meteorological data – a case study in Nanchang station, China. Energy Convers. Manage. 48, 2447–2452.
Wu, J. et al., 2014. Prediction of solar radiation with genetic approach combing multi-model framework. Renew. Energ. 66 (11), 132–139.
Yang, X.-S., 2010a. Firefly algorithm, stochastic test functions and design optimisation. Int. J. Bio-Inspired Comput. 2, 78–84.
Yang, X.-S., 2010. Firefly algorithm, levy flights and global optimization. In: Research and Development in Intelligent Systems XXVI, Springer, pp. 209–218.
Yang, H., Huang, K., King, I., Lyu, M.R., 2009. Localized support vector regression for time series prediction. Neurocomputing 72, 2659–2669.
Yohanna, J.K., Itodo, I.N., Umogbai, V.I., 2011. A model for determining the global solar radiation for Makurdi, Nigeria. Renew. Energy 36, 1989–1992.