Machine Learning Models for Prediction of Meteorological Variables for Weather Forecasting

1 **ABSTRACT**

2 This study trained six machine learning models to predict meteorological variables at a

3 tropical location. The models used are: Multiple linear regression, Decision tree, Random

4 forest, Support vector machine, Extreme gradient boosting and Multilayer perceptron. This

5 was with the aim of determining the best machine learning model for weather forecasting in a

6 tropical location. The meteorological variables that were predicted are: Temperature, Solar

7 radiation, Relative humidity and Wind speed. To identify the efficiency and to quantify the

8 predictive capacity of each models, evaluation metrics such as coefficient of determination

9 (R^2), mean absolute error (MAE), mean absolute percentage error (MAPE) and mean square

10 error (RMSE) were employed. The best performed model for temperature is the Random

11 Forest which has R^2 of 0.93, MAE of 0.78 ^{0}C, MAPE of 2.84 % and RMSE of 1.13 ^{0}C. Also,

12 the best performed model for solar radiation is the Random Forest having an R^2 value of 0.72,

13 MAE of 85.34 W/m^2 and RMSE of 19008.45 W/m^2. For relative humidity, Random Forest also

14 has the best performance. From the evaluation metrics, it has R^2 of 0.92, MAE of 3.41 %,

15 MAPE of 0.75 % and RMSE of 24.71 %. The best performed technique for predicting the wind

16 speed was also the Random Forest having an R^2 value of 0.79, MAE of 0.16 m/s and RMSE

17 of 0.044 m/s. The study concluded that the best machine learning model for predicting

18 meteorological variables in a tropical location is the Random Forest.

19
20 *Keywords: Machine Learning, Predicting models, Meteorological variables, Weather*

21 *forecasting*

22 **1. INTRODUCTION**

Meteorological variables are important for the interpretation of physical processes in the lower atmosphere, agricultural meteorology, monitoring and prediction of weather and climate and for the management of natural resources [1]. These variables are critical for making short-term and long-term decisions on activities such as monitoring of extreme weather events, farming response, early warning of pests and diseases, and so on.

Weather which is the short-term condition of the atmosphere is characterised by wind, temperature, humidity and solar radiation variables forced by radiative fluxes, surface latent and sensible heat fluxes. Climate is the long-term atmospheric condition of a specific location over a long period of time, usually 30 years. It is a continuous, data-intensive, multidimensional, dynamic and chaotic process. These characteristics make weather/climate forecasting a difficult task. As a result, consistent and high-quality observations of climatic variables are critical [2].

Weather forecasting is the use of science and technology to predict the state of the atmosphere at a specific area. Weather predictions are formed by collecting quantitative data about the current condition of the atmosphere and projecting how the atmosphere will evolve using scientific understanding of atmospheric dynamics.

With the current global climate change, there is a need to develop a dependable model capable of accurately capturing fluctuations in weather variables. Weather variables are typically modeled using computational, numerical, and statistical techniques, the majority of which are nonlinear [3]. Many researchers have developed statistically based models for predicting meteorological time series of weather variables [4,5]. The challenge is attributed to the obvious ambient stochastic variables, as well as the fact that future returns cannot be projected with adequate precision when modeling such high-uncertainty conditions [6].

Weather forecasting methods such as numerical weather prediction model, ensemble forecasting, among others, rely on sophisticated physical models and equations however, machine learning models provide a more data-driven approach, recognizing patterns and links in historical weather data to estimate future conditions.

Machine learning is a sub-field of artificial intelligence that focuses on creating algorithms and techniques that allow computers to learn and make predictions without being explicitly

programmed. It entails the study of algorithms and statistical models that enable computers to learn from and predict or act on data. The fundamental principle behind machine learning is to allow computers to learn patterns or correlations from data and then generalize that knowledge to make predictions or decisions on new, previously unseen data. Machine learning algorithms, rather than following a fixed set of rules, learn iteratively from instances or experiences, continuously improving their performance over time [7]. Machine learning has numerous applications in fields such as image [8] and speech recognition, recommendation systems, autonomous vehicles [9], finance, healthcare, weather forecasting [10,11], and many more. It has transformed numerous industries and continues to grow rapidly as massive datasets, improved computer power, and breakthroughs in algorithms and methodologies become available. Machine learning algorithms have shown potential in enhancing weather prediction efficiency and natural disaster forecasts, which may aid in disaster preparedness and response operations [12].

The aim of this study is to predict meteorological variables with selected machine learning models and evaluate their performances in order to identify the best performing model for weather prediction in a tropical location.

Table 1 shows similar studies of using machine learning algorithms for weather and climate prediction.

Table 1: Similar studies of using machine learning algorithms for weather and climate prediction

Authors	Research Topics	Models	Tools
Anton *et al.* [13]	Collaborative data mining in agriculture for prediction of soil moisture and temperature	k-nearest neighbor model (k-NN) local polynomial regression (LPR) neural net model (NN) and support vector machine (SVM)	The data analysis was conducted with the aid of a SQL command and a Microsoft Access database
Cortez and Morais [14]	A data mining approach to predict forest fires using meteorological data	Multiple regression (MR) Decision trees (DT) and Random forests (RF) Neural networks (NN)	The open-source library Rainer (for the R statistical environment) was used.

		Support vector machines (SVM)	
Joshi *et al.* [15]	Weather forecasting and climate changing using data mining application	Decision tree classifiers	Both decision trees and decision tree rules were created using the See5 program
Olaiya and Adeyemo [16]	Application of data mining techniques in weather prediction and climate change studies	Artificial neural network Decision tree algorithm	C5 Decision Tree classifier algorithm using the See5 was implemented
Oladipo *et al.* [17]	Prediction and analysis of student performance by data mining in WEKA	Classification Association	WEKA tool was used as the software for data mining
Segovia et al [18]	Meteorological variables forecasting system using Machine Learning and open-source software	Multiple regression (MR) Polynomial regression, Decision trees (DT) and Random forests (RF) XGBoost, multilayer perceptron neural network (MLP)	Python open-source software
Shivang and Sridhar [19]	Weather Prediction for Indian location using machine learning	Linear regression Functional regression Neural network	Python
Zaman [20]	Machine learning model on rainfall - a predicted approach for Bangladesh	Classification algorithms (Naive Bayes, random forest classifier, and decision tree algorithm)	The machine learning library was Apache Spark

		Regression algorithm (linear regression, random forest regression)	

2. METHODOLOGY

2.1 Design and Prediction Models for Meteorological Variables

In this research, six machine learning models were used to predict the following meteorological variables: temperature, solar radiation, wind speed and relative humidity. The models that were used are: Multiple Linear Regression, Decision Tree, Random Forest, Support Vector Machine, Extreme Gradient Boosting, and Multilayer Perceptron.

Evaluation metrics such as Coefficient of Determination (R^2), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and the Root Mean Square Error (RMSE) were used to identify the best performing algorithm. To predict the meteorological variables, the design methodology shown in Figure 1 was performed.

2.2 Data Acquisition

The meteorological data used for the forecasting models were collected from the Obafemi Awolowo University Meteorological Station (7.53 ^{0}N; 4.54 ^{0}E), Nigeria. The sensors employed for the measurements of the meteorological parameters were mounted on a 6-metre meteorological mast. A wind cup anemometer (034B) was installed at a height of 6 metres to measure wind speed. A temperature and relative humidity probe (HMP45) was mounted at a height of 4 metres to measure air temperature and relative humidity. At a height of 2 metres, a pyranometer (CS300) was mounted for the measurement of incoming solar radiation.
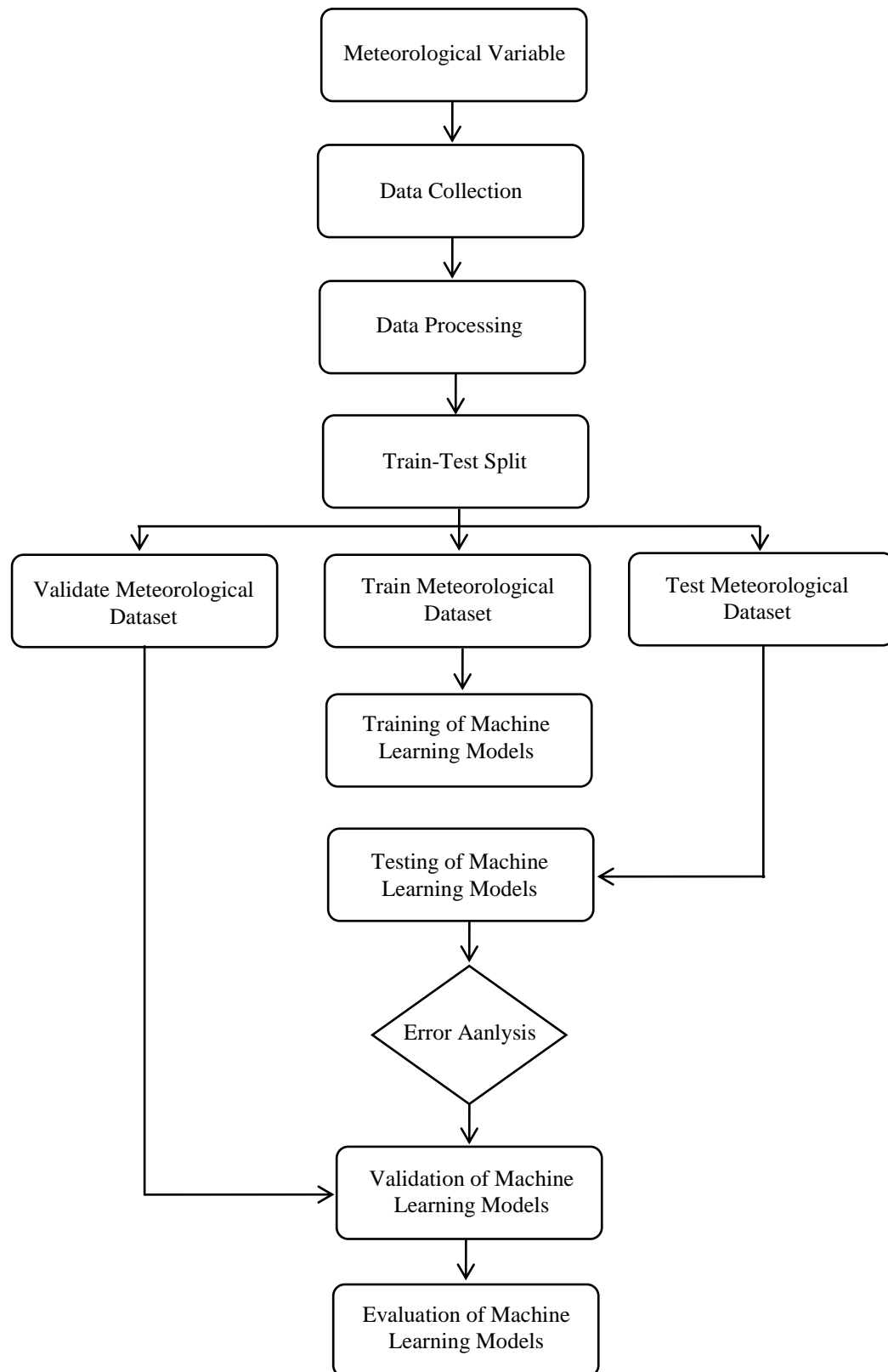
```
┌─────────────────────────┐
│ Meteorological Variable │
└─────────────────────────┘
            │
            ▼
┌─────────────────────────┐
│     Data Collection     │
└─────────────────────────┘
            │
            ▼
┌─────────────────────────┐
│     Data Processing     │
└─────────────────────────┘
            │
            ▼
┌─────────────────────────┐
│     Train-Test Split    │
└─────────────────────────┘
```

