



A GEOSPATIALLY-ENHANCED MACHINE LEARNING FRAMEWORK FOR SOLAR RADIATION FORECASTING IN NIGERIA

¹Ojo, Gabriel Olugbenga, ²Alausa Olalekan Akeem and ¹Omotola, Oluwatobi Enitan

¹Department of Mathematics and Statistics, The Federal Polytechnic Ilaro, Ogun State, Nigeria.

²Department of Surveying and Geo-informatics, The Federal Polytechnic Ilaro, Ogun State, Nigeria.

*Corresponding authors' email: gabriel.ojo@federalpolyilaro.edu.ng

ABSTRACT

Precise forecast of solar radiation is essential for renewable energy development, environmental sustainability, and climate-informed decision-making, particularly in tropical regions with high atmospheric variability. This study evaluates the predictive performance of two advanced machine learning models Artificial Neural Network (ANN) and Extreme Gradient Boosting (XG-Boost) in estimating solar radiation (RADIA) using long-term meteorological data from Nigeria and Digital Elevation Model (DEM)-derived radiation values obtained via ArcGIS 10.6. The data span 29 years with major climatic predictors such as maximum temperature (TMAX), minimum temperature (TMIN), relative humidity (REHU), length of sunshine (SUNH), wind speed (WINS), (DEM), and evaporation pitch (EVPI). DEM-calculated values of ranged between 1200 to 2200 kWh per meter square per year and were divided into 5 solar potential classes. An extensive analytical model was used, which included correlation analysis, Variance Inflation Factor (VIF), residual diagnostics, predicted and observed comparisons and a variety of performance indicators. Results indicated significant relationships between solar radiation and atmospheric variables, with low VIF values confirming the absence of serious multi-collinearity. Comparative analysis showed that XG-Boost outperformed ANN, exhibiting superior residual behavior and closer alignment with the ideal prediction line. Its predictive accuracy was confirmed by strong performance metrics (MAE = 0.0282, MSE = 0.0010, RMSE = 0.0316, MAPE = 0.5412). Feature importance analysis identified maximum temperature and relative humidity as the most influential predictors. Thus, XG-Boost demonstrated strong predictive capability and reliability in forecasting solar radiation (1,200-2,200 kWh/m²/year), which can be used in the renewable energy planning, climate-wise agriculture, and sustainable policy development.

Keywords: Solar Radiation, XG-Boost, Artificial Neural Network, Meteorology, Predictive Modelling, Environmental Sustainability

INTRODUCTION

The increasing global population and industrialization have increased the volume of electrical power produced as the energy demand and exploration of different sources of energy all across the globe intensify (Antonanzas et al., 2016). Fossil fuels are taken to be the alternative to renewable energy. Notably, renewable energy, including solar energy, wind power, hydropower, biomass energy, and geothermal energy, is replenish able and is obtained directly in nature without destruction of the environment. Solar energy is one of such sources and has been given considerable attention because of its abundance and sustainability (Sharma et al., 2018). Proper use of solar power is important in solving the energy crisis, especially in areas with high solar power potential, such as in Nigeria. The shift to renewable energy in the world has been influenced by a mixture of factors, among them being climate change issues, depletion of fossil fuel deposits, and energy security (REN21, 2021). Solar power is one of the renewable energy sources that is distinguished by its enormous potential and the technological development in photovoltaic (PV) systems and solar thermal exploitation. In the world, solar energy is increasingly regarded as a suitable option in generating sustainable power because of the continuous increase in the efficiency of solar panels and energy storage technology (Sampaio and Gonzalez, 2017). Nigeria is a tropical country that is highly favorable to developing solar energy, because the average daily insolation per square meter is between 4.0 and 7.0 kWh/m² (Akinyele and Rayudu, 2016). Nevertheless, regardless of its vast potential for solar power, the nation continues to be dependent on fossil fuels, with little inclusion of solar energy in the country's energy

portfolio (Adejumobi et al., 2019). The low use of solar energy is significantly explained by a lack of proper infrastructures, policy limitations, and the absence of precise solar radiation prediction models that can improve planning and implementation of solar energy (Olusegun et al., 2020). Correct prediction of solar radiation is essential in planning solar energy use and grid integration, as well as increasing the efficiency of energy. Solar radiation is subject to various weather conditions (temperature, humidity, speed of wind, and cloud cover). Solar radiation estimation has been done using traditional empirical models, including the Angstrom-Prezscott model and other regression-based models, but the models rarely achieve the nonlinear and complex relationships among these meteorological variables (Rehman et al., 2017). As a result of the developments in artificial intelligence (AI), machine learning models, especially Artificial Neural Networks (ANNs) and Extreme Gradient Boosting (XG-BOOST) networks, have become popular in solar radiation prediction because they can model complex nonlinear relationships (Yadav and Chandel, 2014). The models provided are capable of better predictive power than the traditional statistical methods as well as have been effectively applied in various climatic areas. The creation of powerful ANN and XG-Boost models that will be specific to the climate of Nigeria will contribute to the effective solar radiation prediction and lead to the effective management of energy in the country (Rahman et al., 2020).

According to Mahmoud Sabo et al., (2016), spatial data has been used to estimate the monthly solar radiations in areas with complicated topography using Geographic Information Systems (GIS) and Digital Elevation Models (DEM). The

findings demonstrate that the changes in topographical factors have a substantial impact on the level of solar radiation as opposed to the data collected by NASA. The given strategy facilitates effective energy planning, makes solar photovoltaic systems a better location choice and enhances the evaluation of geographic variability.

