Communication in Physical Sciences, 2024, 11(2): 233-247

Modelling Nonseasonal Daily Clearness Index for Solar Energy Estimation in Ilorin, Nigeria Using Support Vector Regression

Nsikan Ime Obot*, Okwisilieze Uwadoka, and Oluwasegun Israel Ayayi

Received: 23 August 2023/Accepted: 17 April 2024 /Published: 24 April 2024

Abstract: *Solar radiation is the primary energy source for the planet and is crucial for energy generation in technologies such as photovoltaic systems and solar thermal food dryers. However, accurately quantifying solar radiation poses challenges due to its variability and the lack of appropriate instrumentation, among other factors. To address this, support vector regression (SVR), a machine learning (ML) algorithm, was employed using various kernel functions such as linear, radial basis function (RBF), and sigmoid, with hyperparameter tuning. This approach aimed to estimate the daily clearness index (K_T), which is a key metric for estimating global solar radiation at Ilorin (8° 32′ N, 4° 34′ E), Nigeria. The SVR models were developed and assessed by considering statistical measures such as the correlation coefficient and mean absolute error. The input parameters used in the model included sunshine hours, maximum temperature, minimum temperature, and the ratio of both temperatures. The correlations between K_T and its estimators, and between its actual and calculated values were all below 70%. SVR-RBF outperformed the others, including the traditional regression model, under the statistical assessment measures, both with the training dataset and the testing dataset. Although the regression model obtained under the same conditions surpassed the other kernel functions in some areas and is highly competitive, SVR-RBF is recommended for the estimation of the daily clearness index in this vicinity.*

Keywords: *Clearness index, support vector regression, global solar radiation, radial basic function, kernel function*

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

Okwisilieze Uwadoka
Mechanical Engineering Department, Faculty of Engineering, University of Lagos,
Akoka, Lagos, Nigeria.
Email: okwy.uwadoka@gmail.com

Oluwasegun Israel Ayayi
Physics Department, Faculty of Science,
University of Lagos, Akoka, Lagos, Nigeria.
Email: aoluwasegun90@yahoo.com

1.0 Introduction

Solar radiation is indispensable for powering various solar electronic devices like photovoltaic (PV) cells and solar thermal technologies, besides facilitating plant photosynthesis and producing body heat in living organisms. Being the primary energy source for Earth, solar radiation energises systems such as air circulation and water movements, playing a crucial role in the Earth's climate dynamics. Accurate measurement of solar radiation is essential for effectively harnessing solar energy, as devices rely on the power they receive, with PV cells operating based on incident photon energy. Direct measurement methods using instruments like pyranometers and solarimeters have limitations such as space coverage, location specificity, accessibility, operational expertise required, and relative cost. Additionally, while measuring instruments are relatively scarce, another recognisable data acquisition technique, the satellite-derived method combines nearby locations into square boxes, leading to data inaccuracies where affected locations may have identical values, though this may not reflect the actual conditions. Besides the pyranometers, ground-based weather stations may use specialised instruments to directly measure global solar radiation (Al-Waeliet al., 2021; Paulescu et al., 2013; Udo and Aro, 1999; Grojean et al., 1980).

The space between the sun and Earth comprises various layers, each containing different types of particles and fields. For example, the troposphere, closest to Earth, contains gases such as nitrogen, oxygen, and water vapour, while the ionosphere, primarily in the thermosphere but extending slightly into the mesosphere and exosphere, contains electron jets. Consequently, electromagnetic radiation from the sun interacts with numerous constituents like neutral and ionised gases, particles, and fields before reaching Earth. Solar radiation interacts with matter through transmission, absorption, emission, and scattering processes. Furthermore, solar activities, such as variations in sunspot numbers, and Earth's revolution around the sun, regulate the intensity of solar radiation reaching Earth (Singh, 2024; Eltbaakh et al., 2012).

In Nigeria, as in other developing countries, data on solar radiation are relatively scarce, crucial for accurately depicting the climate of a given site and for various applications such as sizing photovoltaic systems, food processing, and water purification (Obot et al., 2022; Okoye et al., 2016). Challenges in obtaining this data are often attributed to outdated instruments and insufficient funding. Ilorin, Nigeria, is one of the few sites in the country with long-term measurements of global solar radiation. For instance, a pyranometer was once connected to a data logger at a university campus in Ilorin to monitor solar radiation (Udo and Aro, 1999; Udo, 2000). Unfortunately, this setup is no longer operational due to funding constraints.

As an alternative to direct measurement methods, models are used to estimate solar radiation data. However, these mathematical models must exhibit accuracy when applied to different locations, considering their reliance on data from specific sites. Using the clearness index to predict global solar radiation offers advantages because it can directly or indirectly incorporate other factors, such as extraterrestrial solar radiation and diffuse solar radiation. Additionally, it provides insights into atmospheric conditions, particularly cloud status (Udo, 2000; Babatunde and Aro, 1995; Hinrichsen, 1994).

The concept of the clearness index is elegantly expressed through the pioneering work of Angstrom (1924) in the form of the equation:

$$\frac{H}{H_o} = a(S_r) + b \qquad (1)$$

where H represents total solar irradiation, H_o is the extraterrestrial irradiation at the top of the atmosphere, a and b are empirical constants, and S_r denotes sunshine hours, a ratio of bright sunshine hours (s) to total potential below as: sunshine hours (S_o).

Thus Equation 1 when rewritten in the form known as the Angstrom – Prescott model is given as:

$$\frac{H}{H_o} = a\left(\frac{s}{S_o}\right) + b \qquad (2)$$

The total potential sunshine hours S_o is given as equation 3

$$S_o = \frac{2}{15} w_s \qquad (3)$$

where w_s is the solar hour angle.

Furthermore, the solar hour angle, which depends on the location's latitude ϕ and the solar declination angle δ, is given as:

$$w_s = \cos^{-1}(- tan\phi \; tan\delta) \qquad (4)$$

Depending on the day, the solar declination angle can be calculated as:

$$\delta = 23.45 \, sin\left[\frac{(284+n)360}{365}\right] \qquad (5)$$

Here, n represents the Julian day number.

