DOI:ttps://doi.org/10.57233/ijsgs.v12i1.1021

ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL. 12(1) WEBSITE: https://fugus-ijsgs.com.ng

INTERNATIONAL JOURNAL OF SCIENCE FOR GLOBAL SUSTAINABILITY

(A PUBLICATION OF FACULTY OF SCIENCE, FEDERAL UNIVERSITY GUSAU, NIGERIA)

Machine Learning Based Diagnostic Comparative Modeling of Region-Specific Solar Irradiance in Nigeria

G.O. Ojo[1*], O.L. Aako[2], and S.O. Are[3]

Department of Mathematics & Statistics, Federal Polytechnic, Ilaro, Nigeria.

***Corresponding author's e-mail:** gabriel.ojo@federalpolyilaro.edu.ng, 0000-0002-9454-541X

Received on: January, 2026 **Revised and Accepted on:** February, 2026 **Published on:** March, 2026

ABSTRACT

This study evaluates and compares the predictive performance of Multiple Linear Regression (MLR), Support Vector Regression (SVR), and Random Forest (RF) models for estimating solar irradiance across two climatically distinct regions of Nigeria, namely the North and the South. The models were adopted using important meteorological predictors, including minimum and maximum temperatures (tmin, tmax), relative humidity (rehu), wind speed (wins), evaporation pitch (evpi), sunshine hours (sunh), and solar radiation (radi). The analytical framework integrates Pearson correlation analysis, Variance Inflation Factor (VIF) assessment, regression diagnostics, residual distribution analysis, variable importance measures, and performance evaluation using Root Mean Square Error (RMSE) and the coefficient of determination (R^2). Correlation analysis revealed weak to moderate relationships among the predictors, while all VIF values were below 2, indicating the absence of multi-collinearity and supporting the inclusion of all variables in the multivariate models. Model adequacy was further confirmed through diagnostic and residual analyses, which also highlighted clear regional variations in model performance. In the northern region, the RF model demonstrated superior predictive capability, achieving the lowest RMSE (0.0179) and the highest R^2 (0.93), whereas SVR and MLR exhibited substantially weaker performance with R^2 values of 0.35 and 0.23, respectively. In contrast, in the southern region, SVR emerged as the most accurate model (RMSE = 0.0078, R^2 = 0.89), marginally outperforming RF (RMSE = 0.0099), while MLR remained the least effective (RMSE = 0.0132, R^2 = 0.62). These findings were supported by residual and scatter plot analyses, which showed closer agreement between observed and predicted values for SVR and RF. Variable importance analysis consistently identified maximum temperature (tmax) as the most influential predictor in both regions. Overall, the results emphasize the importance of region-specific model selection and rigorous diagnostic validation for producing reliable and interpretable solar irradiance forecasts.

Keywords: Model, Region-specific, Diagnostic, Validation, Prediction.

1.0 INTRODUCTION

The rising demand of cleaner, more reliable and renewable sources of energy has made solar energy a key element of the global sustainable power sources. Precise prediction of solar irradiance is critical to the performance of photovoltaic (PV) systems, grid stability, and the energy planning of future generations. Nevertheless, the precision of forecasting models is limited in many cases by the variability of climatic conditions in the different regions, composition of the atmosphere, atmospheric temperature trends, distribution of clouds, among other environmental dynamics. In other countries such as Nigeria where there is a clear distinction of the north and southern climates, these differences have a great impact on irradiance behavior and make it hard to predict (Aweda et al., 2018; Tor et al., 2022).

Recent international research indicates that machine learning methods have the authority to identify non-linear interactions of solar data, and that they are better than traditional empirical and regression models. As an example, Sehrawat and Vashisht (2023) combined both machine learning algorithms and digital twin technology to provide better forecasting and Demir (2021) showed the stronger predictability of AI-based solar radiation models in Turkiye. On the same note, Kushwaha and Jhingran (2022) emphasized the opportunity of deep-learning models in sustainable community energy systems, which underpins the usefulness of the computational intelligence in solar prediction activities.

In Africa, there is a lot of research on solar irradiance modelling and prediction. The method used by Kassem et al. (2022) to estimate solar irradiation in selected regions through hybrid linear-nonlinear models puts special focus on the sensitivity of the prediction models to the climatic conditions in the respective regions. Research has been done in Nigeria that examined empirical, statistical, and machine learning techniques to predict solar radiation, in various geographical climatic regions. As an illustration, Okono et al. (2023) statistically examined the distribution of solar radiation in Southern Nigeria; Bawonda et al. (2021) used empirical models to characterise global horizontal irradiance; and Nwanze et al. (2022) used both least squares regression and machine learning to analyse the thermodynamic characteristics of solar radiation in Asaba. Other studies like Ofose et al. (2023) that compared hybrid machine learning models to PV output in the South-South Nigeria area and Trull et al. (2020) that provided a comparative analysis of both statistical and ML-based irradiance forecasting models also employed machine learning-based methods.

DOI:ttps://doi.org/10.57233/ijsgs.v12i1.1021

ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL. 12(1)

WEBSITE: https://fugus-ijsgs.com.ng

Nevertheless, the advancement is still very poor, as there is very little research that compares the traditional regression models with machine learning methods in different climatic areas of Nigeria using the same predictors and methodological framework. Majority of the current research on prediction of solar irradiance is either restricted to only one region, only use empirical models or simply use a single machine learning method without comparison, but most studies have greatly overlooked full diagnostic validation. Most often, predictive accuracy indicators such as RMSE and R2 are the primary evaluations of model performance, and correlation analysis, multi-collinearity testing with the Variance Inflation Factor (VIF), residual diagnostics, scatter plot analysis, and/or variable importance analysis are given little consideration, resulting in reports of good accuracy with insufficient statistical support. In order to overcome these limitations, this study uses a combined, diagnostic-based analytical framework, which incorporates Pearson correlation analysis, VIF, residual diagnostics, variable importance analysis, and performance evaluation to make sure that predictions are not merely accurate, but also statistically reliable and interpretable. The framework is implemented on a comparative evaluation of Multiple Linear Regression (MLR), Support Vector Regression (SVR), and Random Forest (RF) models on prediction of solar irradiance in distinct climatic regions of the country, namely north and south, of Nigeria, with the use of the key climatic variables namely minimum and maximum temperatures (tmin, tmax), relative humidity (rehu), wind speed (wins), evaporation pitch (evpi), sunshine hours (sunh), and solar radiation (radi). The modeling of solar irradiance prediction over the last ten years observed is that the traditional statistical methods have been replaced by the sophisticated machine learning and hybrid modelling systems to more accurately reflect the nonlinear atmospheric processes. The initial preparatory research conducted by Voyant et al. (2017) offered the extensive review of machine learning techniques to forecast solar radiation and noted that the traditional linear regression models were not able to consider complex meteorological interactions. Their study has shown that nonlinear models including Support Vector Regression (SVR), Artificial Neural Networks (ANN) and Random Forest (RF) are always better than the classical statistical mode of operation because they are able to capture nonlinear relationships and high-dimensional inputs. This paper provided theoretical foundations to comparative and diagnostic machine learning modelling in solar power application.

