Communication in Physical Sciences, 2023, 9(2):111-124

Equatorial All-Sky Downward Longwave Radiation Modelling

Nsikan Ime Obot*, Busola Olugbon, Ibifubara Humprey and Ridwanulahi Abidemi Akeem

Received: 02 March 2023/Accepted 02 May 2023/Published online: 05 May 2023

Abstract: *Machine learning algorithms, such as random forests (RF), artificial neural networks (ANN), and support vector regression (SVR), are viable modelling tools because they can learn and replicate data patterns. However, linear regression models are relatively easy to implement. Downward longwave radiation (DLR) is rarely measured due to complications of its measuring instrument, notwithstanding the importance of the radiation in the atmosphere and the energy balance of the Earth's surface. Besides linear regression, several machine learning modes, such as SVR and RF, were also used to model daily cloudless and all-sky DLR at Ilorin (8.53° N, 4.57° E), Nigeria. We further sought an appropriate ANN unit for estimating the radiation in this study. Predictors comprised the period, clearness index, air temperature, water vapour pressure, relative humidity, global solar radiation, and solar hour angle. We found that solar hour angle actively predicts all-sky DLR. The most vital variables used for an all-sky DLR linear regression model for this clime are water vapour pressure, relative humidity, and solar hour angle. Machine learning systems generalise better with vast data having well-correlated inputs. The results also reveal that several machine learning algorithms, like SVR with Pearson VII kernel function, can be used for modelling DLR.*

Keywords*: Machine learning, linear regression, all-sky downward longwave radiation, solar hour angle*

Nsikan Ime Obot*
Physics Department, Faculty of Science,
University of Lagos,
Akoka, Lagos, Nigeria.
Email: nobot@unilag.edu.ng
Orcid id: 0000-0003-2289-3711

Busola Olugbon
Physics Department, Faculty of Science,
University of Lagos,
Akoka, Lagos, Nigeria.
Email: bolugbon@unilag.edu.ng
Orcid id: 0000-0002-0773-9508

Ibifubara Humprey
Physics Department, Faculty of Science,
University of Lagos,
Akoka, Lagos, Nigeria.
Email: ihumprey@unilag.edu.ng
Orcid id: 0000-0002-9199-4586

Ridwanulahi Abidemi Akeem
Physics Department, Faculty of Science,
University of Lagos,
Akoka, Lagos, Nigeria.
Email: philosophermann@gmail.com

1. 0 Introduction

In addition to the components of solar shortwave radiation, longwave radiation components also account for the energy balance of the Earth's surface. Without both types of radiation, atmospheric heat energy would be inexistent or extremely low. Because the energetic shortwave radiation balance occurs with sunshine, it vastly controls the daytime atmospheric heat. However, less powerful longwave radiation is available, with or without solar rays from the sun. Consequently, longwave fluxes sustain both day and night atmospheric energies. Several atmospheric phenomena, such as land-sea energy exchange, nocturnal cooling activities like frosts and fog, snow-cover levels, and ambient temperature variation, relate to DLR (Sellers, 1965; Ohmura, 1982; Zhang et al., 2001). The net longwave radiation is the difference between the downward longwave radiation (DLR) and the upward longwave radiation, which is the DLR reflected from the Earth's surface back

to the sky, depending on the ground albedo. While the downwelling longwave radiation is designated positive, the upward component is assigned negative in the evaluation of the net longwave radiation. Due to its thermal nature, DLR is beneficial for soil evapotranspiration, solar collectors, and agricultural and architectural applications.

DLR is rarely measured on the ground because the pyrgeometer, the instrument for measuring DLR, is fragile due to its high sensor sensitivity and requires daytime data correction and regular recalibration (Udo, 2000). Researchers have modelled the radiation using various techniques to ameliorate the paucity of data (Sugita and Brutsaert, 1993; Soares et al., 2004). In Africa, intensive ground measurements of DLR have occurred in only four sites: Tamanrasset (22.78° N, 5.51° E) Algeria, Ilorin (8.53° N, 4.57° E) Nigeria, Namib Desert (23.56° S, 15.04° E) Namibia, and De Aar, Pretoria (30.67° S, 23.99° E), South Africa (Culf and Nash, 1993; Marthews et al., 2012; Obot, 2019). Despite notable studies, none exist yet on using machine learning to model all-sky DLR at Ilorin, Nigeria. In addition to relative humidity, water vapour pressure, and air temperature, the impacts of some parameters, namely global solar radiation and solar zenith angle, have not yet been considered in modelling the radiation at the site (Udo 2003; Obot *et al*., 2008, 2019).

The two main approaches for modelling are the traditional linear regression method and machine learning (ML) procedures. Linear regression is a curve-fitting technique. It is a statistical method for modelling the relationship between a dependent variable and one or more independent variables. The best-fitting straight line through the observed data points comes forth by minimising the sum of the squared differences between the observed values and those predicted. However, ML uses mathematical-based methods, such as statistical mechanics, neurology, and other principles, often inspired by nature to function. Although machine learning can handle problems with well-defined mathematical expressions, its uniqueness lies in its ability to solve tasks without any equation or overt programme processes. Machine learning is a subset of artificial intelligence (AI) that functions based on learning. AI, in general, are machines, such as computers, robots, and other mechanical and electronic units, that perform human tasks, such as walking, talking, reasoning, and calculating; however, they are separable by their techniques. Machine learning is a distinctively experience-based algorithm, whereas not every AI requires experience. Several algorithms, such as artificial neural networks, support vector machines, decision trees, random forests, adaptive neuro-fuzzy inference system, Naive Bayes, logistic regression, k-means, k-nearest-neighbour, etc., fall under the umbrella of machine learning. They perform regression, classification, clustering, filtering, content memory, and dimensionality reduction.

