IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 6, 2021, accepted December 19, 2021, date of publication December 28, 2021,
date of current version January 14, 2022.

Digital Object Identifier 10.1109/ACCESS.2021.3139312

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

NANA KOFI AHOI APPIAH-BADU[1,2], YAW MARFO MISSAH[1], LEONARD K. AMEKUDZI[3], NAJIM USSIPH[1], TWUM FRIMPONG[1], AND EMMANUEL AHENE[1]

[1]Department of Computer Science, Kwame Nkrumah University of Science and Technology (KNUST), Kumasi, Ghana
[2]Department of Computer Science, Christian Service University College, Kumasi, Ghana
[3]Meteorology and Climate Science Unit, Department of Physics, KNUST, Kumasi, Ghana

Corresponding author: Emmanuel Ahene (aheneemmanuel@knust.edu.gh)

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 Rainfall prediction, classification algorithms, ecological zones, rain/no-rain 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]–[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

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

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 K-Nearest 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. Therefore in this paper, we put forward the following contributions:

1) We employ data preprocessing techniques on Ghana Meteorological Agency (GMet) climatic data from the 22 synoptic stations across the four ecological zones of Ghana.
2) We employ classification algorithm such as Decision Tree (DT), Random Forest (RF), Multilayer Perceptron (MLP), Extreme Gradient Boosting (XGB) and K-Nearest Neighbour (KNN) to build the models for rainfall prediction in the different ecological zones.
3) We perform the evaluation of the models based on precision, recall, accuracy and f1 score.

The remaining sections of this paper are as follows: The next section focuses on a brief description of the study area and the data source. The methodology utilized is given in section III. Section IV constitutes the results and discussions, and finally, the study's conclusions are given in section V.

II. STUDY AREA AND DATA SOURCE

A. 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 uni-modal 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.

TABLE 1. 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

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]. The datasets are listed in Table 1.

The datasets used in this study include: Temperature, Rainfall, relative humidity, Sunshine hours and wind speed as listed in Table 1.

III. METHODOLOGY

A. CLASSIFICATION FRAMEWORK

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

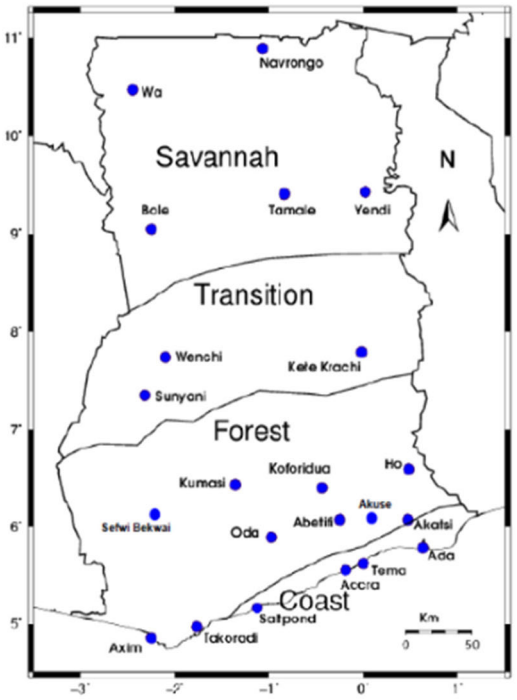


FIGURE 1. Map of Ghana showing all twenty-two Ghana Meteorological Agency Synoptic stations. Adapted from [17].

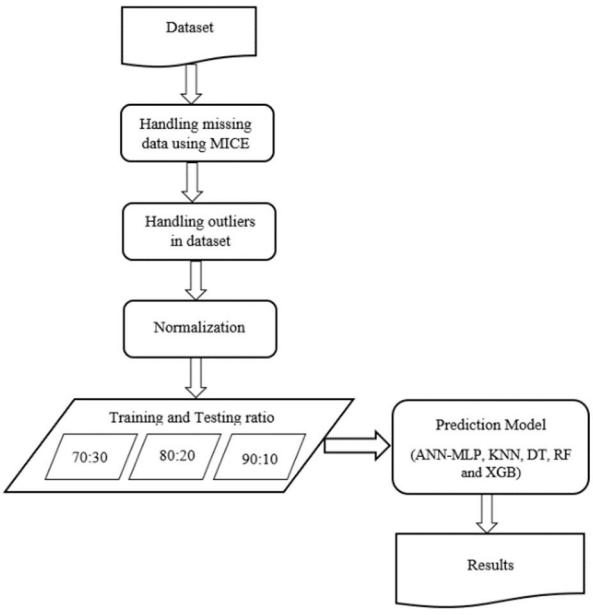


FIGURE 2. Classification framework.

rainfall prediction. The classification framework used in this paper as shown in Figure 2 involves handling of missing data, handling of outliers, normalization, training and testing, the implementation of the prediction model and the results of the performance of the models.

B. 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 pre-processing 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:

$$IQR = Q3 - Q1 \tag{1}$$

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, as shown in Figure 3, 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]. 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:

$$z = \frac{x - min(x)}{max(x) - min(x)} \tag{2}$$

where x, min (x) and max (x) represents the value to be scaled, minimum and maximum values respectively.