Some of the major factors that have contributed to the low adoption of solar energy in the power infrastructure in Nigeria include the inability to allocate sufficient funds, lack of technology and presence of ineffective policy execution. Moreover, conventional solar radiation prediction techniques tend to have difficulties with the representation of complex and nonlinear correlative relationships between meteorological factors and solar irradiance (Rehman et al., 2017). This shortcoming has the disadvantage of making wrong predictions, compromising strategic planning, and hindering the successful implementation of solar energy projects. To combat this gap, it is highly necessary to create and deploy enhanced machine learning models like Artificial Neural Networks (ANNs) and Extreme Gradient Boosting (XG-Boost), which can be used to learn complex patterns in meteorological data. These models have the potential to boost solar irradiance prediction and largely contribute to the development of sustainable energy solutions in Nigeria. The proper planning, design, and operation of solar energy systems requires an accurate prediction of solar radiation. Dependable forecasts increase the efficiency of energy, improve power production, and enable grid adoption (Yadav and Chandel, 2014). The combination of AI-reliant solutions, especially ANN and XG-Boost, can potentially help to increase the accuracy of predictions and manage energy consumption. The results of the research will serve as the foundation of better policies and infrastructure related to renewable energy in Nigeria and will lead to energy sustainability and economic growth. This research intends to compare the predictability of the Artificial Neural Network (ANN) and XG-Boost models in predicting solar irradiation using meteorological variables, and Digital Elevation Model (DEM). In the past, solar radiation was predicted using empirical and statistical models. Rehman et al. (2017) deduced that the efficiency of various empirical models in varying climates was highly constrained in a tropical setting because of the fluctuation of cloud cover and humidity. This emphasizes the necessity of more nonlinear and flexible methods. The use of machine learning (ML) and artificial intelligence (AI) has revolutionized predicting the resources of renewable energy in the recent past. ANNs and XG-Boost, specifically, are now favored due to their ability to learn data without necessarily having explicit physical modeling. Furthermore, Yadav and Chandel (2014) reviewed different ANN structures and concluded that the multilayer perceptron (MLP) showed much better performance in comparison with linear regression models. Their predictive capability in nonlinear relationships makes them applicable in the prediction of solar radiation, where the interactions between the variables of meteorology are complicated. XG-Boost is especially useful in predicting solar radiation because it can capture complicated nonlinear relationships between input variables and the target variables. In contrast to the time-series models based on a sequential data structure, XG-Boost uses gradient boosting decision trees to learn on past data effectively and does not require a temporal recurrence mechanism. This ensures that it is very effective in the management of diverse and high-dimensional meteorological inputs. XG-Boost has been used to forecast solar radiation, and the results indicate that the model can capture interactions between variables like temperature, humidity, and cloud

cover, enabling the model to be seasonally and diurnally accurate (Rahman et al., 2020). In Nigeria, Oyedepo et al., (2015) compared the output of the ANN models in estimating solar radiation in three cities based on meteorological data. Their findings revealed encouraging accuracy, yet the experiment was only performed with feed forward architectures and only brief training periods.

Ghimire et al., (2019) employed XG-Boost on hourly solar radiation prediction in California, showing better results over support vector machines (SVM) and random forests (RF). Their results demonstrated that XG-Boost can use high-dimensional meteorological data and giving accurate predictions with smaller error margins. Similarly, Kong et al., (2017) applied XG-Boost to multi-step solar radiation predictions and discovered that the model exhibited high predictive control with a longer forecasting horizon compared to the few conventional models, which failed with the extended time gap. These developments notwithstanding, research on the use of XG-Boost models in African countries, especially Nigeria, is limited. Due to heterogeneous and uneven climatic conditions in the tropical areas, the use of XG-Boost on this topic is a good research proposal with the likelihood of making significant contributions to solar energy planning and integration. Kumar and Sharma (2020) employed a comparative study of methods in India and discovered that ANN-based models were effective in general trends, whereas XG-Boost-based models would generate more reliable predictions, especially when there are complicated interrelations among meteorological factors. Firoz and Moghimi (2021) employed principal component analysis (PCA) to narrow down on the meteorological variables and discovered that the highest predictors of solar radiation in semi-arid areas were temperature and sunshine hours. The humidity and clouds in tropical areas such as Nigeria cause extra noise, and to be able to note them properly, a strong model such as XG-Boost is needed to be used. Although solar radiation prediction has been investigated in many studies around the world, the number of studies that use Artificial Neural Networks (ANN) and XG-Boost in the Nigerian environment is small. Nigeria has many models that are based on imported data or that are not thoroughly localized, meaning the models are not quite relevant to the unique climatic factors in Nigeria. Such weaknesses as limited implementation of deep learning methods of solar radiation prediction in the long term, the lack of comparative studies between ANN and XG-Boost with real meteorological data of cities in Nigeria, and inaccessibility to high-quality data of diverse regions and weather stations in Nigeria can be identified. As such, the proposed research aims to address such gaps by training and validating ANN and XG-Boost models with locally obtained meteorological data to enhance the accuracy and contextual validity of the solar radiation prediction in Nigeria.

MATERIALS AND METHODS

The Nigerian Meteorological Agency's (NiMet) historical climatic data served as the basis for this investigation. Three decades of meteorological data, including solar irradiation, minimum and maximum temperatures, wind speed, relative humidity, daylight hours, and evaporation pan, were recorded at the Ilorin weather station between 1994 and 2023. Spatial variability was also included by using a Digital Elevation Model (DEM) to come up with more solar radiations in ArcGIS 10.6. (Choi et al., 2019) The raw data before being analyzed were screened, cleaned, and tabulated using Microsoft Excel with care to provide consistency, eliminate any missing or erroneous data, and to provide more reliability

to the data in terms of predictive modelling. The further statistical and machine learning procedures were carried out in the R software (version 4.3.3), which is an efficient platform for preprocessing data, model development, validation, and performance assessment.

