APPLICATIONS OF ARTIFICIAL NEURAL NETWORKS (ANNs) AND RANDOM FOREST (RF) FOR TEMPERATURE PREDICTION

A THESIS

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE

MASTER OF SCIENCE

BY

SAMPSON GYAMFI

DR. FAEZEH SOLEIMANI – ADVISOR

MUNCIE, INDIANA

MAY 2023

Acknowledgment

I am extremely grateful to my Supervisor, Dr. Faezeh Soleimani, who contributed immensely to the completion and success of my thesis. This would not have been possible without her support and nurturing. I am deeply indebted to my committee members; Dr. Drew Michael Lazar and Professor Rahmatullah Imon for standing by me throughout this work. I would like to extend my deepest appreciation to my family, especially to my wife Mrs. Mary Akyere Gyamfi, for her emotional support throughout this study.

Abstract

Artificial Neural Networks (ANNs) and Random Forest (RF) have been used in various domains for modeling and prediction with high accuracy due to their ability to learn and adapt. One of the major applications of ANNs and RF is temperature prediction. For instance, neural networks and random forest have been trained to build models to predict soil temperature, air temperature, sea surface temperature, etc., with minimal errors, and significant accuracy. This thesis concentrates on designing ANNs and RF prediction models to predict the surface temperature of a bridge.

Surface temperature plays an important role in homes, agriculture, industry, and engineering fields. Accurate surface temperature prediction can help improve productivity and avoid risks on many roads and bridges. ANNs and RF have been largely used by engineers for the measurement of diffusion-based overlay detection of structural damage fault, medical image processing for diagnosis of any symptoms, wind turbine based which is Simulink and fast, and others. Our goal is to research, put into practice, and recommend Artificial Neural Networks and Random Forest as more accurate machine learning approaches for temperature prediction.

Currently, many machine learning methods such as Multiple Linear Regression (MLR), Survival Forest (SF), Decision Tree, Support Vector Machine (SVM), and others, have been applied to surface temperature prediction. In this study, we will use artificial neural networks and random forest to the predict average surface temperature given the Direct Normal Irradiance (DNI), air temperature, humidity, and wind speed of a non-heated zone of a bridge as features in our dataset. To be precise, we intend to use the existing dataset and apply the Gradient Descent Backpropagation (GDBP) and the RF techniques to build a predictive model for average surface temperature. To demonstrate how powerful ANNs and RF are as key machine learning methods for better temperature prediction, the predicted temperature values were compared with the actual (true) temperature values. The results showed that both ANNs and RF algorithms can predict the average surface temperature with high accuracy, but RF performed better than ANNs.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

In numerous fields, such as climate research, energy, agriculture, transportation, and health, long-term surface temperature forecasting is crucial. Previously, many important studies have used machine learning algorithms to predict temperatures. Among these studies, the predictive abilities of different machine learning algorithms were compared using different error criteria.

The Gradient Boosting Decision Tree (GBDT), Random Forest (RF), and Linear Regression (LR) were used for temperature prediction models of asphalt pavements in winter [1]. The objective of this study was to explore the correlation between the pavement temperature of asphalt pavements and meteorological factors and implement an accurate trend prediction of the asphalt pavement temperature. The mean-square error of the GBDT predicting results has a lower value when compared with the Random Forest and Linear Regression owing to the high robustness and the good generalization ability, which reflects the GBDT model has good applicability in the field of prediction. The results indicated that GBDT would perform excellent ability on prediction.

The Convolutional Neural Network Based on Ensemble Empirical Mode Decomposition (EEMD-CNN) was used to predict soil temperature [2]. It was concluded that the proposed EEMD-CNN model in this study is a suitable tool for soil temperature prediction, with minimal errors.

Global warming has recently drawn scientists' attention since it is correlated with the rise in air temperature. A comprehensive review of artificial neural networks (ANNs)-based approaches (such as recurrent neural network (RNN), long short-term memory (LSTM), etc.), were used to forecast air temperature [3]. The review showed that the neural network models can be employed as promising tools to forecast air temperature.

Though various ANNs-based approaches performed successfully in many problems, such as flood [4], rainfall [5], water quality [6], air temperature [7], and surface temperature predictions, the random forest has also proven to be one of the most effective machine-learning algorithms for temperature prediction. For instance, explainable machine learning algorithms were used to predict glass transition temperatures [8]. The work investigated how different machine learning algorithms can be used to predict the transition temperatures of glasses based on their chemical composition. To assess the predictive performance obtained by machine learning algorithms, the possible gains

by tuning the hyperparameters of these algorithms were investigated. The results show that the best machine learning algorithm for predicting glass transition temperatures is the Random Forest (RF).

The performance of other machine learning algorithms such as the Support Vector Regression (SVR) in temperature prediction cannot be overlooked. A study on the application of SVR for weather prediction was conducted [9]. The results were compared with Multi-Layer Perceptron (MLP) trained with a back-propagation algorithm and the performance of SVR was found to be consistently better.

Additionally, the performance of two machine learning algorithms, Support Vector Regression (SVR) and Multi-layer Perceptron (MLP), in a problem of monthly mean air temperature prediction from the previously measured values in observational stations of Australia and New Zealand, and climate indices of importance in the region were examined [10]. The performance of the two considered algorithms was discussed in the paper and compared to alternative approaches. The results indicated that the SVR algorithm can obtain the best prediction performance among all the algorithms compared in the paper. Moreover, the results obtained have shown that the mean absolute error made by the two algorithms considered was significantly larger for the last 20 years than in the previous decades, which can be interpreted as a change in the relationship among the prediction variables involved in the training of the algorithms.

The statistical methods that have been used to support surface temperature prediction are numerous. A common theme among data-intensive methods is the use of machine-learning algorithms where the primary objective is to identify emergent patterns and make predictions with the minimal human intervention [11]. Neural networks, in particular, are designed to mimic the neuronal structure of the human brain by "learning" inherent data structures through adaptive algorithms [12, 13]. The most popular form of a neural network is the feed-forward multilayer perceptron (MLP) trained using the backpropagation algorithm [12]. This model is typically used to predict the response of one or more variables given one or many explanatory variables. The hallmark feature of the MLP is the characterization of relationships using an arbitrary number of parameters (i.e., the hidden layer) that are chosen through iterative training with the backpropagation algorithm. Conceptually, the MLP is a hyper-parameterized non-linear model that can fit a smooth function to any dataset with minimal residual error [14].

However, as the amount of data grows, the model-building process becomes challenging, and it becomes difficult to ensure the accuracy of the prediction model. Therefore, establishing a high-accuracy average surface temperature prediction model is the research emphasis. The goal of the research described in this thesis is to use artificial neural networks (ANNs) and random forest (RF), to predict average surface temperature suitable for bridges in a non-heated zone with radiation from the sun (Direct Normal Irradiation, w/m^2), air temperature (℃), humidity (%), and wind speed (m/s), as the explanatory variables.

Organization of chapters: Chapter 2 describes the features of the dataset used, the response, and the explanatory variables. This chapter also reports the results from an Exploratory Data Analysis (EDA) performed on the dataset as well as the error criteria used in selecting a suitable model by both ANNs and RF algorithms.

In Chapter 3, the assumptions and structures of ANNs and RF algorithms will be discussed. Also, this chapter provides the performances and results from the ANNs and RF algorithms. The performance of both algorithms at the testing phase will be discussed using tables and scatter plots, and conclusions from the results are outlined.

Chapter 4 summarizes the research conducted in this study and presents the conclusions. Future work will also be discussed in this chapter.

Computational details: All computations in our study are done using R (version 3.6.1). For the neural network, the **neuralnet** package (version 1.44.2), and for the random forest, the **random forest** package (version 4.7-1), are used.

CHAPTER 2

This chapter describes the characteristics of the dataset used, results from an exploratory data analysis, and error criteria used in the testing phase of the models built using both neural networks and random forest algorithms.

2.1 Dataset and validation of the models

For this study, the data has been obtained from a geothermal bridge de-icing project from the non-heated zone, consisting of 807 observations [15]. The response variable is the average surface temperature with four covariates or features. The covariates include radiation from the sun, measured in (w/m^2), air temperature (℃), humidity (%) and wind speed (m/s). The dataset is divided into a 70 % /30% training/testing set and an 80%/20% training/testing set to avoid possible overfitting of the models. The 70% and 80% training sets are used to train the neural network and random forest algorithms while the 30% and 20% test sets were used to test the accuracy of the model. The data was partitioned and shuffled 20 times for both ANNs and RF algorithms and the best result based on the lowest Mean Square Prediction Error (MSPE) was chosen.