Later works were more concerned with the accuracy of models by selecting and optimizing features. Chaibi et al. (2022) proposed a model that used the random forest-based feature selection and Bayesian optimization for the prediction of daily global solar radiation. Their findings revealed significant improvement in the prediction error and they concluded that when the models are well-selected in terms of variables as well as parameter tuning, it is essential to enhance the model robustness and generalization. This study strengthened the significance of diagnostic assessment in machine learning-solar modelling of irradiance. Similarly, Vashisht et al. (2023) used several machine learning algorithms such as SVR, RF, Decision Tree (DT) and ANN and a digital twin system to forecast solar irradiance based on meteorological predictors. Their results showed that ANN and SVR acquired maximally accurate results in highly varying cloud conditions, whereas RF had more consistent and stable results in more modulated environments. Moreover, a digital twin was also an important factor in improving the accuracy of real-time prediction through the dynamic capture of atmospheric-system interactions. The research concluded that model performance was mainly reliant on local weather dynamics hence the need to have region and diagnostic comparative modelling.

In their study, Nwanze et al. (2024) use the least square regression and machine learning models to compare the results of solar radiation forecasting in Asaba using thermodynamic analysis in Nigerian. They had found that machine learning models perform much better than the traditional regression methods and that thermodynamic diagnostics demonstrates that local weather conditions affect the error behavior and model stability. The current research paper presented empirical data that hybrid physical statistical knowledge is applicable in modeling solar irradiance in tropical areas.

More recent developments have been concerned with predicting the horizons and variability of climatic conditions. Taha et al. (2025) examined the short-term and long-term predictions of solar irradiance with machine learning models based on feature selection. The results obtained showed that various algorithms are best suited to specific forecasting horizons, and selection of features would enhance short-term responsiveness and long-term seasonal predictability. This research also confirmed the diagnostic assessment requirement in the process of choosing predictive models.

In particular, the study of Okorafor et al. (2025) specifically focused on the investigation of the interdependence between solar irradiance and air temperature in the key climatic regions of Nigeria through the regression analysis applied to the data on the country. They found that their findings suggested a powerful yet zone-specific effect of temperature on solar irradiance with maximum model performance recorded in Sahel and Sudan Savanna regions. The paper noted that climatic zone attributes had a large impact on irradiance temperature associations, which supports the necessity of modelling regionally. Likewise, Ohijeagbon (2025) evaluated several machine learning-based solar-radiation forecasting models in Nigeria to improve the level of integration of renewable energy. The research indicated that the performance of the models differs greatly with the climatic regions, and suits no particular algorithm to be the

DOI:ttps://doi.org/10.57233/ijsgs.v12i1.1021

ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL. 12(1)

WEBSITE: https://fugus-ijsgs.com.ng

best one will be selected at the national level. The results are very positive in the comparative diagnostic modelling and emphasize the necessity to choose models in accordance with regional climatic conditions.

2.0 METHODS

This section presents the methodological framework, used to analyze and model the variability of solar irradiance in the regions of Nigeria. The paper uses a data-driven research methodology, which is quantitative, to investigate the associations between the significant meteorological variables and the solar irradiance in Northern and Southern Nigeria. The data used is a series of past annual climate records between 2015-2023, which were taken by the Nigerian Meteorological Agency (NiMet). The modelling framework combines Multiple Linear Regression (MLR), Support Vector Regression (SVR) and the Random Forest Regression (RFR) models in embedding the linear and nonlinear relationships between solar irradiance and the chosen meteorological predictors. These modelling methods are selected due to their popularity and proven performance in the prediction of solar irradiance as is constantly being reported in the literature. These approaches are adopted in accordance with the prior empirical literature (Voyant et al., 2017; Chaibi et al., 2022), they are effective in describing the intricate physical mechanisms in the atmosphere.

2.1 Research Design

A comparative quantitative research design was implemented in order to study the solar irradiance (radi)in two climatically different regions in Nigeria (North and South). The dependent variable is solar irradiance, whereas minimum temperature (tmin), maximum temperature (tmax), relative humidity (rehu), wind speed (wins), evaporation (evpi) and sunshine hours (sunh) are considered as the independent variables. MLR, SVR and RFR are used to measure the predictive accuracy and geographic variability. Statistical diagnostics and performance metrics were applied to give robustness and reliability of the model results.

2.2 Multi-collinearity Assessment Using Variance Inflation Factor (VIF)

To assess the presence of multi-collinearity among the meteorological predictors used in modelling solar irradiance, the Variance Inflation Factor (VIF) was computed. Multi-collinearity occurs when explanatory variables are highly correlated and can inflate the variances of estimated regression coefficients, thereby reducing the statistical reliability and interpretability of model estimates (Kutner et al., 2005). For each predictor variable X_j, the VIF is expressed as;

$$VIF_j = \frac{1}{1-R_j^2} \quad (1)$$

Where R_j^2 is the coefficient of determination obtained by regressing the j^{th} predictor on all remaining predictors. In this study, VIF values were computed for the following meteorological variables: minimum temperature (tmin), maximum temperature (tmax), relative humidity (rehu), wind speed (wins), evaporation pitch (evpi), sunshine hours (sunh), and solar radiation (radi).

VIF for minimum temperature (tmin);

$$VIF_{tmin} = \frac{1}{1-R_{tmin}^2} \quad (2)$$

Where R_{tmin}^2 is obtained from $tmin\ on\ tmax, rehu, wins, evpi, sunh, and\ radi.$

A VIF value close to 1 indicates the absence of multi-collinearity, values between 1 and 5 suggest moderate correlation, while values exceeding 10 indicate severe multi-collinearity (Montgomery et al., 2012).