Regression algorithms in machine learning, such as the Alaike criterion and the least median squared method, take diverse forms slightly different from the traditional linear regression format. Sometimes, a regression method in ML eliminates predictors with little or no relationship with the dependent variable(s). The Akaike criterion is a commonly used measure of the goodness of fit for linear regression models. It considers the goodness of fit and the complexity of the model and provides a way to compare models using different numbers of parameters to avoid overfitting the data. However, the least median squared regression fits a linear model to data with high noise levels or outliers. It minimises the median sum of squared residuals contrary to the least square regression, which minimises the sum of squared residuals.

There are three main learning procedures in ML: supervised, unsupervised, and reinforcement. During training, the input-to-output data pair is shown to the system to learn their relationship in supervised learning, and the test for an accurate response is with new and likely less input. Whereas the training data under unsupervised learning

have no output, the system has to learn the inherently plausible description and quantify an object, training in reinforcement learning has little output description during training with a reward-oriented approach. However, another method, the semi-supervised algorithm, deploys little labelled training data while its target combines a vast unlabelled set with the meagre labelled group for optimal generalisation.

Several studies have modelled clear-sky and all-sky DLR using linear regression and a few machine learning algorithms like neural networks, support vector regression, and adaptive neuro-fuzzy inference system (Sugita and Brutsaret, 1993; Crawford and Duchon, 1999; Udo, 2003; Soares *et al*., 2004; Bilbao and De Miguel, 2007; Obot *et al*., 2008, 2019; Krut *et al*., 2010). To our knowledge, no work has estimated DLR with Weka, an open-source ML software with several packages, or established a relationship between solar hour angle and the radiation before now. In addition to modelling all-sky DLR with linear regression, this study also aims to use machine learning schemes to model cloudless and all-sky DLR from inputs such as clearness index, water vapour pressure, global solar radiation, temperature, and relative humidity using the supervised learning technique. Besides, we seek an appropriate artificial neural network structure for estimating the radiation

2. 0 Materials and Methods

2.1 Data and Statistical Indexes

The main data used in this study belong to the University of Ilorin, Ilorin (8.53° N, 4.57° E), the only known station in Nigeria with ground measurements of DLR so far. They were, however, obtained from the Baseline Surface Radiation Network (BSRN) at https://www.pangaea.de/PHP/BSRN_Status.php. Other atmospheric factors like the air temperature, relative humidity, and water vapour pressure, when not available from BSRN, were obtained from the Nigeria Meteorological Agency (NIMET). The data archived by NIMET pertain to the airport approximately 12 km from the University. DLR measurements were from September 1992 to August 1994 and June 1995 to April 1998 (Udo, 2003; Obot et at., 2019). The calibration information for the instruments used at the campus is in Table 1.

Table 1: Calibration information of the pyrgeometers used at the site of this study

Instrument No.	Year of Calibration	Calibration Constant (μV/Wm^{-2})	Place of Calibration
PIR 20468F3	1980	4.21	New Eppley Lab, New York
PIR 20468F3	1988	3.92	New Eppley Lab, New York
PIR 20468F3	1994	3.75	World Radiation Center, Davos
PIR 28898F3	1992	4.00	New Eppley Lab, New York
PIR 32225F3	1998	4.07	New Eppley Lab, New York

The sky is clear under certain conditions, but mainly, the daily clearness index (k_T) should be greater than or equal to 0.60. It is mathematically expressed as;

$$k_T = \frac{H_G}{H_E} \geq 0.60 \quad (1)$$

H_G is the global solar irradiation (J/m^2), and H_E is the extraterrestrial solar irradiation intercepted by a plane parallel to the Earth's surface (J/m^2).

Input parameters for modelling DLR comprised day, month, year, clearness index, water vapour pressure, air temperature, relative humidity, global solar radiation, and solar hour angle. The solar hour angle was calculated from;

$$\omega_s = cos^{-1}(-tan\emptyset sin\delta) \quad (2)$$

where $\emptyset$ and δ represent the latitude and declination in radians.

Error statistics for assessing the estimation capability of the models include correlation coefficient (r), mean absolute error (MAE), and root mean square error ($RMSE$). The correlation coefficient can be expressed as;

$$r = \frac{\sum(DLR_e - \overline{DLR_e})\sum(DLR_m - \overline{DLR_m})}{\sqrt{\sum(DLR_e - \overline{DLR_e})^2 \sum(DLR_m - \overline{DLR_m})^2}} \quad (3)$$

where DLR_e is the estimated radiation, $\overline{DLR_e}$ is the mean estimated radiation, DLR_m is the ground-measured radiation, and $\overline{DLR_m}$ is the mean ground-measured radiation.

$$MAE = \frac{\sum|DLR_e - DLR_m|}{n} \quad (4)$$

The total number of cases is given by n, and

$$RMSE = \sqrt{\frac{\sum(DLR_e - DLR_m)^2}{n}} \quad (5)$$

2.2 Artificial Neural Networks

Civilisation is sustainable from continuous human inventions, and in some cases, nature replications are in place. Aeroplanes and ships, respectively, have shapes like birds and fishes for optimal speed in the air and water. Hearing aid mimics the hearing mechanism in birds (Bar-Cohen, 2000). Artificial neural networks (ANN) exclusively function using cruelly approximated mathematical expressions of the human nervous system. For an artificial neuron, the relationship between the activation function, f and the output, y is given as:

$$y = f\left(\sum_{i=1}^{n} w_i x_i\right) \quad (6)$$

where w_i represents the weights for all inputs, and x_i represents the neuron input values.

The peculiarities of ANN, like in the biological neuron, are that information processing is parallel in contrast to the series procedure format of the Von Neumann machine. Thus, ANN can tackle numerous problems. ANN is also robust and tolerant to error failure because it can still perform its overall function even if some neurons are inactive and can easily handle fuzzy logic computing tasks.