2.2 Exploratory Data Analysis (EDA)

An Exploratory Data Analysis (EDA) which includes summary statistics, histograms, box plots, and scatter plots was performed to visualize the relationship between the outcome and the features. To check the averages of all the variables, a summary statistic was performed as shown in Table 2.1 below.

Table 2.1: Summary statistics of the variables

Statistic	DNI	Air temp.	Humidity	Wind speed	Surface temp.
Minimum	0.0	-0.5494	10.00	0.000	1.271
1st Quartile	0.0	2.5990	66.00	3.129	2.861
Median	0.0	4.3723	76.00	4.470	4.646
Mean	133.9	7.2830	73.15	4.607	6.053
3rd Quartile	26.0	11.4389	89.00	6.259	8.557
Maximum	965.0	27.4687	97.00	10.729	15.717

From Table 2.1, the minimum surface temperature of the bridges recorded in the non-heated zone was 1.271°C while the maximum was 15.717°C. The mean temperature was 6.053°C while the median was 4.646°C. Direct Normal Irradiation (DNI) is the amount of solar radiation received per unit area by a surface that is always held perpendicular (or normal) to the rays that come in a straight line from the direction of the sun at its current position in the sky [16]. Most of the DNIs recorded were $0\ \mathrm{W/m^2}$ with a maximum of $965\ \mathrm{W/m^2}$. The mean DNI was $133.9\ \mathrm{W/m^2}$. The air temperature was recorded. The minimum air temperature was -0.5494°C while the maximum was 27.4687°C. The mean air temperature was 7.2830°C with a median of 4.3723°C. The mean atmospheric moisture or humidity recorded was 73.15% with a minimum and maximum of 10% and 97% receptively. The mean wind speed recorded was 4.607 miles per hour, with minimum and maximum of 0 and 10.729 miles per hour respectively.

To get a virtual display of the type of distribution of the variables, a histogram with the normal curve overlay was used. Figure 2.1 below shows the histogram of the average surface temperatures recorded.

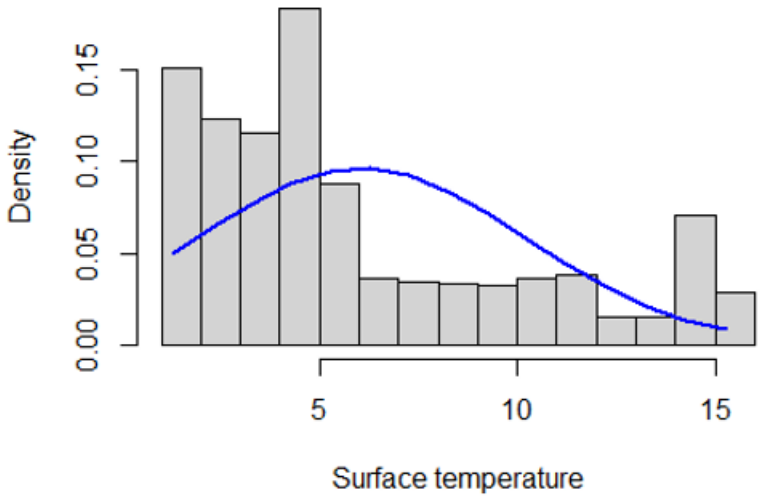

Figure 2.1: Histogram of average surface temperature with normal curve

The histogram above shows that the outcome, average surface temperature, was approximately skewed to right with most of the values below 5 degrees Celsius. We continued to explore the data with the histogram with the normal curve overlay of the predictors of the surface temperatures.

Figure 2.2 shows the histogram of DNI with a normal curve overlay. The histogram shows that the DNI distribution was approximately zero-inflated as most of the values recorded were zero. The distribution of the air temperatures was approximately right skewed with most values between 0 and 5 degrees Celsius.

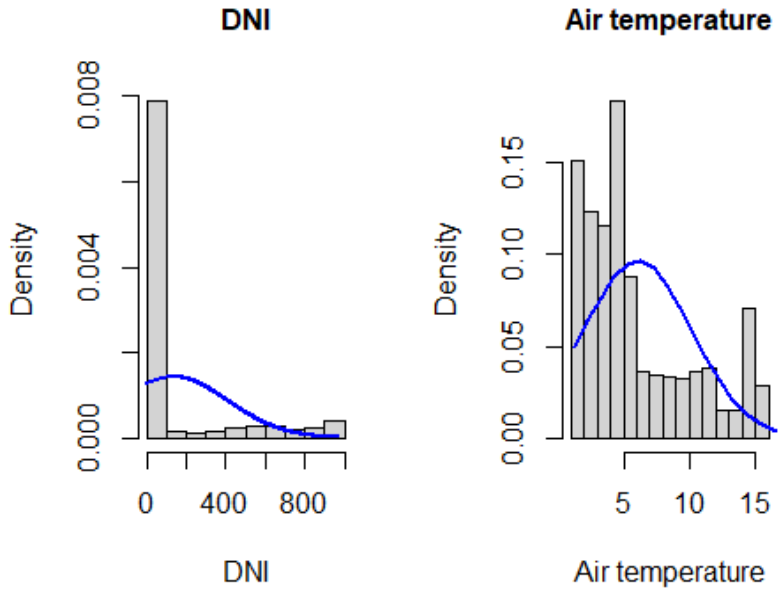

Figure 2.2: Histogram of DNI and air temperature

Figure 2.3 shows the histogram of the humidity recorded. Most of the humidity recorded was between 70-100%, which indicates a high level of atmospheric moisture in the areas. The figure also shows that the distribution of wind speed is approximately normal.

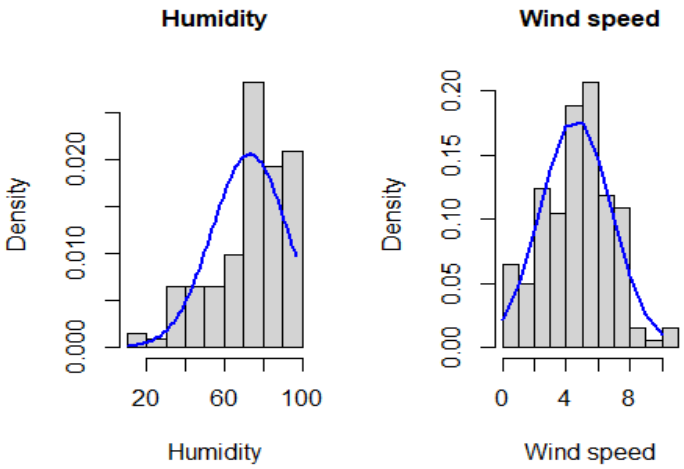

Figure 2.3: Histogram of humidity and wind speed

To further explore the variables used in this study, box plots were used to check possible outliers or extreme values in the dataset. The box plot of the surface temperature is shown in Figure 2.4. The box plot shows no possible outliers in the response.

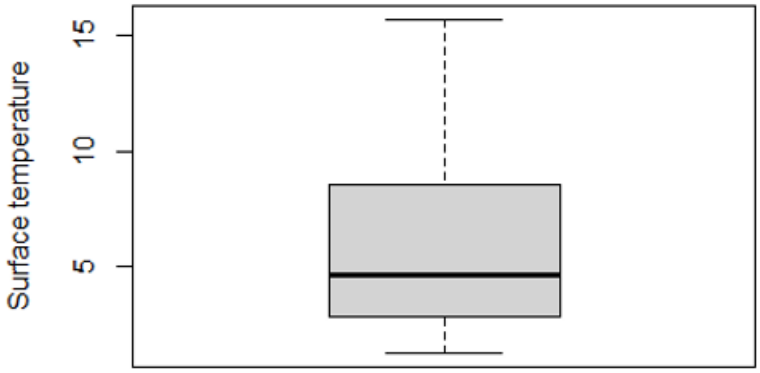

Figure 2.4: Box plot of average surface temperature

Figure 2.5 below shows the box plots of DNI, air temperature, humidity, and wind speed. There are a high number of outliers or extreme values found in the values of DNI. About 68.5% of the

DNI values recorded were $0\ W/m^2$ together with a few high values. A few outliers were also found in the values of humidity and air temperature. However, the wind speed had no outliers.