Validate Meteorological Dataset   Train Meteorological Dataset   Test Meteorological Dataset

Training of Machine Learning Models

Testing of Machine Learning Models

Error Aanlysis

Validation of Machine Learning Models

Evaluation of Machine Learning Models

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125 **Figure 1: Flowchart of the Design and Implementation of the Prediction Models for Meteorological**

126

127 **2.3 Data Processing**

128 A one-year data (from 1 January, 2020 to 31 December, 2021) was obtained from the station.

129 The data was sampled every 10 seconds and saved as 1-minute averaged values before

130 being reduced to produce 30-minute statistics. Following data preprocessing, a total of 17,284

131 data points for each <mark>variable</mark> (temperature, relative humidity, wind speed, and solar radiation)

132 were produced.

In forecasting, it is important to ensure that there are no missing data points in the measurements or to execute a data filling method. In this study, a Python algorithm was used to compute the average of the existing list of data points and automatically fill up the missing data points.

2.4 Division of Data

The database was divided into three categories to ensure that the models perform properly: training set, test set, and validation set. The first, as the name implies, was used to train the forecasting models, the second to evaluate the test set, and the third to validate each of the implemented models. From the total of 17,284 data points obtained for each variable, with 80% of the database (13,825 data points) used to train the models, 20% (3,457 data points) used to test the models, and 5 days (210 data points) used to validate the models.

2.5 Selected Machine Learning Models

2.5.1 Multiple Linear Regression

Multiple Linear Regression (MLR) is a statistical method used to model the relationship between a dependent variable and two or more independent variables. The modeled variables are called the predicted or dependent variables (y), while the independent variables are known as predictors, or features (X) [21]. The general form of the multiple linear regression model is:

$$y = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n \tag{1}$$

where X_1, X_2, …, X_n are the independent variables; b_1, b_2, …, b_n are the coefficients representing the relationship between the independent variables and the dependent variable; a is the constant of the relationship between the dependent and independent variable; and y is the predicted or dependent variable.

2.5.2 Decision Tree

A decision tree is a flow chart that operates by recursively splitting the dataset into subsets based on input feature values. The aim is to build a tree-like model in which each internal node represents a decision based on a feature, each branch indicates the decision's outcome, and each leaf node represents the final forecast. The mean or mode of the responses of the training dataset which are within the new dataset is used for prediction [22]. Gini impurity is

often employed as a criterion for separating nodes during tree construction. The Gini impurity for a node is determined by the probability of each class being present in that node. The equation for a Gini impurity is represented by:

$$G_i = 1 - \sum_{k=1}^{m}(P_{i,k})^2 \qquad (2)$$

where G_i is the Gini impurity; m is the number of class and $P_{i,k}$ is the probability of class i, given node k.

2.5.3 Random Forest

Random Forest is an ensemble learning method that integrates predictions from numerous decision trees to improve the model's overall performance and robustness. Random Forest trains multiple decision trees independently on a random portion of the data. This is accomplished using bootstrapping, which is sampling with replacement. As a result, each tree has a different subset of the data. Decision trees are prone to overfitting, but random forests circumvent this by creating random selections of data and using those subsets to construct smaller trees. The error for random forest is determined by the strength of the individual generated trees and their correlation [23].

2.5.4 Support Vector Machine

Support vector machines are supervised learning models that use learning techniques to examine data for classification and regression. To categorize unlabeled data, support vector clustering algorithm uses the statistics of support vectors obtained in the support vector machines method. These data sets necessitate unsupervised learning algorithms that seek natural clustering of data into groups and then map additional data to these clusters. The aim of the SVM algorithm is to determine the best hyperplane in an N-dimensional space that can split data points into different classes in the feature space. The hyperplane attempts to maximize the margin between the closest points of various classes. The size of the hyperplane is determined by the number of features.

2.5.5 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is an effective algorithm for regression and classification problems. XGBoost performs a second-order Taylor expansion on the loss function, incorporates a regularization term to avoid overfitting and improves the generalization performance of the model. The algorithm works by consecutively building a

193 succession of decision trees, and combining their predictions to form a strong prediction
194 model. In each iteration, XGBoost fits a new tree to the residuals (the difference between the
195 actual and predicted values) of the previous set, focusing on minimizing both the loss function
196 and the term regularization.

197 **2.5.6 Multilayer Perceptron Neural Network**

198 A Multilayer Perceptron (MLP) is a type of artificial neural network distinguished by its layered
199 architecture, consisting of several layers of fully interconnected neurons which includes an
200 input layer, one or more hidden layers, and an output layer. Figure 2 shows the structure of a
201 multilayer perceptron neural network. The input layer is made up of n units that distribute the
202 input signals to the next layer. The hidden layer is made up of neurons k, that have no
203 physical contact with the outside; and the output layer is made up of m eurons whose outputs
204 comprise the vector of external outputs of the multilayer perceptron.

205 The neural network is trained by calculating the linear combination of a set of input variables
206 with a bias term, then applying an activation function, typically the threshold or sign function,
207 to produce the network output. Thus, the network weights are modified using the supervised
208 learning by error correction (back propagation) approach, so that the predicted output is
209 compared to the value of the output variable to be acquired, with the difference being the
210 error or residual. Each neuron acts independently of the others: each neuron gets a set of
211 input values (an input vector), computes the scalar product of this vector and the vector of
212 weights, adds its own bias to the result, and returns the final result obtained [24].

213
214
215
216
217 Input Layer   Hidden Layer   Output Layer

223

224

225

226
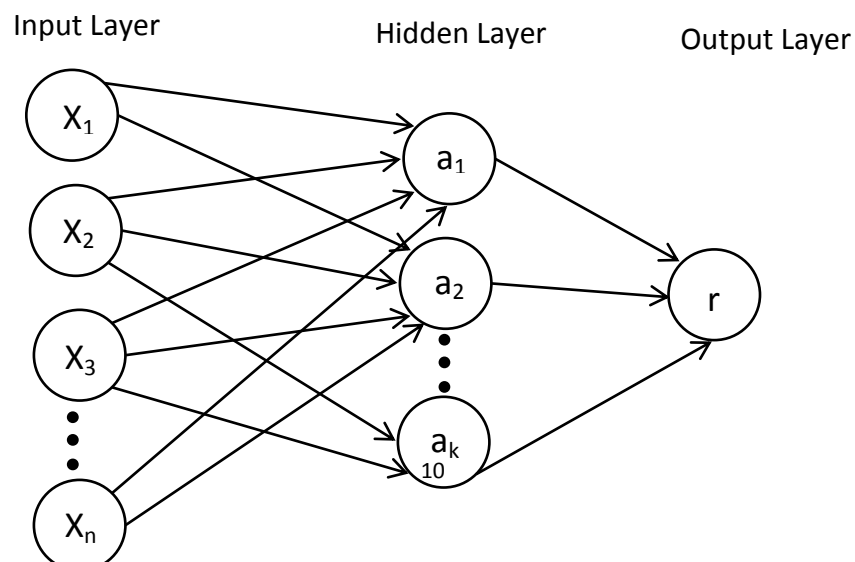
Figure 2: Structure of a Multilayer Perceptron Neural Network

227

228 After carrying out heuristics testing on the above mentioned models, the best tuning

229 parameters for each variable are listed in Table 2.

230

231 **Table 2: Tuning Parameters for the Different Machine Learning Models**

Predicted Variables	Input Variables
Temperature	Solar radiation, Relative humidity, Wind speed
Solar radiation	Temperature, Relative humidity, Wind speed
Wind speed	Temperature, Relative humidity, Solar radiation
Relative humidity	Temperature, Wind speed, Solar radiation

232

233

234 **2.6 Metrics for Accessing the Performances of the Selected Machine Learning**

235 **Models**

236 In order to determine the forecasting accuracy of the weather models i.e. to identify the model

237 that is more efficient in prediction, evaluation metrics such as the mean square error (RMSE),

238 mean absolute percentage error (MAPE), mean absolute error (MAE) were employed. To

239 determine if the models perform well in training and to quantify their predictive capacity, the

240 coefficient of determination (R^2) was used.

241 **2.6.1 Root Mean Square (RMSE)**

242 Root mean Square (RMSE) is used to measure the average magnitude of the errors between

243 the predicted values and actual values. It measures the spread of errors, with lower values

244 suggesting better model performance.

245 $$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad\qquad (3)$$

where: n is the number of data points; y_i is the actual (observed) value for the i-th data point and $\hat{y}_i$ is the predicted value for the i-th data point.

2.6.2 Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) is used to find the average of the percentage errors between the predicted and actual values. It is given by:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{4}$$

where: n is the number of data points; y_i is the actual (observed) value for the i-th data point and $\hat{y}_i$ is the predicted value for the i-th data point.

One of the major limitations of using MAPE is that when using actual numbers close to or at 0, the MAPE score will be off by a factor of 0 or excessively high. As a result, it is not suggested to use MAPE when the real values are near to 0.

2.6.3 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) calculates mean of the absolute differences between the actual and predicted values. It indicates how far the predictions are from the actual values on average. A lower MAE suggests more accuracy because it indicates that the model's predictions are closer to the actual values.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5}$$

where: n is the number of data points; y_i is the actual (observed) value for the i-th data point and $\hat{y}_i$ is the predicted value for the i-th data point. When compared to other error metrics, such as Mean Squared Error (MSE), MAE is less sensitive to outliers.

