Weather Forecasting Using Machine Learning Techniques: Rainfall and Temperature Analysis

Adil Hussain * , Ayesha Aslam , Sajib Tripura , VIneet Dhanawat , Varun Shinde

Posted Date: 13 September 2024

doi: 10.20944/preprints202402.1566.v2

Keywords: rainfall prediction; temperature prediction; ensemble classifier; rain prediction; weather prediction

Article

Weather Forecasting Using Machine Learning Techniques: Rainfall and Temperature Analysis

Adil Hussain [1,*], Ayesha Aslam [2], Sajib Tripura [1], Vineet Dhanawat [3] and Varun Shinde [4]

[1] School of Electronics and Control Engineering, Chang'an University, Xi'an, China
[2] School of Information Engineering, Chang'an University, Xi'an, China
[3] Meta Platforms Inc., California, United States
[4] Cloudera, Inc., Austin, Texas, United States
* Correspondence: 2022032907@chd.edu.cn

Abstract: Heavy rains result in significant threats to human health and life. Floods and other natural disasters, which have a global impact annually, can be attributed to extended periods of intense precipitation. Accurate rainfall prediction is crucial in nations such as Bangladesh, where agriculture is the predominant field of occupation. The efficiency of machine learning methods is enhanced by the nonlinearity of rainfall, surpassing the effectiveness of other approaches. This study proposes the novel combination of rainfall occurrence prediction, rainfall amount prediction, and daily average temperature prediction. This research implements machine learning techniques and an ensemble-based classifier to predict rainfall occurrence, as well as machine learning regressor models and an ensemble-based regressor to predict the rainfall amount and daily average temperature, using the Bangladesh Weather Dataset. The ensemble classifier demonstrated an accuracy of 83.41% and a recall of 78.17%, exhibiting the best performance in predicting when it will rain, but its precision was the lowest, at 51.16%. The ensemble regression model outperformed the base models, including linear regression, random forest, and support vector regression in rainfall amount prediction, with the lowest mean absolute error of 0.36 and root mean squared error of 0.90. Additionally, this model provided the most precise daily average temperature prediction results with the lowest mean absolute error of 0.42 and root mean squared error of 0.54, highlighting its superiority over the other regression models in forecasting temperature. Ensemble approaches consistently exhibit superior task performance metrics.

Keywords: rainfall prediction; temperature prediction; ensemble classifier; rain prediction; weather prediction

1. Introduction

Rainfall has always been an important factor in both historical and contemporary global contexts. Various elements influence rainfall, including humidity, temperature, and water levels [1]. Excessive rainfall has detrimental effects on crops within a given geographical area, ultimately leading to the complete disruption of agriculture in that location. Heavy precipitation can lead to many natural disasters, such as floods, droughts, and cloud bursts, which are often triggered by intense rainfall and rapid landslides [2]. Rainfall prediction involves anticipating long-term precipitation patterns within a specific geographical area. The ability to accurately forecast rainfall can significantly contribute to the success of the agricultural industry and foster economic growth. Ensuring the precision of rainfall measurements is crucial for mitigating the occurrence of landslides, which frequently obstruct river channels [3]. The variability of precipitation is a significant and intricate issue. Existing rainfall forecasting methods have a limited ability to identify precise concealed patterns or nonlinear trends in rainfall data, which are essential for achieving accurate rainfall predictions. Many rainfall forecasts are inaccurate, and such inaccuracies can lead to substantial economic losses. Diverse climatic conditions significantly influence the deterioration of

Manuscript received June 21, 2024; revised August 19, 2024; accepted September 2, 2024.

 doi:10.20944/preprints202402.1566.v2

infrastructure and the occurrence of injuries and fatalities. Hence, it is imperative to obtain precise rainfall forecasts to predict the impact of weather conditions on large-scale activities [4].

Weather forecasting constitutes a subcategory of climate change research that predicts the state of the atmosphere at certain times and locations [5]. Rainfall prediction is key to weather forecasting in large-scale water-dependent operations, including food production planning and water resource management. Rainfall projections must be improved, particularly in terms of accuracy and predictive performance, to properly prepare and plan large-scale activities. Significant advancements have been made in machine learning, which has become a fundamental sub-discipline within the broader domain of artificial intelligence [6–9]. Furthermore, this technology allows computers to autonomously acquire knowledge and understanding, without explicit programming. Machine learning algorithms can be employed to derive significant insights from data, thereby facilitating the efficient intrusion detection [9].

Nevertheless, the current level of machine learning achievement remains significantly inferior to human-level abilities. Human intervention is still required to predefine the algorithms during the initialization process. Various machine learning methodologies have been investigated for rainfall prediction, focusing on diverse geographical regions including South Africa, China, and other nations [10–13]. Various classifiers have been employed for rainfall prediction, including the Random Forest (RF), Decision Trees (DT), support vector machine, K-Nearest Neighbors (KNN), and naïve Bayes classifiers [14–19].

Accurate trend detection and prediction are vital. Variations in rainfall, humidity, wind speed, and temperature over time, space, and aggregate can significantly impact a country's agriculture, potentially causing substantial economic setbacks [20]. Countries with diverse landscapes, such as Bangladesh, face the challenge of predicting ever-changing weather conditions involving key parameters such as wind speed, humidity, temperature, and rainfall [21]. Therefore, rainfall trend detection and prediction remain important fields of study for Bangladesh. Weather factors, including temperature, wind direction, wind speed, and amount of rainfall, can be predicted using machine learning algorithms [22].

Weather patterns in Bangladesh are significantly changing because of its proximity to the equator and rising world temperatures. The nation has experienced extreme weather variations due to these changes, including frequent flooding and other disasters [23]. Existing research works focus on rainfall and temperature predictions separately and ignore the prediction of rainfall occurrence; that is, whether it will be a rainy day or not. This study combines multiple predictions and focuses on the prediction of rainfall occurrence, rainfall amount, and daily average temperature.

This study uses machine learning algorithms and implements ensemble-based classifier and regressor models using the Bangladesh weather dataset to predict weather, including rainfall occurrence, amount, and daily average temperature. The novelty and main contributions of this study are as follows:

- An ensemble-based predictive classifier was implemented to predict whether rain will occur on a particular day.
- An ensemble-based predictive regressor was implemented to predict the rainfall amount and daily average temperature.
- The performance of the ensemble-based models with basic machine learning algorithms was evaluated using metrics including accuracy, precision, recall, F1 score, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

The remainder of this paper is organized as follows: Section II reviews the related work. Section III describes the methodology of the study. Section IV describes the design and implementation of the proposed approach. Section V presents the results and analysis. Finally, Section VI provides the conclusion.

