INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

Rainfall Prediction Using Machine Learning Algorithms for the Various Ecological Zones of Ghana

K. Siva Krishna [1], K. Narayana Rao [2]
[1] Assistant professor, Professor [2]
Department of Computer Science Engineering
RISE Krishna Sai Prakasam Group of Institutions

ABSTRACT- Accurate rainfall prediction has become very complicated in recent times due to climate change and variability. The efficiency of classification algorithms in rainfall prediction has flourished. The study contributes to using various classification algorithms for rainfall prediction in the different ecological zones of Ghana. The classification algorithms include Decision Tree (DT), Random Forest (RF), Multilayer Perceptron (MLP), Extreme Gradient Boosting (XGB) and K-Nearest Neighbour (KNN). The dataset, consisting of various climatic attributes, was sourced from the Ghana Meteorological Agency spanning 1980 – 2019. The performance of the classification algorithms was examined based on precision, recall, f1-score, accuracy, and execution time with various training and testing data ratios. On all three training and testing ratios: 70:30, 80:20 and 90:10, RF, XGB and MLP performed well, whereas KNN performed least across all zones. In terms of the execution time of the models, Decision Tree is consistently portrayed as the fastest, whereas MLP used the most run time.

INDEX TERMS Machine learning, rainfall prediction; classification algorithms; ecological zones; rain/norain class.

I. INTRODUCTION

Accurate and timely rainfall prediction is expected to inject a new intervention phase to the affected sectors accosted with the negative propensities of rainfall extremes. These critical sectors include but are not limited to energy, agriculture, and others, which are greatly affected by rainfall. A plethora of scholarly research has demonstrated that the duration and intensity of rainfall cause major climate-related disasters [1] [2]. The manifestation of the impact of rainfall includes drought [3], floods [4], among others and its associated effects. For example, in 2009, torrential rains affected almost 600,000 people in Senegal, Niger, Burkina Faso and Ghana [1]. In addition, almost half a million people died through floods in 2007 recorded over Ethiopia, Uganda, Togo, Niger, Sudan, Mali and Burkina Faso [1]. Furthermore, findings from [5] project that the death of 30,000 to 50,000 children due to malnutrition in 2009 in sub-Saharan Africa may be worst due to the changes in the variability of rainfall coupled with the acute weather episodes affecting the agricultural sector. Apart from precious lives lost through floods, an ample body of literature has reported the impact of rainfall on other vital sectors of the Ghanaian economy [6][7][8]. In [8], it is reported that two major hydroelectricity plants which cater for over 70% of the electricity demand in Ghana is rainfall reliant. This presupposes that a decrease in rainfall has dire consequences on the electricity generation of the country. Agriculture, which employs about 44.7% of the Ghanaian labour force and contributes significantly to the nation's economy, is pivotal to the growth of the economy [9]. Despite its recent decline in performance, the agriculture sector remains a crucial element for poverty reduction and food security in Ghana [9]. However, Ghana's agriculture sector is mainly rain-fed, with about 3% of the cultivable land supported with irrigation [9]. Meanwhile, in developing countries, including Ghana, the primary water source for agriculture, hydropower generation, and others is rainfall. Many classification algorithms such as Random Forest (RF), Decision Tree (DT), Neural Network (NN), K-Nearest Neighbour (KNN) and others have been investigated for the prediction of rainfall. The performance among these algorithms widely varies, leaving room for enhancement by varying training and testing ratios or combining different techniques. However, rainfall prediction continues to be a challenging task. Therefore, selecting suitable methods in classifying rainfall over a region is vital. Meanwhile, machine learning algorithms have been proposed to enhance rainfall prediction accuracy [10]. For this reason, rainfall prediction based on various techniques for countless locations such as Malaysia, India, Egypt and others is replete. For instance, [11] used machine learning techniques to build rainfall prediction models in some major cities in Australia by comparing Decision trees, Random Forest, Logistic regression, AdaBoost, Gradient boosting and KNearest Neighbour. In a similar comparative study, [12] reported that random forest showed an accuracy of 87.1% for a weather prediction model compared to the C4.5