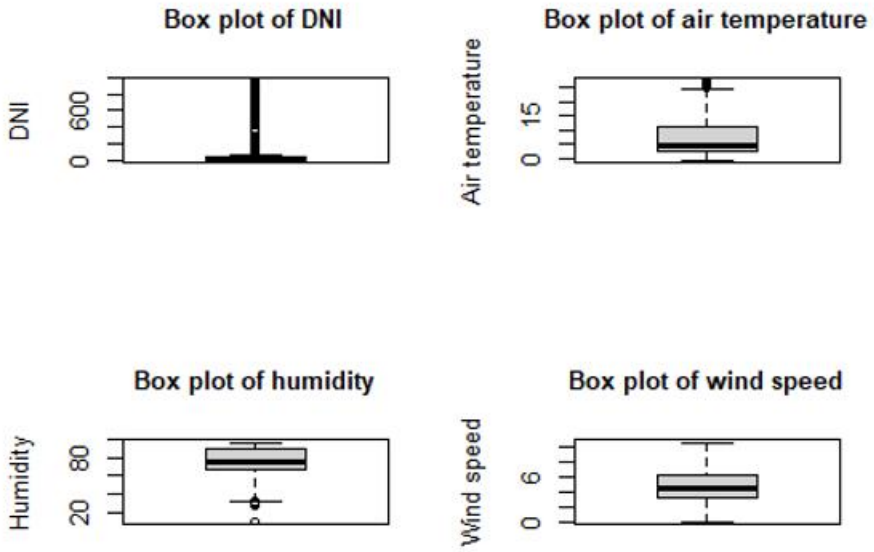

Figure 2.5: Box plot of the predictors

To check the relationship between the response, surface temperatures, and the predictors (DNI, air temperature, humidity, and wind speed), Pearson's correlation coefficients and scatter plots were used. Table 2.2 below shows the correlation coefficients of the variables while Figure 2.6 shows the scatter plot matrix of the surface temperature versus the predictors.

Table 2.2: Correlation coefficients of the variables

	DNI	**Air temp.**	**Humidity**	**Wind speed**	**Surface temp.**
DNI	1.000	0.574	-0.558	-0.135	0.288
Air temp.	0.574	1.000	-0.333	0.113	0.850
Humidity	-0.558	-0.332	1.000	0.201	-0.189
Wind speed	-0.135	0.113	0.201	1.000	0.162
Surface temp.	0.288	0.850	-0.1850	0.162	1.000

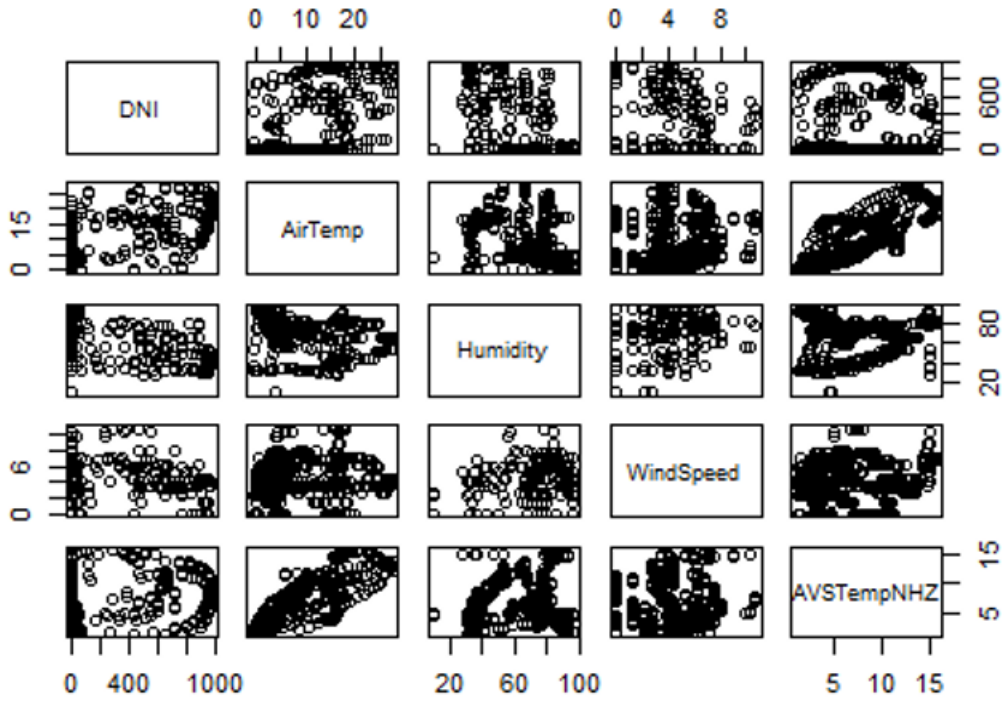

Figure 2.6: Correlation matrix of all variables

The scatter plots above show a strong positive linear relationship between surface temperatures and air temperature with a correlation coefficient of 0.850. However, there was a weak to moderate relationship between the surface temperatures and DNI, humidity, and wind speed.

The p-values of predictors were calculated to determine which variables among the features are statistically significant in predicting the surface temperatures of the bridges. The Variance Inflation Factor (VIF) calculates the degree to which the behavior (variance) of an independent variable is inflated by its interaction or connection with other independent variables. Variance inflation factors make it simple to gauge how much a given variable contributes to the standard error of analysis. To check for possible multicollinearity or strong linear relationship between the predictors, the VIF was calculated for each predictor as shown in Table 2.3.

Table 2.3: P-value and VIF of the predictors

Variable	t-value	p-value	VIF
DNI	-13.407	< 2e-16	1.969
Air temp.	49.159	< 2e-16	1.585
Humidity	-1.217	0.224	1.491
Wind speed	0.567	0.571	1.107

From Table 2.3, the p-values of DNI and air temperature are statistically significant in determining the surface temperatures of bridges. Thus, DNI and air temperature are statistically significant. However, air temperature appears to play a more significant role in surface temperature prediction since the |t-value| of air temperature (49.159) is greater than the |t-value| of the DNI (13.407). The VIFs appear to show less to no evidence of multicollinearity among the predictors with all the VIFs less than 2.00.

2.3 Model Evaluation criteria

In this study, Mean Square Prediction Error (MSPE), the model accuracy based on the Mean Square Deviation (R^2), the Mean Square Deviation (MAD), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) are used to evaluate the performance of both neural network and random forest models. The MSPE was used to evaluate the accuracy of the predictions.

CHAPTER 3

This chapter explains the structure and assumptions of both the neural networks and random forest algorithms. Various variable importance criteria used on the data set given will also be discussed. Finally, the performance of both algorithms at the testing phase is discussed using tables and scatter plots (actual versus predicted), and conclusions from the results are outlined.

3.1 Temperature prediction using Artificial Neural Networks (ANNs)

3.1.1 Assumptions of ANNs

The regression of a dependent variable, *Y*, on an independent variable, *X*, is the computation of the most probable value of *Y* for each value of *X* based on a finite number of possibly noisy measurements of *X* and the associated values of *Y* [17]. For instance, in the case of linear regression, it is assumed that the output, *Y*, is a linear function of the input, *X*, and the unknown parameter, β_i. ANN does not rely on distributional assumptions about the response variable and does not require the assumption of any functional form.

Artificial Neural Networks (ANNs), also known as Neural Network models (NN), are mathematical or computational models based on neural networks (neurons) [18]. Because ANNs are designed based on human biological systems, these models are built on the relationship seen in the training data set by adapting synaptic connections that exist between neurons. An ANN consists of artificially linked artificial neurons and processes information using a linking approach for calculation. An ANN model consists of an input, hidden, and output layer. The hidden layer in ANN contains some nodes that calculate the weights of input based on external or internal information that gets into the network during the learning process. ANN can derive the desired information from complicated or incorrect data, to determine patterns and recognize trends that are very complex or almost impossible to identify by humans or other statistical techniques. ANN is the best example of adaptive learning that can be designed to perform real-time operations with a high level of fault tolerance.

From Figure 2.1, the distribution of the average surface temperature was approximately skewed to right with most of the values below 5 degrees Celsius. A lot of research has been done to investigate the relationship between data skewness in the response variable and the accuracy of artificial neural network predictive models. Data skewness does not have a significant effect on

the accuracy of the artificial neural network predictive model [19]. This implies that the artificial neural network predictive model has a higher capability to cope with skewed data due to its complexity in the hidden layer. The application of machine learning algorithms, such as the artificial neural networks model (ANNs), becomes more popular when it links to the problem of skewed data.