Model Identification

Prediction of the XG-Boost model on a specific instance x_i is given as:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in \mathcal{F} \quad (1)$$

Where: \hat{y}_i : the predicted value for the i^{th} instance, K : the number of trees, f_k : the k^{th} decision tree (a function that maps input x_i to a prediction), \mathcal{F} : the space of regression trees (Classification and Regression Trees).

The objective function to minimize is:

$$\mathcal{L}(\phi) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (2)$$

Where: $l(y_i, \hat{y}_i)$: the **loss function** (e.g., mean squared error) that measures the difference between predicted and actual values,

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (3)$$

while, the regularization term to control model complexity, T : number of leaves in the tree, w : leaf weights, γ : complexity penalty for adding a leaf, λ : L2 regularization term on leaf weights.

Model Specification

This stage expresses the statistical learning structures of the ANN and XG-BOOST models used to map meteorological variables to solar irradiation.

Artificial Neural Networks-Based Approach

According to (Ojo and Udomboso, 2021), artificial neural networks (ANN), often traced back to McCulloch and Pitts (1943), are inspired by the learning processes of the human cognitive system and the neurological functions of the brain with a capability for self-learning and automatic abstracting and with a possible benefit of reducing modeling times. One application of ANNs is an alternative modeling strategy to traditional methods of data and time series analysis.

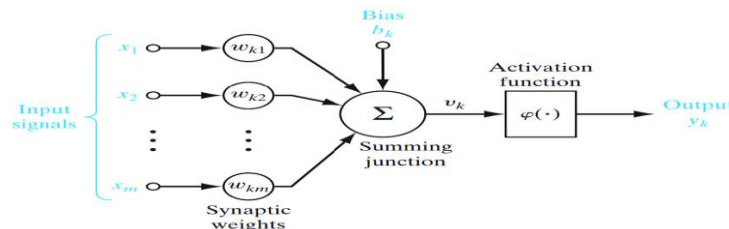


Figure 1: Nonlinear Model of a Neuron (Ojo & Udomboso, 2021)

In Figure 1, the model usually consists of three layers: where the first layer is the input layer where the introduction of data the network is involved, the second layer is the hidden layer where data are processed and the last layer is output layer where the results of given input are produced (Samsudin et al., 2010). As far as the artificial neural network approach is involved, it is fundamental processing elements such as the neurons, the architecture of the network that connects the neurons and the training algorithm used to find values of the network parameters for performing a particular task. The ANN concept in many forms has been adopted such as business and management, medicine, speech processing, computer vision, control systems, tourism and demand, (Law, 2000), Hydrology, Radar, satellite and meteorology concept (Koizumi, 1999) among many other fields. Over time, different types of ANNs have been designed to address specific problems, applications, and data sets. These include MLP, RNN, CNN, and hybrid neural networks. Furthermore, two types of RNNs, LSTM and GRU, are also widely researched due to their exceptional performance when applied to time series data. Hence, this section will study these five ANN models' internal structure, design, and operation.

Multilayer Perceptron

The simplest and most efficient data-based modelling device of ANNs is MLP. A MLP with a single input layer, two hidden layers and a single output layer can be represented through Fig.1. The first process of data input in the MLP is the input layer. This layer takes the outside input variables. Environment, for instance; the sun rays, surrounding temperature, wind velocity, humidity, clarity, dust, and the cover of clouds by the external environment. These variables enable the MLP to grasp and acquire the complicated association linking environmental parameters and the PV generation of power. It is therefore necessary to pre-process

and arrange these input variables and then feed them in the MLP. Then the input variables are sent to the hidden layer which is situated in between the input and output layer. The concealed layers are indispensable to the best MLP facilities particularly when the main concern is accuracy and time complexity. The neurons on the hidden layers are fed by the value of the input layer and compute products of the weighted summation of the value on the input layer. Then the weighted sum is processed by each neuron with an activation function, e.g. the sigmoid, softmax, rectified linear unit (ReLU), parametric ReLU, swish, or Gaussian error linear unit (GELU) to the weighted sum. The significance of the input of a neuron can influence the decision of the neuron to be activated or not through an activation function. In addition, the activation function has made the model include nonlinearity, which enables the MLP to identify and training complex data patterns. The combination of the activation outputs of all the neurons is then summed to the output which is sent to the output layer of the model; the output is the prediction that is given out by the MLP which is the PV power output. Take a four-layer MLP model (one input, two hidden and one output) that is shown in figure 1. The Multilayer Perceptron (MLP) model consists of: m input variables denoted as x_i , where $i = 1, 2, 3, \dots, m$ and p output variables denoted as y_l , where $l = 1, 2, 3, \dots, p$. While the network contains two hidden layers: the first hidden layer has n nodes, denoted as h_j , where $j = 1, 2, 3, \dots, n$, and the second hidden layer has o nodes, denoted as h_k , where $k = 1, 2, 3, \dots, o$. The statistical representation of the MLP is given by the following equations:

First hidden layer output

$$h_j = f_h(\sum_{i=1}^m x_i w_{ji} + a_j) \quad (4)$$

Second hidden layer output:

$$h_k = f_h(\sum_{j=1}^n h_j w_{kj} + b_k) \quad (5)$$

Output layer

$$y_p = f_y(\sum_{k=1}^p h_k w_{lk} + c_p) \tag{6}$$

where:

w_{ji}, w_{kj}, w_{lk} are the weight parameters, a_j, b_k, c_p are the bias terms, f_h and f_y are the activation functions for the hidden and output layers, respectively. Furthermore, a structured feature selection procedure will be implemented to identify the most relevant and non-redundant predictors by combining filter-based methods (Pearson correlation analysis to assess relationships between solar radiation (RADIA) and candidate variables, with a threshold of $|r| < 0.3$ for exclusion, and Variance Inflation Factor (VIF) analysis with a threshold of 10 to address multi-collinearity) and an embedded method (Extreme Gradient Boosting (XG-Boost) algorithm to rank feature importance), ensuring that the selected predictors are statistically significant, non-collinear, and practically relevant for improving model accuracy and generalization.