decision tree algorithm, which gave an accuracy of 82.4%. In Malaysia [13], a rainfall prediction model involving different classification algorithms revealed Neural Networks as the best based on the performance of the evaluation metrics compared to the others. The findings from the study showed Neural Networks with an F-score of 73.2%, which was the highest. Insights from these studies suggest that machine learning algorithms perform well regarding rainfall prediction accuracy and timeliness. Therefore, it is imperative to investigate various classification algorithms to establish the best performing techniques for predicting rainfall in Ghana. The current study seeks to employ multiple rainfall prediction techniques from Ghana Meteorological Agency (GMet) rainfall data. After that, a comparative analysis of rainfall prediction from Random Forest, Decision Tree, KNearest Neighbour, Extreme Gradient Boosting and Multilayer Perceptron. The remaining sections of this paper are as follows: a brief description of the study area and the data source are presented in section 2. The methodology utilized is given in section 3. Section 4 constitutes the results and discussions, and finally, the study's conclusions are given in section 5.

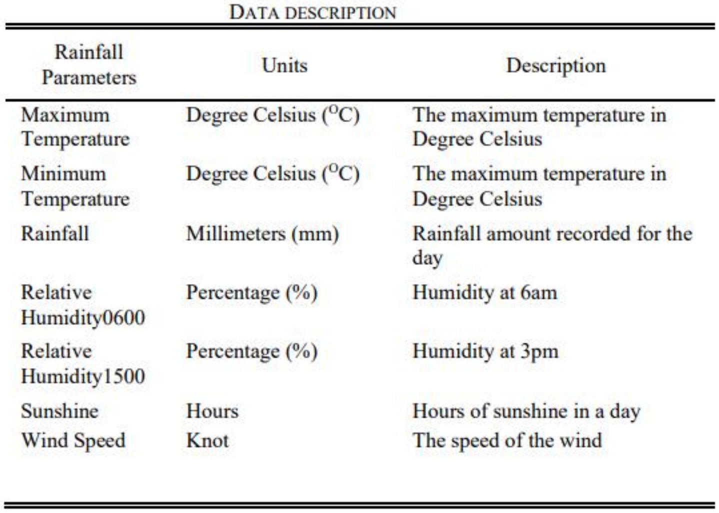
II. STUDY AREA AND DATA SOURCE

STUDY AREA Ghana has been grouped into four (4) agro-ecological zones according to the Ghana Meteorological Agency classification. Namely: Coastal, Forest, Transition and Savannah zones [14]. The zones are specified by distinct climate conditions [15]. Ghana is characterized by two main rainfall regimes resulting from the Inter-Tropical Discontinuity (ITD) [16]. The two rainfall regimes are the bi-modal and uni-modal rainfall patterns. The coastal and forest zones experience a bi-modal rainfall pattern, whereas the transition and savannah zones are characterized by a unimodal rainfall pattern [17]. The mean annual rainfall of the Savannah zone per year is about 1100 mm and in comparison, with the other zones, the Savannah is characterized with warm temperatures all year round. In view of the climatic condition over the zone, leading crops cultivated in this zone includes but not limited to sorghum and millet [18]. Meanwhile, the Transitional zone which is sandwiched between the Savannah and the Forest zones obtains a mean annual rainfall per year of about 1300 mm. The climate of this zone exhibits climatic conditions of both Savannah and Forest zones due to its location. Annual food crops such as plantain and maize are dominant in this zone [18]. The highest mean annual rainfall is recoded over the Forest zone which is 2200 mm per year. This zone is located in the Southwestern part of Ghana and predominantly wet throughout the year. The mean annual rainfall received in the Coastal zone per year which is largely modulated by the circulation of land-sea breeze is about 900 mm.

B. DATA SOURCE

The study will employ rainfall, temperature (minimum and maximum), relative humidity (at 1500 and 0600), Sunshine hours and wind speed data from the 22 synoptic stations across the four ecological zones spanning 1980 – 2019 sourced from the Ghana Meteorological Agency. Figure 1 shows the location of all the 22 stations across Ghana. Weather parameters measured by the GMet is according to the World Meteorological Organization (WMO) standards [14].