3.1.2 Temperature prediction using ANNs

The feed-forward multilayer perceptron (MLP), trained using the backpropagation method, is the most widely used type of neural networks. Supervised neural networks (e.g., multilayer feed-forward networks) are generally used for prediction [20]. This model is typically used to predict the response of one or more variables given one to many explanatory variables. Characterizing relationships with an arbitrary number of parameters (i.e., the hidden layer) selected through iterative training with the backpropagation algorithm is the distinguishing characteristic of the MLP. The MLP is a theoretically hyper-parameterized non-linear model that can apply a smooth function to any dataset with a small amount of residual error. An arbitrarily large number of parameters to fit a neural network provides obvious predictive advantages but complicates the extraction of model information.

The typical MLP network is composed of multiple layers that define the transfer of information between input and response layers. Information travels in one direction where a set of values for variables in the input layer propagates through one or more hidden layers to the final layer of the response variables. Hidden layers between the input and response layers are key components of a neural network that mediate the transfer of information. Just as the input and response layers are composed of variables or nodes, each hidden layer is composed of nodes with weighted connections that define the strength of information flow between layers [21]. The weights that connect variables in a neural network are partially analogous to parameter coefficients in a standard regression model and can be used to describe relationships between variables [20]. Bias layers connected to hidden and response layers may also be used that are analogous to intercept terms in a standard regression model [21].

Diagnostic information such as variable importance or model sensitivity is a necessary aspect of exploratory data analysis**.** In this study, the Neural Interpretation Diagram (NID) and Olden's algorithm were used to visualize our neural network.

3.1.3 Variable importance using the Neural Interpretation Diagram (NID)

A NID is a modification of the standard conceptual illustration of the MLP network that changes the thickness and color of the weight connections based on magnitude and sign, respectively [21]. Positive weights between layers are shown as black lines and negative weights as gray lines. Line thickness is proportional to the absolute magnitude of each weight**.** The hidden labels outside of the nodes represent variable names and labels within the nodes indicate the layer and node (I: input, H: hidden, O: output, B: bias). The rationale for use of NID is to provide insight into variable importance by visually examining the weights between the layers [20]. For example, input (explanatory) variables that have strong positive associations with response variables are expected to have many thick black connections between the layers.

With this study, the neural network algorithm was trained with inputs or observations of four (4) explanatory variables (DNI, Air temperature, humidity, and wind speed) with one output (Avg. surface temperature). There were two hidden layers made up of 6 and 3 nodes in succession as shown in Figure 3.1.

Figure 3.1 represents the NID for the entire dataset. The thick black line between the air temperature node (I2) in the input layer and node H4 in the first hidden layer indicates a strong positive weight (importance) between wind speed and node H4 in the first hidden layer.

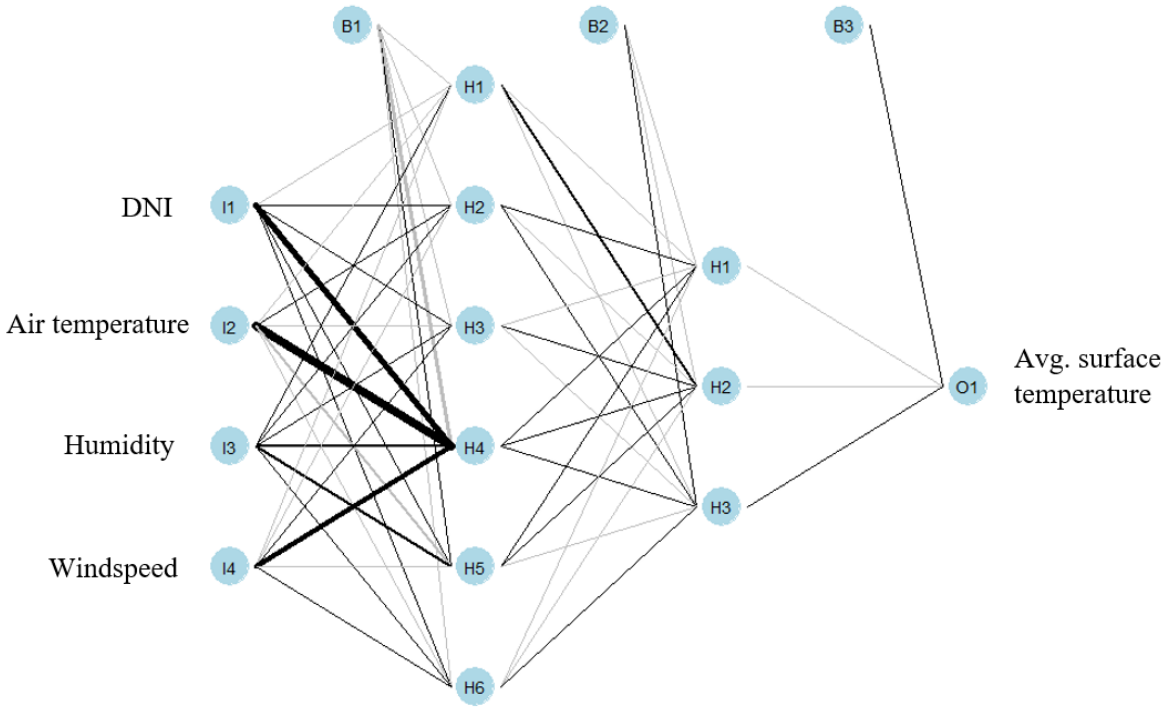

Figure 3.1: Neural Interpretation Diagram

This qualitative interpretation can be very challenging for large models, particularly if the sign of the weights switches after passing the hidden layer, thus, NID provides a better interpretation of variable importance in the simplest models.

3.1.4 Variable importance using Olden's connection weights algorithm

The primary benefit of visualizing a NID is the ability to evaluate network architecture and the variation in connections between the layers [21]. Olden's connection weights algorithm is an alternative method to quantitatively describe a neural network, deconstructing the model weights to determine variable importance. This method calculates importance as the summed product of the raw input-hidden and hidden-output connection weights between each input and output node. An advantage over other methods (such as Garson's algorithm for relative importance) is that the olden function can evaluate neural networks with multiple hidden layers and response variables,

which is important to this study. The importance values assigned to each variable are in units based on the summed product of the connection weights [22].

Figure 3.2 represents the feature importance using Olden's algorithm for the entire data set. The output from the function and the bar plot tells us that the air temperature and humidity have the strongest positive and negative relationships, respectively, with the surface temperature. Similarly, variables that have relative importance close to zero, such as DNI and wind speed do not have any substantial importance for surface temperature prediction.