2.6.4 Coefficient of Determination (R^2)

The Coefficient of Determination (R^2) is a statistical measure that examines how much of the variance in the dependent variable is explained by the independent variables in a regression model. In other words, it assesses the model's goodness of fit. Its formula as written in equation (6) is described by 1 minus the ratio of the sum of the squared differences between

271 the observed values and the predicted values to the sum of the squared differences between

272 the observed values and the mean of the observed values.

273
$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{6}$$

274 where y_i are the observed values, $\hat{y}_i$ are the predicted values and $\bar{y}$ is the mean of the

275 observed values.

276 The values of R^2 range between 0 and 1. R^2 of 1 indicates that the model predicts the

277 dependent variable completely based on the independent variables. A R^2 of 0 implies that the

278 model has no explanatory power.

279 **3. RESULTS AND DISCUSSION**

280 **3.1 Temperature Prediction**

281 The machine learning models that were used to predict temperature and the metrics used to

282 evaluate the performances of each models are shown in Table 3.

283 The coefficient of determination, R^2 for the models, Multiple linear regression, Decision tree,

284 Random Forest, Support Vector Machine, Extreme Gradient Boosting and Multilayer

285 Perceptron are close to 1 (0.86, 0.89, 0.93, 0.86, 0.91 and 0.91 respectively). These obtained

286 values of R^2 indicate that the predicted values obtained from the models show close

287 agreements with the actual values. Thus, implying that the models are good fits for estimating

288 temperatureat the study location.

289 The MAE indicates that all the models' predictions are closer to the actual values, with

290 Random Forest having the best accuracy. The MAPE indicates the percentage of accuracy

291 prediction of the models. Random Forest has the least MAPE which implies that it has the

292 least percentage of errors. The RMSE values produced by the models show that the models

293 produced relatively low values of scatter points as indicated in Figure 3 with Random Forest

294 having the least scatter points and MLR having the largest scatter points. This shows that the

295 models performed well in predicting temperature.

296

297 **Table 3: Evaluation Metrics for Temperature Prediction**

Models	Coefficient of Determination (R^2)	Mean Absolute Error (MAE) [^{0}C]	Mean Absolute Percentage Error (MAPE) [%]	Mean Square Error (RMSE) [^{0}C]
Multiple linear	0.8583	1.17	4.45	2.2148

	regression			
Decision tree	0.8866	0.93	3.39	1.8121
Random forest	0.9293	0.78	2.84	1.1298
Support Vector Machine	0.8613	1.12	4.33	2.2638
Extreme Gradient Boosting	0.9140	0.85	3.10	1.3740
Multilayer perceptron	0.9098	0.94	3.52	1.4406

298

299   Random Forest is the best performing technique for predicting the temperature variable

300   having the highest R^2, lowest MAE, lowest MAPE and the least RMSE. The least performed

301   technique is the Multiple Linear Regression which has the lowest R^2, highest MAE, highest

302   MAPE and highest RMSE.

303   Figure 4 shows the time series plot of the actual (black) and the predicted (red) values of

304   temperature for a representative of five days, using the different Machine Learning techniques.

305   The Figure validates that the best performing technique is the Random Forest.

306

307

308

309

310

311

312

313

314

315

316
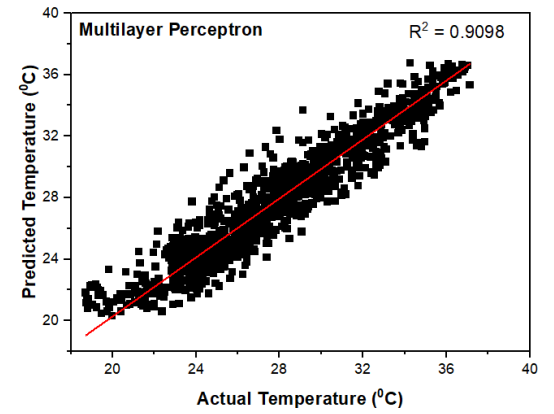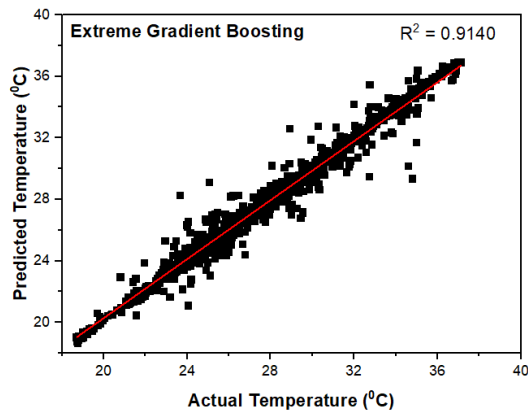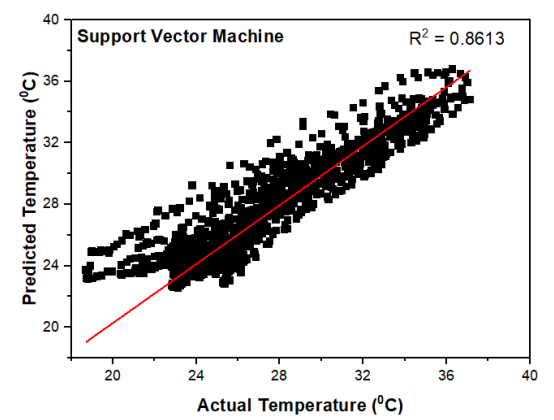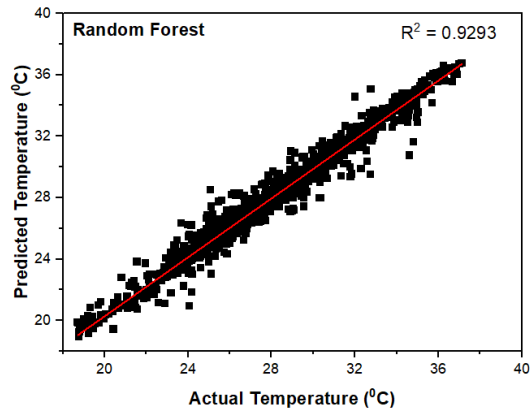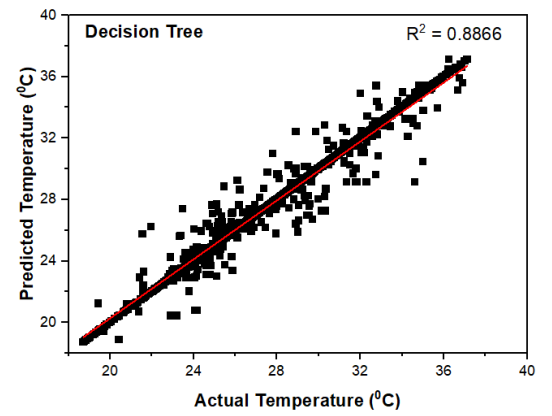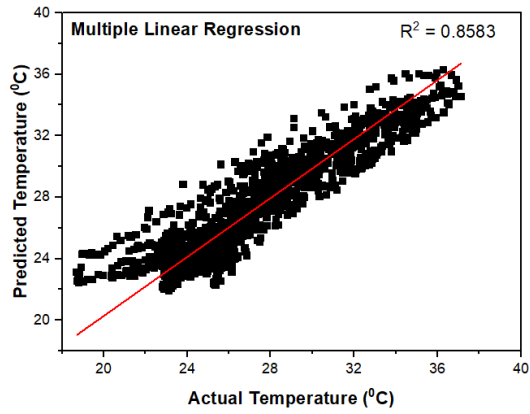
317
318
319
320
321
322
323
324

15

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340 **Figure 3: Scatter Plots of the Actual and Predicted Temperature using the Different Models**