2.3 Multiple Linear Regression (MLR) Model Specification

Multiple Linear Regression has been used as a control statistical model to test the linear correlation between solar irradiance and various meteorological predictors. MLR presupposes that there is linearity, errors independence, homoscedasticity and normally distributed residual (Montgomery et al., 2012). A multivariate linear regression (MLR) model was defined to be used in quantifying the linear relationship between solar irradiance and important meteorological variables. Solar irradiance (SI) is a dependent variable whereas the explanatory variables are minimum temperature (tmin), maximum temperature (tmax), relative humidity (rehu), wind speed (wins), evaporation (evpi), sunshine hours (sunh), and incoming solar radiation (radi). The MLR model for the i^{th} observation is expressed as;

$$SI_i = \beta_0 + \beta_1 tmin_i + \beta_2 tmax_i + \beta_3 rehu_i + \beta_4 wins_i + \beta_5 evpi_i + \beta_6 sunh_i + \beta_7 radi_i + \varepsilon \quad (3)$$

Where; SI_i represents observed solar irradiance, β_0 is the intercept term, $\beta_1,\ \beta_2, \dots, \beta_7$ are regression coefficients, $tmin_i$, $tmax_i$, $rehu_i$, $wins_i$, $evpi_i$, $sunh_i$, and $radi_i$ are the meteorological predictors for observation i, $\varepsilon_i$is the random error term assured to be independently and identically distributed with zero mean and constant variance.

For compactness and computational efficiency, the regression model is written in matrix form as;

$$Y = X\beta + \varepsilon \quad (4)$$

Where; $Y$is an $n \times 1$ vector of observed solar irradiance values, X is an $n \times 8$ design matrix consisting of a column of ones (intercept) and the seven meteorological predictors , ($tmin_i$, $tmax_i$, $rehu_i$, $wins_i$, $evpi_i$, $sunh_i$, and $radi_i$), β is an $n \times 1$ vector of regression coefficients, ε is an $n \times 1$ vector of random errors.

2.3.1 Estimation of Regression Coefficients

The regression coefficients were estimated using the Ordinary Least Squares (OLS) method, which minimizes the sum of squared residuals between observed and predicted solar irradiance values:

$$min_\beta (Y - X\beta)^T (Y - X\beta) \quad (5)$$

DOI:ttps://doi.org/10.57233/ijsgs.v12i1.1021

ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL. 12(1)

WEBSITE: https://fugus-ijsgs.com.ng

Taking the derivative with respect to β and equating to zero yields:

$$\frac{\partial}{\partial\beta}[(Y-X\beta)^T(Y-X\beta)] = -2X^T(Y-X\beta) = 0 \quad (6)$$

Solving for β gives the closed form OLS estimator:

$$\hat{\beta} = (X^TX)^{-1}X^TY \quad (7)$$

2.3.2 Coefficient of Determination (R^2)

The explanatory power of the MLR model was evaluated using the coefficient of determination (R^2), which measures the proportion of variance in solar irradiance explained by the meteorological predictors:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (8)$$

Where;

$$SS_{res} = \sum_{i=1}^{n}\left(SI_i - \widehat{SI}_i\right)^2 \quad (9)$$

$$SS_{tot} = \sum_{i=1}^{n}(SI_i - \overline{SI}_i)^2, \text{ and} \quad (10)$$

$\widehat{SI}_i$ is the predicted solar irradiance, $\overline{SI}_i$ is the mean observed solar irradiance.

2.4 Support Vector Regression (SVR)

Support Vector Regression (SVR) is a supervised learning algorithm that has the potential of modeling nonlinear relationships based on minimizing structural risk instead of empirical risk (Vapnik, 1995). The ac im of SVR is to predict a model which differs the observed values of solar irradiance by at most an e-insignificant margin and ensures model flatness.

Let the input vector for observation i be expressed as;

$$X_i = [tmin_i, tmax_i, rehu_i, wins_i, evpi_i, sunh_i, radi_i]^T \quad (11)$$

Where the variable names are expressed as in equation (1).

The SVR regression function is given as:

$$f(x_i) = (w, x_i) + b \quad (12)$$

Where; w is the weight vector, b is the bias term, $(.,.)$ denotes the inner product.

Subject to the constraints:

$$y_i - \langle w, x_i\rangle - b \leq \varepsilon + \xi_i$$
$$\langle w, x_i\rangle + b - y_i \leq \varepsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \geq 0$$

The optimization problem is expressed as:

$$\min\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)\right) \quad (13)$$

Where;

y_i is the observed solar irradiance, ε is the width of the insensitive margin, ξ_i and ξ_i^* are slack variables capturing deviations outside the margin, C controls the trade-off between model complexity and prediction error.

2.5 Random Forest Regression (RFR)

Random Forest Regression (RFR) is a technique of ensemble learning which creates a series of decision trees based on bootstrap sampling and random selection of features and then methods their predictions to improve accuracy and reduce overfitting (Breiman, 2001). For n input vector;

$X_i = [tmin_i, tmax_i, rehu_i, wins_i, evpi_i, sunh_i, radi_i]$, the random forest prediction is expressed as:

$$\hat{f}(X_i) = \frac{1}{B}\sum_{b=1}^{B} T_b\,(X_i) \quad (14)$$

Where;

B is the total number of decision trees, $T_b(X_i)$ denotes the prediction from the b^{th} tree.

2.6 Model Evaluation Metrics

The Root Mean Square Error (RMSE) and Coefficient of Determination (R^2) are the evaluation criteria of model performance that have been commonly used to evaluate solar irradiance forecasting (Voyant et al., 2017).

2.6.1 Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (15)$$

Where, y_i is the observed solar irradiance, $\hat{y}_i = f(X_i)$ is the predicted solar irradiance based on $[tmin_i, tmax_i, rehu_i, wins_i, evpi_i, sunh_i, radi_i]$, n is the number of observations.

2.6.2 Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (16)$$

Where, $\bar{y}$ is the mean observed solar irradiance. However, the higher R^2 values and lower RMSE values indicate superior predictive performance.