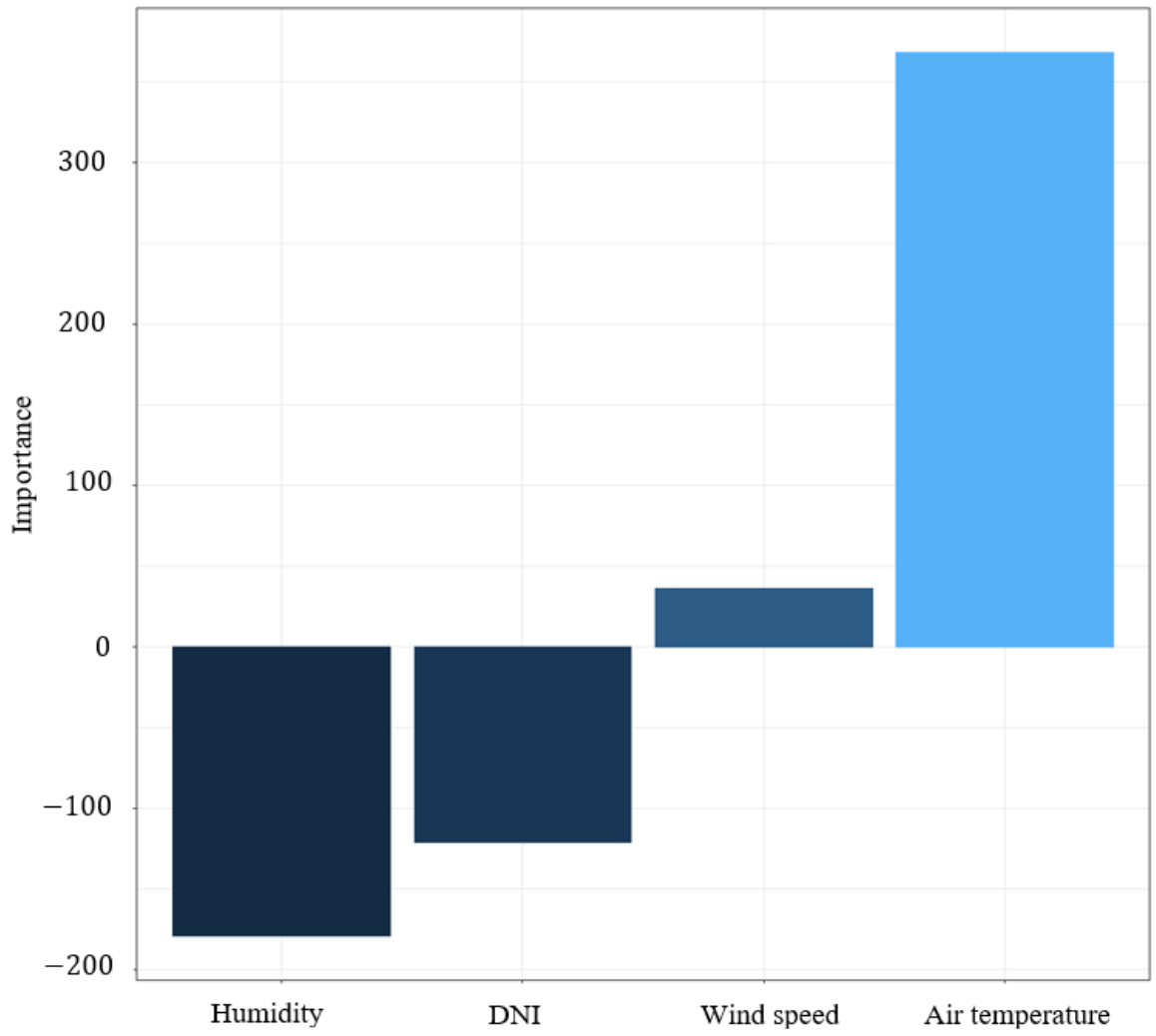

Figure 3.2: Feature importance using Olden's algorithm.

3.1.5 Performance of ANNs in the testing phase

The neural network was performed using all 4 features available. Error values between observed and predicted data were determined by Mean Square Prediction Error (MSPE), the model accuracy based on the Mean Square Deviation (R^2), the Mean Square Deviation (MAD, the Mean Absolute Percentage Error (MAPE), the Root Mean Square Error (RMSE), and the Mean Absolute Error (MAE). The MSPE was used to evaluate the accuracy of the predictions. The MSPE measures the expected squared distance between what the model predicts for the average surface temperature value and the true value. Models with low MSPE are viewed more favorably.

The mean absolute deviation MAD of a dataset is the average distance between each data point and the mean. $1 - \text{MAD}$ gives us an idea about the variability in a dataset which is represented by R^2 (Coefficient of determination). This coefficient varies from 0 to 1, where coefficients close to 1 indicate that most of the variabilities in the response variable can be explained by the predictors.

Table 3.1 shows the performance of the ANN models in the testing face using the two different samples: 70%/30% and 80%/20% train/test sets. The model with 70%/30% train/test set had MSPE of 0.9170 yielding an accuracy of 95.67% on a mean absolute deviation basis (i.e., the average deviation between estimated and actual temperature stands at a mean of 4.33%). Also, the model with an 80%/20% train/test set had MSPE of 0.691 yielding an accuracy of 95.24% on a mean absolute deviation basis (i.e., the average deviation between estimated and actual temperature stands at a mean of 4.76%).

Table 3.1: Evaluation of the performance of ANNs models in the testing phase

MODELS	***MSPE***	R^2	**MAD**	***MAPE***	***RMSE***	***MAE***
70%/30%	0.9170	95.67%	0.0433	0.148	0.958	0.647
80%/20%	0.691	95.24%	0.0476	0.132	0.831	0.595

3.1.6 Data visualization for ANNs model: Actual average surface temperature versus predicted average surface temperature.

To provide a virtual representation of the performance of the ANN model, scatter plots with smooth line markers were used to compare the actual and the predicted average surface temperature values. The blue trend indicates that of the predicted observations and the orange trend represents the actual values taking into consideration the number of features in each model. The figures show the number of data points in the testing set. Though there were 242 observations and 161 observations for the 30% and 20% test sets respectively, the data visualization diagram was divided into smaller sets for easier analysis and comparison. The trend for observations 1-40, 41-80, 81-120, 121-160, 161-200, and 201-242 were shown considering the ratio 70%/30%, and the trend for observations 1-40, 41-80, 81-120 and 121-161, were shown considering the ratio 80%/20%. This gives a clear trend of how the predicted values are close to the actual observations for each set of features. Figures 3.3 and 3.4 give a trend of the actual and predicted values considering all 4 features for ratios 70%/30% and 80%/20% respectively.

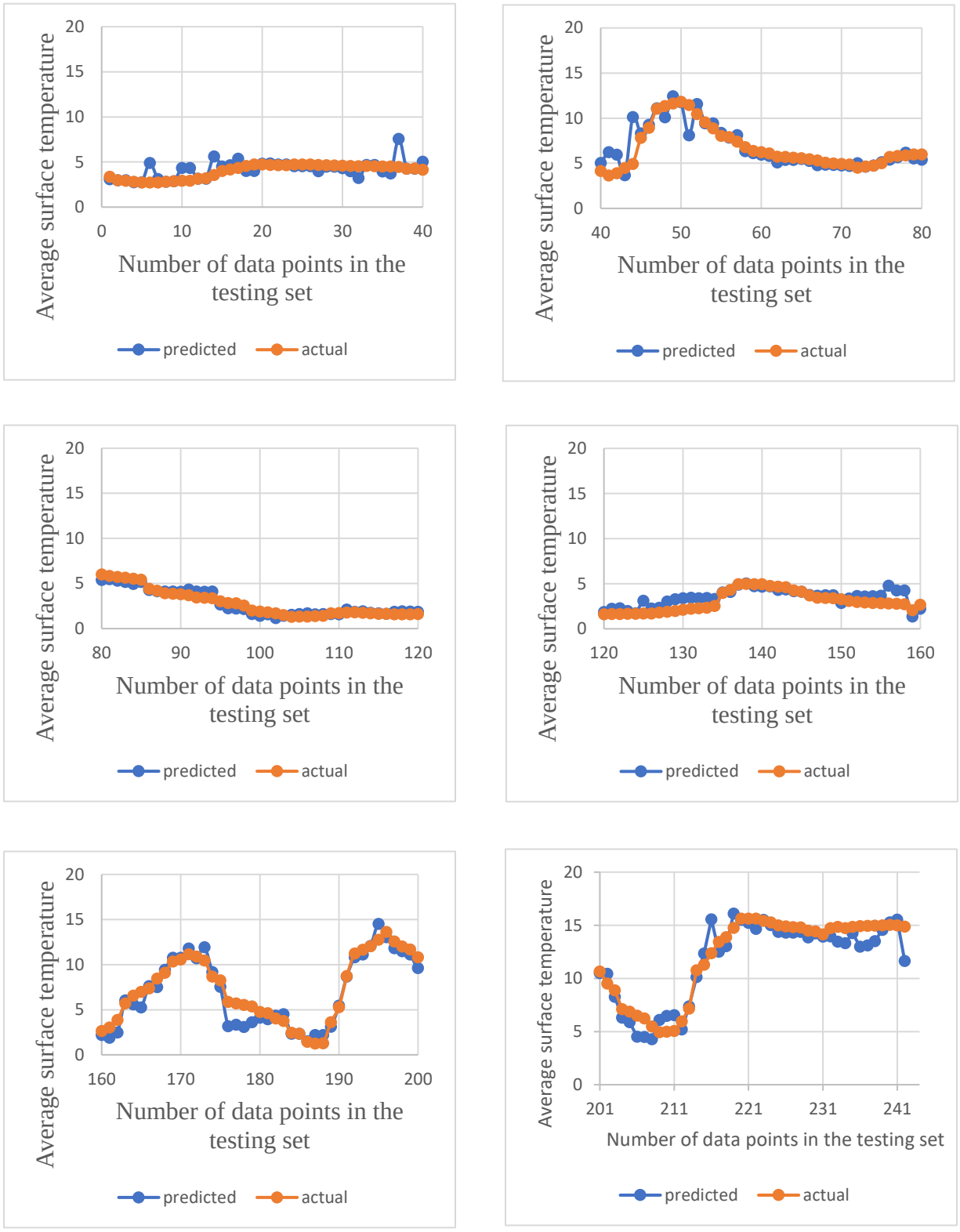

Figure 3.3: Actual versus predicted results with ANNs for the model with all features (70%/30%)