2. Related Work

 doi:10.20944/preprints202402.1566.v2

Intelligent weather prediction techniques can provide valuable insights, enabling us to make effective decisions that save lives, time, and property. Bosu *et al.* [24] analyzed recent changes in temperature and rainfall in different areas of Bangladesh from 1981 to 2019 using the CMIP5 dataset. In [4], the authors examined the trends and variations in Bangladesh's inter-annual, monthly, and dry-season rainfall patterns by applying ARIMA predictions and conducting Mann-Kendall and Spearman's rho tests. Mahabub and Habib [21] experimented with the raw dataset collected from the Bangladesh Meteorological Division (BMD), applied regression algorithms in machine learning models, and achieved more precise results than traditional weather forecasting approaches. The regression algorithm-based machine learning model predicted more accurate results than traditional weather forecasting approaches [21]. Hashim *et al.* [25] successfully predicted precipitation using wind, temperature, pressure, and relative humidity as meteorological factors in a backpropagation neural network model. Dong *et al.* [26] enhanced the short-term forecasting of daily precipitation using the XGBoost model combined with multifactor bias correction for Numerical Weather Prediction (NWP). Paul and Roy [27] developed a machine-learning-based time-series forecasting model to predict temperatures in Bangladesh in future years.

In the past, individual classifiers such as DT, Multilayer Perceptrons (MLP), Naïve Bayes (NB), KNN, Neural Networks (NN), and Support Vector Machines (SVM) have been used to create prediction models on prelabeled datasets for rainfall prediction [15,28–31]. However, individual classifiers face certain constraints. For instance, when data exhibit unstructured and intricate characteristics, together with numerous features, the problem may be classified as nonlinear. The data are high-dimensional in nature, whereas the dataset is small. This combination of factors can result in overfitting and a lack of interpretability when training a model. The principal disadvantage of this approach is the limitation of using an individual classifier in a prediction model. The reliability of individual classifiers is lower than that achieved by combining numerous classifiers [32]. To address this limitation, numerous scholars have proposed that ensemble learning techniques offer superior classification accuracy compared to that of individual classifiers. Ensemble learning is a machine learning methodology used to enhance the accuracy of predictions [33]. Ensemble approaches are commonly considered to be the most sophisticated methods for forecasting precipitation. These strategies enhance the predictive accuracy of a single model and aggregate their forecasts. Ensemble learning has consistently produced robust models, and these techniques have been successfully employed for rainfall prediction, resulting in notable improvements in prediction accuracy. This advancement in rainfall prediction could mitigate the risk of substantial losses. Ensemble learning combines the predictive capacities of numerous classifiers to provide enhanced prediction results for a given dataset [34,35]. An overview of related work is provided in Table I.

Table I. MAE AND RMSE COMPARISON FOR RAINFALL AMOUNT PREDICTION

Refs.	Models	Prediction	Limitation
[11]	Genetic programming, Support Vector Regression (SVR), M5 rules, M5 model trees, radial basis neural network	Rainfall amount	Using traditional machine learning techniques
[17]	SVR, linear regression, ridge regression, Bayesian ridge, gradient boosting, XGBoost, CatBoost,	Windspeed, humidity, temperature, and rainfall amount	• Rainfall occurrence prediction is not implemented • Only regression-based algorithms are used

 doi:10.20944/preprints202402.1566.v2

Refs.	Models	Prediction	Limitation
	AdaBoost, KNN, decision trees		
[20]	Multi-Layer Perceptron (MLP)	Rainfall prediction using temperature, pressure, and humidity	• Uses only a single model • Rainfall occurrence prediction and temperature prediction are not implemented
[21]	XGBoost	Rainfall amount prediction	Rainfall occurrence prediction is not implemented
[22]	Linear regression, polynomial regression, and SVR	Daily min, max, and average temperature prediction	• Rainfall occurrence prediction and rainfall amount prediction are not implemented • Traditional techniques
[23]	Artificial Neural Network (ANN)	Rainfall amount prediction	Rainfall occurrence prediction is not implemented

Most of the related work uses traditional machine learning models only for a single prediction category (rainfall amount, rainfall prediction, or average temperature) and uses smaller datasets containing shorter periods of data. However, rainfall occurrence prediction is necessary in countries such as Bangladesh to mitigate flood-related challenges. In this study, to increase the performance accuracy of machine learning models, ensemble-based models are proposed for rainfall occurrence prediction using five machine learning techniques, along with rainfall amount and daily average temperature prediction using an ensemble-based regressor with three regression algorithms.

It is difficult to guarantee absolute independence among basic classifiers. Ensemble learning reduces interpretability, and the ensemble approach is challenging to predict and explain. Mastering ensemble learning is challenging because any errors made in rainfall prediction could potentially lead to a model that exhibits worse predictive accuracy than an individual model. Hence, it is also possible for losses to occur. Two distinct types of ensemble learning methods exist: heterogeneous and homogeneous ensemble learning. In ensemble methods, heterogenous approaches use identical base learners for distinct subsets of samples within a given dataset [36]. Bagging, boosting, and RF are among the various examples that can be cited. Heterogeneous ensemble strategies involve diverse base learners, constructed by either applying statistical methods or aggregating the predictions of individual base learners. Ensemble learning has emerged as a prominent approach to rainfall prediction.

3. Methodology

This study used machine learning models to conduct predictive analyses of rainfall occurrence, rainfall amounts, and daily average temperatures. These models are characterized by their unique

 doi:10.20944/preprints202402.1566.v2

methodologies, each possessing specific strengths and capabilities. Our approach incorporates both classification and regression algorithms, thereby encompassing a comprehensive array of techniques to effectively address the intricacies associated with this multifaceted task. Logistic regression, DT classification, RF classification, KNN, and Support Vector Classifier (SVC) are the models that were used to predict rainfall occurrence. Furthermore, the regression-based algorithms used to predict the rainfall amount and daily average temperature were linear regression, RF regression, and Support Vector Regression (SVR). To further improve the results, this research also implemented an ensemble classifier using these machine learning models for rainfall occurrence prediction and regression algorithms based on ensemble regressors for rainfall amount and daily average temperature prediction. SVC was used for rainfall prediction, as it performs binary classification, and SVM, which performs regression, was used for rainfall amount and daily average temperature prediction. This methodology is illustrated in Figure 1 below.

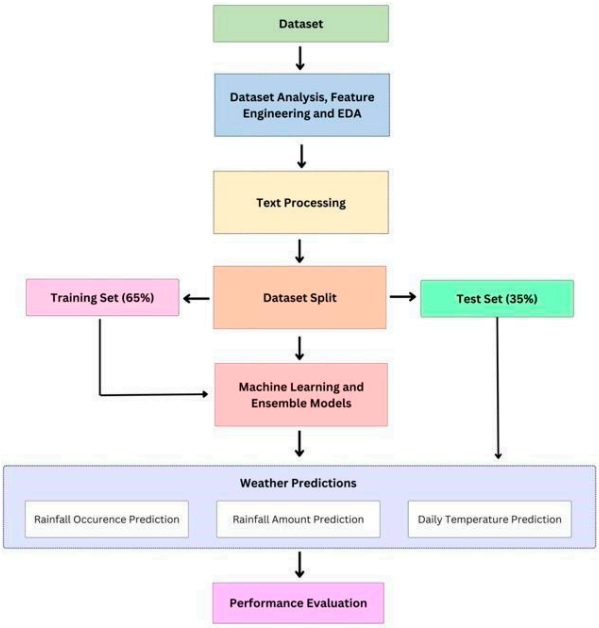


Figure 1. Methodology.