Weka, an open-source ML software, was used for the experiments because of its relative ease of designing ANN. During the design, the hidden layers of the feedforward multilayer perceptron neural network were turned, while other system parameters were at default. In addition to the entire training set, further system assessment was with some percentages of the training data.

2.3 Support Vector Regression

Support vector regression (SVR) is the other machine learning algorithm besides the support vector classifier (SVC) under the support vector machines. Whereas SVR is highly accurate and efficient in solving regression problems, its counterpart is for tackling classification issues. SVR is a supervised learning algorithm for mapping data into a higher dimensional feature space. It uses a suitable kernel function to find the optimal hyperplane that maximises the margin between the hyperplane and the closest data points, known as support vectors. Additionally, SVR handles noisy data and outliers by incorporating a soft margin, which allows some data points to violate the margin. Hence, SVR can effectively deal with high-dimensional data with complex nonlinear relationships. There are several kernel functions, such as linear kernel, polynomial kernel, radial basis function (RBF) kernel, and sigmoid kernel. A kernel function has its own set of hyperparameters that can produce the best performance for a given problem when turned. Unlike other kernel functions, the Pearson VII Universal-based function (PUF) kernel is versatile because it can change from Gaussian to Lorentzian shape and beyond when its parametric properties for curve fitting are varied. Thus, it can replace other kernel functions. The PUF kernel (Üstün et al., 2006) can be expressed as;

$$K(x_i, x_j) = 1\Big/\left[1 + \left(\frac{2\sqrt{\|x_i - x_j\|^2}\sqrt{2^{(1/\omega)}-1}}{\sigma}\right)^2\right]^{\omega} \quad (7)$$

where x_i and x_j are the input vectors, σ and ω are the parametric properties for curve fitting.

2.4 Random Forests

Random Forest (RF) is a machine learning algorithm for classification and regression tasks. It is an ensemble learning method that constructs a forest or multitude of decision trees at training time and outputs the predicted class or average. A forest consists of randomly selected regression-based trees. RF algorithm constructs the multiple decision trees using bootstrapped datasets and

randomly-selected subsets of input variables. At prediction, each tree in the forest independently predicts the class or value of a given input. Then, the class or mean that receives the most votes from all the trees gives the final prediction. This approach helps to reduce overfitting and improve the accuracy and generalisation of the model. Because RF can accurately handle missing values, it is useful when dealing with large datasets, high-dimensional spaces, and complex feature relationships. Moreover, the RF algorithm can also provide information on the relative importance of each feature in predicting the target variable, making it a valuable tool in dimensionality reduction and exploratory data analysis (Breiman, 2001).

The experiments in this work involved evaluating all the modes in the classifier folder in Weka version 3.9.5 under various conditions. Besides the unsuitable machine learning algorithms for numerical data, overfitted or unfitted models were also overlooked since they had a correlation coefficient value of either 100% or 0% during training. Further sieving criteria were that no mode should have a correlation coefficient value lower than 85% and $RMSE$ should be less than or equal to 10.0 W/m^2 during the training phase under clear skies DLR estimation. The software also determined the statistical indexes from its packages.

3.0 Results and Discussion

3.1 Linear Regression Model

During the periods under investigation, 206 days were clear. The retained data for all-sky DLR modelling covers September 1992 – August 1994, while days with missing data for variables, such as water vapour pressure, RH, and temperature, in the archives of NIMET, were eliminated. Table 2 shows the correlation between the independent variable (DLR) and the dependent factors. The trends between the predictors and DLR, under clear skies and all skies, are similar but with different correlation coefficient values. If the clearness index, which sometimes serves as a sieving parameter is ignored, then air temperature and global solar radiation are poor predictors of DLR in this clime. The main variables are water vapour pressure, relative humidity, and solar hour angle. Under the all-sky situation, the clearness index and global solar radiation have negative correlations. Additionally, global solar radiation correlation with DLR is lower than air temperature correlation with the radiation, unlike under clear skies.

Table 2: Correlation coefficient values between DLR and its predictors

Parameters	Cloudless DLR (n = 206)	All-sky DLR (n =721)
Clearness index	0.0983	- 0.0568
Air temperature	0.2458	0.1693
Water vapour pressure	0.7871	0.8057
Relative humidity	0.6765	0.6998
Solar hour angle	0.3939	0.5480
Global solar radiation	0.2854	-0.1081

The linear expression for the daily all-sky DLR was obtained using adequately correlating factors as;

$$3.03e - 0.045RH + 0.727\omega_s + 265.349 \quad (8)$$

where e is water vapour pressure, RH is relative humidity, and ω_s is solar hour angle. The correlation coefficient, mean absolute error, and root mean square error accrued from Equation 8 were 0.8086, 8.4219 (W/m^2), and 12.5557 (W/m^2), respectively. However, those respective statistics stood at 0.7784, 10.3633 (W/m^2), and 15.3133 (W/m^2) when the equation was tested on clear-sky DLR.

3.2 Machine Learning Modes

Regarding the designed ANN, descend order of the number of nodes in three hidden layers gave optimal results (Fig 1). The same ANN was suitable for both clear skies and all-sky

DLR estimations while setting network viewing to false reduced the running time.

The results for modelling clear-sky DLR are in Appendix 1. After sieving with the two earlier-mentioned criteria, 10 out of 27 modes were retained (Table 3). During training, the correlation coefficient, mean absolute error, and root mean square error for the selected modes ranged between approximately 0.987 and 0.888, 1.485 and 6.555 (W/m^2), and 3.565 and 9.458 (W/m^2), respectively (Table 3). Though RF had the best correlation coefficient and root mean square values, support vector regression (SVR) with the Pearson VII universal-based function (PUF) kernel had the least mean absolute error. In the testing section, the ranges for r, MAE, and $RMSE$ were between approximately 69% and 85%, 7.19 W/m^2 and 10.07 W/m^2, and 10.42 W/m^2 and 15.35 W/m^2, respectively. However, RF had the best results in the segment (Table 3).