The average daily extraterrestrial radiation at the horizontal surface can be evaluated as:

$$H_o = \frac{24}{\pi} I_{sc}\left[1 + 0.033\ \cos\frac{360n}{365}\right]\left[\sin w_s \cos\phi \cos\delta + \frac{\pi}{180} w_s \sin\phi \sin\delta\right] \qquad (6)$$

The daily clearness index symbolised as K_T, represents the ratio of daily total solar irradiation to the daily extraterrestrial irradiation (at the horizontal surface). This allows for the computation of parameters such as cloudiness index and diffuse solar irradiation, which are associated with K_T. This method offers a practical approach to estimating diffuse solar radiation in situations where direct measurements are either unavailable or restricted (Udo, 2000; Babatunde and Aro, 1995; Hinrichsen, 1994).

In the same vein as Angstrom (1924) and in attempts to achieve more accurate results considering the bias of solar radiation to specific regions, topography, and seasons globally, several other regression-based models have been proposed by researchers. These models seek appropriate empirical constants and incorporate other meteorological variables such as various forms of air temperatures, relative humidity, atmospheric pressure, cloud factor, precipitation, evaporation, and wind speed, among others. Due to technical know-how, sensor calibration, and cost-related instrument issues, global solar radiation data are unavailable at various stations, often necessitating the extension of results obtained elsewhere to regions with similar climates without measurements (Besharat et al., 2013; Mohanty et al., 2016; Chukwujindu, 2017; Nwokolo and Ogbulezie, 2018).

Machine learning (ML) schemes, a subset of artificial intelligence, often provide superior and faster solutions to problems compared to traditional statistical or regression methods, depending on the specific issue at hand. ML utilises methods inspired by nature or their replicas, such as neural networks that mimic the human nervous system. ML models come in diverse forms, including standalone algorithms, as well as hybrids that combine these approaches with metaheuristics and traditional mathematical models. Soft computing methods like artificial neural networks (ANN), k-nearest neighbour (KNN), adaptive neuro-fuzzy inference system (ANFIS), support vector regression (SVR), and hybrids of intelligent-intelligent and intelligent-traditional systems have been successfully deployed in estimating global solar radiation. Specifically, SVR is comparatively flexible, reliable, fast, accurate, and easy to use, making it widely applied in modelling global solar radiation and other related parameters like wind energy (Obot et al., 2023; Martins and Giesbrecht, 2021; Zendehboudi et al., 2018; Belaid and Mellit, 2016; Ramli et al., 2015; Fonseca Jr. et al., 2011).

Studies using SVR to model global solar radiation in Nigeria are relatively rare. For instance, Ayodele et al. (2019) combined k-means with SVR to estimate the radiation in Ibadan with data from 2010 – 2015 as the training set, while data from 2016 – 2017 served as the testing set. They found the hybrid better than the Anstrom-Prescott, autoregressive moving average, and ANN models. Moreover, Olatomiwa et al. (2015) hybridised the firefly algorithm with the SVR system (ff-SVR) and compared it with ANN and genetic algorithm (GA) in predicting solar radiation for three Nigerian cities, namely Maiduguri, Iseyin, and Jos. Likewise, the study also established that ff-SVR is superior to ANN and GA.

While both traditional methods and machine learning schemes, such as neural networks incorporating parameters like sunshine hours, temperature, and relative humidity, have been used to model global solar radiation in Nigeria, specific machine learning models like SVR have not yet been applied to this particular radiation in Ilorin, Nigeria, to our knowledge. This study aims to model the daily clearness index using SVR with various kernel functions at the selected location in Nigeria. There will be a comparison of results with the traditional regression method to determine the extent of support vector regression superiority, if any.

2. 0 Methodology

2.1. Site and Data

Ilorin (8° 32′ N, 4° 34′ E), Nigeria was chosen as the case study due to the availability and quality of data. Situated in the North Central geopolitical zone of Nigeria, Ilorin experiences three seasons, namely the dry season, the Harmattan season, and the rainy season. Solar radiation data from the University of Ilorin, obtained through the Baseline Surface Radiation Network in collaboration with PANGAEA, were downloaded from https://dataportals.pangaea.de/bsrn/.

Additionally, measurements such as sunshine hours, and maximum and minimum temperatures were sourced from the Nigerian Meteorological Agency (NIMET) in Oshodi, Lagos. The NIMET records were collected at the city airport, approximately 12 km from the university campus. Specific days with incomplete data, particularly regarding global solar radiation, were excluded from the analysis. Further data pruning was done based on quality checks spanning the period from 1992 to 2006.

2.2 Support Vector Regression

Support vector regression (SVR) is one of the two types of support vector machines (SVM), the other being support vector classification (SVC), which is a binary classifier. Both SVR and SVC originate from the support vector theory of nonlinear statistical learning algorithms (Basak et al., 2007; Smola and Schölkopf, 2004; Vapnik and Chervonenkis, 1964). The fundamental principle of SVM is to solve classification and regression problems through convex optimisation, aiming to minimise the separation of data points in a hyperplane. This is achieved by using a loss function to measure the distance of data cases from a reference plane. The core principles of SVM apply to both SVC and SVR. However, SVR employs a specific loss function with a distance measure that focuses on determining the sparseness of the support vectors. While SVR primarily addresses estimation and prediction tasks using a simpler and more efficient linear-functions algorithm for optimisation during the training stage, SVC utilises quadratic-expressions programming in the same phase to manage classification and identification problems.

SVM is an intelligent system that employs a generalisation mapping approach akin to ANN. However, the distinction lies in their optimisation objectives during training. While ANN seeks to minimise the training error, SVM maximises the feature space boundary to minimise the outermost bound. Consequently, depending on the features, SVM may optimise more aggressively than ANN and other ML algorithms. Thus, SVM stands out as a potent tool in machine learning, adept at optimising the separation of data points in a hyperplane, rendering it suitable for classification and regression tasks with unique optimisation strategies.

Suppose the data combination of targets and inputs is represented as $(y_i,\ x_i)$, where $i = 1,2,3,\ldots,N,\ x_i \in \{-N, N\},\ y \in \Re^D$. Here, y_i denotes the output values and x_i the input values in the training section. Linear regression aims to find the function f for the best fit, $f\colon \Re^N \to \Re$. The hyperplane for the data separation is represented as $f(x) = w^T x + b$, where w is the coefficient vector of the D-dimensional space, x is the input vector, and b is the bias at the origin. Generally, the loss function of the constructed hyperplane seeks to minimise error deviation between the two data sets of input and target. Regularising the solution involves imposing a penalty on large separations from the reference plane by minimising the Euclidean norm, $||w||$ using the convex optimisation technique as follows: minimise $\frac{1}{2}||w||^2$ subject to

$$\begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon \end{cases} \tag{7}$$

ε is the precision level.