Model Accuracy Assessment

The comparison of the performance and accuracy of the time series forecasting models is carried out by comparing Artificial Neural Network (ANN) and Extreme Gradient Boosting (XG-BOOST) algorithms. Three established error metrics, which are Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) were used in the analysis. These measures provide a full picture of the quality of model prediction because they measure various characteristics of forecasting errors. The metrics of errors are as follows:

Mean Absolute Error (MAE):

$$MAE = \frac{1}{n} \sum_{t=1}^n |e(t)| \tag{7}$$

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction.

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n e(t)^2} \tag{8}$$

RMSE penalizes large errors more heavily than MAE, making it sensitive to outliers in the dataset.

Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{100}{n} \sum_{t=1}^n \left| \frac{X_{real,t} - X_{pred,t}}{X_{real,t}} \right| \tag{9}$$

MAPE provides error as a percentage, making it easier to interpret across different scales. Here; error at time t ; $e(t) = X_{real,t} - X_{pred,t}$, $X_{real,t}$ is the actual (real) value at time t , $X_{pred,t}$ is the predicted value at time t , n is the total number of observations. These evaluation criteria are widely used in time series forecasting literature due to their interpretability and reliability (Hyndman & Koehler, 2006; Zhang et al., 1998).

DEM Processing and Solar Radiation Estimation

The SRTM digital elevation model was downloaded from the United States Geological Survey (USGS) website and was

overlaid on the study area shape file to extract by mask the area of interest. This was later segmented to different towns within the study area, and the solar radiation analysis tool was used to generate the solar radiation range, which was later reclassified into five classes. These value range were then extracted from the DEM as raw solar radiation values in MJ/m² (mega joules per square meter), which need to be converted to kWh/m²/year. Were

$$1kWh = 3.6 MJ \Rightarrow kWh = \frac{MJ}{3.6}$$

Hence, 1,038,898.779 MJ \rightarrow kWh

$$\frac{1,038,898.779}{3.6} = \frac{288,583kWh}{year}$$

$$\frac{1,904,627.522}{3.6} = 529,063kWh/m^2/year$$

$$\frac{2,885,786.765}{3.6} = \frac{801,607kWh}{year}$$

$$\frac{4,136,283.839}{3.6} = \frac{1,149,023kWh}{year}$$

$$\frac{4,905,820.5}{3.6} = \frac{1,362,728kWh}{year}$$

The range was then noted to be

$$\text{Min} = 1,038,898.779, \text{Max} = 4,905,820.5$$

Hence, to normalize the values to Nigeria annual solar generation rang

$$\text{Annual Min} = 1,200 kWh/m^2/year, \text{Annual Max} = 2,200 kWh/m^2/year$$

$$\text{Normalized Value} = \frac{\text{Target Min} + (\text{DEM Value} - \text{DEM Min}) \times (\text{Target Max} - \text{Target Min})}{\text{DEM Max} - \text{DEM Min}}$$

Normalizing each DEM class boundary

$$1,038,898.779 \rightarrow kWh/year$$

$$1,200 + \frac{(1,038,898.779 - 1,038,898.779) \times (2,200 - 1,200)}{4,905,820.5 - 1,038,898.779} = 1,200$$

$$1,904,627.522 \rightarrow kWh/year$$

$$1,200 + \frac{(1,904,627.522 - 1,038,898.779) \times 1,000}{4,905,820.5 - 1,038,898.779} \approx 1,200 + 224 = 1,424$$

$$2,885,786.765 \rightarrow kWh/year$$

$$1,200 + \frac{(2,885,786.765 - 1,038,898.779) \times 1,000}{3,866,921.721} \approx 1,200 + 471 = 1,671$$

$$4,136,283.839 \rightarrow kWh/year$$

$$1,200 + \frac{(4,136,283.839 - 1,038,898.779) \times 1,000}{3,866,921.721} \approx 1,200 + 79 = 1,279$$

$$4,905,820.5 \rightarrow kWh/year$$

$$1,200 + \frac{(4,905,820.5 - 1,038,898.779) \times 1,000}{3,866,921.721} = 2,200$$

RESULTS AND DISCUSSION

This paper shows the statistical and environmental explanation of the findings derived after the predictive modelling of solar irradiation (RADIA) of Artificial Neural Network (ANN) and Extreme Gradient Boosting (XG-Boost) models. The discourse incorporates both statistical diagnostics and environmental arguments in its explanation of the links between the meteorological variables and the solar radiation dynamics in the Nigerian climatic setting.