3.0 RESULTS

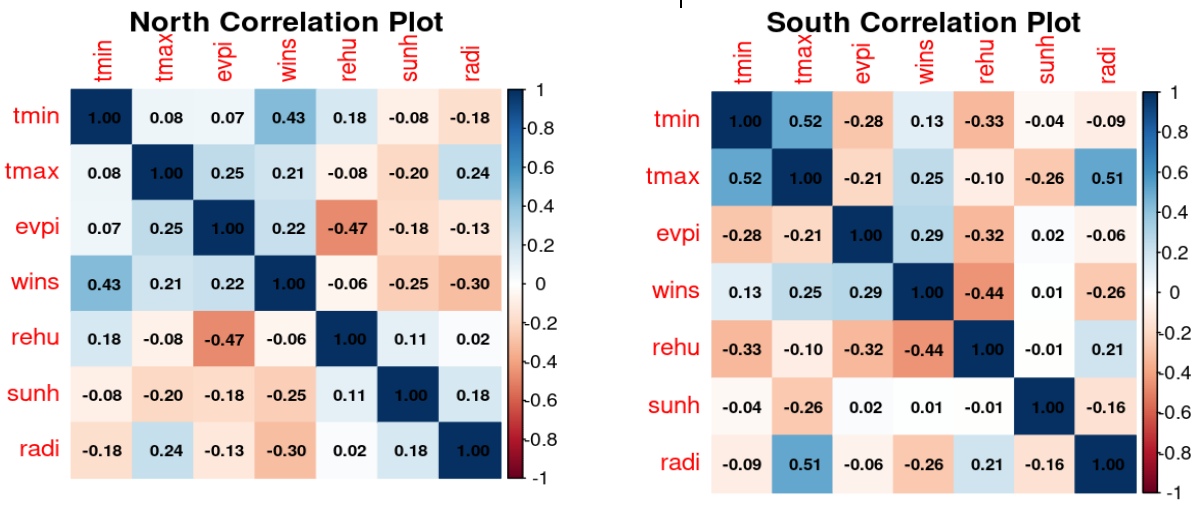
The findings of the comparative analysis of statistical and machine learning models in predicting solar irradiance in the regions under study are given and interpreted in the figures below systematically.

DOI:ttps://doi.org/10.57233/ijsgs.v12i1.1021

ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL. 12(1)

WEBSITE: https://fugus-ijsgs.com.ng



3.1 Correlation Plots

Figure 1: North and South Correlation Plots

Figure 1 presents Pearson correlation plots of climatic variables in the North and South and concludes that in general weak to moderate relationships exist among variables with no strong multi-collinearity relationships that support multivariate modeling. In the North, there exists a moderate positive relationship between minimum temperature (tmin) and wind speed (wins, r = 0.43) and negatively between evaporation pitch (evpi) and relative humidity (rehu, r = -0.47), indicating that more humid areas have a higher evaporation pitch. Solar radiation (radi) has weak to moderate negative relationships with wins (r = -0.30), maximum temperature (tmax, r = -0.24) and rehu (r = -0.12), which suggests that increased radiations could be associated with less wind, humidity and slightly low temperatures. The maximum temperature and minimum temperature in the South show stronger positive relationship (r = 0.52) where the higher the night-time temperature, the higher the day-time temperature, whereas relative humidity has moderate negative relationships with tmin (r = -0.33), evpi (r = -0.32) and with wins (r = -0.44), indicating reverse relationships between humidity and temperature (and wind). Solar radiation(radi) and sunshine hours (sunh) are also weakly correlated with other variables in both areas.

Table 1: Variance Inflation Factors (VIF) for Climatic Variables in the North and South Regions

Variable	VIF (North)	VIF (South)
tmin	1.3012	1.8034
tmax	1.1187	1.6688
evpi	1.4348	1.4328
wins	1.3677	1.4273
rehu	1.3731	1.6358
sunh	1.1085	1.1018

Source: Rstudio computations

The Variance Inflation Factors (VIF) of six climatic variables in the North and South regions are given by Table 1 as an evaluation of the possibility of multi-collinearity in the regression modeling. None of the VIF values exceed the usual threshold of 5, which means that none of the multi-collinearity problems are severe. The VIF of tmin (1.8034) and tmax (1.6688) are the highest in the South, but the North has equally low VIFs: evpi has the highest value of 1.4348. These values indicate that each variable will provide unique information to the model with little or no linear dependence with the other variables hence the support of the statistical validity of the inclusion of all the variables in the predictive model of the two regions.

DOI:ttps://doi.org/10.57233/ijsgs.v12i1.1021

ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL. 12(1) WEBSITE: https://fugus-ijsgs.com.ng

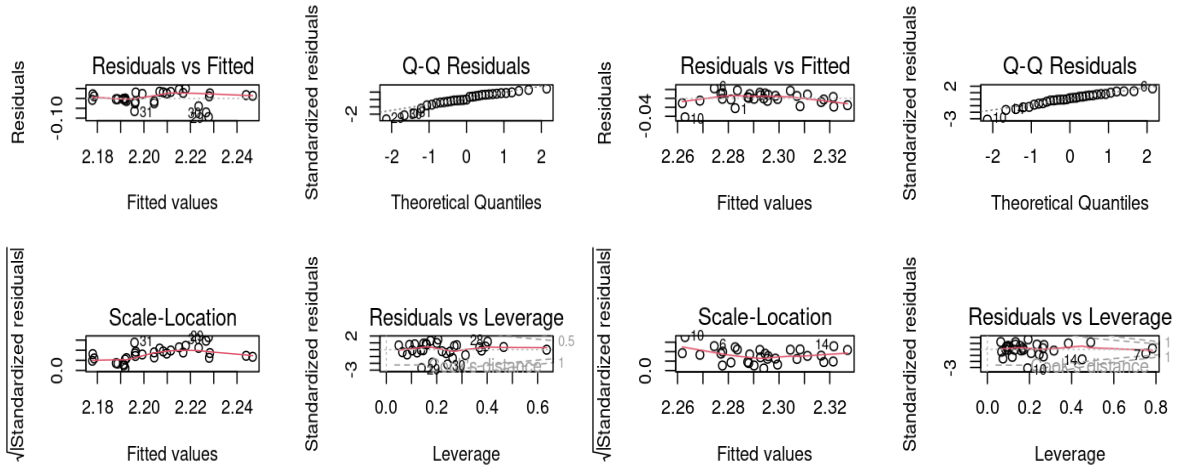


Figure 2: Diagnostic Plots for the North Region

Figure 3: Diagnostic Plots for the South Region

In Figure 2 and Figure 3, the diagnostic plots of the North and South regions reveal that the regression models are largely considered acceptable to the most common statistical assumptions. In the north, the Residuals vs Fitted plot indicates linearity and homoscedasticity, the Q-Q plot indicates the approximate normality of the residuals, and the Scale-Location and Residuals vs Leverage plots indicate no influential points and constant variance, respectively indicate the well-specified model and none of the assumptions are violated. In the South, model adequacy is also facilitated by the diagnostics: the residuals are randomly distributed, which is an indication of linearity and homoscedasticity, and Scale-Location and leverage plots do not raise any issues. Nonetheless, the Q-Q plot deduced some deviations in the tail, implying mild non-normality of residuals. Despite this minor issue, the general model fit in the two regions is statistically sound and reliable for inferences and predictions.

