RESEARCH ARTICLE OPEN ACCESS

An Offline Rainfall Prediction Using MLR and Ensemble Learning

Hari Prasad Chandika[1], Eswar Maddi[2], Sony Kattepogu[3], Venkata Siva Reddy Kandula[4], Purna Kommalapati[5]

1(Assistant Professor, Department of CSE, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, Email: hari.chandika@gmail.com)

2(UG Student, Department of CSE, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, Email: 2002eswar@gmail.com)

3(UG Student, Department of CSE, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, Email: sonykattepogu3@gmail.com)

4(UG Student, Department of CSE, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, Email: kandulasiva369@gmail.com)

5(UG Student, Department of CSE, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, Email: kommalapatipurna@gmail.com)

Abstract:

Accurate rainfall prediction is vital for various sectors in Ghana, including agriculture, hydrology, and disaster management. In this study, various regressor models like Multiple Linear Regression (MLR), Gradient Boosting (GB), Extreme Gradient Boost (XGB), and a Voting Regressor ensemble approach are included. The dataset, consisting of various climatic attributes, was sourced from NASA Power spanning 1982 – 2023. The Mean Squared Error (MSE) and explained variance were used as evaluation metrics. The default splitting ratio for training and testing data included is 75:25, which is applied to all the machine learning models. Our results indicate that the advanced ensemble methods, Gradient Boosting, Extreme Gradient Boosting, and Voting Regressor, outperform the traditional Multiple Linear Regression model in terms of both MSE and explained variance. These models exhibit superior predictive capabilities, capturing nonlinear relationships and interactions among predictor variables more effectively. Furthermore, the ensemble approach, by combining the strengths of individual models, enhances the overall predictive performance, resulting in a lower error rate for the target location. Our research highlights the importance of utilizing advanced regression techniques, as evidenced by a mean squared error (MSE) score of 11.56.

Keywords: **Multiple Linear Regression, Gradient Boosting, Extreme Gradient Boosting, Voting Regressor, MSE, Machine Learning**

I. INTRODUCTION

Rainfall prediction plays a crucial role in various sectors of Ghana's economy, including [7]agriculture, water resource management, and disaster preparedness. Accurate forecasts enable stakeholders to make informed decisions, mitigate risks, and optimize resource allocation. In recent years, advancements in machine learning algorithms have revolutionized the field of meteorology, offering enhanced predictive capabilities compared to traditional statistical methods. This study aims to investigate the effectiveness of multiple regression models, including [8]Multiple Linear Regression (MLR), Gradient Boosting, XGBoost, and a Voting Regressor ensemble approach, in predicting rainfall for a specific location in Ghana. [7],[8]Multiple Linear Regression (MLR) is a fundamental a statistical technique widely used for modeling linear relationships between predictor variables and a target variable.

However, its limitations in capturing nonlinear dependencies and complex interactions among predictors necessitate exploring more sophisticated regression algorithms. Gradient Boosting and [6]XGBoost are ensemble learning techniques that combine the predictive power of multiple weak learners to construct a robust predictive model. These algorithms iteratively fit new models to the residuals of previous models, gradually improving prediction. Additionally, the Voting Regressor aggregates predictions from multiple base estimators to produce a final prediction, leveraging the diversity of individual models to enhance overall performance.

In this research, historical meteorological data from reliable sources in Ghana will be utilized for model training and evaluation. The dataset will include parameters such as temperature, maximum temperature, minimum temperature, specific humidity, relative humidity, wind speed, surface pressure, wind direction, and dew point. The predictive performance of each regression model will be assessed using the Mean Squared Error (MSE) and explained variance metrics. MSE quantifies the average squared difference between predicted and observed rainfall values, providing a measure of prediction accuracy. Explained variance measures the proportion of variance in the target variable that is explained by the regression model, indicating its overall goodness of fit.

By comparing the performance of different regression models, this study aims to identify the most effective approach for rainfall prediction in the specific location under consideration. The findings will contribute to advancing the understanding of meteorological modeling techniques and provide valuable insights for stakeholders involved in weather forecasting and decision-making processes in Ghana. Ultimately, accurate rainfall predictions derived from advanced regression models can support sustainable development initiatives and improve resilience to climate-related challenges in the region.