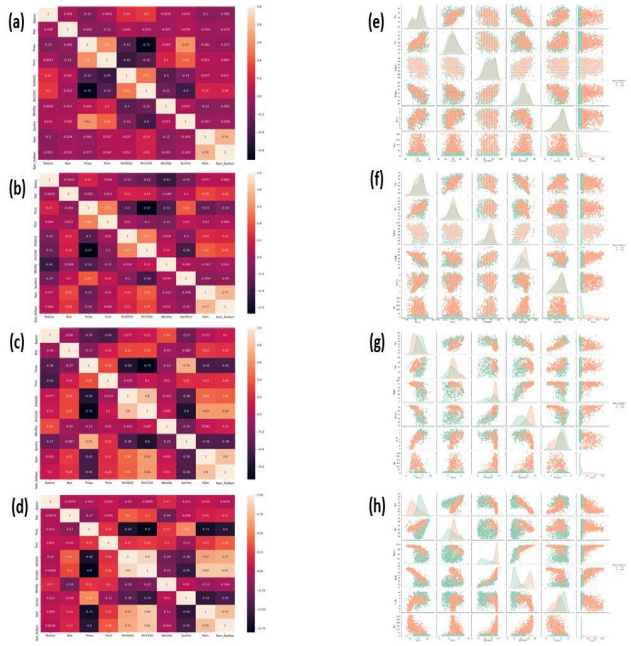


FIGURE 3. Correlation heatmap and pairplots of variables across the 4 ecological zones of Ghana. Panels a – d are the correlation matrix depicting the correlation among the variables. With a, b, c and d representing Coastal, Forest, Transition and Savannah zones respectively. Panels e – h are the corresponding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

C. 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:

1) 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 $x = (x_1, x_2, \ldots x_n)$ with a vector of weights $w_j = (w_{1j}, w_{2j}, \ldots w_{nj})$ for a given node j. To determine the simulated y_j at the node j is given by the equation below [24].

$$y_j = f(x.w_j - b_j) \tag{3}$$

where f is the activation function, w_j as the weight vector and the bias related to the node represented as b_j.

Therefore the output y_k is computed by the equation below:

$$Y_k = f_2\left[\sum_{i=1}^{m} w_{ki} f_1\left(\sum_{j=1}^{n} w_{ij}x_j + b_i\right) + b_k\right] \tag{4}$$

where f_1 and f_2 represents the activation functions, also, j, i and k refers to the input, hidden and output layers respectively. x_j indicates the inputs, the bias linked to the hidden and output layers are assigned with b_i and b_k 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 w_{ij} whereas the weights between the hidden and output layers denoted by w_{ki}. Y_k 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.

2) 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 x_{ij} and x_{io} refers to the ith data point in the jth predictor and predictand.

$$d_j = \sqrt{\sum_{i=1}^{n} (x_{ij} - x_{io})^2} \tag{5}$$

To calculate the KNN value, equation. 5 is used

$$Z_r = \sum_{k=1}^{K} f_k(d_j \times Z_k) \tag{6}$$

Z_r and Z_k represents the predicted and neighboring data respectively whereas $f_k\left(d_j\right)$ 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.

3) 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.

4) 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.

5) 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.

D. 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:

1) CONFUSION MATRIX

The confusion matrix as shown in Table 3, yields an output in a matrix form which details the model's performance. Where:

1. TN represent the overall number of negatively classified data that has been classified as correct.
2. FN is the overall number of positively classified data that has been identified as 'negative' falsely.
3. FP represent the total number negatively classified data that is classified as 'positive' falsely.
4. TP is the total number of positively classified data that is classified as correct.

By extension, it suffice to deduce the mathematical expression for the evaluation metrics such as recall, precision, accuracy and F1 score in the equations 7, 8, 9 and 10 respectively.

$$Recall = \frac{true\ positive}{true\ positive + false\ negative} \tag{7}$$

$$Precision = \frac{true\ positive}{true\ positive + false\ positive} \tag{8}$$

TABLE 3. Confusion matrix.

		CLASS	
		Negative (N)	Positive (P)
Prediction	Negative (N)	True Negative (TN)	False Negative (FN)
	Positive (P)	False Positive (FP)	True Positive (TP)

$$Accuracy = \frac{number\ of\ correct\ predictions}{total\ number\ of\ input\ samples} \tag{9}$$

$$F1\ Score = \frac{2 \times recall \times precision}{recall + precision} \tag{10}$$

IV. RESULTS AND DISCUSSION

The five (5) models namely: Decision Tree (DT), Multilayer Perceptron (MLP), Random Forest (RF), Extreme Gradient Boosting (XGB) and K-Nearest Neighbour (KNN) were developed using Python 3. Using Jupyter Notebook, libraries such as Pandas, Numpy, Sckit Learn, Seaborn and Matplotlib were employed. For the training and testing purposes, we used three (3) different ratios. The datasets were splitted into the ratio of 70:30, 80:20 and 90:10 for training and testing (training: testing) purposes. The performance of the models on the 3 different ratios covering all the ecological zones are presented comprehensively in Tables 4 - 7 and pictorially in Figure 4 – 6 respectively.

A. ANALYSIS OF CLASSIFICATION ALGORITHMS AT THE COASTAL ZONE OF GHANA

The results for the coastal zone are shown in Table 4. Firstly, results for no-rain class with DT, RF and XGB, in precision, recall and *f1* score were 1, 0.993 and 0.996 respectively at the ratio 70:30. However, with the MLP, performed better in precision at the ratio 90:10. Consistently, the KNN classifier showed low results on all three ratios.

Secondly, with rain class, the best performance in precision, recall and *f1* score were recorded at ratio 80:20 with DT, RF and XGB. However, MLP performed better in recall at 90:10. Again, KNN performed worst on all ratios with the rain class whereas XBG performed best on both classes in precision, recall and *f1* score at 80:20.