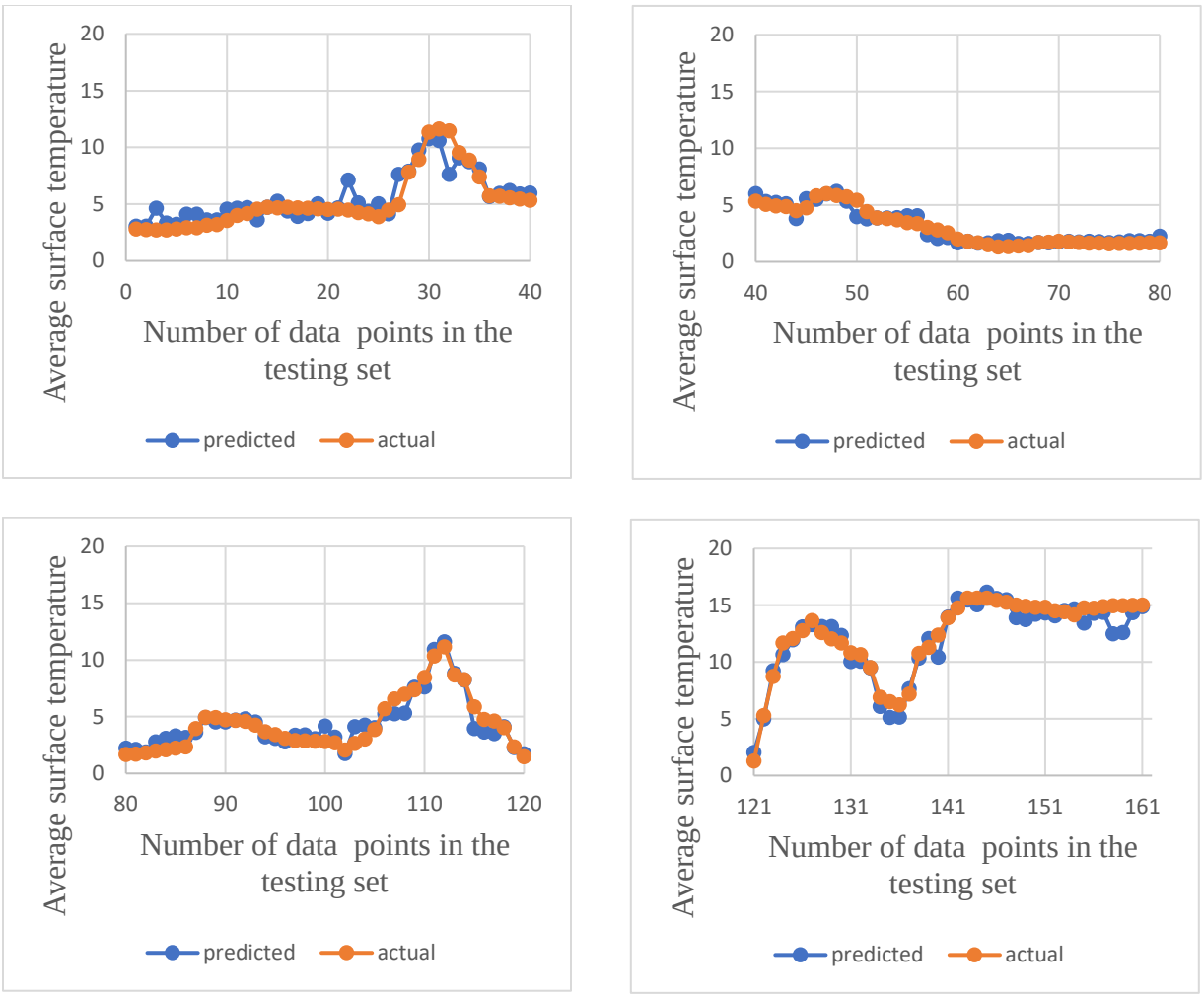

Figure 3.4: Actual versus predicted results with ANNs for the model with all features (80%/20%)

3.2 Temperature prediction using Random Forest

Random Forest (RF) is a supervised learning algorithm that uses an ensemble learning method for both regression and classification. The ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model [23]. RF is an integrated machine-learning algorithm that consists of decision trees and bagging [24, 25]. In bagging, each tree is constructed based on all possible features.

The RF method builds several decision trees on bootstrapped training samples. But when building

these decision trees, each time a split in a tree is considered, a random sample of *m* predictors is chosen as split candidates from the full set of *p* predictors.

In this study, the RF was performed using the full set of predictors, $p = 4$. When building the decision trees, a random sample of $m = 2$ predictors was chosen as a split candidate. The results of the variable importance using variance important plots for the dataset are shown in Table 3.2 /Figure 3.5. The air temperature was found to be the most important feature in predicting the average surface temperature. Additionally, the model accuracy using the RF methods was reported in Table 3.3. The reduction in the MSPE from the ANNs methods (Table 3.1) shows how strong RF methods are in predicting the average surface temperature.

3.2.1 Variable importance using the Random Forest

Two variable importance measures are reported based on out-of-bag (OOB) observations. The first one indicates a mean decrease in prediction error, and the second measure indicates a decrease in node impurity.

Table 3.2 and Figure 3.5 shows feature importance using variance important plots for the entire dataset. Air temperature and humidity are placed at the top of the variance importance plots indicating two top important features for the response variable average surface temperature.

Table 3.2: Feature importance using variance important plot.

Variable	$\%IncMSE$	***Variable***	***IncNodePurity***
Air Temp.	150.45786	Air Temp.	46.046644
Humidity	55.32753	Humidity	11.653058
Wind speed	47.00688	DNI	4.229818
DNI	22.32849	Wind speed	3.871396

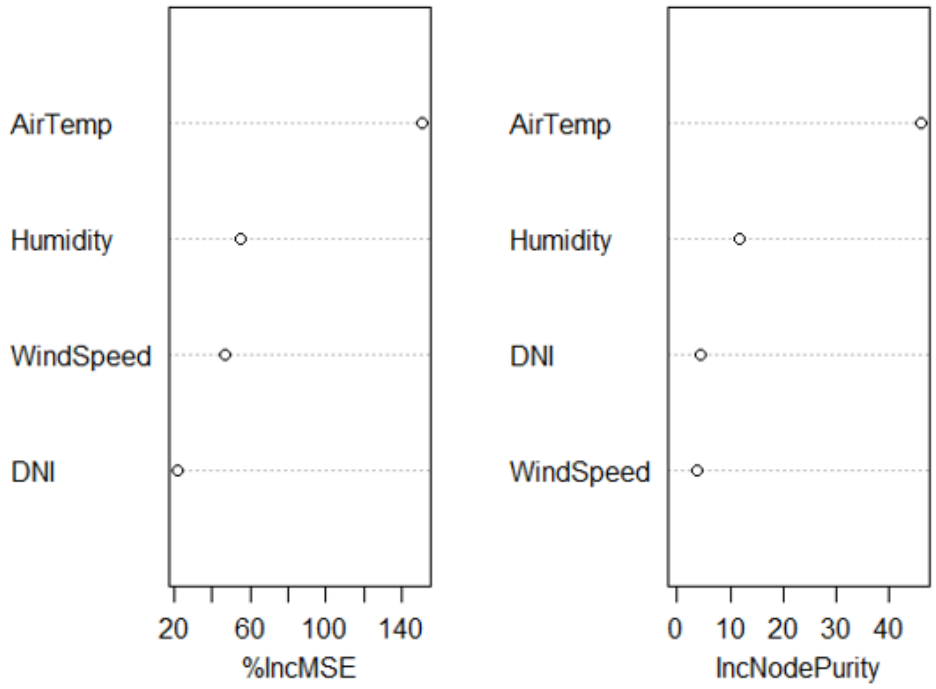

Figure 3.5: Feature importance using variance important plot.

3.2.2 Performance of Random Forest in the testing phase

The random forest was performed using all 4 features available. Error values between observed and predicted data were determined by Mean Square Prediction Error (MSPE), the model accuracy based on the Mean Square Deviation (R^2), the Mean Square Deviation (MAD, the Mean Absolute Percentage Error (MAPE), the Root Mean Square Error (RMSE), and the Mean Absolute Error (MAE) were reported. The MSPE was used to evaluate the accuracy of the predictions.

The mean absolute deviation MAD of a dataset is the average distance between each data point and the mean. $1 - \text{MAD}$ gives us an idea about the variability in a dataset which is represented by R^2 (Coefficient of determination). This coefficient varies from 0 to 1, where coefficients close to 1 indicate that most of the variabilities in the response variable can be explained by the predictors.