TABLE I
DATA DESCRIPTION



Rainfall Parameters	Units	Description
Maximum Temperature	Degree Celsius (^{O}C)	The maximum temperature in Degree Celsius
Minimum Temperature	Degree Celsius (^{O}C)	The maximum temperature in Degree Celsius
Rainfall	Millimeters (mm)	Rainfall amount recorded for the day
Relative Humidity0600	Percentage (%)	Humidity at 6am
Relative Humidity1500	Percentage (%)	Humidity at 3pm
Sunshine	Hours	Hours of sunshine in a day
Wind Speed	Knot	The speed of the wind

III. METHODOLOGY

In this section we describe the techniques and tools that have been employed to utilize the various weather features for rainfall prediction.

A. CLASSIFICATION FRAMEWORK

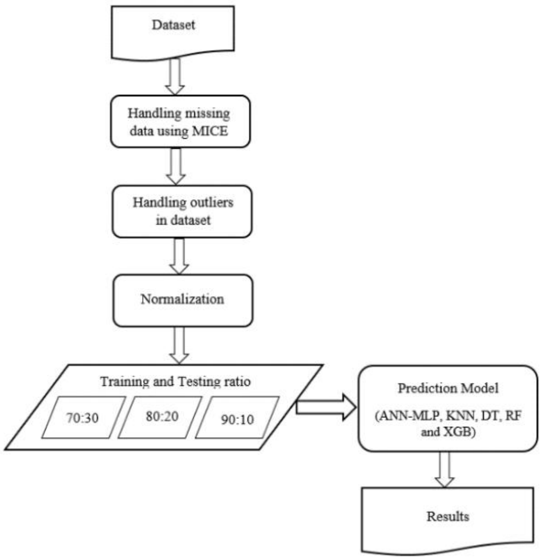

Dataset
Handling missing data using MICE
Handling outliers in dataset
Normalization
Training and Testing ratio
70:30
80:20
90:10
Prediction Model
(ANN-MLP, KNN, DT, RF and XGB)
Results

FIGURE 2. Classification framework.

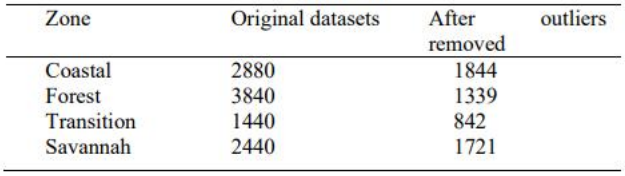
B. I. DATA EXPLORATORY AND ANALYSIS

To achieve high certainty of the validity of future results, Exploratory Data Analysis (EDA) is key to machine learning tasks [19]. Pre-processing techniques are essential for a steady classification proceeding in addition to the generation of satisfactory results. In view of this, various data preprocessing techniques were performed on the dataset. Firstly, the Multiple Imputation by Chained Equations (MICE) package was utilized to impute missing values. MICE is a robust means of filling in missing data in a dataset through an iterative process. A significant advantage of MICE over other missing value approaches is that, by multiple imputations it fills in the missing values multiple times leading to achieving a complete dataset [20]. Secondly, another most important phase of the EDA is to identify and remove outliers in the datasets that can affect the model. The study employed the mathematical function approach in doing this. Specifically, the Inter Quartile Range (IQR) score was used to detect and remove outliers from the datasets using the relation:

$$IQR = Q3 - Q1$$

Where IQR is the interquartile range which equal to the first quartile Q1 subtracted from the third quartile Q3. In other words, the IQR is equal to the difference between the 75th and the 25th percentiles. Table 2 is the summary of the raw datasets and after outliers are removed for all the ecological zones.

TABLE 2
SUMMARY OF ORIGINAL DATASETS AND SHAPE OF DATA AFTER OUTLIERS REMOVED



Zone	Original datasets	After outliers removed
Coastal	2880	1844
Forest	3840	1339
Transition	1440	842
Savannah	2440	1721

Further, to check for multi-collinearity among the variables, pair-wise correlation matrix and a corresponding pairplot were constructed across all ecological zones. Also, oversampling of minority class was employed to handle the issue of class imbalance in the target variable. Since imbalanced data can generate biased results in the model due to model's inability to learn much about the minority class [11].