II. RELATED WORK

Several studies have explored the application of various regression models for rainfall prediction in different geographical regions, contributing to the advancement of meteorological forecasting techniques.

Predicting rainfall accurately is the most difficult task. Prediction includes things like observation of previous models and knowledge of current trends. With all these, we can make accurate predictions on rainfall patterns. Already there are previous existing models used to predict the weather and climatic conditions.

Reference [1] study shows distinct characteristics of classification of the rain and no-rain classes in the ecological zones of Ghana. No-rain class was well classified by classifiers in the coastal zone as compared to the rain class.

Reference [2] shows that the system will output one of the 5 rainfall categories depending on the results of the regression method. The system tested in 20086 recorded weather information of eleven cities for the period from 2018 to 2022. Using SoftMax logistic regression they found 83% is correctly predicted.

Most of the rainfall datasets available have irregular patterns so the classifier implemented in this is a Random Forest classifier because it can handle missing values and a large dataset with less training time compared to other algorithms. It has given the best accurate result of 85%[3].

In this system, In ANN if the number of neurons increases, the MSE decreases. Here, BPA is the best algorithm as compared to other models. The best learning function to train the data is LEARNGDM. LEARNGD is a bit time-consuming. The best training function is TRAINLM[4].

Reference [5] Models used are based on LSTM and RNN has been developed to predict the amount of rainfall on a monthly basis and on an annual basis.

The research addresses the relationship between independent and dependent variables to capture which variables affect the rainfall to rain or not to rain[7],[8].

Overall, these related works underscore the importance of leveraging advanced regression models and ensemble learning techniques for rainfall prediction in Ghana. By employing MSE and explained variance metrics for model evaluation, researchers can assess predictions.

III. METHODOLOGY

This paper reveals the observation on the models of XGBoost, Gradient Boost, Multiple Linear Regression, and Voting regressor(Hybrid model which includes XGBoost, Gradient Boost, and Multiple Linear Regression). A voting regressor is used for predicting rainfall using important atmospheric features by exposing the relationship between atmospheric variables that show some effect on the rainfall patterns.

A. Multiple Linear Regression(MLR)

MLR is a statistical method that makes us to analyze the relationship between dependent variables and independent variables. It models the relationship between various environmental parameters(independent variables) and the amount of rainfall(dependent variable).

$Y = \beta0+\beta1X1+\beta2X2+...+\beta nXn+E$

Here, Y = dependent variable
X1,X2,...,Xn = independent variables
β0=intercept term
β1,β2,...,βn = coefficients
E = error term

1. It uses the training data set to find the coefficients.
2. Test the model performance on testing dataset to ensure it works good for new and unseen data.
3. Once the model is trained, tested and evaluated, we can use this to make predictions on new and unseen data by giving new values for the independent variables.
4. It uses historical data to estimate coefficients that reduce the difference between predicted values and actual values.

Success of the MLR is based on the quality of chosen independent variables.

B. Gradient Boosting

Gradient Boosting is an ensemble learning model that combines predictions of multiple weak learners to make a better model. The model is represented in Fig.1.

1. Collect the historical data that contains meteorological parameters.
2. Clean the data by handling missing values and outliers.
3. Now Initialize the model with a weak learner.
4. Train the weak learner.
5. Calculate the difference between predicted and actual values.
6. In the gradient boost model, decision trees are mostly used as weak learners. The algorithm makes trees sequentially. After training each tree, it adds to the ensemble.
7. The algorithm focuses on the residuals i.e., the differences between actual and predicted values of the combined ensemble.
8. A loss function(MSE) is made to identify the difference between actual and predicted values. It uses gradient descent optimization to reduce the loss function. It can provide the accurate predictions for rainfall based on the historical data.
9. Evaluate the performance of the model on the testing dataset by using evaluation metrics.