B. ANALYSIS OF CLASSIFICATION ALGORITHMS AT THE FOREST ZONE OF GHANA

The results for forest zone are presented in Table 5. It can be seen with no-rain class, RF and XGB performed well with same results in precision, recall and *f1* score at 70:30 and 80:20. However, with DT, with no-rain class the 70:30 and 90:10 performed better in recall and 70:30 in *f1* score. Further, with MLP, no-rain class at 90:10 performed better in recall

TABLE 4. Coastal zone results.

Model	Proportion	Class	Precision	Recall	F1 Score
		No Rain	1.00000	0.99313	0.99655
	70:30	Rain	0.99245	1.00000	0.99621
Decision Tree		No Rain	0.98507	1.00000	0.99248
	80:20	Rain	1.00000	0.98246	0.99115
		No Rain	1.00000	0.92453	0.96078
	90:10	Rain	0.90805	1.00000	0.95181
		No Rain	0.99656	0.99656	0.99656
	70:30	Rain	0.99620	0.99620	0.99620
MLP		No Rain	0.99495	0.99495	0.99495
	80:20	Rain	0.99415	0.99415	0.99415
		No Rain	1.00000	0.94340	0.97087
	90:10	Rain	0.92941	1.00000	0.96341
		No Rain	1.00000	0.99313	0.99655
	70:30	Rain	0.99245	1.00000	0.99621
Random Forest		No Rain	0.99497	1.00000	0.99748
	80:20	Rain	1.00000	0.99415	0.99707
		No Rain	1.00000	0.92453	0.96078
	90:10	Rain	0.90805	1.00000	0.95181
		No Rain	1.00000	0.99313	0.99655
	70:30	Rain	0.99245	1.00000	0.99621
XGBoost		No Rain	1.00000	1.00000	1.00000
	80:20	Rain	1.00000	1.00000	1.00000
		No Rain	1.00000	0.92453	0.92453
	90:10	Rain	0.90805	1.00000	0.95181
		No Rain	0.90584	0.95876	0.93155
	70:30	Rain	0.95122	0.88973	0.88973
KNN		No Rain	0.90732	0.93939	0.92308
	80:20	Rain	0.92683	0.88889	0.90746
		No Rain	0.91509	0.91509	0.91509
	90:10	Rain	0.88608	0.88608	0.88608

and 70:30 in *f1* score. KNN performed well with no-rain class at 90:10 in recall. With rain class, with DT, RF and XGB, 70:30 performed better in precision, recall and *f1* score. The notable point is that, all 3 classifiers showed same results. The best performance with MLP in precision is at 90:10. KNN performed worst.

TABLE 5. Forest zone results.

Model	Proportion	Class	Precision	Recall	F1 Score
		No Rain	0.98901	1.00000	0.99448
	70:30	Rain	1.00000	0.99099	0.99548
Decision Tree		No Rain	0.99174	0.95238	0.97166
	80:20	Rain	0.95918	0.99296	0.97578
		No Rain	0.93220	1.00000	0.96491
	90:10	Rain	1.00000	0.94937	0.97403
		No Rain	0.97253	0.98333	0.97790
	70:30	Rain	0.98636	0.97748	0.98190
MLP		No Rain	0.99167	0.94444	0.96748
	80:20	Rain	0.95270	0.99296	0.97241
		No Rain	0.94828	1.00000	0.97345
	90:10	Rain	1.00000	0.96203	0.98065
		No Rain	0.98901	1.00000	0.99448
	70:30	Rain	1.00000	0.99099	0.99548
Random Forest		No Rain	1.00000	0.96032	0.97976
	80:20	Rain	0.96599	1.00000	0.98270
		No Rain	0.93220	0.93220	0.96491
	90:10	Rain	1.00000	0.94937	0.97403
		No Rain	0.98901	1.00000	0.99448
	70:30	Rain	1.00000	0.99099	0.99548
XGBoost		No Rain	1.00000	0.96032	0.97976
	80:20	Rain	0.96599	1.00000	0.98270
		No Rain	0.93220	1.00000	0.96491
	90:10	Rain	1.00000	0.94937	0.97403
		No Rain	0.80556	0.96667	0.87879
	70:30	Rain	0.96774	0.81081	0.88235
KNN		No Rain	0.81208	0.96032	0.88000
	80:20	Rain	0.95798	0.80282	0.87356
		No Rain	0.76056	0.98182	0.85714
	90:10	Rain	0.98413	0.78481	0.87324

C. ANALYSIS OF CLASSIFICATION ALGORITHMS AT THE TRANSITIONAL ZONE OF GHANA

The results for the transitional zone is shown in Table 6. As shown, both rain and no-rain classes with RF and XGB, all 3 ratios performed best in precision, recall and f1 score. It can be noted both RF and XGB classified all instances correctly. Similar performance was seen with MLP with precision, recall and f1 score achieving a result of 1. However, this performance by MLP was only on the ratio 90:10. Comparatively, on all ratios in terms of performance in precision, recall and f1 score, DT outperformed KNN for both rain and no-rain classes.

TABLE 6. Transitional zone result.