3.1. Ensemble Models

Our methodology also involves an ensemble classifier to perform comprehensive predictions. The ensemble classifier operates on the principles of consensus and voting, drawing upon a collective wisdom that transcends the limitations of any single model. As the classifiers generate independent predictions, the ensemble classifier combines these predictions, leading to a dynamic and balanced output that represents the collective insight of the entire ensemble. Specifically, the ensemble classifier predicts rainfall occurrence based on historical data patterns. This aggregated result is achieved using intricate mechanisms that prioritize reliability, accuracy, and robustness. By cultivating harmony among the classifiers, the ensemble classifier fortifies its predictive ability and diminishes the influence of potential outliers or biases in individual models. Rainfall occurrence prediction is ideally suited for classifier algorithms such as SVC, logistic regression, KNN, DT, and RF, owing to their classification characteristics. This task classifies occurrences as rainfall or non-rainfall based on the input features. Classifier algorithms are used to generate discrete class labels in this situation. They can detect patterns and class boundaries, making them good rainfall predictors. Their efficient classification ability facilitates accurate rainfall forecasts, making them the best choice for this meteorological prediction task. The ensemble-based classifier model used in this study for rainfall occurrence prediction was based on five machine-learning classifiers, as shown in Figure 2.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 September 2024 doi:10.20944/preprints202402.1566.v2

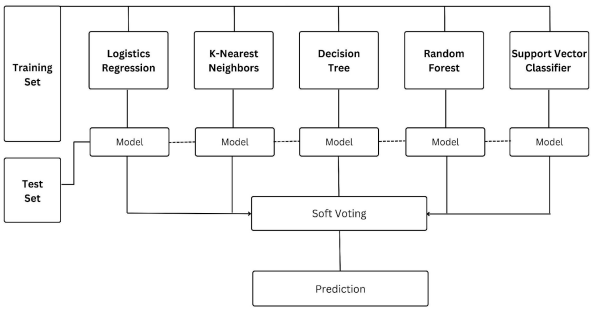


Figure 2. Ensemble-based classifier for rainfall occurrence prediction.

The ensemble-based regressor model used for the rainfall amount and daily average temperature prediction was based on three machine-learning regressor algorithms, as shown in Figure 3. Owing to their nature, regression techniques such as linear regression, RF, and SVR are suitable for rainfall and daily average temperature prediction. Both aim to estimate continuous numerical quantities (rainfall or temperature) from input features. For this reason, regression algorithms are designed to capture and model data patterns and correlations. Their capacity to handle continuous output variables and adapt to complex data patterns makes them the best choices for regression-based meteorological predictions of rainfall and daily temperatures.

Ensemble-based forecasting techniques have emerged as crucial tools for improving the accuracy and reliability of weather predictions, particularly in regions characterized by dynamic and complex climatic patterns, such as Bangladesh. The proposed approach integrates multiple models and their outputs to generate robust and comprehensive forecasts. In Bangladesh, accurate rainfall and temperature predictions are paramount, as they directly impact various sectors such as agriculture, water resource management, and disaster preparedness. This approach explores the significance and applications of ensemble-based forecasting methods in addressing Bangladesh's unique meteorological challenges, and highlights the potential benefits and contributions of such models in improving the country's resilience to changing weather patterns.

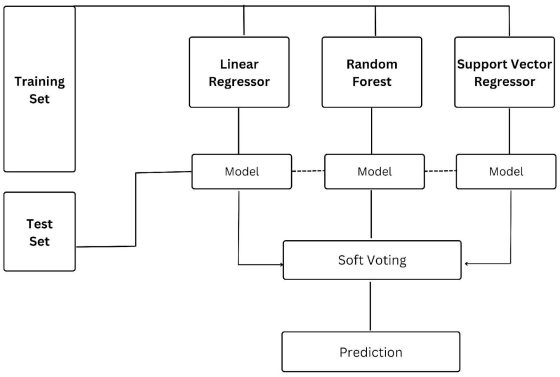


Figure 3. Ensemble-based regressor for rainfall amount and daily average temperature prediction.

3.2. Dataset

Weather Data Bangladesh is an open-source dataset that contains 10 years of daily weather observations from many locations across Bangladesh [37]. It contains observations of weather metrics for each day from 2013 to 2022. The dataset includes the columns Date, MinTemp, MaxTemp, WindDir9am, WindDir3pm, Windspeed9am, windspeed3pm, humidity9am, humidity3pm, pressure9am, pressure3pm, cloud9am, cloud3pm, temp9am, temp3pm, and rainToday. If a given day is rainy, then this value is "Yes." Otherwise, it is "No."

3.3. Evaluation Metrics

 doi:10.20944/preprints202402.1566.v2

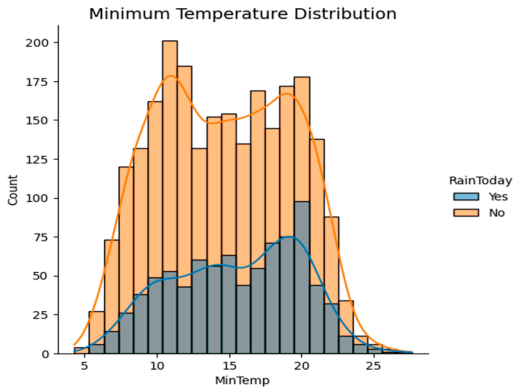
The machine learning models implemented in this study were evaluated using the accuracy Score, F1 score, MAE, and Root Mean Squared Error (RMSE). The accuracy and F1 score were used to predict rainfall occurrence. However, the MAE and RMSE were used to predict rainfall amounts and daily average temperatures.

4. Data Analysis

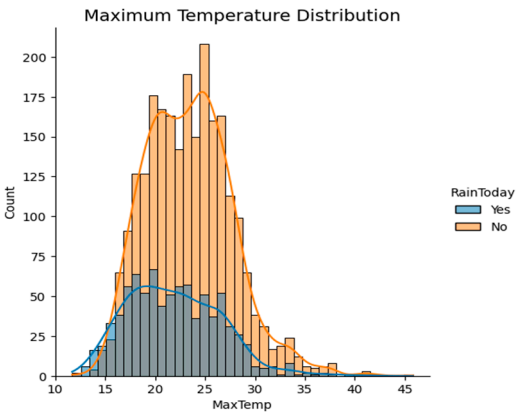
To further analyze the data, a comprehensive analysis was performed for various attributes, including the minimum and maximum temperature, wind speed, humidity, pressure, clouds, and temperature. The dataset analysis included wind speed, humidity, pressure, clouds, and temperature values at only 9 AM and 3 PM.

4.1. Feature Distribution

The dataset's minimum temperature ranges from 4.3 to 27.6 degrees Celsius, with the largest frequency at 11 and 20 degrees Celsius, as illustrated in Figure 4(a). However, as Figure 4(b) illustrates, the maximum temperature ranges from 11.7 degrees Celsius to 45.8 degrees Celsius, with the largest frequency at 25 degrees Celsius.