10. Once the model is trained and tested, we can use it to make predictions on new and unseen data.

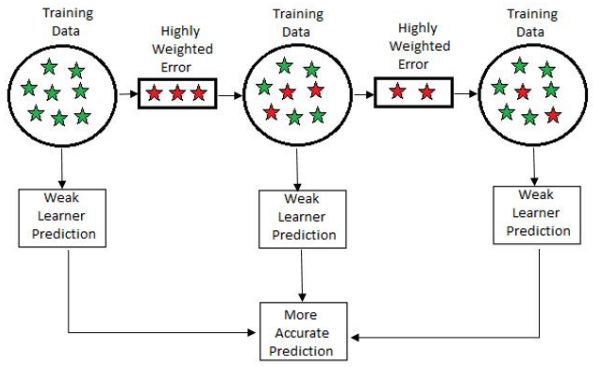

Fig. 1 Gradient Boosting Model

C. *Extreme Gradient Boosting*

Extreme Gradient Boosting is popular in its efficiency, performance, and scalability.

1. Gather historical data that contains metrological parameters.
2. Preprocess data by handling missing values and outliers.
3. Initialize the model with selected features.
4. Train the model on the training dataset using rainfall values.
5. It uses a function that combines a loss function measuring the difference between predicted values and actual values. It also builds trees sequentially with each tree trained.
6. XGBoost indicates which features have the most impact on the models predictions. Trees are added one by one till we get sufficient improvement. Decision tree in XGBoost is made as a sum of the leaf scores and optimization finds the optimal scores for every leaf. It is used for good parallel processing, highly scalable and faster performance than other gradient boosting models.
7. Once the model is trained and evaluated, we can use it to make predictions.

D. *Voting Regressor*

The idea is to make the predictions of individual regression models to get a more robust and accurate prediction. We need to choose a set of regression models as base regressors like XGBoost, Gradient Boost, and MLR. Each model is trained independently on the training dataset using the selected parameters. The predictions from every base regressor model are combined to make a final prediction.

$$\text{Final Prediction} = (1/N)\sum \text{Prediction}_i$$

Here N is the number of base regressors.

1. Gather historical data that contains metrological parameters.
2. Preprocess the data by handling missing values and outliers.
3. Split the data into training data and testing data.
4. Choose a set of models that you want to combine.
5. Initialize each model with its selected parameters.
6. Create a 'Voting Regressor' instance and pass a list of our selected models.
7. Train the voting regressor model and the model combines the predictions of all the selected models to give a final prediction.
8. Once the model is trained and evaluated, we can use it to make predictions on new and unseen data.

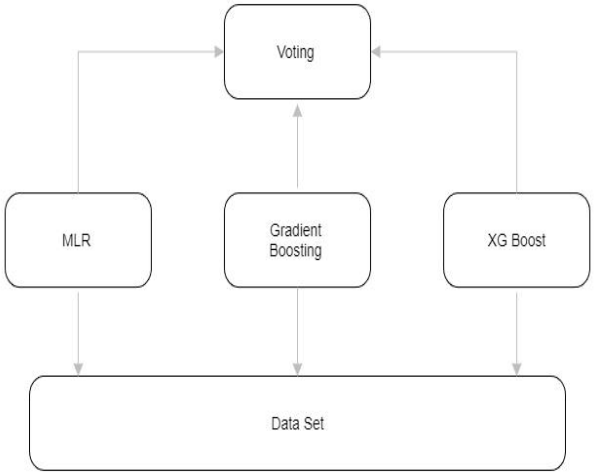

Fig. 2 Voting Regressor Model

IV. PROPOSED MODEL

Ghana is grouped into four (4) agroecological zones according to the Ghana Meteorological Agency classification. They are Coastal, Forest, Transition, and Savannah zones. 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 by warm temperatures all year round.

Fig. 3 Ghana Map

A. Data collection:

For this study, the raw data was collected from NASA Power a particular location in Ghana its capital Accra from the year 1982 to 2023. The dataset consists of 15,340 records. Features like the year, day, maximum temperature, minimum temperature, Dew/frost point, Earth skin temperature, Specific humidity, Relative humidity, Surface pressure, wind speed, and wind direction at 2 meters and 10 meters are included. They are represented in Table I.