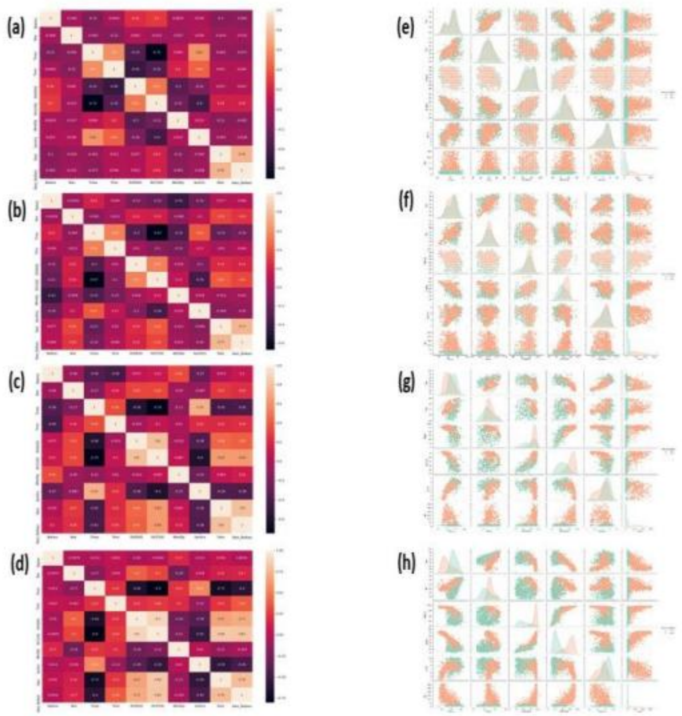


FIGURE 3. Correlation heatmap and pairplots of variables across the 4 ecological zones of Ghana. Panels a – d are the correlation matrix depicting the correlation among the variables. With a, b, c and d representing Coastal, Forest, Transition and Savannah zones respectively. Panels e – h are the corresponding pairplots

Data normalization or feature scaling is essential since the mathematical processes in machine learning relies on Euclidean distance between two data points. This is important since features used in this current study are characterized with varying magnitudes. Adopting normalization will scale the dataset into weight lying between 0 and 1.

This enhances the data to utilize a standard scale with distortions or loss of vital information. The Min-Max normalization scaler was calculated by using:

$$z = \frac{x - min\,(x)}{max\,(x) - min\,(x)}$$

Where x, min (x) and max (x) represents the value to be scaled, minimum and maximum values respectively. II. Models The study employed 5 classification algorithms. In selecting the classifiers, this current study adopted the model family approach in [11]. These include Tree-based, distance-based, ensemble and a deep learning model. Scikit-learn was used to implement the classifiers. The details of the classification algorithms employed are given below: Artificial Neural Network-Multi-Layer Perceptron: As the most extensively used machine learning algorithm [21], Artificial Neural Network (ANN) is characterized with various types which has been employed in various aspects of research [21]. One of such is the Multi-Layer Perceptron (MLP). In hydro-climatological research, MLP is employed to establish the relationship between predictors and predictands [22]. MLP is a classical structure of Deep Neural Network (DNN) [23]characterized with several layers with numerous neurons. The first layer of MLP is known as the input layer whereas the last layer represents the output layer. The layers which are located in the middle are known as the hidden layers. Using an activation function, the hidden layers merges weights and bias terms with inputs to generate the output. With n number of inputs x= (x1 ,x2 ,… xn) with a vector of weights wj= (w 1j ,w2j ,… wnj) for a given node j. To determine the simulated y j at the node j is given by the equation below [24].

$$y_j = f\,(x.\,w_j - b_j)$$

Where f (.) is the activation function, *w*j as the weight vector and the bias related to the node represented as bj . Therefore the output y k is computed by the equation below:

$$Y_k = f_2\left[\sum_{i=1}^{m} w_{ki}\; f_1\left(\sum_{j=1}^{n} w_{ij}\; x_j \;+\; b_i\right) + b_k\right]$$