Correlation Plot

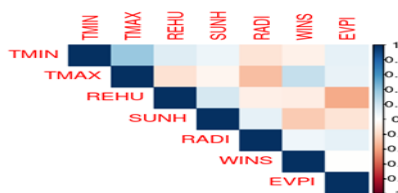


Figure 1: Correlation Matrix of Meteorological Variables

The Pearson correlation table (Figure 1) shows the association between the meteorological variables employed in the research. The color gradient will be between -1 (strong negative) and +1 (strong positive), darker color shades will reflect a stronger correlation. The outcomes indicate that there is a moderate positive relationship between the day temperature (TMIN) and maximum temperature (TMAX), which is consistent with the behavior of the thermal processes in the study area where daily extremes in the temperature can be observed to be consistent. Solar irradiation (RADI) has weak-moderate positive relationships with TMAX and sunshine duration (SUNH), affirms that more atmospheric heating and long exposure to the sun is a positive factor in solar energy provision. This relationship would be expected in the tropical climatic conditions like Nigeria where higher

temperatures are usually related with less cloud cover and more transparency in the atmosphere which permits entry of more solar radiation. There is moderate positive correlation between relative humidity (REHU) and evaporation (EVPI), which shows interaction between the moisture content of the atmosphere and the surface water loss processes. Largely, wind speed (WINS) demonstrates weak correlations with most variables meaning that it does not have a direct impact on solar irradiation as much as thermodynamic and other moisture-related variables. Therefore, the correlation matrix gives valuable initial knowledge regarding the interdependencies between variables and facilitates an appropriate choice of features and reduces the redundancy of predictive modelling.

Table 1: Variance Inflation Factor (VIF) for Predictor Variables in the Linear Model for Solar Irradiation Multi-Collinearity Assessment

| Predictor Variable | VIF Value |
|--------------------|-----------|
| TMIN | 1.280704 |
| TMAX | 1.318266 |
| REHU | 1.277769 |
| SUNH | 1.109078 |
| WINS | 1.164049 |
| EVPI | 1.216347 |

Source: Rstudio Computations

Table 1 depicts the Value Inflation Factor (VIF) of the predictor variables that were used in the solar irradiation model. All the predictors have the VIF between 1.11 and 1.32, which is significantly lower than the traditional value of 5. It means that there is no serious multi-collinearity of explanatory variables. TMAX has the largest VIF value (1.318) and is then followed by TMIN and REHU but these three values indicate insignificant variance inflation. The low VIF of SUNH, WINS, and EVPI also prove the statistical

independence of predictors. As a result, the regression structure becomes stable so that the estimation of the coefficients is reliable and the future machine learning models become less vulnerable. Put differently, environmental standpoint arrived at the conclusion that the low multi-collinearity indicates that each meteorological factor measures different atmospheric processes that affect solar radiation, such as thermal conditions, and moisture variability as well as energy exchange in the surface atmosphere.

Diagnostic Evaluation Plots

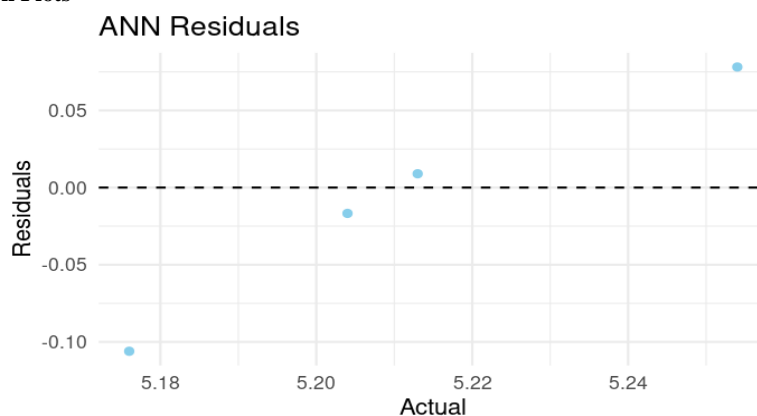


Figure 2: Residual Plot of the Artificial Neural Network (ANN) Model

Figure 2 presents the distribution of residues of the Artificial Neural Network (ANN) model. The residuals are seen to be randomly scattered around the zero-reference line, meaning that the underlying data structure is quite well represented by the model and there is no strong systematic bias in it. The overall small sizes of the residuals indicate the tolerable predictive performance. Nevertheless, the fact that we have a

slightly larger negative residual implies that the ANN model is underperforming on localized underestimation, which implies that the ANN model cannot capture all nonlinear atmospheric interactions with solar radiation. The limitation can be caused by sensitivity to train setup and network topology which may affect the generalization performance of such complex climatic data.

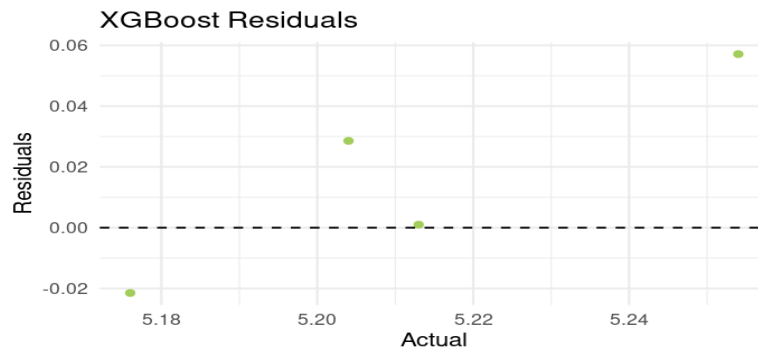


Figure 3: Residual Plot of the XG-Boost Model

Figure 3 illustrates the residual plot that shows the performance of XG-Boost model. Even though there is moderate rise in residual spread at high actual value, which is evidence of a slight heteroscedasticity, the residuals are well-centered about zero. This trend indicates better learning in nonlinear relationships than ANN. Owing to smaller levels of underestimation in low radiation levels and minor

overestimation in higher radiation levels is also observed but within acceptable predictive margins. Generally, residual behavior supports high model stability and aptness. This performance is also more environmentally friendly to show that XG-Boost can capture the multifaceted interaction of temperature, humidity, and atmospheric energy balance to cause changes in solar radiation.