Operationally, no penalty is imposed if the data cases are within the tolerance region regarded as ε-tube. Errors arising from mismatched inputs and targets in the feature space and non-optimal learning procedures in the optimisation are accommodated by the slack variables ξ_i and ξ_i^* respectively. So, the new problem is formulated as:

minimise $\frac{1}{2}\|w\|^2 + C\sum(\xi_i + \xi_i^*)$ subject to

$$\begin{cases} y_i - \langle w, x_i \rangle - b \ \leq \ \xi_i \ + \xi_i^* \\ \langle w, x_i \rangle + b \ - y_i \ \leq \ \xi_i + \ \xi_i^* \\ \xi_i, \ \xi_i^* \qquad\qquad \geq \ 0 \end{cases} \tag{8}$$

Here, C is the regularisation parameter.

The training error in the modelling can be characterised using the ε-insensitive loss function, which minimises the problem by maximising the two separation margins on either side of the reference line of the dimensional feature space. It takes the form of:

$$|\xi|_\varepsilon := \begin{cases} 0 & \text{if } |\xi| < \varepsilon \\ |\xi| - \varepsilon & \text{otherwise} \end{cases} \tag{9}$$

For robustness, tiny errors are assumed to have no impact, and changes in the input-output paired positions do not affect the overall outcome. However, the solution to the optimisation problem is relatively complicated. One technique is using linear regression, where convex optimisation is handled via the Lagrange function. This function uses dual variables, n_i and n_i^* along with its peculiar multipliers, α_i^* and α_i represented as:

$$L_f \ := \frac{1}{2}\|w\|^2 + C\sum(\xi_i + \xi_i^*) + \ L_{3rd} + L_{4th} + \ L_{5th} \tag{10}$$

where

$$L_{3rd} = \ \sum(n_i\xi_i - \ n_i^*\xi_i^*)$$

$$L_{4th} = \ \sum\alpha_i\left(\frac{1}{2}\|w\|^2\langle w, x_i \rangle + b - \ y_i - \varepsilon - \xi_i\right)$$

$$L_{5th} = \ \sum\alpha_i^*\left(y_i - \frac{1}{2}\|w\|^2\langle w, x_i \rangle - b - \varepsilon - \xi_i^*\right)$$

where $\alpha_i, \alpha_i^*, n_i, n_i^* \geq 0$ for every instance.

For optimality, the partial derivative of the Lagrange function with respect to the primal variables w, b, ξ_i, and ξ_i^* vanishes as dictated by the saddle point condition:

$$\frac{\partial L_f}{\partial w} = 0, \ w = \ \sum x_i(\alpha_i^* - \alpha_i) \tag{11}$$

$$\frac{\partial L_f}{\partial \xi_i} = 0, \ n_i \ = C - \alpha_i \tag{12}$$

$$\frac{\partial L_f}{\partial \xi_i^*} = 0, \ n_i^* = C - \alpha_i^* \tag{13}$$

$$\frac{\partial L_f}{\partial n_i} = 0, \ \sum\xi_i = 0 \tag{14}$$

$$\frac{\partial L_f}{\partial n_i^*} = 0, \ \sum\xi_i^* = 0 \tag{15}$$

and

$$\frac{\partial L_f}{\partial b} = 0, \ \sum(\alpha_i^* - \alpha_i) = 0 \tag{16}$$

The Karush-Kuhn-Tucker (KKT) conditions require that for optimal prime variables w and b, the product of the Lagrange multipliers and the constants equals zero (i.e., $\alpha_i\alpha_i^* = 0$). This satisfies the constraints, transforming the constrained problem to an unconstraint one when plucking the expressions for w, n_i, and n_i^* into the Lagrange function. This leads to a different feature space given as $f(x) = \sum(\alpha_i - \alpha_i^*)\,(x_i.x) + b$.

The expansion is resolved as follows:

$$\text{maximise } -\frac{1}{2}\sum(\alpha_i - \alpha_i^*)\left(\alpha_j - \alpha_j^*\right)\left(x_i.x_j\right) - \varepsilon\sum(\alpha_i + \alpha_i^*) + \sum y_i(\alpha_i - \alpha_i^*)$$

$$\text{subject to } \sum(\alpha_i - \alpha_i^*) = 0 \text{ and } \alpha_i\alpha_i^* \ \in [0 \ C] \tag{17}$$

The relationships between the regularisation parameter and the slack variables are $\xi_i = \alpha_i/C$ and $\xi_i^* = \ \alpha_i^*/C$

Again, applying the Lagrange and KKT processes to the new optimisation problem yields a solution in the form of:

$$\langle w, x_i \rangle + b - \ y_i + \varepsilon + \ \xi_i = 0$$

$$y_i - \langle w, x_i \rangle - b + \varepsilon + \xi_i^* = 0$$
$$\xi_i \xi_i^* = 0,, \ \alpha_i \alpha_i^* = 0 \qquad (18)$$

At a certain stage, the solutions for b yield non-vanishing coefficients known as the support vectors, and the inverse of the square root of sum α_i for the margin. Furthermore, nonlinear mapping to a higher dimensional space in SVR can be achieved with a kernel function. Instead of explicitly solving the mapping, the trick is to introduce the kernel function and pre-process the training data in the feature space. A kernel function is represented as $K(x_i, x_j)$, whereas the reconsidered hyperplane function becomes; $f(x) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(x_i, x_j) + b$. Now, the optimisation takes the form of:

maximise $\frac{1}{2} \sum (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(x_i, x_j) + \sum y_i (\alpha_i - \alpha_i^*) - \varepsilon \sum (\alpha_i - \alpha_i^*)$ subject to $\sum (\alpha_i - \alpha_i^*) = 0,$ $\alpha_i \alpha_i^* = 0$ (19)

The procedure for applying the Lagrange function and ensuring that the KKT conditions are met can be employed to solve the optimisation problem involving the kernel and obtain the solution for SVR. In addition to the linear regression loss function, other SVR solution processes include quadratic loss, Huber loss, and nonlinear regression loss functions. Furthermore, there are various kernel functions available, including the linear, sigmoid, and radial basic functions (RBF), among others. The SVR algorithm considers factors such as the risk factor and model density of the solution, maximum likelihood for the cost function, and conditions like convergence, expansibility, and integrability that the kernel function must satisfy. Nevertheless, the kernel function can be customised for absolute use in SVM. For this specific study, the kernel functions of interest are linear, RBF, and sigmoid. These functions are expressed as follows;

The linear kernel function:

$$K(x_i, x_j) = x_i^T \cdot x_j \qquad (20)$$

where x_i, x_j are the input pairs or feature vectors, and x_i^T denotes the transposition of x_i.