Model	Proportion	Class	Precision	Recall	F1 Score
		No Rain	0.93939	0.97895	0.95876
	70:30	Rain	0.98182	0.94737	0.96429
Decision Tree		No Rain	0.98438	0.98438	0.98438
	80:20	Rain	0.98667	0.98667	0.98667
		No Rain	0.96875	0.93939	0.95385
	90:10	Rain	0.94737	0.97297	0.96000
		No Rain	0.98958	1.00000	0.99476
	70:30	Rain	1.00000	0.99123	0.99559
MLP		No Rain	1.00000	0.98438	0.99213
	80:20	Rain	0.98684	1.00000	0.99338
		No Rain	1.00000	1.00000	1.00000
	90:10	Rain	1.00000	1.00000	1.00000
		No Rain	1.00000	1.00000	1.00000
	70:30	Rain	1.00000	1.00000	1.00000
Random Forest		No Rain	1.00000	1.00000	1.00000
	80:20	Rain	1.00000	1.00000	1.00000
		No Rain	1.00000	1.00000	1.00000
	90:10	Rain	1.00000	1.00000	1.00000
		No Rain	1.00000	1.00000	1.00000
	70:30	Rain	1.00000	1.00000	1.00000
XGBoost		No Rain	1.00000	1.00000	1.00000
	80:20	Rain	1.00000	1.00000	1.00000
		No Rain	1.00000	1.00000	1.00000
	90:10	Rain	1.00000	1.00000	1.00000
		No Rain	0.89423	0.97895	0.93467
	70:30	Rain	0.98095	0.90351	0.94064
KNN		No Rain	0.88571	0.96875	0.92537
	80:20	Rain	0.97101	0.89333	0.93056
		No Rain	0.86111	0.93939	0.89855
	90:10	Rain	0.94118	0.86486	0.90141

D. ANALYSIS OF CLASSIFICATION ALGORITHMS AT THE SAVANNAH ZONE OF GHANA

The results for the savannah zone is shown in Table 7. Consistently, RF and XGB with no-rain class, all 3 ratios; 70:30, 80:20 and 90:10 performed best in precision, recall and f1 score. At 90:10, DT performed better in precision with regards to no-rain class. KNN no-ran class at both 70:30 and 80:20 performed better in recall. With rain class, RF and XGB performed the best in recall for all 3 ratios. On the other hand, MLP exhibited similar performance in recall but only with

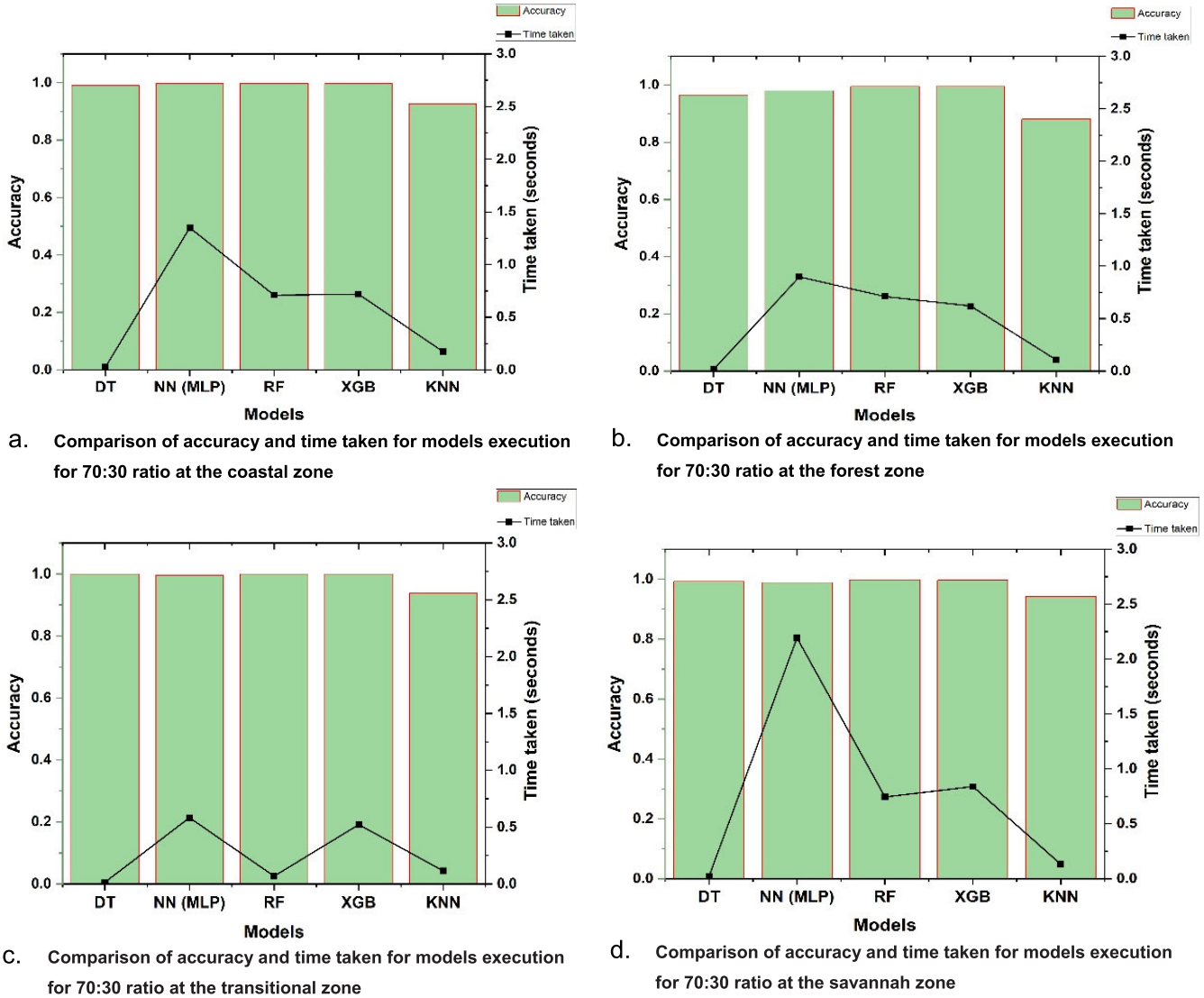
TABLE 7. Savannah zone results.