341

342

343

344

345
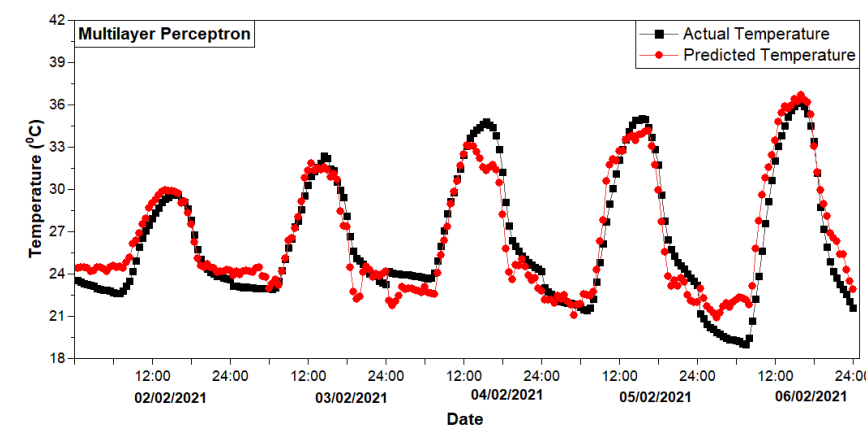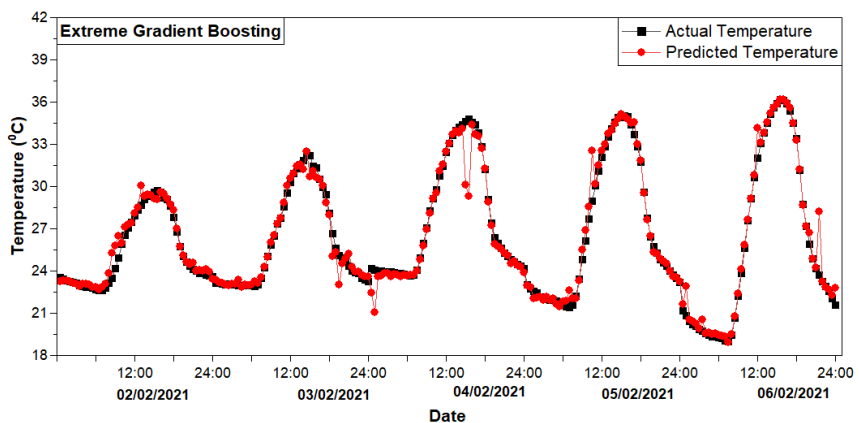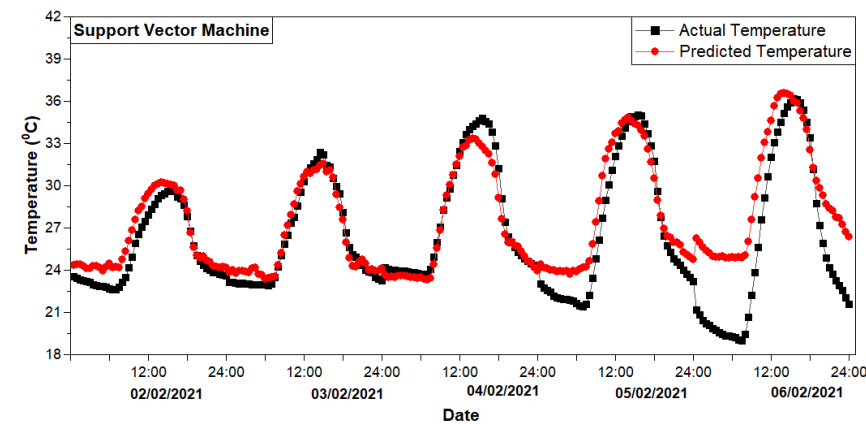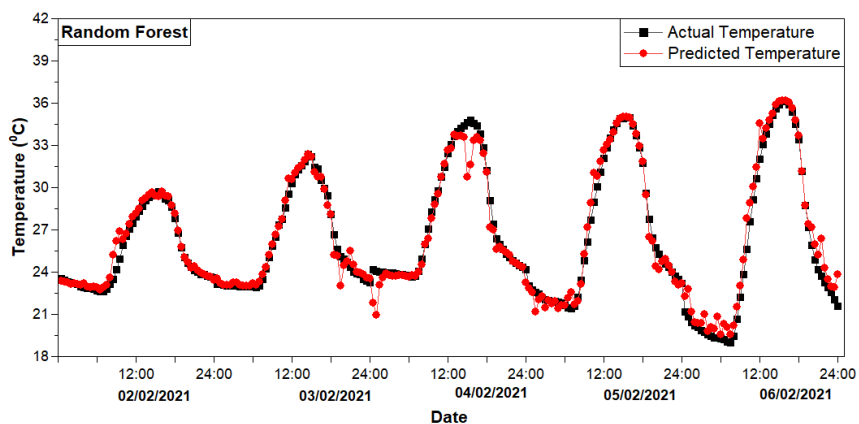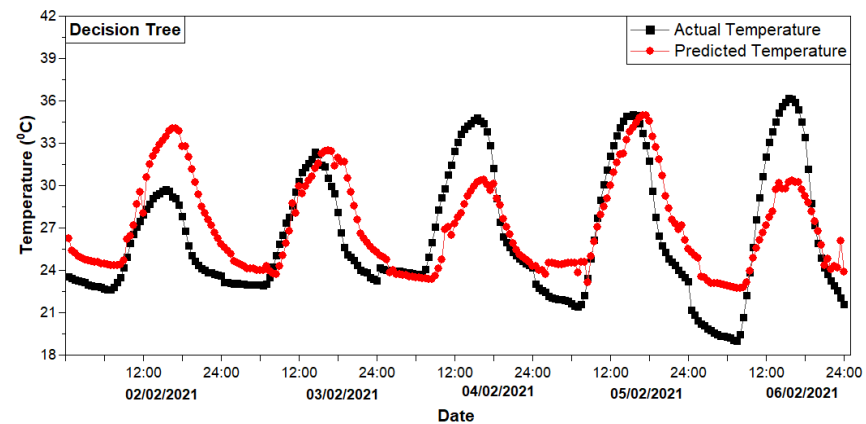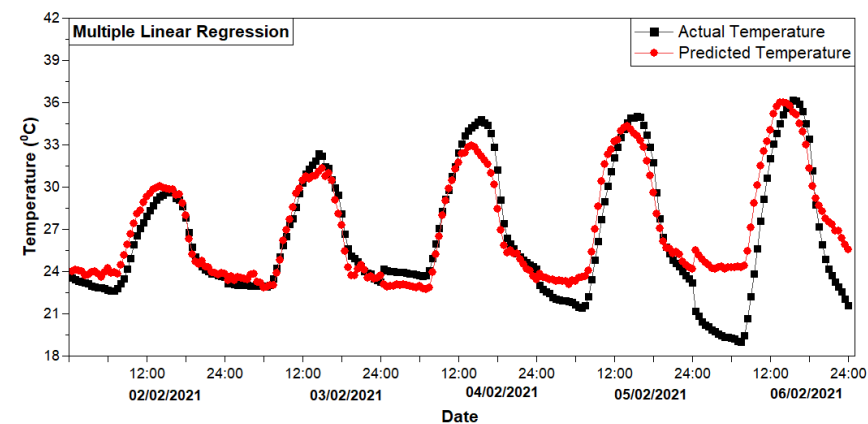
Figure 4: Time Series Plots of the Actual and Predicted Temperature using the Different Models

3.2 Solar Radiation Prediction

Table 4 shows the machine learning models and the metrics used for evaluating the performances of the models for the prediction of solar radiation. The obtained value of coefficient of determination, R^2 for Random forest and Multilayer perceptron are 0.72 and 0.70 respectively. The value of R^2 for Multiple linear regression, Decision tree, Support Vector Machine and XGboostare 0.66, 0.51, 0.67 and 0.67 respectively. These values indicate that Multiple linear regression, Decision tree, Support Vector Machine and XGboost do not perform well in their training, hence their predictive ability for solar radiation variable is low.Random forest and Multilayer perceptron showed relatively close agreements with the actual values. Thus, implying that the models are relatively good fits for estimating solar radiation. The MAE values obtained from the models are large numbers which shows that the models have large errors. The RMSE values show large scatter points as shown in Figure 5. The large numbers obtained for MAE and RMSE are due to the negative values present in the actual values of solar radiation. These negative values which are obtained in the early hours of the morning and late hours of the evening are results of the radiative cooling of the earth's surface [25]. Using the evaluation metrics for the performance of the models, Random Forest is the best performed technique, seconded by Multilayer perceptron while Decision Tree is the least performed technique.

Figure 6 shows the time series plot of the actual (black) and the predicted (red) values of solar radiation for a representative of five days, using the different Machine Learning techniques. The figure validates that the best performing technique is the Random Forest.