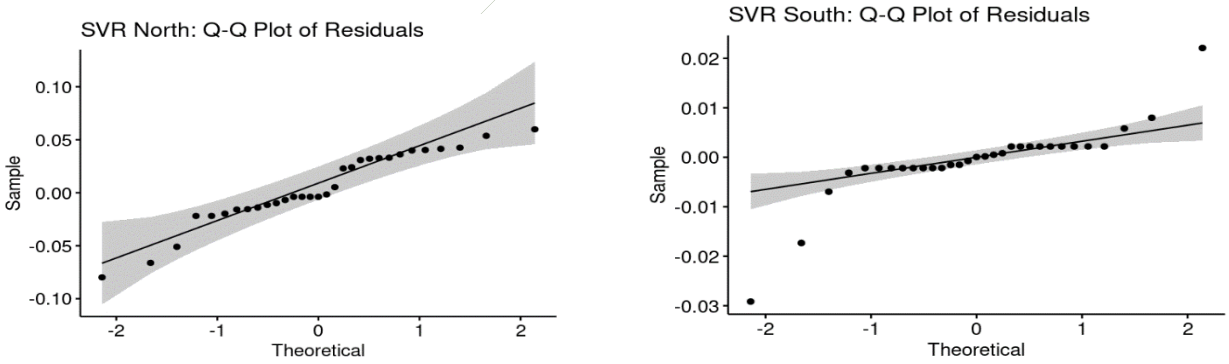


Figure 4: Q-Q Plots of Residuals from SVR Models for the North and South Regions

Figure 4 displays Q-Q plots that evaluated the normality of the residues of the Support Vector Regression (SVR) models of the North and South regions. The residuals in the North are very close to the 45-degree reference line with a few deviations at the ends, which all fall within the confidence bands of approximate normality and symmetrical distribution of errors, hence, model validity is met. In the same manner, residuals of the South model are mostly distributed around the normal distribution line with minor deviations on the lower and upper ends. These tail deviations are small and acceptable and there is an indication that the normality assumption is sufficiently met. Therefore, the two Q-Q plots render the statistical validity of the SVR models, and thus make both the predictive and inferential performances reliable.

DOI:ttps://doi.org/10.57233/ijsgs.v12i1.1021

ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL. 12(1) WEBSITE: https://fugus-ijsgs.com.ng

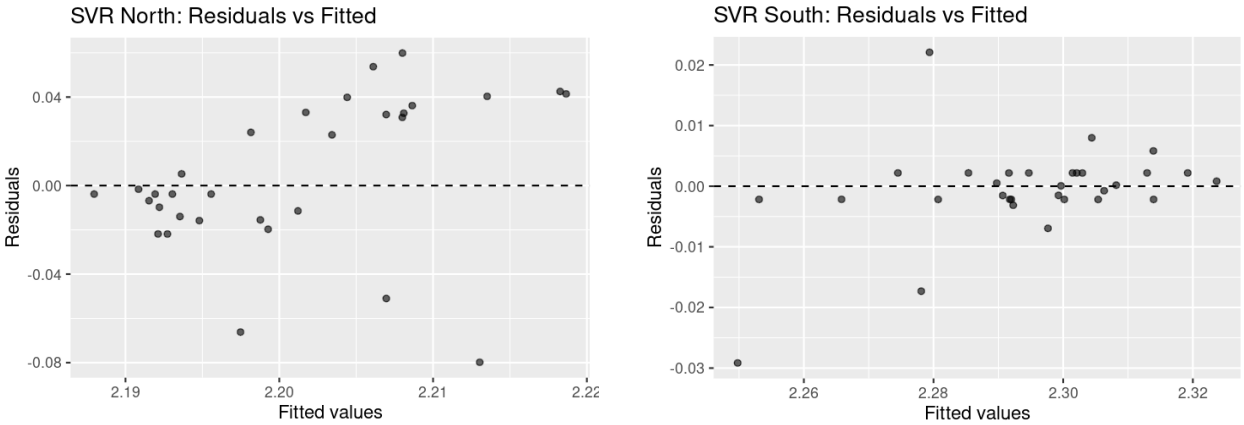


Figure 5: Residuals vs. Fitted Values Plots for SVR Models in the North and South Region

Figure 5 shows the plots of the residuals vs. fitted values of the SVR models in the North and the South which gives an idea of the model fit and the validity of the assumptions. The plot in the North region takes a unique fan shaped form, as the values of residuals are moving consistently towards negative values at the lower fitted values and towards positive values at the higher fitted values. This shows that there may be a lack of linearity and homoscedasticity. This implies that the model does not represent non-linear relationships in the data and the sizes of the residuals are relatively large between about +0.05 and -0.08 indicating greater error in predictions. That is, the plot of the South region indicates that the residuals are randomly distributed around zero and there is no clear pattern, which is an indication of the assumption of linearity and constant variance. The smaller range of residual values (around -0.03 to +0.02) shows a closer fit and more objective fit. These findings are statistically significant to suggest that the SVR model is more appropriate and reliable in South region and not necessary in North model.

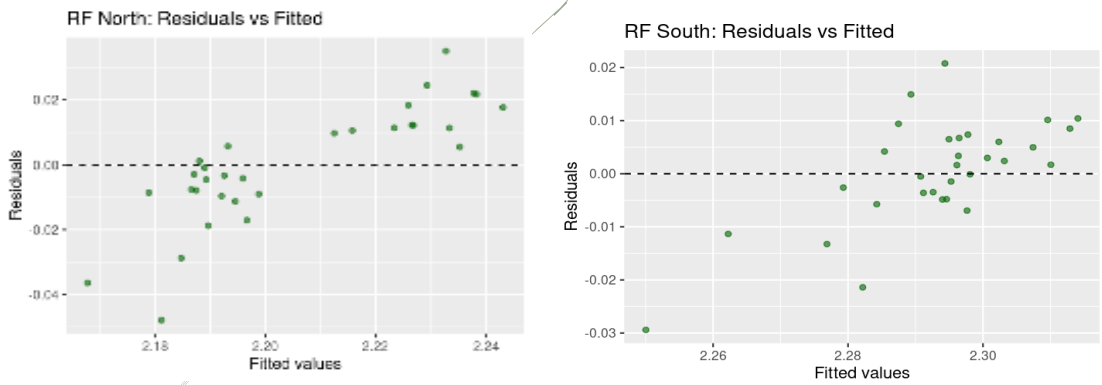


Figure 6: Residuals vs. Fitted Values Plots for Random Forest Models in the North and South Regions

Figure 6 shows the residuals vs. fitted values graphs of the Random Forest (RF) models in the North and the South region, which gives some understanding of the results of the model and the structure of variances. In the North, the residual value has evident upward trend with fitting values and so the heteroscedasticity and under-fitting may occur in some part of the data. Such unevenness in the distribution of residuals about zero indicates the lack of independence, which means that the RF model might fail to capture the significant patterns in the data, and it might need to be tuned or employ different approaches to modeling. General model adequacy is supported in the South where the residuals are randomly distributed about the zero line. An increasing dispersion of residuals at the larger fitted values, however, points to slight heteroscedasticity and potential overfitting or un-modeled higher variance. Nonetheless, the magnitude of the residuals is rather small, and they do not exhibit a strong pattern, which indicates that the model of the South RF remains quite predictive validity but with certain reservations about the consistency of the errors across the range of responses.