Model	Proportion	Class	Precision	Recall	F1 Score
		No Rain	0.96629	0.95203	0.95911
	70:30	Rain	0.94800	0.94800	0.95565
Decision Tree		No Rain	0.97110	0.93855	0.95455
	80:20	Rain	0.93605	0.96988	0.95266
		No Rain	0.98864	0.91579	0.95082
	90:10	Rain	0.90588	0.98718	0.94479
		No Rain	0.99237	0.95941	0.97561
	70:30	Rain	0.95686	0.99187	0.97405
MLP		No Rain	0.99438	0.98883	0.99160
	80:20	Rain	0.98802	0.99398	0.99099
		No Rain	1.00000	0.95789	0.97849
	90:10	Rain	0.95122	1.00000	0.97500
		No Rain	1.00000	0.99631	0.99815
	70:30	Rain	0.99595	1.00000	0.99797
Random Forest		No Rain	1.00000	0.99441	0.99720
	80:20	Rain	0.99401	1.00000	0.99700
		No Rain	1.00000	0.92632	0.96175
	90:10	Rain	0.91765	1.00000	0.95706
		No Rain	1.00000	0.99262	0.99630
	70:30	Rain	0.99194	1.00000	0.99595
XGBoost		No Rain	1.00000	0.98883	0.99438
	80:20	Rain	0.98810	1.00000	0.99401
		No Rain	1.00000	0.92632	0.96175
	90:10	Rain	0.91765	1.00000	0.95706
		No Rain	0.93190	0.95941	0.94545
	70:30	Rain	0.95378	0.92276	0.93802
KNN		No Rain	0.94054	0.97207	0.95604
	80:20	Rain	0.96875	0.93373	0.95092
		No Rain	0.94737	0.94737	0.94737
	90:10	Rain	0.93590	0.93590	0.93590

90:10 ratio. DT at 70:30 performed better with rain class in f1 score. On the savannah zone, KNN outperformed DT with 80:20 ratio in precision.

1) CRITICAL ANALYSIS AND SUMMARY OF THE RESULTS OF THE CLASSIFICATION ALGORITHMS

Machine learning classifiers used in this study showed good results on all three different training and testing ratios for the no-rain class in the coastal zone. F1- score, also known as F-measure is an excellent evaluation measure which is based on the average of recall and precision. As clearly shown in Table 4 with regards to f1-score, all the classifiers performed well in classifying the no-rain class as compared to the rain class. Several reasons could be attributed to the lower results with the rain class on the all ratios. These may include the absence of other climatic features and the low



a. Comparison of accuracy and time taken for models execution for 70:30 ratio at the coastal zone

b. Comparison of accuracy and time taken for models execution for 70:30 ratio at the forest zone

c. Comparison of accuracy and time taken for models execution for 70:30 ratio at the transitional zone

d. Comparison of accuracy and time taken for models execution for 70:30 ratio at the savannah zone

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.

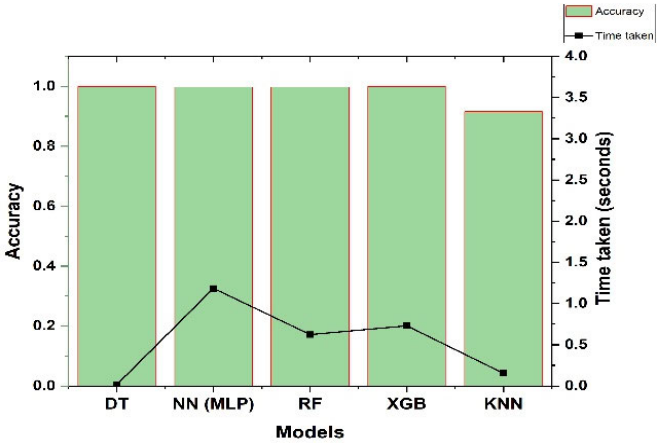
rainfall rate over the zone which is consistent with findings in [18].

In contrast, results in Table 5 shows all the classifiers classified the rain class better than the no-rain class on the basis of the f1-score. The zone which is characterized with high rainfall amounts throughout the year, performed well on the 70:30 training and testing ratio respectively for all five (5) classifiers. However, among the five classifiers, Random Forest and XGBoost outperformed the other classifiers with regards to the rain class classification. Furthermore, f1-score at the savannah zone, is evident that all the classifiers showed good results for the rain class on all the three (3) training and testing ratios.