Table 4: Evaluation Metrics for Solar Radiation Prediction

Models	Coefficient of Determination (R^2)	Mean Absolute Error (MAE) [W/m^2]	Mean Square Error (RMSE) [W/m^2]
Multiple linear regression	0.6626	109.78	21427.0687
Decision tree	0.5050	836.64	31436.6249
Random forest	0.7220	85.34	19008.4522
Support Vector Machine	0.6694	108.36	20995.9538
Extreme Gradient Boosting	0.6740	87.97	20703.7786
Multilayer perceptron	0.7007	88.33	17655.6951
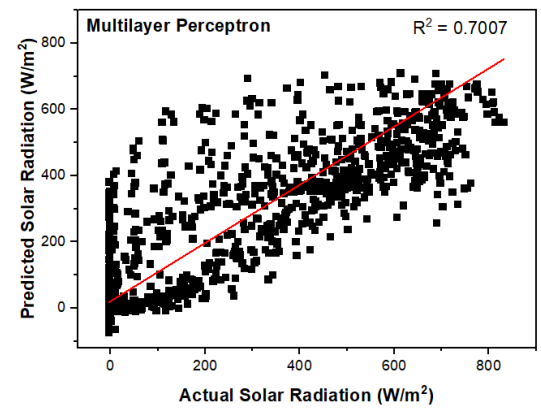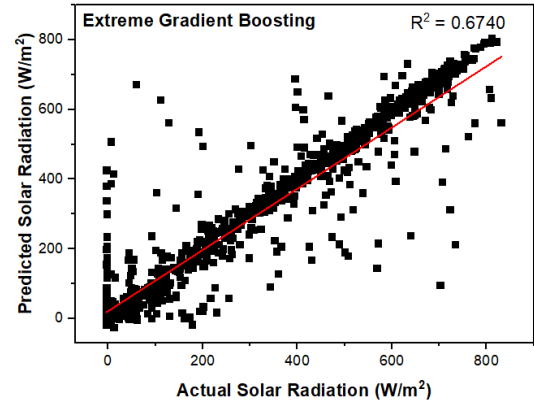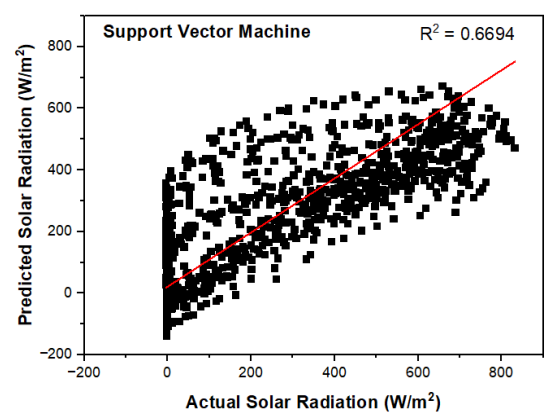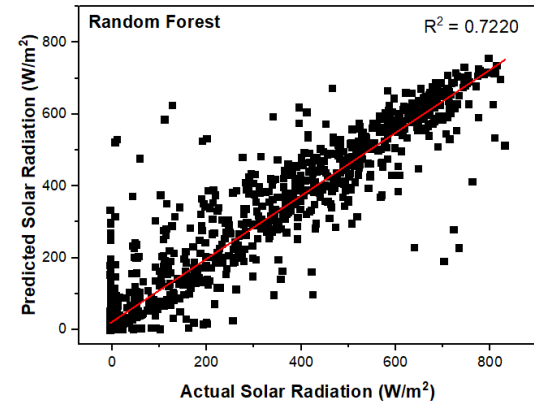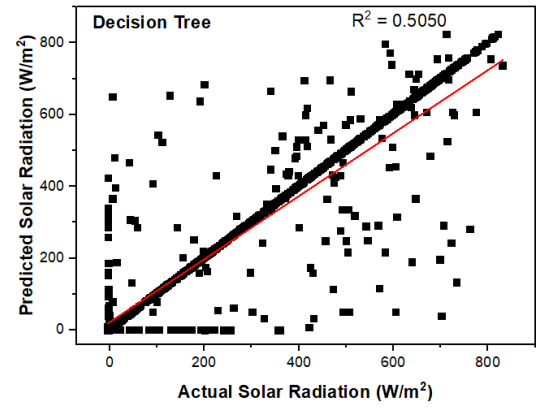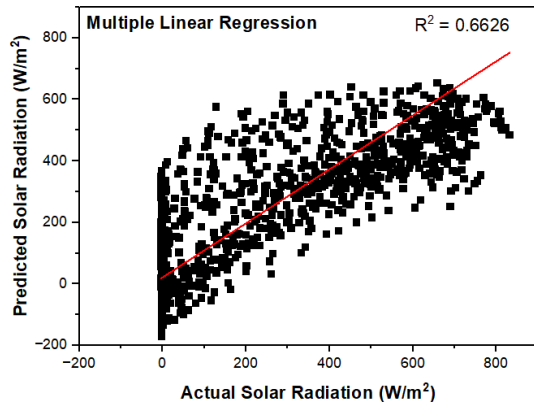
Figure 5: Scatter Plots of the Actual and Predicted Solar Radiation using the Different Models
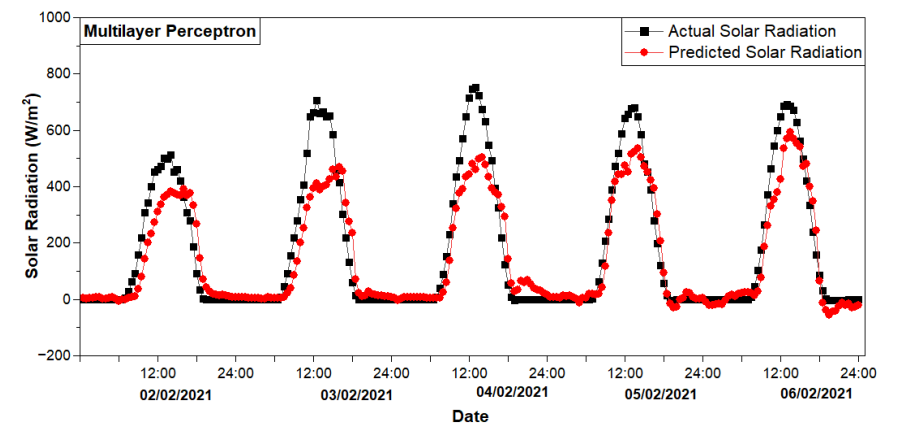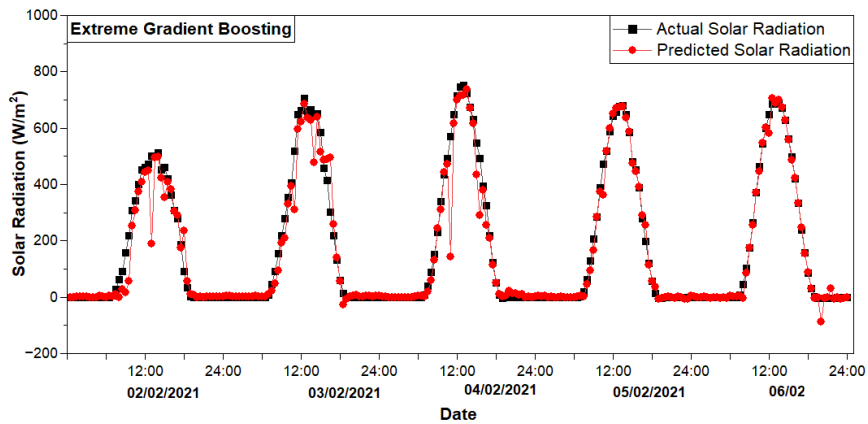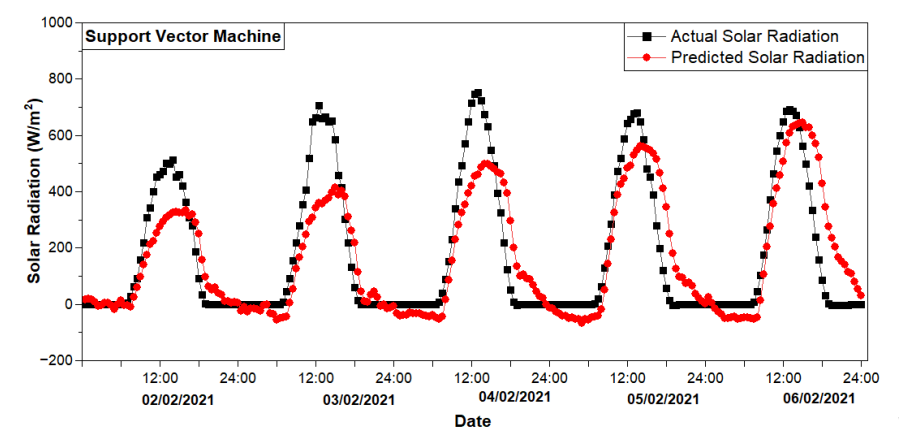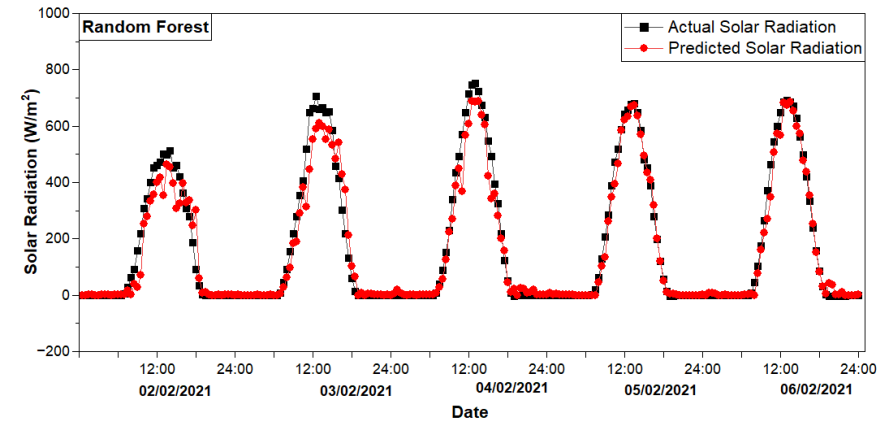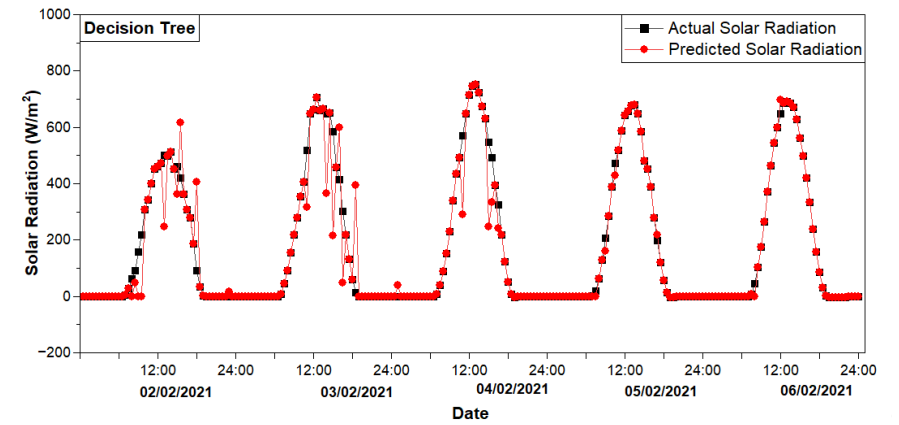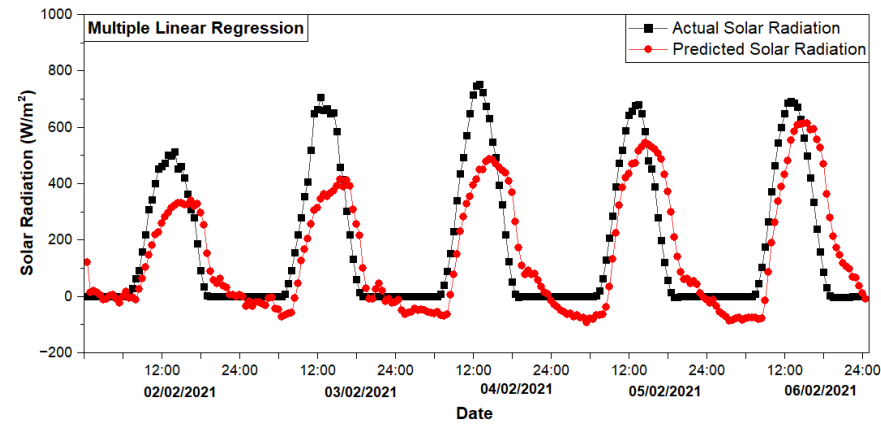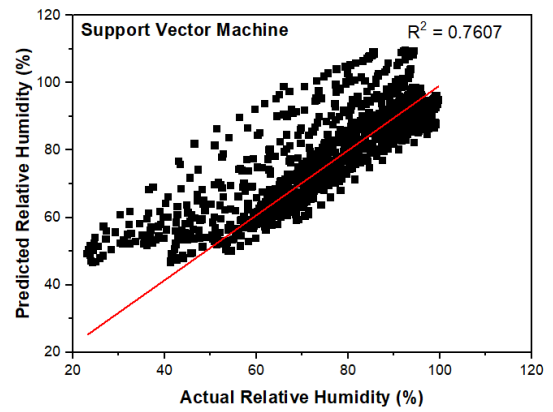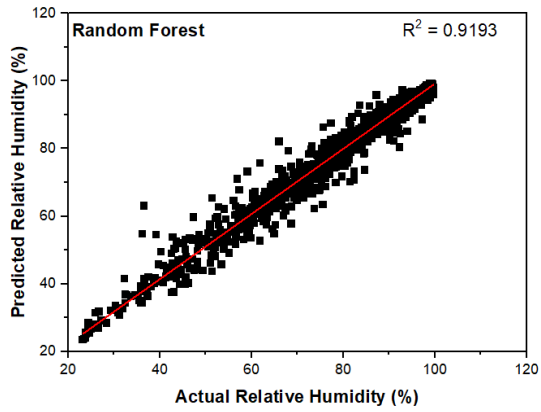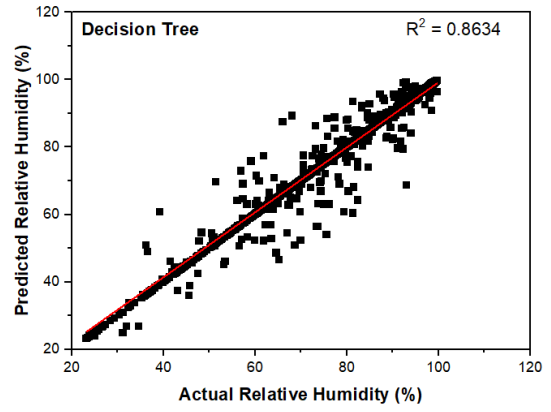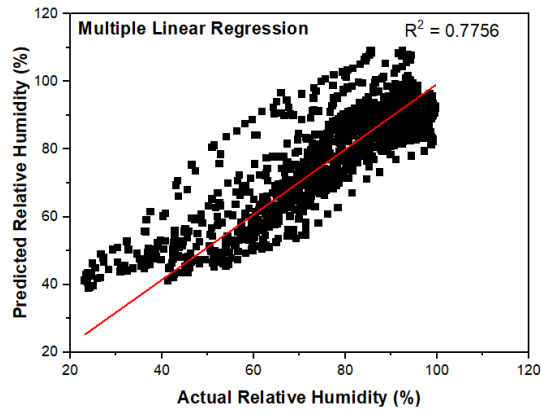
Figure 6: Time Series Plots of the Actual and Predicted Solar Radiation using the Different Models
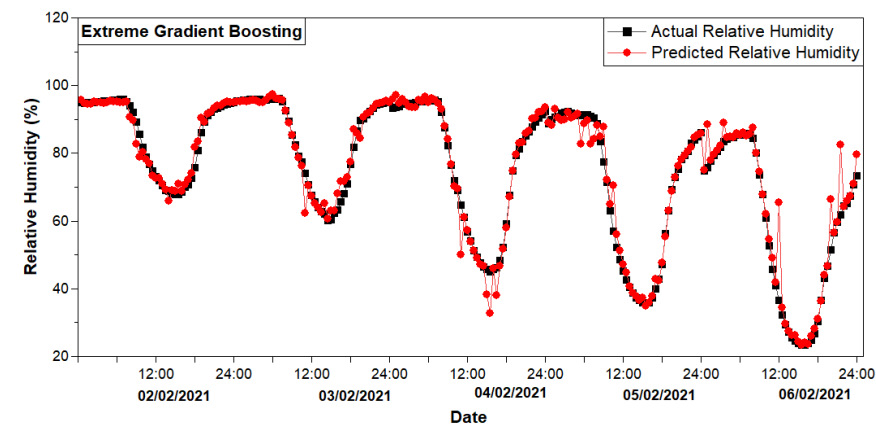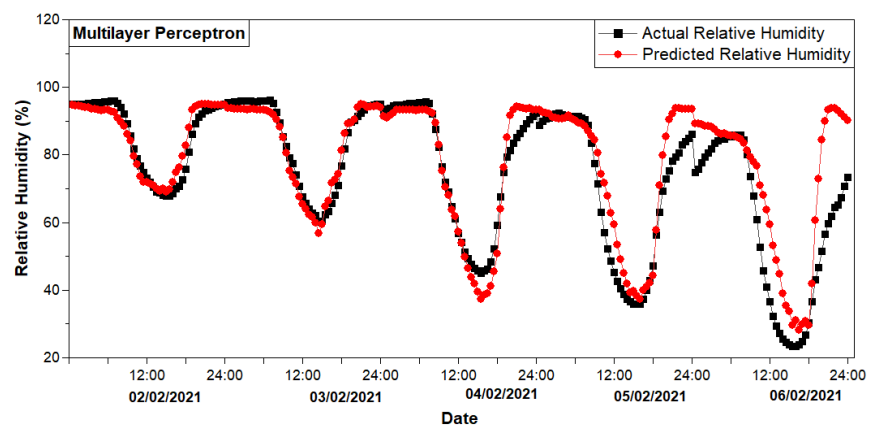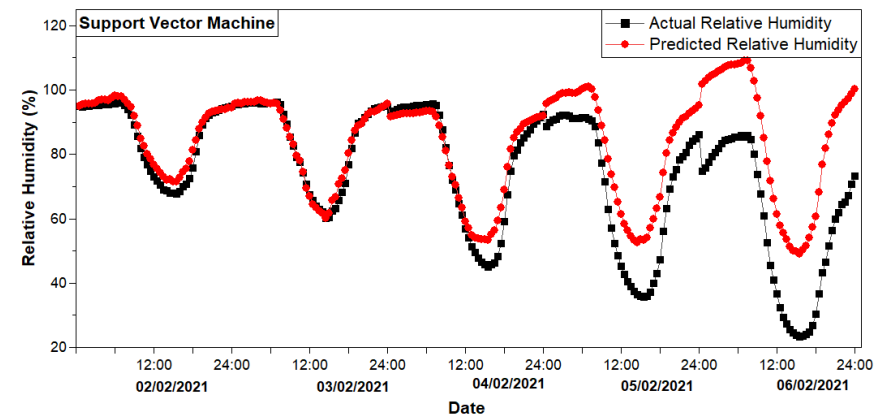
3.3 Relative Humidity Prediction