(a)



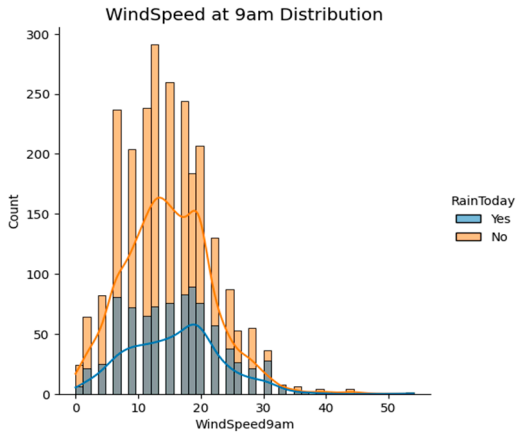
(b)

Figure 4. Temperature distribution. (a) Minimum temperature; (b) Maximum temperature.

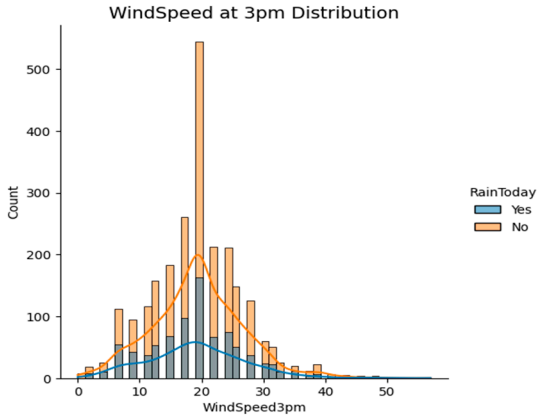
At 9 AM, the wind speed ranges from 0 to 57 km/h, with 12 km/h having the largest frequency in the dataset, as shown in Figure 5(a). Figure 5(b) shows that the wind speed range at 3 PM is also 0 to 57 km/h, with the highest frequency occurring at 19 km/h.

The humidity ranges from 19% to 100% between 9 AM and 3 PM, with 70% humidity at 9 AM exhibiting the highest frequency in the dataset, as shown in Figure 6(a). In comparison, Figure 6(b) shows that 47.58% humidity at 3 PM has the highest frequency.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 September 2024 doi:10.20944/preprints202402.1566.v2

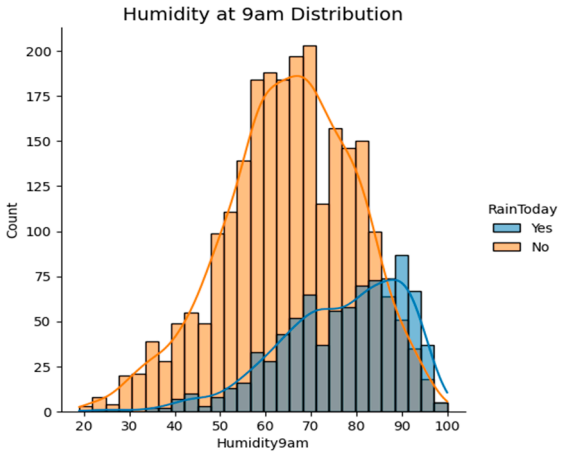


(a)



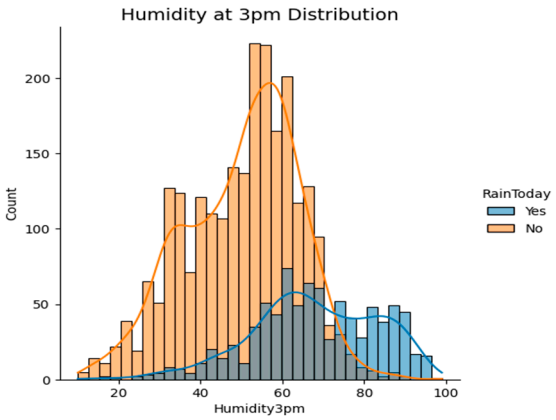
(b)

Figure 5. Wind speed distribution. (a) Wind speed at 9 AM; (b) Wind speed at 3 PM.



(a)

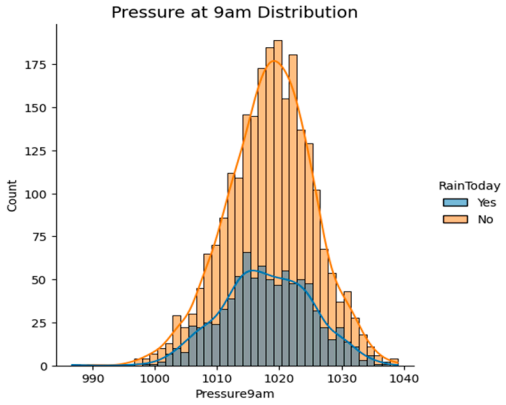
 doi:10.20944/preprints202402.1566.v2



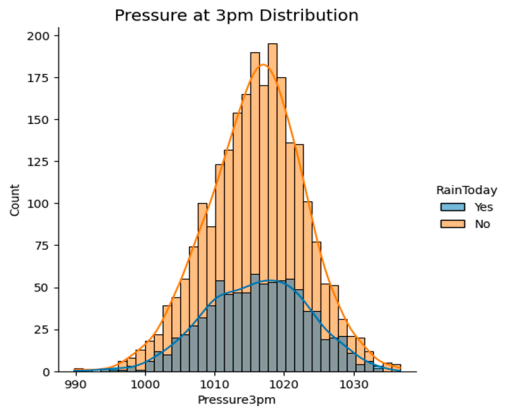
(b)

Figure 6. Humidity distribution. (a) Humidity at 9 AM; (b) Humidity at 3 PM.

The wind pressure ranges from 980.5 to 1,042 hPa at 9:00 AM. Figure 7(a) illustrates that 1024.68 hPa pressure has the highest frequency in the dataset. Figure 7(b) illustrates that the pressure range at 3 PM is 988.2 to 1,039.6 hPa, with 1,015.28 hPa pressure having the highest frequency in the dataset.



(a)



(b)

Figure 7. Wind pressure distribution. (a) Pressure at 9 AM; (b) Pressure at 3 PM.