The RBF kernel function:

$$K(x_i, x_j) = e^{\left(-\gamma \|x_i - x_j\|^2\right)} \qquad (21)$$

where γ is a constant or hyperparameter that controls the local decision boundary for each training sample, and the double bar bracket indicates the Euclidean distance between the input feature vectors.

And, the sigmoid kernel function:

$$K(x_i, x_j) = \tanh\left(\alpha x_i^T \cdot x_j + C\right) \qquad (22)$$

where α is a constant that controls the slope of the sigmoid kernel function.

The estimation ofthe daily clearness index, using support vector regression, was carried out using Python, an open-source computing environment. Relevant libraries such as sci-kit-learn and Pandas were utilised for this purpose. Initially, the skewness of estimators was identified, with a specific focus on the ratio of maximum temperature (T_{max}) to minimum temperature (T_{min}), denoted as T_r. To address this, power transformation was applied to reshape the values within the range of -1 to 1, followed by standard scaling to revert to the original data range.

A portion of the data (28%) was set aside for testing, while the remaining data was used for training the SVR models. Cross-validation was performed between 2 and 10 folds based on obtained optimal values of the regularisation parameters (hyperparameters) C and γ (gamma). The search for these optimal values was based on error terms such as correlation coefficient, mean bias error, and mean absolute error. The best-performing model was determined by presetting the mentioned hyperparameters to values ranging from 0.001 to 100 and assessing their performance metrics.

The correlation coefficient (r) can be calculated using the formula shown as equation 23 below

$$r = \frac{\sum_{i=1}^{N} (V_{meas,i} - \bar{V}_{meas,i})(V_{pred,i} - \bar{V}_{pred,i})}{\sqrt{\sum_{i=1}^{N} (V_{meas,i} - \bar{V}_{meas,i})^2 \sum_{i=N}^{N} (V_{pred,i} - \bar{V}_{pred,i})^2}} \qquad (23)$$

where $V_{meas,i}$ represents K_T values obtained from the ground measurements of global solar radiation, $V_{pred,i}$ represents K_T values obtained from SVR models, the bar indicates the averages of the affected parameters, and N is the total number of cases.

The mean absolute error (MAE) and the root mean square error ($RMSE$) are given respectively as:

$$MAE = \frac{\sum_{i=1}^{N}\left|V_{pred,i} - V_{meas,i}\right|}{N} \qquad (24)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(V_{pred,i} - V_{meas,i}\right)^2}{N}} \qquad (25)$$

3. 0 Results and Discussion

3.1 Data Distribution

Due to the huge number of missing bright sunshine hours in the records obtained from NIMET, coupled with the quality check and the periods where there were no measurements at the university campus, only 979 days were retained between 1992 and 2006. Table 1 gives an overview of the distribution of several descriptive statistics such as maximum, minimum, mean, standard deviation, range, median, standard error, and mode. The values for clearness index, sunshine hours, maximum temperature, minimum temperature, and the ratio of maximum temperature to minimum temperature range between 0.73 and 0.20, 0.99 and 0.01, 39 and 20, 27 and 12, and 2.75 and 1.04. respectively. The average values together with the standard deviation are 0.47 ± 0.1 for K_T, 32.17 ± 3.0 °C for T_{max}, 0.54 ± 0.23 for S, 21.51 ± 2.19 °C for T_{min}, and 1.51 ± 0.22 for T_r. Furthermore, since standard deviation indicates the degree of dispersion around the mean, it implies that T_{max}, with the highest value of 3.0, has the widest spread compared to K_T, which has the lowest spread of 0.1. While the middle values (median) stand at 0.48 for clearness index, 32 °C for maximum temperature, 0.58 for sunshine hours, 22 °C for minimum temperature, and 1.48 for the ratio of max temperature to min temperature, the most frequently occurring values, which is the mode, stand at 0.5, 31 °C, 0.71, 22 °C, and 1.52, for the respective variables.

Among these variables, maximum temperature has the most influence on the clearness index in this region (Fig 1). Additionally, the range of the correlation coefficient between K_T and the estimators is from 6.8% to 47%, which is relatively poor. Concerning themselves, only two instances indicate negative correlations among the lot, which occur in the relationship between S_r and T_{min} (-7.1%), and between T_r and T_{min} (-73%). Although not explicitly shown, upon reviewing the values, it was observed that instances where the maximum temperature remained relatively high while the clearness index was comparatively low, were associated with reasonably low values of sunshine hours.

Fig 2 confirms the necessity of transposing the original data in the adopted ML algorithm, particularly regarding T_r, where most values align to the left-hand side with a long tail of the trend line towards the right-hand side. Whereas T_{max} is centrally distributed, both K_T and S_r are skewed to the left due to the prevalence of cloudy days over clear days at the site (Udo, 2000). It appears that the mathematical division of centrally distributed data (i.e., T_{max}) by another dataset that is negatively skewed or primarily distributed to the right-hand side (i.e., T_{min}) leads to a resulting dataset that is positively skewed or predominantly distributed to the left-hand side (i.e., T_r).

3.2 SVR Models Estimations

Although a total of 979 cases were initially randomly split into a ratio of 72:28 for the training and testing groups respectively, eventually only 972 cases were retained. The training set consisted of 777 cases, while the testing set comprised 195 cases, indicating that 6 cases with identical variables were automatically eliminated during the split. In this study, the polynomial kernel function, though available alongside other kernel functions in Python, was not considered due to time constraints that prevented the tuning of hyperparameters on the computer. Consequently, the eventually retained values of

C and gamma were deemed optimal (Fig 3). Despite similar outputs observed within the considered cross-validation folds of 2 and 10, the 4th fold was retained, albeit yielding the same results as those from the 5th to 8th fold. In addition to hyperparameters,

Table 1: comprehensive descriptive statistics for the variables used in this study.