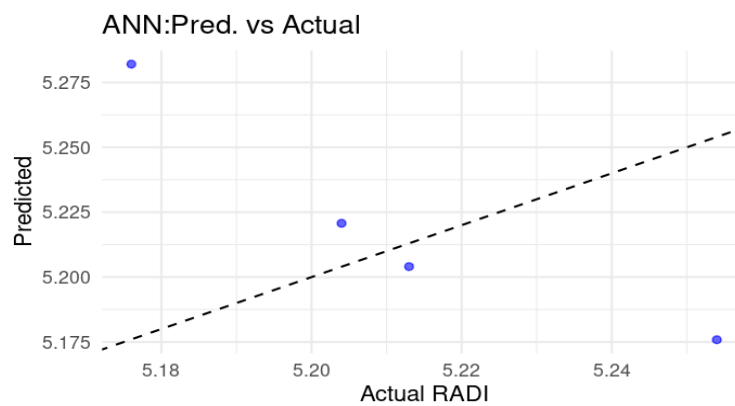


Figure 4: Predicted vs. Actual Plot for ANN Model in Predicting RADIA

The results in the ANN model in forecasting and observed solar irradiation are compared in figure 4. Preferably, the correct predictions must be close to the 45deg line that is the expression of perfect agreement. Nonetheless, visible scatter about the line can be considered as a sign of inconsistencies in predictions, both underestimation and overestimation. This

bias indicates that ANN is only showing broad trends but not the nonlinear climatic processes that determined the character of solar radiation within the dataset. As a result, the ANN model has moderate predictive accuracy, which is not the best.

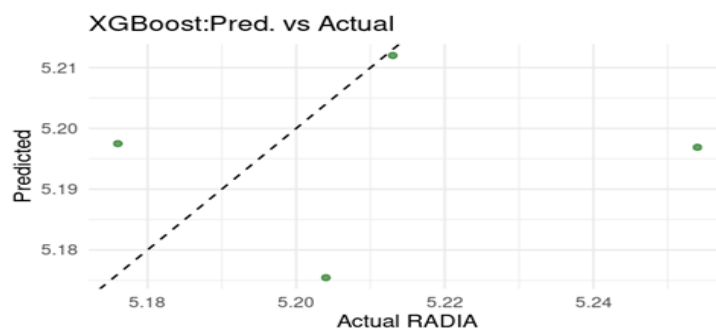


Figure 5: Predicted vs. Actual Plot for XG-Boost Model in Predicting RADIA

Figure 5, conversely, demonstrates that there is a high correlation between predicted and actual values of the XG-Boost model. The data points are clustered closely around the theoretical reference line which implies that the error of prediction is low and the bias is less. The enhanced clustering

attests to the fact that XG-Boost has a better capacity to model the intricate relationships existing in the environment utilizing ensemble learning and gradient optimization. This proves it to be useful in meteorological prediction activities with non-linear atmospheric processes.



Figure 6: Comparative Performance Metrics of ANN and XG-Boost Models

Table 2: Performance Comparison of ANN and XGBoost Models

| Model | MAE | MSE | RMSE | MAPE |
|----------|--------|--------|--------|--------|
| ANN | 0.0569 | 0.0061 | 0.0782 | 1.0923 |
| XG-Boost | 0.0282 | 0.0010 | 0.0316 | 0.5412 |

Source: Rstudio Computations

The comparative performance analysis of channels ANN and XG-Boost based on MAE, MSE, RMSE, and MAPE are available in figure 6 and table 2. XG-Boost always obtains smaller figures of error (MAE = 0.0282; MSE = 0.0010; RMSE = 0.0316; MAPE = 0.5412) than ANN on all the evaluation criteria. With a lower error though, it suggests that the predictive precision and variance of the observed and

estimated values is enhanced statistically. The significantly smaller MAPE value indicates that XG-Boost generates more precise percentage-based predictions, which further confirms its usefulness in estimating the solar radiation. As such, they indicate that ensemble tree-based learning models can be more effective with neural networks when they are used with structured meteorological data with moderate sample sizes.

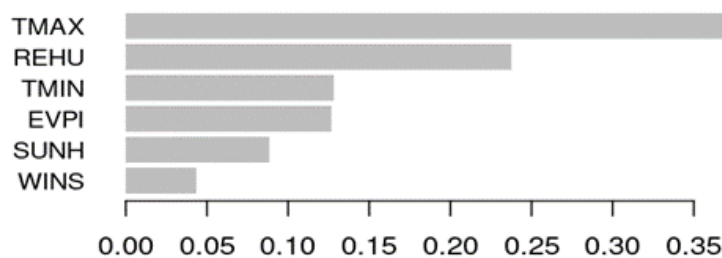


Figure 7: Feature Importance and Environmental Interpretation from XG-Boost Model

The importance of features obtained using the XG-Boost model is shown in Figure 7. The strongest predictor is the maximum temperature (TMAX) which has more than 35 percent of the contribution to the performance of prediction. Relative humidity (REHU), minimum temperature (TMIN) and evaporation (EVPI) follow behind with sunshine length and wind speed being less important. On environmental side, the temperature variables are dominating, and it signifies the

high sensitivity of solar radiation to the thermal conditions of the atmospheric environment that control the formation of clouds and the transmissivity of the atmosphere. Relative humidity is also significant, which also shows the significance of moisture content in scattering and absorbing solar energy. Damp weather is normally associated with cloud cover meaning fewer incoming rays, and drier weather reflects more sunlight.