4.2. EDA

In this section, the average speed, humidity, and temperature per month are analyzed.
Average wind speed analysis

 doi:10.20944/preprints202402.1566.v2

The monthly average wind speed data help us understand how wind patterns change seasonally. The results of this study show that wind speeds are usually not too high in the first few months of the year. At 9 AM in January, the average wind speed was approximately 15.29 km/h. It increased slightly in February, reaching 15.47 km/h. There was a small increase in the average wind speed at 9 AM in March, reaching approximately 15.99 km/h. Based on this view, it appears that spring has begun. As spring turns into late spring and early summer, the wind speeds start to increase. The average wind speed at 9 AM in April increased, hitting 16.47 km/h. Subsequently, in May, the wind speed jumped to 16.58 km/h. There was a noticeable drop in wind speed in June, with a recorded speed of 15.08 km/h at 9 AM, suggesting a generally calm atmosphere. As the spring months give way to summer, particularly in July and August, the wind speeds tend to increase even more. At 9 AM in July, the wind speed reached 14.61 km/h. Following this, the wind speed increased in August, reaching 13.65 km/h. Based on these values, the wind was stronger during this period. As summer turned into autumn, the wind speed increased greatly in September, with an average of 13.82 kilometers per hour at 9 AM. At 9 AM. in October, the wind usually blew at 13.90 km/h. In the last few months of the year, especially in November and December, wind speeds dropped from their highest point in the summer to very low levels. The average wind speed at 9 AM in November was 14.92 km/h, and at the same time in December, it was 15.21 km/h, marking the end of the year. The results of the average wind speed analysis are shown in Figure 8.

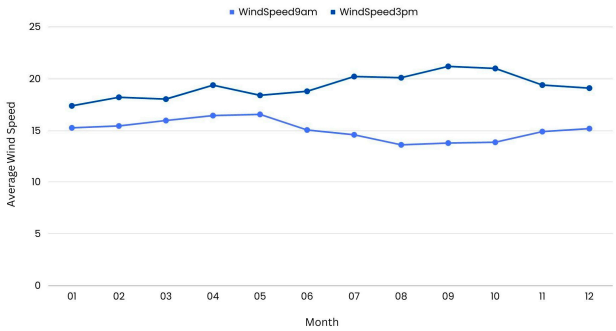


Figure 8. Average wind speed per month.

Average humidity analysis

According to the dataset, there were clear regular patterns in the average relative humidity in Bangladesh during each season. In January, the humidity started relatively high, averaging 73.57% at 9 AM and 57.14% at 3 PM. During the winter months of February and March, the humidity gradually decreased, reaching approximately 68.52% at 9 AM and 52.70% at 3 PM. The humidity decreased even more in April and May, which are spring months. In April, it was 67.29% at 9 AM. and 52.37% at 3 PM.; in May, it was 64.01% at 9 AM. and 49.91% at 3 PM. When summer returns to June, the humidity begins to increase again. At 9 AM, it was approximately 66.43%, and at 3 PM, it was only 54%. This growth trend is maintained in August and July. In August, the average humidity was 52.33% at 3 PM, 65.18% at 9 AM, 64.16% at 9 AM and 52.87% at 3 PM. The wettest month of summer was September, when the humidity was approximately 65.84% at 9 AM and 56.10% at 3 PM. In autumn, however, the humidity levels slowly decreased. On average, they dropped from approximately 71.20% at 9 AM to 59.14% at 3 PM. in October and November to 70.50% at 9 AM and 58.51% at 3 PM. December saw a return to lower humidity levels, with averages of 70.01% at 9 AM and 55.21% at 3 PM. The average monthly humidity analysis is illustrated in Figure 9.

 doi:10.20944/preprints202402.1566.v2

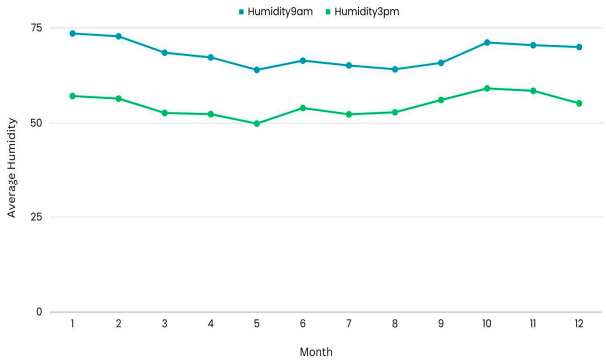


Figure 9. Average humidity per month.

Average temperature analysis

The average temperature data analysis revealed unique seasonal patterns. The weather is warm in January and February, with an average high temperature of 22.6 °C and pleasant afternoon temperatures. In March and April, the start of spring, temperatures rose to a modest level, with a maximum average increase of about 21.3 °C. In May, the temperature increased even more, reaching a high of 21.9 °C. There was a gradual increase in temperature from June to August, which is summer. July and August had the highest normal maximum temperatures at approximately 22.5 °C. The temperature steadily dropped in September and October, changing from 24.8 °C to 24.4 °C in October. This is because it was fall. November and December are the last two months of the year. The weather is usually mild during these months, with normal high temperatures of about 24.4 °C and 24.1 °C. The results of the monthly average temperature analysis are shown in Figure 10.

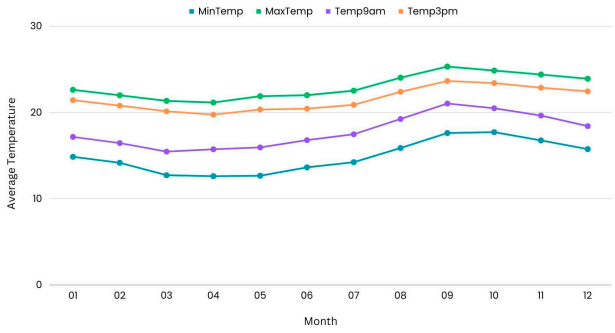


Figure 10. Minimum, maximum, and average temperature per month.

5. Implementation

5.1. Data Processing

The dataset used in this study was obtained from Kaggle. It consists of 10 years of Bangladesh weather data, with daily weather observations from many locations across the country for each day from 2013 to 2022. The preprocessing steps used to process the data before model training included various steps, including standardize variables and transforming categorical variables.

5.2. Standardizing the Variables

The scale of the variables is significant because the classifier and regressor use the identification of the nearest test observations to predict the class and values of a given test observation. Variables with a large scale exert a considerably greater influence on the distance between observations. Standard Scaler was implemented to standardize the variables during preprocessing. This process involves calculating the mean and standard deviation for each feature. The scaler then subtracts the mean from each feature and divides the result by the standard deviation. Furthermore, the transform

 doi:10.20944/preprints202402.1566.v2

method of Standard Scaler was used for scaling transformation based on the mean and standard deviation of the parameters.

5.3. Transforming Categorical Variables

Initially, it is important to transform categorical values into binary variables. The get_dummies() method in pandas was used for this purpose. Then, the categorical values within the "RainToday" column were substituted with binary values, transforming the column from a categorical representation to a binary one. The get_dummies method was not utilized to avoid the creation of duplicate columns for the variable "RainToday," the target variable of interest.

5.4. Rainfall Occurrence Prediction

The rainfall occurrence prediction uses machine learning techniques including logistic regression, KNN, DT, RF, SVC, and ensemble-based classifiers. The RainToday column was selected as the target for predicting the occurrence of precipitation. With a training set size of 0.65, a test set size of 0.35, and a random state of 101, the train_test_split function splits the features and Y data frames for training and testing all models, including the ensemble classifier.

5.5. Rainfall Amount Prediction

The rainfall amount prediction uses regression algorithms such as logistic regression, RF, SVR, and ensemble-based regression. The rainfall column is the target variable for predicting the amount of rainfall in millimeters on a given day. With a training size of 0.65, a test_size of 0.35, and a random state of 101, the train_test_split function splits the features and Y data frames for training and testing all regressor algorithms, including the ensemble regressor.