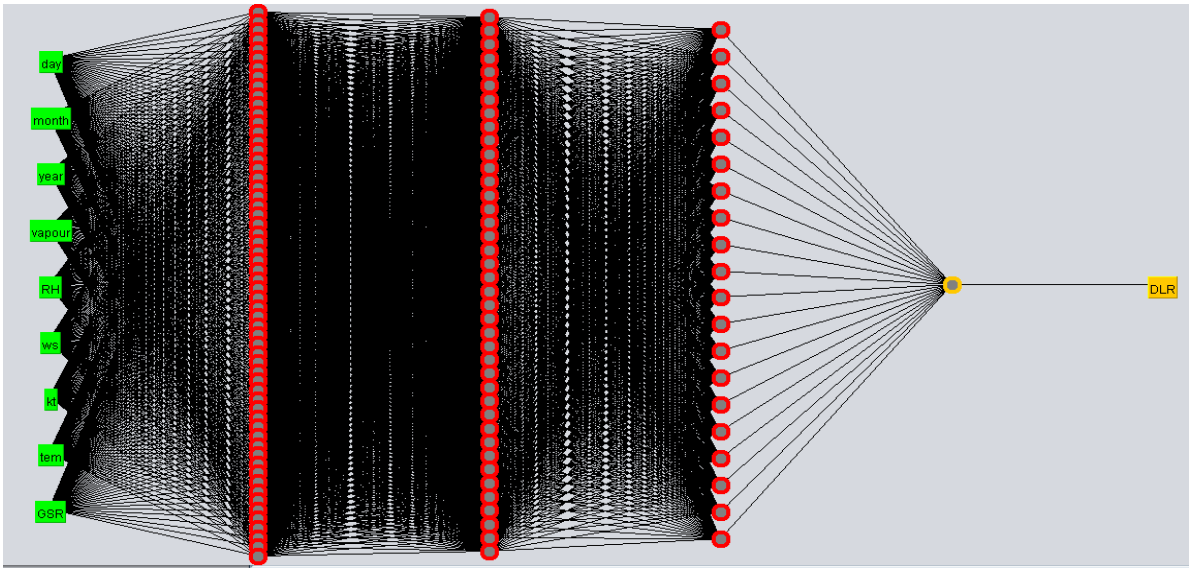

Fig 1: The feedforward multi-perceptron network of three hidden layers of 60, 40, and 20 nodes in the respective layers used for estimating DLR. While DLR was the output, the input comprised: day, month, year, water vapour pressure (vapour), relative humidity (RH), solar hour angle (ω_s), clearness index (kt), temperature (tem), and global solar radiation (GSR).

Under clear skies DLR modelling, Table 3 and Table 4 have nine and six predictors, respectively, while Table 5 has three predictors that highly relate to DLR. Tables 6, 7, and 8 likewise have the results for modelling all-sky DLR with the same nine, six, and three predictors, respectively. Deploying many input variables in machine learning algorithms does not always result in better performance though it could be vital to use a limited number of reliable variables. Whereas there is substantial depreciation in the results as the input variables decrease from nine to six during the training phase, the reserve sometimes is the case at the testing phase for some modes (Tables 3 and 4).

Unlike during training, the testing phase indicates the generalisation capability of models.

The period stamp, which comprises the day, month, and year, does not always affect the model performance positively under cloudless and all-sky conditions. During the training phase under clear skies, r for ML algorithms like ANN and SVR depreciates from approximately 92% and 96% for nine predictors (Table 3) to 86% and 88%, respectively, when six predictors were considered (Table 4).

Table 3: Clear skies estimation of DLR using nine predictors namely day, month, year, water vapour pressure, relative humidity, solar hour angle, clearness index, temperature, and global solar radiation

Model	r	MAE (W/m^2)	$RMSE$ (W/m^2)	r	MAE (W/m^2)	$RMSE$ (W/m^2)
	Training set (n = 206)			Testing set (n = 103)		
Multilayer Perceptron (designed, Fig 1)	0.9209	6.4318	8.5451	0.6945	10.0689	15.3468
SVR (PUF)	0.9558	1.4852	6.0068	0.7880	8.1919	12.2915
Additive Regression	0.8921	6.3449	9.1543	0.7808	8.0839	13.0347
Bragging	0.9025	5.5418	8.7781	0.8308	8.3542	11.4209
Random Sub Space	0.8958	6.5547	9.4582	0.7801	8.9996	12.4892
Regression by Discretisation	0.9763	3.2468	4.356	0.6969	9.8142	15.0395
Decision Table	0.9730	3.0706	4.6449	0.7737	8.7606	12.7093
M5 Rules	0.9325	5.522	7.2693	0.7797	8.7392	12.7032
M5 Tree	0.8875	6.3097	9.2939	0.8092	7.5869	11.6638
Random Forest	0.9874	2.4612	3.5651	0.8527	7.1939	10.4212

However, the reserve is the case under the same conditions at the testing phase, where r appreciates from around 69% and 79% to 81% and 80%, respectively, for ANN and SVR. MAE and $RMSE$ also have similar trends with both machine learning algorithms. Additionally, three to four other algorithms, such as Additive Regression and Bragging, behave likewise though the Additive Regression mode somewhat improved during the training phase, which was exceptional.