DOI:ttps://doi.org/10.57233/ijsgs.v12i1.1021 ISSNp: 2488-9229; ISSNe: 3027-1118
IJSGS FUGUSAU VOL. 12(1) WEBSITE: https://fugus-ijsgs.com.ng

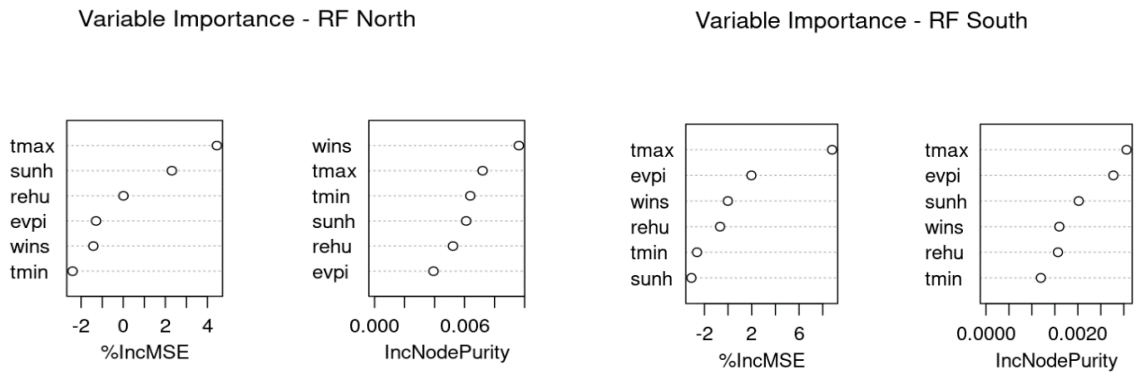


Figure 7: Variable Importance Plots for Random Forest Models in the North and South Region

Figure 7 shows the rankings of the variables-of importance in the North and South regions based on two measures, which are, 1) the percentage change in mean squared error (%IncMSE), and 2) the percentage change in node purity (IncNodePurity). In the North, tmax (maximum temperature) turns out to be the most critical predictor, with the largest positive effect on the model accuracy and node purity. Other variables, such as, "sunh" (sunshine hours) and rehu (relative humidity) also have moderate effect in model accuracy whereas tmin (minimum temperature) and wins (wind speed) show weak or even negative effects on model accuracy indicating low predictive efficiency. Nevertheless, the wins is one of this highest in IncNodePurity which means that it can be useful in splitting nodes. tmax in the South is the most useful variable with both metrics, which proves that tmax is a powerful predictor. Other variables include: "evpi" (evaporation pitch), "wins" and "rehu" which have significant contributions, with other variables like; sunh and tmin having lower positions, with sunh particularly having a negative value of the contributions in terms of degree of importance to the model. Both models, statistically, emphasize that tmax is a highly dominant predictor with other variables showing regional variation in their predictive usefulness to the model.

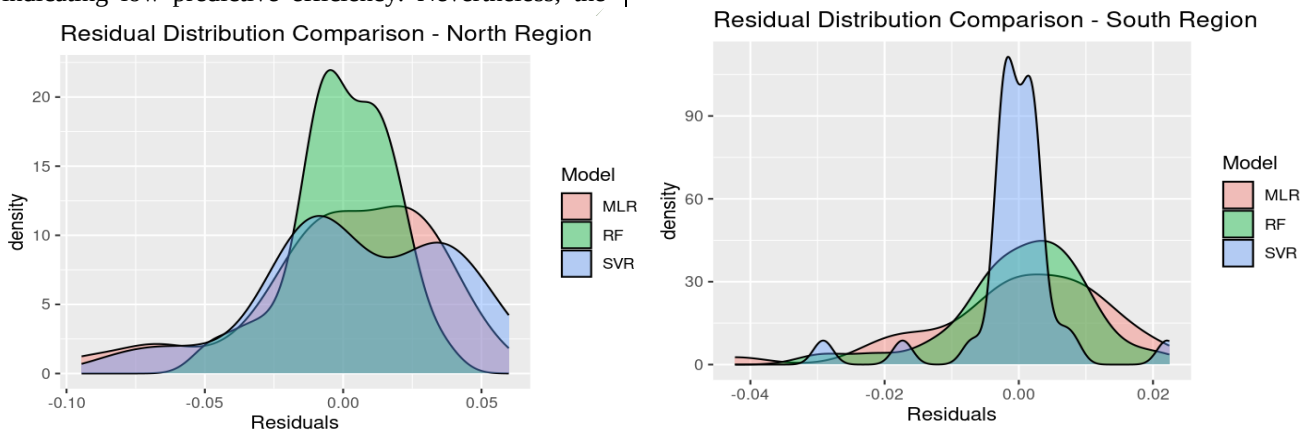


Figure 8: Residual Distribution Comparison of MLR, RF, and SVR Models in the North and South Regions

Figure 8 depict the plots of the residual distributions of the three models Multiple Linear Regression (MLR), Random Forest (RF) and Support Vector Regression (SVR) of the prediction errors between the North and South regions. The RF model has the highest prediction accuracy and lowest error variance, as they show the most peaked and narrow residual distribution around the zero that is located in the North. MLR and SVR, by contrast, tend to be more widely spread, with SVR having slight right skewness and MLR being more symmetric implying higher residual dispersion and lower accuracy. The SVR model exhibits the highest level of concentrated and narrow distribution of residual at zero and better predictive consistency and low error variance than RF and MLR do, in the South. The RF model is also centered around zero, but is spread more widely, and the residuals are the most widely spread in MLR. These trends are statistically significant with RF being the most accurate model in the North and SVR being the most effective model in the South, and MLR always demonstrating the worst predictive performance in both regions.