The meteorology station recorded the values of the environmental variables for each year directly from the devices in the station. The data were placed in a comma-separated file. The environmental variables are in the columns of the table. The raw data recorded at the station for 41 years (1982–2023) were used for the study.

TABLE I
Attributes in data

Short Hand Notation	Parameter	Definition
T2M	Temperature at 2 meters (C)	Air temperature is measured at a height of 2 meters above the ground level
T2MDEW	Dew/Frost point at 2 Meters(C)	Temperature at which air becomes saturated with moisture leads to condensation
TS	Earth Skin Temperature (C)	Temperature of the ground or Earth's surface
T2M_RANGE	Temperature at 2 Meters Range (C)	Difference between the maximum and minimum temperatures
T2M_MAX	Temperature at 2 Meters Maximum (C)	Highest recorded air temperature
T2M_MIN	Temperature at 2 Meters Minimum (C)	Lowest recorded air temperature
QV2M	Specific Humidity at 2 Meters(g/kg)	Amount of water vapor present in the air per unit mass of dry air
RH2M	Relative Humidity at 2 Meters(%)	Percentage of the maximum amount of water vapor the air can hold at the same temperature and pressure
PS	Surface pressure(kPa)	Atmospheric pressure exerted by the air molecules at ground level
WS2M	Wind Speed at 2 Meters(m/s)	Rate at which air moves horizontally at height of 2 meters
WD2M	Wind Direction at 2 Meters(Degrees)	Direction from which the wind is blowing at 2 meters
WS10M	Wind Speed at 10 Meters(m/s)	Rate at which air moves horizontally at height of 10 meters
WD10M	Wind Direction at 10 Meters(Degrees)	Direction from which the wind is blowing at 10 meters

B. Data preprocessing:

The data preprocessing includes data conversion, managing missing values, and splitting datasets for training and testing datasets. The dataset doesn't contain missing values and null data values in it. The important features for prediction are selected and the dataset splitting as 75% for training data and 25% for testing data were taken as an input for the model which is the default splitting ratio.

To know the correlations between the attributes we used a heatmap that describes the correlation of the attributes from the values 0 to 1. It is represented in Fig.4. The most correlated attributes are T2M and TS, QV2M, T2MDEW. The least correlated attributes are T2M and PS, PS and T2M.

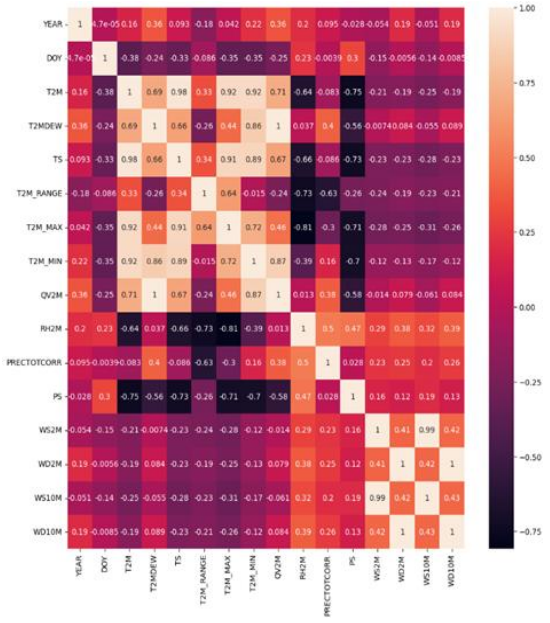

Fig. 4 Correlation of attributes

C. Model Architecture:

Splitting the dataset of rainfall historical data into input (X) and output (Y) variables. The input(X) has the features of the data and output(Y) is the amount of rainfall measured in millimeters per day. The process flow of the rainfall prediction system is described in Fig. 5.