In contrast to when predictors reduce from nine to six, the reduction from six to three variables does not improve clear-sky DLR estimations. Whereas every mode had a decreased performance during the training segment when the predictors reduced from nine to six, some algorithms improved as predictors reduced from six to three. Such an exception includes the Random Sub Space, whose r, MAE, and $RMSE$ improved from around 0.85, 8.56 (W/m^2), and 11.81(W/m^2), respectively, to 0.86, 7.20 (W/m^2), and 10.53 (W/m^2) when three instead of six predictors were used (Table 4–5). However, the Decision Table algorithm has no difference between Table 4 and Table 5 in the training phases. M5 Rules and M5 Tree occasionally have indistinguishable characteristics; for instance, they have the same correlation coefficient, mean absolute error, and root mean square error magnitudes during the training phase in Table 4 and the testing phase in Table 5. This attribute extends to the all-sky estimations of DLR (Table 7). Therefore, both algorithms may be considered unreliable in an exercise of this nature.

Though not absolute, the volume of data can inversely or directly impact the modes during the testing phase. As such, instead of 70%, 50% of the training data was retained for testing to avoid misleading results with little data (Appendix 2). Appendix 2 reveals that models have an indirect relationship with data as they perform better with little data volume during testing. However, in most instances, data quantity can also impact machine learning models positively at testing. Comparing Table 3 to Table 6, since both have the same number of predictors, shows that although r, MAE, and $RMSE$ degenerate as the data size increases from 206 (Table 3) to 721 (Table 6) during training, the statistics generally improve across the testing phases.

A similar trend exists while comparing the respective results of Tables 4 and 5 to those of Tables 7 and 8. Nonetheless, a remarkable deviation from that pattern is the Random

Forest. It seems RF overfits when handling estimation problems because despite performing exceptionally during training on every occasion, it rarely sustains such during testing (Tables 3-8).

Table 4: Clear-sky DLR machine learning modelling with six inputs of water vapour pressure, relative humidity, solar hour angle, clearness index, temperature, and global solar radiation using Weka

Model	r	MAE (W/m^2)	$RMSE$ (W/m^2)	r	MAE (W/m^2)	$RMSE$ (W/m^2)
	Training set (n = 206)			Testing set (n = 103)		
Multilayer Perceptron (designed, Fig 1)	0.8551	7.1367	10.7185	0.8128	7.8614	12.0958
SVR(Puk)	0.8810	3.9313	9.5804	0.8047	7.5763	11.8232
Additive Regression	0.8983	6.4072	8.8514	0.7750	8.3463	13.2829
Bragging	0.8988	5.5954	8.9502	0.8408	8.0855	11.0883
Random Sub Space	0.8465	8.5599	11.8117	0.7835	8.8804	12.4925
Regression by Discretisation	0.9515	3.7299	6.1887	0.8043	8.8523	12.0423
Decision Table	0.7969	8.0275	12.1514	0.7632	9.0217	12.8857
M5 Rules	0.8307	7.4478	11.1994	0.7713	8.1934	12.7826
M5 Tree	0.8307	7.4478	11.1994	0.8221	7.4723	11.2949
Random Forest	0.9855	2.4159	3.6956	0.8382	6.9752	10.819

Table 5: Three inputs of water vapour pressure, relative humidity, and solar hour angle usage to model clear-sky DLR

Model	r	MAE (W/m^2)	$RMSE$ (W/m^2)	r	MAE (W/m^2)	$RMSE$ (W/m^2)
	Training set (n = 206)			Testing set (n = 103)		
Multilayer Perceptron (designed, Fig 1)	0.8327	7.7528	11.1377	0.8060	8.4981	12.6553
SVR(PUF)	0.8488	6.0802	10.7642	0.8181	7.3834	11.5019
Additive Regression	0.8807	6.9428	9.5287	0.8019	8.5157	12.2797
Bragging	0.8896	5.9788	9.3010	0.8316	8.0596	11.2266
Random Sub Space	0.8564	7.2025	10.5345	0.7671	9.1384	12.7197
Regression by Discretisation	0.9097	5.1119	8.3534	0.7159	9.2296	14.3809
Decision Table	0.7969	8.0275	12.1514	0.7632	9.0217	12.8857
M5 Rules	0.8144	7.9802	11.6761	0.7954	8.1919	11.9933
M5 Tree	0.8283	7.7525	11.2746	0.7954	8.1919	11.9933
Random Forest	0.9848	2.4953	3.7301	0.8324	7.4413	10.9806

Since ambient temperature is not a significant nor an active predictor of DLR in this vicinity, it may not be included in parametrising the radiation, though it is considered valuable. Since the sky acts like a blackbody, the formats for modelling both clear-sky and all-sky DLR are often in line with the Stefan Boltzmann equation, which incorporates temperature (Markut and Church, 1973; Sugita and Brutsaret, 1993; Crawford and Duchon, 1999; Bilbao and de Miguel, 2007; Krut et al., 2010). The expression for radiated

energy of bodies in the referred Stefan Boltzmann law is;

$R = \varepsilon\sigma T^4$ (9)

where ε is the emissivity of the body, which is equal to 1 for a blackbody but less than 1 for a grey body (like the sky), σ is the Stefan Boltzmann constant, and T is the temperature of the body.

Table 6: All-sky estimation of DLR using nine predictors namely day, month, year, water vapour pressure, relative humidity, solar hour angle, clearness index, temperature, and global solar radiation

Model	r	MAE (W/m^2)	$RMSE$ (W/m^2)	r	MAE (W/m^2)	$RMSE$ (W/m^2)
	Training set (n = 721)			Testing set (n = 306)		
Multilayer Perceptron (designed, Fig 1)	0.8525	7.1375	11.1683	0.8269	8.8831	13.7353
SVR (PUF)	0.9179	2.1576	8.0537	0.8603	6.9599	10.4850
Additive Regression	0.8328	7.0617	11.2195	0.8627	6.9662	10.3489
Bragging	0.8942	5.3577	9.1428	0.8576	7.1638	10.5340
Random Sub Space	0.8739	5.8077	9.9337	0.8575	7.2331	10.5981
Regression by Discretisation	0.8890	5.7418	9.2871	0.8085	8.8755	12.1022
Decision Table	0.8410	6.5758	10.9577	0.8242	7.8521	11.5926
M5 Rules	0.8531	6.1882	10.5693	0.8575	6.6942	10.5866
M5 Tree	0.8529	6.2025	10.5745	0.8579	6.7131	10.5666
Random Forest	0.9825	2.3111	4.0354	0.8413	6.8576	11.1234