DOI:ttps://doi.org/10.57233/ijsgs.v12i1.1021 ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL. 12(1) WEBSITE: https://fugus-ijsgs.com.ng

Table 2: Comparative Performance Metrics (RMSE and R^2) of MLR, SVR, and RF Models in the North and South Regions

Model	Region	RMSE	R^2
MLR	**North**	0.033191170	0.2276668
MLR	**South**	0.013207156	0.6208936
SVR	**North**	0.033841863	0.3475058
SVR	**South**	***0.007788683***	***0.8894992***
RF	**North**	***0.017893520***	***0.9306523***
RF	**South**	0.009945155	0.8877250

Source: Rstudio Computations

Table 2 evaluates the performance of the models (MLR, SVR, RF) in North and South considering RMSE (Root Mean Square Error) and R2 (coefficient of determination). Random Forest (RF) has statistically better performance in the North, with the lowest RMSE (0.0179), and highest R2 (0.93), which is the best predictive performance and a good fit. Whereas in the South, SVR also slightly outperforms RF with both less RMSE (0.0078 vs. 0.0099) and slightly higher R2 (0.8895 vs. 0.8877) and is therefore the best model there. Both regions show a low level of MLR, particularly in the North (low R2 0.23), indicating that the model is not able to fit non-linear or complex relationships modeled more effectively by RF and SVR.

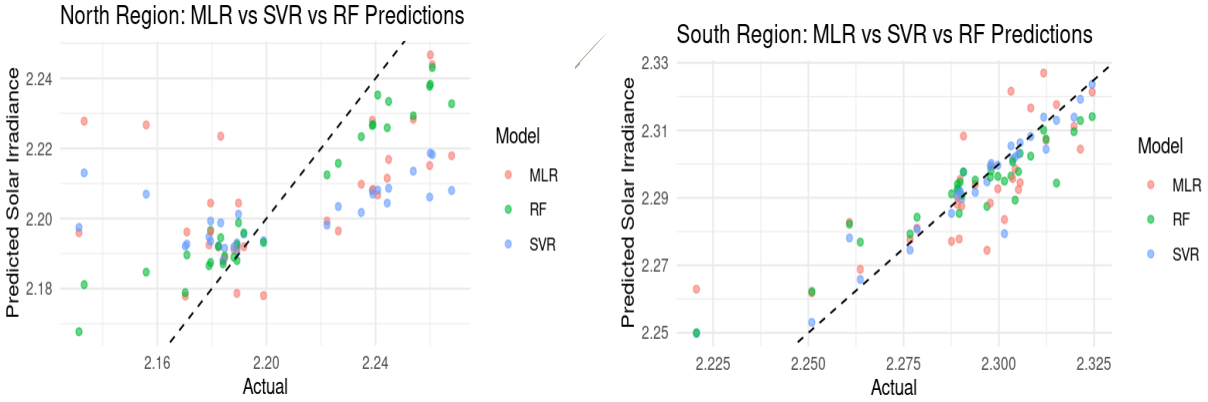


Figure 9: Predicted vs. Actual Solar Irradiance Scatter Plots for MLR, SVR, and RF Models in the North and South Region.

DOI:ttps://doi.org/10.57233/ijsgs.v12i1.1021

ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL. 12(1)

WEBSITE: https://fugus-ijsgs.com.ng

Figure 9 gives scatter plot of the predicted and actual values of solar irradiance of the MLR (red), SVR (blue), and RF (green) models at the North and South regions and the 45-degree dotted line indicates the perfect prediction. All models in the North are characterized by apparent dispersion around the diagonal, which implies error in prediction especially at the ends of the values observed. There is a slight better fit indicated by SVR and RF than there is in MLR, but all models exhibit some central bias. Conversely, the south region models are more aligned with the model showing that the SVR predictions are tightly clustered along the diagonal which implies the model is highly predictive with a low error. The RF model would come next with moderate dispersion whereas the MLR predictions are the most dispersed as they depict the weakest performance. The plots are statistically confirmed to show that the SVR has the best performance in the South, whereas the performance of the models is more comparable and less predictable in the North as there is no decisive dominance model. The models of nonlinear machine learning Support Vector Regression (SVR) and Random Forest (RF) have demonstrated a higher level of predictive power compared to Multiple Linear Regression (MLR), which establishes its efficiency in predicting the nonlinear relationships among solar irradiance. The low values of R 2 obtained when using MLR in both regions illuminate the weakness of linear assumptions in the representation of the complex atmospheric processes, which is in line with Voyant et al. (2017). All the predictors were statistically appropriate, which was confirmed by correlation and VIF analysis, in line with Chaibi et al. (2022). There were also clear regional differences, as RF did the best in the North and SVR did the best in the South, which was supported by Vashisht et al. (2023). Diagnostic analyses also confirmed the strength of SVR and RF, and variable importance findings revealed that maximum temperature (tmax) was the main contributor to solar irradiance as was also found by Okorafor et al. (2025). These findings support the need to have region cognizant, nonlinear modelling frameworks to have credible solar energy forecasts and policy planning.

4.0 DISCUSSION

In this research, Multiple Linear Regression (MLR), Support Vector Regression (SVR), and Random Forest models were compared in predicting solar irradiance covering two climatic distinct areas, i.e. the North and South using main meteorological predictors, i.e. minimum and maximum temperature (tmin, tmax), relative humidity (rehu), wind speed (wins), evaporation pitch (evpi), sunshine hours (sunh), and solar radiation (radi). The first statistical analysis based on Pearson correlation analysis showed weak to moderate correlations between predictors, whereas the Variance Inflation Factor (VIF) of less than 2 was the sign of the lack of multi-collinearity and justified the use of all variables in the multivariate modelling model. The diagnostic checks on regression and the analysis of residual also justified the reasonableness and statistical reliability of the developed models, which means that the dissimilarity in the performance depended mainly not on model misspecification but on regional climatic conditions. The findings showed a definite geographical difference in the effectiveness of the models. Further, the RF model was found to have the highest predictive performance in the Northern region with the lowest root mean square error (RMSE = 0.0179) and coefficient of determination (R2 = 0.93), which showed that it has powerful predictive ability. By contrast, SVR and MLR have significantly lower R2 values of 0.35 and 0.23 respectively, indicating that they are less able to capture the complicated atmospheric processes that control the solar irradiance in this area. On the other hand, SVR was the best model in the Southern region with RMSE = 0.0078 and R 2 = 0.89 slightly beating RF (RMSE = 0.0099) but once again with the lowest performance MLR (RMSE = 0.0132, R 2 = 0.62). Scatter as well as residual plots proved these findings with the SVR and RF predictions being closer to observed ones and the use of nonlinear machine learning methods proving better than the traditional linear regression. The analysis of the variable importance revealed that the most significant predictor of solar irradiance in the two areas was maximum temperature (tmax), which validated the significance of atmospheric dynamics associated with temperature in variability of solar energy. The variation in RF and SVR dominance in the North and South, respectively, indicates that universal predictive models are inappropriate and that one should select a model that would be more effective within a specific region. In general, the research shows that nonlinear models offer better flexibility in modeling complex relationships in the environment, but diagnostic validation is necessary to obtain reliable predictions. These results highlight the need to embrace region-conscious modelling systems in order to improve the precision of solar irradiance forecasts and to aid in the successful planning and policy development of renewable energy.