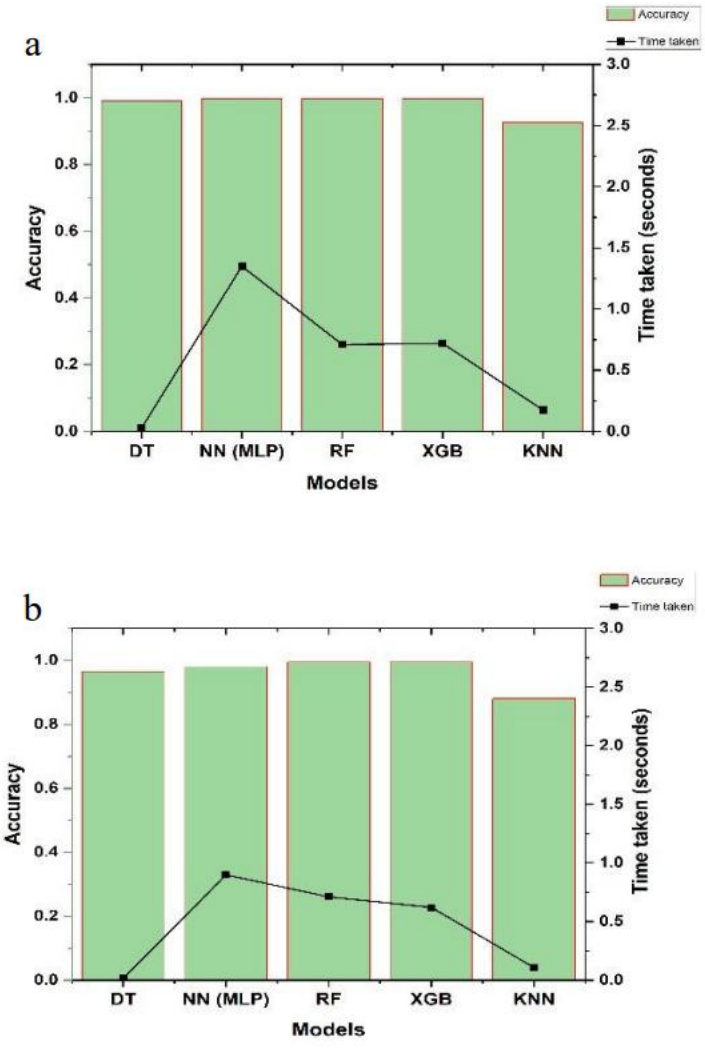
Where f 1 and f 2 represents the activation functions, also, j,i and k refers to the input, hidden and output layers respectively. xj indicates the inputs, the bias linked to the hidden and output layers are assigned with bi and bk respectively. Further, n and m point out to the neurons in the input and hidden layers respectively. The weights between the hidden and input layers is denoted by wij whereas the weights between the hidden and output layers denoted by wki . Yk represents the output of the network. Hereafter, Artificial Neural Network-Multi-Layer Perceptron used in this current study is referred to as Multi-Layer Perceptron (MLP). MLP algorithm was applied on the three training and testing ratios. K-Nearest Neighbour: K-Nearest Neighbour (KNN) is a non-parametric learning algorithm which uses euclidean, manhattan and minkowski distances approach in making classification [25]. It's been reported that KNN performs better with minimal number of features [11]. The Euclidean distance is calculated using equation 4 as shown below. Where xij and xio refers to the i ith data point in the jth predictor and predictand.

Zr and Zk represents the predicted and neighboring data respectively whereas f k (dj) is the kernel function. This study employs 7 meteorological features which makes KNN a favorable candidate for this current study. According to [26], the performance of KNN in modelling is dependent on the number of neighbors (K) utilized. The value of K was set to 5 in this current study after preliminary assessment. KNN with K = 5 was applied on all three training and testing ratios. Decision Tree: Decision tree algorithm is used for both classification and regression in machine learning. In a decision tree, each node in a branch serves as a choice of alternative whereas leaf nodes signifies a decision. Decision performs well with both categorical and continuous variables which fits well with in this current study since our target variable (rainfall) is binary categorical. In building decision trees, the known algorithms utilized include C5.0, Chi-squared Automatic Interaction Detection (CHAID), ID3, Quest, Classification and Regression Trees (CART) and C4.5. The C5.0 was selected for this current study and applied on the three training and testing ratios. C5.0 is an enhanced algorithm from the previous C4.5 and ID3.