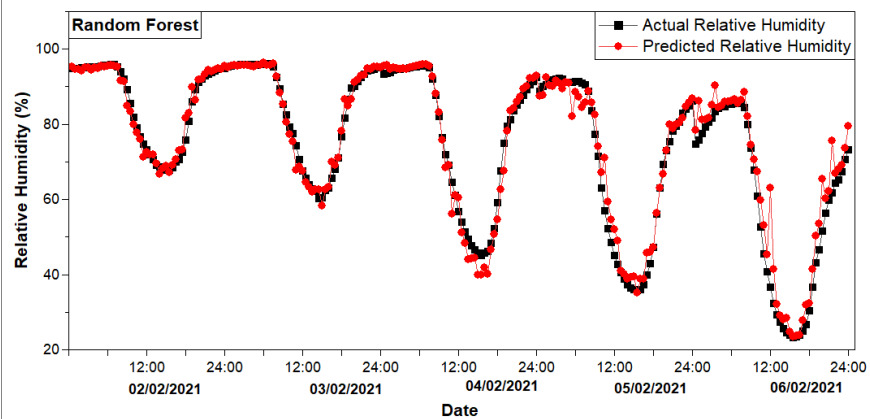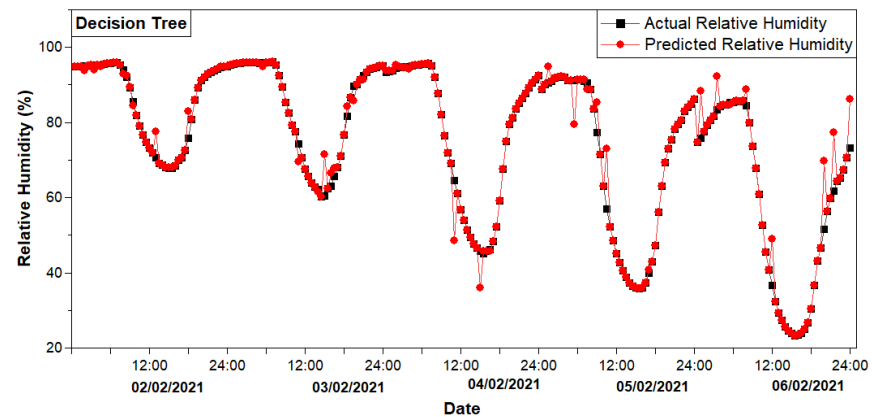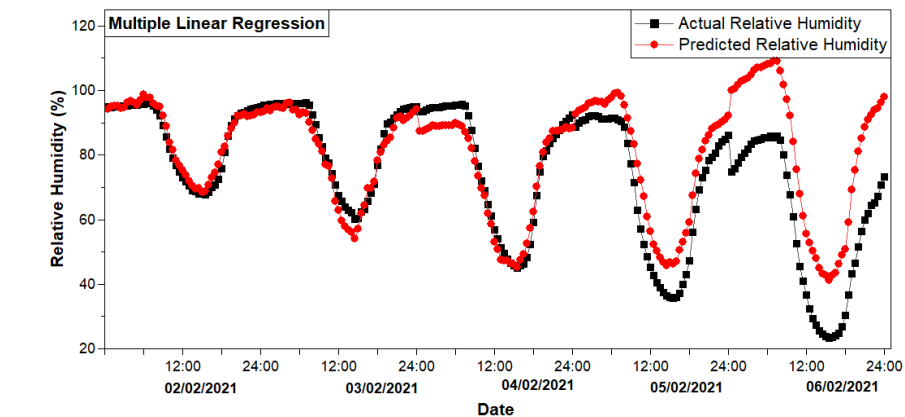Table 5 shows the performances of the different models used to predict the relative humidity. The coefficient of determination (R^2) for the models are close to 1 which implies that the trained models have good fittings with the actual values. The MAE values show that the models do not have large errors and MAPE shows the percentage of the errors with Random Forest having the least MAPE. The RMSE shows that the models have moderate scatter points (as shown in Figure 7) and Random Forest has the least RMSE. Evaluation of the metrics shows that Random Forest has the best performance seconded by Extreme Gradient Boosting while the least performed technique is the Support Vector Machine. Figure 8 which shows the time series plot of the actual (black) and the predicted (red) values of relative humidity for a representative of five days, confirms that the best performed model is the Random Forest among the different Machine Learning Models.

Table 5: Evaluation Metrics for Relative Humidity Prediction

Model	Coefficient of Determination (R^2)	Mean Absolute Error (MAE) [%]	Mean Absolute Percentage Error (MAPE) [%]	Mean Square Error (RMSE) [%]
Multiple linear regression	0.7756	6.50	9.43	68.7015
Decision tree	0.8634	4.09	5.97	41.8424
Random forest	0.9193	3.41	0.75	24.7115
Support Vector Machine	0.7607	6.04	5.01	73.2893
Extreme Gradient Boosting	0.8944	3.96	5.73	32.3235
Multilayer perceptron	0.8660	4.42	6.35	41.0369
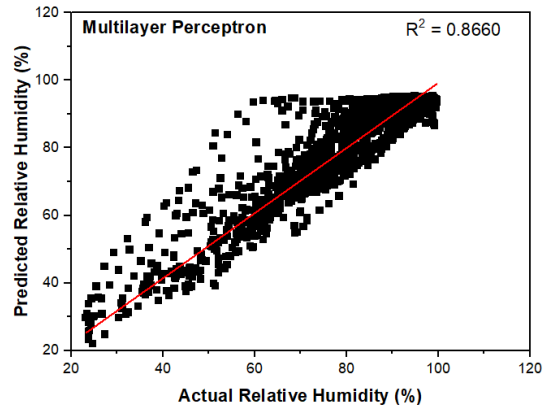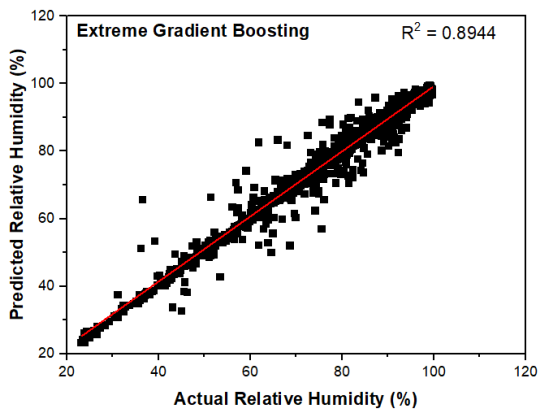
Figure 7: Scatter Plots of the Actual and Predicted Relative Humidity using the Different Models