Statistic	K_T	T_{max} (℃)	S_r	T_{min} (℃)	T_r
Maximum	0.73	39	0.99	27	2.75
Mean	0.4660	32.1665	0.5396	21.5077	1.5118
Median	0.48	32	0.58	22	1.46
Minimum	0.20	24	0.01	12	1.04
Mode	0.5	31	0.71	22	1.52
Range	0.53	15	0.98	15	1.71
Standard deviation	0.1006	3.0007	0.2254	2.1928	0.2201
Standard error	0.0032	0.0959	0.0072	0.0701	0.0070

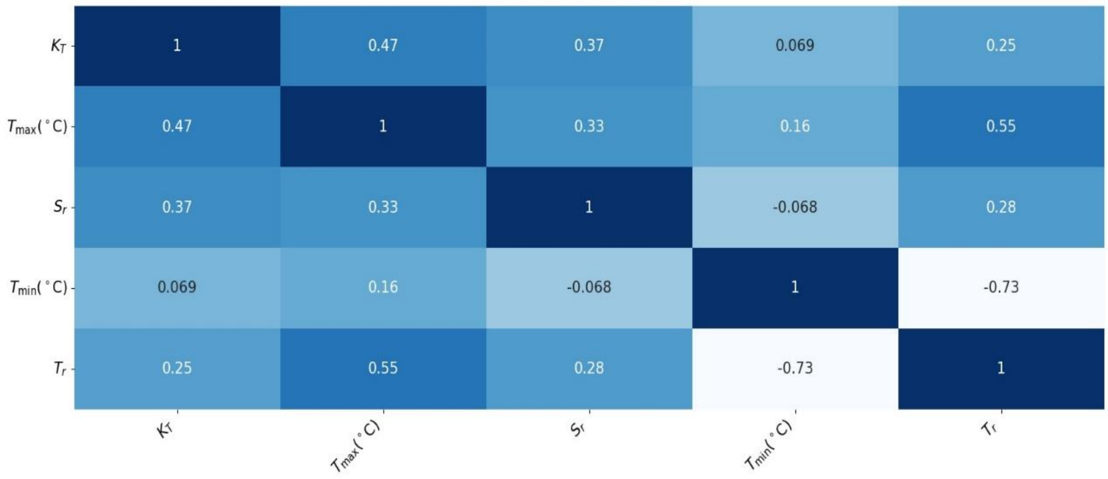

Fig Fig. 1: correlation coefficient between the variables considered in this study.

Table 2 presents the statistical measures' results for each kernel function. During the training phase, the MAE values were 0.0673 for the linear kernel function, 0.0631 for the RBF kernel function, and 0.0673 for the sigmoid kernel function, with testing phase errors standing at 0.0658, 0.0661, and 0.0658, respectively. The r values were approximately 54%, 62%, and 53%, respectively for the linear kernel function, RBF kernel function, and sigmoid kernel function during the training phase, though they reduced respectively to 47%, 50%, and 47% during the testing phase. However, comparing the training data set to the testing data set, $RMSE$ values increased marginally from 0.0854 to 0.0858 for the linear kernel function, from 0.0793 to 0.0838 for the RBF kernel function, and from 0.0854 to 0.0857 for the sigmoid kernel function. From the above, the best-performing kernel function of SVR in the estimation of the nonseasonal daily clearness index in Ilorin, Nigeria, is the RBF. This kernel function has the most desirable outcomes from all the statistical measures during both the training and testing phases, except under MAE in the testing phase. While low MAE and $RMSE$ values are desirable, high r values are preferable.