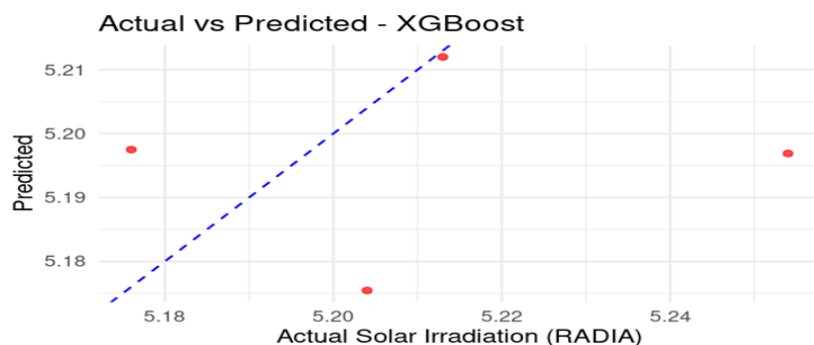


Figure 8: XG-Boost Model Validation

Table 3: Actual vs. Predicted Solar Radiation (RADIA) Values

| Observation | Actual_RADIA | Predicted_RADIA |
|-------------|--------------|-----------------|
| 1 | 5.213 | 5.212 |
| 2 | 5.204 | 5.175 |
| 3 | 5.254 | 5.197 |
| 4 | 5.176 | 5.198 |

Source: Rstudio Computations

Table 3 and Figure 8 are comparative on actual and predicted values of solar irradiation generated by the XG-Boost model. The fact that the difference between the observed and the predicted values are so close, demonstrates high levels of generalization of the model. The deviations are low and within the acceptable prediction limits which prove the

predictive reliability of the model. Both low variance and high forecasting accuracy indicated by the close clustering of the prediction points around the ideal 1:1 relationship argue in favor of the appropriateness of XG-Boost in predicting solar radiations in the Nigerian weather conditions.

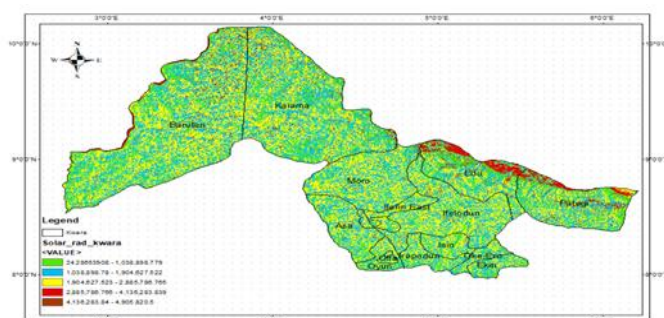


Figure 9: Solar Radiation Map

Table 4: Solar Radiation Classification

| Class | Normalized Annual Solar Radiation (kWh/m ² /year) | DEM Value (average of range) |
|---------------------------|--|---|
| Low Radiation | 1,200 – 1,400 | 24.287 |
| Moderately Low Radiation | 1,401 – 1,600 | 1,038,898.78 |
| Medium Radiation | 1,601 – 1,800 | 1,904,627.52 |
| Moderately High Radiation | 1,801 – 2,000 | 2,885,786.77 |
| High Radiation | 2,001 – 2,200 | 4,521,052.1695 (average of 4,136,283.839 and 4,905,820.5) |

Source: DEM (Arc GIS 10.6) output

Figure 9 and Table 4 combine normalized annual solar radiation values (kWh/m²/year) with outputs derived by the use of DEM to bring terrain-specific estimates of solar radiation in Nigeria in line with standard values of solar radiation. Terrain limitations (shading, poor slope (aspect), and elevation) which are usually located in valleys or depressions are linked to low radiation (1,200 - 1,400 kWh/m²/year). Moderate classes are a reflection of the impact which terrain has had on the variability of solar energy, and higher classes are a sign of the region having good exposure, low topographic obstacles, and good slope. Nevertheless, the results of DEM basically reflect relative spatial patterns that are in need of normalization as they do not take into consideration such atmospheric variables as cloud cover and humidity. As an example, Ilorin is within 1,200-1,800 kWh/m²/year, indicating that although DEM is able to determine the effects of the terrain, it does not include atmospheric factors. As such, although DEM can be utilized in spatial visualization, other models such as ANN and XG-Boost, offer better and more comparative results in terms of solar radiation, thus Table 4 can be used to compare terrain-based and data-driven models.

CONCLUSION

This study compared the predictive abilities of Artificial Neural Networks (ANN) and Extreme Gradient Boosting

(XG-Boost) in predicting the Number of sun rays and meteorological data on long term meteorological data of Nigeria. Correlation, multi-collinearity, diagnostic testing, and predictive accuracy were all included in the analysis to get a clue of the relationship between solar radiation and the major climatic factors. Findings demonstrated that the maximum temperature and sunshine hours are very strong predictors of solar irradiation and multi-collinearity tests indicated relevance of all the chosen predictors. Both of these models represented the nonlinear characteristics of the data, though their performance varied: the ANN model demonstrated the acceptable results but had visible deviations to the ideal prediction, that is, it showed the moderate accuracy, whereas the XG-Boost model demonstrated the high level of performance with lower values of error and a better correlation between the predicted and the observed data. The residual analysis also verified the strengths and trustworthiness of XG-Boost. The analysis of feature importance determined that maximum temperature and relative humidity were the most significant variables, which implied the great role of atmospheric heat and moisture in solar radiation dynamics where solar energy moves larger in regions with good exposure of their terrain. Besides, the analysis of solar radiation in ArcGIS through the use of a DEM showed that topographical parameters, including elevation, slope, and aspect, played an essential role in solar

radiation distribution, with a greater amount of radiation being observed where a terrain has ideal exposure. Nevertheless, the outputs of DEM had to be normalized to standard units, which implies that even though GIS-based models are effective in producing a spatial variation of the desired element, they cannot be used to achieve a more accurate quantitative forecasting as easily as machine learning methods. In general, the results are consistent with the literature and also provide an indication of the benefit of the ensemble learning technique such as XG-Boost, particularly when being used on locally acquired data, which improves the predictive capacity of the models and their applicability to the context. The research comes to the conclusion that XG-Boost is a more precise and reliable framework compared to ANN to predict the solar radiation in Nigeria, whereas the DEM analysis can be considered a useful tool to gain spatial knowledge. Combining machine learning with locally obtained environmental variables provides a good route towards better predicting renewable energy and contributing to plans of sustainable energy. Nevertheless, future research ought to incorporate variables of machine learning models based on DEM (elevation, slope, and aspect) to enhance the accuracy of predictions, and devise universal ways to transform the radiation outputs of DEM models into standardized physical units like kWh/m²/year.