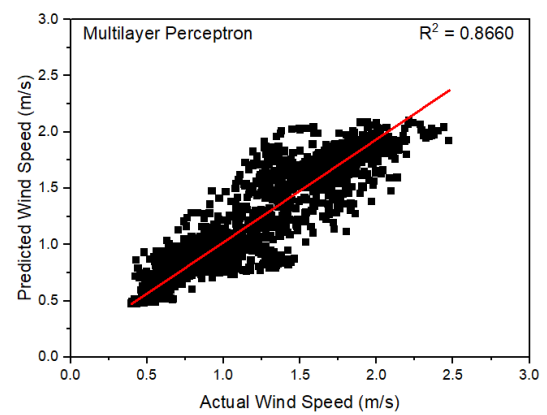**Figure 8: Time Series Plots of the Actual and Predicted Relative Humidity using the Different Models**
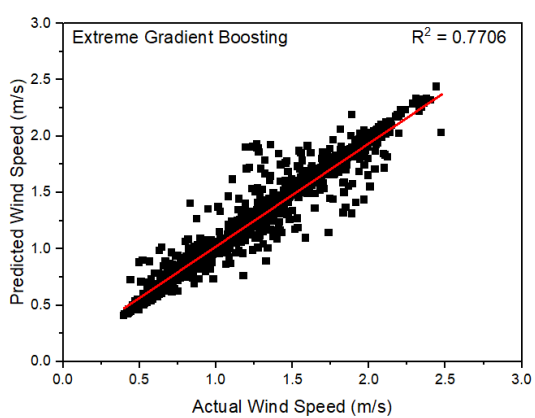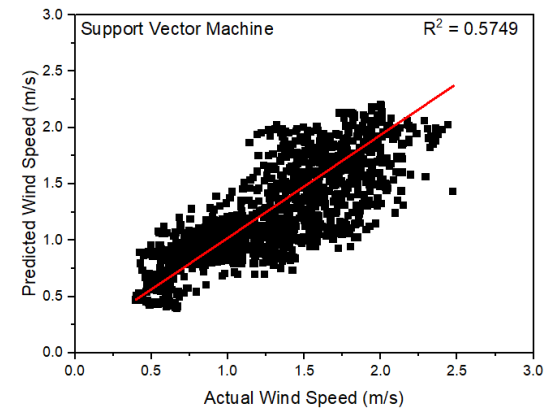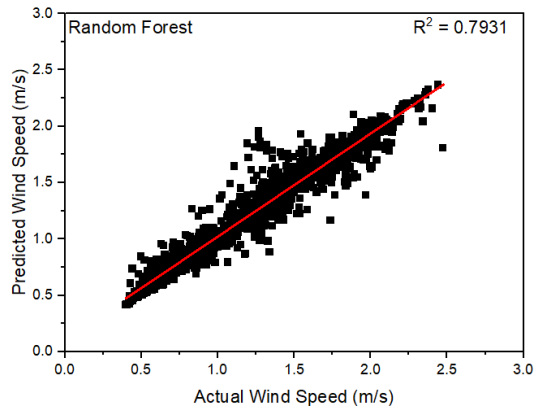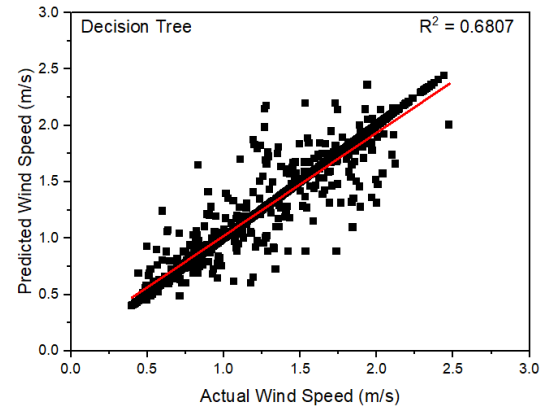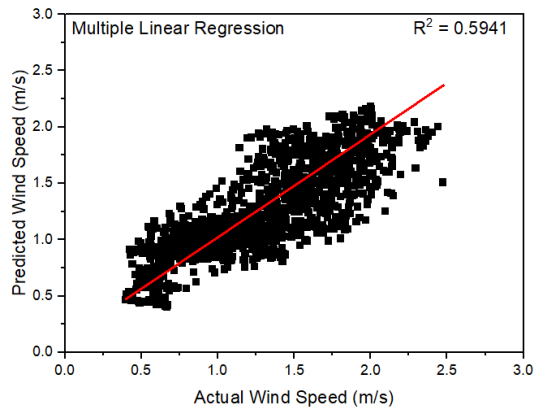
3.4 Wind Speed Prediction

Table 6 shows the performances of the different models used to predict the wind speed. The coefficient of determination (R^2) obtained for the models shows that Random Forest, Extreme Gradient Boosting and Multilayer Perceptron are 0.79, 0.77 and 0.76 implying that the models are good fits for prediction of wind speed. The other models: Multilinear Regression, Decision Tree and Support Vector Regression have R^2 of 0.59, 0.68 and 0.57 respectively, implying that they have moderately good fittings with the actual values. The models have low MAE and RMSE values which means that they have low errors and low spread of scatter points as shown in Figure 9. This suggests that the models have good performances. The values of MAPE were not taken into account because MAPE is used when the values are higher than 0. The best performed model is Random Forest followed by Extreme Gradient Boosting while the least performed is Multiple Linear Regression. Figure 10 shows the time series plot of the actual (black) and the predicted (red) values of wind speed for a representative of five days, using the different Machine Learning techniques. The figure confirms that the best performed model is the Random Forest.

Table 6: Evaluation Metrics for Wind Speed Prediction

Model	Coefficient of Determination (R^2)	Mean Absolute Error (MAE) [m/s]	Mean Square Error (RMSE) [m/s]
Multiple linear regression	0.5941	0.24	0.08716
Decision tree	0.6807	0.19	0.06857
Random forest	0.7931	0.16	0.04440
Support Vector Machine	0.5749	0.24	0.09128
Extreme Gradient Boosting	0.7706	0.16	0.04926
Multilayer perceptron	0.7631	0.17	0.05080

24

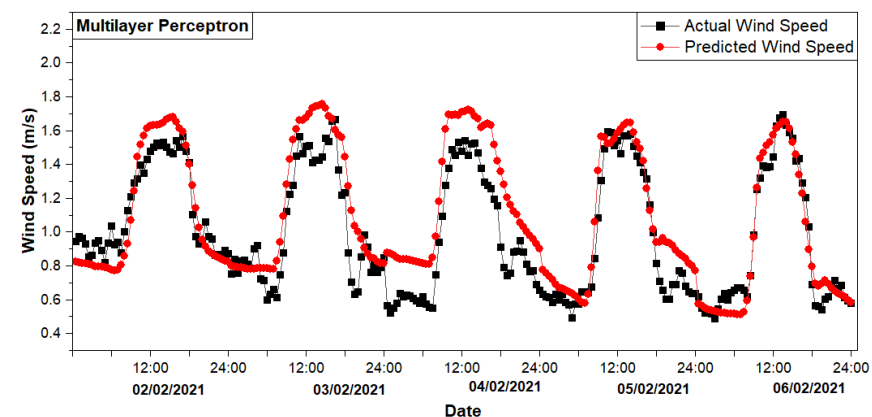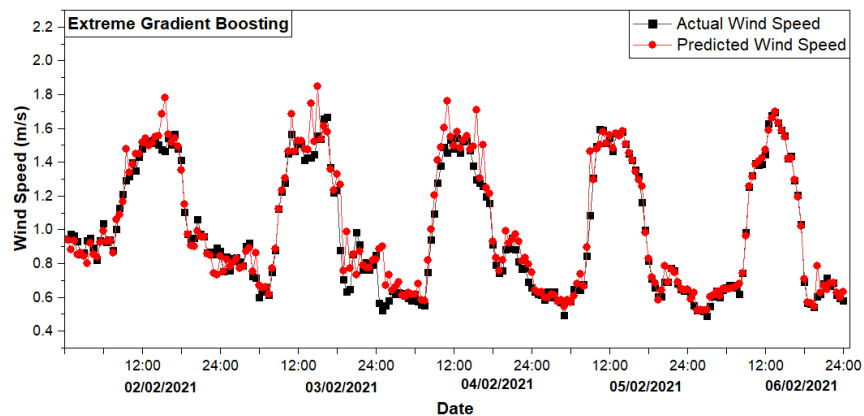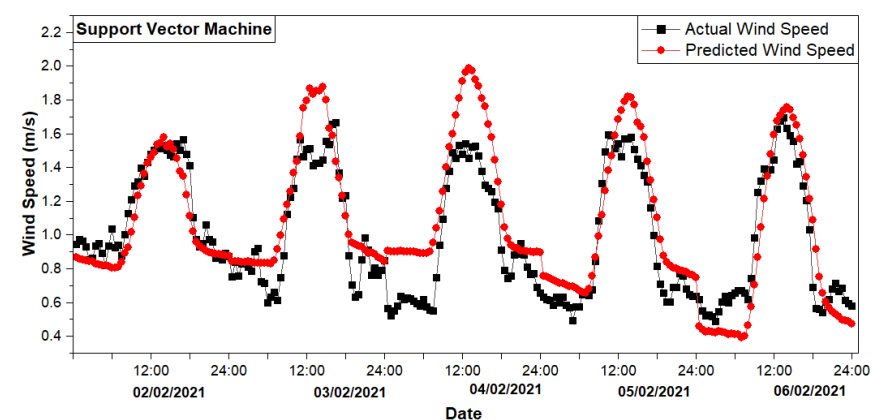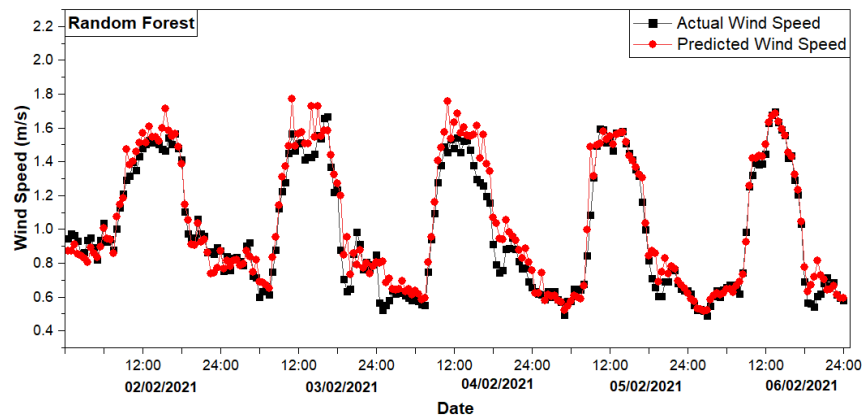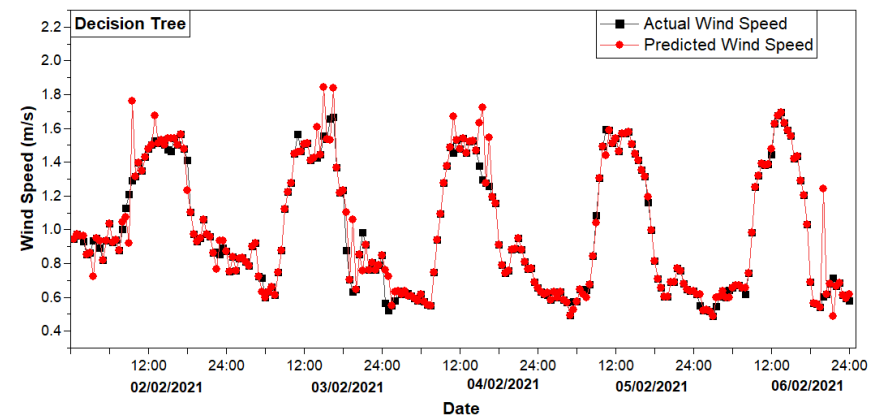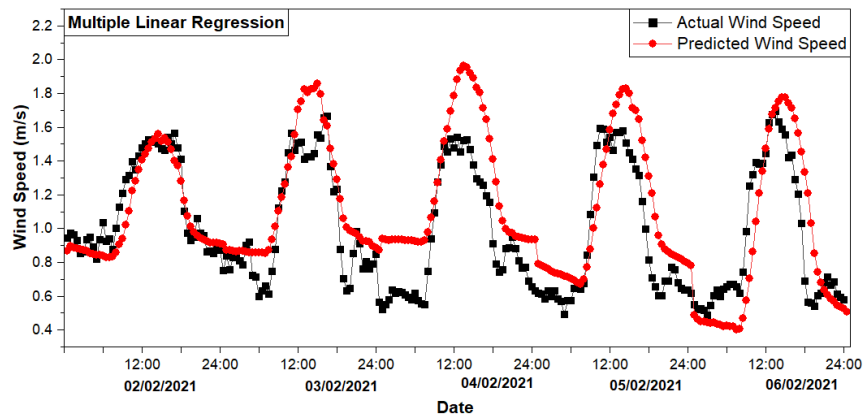**Figure 9: Scatter Plots of the Actual and Predicted Wind Speed using the Different Models**