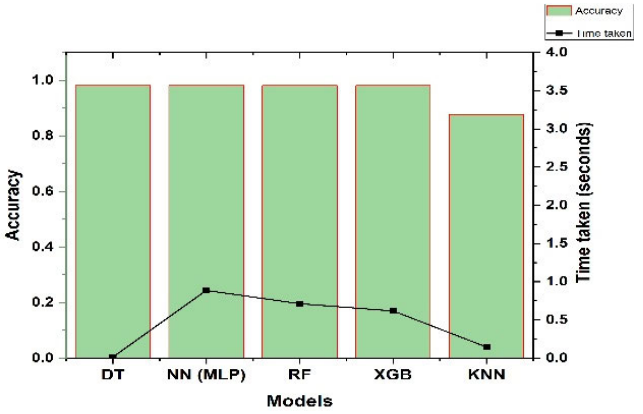
As compared to classifiers at the forest zone whose best performance relied on Random Forest and XGBoost at 70:30, the best performing classifiers at the transitional zone include the MLP, RF and XGB (See table 6). The performance of the classifiers were independent on the ratio of training and testing as all three classifiers were consistent in performance on the three different ratios. From table 7, it is observed that the performance of classifiers for the no-rain class is dominant on all three training and testing ratios which is indicative of low rainfall at the savannah zone which consistent with the findings in [18]. Generally, f1-score for all classifiers performed better for the no-rain class as compared to the rain class.

2) COMPARISON OF TIME OF EXECUTION AND ACCURACY OF MODELS OVER THE VARIOUS ECOLOGICAL ZONES OF GHANA ON THREE (3) TRAINING AND TESTING RATIOS; 70:30, 80:20 AND 90:10

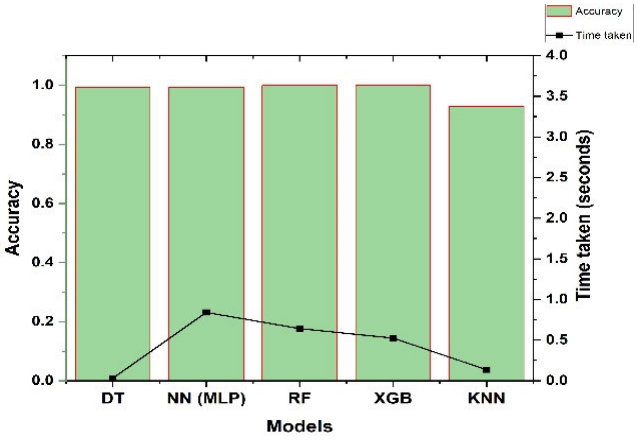
Timeliness is crucial in rainfall prediction. As the adverse effect of rainfall such as floods can be prevented if timely rainfall prediction is achieved. Regardless of a models



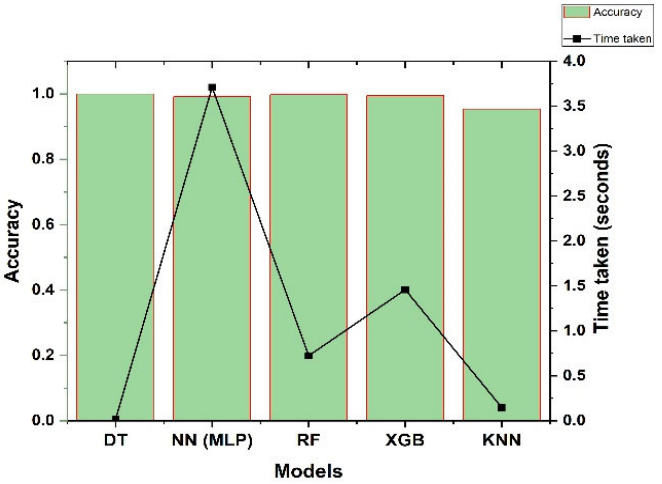
a. Comparison of accuracy and time taken for models execution for 80:20 ratio at the coastal zone



b. Comparison of accuracy and time taken for models execution for 80:20 ratio at the forest zone



c. Comparison of accuracy and time taken for models execution for 80:20 ratio at the transitional zone