Random Forest: Random Forest (RF) was developed by Breiman in 2001 for classification. It's an ensemble machine learning algorithm that uses numerous classification trees thereby gaining the name random forest. In the computation of regression, it merges different decision trees for regression and classification purposes [25]. In spite of its bias towards variables with high levels amongst categorical variables with varying levels, RF algorithm is weighed to be an extremely rigorous learning algorithm in recent times [27]. The function of the RF algorithm involves first of all the collection of random samples from the given data. A decision tree will then be created for each sample in the second stage. Afterwards, voting is done for the predicted results and finally, the classification with most voted prediction is chosen. The RF algorithm was applied on all the three training and testing ratios. Our model configuration used in this current study adopted the number of weak learners to be 100 and maximum depth of tree to be 16. Extreme Gradient Boosting: The Extreme Gradient Boosting (XGBoost) is an advanced machine learning technique which hinges on the gradient boosting algorithm developed by [28]. XGBoost is a better handler of overfitting through model formalization. This algorithm was selected for this current study due to its high execution speed. XGB was applied on all three training and testing ratios. III. Evaluation metrics To evaluate the efficiency of the various algorithms can be done by numerous evaluation metrics. However, this current study focuses on accuracy, precision, recall, f-measure and the confusion matrix which is the basis for the previous metrics. The metrics are therefore defined as: Confusion matrix: The confusion matrix yields an output in a matrix form which details the model's performance.

IV. Results

However, at the coastal zone, the execution time of xgboost and random forest are comparable. On the same 70:30 ratio, transitional and savannah zones also showed the consistency of the decision tree in terms of speed in model execution (See Figures 4c and 4d). Meanwhile, as compared to coastal and forest zones, random forest used more time to execute the model at the transitional zone whereas at the savannah zone, MLP regained its position as the model with the longest execution time. Generally, from Fig. 4, it can be seen that the overall model accuracy of MLP, RF and XGB were high at all the zones on the 70:30 ratio whereas consistently KNN performed least. Comparing the execution times and model accuracy levels at all the zones on 80:20 training and testing ratio, decision tree shows consistency as the model with the fastest time of execution at all zones (See Figure 5). Meanwhile, in terms of model accuracy, DT exhibited good model accuracy levels with the exception of the savannah zone where it performance was low. (See Figure 5d). From Fig. 6, it can be observed that, in spite of the longer time of execution of MLP at all zones, it performed best in terms of model accuracy as compared to the other models. Again, on the 90:10 ratio, decision tree stood out as the fastest model execution. Overall, decision tree has shown to be a good candidate in terms of timeliness in rainfall prediction over all ecological zones of Ghana. However, in terms of accuracy the MLP, XGB and RF have been pronounced.