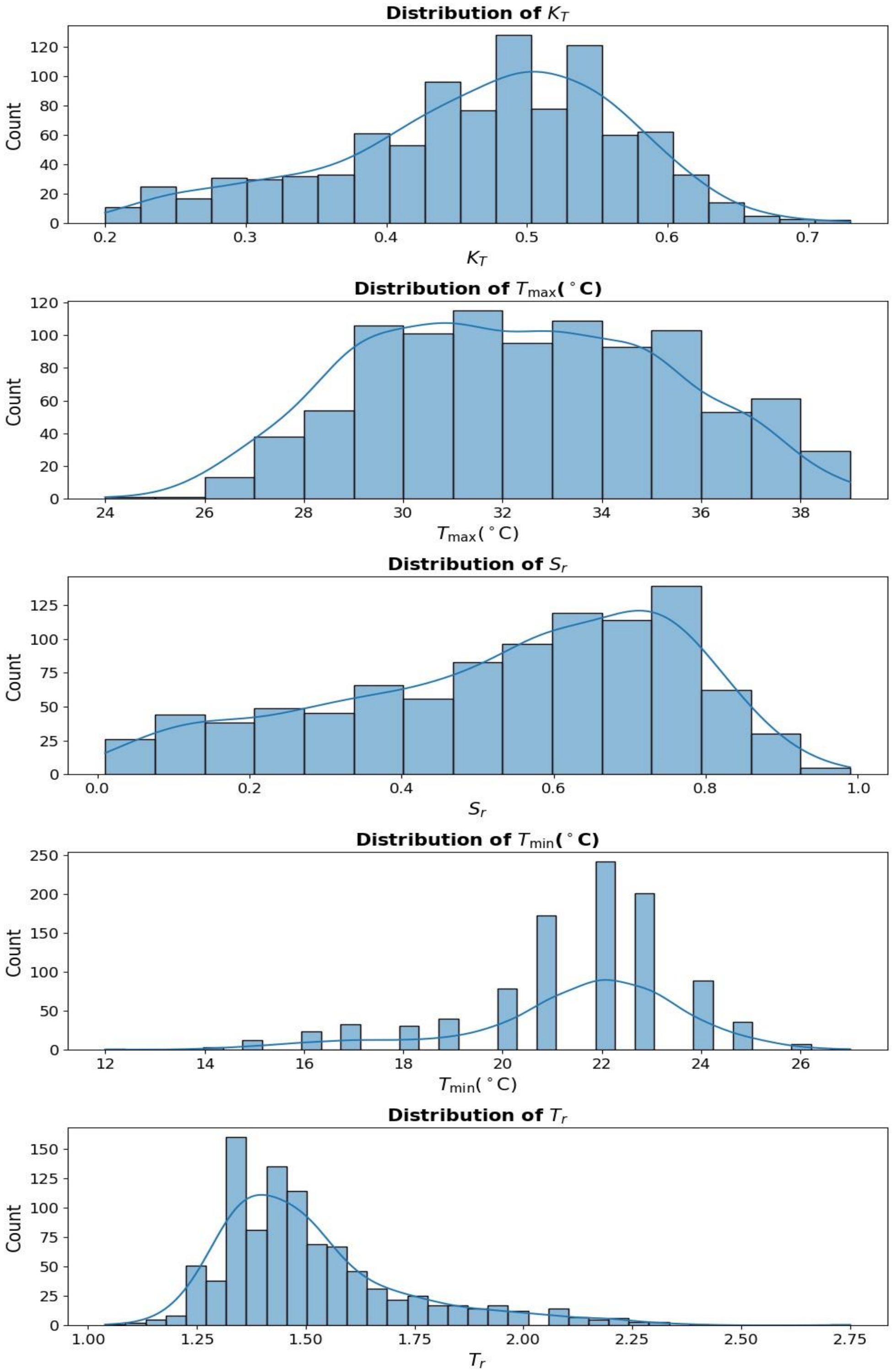

Fig 2: The distribution of the variables in this study

3.3 Traditional Regression Models

Based on daily clearness index values, and concerning two commonly used variables in this region, namely sunshine hours and maximum temperature, the regression models for Ilorin were obtained as:

$$K_T = 0.167\,(S_r) + 0.376 \qquad (26)$$

$$K_T = 0.016(T_{max}) - 0.044 \qquad (27)$$

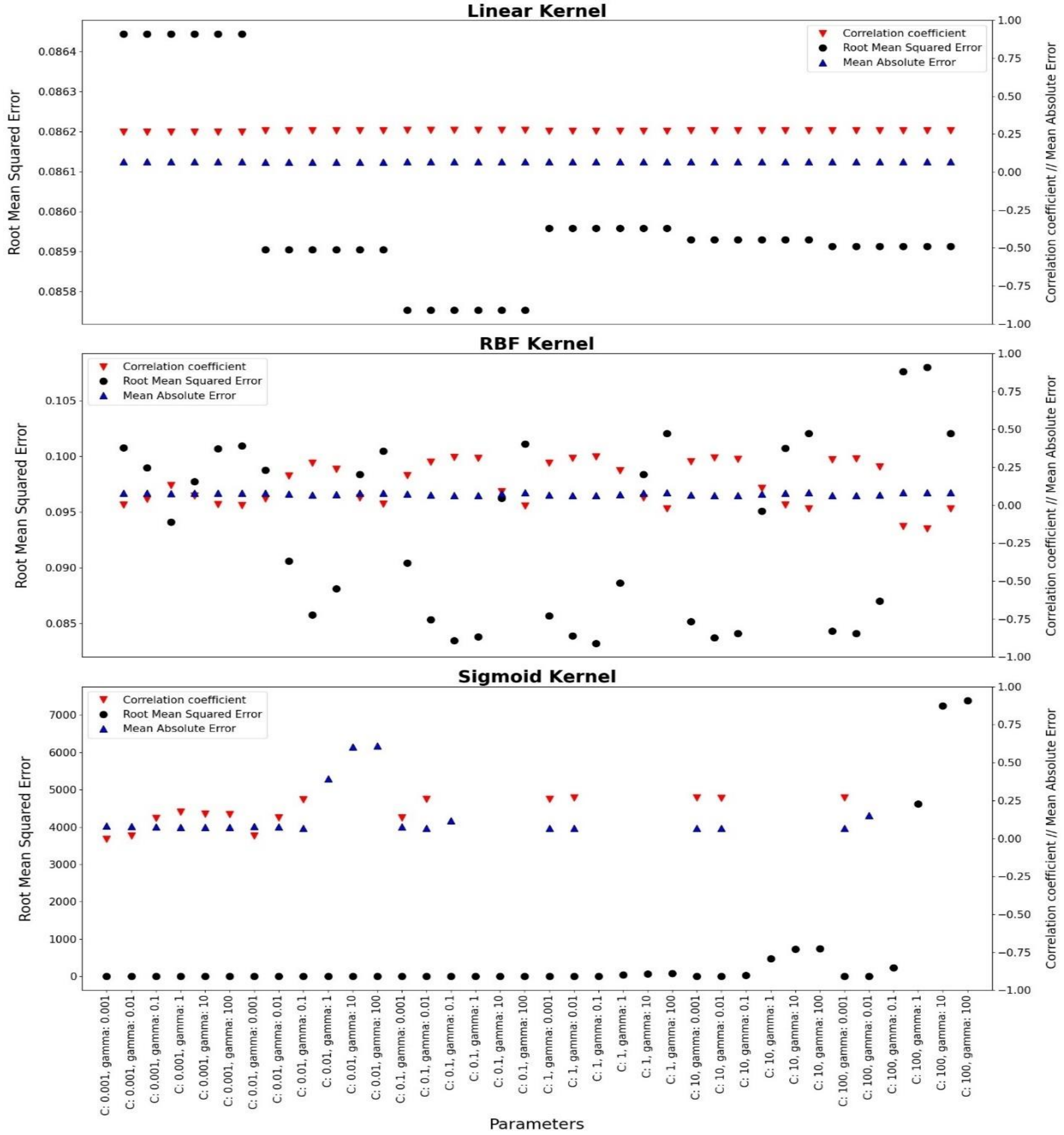

Fig 3: The statistical outcomes for tuning the hyperparameters of SVR kernel function

The correlation coefficient, mean absolute error, and root mean square error values for Equations 26 and 27 were 37%, 0.0723, and 0.0931, and 47%, 0.0700, and 0.0885, respectively. However, since the essence of the traditional regression models is to assess the relative performance of SVR, Equation 28 below was developed based on the training set used for SVR, which had 777 cases and four variables:

$$K_T = a(T_{max}) + b(S_r) + c(T_{min}) + d(T_r) + e \quad (28)$$

The constants were obtained as follows: a = 0.025, b = 0.121, c = - 0.018, d = - 0.237, and e = 0.365. The r, MAE, and $RMSE$ values were 54%, 0.0680, and 0.0877 for the training group and 48%, 0.0675, and 0.0900 for the testing group. A comparison of these values reveals that SVR with the RBF kernel function surpasses the regression models in all divisions, though the regression model (Equation 28) outperforms other kernel functions under some categories. For instance, during the training phase, the r value of Equation 28 is the same as that of the linear kernel function (54%), which is higher than that of the sigmoid kernel function at 53%. However, during testing, the correlation coefficient of the multiple regression model (48%) is higher than that of both the linear kernel function and sigmoid kernel function, each at 47%. Remarkably, the r values for RBF are 62% and 50%, respectively, during the training phase and the testing phase.