REFERENCES

- Adejumobi, I. A., Oyagbola, R. B., & Akinyele, D. O. (2019). Prospects and challenges of solar energy in Nigeria's energy mix. *Renewable and Sustainable Energy Reviews*, 103, 328–336. <https://doi.org/10.1016/j.rser.2018.12.033>
- Akinyele, D. O., & Rayudu, R. K. (2016). Strategy for developing energy systems for rural electrification in Nigeria. *Renewable and Sustainable Energy Reviews*, 66, 191–204. <https://doi.org/10.1016/j.rser.2016.07.036>
- Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de-Pison, F. J., & Antonanzas-Torres, F. (2016). *Review of photovoltaic power forecasting*. *Solar Energy*, 136, 78–111. <https://doi.org/10.1016/j.solener.2016.06.069>
- Choi, Y., Suh, J., & Kim, S. M. (2019). GIS-based solar radiation mapping, site evaluation, and potential assessment: A review. *Applied Sciences*, 9(9), 1960. <https://doi.org/10.3390/app9091960>
- Firoz, A., & Moghimi, M. (2021). Meteorological predictors of solar radiation: A PCA approach. *Energy Reports*, 7, 1229–1240. <https://doi.org/10.1016/j.egyr.2021.02.015>
- Ghimire, S., Deo, R. C., Downs, N. J., & Raj, N. (2019) (2019). Global solar radiation prediction using XGBoost and random forests algorithms. *Renewable and Sustainable Energy Reviews*, 111, 446–456. <https://doi.org/10.1016/j.rser.2019.05.025>
- Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688. <https://doi.org/10.1016/j.ijforecast.2006.03.001>
- Koizumi, K. (1999). Forecasting precipitation systems. *Journal of the Meteorological Society of Japan*, 77(1), 145–165.
- Kong, W., Dong, Z. Y., Jia, Y., Hill, D. J., Xu, Y., & Zhang, Y. (2017). Short-term load forecasting based on long short-term memory network. *IEEE Transactions on Smart Grid*, 10(1), 841–851. <https://doi.org/10.1109/TSG.2017.2753802>
- Kumar, A., & Sharma, R. (2020). Artificial neural network and XGBoost for solar radiation forecasting. *Journal of Cleaner Production*, 276, 124733. <https://doi.org/10.1016/j.jclepro.2020.124733>
- Law, R. (2000). Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting. *Tourism Management*, 21(4), 331–340. [https://doi.org/10.1016/S0261-5177\(99\)00067-9](https://doi.org/10.1016/S0261-5177(99)00067-9)
- Mahmoud Sabo, L., Sambo, A. S., & Bala, E. J. (2016). Estimation of global solar radiation using digital elevation model (DEM). *World Journal of Engineering*, 13(5), 453–460. <https://doi.org/10.1108/WJE-08-2016-0063>
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5(4), 115–133.
- Olusegun, O. O., Akinwale, Y. O., & Ojo, O. O. (2020). Solar energy forecasting in Nigeria: A review. *Energy Policy*, 142, 111567. <https://doi.org/10.1016/j.enpol.2020.111567>
- Oyedepo, S. O., Adaramola, M. S., & Paul, S. S. (2015). Analysis of wind speed data and wind energy potential in three selected locations in south-east Nigeria. *International Journal of Energy and Environmental Engineering*, 6, 147–157. <https://doi.org/10.1007/s40095-015-0161-6>
- Rahman, M. M., Islam, M. R., & Hasanuzzaman, M. (2020). Data-driven solar radiation forecasting using machine learning models. *Applied Energy*, 268, 114999. <https://doi.org/10.1016/j.apenergy.2020.114999>
- Rehman, S., Mohandes, M., & Halawani, T. O. (2017). Artificial intelligence models for global solar radiation prediction. *Solar Energy*, 144, 729–744. <https://doi.org/10.1016/j.solener.2017.01.037>
- REN21. (2021). Renewables 2021 global status report.
- Sampaio, P. G. V., & González, M. O. A. (2017). Photovoltaic solar energy: Conceptual framework. *Renewable and Sustainable Energy Reviews*, 74, 590–601. <https://doi.org/10.1016/j.rser.2017.02.081>
- Samsudin, R., Saad, P., & Shabri, A. (2010). River flow forecasting using hybrid models. *Hydrology and Earth System Sciences*, 14(10), 2053–2067. <https://doi.org/10.5194/hess-14-2053-2010>
- Sharma, R., Goel, S., & Gupta, A. (2018). Solar energy development and challenges in India. *Renewable and Sustainable Energy Reviews*, 82, 994–1004. <https://doi.org/10.1016/j.rser.2017.09.075>
- Yadav, A. K., & Chandel, S. S. (2014). Solar radiation prediction using artificial neural network techniques: A review. *Renewable and Sustainable Energy Reviews*, 33, 772–781. <https://doi.org/10.1016/j.rser.2014.02.016>
- Zhang, G. P., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, 14(1), 35–62. [https://doi.org/10.1016/S0169-2070\(97\)00044-7](https://doi.org/10.1016/S0169-2070(97)00044-7)