Table 3.3 shows the performance of the random forest models in the testing face using the two different samples: 70%/30% and 80%/20% train/test sets. The model with 70%/30% train/test set had MSPE of 0.2233 yielding an accuracy of 97.2% on a mean absolute deviation basis (i.e., the

average deviation between estimated and actual temperature stands at a mean of 2.8%). Also, the model with 80%/20% train/test set had MSPE of 0.2286 yielding an accuracy of 97.6%on a mean absolute deviation basis (i.e., the average deviation between estimated and actual temperature stands at a mean of 2.4%).

Table 3.3: Evaluation of the performance of Random Forest models in the testing phase.

MODELS	***MSPE***	R^2	**MAD**	***MAPE***	***RMSE***	***MAE***
70%/30%	0.2233	97.2%	0.028	0.071	0.473	0.320
80%/20%	0.2286	97.6%	0.024	0.066	0.478	0.298

3.2.3 Data visualization for Random Forest models: Actual average surface temperature versus predicted average surface temperature

To provide a virtual representation of the performance of the random forest models, scatter plots with smooth line markers were used to compare the actual and the predicted average surface temperature values. The blue trend indicates that of the predicted observations and the orange trend represents the actual values taking into consideration the number of features in each model. The figures show the number of data points in the testing set. The trend for observations 1-40, 41-80, 81-120, 121-160,161-200, and 201-242 were shown considering the ratio 70%/30%, and the trend for observations 1-40, 41-80, 81-120, and 121-161, were shown considering the ratio 80%/20%. The result from the random forest gives a trend of how the predicted values are closer to the actual observations for each set of features. Figures 3.6 and 3.7 give a trend of the actual and predicted values considering all 4 features for ratios 70%/30% and 80%/20% respectively.

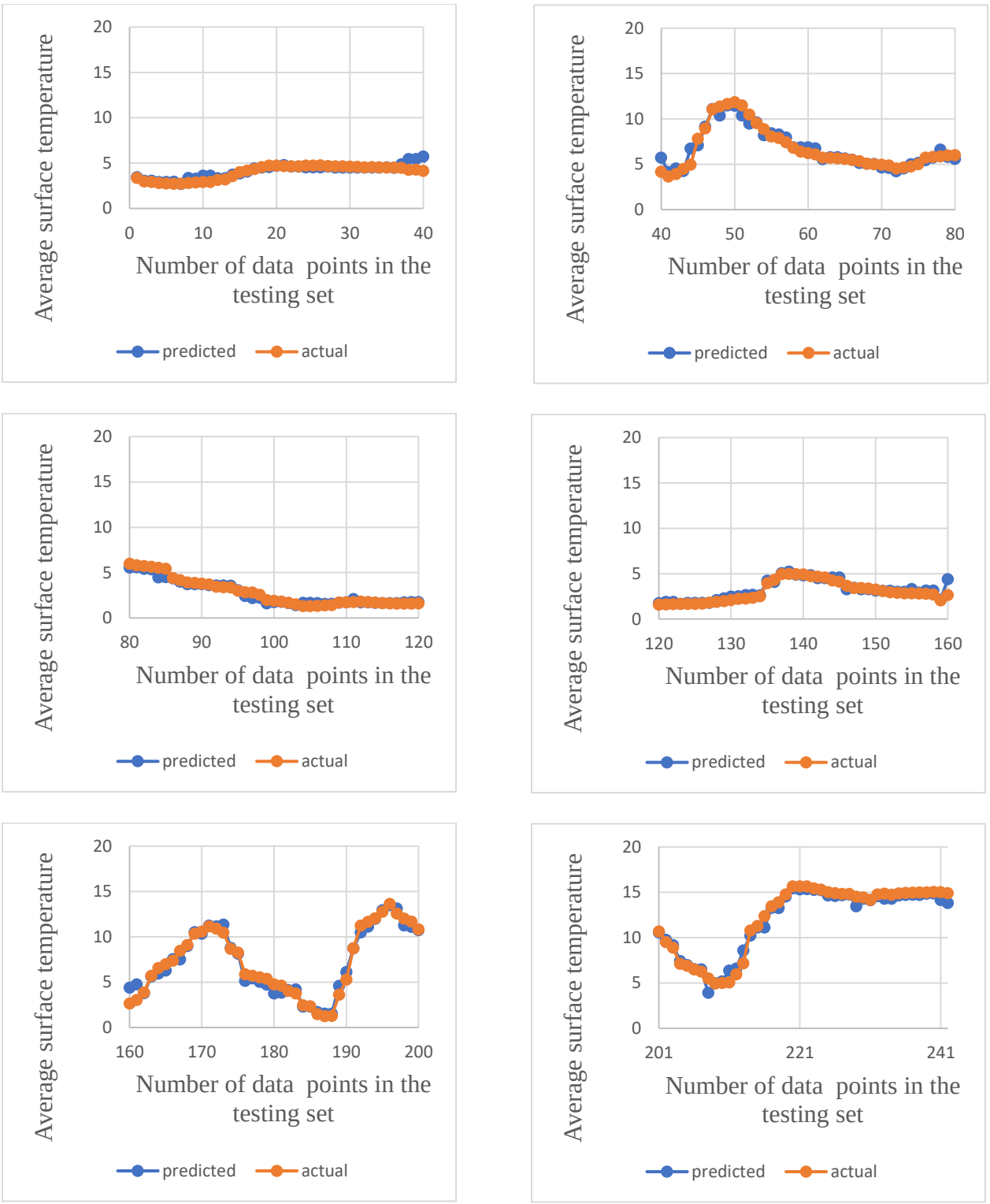

Figure 3.6: Actual versus predicted results with Random Forest for the model with all features (70%/30%)

Figure 3.7: Actual versus predicted results with Random Forest for the model with all features (80%/20%)

3.3 Conclusions

The result of this work for both ANNs and RF shows that the air temperature and humidity have the strongest positive and negative relationships, respectively, with the surface temperature as shown in Figures 3.1 and 3.2. However, the result of this work shows that the model from a ratio of 80%/20% had higher prediction accuracy than the model from a ratio of 70%/30% based on the MSPE, shown in Table 3.1 and Table 3.3. These results can be confirmed from the actual versus predicted values virtualization from Figures 3.3, 3.4, 3.6, and 3.7. The predicted values were closer to the actual values in the model from the ratio of 80%/20% than the model from the ratio of 70%/30% for both ANNs and RF.

However, comparing Tables 3.1 and 3.3, the random forest had a better prediction accuracy than the ANNs based on the MSPE criteria. This result can be confirmed from the data virtualizations of the actual versus predicted results from both ANNs and RF. Looking at Figures 3.3, 3.4, 3.6, and 3.7, the data points of the actual and predicted are closer using the random forest algorithm than the ANNs algorithm. In a nutshell, the random forest models will do a better job of predicting the average surface temperature given air temperature, humidity, wind speed, and DNI.

CHAPTER 4

In this chapter, we present a summary of our analysis and conclusions based on the artificial neural network and the random forest algorithms described in Chapter 2.

4.1 Summary of findings

There were 807 observations in the non-heated zone with the average surface temperature as the output feature and four input features: radiation from the sun (w/m^2), air temperature (℃), humidity (%) and wind speed (m/s). To achieve the most accurate prediction outcomes, both ANN and RF algorithms were used, and the data set was reshuffled 20 times. The data set was split into a training set and a testing set for each iteration of the shuffle. As a result, 70% of the data was used to train the model and 30% was used to test it in each cycle. When the 80% training set and 20% testing set ratio were used, the identical process was repeated. The input features were considered when making predictions. Before using the two chosen machine learning methods, Exploratory Data Analysis was conducted. The performance of Artificial Neural Networks (MLR) and Random Forest (RF) algorithms was evaluated using MSPE, R^2, MAD, MAE, RMSE, and MAPE prediction model evaluation criteria.

The result from the exploratory data analysis indicated that air temperature was the primary factor in forecasting the average surface temperature according to the correlation coefficients, which illustrate the linear relationship between each of the features. The air temperature had the highest correlation coefficient (0.850) with the average surface temperature. The p-values of radiation from the sun (DNI) and air temperature were statistically significant in predicting the surface temperatures of bridges. This implies that radiation from the sun and air temperature play important roles in predicting the average surface temperature of bridges in the non-heated zone.