Table 2: Values for the hyperparameters and statistical measures.

Kernel function	**C**	**Gamma**	**Training set (*N* = 777)**			**Testing set (*N* = 195)**		
			r	MAE	$RMSE$	r	MAE	$RMSE$
Linear	0.01	0.001	0.5360	0.0673	0.0854	0.4693	0.0658	0.0858
RBF	1	0.1	0.6196	0.0631	0.0793	0.5045	0.0661	0.0838
Sigmoid	10	0.001	0.5347	0.0673	0.0854	0.4695	0.0657	0.0857

Because the testing phase assesses generalisation capability by evaluating the performance on unseen data, it is used to determine the best-performing ML model instead of the training phase. If MAE is rounded up to 3 decimal places (Table 2), then all three kernel functions would have the same values of 0.066. Consequently, r and $RMSE$ become the criteria for selecting the best model, and those associated with the RBF kernel function outperform the others. Thus, SVR-RBF generalises better than the linear kernel function and sigmoid kernel function, in addition to its superior performance during the training phase.

The overall relatively low correlations between the actual and calculated daily K_T values, where excellent grades of 70% and above were not achieved in both the training and testing phases, are likely attributed to the low correlation coefficient values of its estimators. However, the low yet desired values of the error terms, which do not reflect on the correlation, could stem from the fractional magnitude of K_T. Unlike other statistical measures, the correlation coefficient is often disregarded in similar studies, possibly due to its unsatisfactory outcomes, except for the yearly average clearness index (Babatunde and Aro, 1995; Udo, 2000; Olayinka, 2000). Utilising an earlier subset of the same dataset used in this current study, spanning from September 1992 to August 1994, Udo (2000) reiterated the seasonal variation of K_T, showcasing higher values during the dry season and lower values during the rainy season (Babatunde and Aro, 1995). However, during the Harmattan season, which is a subset of the dry season and characterised by an extremely cool and dust-laden atmosphere, K_T values are intermediate between those of the dry season and rainy season. Remarkably, the clearness index and sunshine hours exhibit similar seasonal characteristics (Udo, 2000). Despite variances in the volume and duration of the data used, there exists no significant difference between the findings of this study and those of Udo (2000). While the daily average K_T in this research stands at 0.47 ± 0.10, the monthly average clearness index was reported as 0.48 ± 0.06 in Udo (2000).

Although daily values were used in this study, further alignment with established findings suggests that maximum temperature is viable for modelling the clearness index in this region (Nwokolo and Ogbulezie, 2018; Chukwujindu, 2017; Olayinka, 2011). Using maximum temperature, sunshine hours, and other variables, Olayinka (2011) modelled a six-year monthly clearness index with multiple regression for Ilorin and three other cities. Therein, the correlation coefficient values were mostly around 97% and above, with much lower errors. Additionally, apart from the differences in the regression constants, the sunshine hours-based model (Equation 26) slightly outperformed the max temperature-based model (Equation 27), whereas the reverse is observed in this study.

As mentioned earlier, ML can surpass traditional statistical methods depending on certain circumstances. Due to the adopted technique, support vector regression can be a viable tool for modelling physical phenomena such as the intensity of solar radiation, possibly due to its generalisation mapping approach and the deployed kernel function (Obot et al., 2023). However, multiple regression models can be easily utilised unlike ML algorithms, which are expert-based systems.

5.0 Conclusion

In this study, using Python software and hyperparameter tuning, support vector regression with three kernel functions, namely linear, radial basis function, and sigmoid, were employed to estimate the clearness index in Ilorin (8° 32′ N, 4° 34′ E), Nigeria, from maximum temperature, sunshine hours, minimum temperature, and the ratio of both temperatures. In addition to statistical measures such as the correlation coefficient, mean absolute error, and root mean square error, multiple regression models were used to evaluate the performance of the SVR models. Approximately 972 days were selected between 1992 and 2006, with 72% used for training and the remainder for testing the models. Likely due to the large volume of data and the use of daily values, there were some differences between this study and previous studies. For instance, the daily mean K_T is 0.47 ± 0.10, compared to an earlier monthly value of 0.48 ± 0.06 (Udo, 2000). In contrast to Olayinka (2011), the errors are larger and the correlations are smaller in this study. Furthermore, the sunshine hours-based linear regression outperformed the max temperature-based linear regression in Olayinka (2011), which is the opposite here. In addition to the training phase, the performance of RBF mostly surpasses that of the linear kernel function and sigmoid kernel function at the testing phase. Although the multiple regression model, conditioned similarly to the SVR models, competed reasonably well with the other kernel functions, it neither met nor surpassed the RBF kernel function in any capacity. The RBF kernel function of SVR is recommended for the estimation of the nonseasonal daily clearness index in Ilorin, Nigeria, and other places with similar climatic conditions.

Acknowledgements

We, the authors of this manuscript, are grateful to the individuals associated with NIMET, BSRN, and the Physics Department, University of Ilorin, Nigeria, for their contributions in making this study possible through the data they collected and provided to us.

6.0 References

Al-Waeli, A.H., Kazem, H.A., Chaichan, M.T., &Sopian, K. (2021). A review of photovoltaic thermal systems: Achievements and applications. *International Journal of Energy Research,* 45(2), pp. 1269–1308. doi.org/10.1002/er.58728.

Angstrom, A. (1924). Solar and terrestrial radiation. Quarterly Journal of the Royal Meteorological Society 50, pp. 121–125. doi.org/10.1002/qj.49705021008.

Ayodele, T.R., Ogunjuyigbe, A.S.O., Amedu, A. & Munda, J.L. (2019). Prediction of global solar irradiation using hybridized k-means and support vector regression algorithms. *Renewable Energy Focus* , 29, pp. 78–93. doi.org/10.1016/j.ref.2019.03.003.