5.6. Daily Average Temperature Prediction

The daily average temperature prediction was also performed using regression algorithms such as logistic regression, RF, SVR, and ensemble-based regression. The average temperature for a certain day was predicted in degrees Celsius. The Temp9am, Temp3pm, MinTemp, and MaxTemp columns are used to generate a new column, AvgTemp, which is the target variable. With a training size of 0.65, a test_size of 0.35, and a random state of 101, the train_test_split function splits the features and Y data frames for training and testing all regressor algorithms, including the ensemble regressor.

6. Result and Discussion

This section presents the results for predicting rainfall occurrence using machine learning models and the ensemble classifier, as well as for predicting the rainfall amount and daily average temperature using regression algorithms and the ensemble classifier.

6.1. Rainfall Occurrence Prediction

When comparing the different classification algorithms based on their accuracy, precision, recall, and F1 score, the ensemble classifier exhibited the highest accuracy (83.41%). Based on this result, the ensemble classifier can correctly predict rainfall occurrence. However, the model's precision, that is, the number of correct positive guesses, was 51.16%, which is the same as the precision of the RF model. The recall of the ensemble classifier was better, with a value of 78.17%, which means that it found all relevant cases effectively. It is important to note that the DT method has a significantly lower recall rate of 56.82%. There is considerable consistency between all models in terms of the F1 score, which balances precision and recall. With an F1 score of 61.85%, the ensemble classifier performed better than the others. Overall, the logistic regression model performed well, with an F1 score of 62.03%, an accuracy of 82.36%, a precision of 54.82%, and a recall of 71.43%.

Regarding precision and F1 score, SVC works in the same manner as logistic regression. However, it exhibits poor accuracy and recall. The KNN algorithm had an impressively high F1 score

 doi:10.20944/preprints202402.1566.v2

of 82.53%, although it was less accurate at 80%. This F1 score indicates that KNN strikes a good balance between accuracy and recall when correctly guessing events in its environment. The RF accuracy is slightly more accurate than the SVC algorithm (82.97% for RF and 82.89% for the SVC). It is important to note that RF has the lowest precision, at 51.17%. Taken together, these measurements show that the ensemble classifier has better accuracy and recall; however, it is important to consider the loss of precision. The logistic regression and SVC models, on the other hand, performed better across all measures. Table II lists the accuracy, precision, recall, and F1 scores for rainfall occurrence prediction of the machine learning models and ensemble classifier.

Table II. MAE AND RMSE COMPARISON FOR RAINFALL AMOUNT PREDICTION

Models	Accuracy	Precision	Recall	F1 Score
Logistic Regression	82.36%	54.81%	71.42%	62.03%
KNN	80.0%	36.87%	74.0%	82.52%
Decision Tree	77.47%	59.47%	56.82%	58.11%
SVC	82.89%	52.82%	74.64%	61.87%
Random Forest	82.97%	51.17%	76.23%	61.23%
Ensemble Classifier	83.40%	51.17	78.18%	61.84%

Figure 11 compares the machine learning classifiers with the ensemble-based classifier using the accuracy, precision, recall, and F1 score of the ensemble-classifier machine learning models.

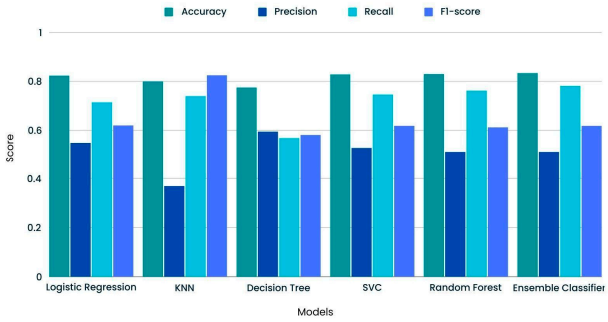


Figure 11. Performance comparison for rainfall amount prediction.

6.2. Comparison with Literature

The results of rainfall prediction using an ensemble classifier were compared with those of a similar study [34], which implemented an ensemble-based model using multiple machine learning methods, including NB, DT, SVM, RF, NN, and artificial performance metrics including accuracy, precision, recall, and F1 core.

Table III compares the performance of the proposed model with that of the existing ensemble-based model.

Table III. PERFORMANCE COMPARISON WITH LITERATURE

 doi:10.20944/preprints202402.1566.v2

Models	Accuracy	Precision	Recall	F1 Score
Combination of (SVM, ANN, NB, C4.5, RF) [28]	75%	53%	73%	61%
Ours	83%	51%	78%	61%

6.3. Rainfall Amount Prediction

The effectiveness of various machine learning regression models can be evaluated using MAE and RMSE, which are two essential metrics for assessing predicted accuracy. Significant variations in the accuracy of the regression algorithms were observed when comparing their performance. Linear regression exhibited comparatively greater MAE and RMSE values, indicating constraints on its predictive accuracy. In contrast, the RF algorithm demonstrated enhanced performance by exhibiting lower MAE and RMSE values, suggesting a higher accuracy in its predictive capabilities. The SVR model had a decreased MAE but a greater RMSE, indicating the possibility of enhancing the predictive accuracy. The ensemble regression model performed better by achieving an optimal trade-off between MAE and RMSE. Consequently, it outperformed all other assessed models in terms of prediction accuracy. Table IV lists the MAE and RMSE for rainfall amount prediction using the regression algorithms and ensemble classifier.

Table IV. MAE AND RMSE COMPARISON FOR RAINFALL AMOUNT PREDICTION

Algorithms	MAE	RMSE
Linear Regression	0.498774	0.948272
Random Forest	0.378243	0.882860
Support Vector Regression (SVR)	0.365070	0.971967
Ensemble Regression	0.363691	0.904688

Figure 12 compares the MAE and RMSE values of the regression algorithms and the ensemble classifier.

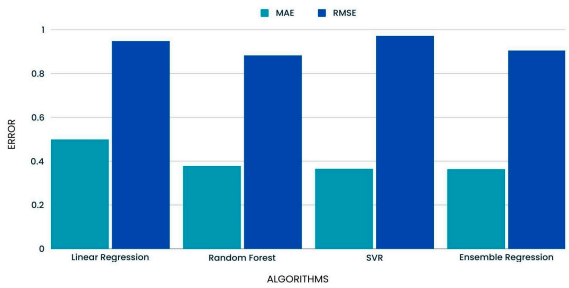


Figure 12. MAE and RMSE comparison for rainfall amount prediction.

6.4. Daily Average Temperature Prediction

The performance of various machine learning regression models was evaluated using two critical metrics: MAE and RMSE. Linear regression, despite exhibiting a certain degree of predictive capability, demonstrated comparatively higher MAE and RMSE values. In contrast, the RF model demonstrated enhanced accuracy, as evidenced by the lower MAE and RMSE values, suggesting

 doi:10.20944/preprints202402.1566.v2

more accurate predictions. The efficiency of SVM in generating accurate predictions is indicated by its lower MAE and RMSE values. Nevertheless, the ensemble regression model performed better than the other assessed models. This model demonstrated a commendable equilibrium between MAE and RMSE, resulting in the most precise predictions among the examined models. Table V lists the MAE and RMSE values for the daily average temperature prediction using the regression algorithms and ensemble classifier.