Table 7: All-sky DLR machine learning modelling with six inputs of water vapour pressure, relative humidity, solar hour angle, clearness index, temperature, and global solar radiation using Weka

Model	r	MAE (W/m^2)	$RMSE$ (W/m^2)	r	MAE (W/m^2)	$RMSE$ (W/m^2)
	Training set (n = 721)			Testing set (n = 306)		
Multilayer Perceptron (designed, Fig 1)	0.8500	7.3719	11.3578	0.8563	8.7092	12.7549
SVR(Puk)	0.8875	4.2722	9.3471	0.8786	6.1771	9.7862
Additive Regression	0.8411	6.8933	10.9675	0.8627	6.9662	10.3489
Bragging	0.8942	5.3665	9.1319	0.8570	7.1777	10.5611
Random Sub Space	0.8909	5.8134	9.3127	0.8568	7.2416	10.5935
Regression by Discretisation	0.8656	5.9512	10.1528	0.8392	7.9343	11.1349
Decision Table	0.8340	6.9937	11.1752	0.8242	7.8521	11.5926
M5 Rules	0.8297	6.7533	11.3058	0.8400	7.0448	11.2301
M5 Tree	0.8297	6.7533	11.3058	0.8400	7.0448	11.2301
Random Forest	0.9819	2.3341	4.0790	0.8518	7.0342	10.7453

Some studies neglect the sky emissivity while applying linear regression to DLR (Udo, 2003; Abramowitz et al., 2012). Similarly, this work deploys commonly measurable and calculatable atmospheric parameters and ignores the sky emissivity to reduce the otherwise cumbersomeness of determining the clearness index and other relatively tedious processes. Water vapour impacts on DLR depend on its pressure, relative

humidity, and solar hour angle. Therefore, these three parameters effectively account for daily DLR. During the rainy season, the atmosphere is damped and moist and is accompanied by heavy cloudiness, unlike the dry season. The distinctiveness of the two main seasons depends on water vapour pressure, relative humidity, and solar hour angle.

Table 8: All-sky DLR Three inputs of water vapour pressure, relative humidity, and solar hour angle

Model	r	MAE (W/m^2)	$RMSE$ (W/m^2)	r	MAE (W/m^2)	$RMSE$ (W/m^2)
	Training set (n = 721)			Testing set (n = 306)		
Multilayer Perceptron (designed, Fig 1)	0.8187	8.1182	12.1543	0.8415	10.2259	13.8373
SVR (PUF)	0.8529	5.8945	10.5942	0.8631	6.9780	10.3803
Additive Regression	0.8220	7.4402	11.5493	0.8412	7.7042	11.0619
Bragging	0.8796	5.9282	9.6949	0.8429	7.4493	11.0217
Random Sub Space	0.8697	6.1003	10.0610	0.8398	7.5682	11.1237
Regression by Discretisation	0.8378	6.8616	11.0686	0.8394	7.8786	11.1181
Decision Table	0.8527	6.7735	10.5809	0.8242	7.8521	11.5926
M5 Rules	0.8400	6.7648	10.9920	0.6353	8.9645	19.1102
M5 Tree	0.8437	6.7055	10.8778	0.8282	7.7997	11.4766
Random Forest	0.9791	2.5799	4.3492	0.8139	7.6634	12.1531

Data acquisition issues could be complicated while using some models; hence, the methods that generalise better with limited variables are preferable. Although machine learning schemes need sufficient data for learning, oversaturated or undependable predictors, such as day, month, and year, lead to poor generalisation. Since there is not much difference between the six and three predictors of DLR, using only three highly correlating factors of water vapour pressure, RH, and solar hour angle is sufficient for estimating DLR. Four of the ten machine learning algorithms, namely RF, Decision Table, M5 rules, and M5 tree, have one problem or another. The six others, including the designed ANN, perform reasonably well, but SVR is the most commendable unit due to its generally low error values. The same SVR can be very effective in regression and classification problems (López-Martín, 2021; Zhang and Ge, 2013).

4.0 Conclusion

This study uses linear regression and machine learning modes to model clear-sky and all-sky DLR using date and other variables like clearness index, air temperature, relative humidity, water vapour pressure, global solar radiation, and solar hour angle. The ML implementation over an Equatorial African site (Ilorin) was with Weka. The followings are the conclusions;

(i) Besides water vapour pressure and relative humidity, the solar hour angle is superior to ambient temperature when predicting DLR in this region.

(ii) Setting aside the cumbersomeness of excessive variables usage, which could lead to excessive estimation errors, and ignoring the sky emissivity, a linear regression model for daily all-sky DLR was obtained with the three most correlated factors namely, water vapour pressure, relative humidity, and solar hour angle as $3.03e - 0.045RH + 0.727\omega_s + 265.349$.

(iii) Machine learning algorithms generalise better with encompassing data sets and dependable inputs. So, it's unnecessary to deploy several variables that will result in excellent training phase performance but poor generalisation at testing. While the training phase should have reasonably large data, the testing should have little data.

(iv) A suitable artificial neural network for estimating DLR in Weka has 1st, 2nd, and 3rd hidden layers, with the number of nodes of descending order of 60, 40, and 20, respectively.

(v) In some circumstances, improved generalisation in machine learning algorithms is accompanied by poor results during training. Though several machine learning modes can handle DLR, SVR(PUK) is relatively suitable for modelling the radiation.