5.0 CONCLUSION

Figure 9 gives scatter plot of the predicted and actual values of solar irradiance of the MLR (red), SVR (blue), and RF (green) models at the North and South regions and the 45-degree dotted line indicates the perfect prediction. All models in the North are characterized by apparent dispersion around the diagonal, which implies

error in prediction especially at the ends of the values observed. There is a slight better fit indicated by SVR and RF than there is in MLR, but all models exhibit some central bias. Conversely, the south region models are more aligned with the model showing that the SVR predictions are tightly clustered along the diagonal which implies the model is highly predictive with a low error. The RF model would come next with moderate dispersion whereas the MLR predictions are the most dispersed as they depict the weakest performance. The plots are statistically confirmed to show that the SVR has the best performance in the South, whereas the performance of the models is more comparable and less predictable in the North as there is no decisive dominance model. The models of nonlinear machine learning Support Vector Regression (SVR) and Random Forest (RF) have demonstrated a higher level of predictive power compared to Multiple Linear Regression (MLR), which establishes its efficiency in predicting the nonlinear relationships among solar irradiance. The low values of R 2 obtained when using MLR in both regions illuminate the weakness of linear assumptions in the representation of the complex atmospheric processes, which is in line with Voyant et al. (2017). All the predictors were statistically appropriate, which was confirmed by correlation and VIF analysis, in line with Chaibi et al. (2022). There were also clear regional differences, as RF did the best in the North and SVR did the best in the South, which was supported by Vashisht et al. (2023). Diagnostic analyses also confirmed the strength of SVR and RF, and variable importance findings revealed that maximum temperature (tmax) was the main contributor to solar irradiance as was also found by Okorafor et al. (2025). These findings support the need to have region cognizant, nonlinear modelling frameworks to have credible solar energy forecasts and policy planning.

REFERENCES

Aweda, A. K., Adeyemi, O. G., & Akinwale, A. A. (2018). Assessment of solar radiation variability across different climatic zones of Nigeria. *Renewable Energy, 122*, 468–478. Bawonda, J. B., Okoye, C. O., & Nwokoye, A. O. (2021). Empirical modeling of global horizontal irradiance in Nigeria. *Renewable Energy Focus*, 38, 14–24.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Chaibi, M., El Makrini, A., & Saadane, R. (2022). Solar irradiance prediction using machine learning techniques: Performance comparison and predictor assessment. *Energy Reports,8,1250–1263.*

Demir, H. (2021). Artificial intelligence-based models for solar radiation prediction in Türkiye. *Solar Energy, 223***,** 328–343.

Kassem, Y., Gökçekuş, H., & Al-Turki, Y. A. (2022). Hybrid linear–nonlinear models for solar irradiation estimation under varying climatic conditions. *Energy Reports, 8***,** 901–915.

Kushwaha, V., & Jhingran, R. (2022). Deep learning opportunities for sustainable community energy systems. *Sustainable Energy Technologies and Assessments, 52*, 102114.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied linear statistical models *(5th ed.). New York, NY*: McGraw-Hill Irwin.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (5th ed.). *Hoboken, NJ:* Wiley.

Nwanze, E. E., Ezenwora, A. E., & Onyekachi, O. J. (2022). Thermodynamic analysis of solar radiation using regression and machine learning techniques in Asaba, Nigeria. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, *44*(4), 925–941. Nwanze, E. E., Okafor, G. C., & Obi, I. C. (2024). Comparison of least squares regression and machine learning approaches for solar radiation forecasting in Asaba, Nigeria. *Journal of Renewable Energy Research, 14**(1)***, 77–89.

Ofose, J. E., Akinpelu, J. A., & Oladipo, A. O. (2023). Hybrid machine learning models for photovoltaic power prediction in South-South Nigeria. *Cleaner Engineering and Technology, 12***,** 100633.

Ohijeagbon, O. D. (2025). Evaluation of machine learning-based solar radiation forecasting models for renewable energy integration in Nigeria. *Renewable and Sustainable Energy Reviews, 182***,** 113394.

Okono, E. E., Udoh, S. S., & Essien, U. A. (2023). Statistical analysis of solar radiation distribution in southern Nigeria. *Journal of Atmospheric and Solar-Terrestrial Physics, 239***,** 105944.

Okorafor, E. C., Adeyemi, O., & Bello, M. O. (2025). Meteorological variable importance in

solar irradiance modeling over West Africa using machine learning. *Renewable and Sustainable EnergyReviews*,*178*,113245.

Sehrawat, R., & Vashisht, A. (2023). Integration of machine learning and digital twin technology for improved solar irradiance forecasting. *Energy Conversion and Management, 275*, 116496.

Taha, M. A., Al-Garni, H. Z., & Rehman, S. (2025). Short-term and long-term solar irradiance prediction using machine learning and feature selection techniques. *Renewable Energy*, *225*, 120–134.

Tor, O. J., Yusuf, S. O., & Salami, A. A. (2022). Climatic zone influence on solar irradiance distribution in Nigeria. *Journal of Renewable and Sustainable Energy, 14****(3),*** 033701.

Trull, O., Pozo-Vázquez, D., & Ruiz-Arias, J. A. (2020). Comparison of statistical and machine learning models for solar irradiance forecasting. *Solar Energy, 206*, 945–958.

Vapnik, V. N. (1995). The nature of statistical learning theory. *New York, NY: Springer*.

Vashisht, A., Sharma, A., & Kumar, R. (2023). Regional assessment of machine learning models for global solar radiation estimation. *Solar Energy*, *247*, 112–125.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M. L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable Energy, 105*,569–582.