Figure 10: Time Series Plots of the Actual and Predicted Wind Speed using the Different Models

4. CONCLUSION

In this study, six machine learning models (Multiple linear regression, Decision tree, Random forest, Support Vector Machine, Extreme Gradient Boosting and Multilayer perceptron) were used to predict meteorological variables. The meteorological variables that were predicted are: Temperature, Solar radiation, Relative humidity and Wind speed. This was with the aim of determining the best machine learning model for weather prediction in a tropical location. The data used in the study was collected at the Meteorological Station located at Obafemi Awolowo University, Nigeria (7.53 ^{0}N; 4.54 ^{0}E).

Evaluation metrics such as the mean square error (RMSE), mean absolute percentage error (MAPE), mean absolute error (MAE) and Coefficient of Determination (R^2) were employed to identify the efficiency and to quantify the predictive capacity of each models.

From the study, Random Forest gave the best performance for predicting temperature. The evaluation metrics obtained from the model showed that it has R^2 of 0.93, MAE of 0.78 ^{0}C, MAPE of 2.84 % and RMSE of 1.13 ^{0}C. Extreme Gradient Boosting also performed well with an R^2 of 0.91, MAE of 0.85 ^{0}C, MAPE of 3.10 % and RMSE of 1.37 ^{0}C. The best performed model for solar radiation is the Random Forest having an R^2 value of 0.72, MAE of 85.34 W/m^2 and RMSE of 19008.45 W/m^2. For relative humidity, Random Forest has the best performance. From the evaluation metrics, it has R^2 of 0.92, MAE of 3.41 %, MAPE of 0.75 % and RMSE of 24.71 %. Extreme Gradient Boosting also performed well with an R^2 of 0.89, MAE of 3.96 %, MAPE of 5.73 % and RMSE of 32.32 %. The best performed technique for predicting the wind speed is the Random Forest having an R^2 value of 0.79, MAE of 0.16 m/s and RMSE of 0.044 m/s. Extreme Gradient Boosting also performed well with R^2 value of 0.77, MAE of 0.16 m/s and RMSE of 0.049 m/s.

Random Forest was adjudged the best performed model having the highest R^2, least MAE, least MAPE and least RMSE from the prediction of all the meteorological variables. The second-best performed model was the Extreme Gradient Boosting.

The study concluded that Random Forest was the best performed machine learning model for the prediction of meteorological variables for weather forecasting in a tropical location. The prediction of the aforementioned meteorological variables will be used for future projects in

the study area such as agricultural meteorology, management of natural resources and

monitoring and prediction of weather and climate.

ACKNOWLEDGEMENT

The author acknowledges the efforts of the members of Atmospheric Physics Research

Group (APRG) at the Department of Physics and Engineering Physics of Obafemi Awolowo

University, Ile-Ife, Nigeria during the field experiment.

REFERENCES

1. Obisesan OE. Measurements of Some Meteorological Variables at a Tropical Location in Nigeria. Advances in Research. 2021; 23(4): 26-36, 2022DOI: 10.9734/AIR/2022/v23i430340

2. Hubbard KG. Measurement systems for agricultural meteorology. Handbook of agricultural meteorology, Oxford University Press, New York and Oxford. 1994;76 – 81.

3. Millán H, Kalauzi A, Cukic M, Biondi R. Nonlinear dynamics of meteorological variables: Multifractality and chaotic invariants in daily records from Pastaza, Ecuador. Theor. Appl. Climatol. 2010;102: 75–85.

4. Ridwan WM *et al.* Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia. Ain Shams Eng. J. 2020; 09: 11https://doi.org/10.1016/j.asej.

5. Chong KL *et al.* Performance enhancement model for rainfall forecasting utilizing integrated wavelet-convolutional neural network. Water Resour. Manag.2020;34: 2371–2387.

6. Hanoon MS, Ahmed AN, Zaini N, Razzaq A, Kumar P , Sherif7 M, Sefelnasr A and El-Shafe A. Developing machine learning algorithms for meteorological temperature and humidity forecasting at Terengganu state in Malaysia. Scientific Report. 2021; 11:18935 https://doi.org/10.1038/s41598-021-96872-w.

7. Soumelidis D, Karoutsos G, Skepastianos N, Tzonichakis N. Optimization of Weather Forecast Data Using Machine Learning Algorithms. Environ. Sci. Proc.2023;26: 49. https://doi.org/10.3390/ environsciproc2023026049.

8. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30. 2016; 770–778.

9. Bachute MR, Subhedar JM. Autonomous Driving Architectures: Insights of Machine Learning and Deep Learning Algorithms. Mach. Learn. Appl.2021;*6:* 100164.

578 10. Abhishek K, Singh MP, Ghosh S, Anand A. Weather Forecasting Model using Artificial
579 Neural Network. Proc. Technol.2012; 4: 311–318.
580

581 11. Bochenek B, Ustrnul Z. Machine Learning in Weather Prediction and Climate Analyses—
582 Applications and Perspectives. Atmosphere. 2022;13: 180.
583

584 12. Powers CJ, Devaraj A, Ashqeen K, Dontul A, Joshi A, Shenoy J, *et al.* Using artificial
585 intelligence to identify emergency messages on social media during a natural disaster: a deep
586 learning approach. Int. J. Inf. Manag. Data Insights. 2023: 3 (1): 100164.
587 doi:10.1016/j.jjimei.2023.100164.
588

589 13. Anton CA, Matei O and Avram A. Collaborative data mining in agriculture for prediction of
590 soil moisture and temperature. Computer Science On-Line Conference, Springer International
591 Publishing. 2019; 141–151.
592

593 14. Cortez P and Morais JR. A data mining approach to predict forest fires using
594 meteorological data. Environmental Science, Computer Science. 2007. Available:
595 http://www3.dsi.uminho.pt/pcortez/fires.pdf.
596

597 15. Joshi A, Kamble B, Joshi V, Kajale K and Dhange N. Weather forecasting and climate
598 changing using data mining application. Int. J. Adv. Res. Comput. Commun. Eng. 2015; 19–
599 21. doi: 10.17148/ijarcce.2015.4305.
600

601 16. Olaiya F and Adeyemo AB. Application of data mining techniques in weather prediction
602 and climate change studies. Int. J. Inf. Eng. Electron. Bus. 2021; 51–59. doi:
603 10.5815/ijieeb.2012.01.07.
604

605 17. Oladipo ID *et al.* An improved course recommendation system based on historical grade
606 data using logistic regression. Communications in Computer and Information Science,
607 Springer International Publishing. 2021; 207–221.
608

609 18. Segovia JA, Toaquiza JF, Llanos JR, Rivas DR. Meteorological Variables Forecasting
610 System Using Machine Learning and Open-Source Software. Electronics.2023; 12:1007.
611 https://doi.org/ 10.3390/electronics12041007
612

613 19. Shivang J and Sridhar SS. Weather prediction for Indian location using Machine learning.
614 Int. J. Pure Appl. Math., vol. 2018; 118:1945–1949.
615

616 20. Zaman Y. Machine learning model on rainfall-a predicted approach for Bangladesh.
617 United International University, 2018.
618
619 21. Montero GR. Modelos de Regresión Lineal Múltiple. Technical Report. Documentos de
620 Trabajoen Economía Aplicada; Universidad de Granada: Granada, Spain. 2006.
621
622 22. Aurélien G. Hands-on Machine Learning with Scikit-Learn &Tensorflow. O'Reilly Media,
623 Inc.: Sebastopol, CA, USA. 2017.
624
625 23. Elbeltagi A, Kumar M, Kushwaha NL, Pande CB, Ditthakit P, Vishwakarma DK, Subeesh
626 A. Drought indicator analysis and forecasting using data driven models: Case study in
627 Jaisalmer, India. Stoch. Environ. Res. Risk Assess.2022;37:113–131.

628

629 24. Menacho CH. Modelos de regresión lineal con redes neuronales. An. Científicos2014;75,
630 253.
631

632 25. Obisesan OE. Evaluation of Selected Empirical Schemes of Calculating Sensible Heat
633 Flux from RoutinelyMeasured Meteorological Parameters in a Tropical Location. Advances in
634 Research. 2022;23(4):11-25.DOI: 10.9734/AIR/2022/v23i430338.
635
636