5.0 Acknowledgements

The authors wish to acknowledge NIMET, Oshodi, Lagos and the scientists at BSRN and the Department of Physics, University of Ilorin for the data used in this study.

6.0 References

Abramowitz, G., Pouyanne, L. & Ajami, H. (2012). On the information content of surface meteorology for downward atmospheric long-wave radiation synthesis. Geophysical Research Letters, 39, L04808. doi.org/10.1029/2011 - GL050726.

Bar-Cohen, Y. (2006). Biomimetics – using nature to inspire human innovation. Bioinspiration & Biometric, 1(1), P1–P12. doi.org/10.1088/1748-3182/1/1/P01.

Bilbao, J. & de Miguel, A. H. (2007). Estimation of daylight downward longwave atmospheric irradiance under clear sky and all-sky conditions. Journal of Applied Meteorology and Climatology, 46, pp. 878–889. doi.org/10.1175/ - JAM2503.1.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), pp. 5–32. doi.org/10.1023/A:1010933404324.

Crawford, T. M. & Duchon, C.E. (1999). An improved parameterization for estimating effective atmospheric emissivity for use in calculating daytime downwelling longwave radiation. Journal of Applied Meteorology, 38, pp. 474–480. doi.10.1175/1520-0450(1999)038<0474:AIPFEE>2.0.CO;2

Culf, A.D. & Nash, J. H. C. (1993). Longwave radiation from clear skies in Niger: a comparison of observations with simple formulas. Journal of Applied Meteorology, 32, pp. 539–547. doi.10.1175/1520-0450(1993)032<0539:LRFCSI>2.0.CO;2

Krut, N. S., Vendrome, I. F., da Rocha, H. R., Chou, S. C., & Cabral, O. (2010). Downward longwave radiation estimates for clear and all-sky conditions in the Sertãozinho region of São Paulo, Brazil. Theoretical and Applied Climatology, 99, pp. 115–123. doi.org/10.1007/s00704-009-0128-7.

López-Martín, C. (2021). Effort prediction for the software project construction phase. Journal of Software: Evolution and Process, 33, e2365. 24. doi.org/10.1002/smr.2365.

Marthews, T. R., Malhi, Y., & Iwata, H. (2012). Calculating downward longwave radiation under clear and cloudy conditions over a tropical lowland forest site: an evaluation of model schemes for hourly data. Theoretical and Applied Climatology, 72, pp. 461–477. doi.org/1 - 0.1007/s00704-011-0486-9.

Maykut, G. A., & Church, P. F. (1973). Radiation climate of Barrow, Alaska, 1962–66. Journal of Applied Meteorology and Climatology, 12, pp. 620–628. doi.org/10.1175/1520-0450(1973)012<0620:RCOBA>2.0.CO;2

Obot, N. I. (2019). Comparison between ground measured and satellite estimates of downward longwave radiation. MAUSAM, 70, pp. 533–540.

Obot, N., Chendo, M., Erusafe, E. & Olutayo O. (2008). Comparison of clear skies downward longwave radiation obtained through equations with that obtained from

artificial neural networks. 10th World Renewable Energy Congress and Exhibition, Scotland, pp. 2023- 2027.

Obot, N. I., Humphrey, I., Chendo, M. A. C., & Udo, S.O. (2019). Deep learning and regression modelling of cloudless downward longwave radiation. Beni-Suef University Journal of Basic and Applied Sciences, 8, 23. doi.org/10.1186/s43088-019-0018-8.

Ohmura, A. (1982). Climate and energy balance of the Arctic tundra. International Journal of Climatology, 2, pp. 65-84. doi.org/10.1002/joc.3370020106.

Sellers, W. D. (1965). Physical Climatology. University of Chicago Press, Chicago.

Soares, J., Oliveira, A. P., Boznar, M. Z., Mlakar, P., Escobedo, J. A. & Machado, A.J. (2004). Modelling hourly diffuse solar-radiation in the city of São Paulo using neural-network technique. Applied Energy, 79, pp. 201–214. doi.org/10.1016/j.apenergy.2003.11.004.

Sugita, M. & Brutsaert, W.H. (1993). Cloud effect in the estimation of instantaneous downward longwave radiation. Water Resources Research, 29, pp. 599–605. doi.org/10.1029/92WR02352.

Udo, S. O. (2000). Quantification of solar heating of the dome of a pyrgeometer for a tropical location: Ilorin, Nigeria. Journal of Atmospheric and Oceanic Technology, 17, pp. 995–1000. doi.org/10.1175/1520-0426(2000)017<0995:QOSHOT>2.0.CO;2.

Udo, S. O. (2003). Modelling of infrared radiation for all skies from easy-to-measure meteorological parameters at tropical location: Ilorin, Nigeria. Turkish Journal of Physics, 27, pp. 61–68.

Üstün, B., Melssen, W. J. & Buydens, L. M. C. (2006). Facilitating the application of Support Vector Regression by using a universal Pearson VII function based kernel. Chemometrics and Intelligent Laboratory Systems, 81, pp. 29–40. doi.org/10.1016/j.chemolab.2005.09.003.

Zhang, G. & Ge, H. (2013). Support vector machine with a Pearson VII function kernel for discriminating halophilic and non-halophilic proteins. Computational Biology and Chemistry, 46, pp. 16–22. doi.org/10.1016/j.compbiolchem.2013.05.001.

Zhang, T., Stamnes, K. & Bowling, S.A. (2001). Impact of the atmospheric thickness on downwelling longwave radiation and snowmelt under clear-sky conditions in the Arctic and Subarctic. Journal of Climate, 14, pp. 920–939. doi.org/10.1175/1520-0442(2001)014<0920:IOTATO>2.0.CO;2.