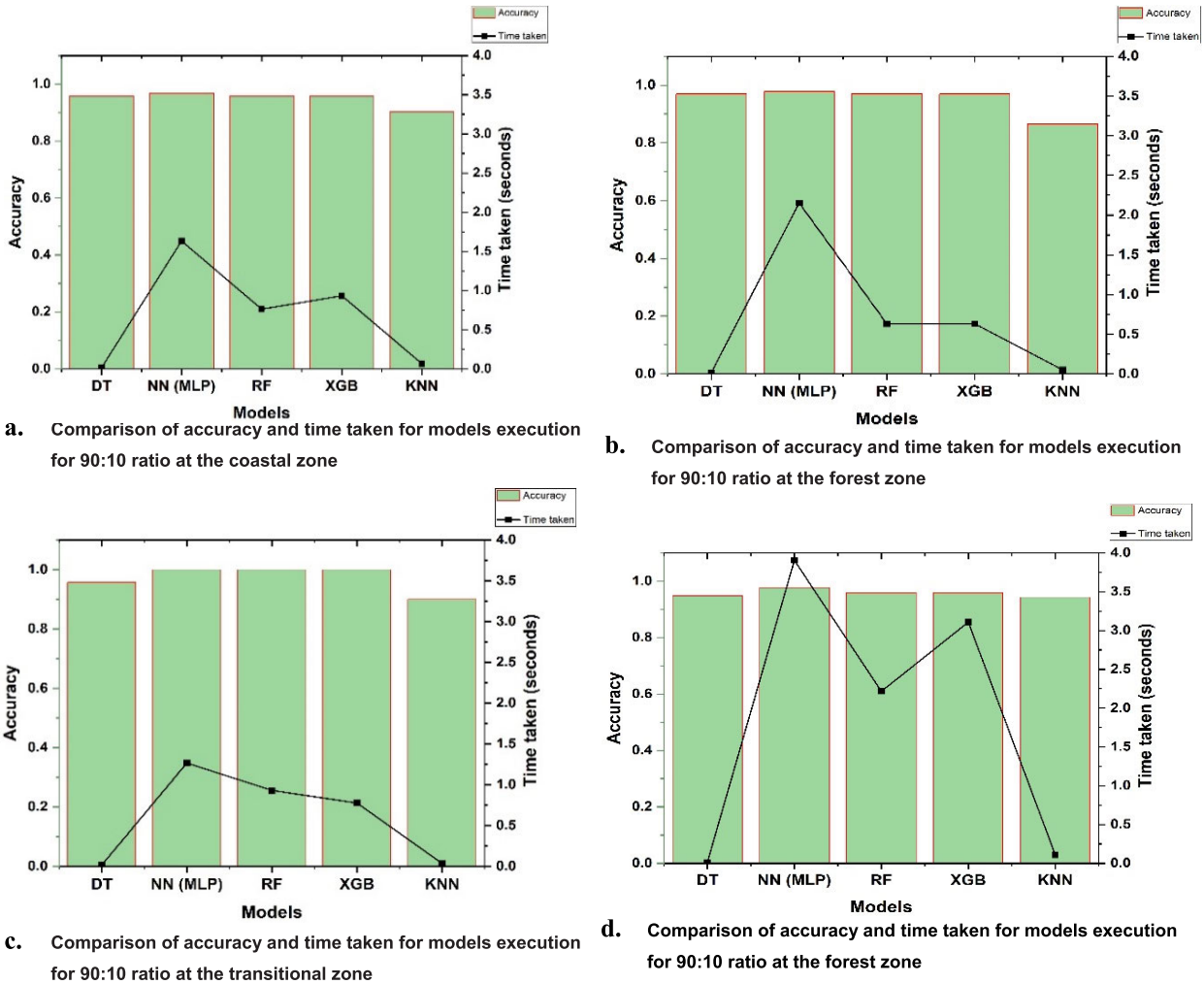
d. Comparison of accuracy and time taken for models execution for 80:20 ratio at the savannah zone

FIGURE 5. Comparison of accuracy and time taken for models execution for 80:20 ratio. Panels a, b, c and d represents Coastal, Forest, Transitional and Savannah zones respectively.

performance in accuracy, its swiftness to rainfall prediction is essential. A visual inspection of Fig. 4a shows that almost all the classifiers exhibited similar overall model accuracy with the exception of KNN showing a low model accuracy on the 70:30 training and testing ratio at the coastal zone. However, with regards to time taken for the execution of these models, decision tree leads as the fastest whereas MLP stands out as the model with the longest execution time. Similarly, the decision tree classifier also performed as the fastest classifier to be executed with the shortest possible time at the forest zone (see Figure 4b.).

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



a. Comparison of accuracy and time taken for models execution for 90:10 ratio at the coastal zone

b. Comparison of accuracy and time taken for models execution for 90:10 ratio at the forest zone

c. Comparison of accuracy and time taken for models execution for 90:10 ratio at the transitional zone

d. Comparison of accuracy and time taken for models execution for 90:10 ratio at the forest zone

FIGURE 6. Comparison of accuracy and time taken for models execution for 90:10 ratio. Panels a, b, c and d represents Coastal, Forest, Transitional and Savannah zones respectively.

ecological zones of Ghana. However, in terms of accuracy the MLP, XGB and RF have been pronounced.

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 K-Nearest 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, pp. 529–546, Nov. 2010.

[2] N. K. Karley, "Flooding and physical planning in urban areas in West Africa: Situational analysis of Accra, Ghana," *Theor. Empirical Res. Urban Manage.*, vol. 4, no. 4, 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*. Amsterdam, The Netherlands: 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, Jun. 2019.

[5] I. Yabi and F. Afouda, "Extreme rainfall years in Benin (West Africa)," *Quaternary Int.*, vol. 262, pp. 39–43, Jun. 2012.

[6] C. Kyei-Mensah, R. Kyerematen, and S. Adu-Acheampong, "Impact of rainfall variability on crop production within the Worobong ecological area of Fanteakwa district, Ghana," *Adv. Agricult.*, vol. 2019, May 2019, Art. no. 7930127.

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

[8] K. Owusu and N. Klutse, "Simulation of the rainfall regime over Ghana from CORDEX," *Int. J. Geosci.*, vol. 4, no. 4, pp. 785–791, 2013, doi: 10.4236/ijg.2013.44072.

[9] S. Mofa, "Agriculture in Ghana: Facts and figures," *Minist. Food Agric. Accra.*, vol. 10, no. 1, pp. 1–58, May 2010.

[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, Mar. 2018.

[11] N. Oswal, "Predicting rainfall using machine learning techniques," 2019, *arXiv:1910.13827*.

[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. Amekudzi, E. Yamba, K. Preko, E. Asare, J. Aryee, M. Baidu, and S. Codjoe, "Variabilities in rainfall onset, cessation and length of rainy season for the various agro-ecological zones of Ghana," *Climate*, vol. 3, no. 2, pp. 416–434, Jun. 2015.

[15] M. New, M. Todd, M. Hulme, and P. Jones, "Precipitation measurements and trends in the twentieth century," *Int. J. Climatol.*, vol. 21, no. 15, pp. 1889–1922, Dec. 2001.