The average surface temperature was predicted using all the input features with both the ANNs and RF algorithms. Using Olden's algorithm for the 70% training set and 30% testing set used to train the neural network showed that the air temperature has the strongest relationship with the average surface temperature with wind speed having the least relationship with the average surface temperature. Using the ANNs algorithm, the model with 70%/30% train/test set had an MSPE of 0.9170 yielding an accuracy of 95.67% on a mean absolute deviation basis (i.e., the average deviation between estimated and actual temperature stands at a mean of 4.33%). Also, the model

with an 80%/20% train/test set had an MSPE of 0.691 yielding an accuracy of 95.24% on a mean absolute deviation basis (i.e., the average deviation between estimated and actual temperature stands at a mean of 4.76%). Data visualization for actual versus predicted average surface temperature gave a clear trend of how the predicted values are closer to the actual observations for the 80%/20% train/test set than the 70%/30% train/test set.

Using the RF algorithm, feature importance using the variance importance plot for both the 80%/20% train/test set and the 70%/30% train/test set showed that the air temperature is the most important feature for the response variable, average surface temperature. The RF Forest performed better in predicting the average surface temperature. The model with 70%/30% train/test set had an MSPE of 0.2233 yielding an accuracy of 97.2% on a mean absolute deviation basis (i.e., the average deviation between estimated and actual temperature stands at a mean of 2.8%). Also, the model with an 80%/20% train/test set had an MSPE of 0.2286 yielding an accuracy of 97.6% on a mean absolute deviation basis (i.e., the average deviation between estimated and actual temperature stands at a mean of 2.4%). Data visualization for actual versus predicted average surface temperature gave a clear trend of how the predicted values are closer to the actual observations for the 80%/20% train/test set than the 70%/30% train/test set.

4.2 Conclusions

Surface temperature plays an important role in homes, agriculture, industry, and engineering fields. Precise temperature prediction can help prevent losses, including monetary and human losses, and predict daily operations.

This study conducted a comparison of the performance of two machine learning algorithms on a problem of temperature prediction. The data set was divided into ratios of 70%/30% and 80%/20%. We were able to identify the input features that have a large impact on predicting the target feature using ANNs and RF. Data visualization of the actual versus the predicted average surface temperature was displayed.

We have demonstrated that it is possible to predict temperature using the ANNs and RF algorithms and the target feature was predicted by both models with high accuracy. However, based on our observations, RF was performing better than ANNs.

4.3 Future work

The black box character of a neural network is its biggest drawback. It has the ability to approximate any function, and study its structure but does not give any insights into the structure of the function being approximated. Additionally, fitting a neural network with an arbitrarily large number of parameters has clear predictive benefits but makes it harder to extract model information. In the future, we would like to apply other machine learning algorithms such as Multiple Linear Regression (MLR), Support Vector Regression (SVR), and the Decision Tree on the dataset to analyze the performances of these algorithms in temperature prediction problems. We would also like to apply the ANNs and RF machine learning algorithms to another temperature prediction problem to compare the performances of these algorithms in the testing phase.

References

[1] Qiu, X., Xu, W. Y., Zhang, Z. H., Li, N. N., & Hong, H. J. (2020). Surface temperature prediction of asphalt pavement based on GBDT. *IOP Conference Series: Materials Science and Engineering*, *758*(1), 012031. https://doi.org/10.1088/1757-899x/758/1/012031

[2] Hao, H., Yu, F., & Li, Q. (2021). Soil temperature prediction using convolutional neural network based on ensemble empirical mode decomposition. *IEEE Access*, *9*, 4084–4096. https://doi.org/10.1109/access.2020.3048028

[3] Tran, T. T., Bateni, S. M., Ki, S. J., & Vosoughifar, H. (2021). A review of neural networks for Air Temperature Forecasting. *Water*, *13*(9), 1294. https://doi.org/10.3390/w13091294

[4] Yang, T., Sun, F., Gentine, P., Liu, W., Wang, H., Yin, J., Du, M., & Liu, C. (2019). Evaluation and machine learning improvement of global hydrological model-based flood simulations. *Environmental Research Letters*, *14*(11), 114027. https://doi.org/10.1088/1748-9326/ab4d5e

[5] Lee, J., Kim, C.-G., Lee, J., Kim, N., & Kim, H. (2018). Application of artificial neural networks to rainfall forecasting in the Geum River basin, Korea. *Water*, *10*(10), 1448. https://doi.org/10.3390/w10101448

[6] Zou, Q., Xiong, Q., Li, Q., Yi, H., Yu, Y., & Wu, C. (2020). A water quality prediction method based on the multi-time scale bidirectional long short-term Memory Network. *Environmental Science and Pollution Research*, *27*(14), 16853–16864. https://doi.org/10.1007/s11356-020-08087-7

[7] Altan Dombaycı, Ö., & Gölcü, M. (2009). Daily means ambient temperature prediction using artificial neural network method: A case study of turkey. *Renewable Energy*, *34*(4), 1158–1161. https://doi.org/10.1016/j.renene.2008.07.007

[8] Alcobaça, E., Mastelini, S. M., Botari, T., Pimentel, B. A., Cassar, D. R., de Carvalho, A. C., & Zanotto, E. D. (2020). Explainable machine learning algorithms for predicting glass transition temperatures. *Acta Materialia*, *188*, 92–100. https://doi.org/10.1016/j.actamat.2020.01.047

[9] Radhika, Y., & Shashi, M. (2009). Atmospheric temperature prediction using support vector machines. *International Journal of Computer Theory and Engineering*, 55–58. https://doi.org/10.7763/ijcte.2009.v1.9

[10] Salcedo-Sanz, S., Deo, R. C., Carro-Calvo, L., & Saavedra-Moreno, B. (2015). Monthly prediction of air temperature in Australia and New Zealand with machine learning algorithms.Theoretical and Applied Climatology, 125(1-2), 13–25. https://doi.org/10.1007/s00704-015-1480-4

[11] Beck, M. W. (2018). NeuralNettools: Visualization and analysis tools for neural networks. *Journal of Statistical Software*, *85*(11). https://doi.org/10.18637/jss.v085.i11

[12] Rumelhart DE, Hinton GE, Williams RJ (1986). "Learning Representations by Back-Propagating Errors." Nature, 323(6088), 533–536. https://doi:10.1038/323533a0

[13] Ripley BD (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge.

[14] Hornik K (1991). "Approximation Capabilities of Multilayer Feedforward Networks." Neural Networks, 4(2), 251–257. https://doi:10.1016/0893-6080(91)90009-t.

[15] Habibzadeh-Bigdarvish, O., Yu, X., Li, T., Lei, G., Banerjee, A., & Puppala, A. J. (2020). A novel full-scale external geothermal heating system for bridge deck de-icing. *Applied Thermal Engineering*, *185*, 116365. https://doi.org/10.1016/j.applthermaleng.2020.116365

[16] Cleveland, C. J., & Morris, C. (2013). Handbook of Energy. *Amsterdam: Elsevier*. https://doi.org/10.1016/C2009-0-16729-6 Pages 405-450

[17] Specht, D. F. (1991). A general regression neural network. *IEEE Transactions on Neural Networks*, *2*(6), 568–576. https://doi:10.1109/72.97934

[18] Gill NS, Mittal P. A 2016 *J Theor Appl Inf Technol*. **87** (1) 1–10.

[19] Larasati, A., Hajji, A. M., & Dwiastuti, A. (2019). The relationship between data skewness and accuracy of Artificial Neural Network Predictive model. *IOP Conference Series: Materials Science and Engineering*, *523*(1), 1–5. https://doi:10.1088/1757-899x/523/1/012070

[20] Beckmw. (2014, June 21). Variable importance in neural networks. *R is my friend*. beckmw. https://beckmw.wordpress.com/2013/08/12/variable-importance-in-neural-networks/. Accessed 1 December 2022

[21] Beck, M. W. (2018). Visualization and analysis tools for neural networks. *Journal of Statistical Software*, *85*(11), 1–20. https://doi:10.18637/jss.v085.i11

[22] Olden JD, Joy MK, Death RG (2004). "An Accurate Comparison of Methods for Quantifying Variable Importance in Artificial Neural Networks Using Simulated Data." *Ecological Modelling*, 178(3–4), 389–397. https://doi:10.1016/j.ecolmodel.2004.03.013

[23] Chaya. (2022, April 14). *Random Forest regression*. Medium. Retrieved January 5, 2023, from https://levelup.gitconnected.com/random-forest-regression-209c0f354c84#:~:text=Random%20Forest%20Regression%20is%20a%20supervised%20learning%20algorithm%20that%20uses,prediction%20than%20a%20single%20model

[24] Breiman L., Random forests, *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[25] Hutengs, C., & Vohland, M. *Downscaling land surface temperatures at regional scales with random forest regression, Remote Sens.* Environ., vol. 178, pp. 127–141, Jun. 2016.