Consent for publication

Not Applicable

Availability of data and materials

The publisher has the right to make the data public

Competing interests

The authors declared no conflict of interest.

Funding

There is no source of external funding

Authors' contributions:

All authors conceptualised the study. BU and RAA handled data processing. IH, BU, and NIO did literature reach. NIO implemented the modelling and wrote the manuscript.

Appendix 1: The training phase of the preliminary experiment using Weka to model clear-sky DLR using nine input variables consisting of day, month, year, water vapour pressure, relative humidity, solar hour angle, clearness index, air temperature, and global solar radiation

S/no	Model	Weka Classifier Location Folder	r	MAE (W/m^2)	$RMSE$ (W/m^2)
			Training set (n = 206)		
1	Isotonic Regression	Functions	0.8083	7.6355	11.8448
2	Gaussian Progresses	Functions	0.7122	10.3086	14.3798
3	Least Median Squared Regression	Functions	0.8135	6.9548	11.9033
4	Linear Regression (Akaike criterion)	Functions	0.8307	7.4478	11.1994
5	ML Regressor	Functions	0.8570	7.0295	10.3681
6	Multilayer Perceptron (default)	Functions	0.8617	7.101	10.9475
7	Multilayer Perceptron (designed, Fig 1)	Functions	0.9209	6.4318	8.5451
8	Pace Regression	Functions	0.8298	7.3421	11.2259
9	RBF Regressor	Functions	0.8348	7.255	11.0834
10	Simple Linear Regression	Functions	0.7871	8.0735	12.4073
11	SVR (PolyKernel)	Functions	0.8157	6.8758	11.7977
12	SVR (Normalised PolyKernel)	Functions	0.7602	8.4166	13.2465
13	SVR (Puk)	Functions	0.9558	1.4852	6.0068
14	SVR (RBFKernel)	Functions	0.7869	8.3831	12.871
15	KNN (LWL)	Lazy	0.7594	9.2315	13.0896
16	Additive Regression	Meta	0.8921	6.3449	9.1543
17	Bragging	Meta	0.9025	5.5418	8.7781
18	Iterative Absolute Error Regression	Meta	0.7436	9.2728	13.7005
19	Random Sub Space	Meta	0.8958	6.5547	9.4582
20	Regression by Discretisation	Meta	0.9763	3.2468	4.356
21	Decision Table	Rules	0.973	3.0706	4.6449
22	M5 Rules	Rules	0.9325	5.522	7.2693
23	Decision Stump	Trees	0.7436	9.5474	13.4491
24	DPC Tree	Trees	0.8564	7.2885	11.0501
25	M5 Tree	Trees	0.8875	6.3097	9.2939
26	Random Forest	Trees	0.9874	2.4612	3.5651
27	REP Tree	Trees	0.8626	7.0288	10.1761

Appendix 2: The testing phase of the models of Appendix 1

S/No	Model	r	MAE (W/m^2)	$RMSE$ (W/m^2)	r	MAE (W/m^2)	$RMSE$ (W/m^2)
		Testing set (n = 103) 50%			Testing at 70% (n = 62)		
1	Isotonic Regression	0.7827	8.4135	12.4014	0.8139	7.2497	10.5638
2	Gaussian Progresses	0.7053	10.6757	14.3980	0.7599	9.3396	12.0141
3	Least Median Squared Regression	0.7952	7.6632	12.1667	0.8195	7.5021	10.5755
4	Linear Regression (Akaike criterion)	0.7828	8.2464	12.3626	0.8449	7.0764	9.691
5	ML Regressor	0.7885	8.2177	12.2507	0.8205	7.8007	10.4409
6	Multilayer Perceptron (default)	0.7203	11.6586	16.5889	0.7815	9.3983	12.5292
7	Multilayer Perceptron (designed, Fig 1)	0.6945	10.0689	15.3468	0.8750	6.7827	9.1125
8	Pace Regression	0.7965	7.9468	11.9866	0.8277	7.4788	10.1699
9	RBF Regressor	0.7549	8.5018	13.1558	0.7798	8.0548	11.4981
10	Simple Linear Regression	0.7917	7.9332	12.0840	0.8143	7.3353	10.5224
11	SVR (PolyKernel)	0.7910	7.6560	12.3460	0.8206	7.4188	10.459
12	SVR (Normalised PolyKernel)	0.7835	8.5194	12.3236	0.8048	7.8846	10.8621
13	SVR (Puk)	0.7880	8.1919	12.2915	0.8572	6.5981	9.4017
14	SVR (RBFKernel)	0.7557	10.0735	14.5119	0.7930	8.2790	11.7624
15	KNN (LWL)	0.7587	9.3178	12.9231	0.7994	7.8899	10.9203
16	Additive Regression	0.7808	8.0839	13.0347	0.8963	6.2758	8.0696
17	Bragging	0.8308	8.3542	11.4209	0.8862	6.4564	8.4577
18	Iterative Absolute Error Regression	0.7481	9.2429	13.3989	0.7926	8.0016	11.6106
19	Random Sub Space	0.7801	8.9996	12.4892	0.8674	7.042	9.3847
20	Random by Discretisation	0.6969	9.8142	15.0395	0.9140	5.8375	7.6885
21	Decision Table	0.7737	8.7606	12.7093	0.8292	7.556	10.1605
22	M5 Rules	0.7797	8.7392	12.7032	0.8449	7.0764	9.691
23	Decision Stump	0.7481	9.478	13.1659	0.7926	7.9103	11.0618
24	DPC Tree	0.8134	8.0113	11.7218	0.8201	8.3306	10.9165
25	M5 Tree	0.8092	7.5869	11.6638	0.8449	7.0764	9.691
26	Random Forest	0.8527	7.1939	10.4212	0.9379	5.0027	6.4262
27	REP Tree	0.7840	8.5277	12.4496	0.7381	8.7607	12.2648