[16] K. Owusu and P. R. Waylen, "The changing rainy season climatology of mid-Ghana," *Theor. Appl. Climatol.*, vol. 112, nos. 3–4, pp. 419–430, May 2013.

[17] C. Mensah, L. Amekudzi, N. Klutse, J. Aryee, and K. Asare, "Comparison of rainy season onset, cessation and duration for Ghana from RegCM4 and GMet datasets," *Atmos. Climate Sci.*, vol. 6, no. 1, pp. 300–309, 2016, doi: 10.4236/acs.2016.62025.

[18] M. Baidu, L. K. Amekudzi, J. Aryee, and T. Annor, "Assessment of long-term spatio-temporal rainfall variability over Ghana using wavelet analysis," *Climate*, vol. 5, no. 2, p. 30, Mar. 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 Psychiatric Res.*, vol. 20, no. 1, pp. 40–49, Mar. 2011.

[21] R. Kaur and S. Sharma, "An ANN based approach for software fault prediction using object oriented metrics," in *Proc. Int. Conf. Adv. Inform. Comput. Res.*, 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, May 2020, Art. no. 104806.

[23] T. Shimobaba, N. Kuwata, M. Homma, T. Takahashi, Y. Nagahama, M. Sano, S. Hasegawa, R. Hirayama, T. Kakue, A. Shiraki, N. Takada, and T. Ito, "Deep-learning-based data page classification for holographic memory," 2017, *arXiv:1707.00684*.

[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, Nov. 2003.

[25] M. Bowles, *Machine Learning with Spark and Python: Essential Techniques for Predictive Analytics*. Hoboken, NJ, USA: Wiley, 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 Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1774–1785, May 2018.

[27] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," in *Ensemble Machine Learning*. New York, NY, USA: Springer, 2012, pp. 157–175.

[28] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

NANA KOFI AHOI APPIAH-BADU received the B.Sc. degree from the University of Ghana, in 2011, and the master's degree from the Kwame Nkrumah University of Science and Technology, in 2016, where he is currently pursuing the Ph.D. degree with the Department of Computer Science. His research interests include machine learning and deep learning.

YAW MARFO MISSAH received the B.Sc. degree in computer science from the University Science and Technology, Kumasi, Ghana, in 2000, the M.Sc. degree in information technology from Clark University, Worcester, MA, USA, in 2004, and the Ph.D. degree in computer science from Colorado Technical University, Colorado Springs, CO, USA, in 2013. His research interests include machine learning, and deep learning, machine learning concepts and its applications in the environment, and artificial intelligence and its applications in enterprise systems.

LEONARD K. AMEKUDZI received the B.Sc. degree in physics with Dip.Ed., and the M.Phil. degree in theoretical physics from the University of Cape Coast, Ghana, in 1997 and 2001, respectively, and the Ph.D. degree in atmospheric physics and satellite remote sensing from the University of Bremen, Germany, in 2005. He is currently a Professor of atmospheric and climate science and provost with the College of Science, Kwame Nkrumah University of Science and Technology (KNUST), Kumasi. After completing one year of M.Sc. course work at the Institute of Environmental Physics and Remote Sensing, University of Bremen, in 2002, the European Space Agency (ESA) ENVISAT (Environmental Satellite) awarded him a Ph.D. Fellowship. He continued as a Postdoctoral Research Scientist on the ESA ENVISAT project. He worked extensively with different research clusters across the globe. Among the research-collaborated projects include ENVISAT SCIAMACHY Limb and Occultation Validation (SCILOV), Quantifying Weather and Climate Impacts on Health in Developing Countries (QWeCI), Building Stronger Universities (BSU), West Africa Science Service Center for Climate

Change and Adapted Land Use (WASCAL), Dynamic-Aerosol-Chemistry-Cloud interactions in West Africa (DACCIWA) and International Development Research Center - Climate Change Adaptation Research and Training Capacity for Development (IDRC-CCARTCD), Global Challenge Research Fund Africa Science for Weather Information and Forecasting Techniques (GCRF African SWIFT), Current and Future risks of Urban and Rural Flooding in West Africa-An integrated analysis and eco-system-based solutions (FURILOOD), and Green gas emissions and mitigation options under climate and land-use change in West Africa-A concerted regional modeling and assessment (CONCERT). He has over 70 publications in high-impact peer-reviewed journals and over 85 oral and poster presentations in international conferences.

NAJIM USSIPH received the B.Sc. and M.Sc. degrees in computer science from the University of Ibadan, Ibadan, Nigeria, in 1987 and 1993, respectively, and the Ph.D. degree from the University of Salford, Manchester, in 2015. He has over 15 years of experience in teaching and research at Polytechnic and University level. He has been a Faculty Member with the Department of Computer Science, Kwame Nkrumah University of Science and Technology, Kumasi, since April 2001, and have taught several courses at both undergraduate and graduate levels. His research interests include information systems and e-learning and learning technologies.

TWUM FRIMPONG received the M.Sc. degree in information systems from Roehampton University, and the Ph.D. degree in computer science from the Kwame Nkrumah University of Science and Technology (KNUST). He is currently a Senior Lecturer with the Department of Computer Science, KNUST. His research interests include computer networks security and machine learning.

EMMANUEL AHENE received the M.Eng. and Ph.D. degrees in computer science and technology from the University of Electronic Science and Technology of China. He is currently a Lecturer with the Department of Computer Science, Kwame Nkrumah University of Science and Technology, Ghana, and also a Co-Founder of Cyberpassconsult, an international cybersecurity consultancy firm. His research interests include information security and machine learning.