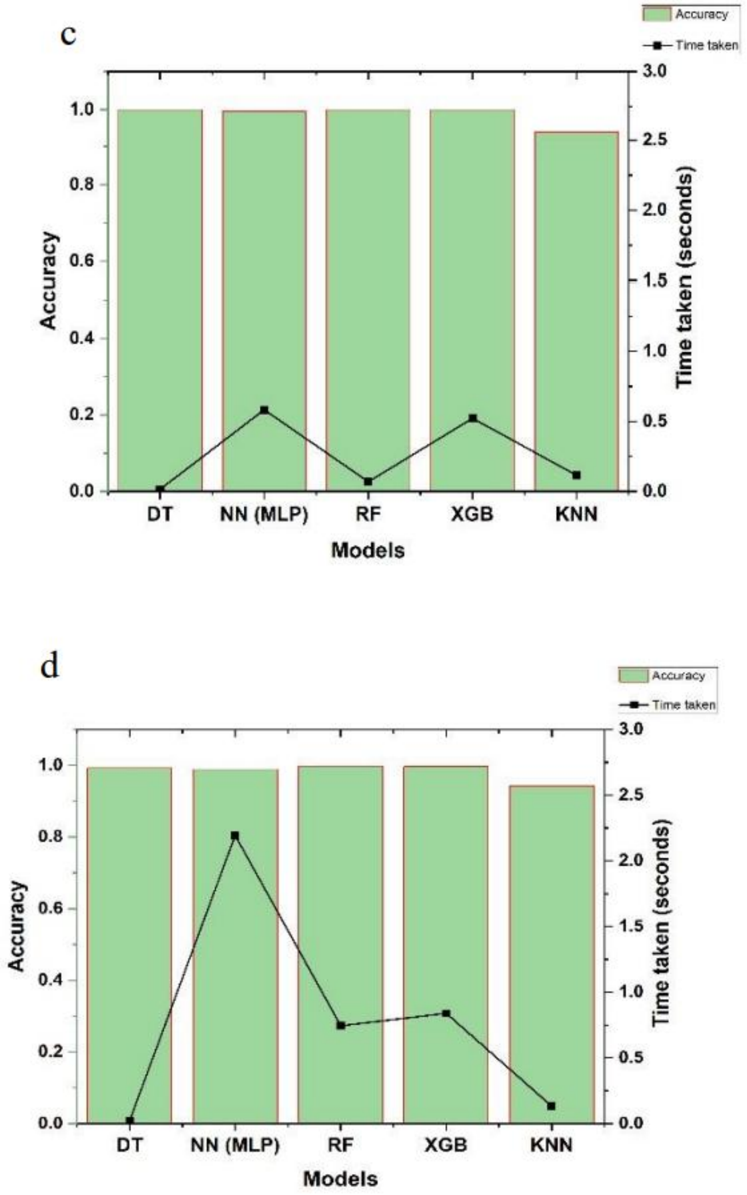


FIGURE 4 Comparison of accuracy and time taken for models execution for 70:30 ratio. Panels a, b, c and d represents Coastal, Forest, Transitional and Savannah zones respectively

V. SUMMARY AND CONCLUSION

This research executed rainfall prediction in Ghana covering all the ecological zones using five (5) classification algorithms namely: Decision Tree, Random Forest, Multilayer Perceptron, Extreme Gradient Boosting and KNearest Neighbour. 41 years of past climatic data spanning 1980 – 2019 from the Ghana Meteorological Service was used for this study. To evaluate the performance of the classifiers, the evaluation metrics employed included precision, f1-score and recall with results presented in tables. Further, the overall accuracy of the model and the execution times of the individual models were also ascertained and the results are shown in figures. To ensure effective rainfall prediction, input datasets went through the exploratory data analysis where the multiple imputation by chained equations algorithm was used replace missing data, outliers were removed from the datasets and normalized before the classification stage. The datasets were splitted into two parts: training datasets and testing datasets. We employed 3 different types of training and testing ratios (training data: testing data): 70:30, 80:20 and 90:10 to analyze the performance of the classification algorithms on different training and testing ratios. Findings from the study showed distinct characteristics of classification of the rain and no-rain classes in the various ecological zones in the country. No- rain class was well classified by classifiers in the coastal zone as compared to the rain class. However, there was an opposite response in the forest zone. In the forest zone, the classifiers performance was best with regards to the rain class. At the savannah zone, all classifiers on the 3 training and testing ratios performed well in classifying the no-rain class which is consistent with low rain pattern observed the region. On all ecological zones and putting together all training and testing ratios, decision tree distinguished itself as the model with the fastest execution time whereas multilayer perceptron performed poorly in terms of time of execution. Generally, random forest, extreme gradient boosting and multilayer perceptron performed well in all instances which is suggestive that ensemble and deep learning models are good candidates for rainfall prediction. However, K-Nearest Neighbour performed worst in all zones on all training and testing ratios which warrants further investigation. Further study using other classification algorithms and a hybrid model at different training and testing ratios for rainfall prediction in all ecological zones of Ghana is under consideration.

REFERENCES

[1] G. Di Baldassarre, A. Montanari, H. Lins, D. Koutsoyiannis, L. Brandimarte, and G. Blöschl, “Flood fatalities in Africa: from diagnosis to mitigation,” Geophys. Res. Lett., vol. 37, no. 22, 2010.

[2] N. K. Karley, “Flooding and physical planning in urban areas in West Africa: situational analysis of Accra, Ghana,” Theor. Empir. Res. Urban Manag., vol. 4, no. 4 (13, pp. 25–41, 2009.

[3] R. C. Deo, S. Salcedo-Sanz, L. Carro-Calvo, and B. Saavedra-Moreno, “Drought prediction with standardized precipitation and evapotranspiration index and support vector regression models,” in Integrating disaster science and management, Elsevier, 2018, pp. 151–174.

[4] D. T. Bui, P. Tsangaratos, P.-T. T. Ngo, T. D. Pham, and B. T. Pham, “Flash flood susceptibility modeling using an optimized fuzzy rule based feature selection technique and tree based ensemble methods,” Sci. Total Environ., vol. 668, pp. 1038–1054, 2019.

[5] I. Yabi and F. Afouda, “Extreme rainfall years in Benin (West Africa),” Quat. Int., vol. 262, pp. 39–43, 2012.