Table V. MAE AND RMSE COMPARISON FOR DAILY AVERAGE TEMPERATURE PREDICTION

Algorithms	MAE	RMSE
Linear Regression	0.470631	0.603241
Random Forest	0.450968	0.570240
Support Vector Regression (SVR)	0.434701	0.560317
Ensemble Regression	0.425209	0.545714

A comparison of the MAE and RMSE of the models and the ensemble classifier is shown in Figure 13.

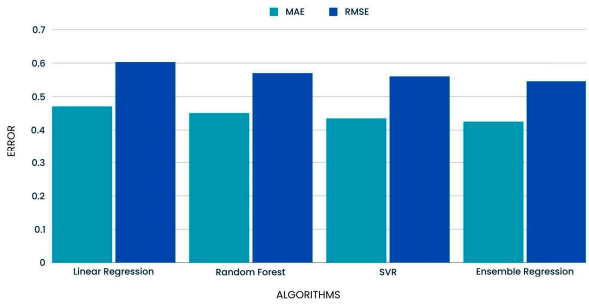


Figure 13. MAE and RMSE comparison for rainfall amount prediction.

The results indicated that the ensemble-based models outperformed the machine learning models in predicting rainfall occurrence, rainfall amount, and daily average temperature. The ensemble classifier had the highest accuracy (83.41%) and recall (78.17%) in predicting rainfall. However, its precision was tied for the lowest, at 51.16%. Despite a lower accuracy of 80%, KNN's high F1 score of 82.53% implies a stable equilibrium. The ensemble regression model surpassed linear regression, RF, and SVR in predicting the amount of precipitation, as evidenced by its lowest MAE of 0.363691 and RMSE of 0.904688. The Ensemble Regression model also outperformed other regression models in daily average temperature prediction, producing the most accurate results, with an MAE of 0.425209 and RMSE of 0.545714 as the lowest error values. Ensemble approaches consistently provided a performance advantage across all tasks.

7. Conclusion

Forecasting rainfall involves considering numerous variables, such as temperature, humidity, wind speed, and water level, to estimate where rainfall might occur. The most popular techniques used in rainfall forecasting are supervised machine learning techniques that use testing data to make predictions after training predetermined example data. Finding appropriate mechanisms, balancing the sensitivity of the objective functions, and handling characteristics all present significant challenges for these systems. These variations result in variable performance, making it difficult to select an appropriate technique for rainfall prediction. This study used the Bangladesh Weather Dataset to implement machine learning algorithms and ensemble-based models for weather

 doi:10.20944/preprints202402.1566.v2

forecasting, including rainfall occurrence prediction, rainfall amount prediction, and daily average temperature prediction. The ensemble-based model was used to improve prediction performance. The ensemble-based model used for rainfall occurrence prediction was based on a voting classifier that uses five machine learning algorithms. However, for the rainfall amount and daily average temperature, the ensemble regressor was formed by combining regression-based algorithms. The models were trained and tested using the dataset.

The results showed that the ensemble-based models performed better than the machine learning models for rainfall occurrence, amount, and daily average temperature prediction. The ensemble classifier exhibited the highest accuracy of 83.41% and recall of 78.17% in predicting the occurrence of rainfall. However, its precision was the lowest (51.16%). Although KNN achieved a lower accuracy of 80%, its high F1 score of 82.53% indicates a robust equilibrium. The ensemble regression model outperformed the linear regression, RF, and SVR models in predicting the amount of precipitation, as evidenced by its lowest MAE of 0.363691 and RMSE of 0.904688. The ensemble regression model also demonstrated its superiority over alternative regression models in daily average temperature prediction by yielding the most accurate results, with an MAE of 0.425209 and RMSE of 0.545714 as the lowest error values. The ensemble methods demonstrated a consistent advantage in terms of performance metrics across all tasks.

The main objective of this study was to accurately predict rainfall occurrence and amount, along with the daily average temperature, using ensemble-based models and machine learning models. An accurate weather forecast helps mitigate the challenges of heavy rainfall, especially in Bangladesh, which has an agriculture-based economy.

In the future, ensemble-based models and other machine learning models can be applied using multiple datasets, and their performance can be evaluated. In addition, advanced deep learning models can be applied for similar predictions and compared with machine learning models.

Author Contributions: The contributions of the authors is as follows: "Conceptualization, A. Hussain. V. Dhanawat; methodology, A. Hussain. A. Aslam and V. Dhanawat; software, S. Tripura.; validation, A. Hussain, and A. Aslam; formal analysis, S. Tripura, V. Dhanawat and V. Shinde; investigation, A. Hussain resources, A. Hussain and A. Aslam; data curation, A. Hussain and A. Aslam; writing—original draft preparation, A. Hussain and A. Aslam; writing—review and editing, A. Hussain and V. Dhanawat; visualization, A. Aslam, S. Tripura and V. Shinde; supervision, V. Dhanawat; project administration, V. Shinde; funding acquisition, A. Hussain, V. Dhanawat, and V. Shinde. All authors have read and agreed to the published version of the manuscript."

Conflict of Interest: The authors declare no conflict of interest.

References

1. S. Badhiye, P. Chatur, and B. Wakode, "Temperature and humidity data analysis for future value prediction using clustering technique: an approach," *International Journal of Emerging Technology and Advanced Engineering,* vol. 2, no. 1, pp. 88–91, 2012.
2. K. Pabreja, "Clustering technique to interpret Numerical Weather Prediction output products for forecast of Cloudburst," *International Journal of Computer Science and Information Technologies (IJCSIT),* vol. 3, no. 1, pp. 2996–2999, 2012.
3. A. Parmar, K. Mistree, and M. Sompura, "Machine learning techniques for rainfall prediction: A review," in *Proc. International conference on innovations in information embedded and communication systems*, vol. 3, 2017.
4. S. Kundu, S. K. Biswas, D. Tripathi, R. Karmakar, S. Majumdar, and S. Mandal, "A review on rainfall forecasting using ensemble learning techniques," *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, 100296, 2023.
5. M. E. Mann and P. H. Gleick, "Climate change and California drought in the 21st century," *Proceedings of the National Academy of Sciences,* vol. 112, no. 13, pp. 3858–3859, 2015.
6. A. Malki, E.-S. Atlam, and I. Gad, "Machine learning approach of detecting anomalies and forecasting time-series of IoT devices," *Alexandria Engineering Journal,* vol. 61, no. 11, pp. 8973–8986, 2022.
7. A. Hussain and A. Aslam, "Cardiovascular disease prediction using risk factors: A comparative performance analysis of machine learning models," *Journal on Artificial Intelligence,* vol. 6, no. 1, pp. 129–152, 2024.
8. A. Aslam and A. Hussain, "A performance analysis of machine learning techniques for credit card fraud detection," *Journal of Artificial Intelligence (2579-0021),* vol. 6, 2024.