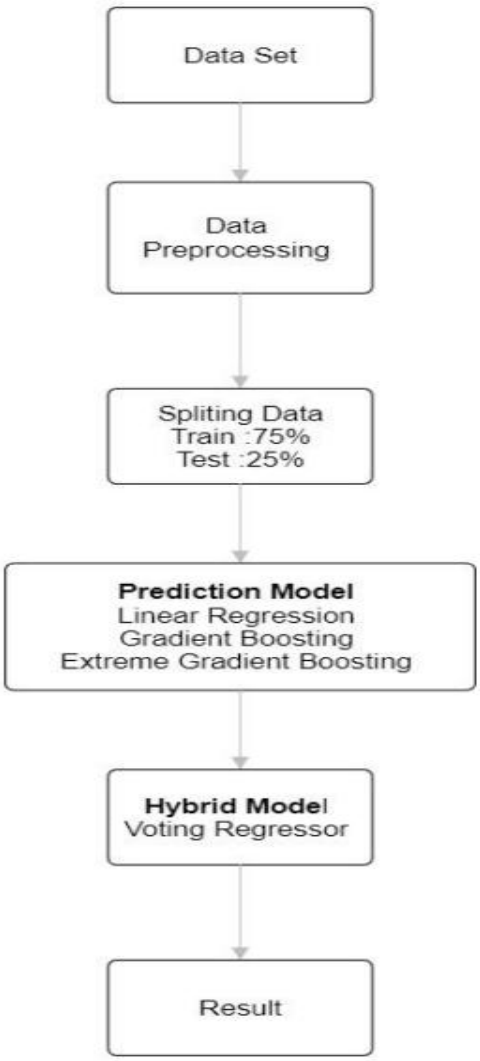

Fig. 5 Process Flow

D. Evaluation of model:

Variance and Mean square error(MSE) are the two metrics that were used to measure the strength and performance of each machine learning model and make us know which learning model can fit better than others.

The mean square error can be described as

$$\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2$$

Here, n is the total number of observations. y_i represents the actual rainfall values. y_i' represents the predicted rainfall values.

The explained variance is calculated as

$$\mathbf{EV} = 1 - \frac{\mathbf{Var}[Y - \bar{Y}]}{\mathbf{Var}[Y]}$$

$$\mathbf{Var}(Y) = \frac{\sum (y_i - \bar{y})^2}{n}$$

Here, y_i represents each predicted rainfall value, $\bar{y}$ represents the mean of the predicted rainfall values and n represents the total number of predictions.

V. RESULT AND DISCUSSION

From this study, experimental results are shown in Fig. 6 and Fig. 7, with an error rate of 13.85 for MLR. Since the error rate is high and negatively impacts the accuracy of the model and is also prone to overfitting when dealing with small datasets, we evaluated other new models. Gradient Boost gave an error rate of 12.29 and the XGBoost model error rate is 11.68. We performed a hybrid model which gave us a better result at the end. Our hybrid model is a combination of XGBoost, Gradient Boost, and MLR. In this model, the results are decent with an error rate of 11.56 which can be considered and performed well when compared with the other Machine learning models. Voting regressor's results showed effectiveness in reducing model variance. So, Voting regressor can be used to predict rainfall patterns with less error rate.