[6] C. Kyei-Mensah, R. Kyerematen, and S. AduAcheampong, “Impact of rainfall variability on crop production within the Worobong Ecological Area of Fanteakwa District, Ghana,” Adv. Agric., vol. 2019, 2019.

[7] P. A. Williams, O. Crespo, C. J. Atkinson, and G. O. Essegbey, “Impact of climate variability on pineapple production in Ghana,” Agric. Food Secur., vol. 6, no. 1, pp. 1–14, 2017.

[8] K. Owusu and N. A. B. Klutse, “Simulation of the Rainfall Regime over Ghana from CORDEX,” 2013.

[9] S. MoFA, “Agriculture in Ghana: facts and figures,” Minist. Food Agric. Accra. 47p, 2012.

[10] H. Meyer, C. Reudenbach, T. Hengl, M. Katurji, and T. Nauss, “Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation,” Environ. Model. Softw., vol. 101, pp. 1–9, 2018.

[11] N. Oswal, “Predicting rainfall using machine learning techniques,” arXiv Prepr. arXiv1910.13827, 2019.

[12] S. Karthick, D. Malathi, and C. Arun, “Weather prediction analysis using random forest algorithm,” Int. J. Pure Appl. Math., vol. 118, no. 20A, pp. 255–262, 2018.

[13] N. SamsiahSani, I. Shlash, M. Hassan, A. Hadi, and M. Aliff, “Enhancing Malaysia Rainfall Prediction Using Classification Techniques,” J. Appl. Environ. Biol. Sci, vol. 7, no. 2S, pp. 20–29, 2017. [14] L. K. Amekudzi et al., “Variabilities in rainfall onset, cessation and length of rainy season for the various agroecological zones of Ghana,” Climate, vol. 3, no. 2, pp. 416–434, 2015.

[15] M. New, M. Todd, M. Hulme, and P. Jones, “Precipitation measurements and trends in the twentieth century,” Int. J. Climatol. A J. R. Meteorol. Soc., vol. 21, no. 15, pp. 1889–1922, 2001.

[16] K. Owusu and P. R. Waylen, “The changing rainy season climatology of mid-Ghana,” Theor. Appl. Climatol., vol. 112, no. 3–4, pp. 419–430, 2013.

[17] C. Mensah, L. K. Amekudzi, N. A. B. Klutse, J. N. A. Aryee, and K. Asare, “Comparison of rainy season onset, cessation and duration for Ghana from RegCM4 and GMet datasets,” 2016.

[18] M. Baidu, L. K. Amekudzi, J. N. A. Aryee, and T. Annor, “Assessment of long-term spatio-temporal rainfall variability over Ghana using wavelet analysis,” Climate, vol. 5, no. 2,

ISSN: 2057-5688

p. 30, 2017. [19] C. Room, “Exploratory Data Analysis,” Mach. Learn., vol. 8, no. 39, p. 23, 2020.

[20] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, “Multiple imputation by chained equations: what is it and how does it work?,” Int. J. Methods Psychiatr. Res., vol. 20, no. 1, pp. 40–49, 2011. [21] R. Kaur and S. Sharma, “An ann based approach for software fault prediction using object oriented metrics,” in International Conference on Advanced Informatics for Computing Research, 2018, pp. 341–354.

[22] K. Ahmed, D. A. Sachindra, S. Shahid, Z. Iqbal, N. Nawaz, and N. Khan, “Multi-model ensemble predictions of precipitation and temperature using machine learning algorithms,” Atmos. Res., vol. 236, p. 104806, 2020.

[23] T. Shimobaba et al., “Deep-learning-based data page classification for holographic memory,” arXiv Prepr. arXiv1707.00684, 2017.

[24] T.-W. Kim and J. B. Valdés, “Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks,” J. Hydrol. Eng., vol. 8, no. 6, pp. 319–328, 2003.

[25] M. Bowles, Machine Learning with Spark and Python: Essential Techniques for Predictive Analytics. John Wiley & Sons, 2019.

[26] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, “Efficient kNN classification with different numbers of nearest neighbors,” IEEE Trans. neural networks Learn. Syst., vol. 29, no. 5, pp. 1774–1785, 2017. [27] A. Cutler, D. R. Cutler, and J. R. Stevens, “Random forests,” in Ensemble machine learning, Springer, 2012, pp. 157–175.

[28] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.