9. A. Hussain, A. Khatoon, A. Aslam, Tariq, and M.-A. Khosa, "A comparative performance analysis of machine learning models for intrusion detection classification," *Journal of Cyber Security,* vol. 6, no. 1, pp. 1–23, 2024.
10. C. C. Stephan, N. P. Klingaman, P. L. Vidale, A. G. Turner, M.-E. Demory, and L. Guo, "A comprehensive analysis of coherent rainfall patterns in China and potential drivers. Part I: Interannual variability," *Climate Dynamics,* vol. 50, pp. 4405–4424, 2018.
11. N. A. B. Klutse, B. J. Abiodun, B. C. Hewitson, W. J. Gutowski, and M. A. Tadross, "Evaluation of two GCMs in simulating rainfall inter-annual variability over Southern Africa," *Theoretical and applied climatology,* vol. 123, pp. 415–436, 2016.
12. K. Sittichok, A. G. Djibo, O. Seidou, H. M. Saley, H. Karambiri, and J. Paturel, "Statistical seasonal rainfall and streamflow forecasting for the Sirba watershed, West Africa, using sea-surface temperatures," *Hydrological Sciences Journal,* vol. 61, no. 5, pp. 805–815, 2016.
13. J. Wu, J. Long, and M. Liu, "Evolving RBF neural networks for rainfall prediction using hybrid particle swarm optimization and genetic algorithm," *Neurocomputing,* vol. 148, pp. 136–142, 2015.
14. N. Singh, S. Chaturvedi, and S. Akhter, "Weather forecasting using machine learning algorithm," in *Proc. 2019 International Conference on Signal Processing and Communication (ICSC)*, 2019, pp. 171–174.
15. S. Cramer, M. Kampouridis, A. A. Freitas, and A. K. Alexandridis, "An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives," *Expert Systems with Applications,* vol. 85, pp. 169–181, 2017.
16. N. Srinu and B. H. Bindu, "A review on machine learning and deep learning based rainfall prediction methods," in *Proc. 2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)*, 2022, pp. 1–4.
17. E. Dritsas, M. Trigka, and P. Mylonas, "A multi-class classification approach for weather forecasting with machine learning techniques," in *Proc. 2022 17th International Workshop on Semantic and Social Media Adaptation & Personalization (SMAP)*, 2022, pp. 1–5.
18. S. Choi and E.-S. Jung, "Optimizing numerical weather prediction model performance using machine learning techniques," *IEEE Access,* 2023.
19. S. Nigam, M. Gupta, A. Shrinivasan, A. V. S. Uttej, C. Kumari, and P. Disha, "Comparative study to determine accuracy for weather prediction using machine learning," in *Proc. 2023 International Conference on Computer Communication and Informatics (ICCCI)*, 2023: IEEE, pp. 1–4.
20. M. A. Rahman, L. Yunsheng, and N. Sultana, "Analysis and prediction of rainfall trends over Bangladesh using Mann–Kendall, Spearman's rho tests and ARIMA model," *Meteorology and Atmospheric Physics,* vol. 129, no. 4, pp. 409–424, 2017.
21. A. Mahabub and A. Habib, "An overview of weather forecasting for Bangladesh using machine learning techniques," *Machine Learning,* pp. 1–36, 2019.
22. H. Shaiba *et al.*, "S," *Computers, Materials and Continua,* vol. 73, no. 2, 2022.
23. M. A. Al Mamun *et al.*, "Identification of influential weather parameters and seasonal drought prediction in Bangladesh using machine learning algorithm," *Scientific reports,* vol. 14, no. 1, p 566, 2024.
24. H. Bosu, T. Rashid, A. Mannan, and J. Meandad, "Trends of rainfall and temperature in bangladesh: A comparative analysis of CMIP5 results and meteorological station data," *The Dhaka University Journal of Earth and Environmental Sciences,* vol. 9, no. 2, pp. 9–18, 2020.
25. F. Hashim, N. N. Daud, K. Ahmad, J. Adnan, and Z. Rizman, "Prediction of rainfall based on weather parameter using artificial neural network," *Journal of Fundamental and Applied Sciences,* vol. 9, no. 3S, pp. 493–502, 2017.
26. J. Dong, W. Zeng, L. Wu, J. Huang, T. Gaiser, and A. K. Srivastava, "Enhancing short-term forecasting of daily precipitation using numerical weather prediction bias correcting with XGBoost in different regions of China," *Engineering Applications of Artificial Intelligence,* vol. 117, 105579, 2023.
27. S. Paul and S. Roy, "Forecasting the average temperature rise in Bangladesh: A time series analysis," *Journal of Engineering Science,* vol. 11, no. 1, pp. 83–91, 2020.
28. J. Sulaiman and S. H. Wahab, "Heavy rainfall forecasting model using artificial neural network for flood prone area," in *Proc. IT Convergence and Security 2017*, Springer, vol. 1, 2018, pp. 68–76.
29. B. T. Pham, D. Tien Bui, M. Dholakia, I. Prakash, and H. V. Pham, "A comparative study of least square support vector machines and multiclass alternating decision trees for spatial prediction of rainfall-induced landslides in a tropical cyclones area," *Geotechnical and Geological Engineering,* vol. 34, pp. 1807–1824, 2016.
30. M. Kim, Y. Kim, H. Kim, W. Piao, and C. Kim, "Evaluation of the k-nearest neighbor method for forecasting the influent characteristics of wastewater treatment plant," *Frontiers of Environmental Science & Engineering,* vol. 10, pp. 299–310, 2016.
31. S. Zainudin, D. S. Jasim, and A. A. Bakar, "Comparative analysis of data mining techniques for Malaysian rainfall prediction," *Int. J. Adv. Sci. Eng. Inf. Technol,* vol. 6, no. 6, pp. 1148–1153, 2016.

32. Y. Zhang, S. Xu, L. Zhang, W. Jiang, S. Alam, and D. Xue, "Short-term multi-step-ahead sector-based traffic flow prediction based on the attention-enhanced graph convolutional LSTM network (AGC-LSTM)," *Neural Computing and Applications*, pp. 1–20, 2024.
33. O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, e1249, 2018.
34. N. S. Sani, A. H. Abd Rahman, A. Adam, I. Shlash, and M. Aliff, "Ensemble learning for rainfall prediction," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, 2020.
35. Y. Ren, L. Zhang, and P. N. Suganthan, "Ensemble classification and regression-recent developments, applications and future directions," *IEEE Computational intelligence magazine*, vol. 11, no. 1, pp. 41–53, 2016.
36. G. Kunapuli, *Ensemble Methods for Machine Learning*, Simon and Schuster, 2023.
37. Kaggle. (September 2023). Weather data bangladesh: Rain and temperature prediction based on weather data using machine learning. [Online]. Available: https://www.kaggle.com/datasets/apurboshahidshawon/weatherdatabangladesh