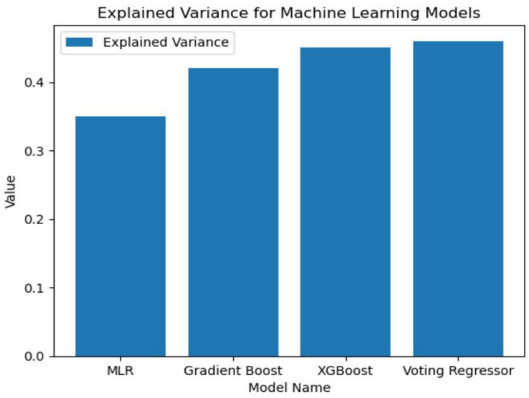

Fig. 6 Explained Variance

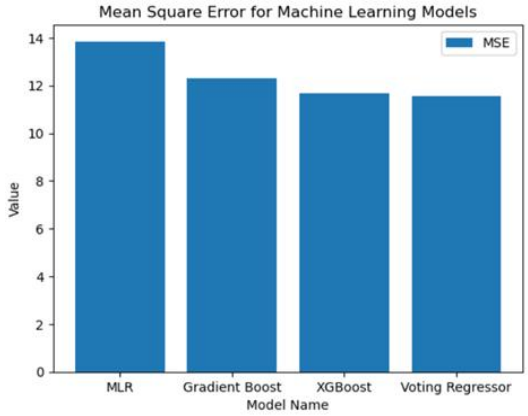

Fig. 7 Mean Square Error

TABLE II
Evaluation Metric Scores

Model	Mean Square Error	Explained Variance Score
Multiple Linear Regression	13.85	0.35
Gradient Boosting	12.29	0.42
Extreme Gradient Boosting	11.68	0.45
Voting Regressor	11.56	0.46

MLR assumes a linear relationship between the predictors and the dependent variable. However, in reality, the relationship between meteorological variables and rainfall may not be strictly linear. Nonlinear relationships may lead to inaccurate predictions if not properly accounted. MLR is sensitive to outliers in the data. Gradient boosting model have several hyperparameters that need tuning for optimal performance. The performance of XGBoost heavily relies on the quality of the input data and the features used for prediction. In rainfall prediction, selecting relevant features and performing effective feature engineering can be challenging due to the complex nature of meteorological data. XGBoost models trained on data from one region or climate may not generalize well to other regions or climates with different meteorological characteristics.

VI. CONCLUSION

In this research rainfall prediction in Ghana for a particular location. 41 years of past climatic data spanning 1982 – 2023 from NASA Power was used for this study. The machine learning model performed well in the Voting Regressor model, achieving an MSE value of 11.56. The MLR model did not perform well. But the other models Gradient Boost and XGBoost performed well. The further study can include transfer learning approaches or domain adaptation techniques that may be necessary to address this issue. It can also be implemented in different locations in Ghana to reduce the MSE value.

REFERENCES

1. *N. K. A. Appiah-Badu, "Rainfall Prediction Using Machine Learning Algorithms for the Various Ecological Zones of Ghana", IEEE Xplore, Vol. 10, Jan 2022, doi: 10.1109/ACCESS.2021.3139312*
2. *Mar Mar Soe, "Rainfall prediction using regression model", ICCA, Feb 2023, doi: 10.1109/ICCA51723.2023.10182116*
3. *Deepika Mahajan, Sandeep Sharma, "Prediction of rainfall using machine learning", ICERECT, Sept 2023, doi: 10.1109/ICERECT56837.2022.10059679*
4. *Kumar Abhishek, Abhay Kumar, "A Rainfall Prediction Model using Artificial Neural Network", ICSGRC, 2012*
5. *Imrus Salehin, Sadia Tamim Dip, Iftakhar Mohammad Talha, Mohd. Saifuzzaman, Md. Mehedi Hasan, Nazmun Nessa Moon, "An Artificial Intelligence Based Rainfall Prediction Using LSTM and Neural Network", WIECON-ECE, 2020, doi: 10.1109/WIECON-ECE52138.2020.9398022*
6. *Chalachew Muluken Liyew & Haileyesus Amsaya Melese, "Machine learning techniques to predict daily rainfall amount", Journal of Big Data, Vol 8, Article number: 153, December 2021*
7. *Thirumalai C, Harsha KS, "Heuristic prediction of rainfall using machine learning techniques", International Conference on Trends in Electronics and Informatics (ICEI), 2017*
8. *Dr. C. k.gomathy, Annapareddy Bala Narasimha Reddy, "A study on rainfall prediction techniques", IJSREM, Vol 05, Oct 2021*
9. *Geetha, A., G. M. Nasira, "Data mining for meteorological applications: Decision trees for modeling rainfall prediction", IEEE International Conference on Computational Intelligence and Computing Research, 2014*
10. *T. Yue, S. Zhang, J. Zhang, B. Zhang, and R. Li, "Variation of representative rainfall time series length for rainwater harvesting modelling in different climatic zones," Journal of Environmental Management, vol. 269, 2020*