Babatunde, E.B. & Aro, T.O. (1995). Relationship between "clearness index" and "cloudiness index" at a tropical station (Ilorin, Nigeria). *Renewable Energy* 6, pp. 801–805. doi.org/10.1016/0960-1481(94)00087-M.

Paulescu, M., Paulescu, E., Gravila, P., Badescu, V., Paulescu, M., Paulescu, E., Gravila, P. & Badescu, V. (2013). *Weather modeling and forecasting of PV systems operation*, Vol. 358, London, Springer.

Basak, D., Pal, S. & Patranabis, D.C. (2007). Support vector regression. *Neural Information Processing-Letters and Reviews*, 11, pp. 203–224.

Belaid, S. & Mellit, A. (2016). Prediction of daily and mean monthly global solar radiation using support vector machine in an arid climate. *Energy Conversion and Management* 118, pp. 105–118. doi.org/10.1016/j.enconman.2016.03.082.

Fonseca Jr., J.G.S., Oozeki, T., Takashima, T., Koshimizu, G., Uchida, Y. & Ogimoto, K. (2011). Use of support vector regression and numerically predicted cloudiness to forecast power output of a photovoltaic power plant in Kitakyushu, Japan. Progress in Photovoltaics: *Research and Applications*, 20, pp. 874–882. doi.org/10.1002/pip.1152.

Besharat, F., Dehghan, A.A. & Faghih, A.R. (2013). Empirical models for estimating global solar radiation: a review and case study. *Renewable and Sustainable Energy Reviews,* 21, pp. 798–821. doi.org/10.1016/j.rser.2012.12.043.

Chukwujindu, N.S. (2017). A comprehensive review of empirical models for estimating global solar radiation in Africa. *Renewable and Sustainable Energy Reviews,* 78, pp. 955–995. doi.org/10.1016/j.rser.2017.04.101.

Nwokolo, S. C. & Ogbulezie, J.C. (2018). A quantitative review and classification of empirical models for predicting global solar radiation in West Africa. Beni-Suef University *Journal of Basic and Applied Sciences* , 7, pp. 367–396. doi.org/10.1016/j.bjbas.2017.05.001.

Smola, A. &Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing* 14, pp. 199–222. doi.org/10.1023/B:STCO.0000035301.49549.88.

Eltbaakh, Y.A., Ruslan, M.H., Alghoul, M.A., Othman, M.Y., Sopian, K. and Razykov, T.M. (2012). Solar attenuation by aerosols: an overview. *Renewable and Sustainable Energy Reviews* 16, pp. 4264–4276. doi.org/10.1016/j.rser.2012.03.053.

Grojean, R.E., Sousa, J.A. & Henry, M.C. (1980). Utilization of solar radiation by polar animals: an optical model for pelts. I 19, pp. 339–348.

Martins, G.S. & Giesbrecht, M. (2021). Clearness index forecasting: a comparative study between a stochastic realization method and a machine learning algorithm. *Renewable Energy* 180, pp. 787–805.

doi.org/10.1016/j.renene.2021.08.094.

Ramli, M.A.M., Twaha, S. & Al-Turki, Y.A. (2015). Investigating the performance of support vector machine and artificial neural networks in predicting solar radiation on a tilted surface: Saudi Arabia case study. *Energy Conversion and Management* 105, pp. 442–452. doi.org/10.1016/j.enconman.2015.07.083.

Singh, V. (2024). *The environment and its components. In: Textbook of Environment and Ecology*, Singapore, Springer, pp. 1–13. doi.org/10.1007/978-981-99-8846-4_1.

Mohanty, S., Patra, P. K. & Sahoo, S. S. (2016). Prediction and application of solar radiation with soft computing over traditional and conventional approach – a comprehensive review. *Renewable and Sustainable Energy Reviews,* 56, pp. 778–796. doi.org/10.1016/j.rser.2015.11.078.

Obot, N.I., Akanbi, S.A., Ajiboye, A.A. & Chendo, M.A.C. (2022). Charcoal and gravel basin lined solar still for brackish water purification. *Journal of Current Science and Technology*, 12, pp. 110–127. doi: 10.14456/jcst.2022.11.

Obot, N.I., Olubgon, B., Humphrey, I. & Akeem, R.A. (2023). Equatorial all-sky downward longwave radiation modelling. Communication in Physical Sciences 9(2), pp. 111–124.

Olayinka, S (2011). Estimation of global and diffuse solar radiations for selected cities in Nigeria. International Journal of Energy and Environment Engineering 3, pp. 13–33.

Okoye, C.O., Taylan, O. & Baker, D.K. (2016). Solar energy potentials in strategically located cities in Nigeria: review, resource assessment and PV system design. Renewable and Sustainable Energy Reviews 55, pp. 550–566. doi.org/10.1016/j.rser.2015.10.154.

Udo, S. O. (2000). Sky conditions at Ilorin as characterized by clearness index and relative sunshine. Solar Energy 69, pp. 45–53. doi.org/10.1016/S0038-092X(00)00008-6.

Udo, S.O. & Aro, T.O. (1999). Measurement of global, solar global photosynthetically active and downward infrared radiations at Ilorin, Nigeria. *Renewable Energy*, 06, pp. 113–122.

Vapnik, V. & Chervonenkis, A.Y. (1964). A class of algorithms for pattern recognition learning. Avtomat. i *Telemekh*, 25, pp. 937–945.

Zendehboudi, A., Baseer, M.A. and Saidur, R. (2018). Application of support vector machine models for forecasting solar and wind energy resources: a review. *Journal of Cleaner Production* 199, pp. 272–285. doi.org/10.1016/j.jclepro.2018.07.164.

Hinrichsen, K. (1994). The Ångström formula with coefficients having a physical meaning. *Solar Energy* 52, pp. 491–495. doi.org/10.1016/0038-092X(92)90656-4.

Compliance with Ethical Standards Declarations

The authors declare that they have no conflict of interest.

Data availability

All data used in this study will be readily available to the public.

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

The authors declared no source of funding.

Authors' Contributions

NIO conceptualised the study, managed the literature search, and wrote the manuscript. OU and NIO were responsible for the ML computation. OIA and NIO handled and processed the initial data before the ML application.