IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 30, No. 3, March, 2025, pp. 655-663
Journal homepage: http://iieta.org/journals/isi

Comparing the Performance of Machine Learning Models for a Day Ahead Rainfall Prediction

Shilpa Hudnurkar[1*], Prashant Patel[2], Sivagami Ponnalagarsamy[3]

[1] Department of Electronics & Telecommunication Engineering, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412115, India
[2] Instrumentation Department, Dr. D. Y. Patil Institute of Technology (DIT), Pune 411018, India
[3] Electrical and Electronics Engineering Department, Sathyabama Institute of Science and Technology, Chennai 600119, India

Corresponding Author Email: Shilpa.hudnurkar@sitpune.edu.in

https://doi.org/10.18280/isi.300310

Received: 7 September 2024
Revised: 14 January 2025
Accepted: 24 March 2025
Available online: 31 March 2025

Keywords:
rainfall, machine learning, MLR, K-NN, RF, DT, root mean square error, mean absolute error

ABSTRACT

Rainfall is a key factor in regulating the water levels in reservoirs. Due to climate change, the unpredictable nature of rainfall can lead to either overflow or drought conditions. This paper explores the use of machine learning techniques to predict a day ahead rainfall for two regions of India. The time series data of daily rainfall was available with three dimensions namely latitude, longitude and time and for 35 longitudes and 33 latitudes covering the Indian region. The data was processed to get the rainfall time series over two regions of India, namely Pune Region and Panchmahal Region from Maharashtra and Gujarat states respectively. The results were obtained for the test dataset from various models such as Multiple Linear Regression (MLR), Decision Tree (DT), K-Nearest Neighbor (K-NN), and Random Forest (RF). The performance of these models was gauged from root mean squared error (MSE), mean absolute error (MAE), correlation coefficient and the model's capability to predict maximum rainfall compared to actual rainfall. The performance analysis revealed that for the Pune region K-NN model outperformed while RF outperformed in the case of Panchmahal region.

1. INTRODUCTION

Weather forecasting, the science of predicting atmospheric conditions in a specific area, plays a crucial role in shaping people's daily lives. It is not only essential for human activities but also impacts non-living systems, influencing decisions ranging from crisis management to addressing agricultural challenges [1].

In recent decades, meteorological science in India has made significant advancements. However, due to the inherent complexity and chaotic nature of weather variables, accurate forecasting remains a challenge. The unpredictability of atmospheric movements makes it difficult to provide reliable forecasts, even for the following day [2].

Although extreme weather events are rare and typically short-lived, they can be highly destructive, leaving long-lasting impacts on the environment. While preventing such occurrences is impossible, effective mitigation strategies can help minimize their effects. Therefore, prioritizing preparedness, prevention, and rescue operations is essential.

The origins of contemporary numerical weather prediction (NWP) may be traced to the 1920s [3]. Today, it is a crucial instrument in economic planning for several industries, including transportation, logistics, agriculture, and energy production. These forecasts have been crucial in saving many lives by giving early warnings of hazardous weather occurrences, with their accuracy having considerably improved over the years. The improvement in weather prediction can be due to better processing power, a better comprehension of micro-scale processes, and higher-quality atmospheric observations, which result in better model initialization through data assimilation. However, NWP also has some challenges and limitations [4]. Nonetheless, this challenge also allows machine learning (ML) models to learn patterns from data. Weather prediction can also be considered a Big Data Analytics issue. Continuous training for ML model often requires enormous data.

The ML techniques used by researchers for weather forecasting differ in many ways. They vary in the way weather data is used, the way in which the weather or climate data is processed, input variables (univariate/multivariate) used for predicting a variable, the scale of prediction, and the region for which the prediction is being made. Some researchers have used classification approaches such as the temperature will be high or normal or low; it is raining or not raining [5]. Some researchers have used a regression approach to predict weather variables. Hence, the results they obtained cannot be compared fairly.

This paper focuses on one weather variable, i.e., rainfall, and fully utilizes ML's potential for forecasting day-ahead rainfall conditions. ML has revolutionized several sectors and is the perfect route for enhancing prediction efficiency and accuracy in weather forecasting. ML is a technique for automating the learning of a task and for enhancing

performance by iteratively improving the learning process. The objective of this study is to explore the basic ML techniques and figure out the correlation between all these ML techniques in forecasting rainfall patterns. The algorithms used here are MLR, DT, RF, and K-NN. The performance of all the models has been analyzed on the basis of MAE, RMSE, and correlation coefficient (R-value). One more parameter for comparison is considered here is the capability of the model to predict the maximum rainfall. As the rainfall pattern varies from place to place, two different regions on the same longitude but on different longitude are considered here to gauge how well ML techniques can predict a day ahead rainfall.

The paper is organized into five sections. After the introduction, the second section reviews the literature; the third section details the methodology. The methodology section covers the dataset description, and a brief description of the ML techniques used in the work. Results are discussed in the fourth section, and the fifth section concludes the paper.

2. LITERATURE REVIEW

Researchers have utilized various ML models to predict weather, especially rainfall. The prediction was either a day ahead, a month ahead, or a seasonal prediction. Gaussian Process Regression (GPR) was used to predict heavy and light rainfall days over a coastal area of India [6]. They used climatological data of daily rainfall in their work. They compared the performance of the GPR model with K-NN, RF, and DT. The performance of GPR was found to be better than that of other models in terms of MAE, MSE, and root mean square error (RMSE). The application of deep learning (DL) for forecasting extreme weather conditions has been suggested [7]. The various deep-learning models and their benefits and drawbacks in predicting extreme weather have been covered in this article. Richardson used a slide rule and a table of logarithms to create the first dynamically modeled numerical weather forecast for a single place in 1922 [3]. It took six weeks to create a 6-hour forecast of the atmosphere.

Adaline, an adaptive method for classifying patterns, has been used in the development of weather forecasting utilizing artificial neural networks and data science. It involves training over a 24-hour period with the help of fluctuating wind direction and sea level atmospheric pressure [8]. In more than 90 distinct, independent scenarios, Adaline was able to forecast "rain vs. no-rain" circumstances for the San Francisco area. The projections did well when measured against real forecasts from the US Weather Bureau. Hu's contributions established the foundation for developing complex and accurate rainfall techniques.

An extensive analysis of data-driven rainfall prediction was conducted using a neural network model to forecast two types of rainfall for the following hour [9]. The effectiveness of a neural network model in combination with linear regressions to estimate and replicate missing rainfall data has also been assessed [10]. Additionally, a decision tree (DT) and the Classification and Regression Tree (CART) algorithm have been used to predict rainfall using data gathered between 2002 and 2005 [11]. Underlying patterns in a sizable meteorological dataset were discovered using a clustering algorithm with K-NN, achieving high accuracy in rainfall, temperature, and humidity predictions [12].

Data mining and data preprocessing are the two major components of any prediction model. Extracting the relevant data for the model and then pre-processing it precisely for the model. The Spatio-Temporal Data Mining (STDM) technique was proposed in the study [13], and they predicted extreme weather events by analyzing various ML techniques after preprocessing the mined data using Z-score normalization. They also employed the anomaly frequency method (AFM) technique to extract the features of extreme weather events only. They have validated their findings using K-means and DBSCAN, both clustering algorithms with extreme weather events like Cyclone BOB, and Cyclone Thane Cyclone Vardah.

In the past, researchers have also explored different machine-learning techniques that can be helpful in weather forecasting and predicting extreme weather events. The study [14] discuss the challenges of predicting extreme rainfall events through a data-driven model employing data mining techniques. This study also encompasses various machine-learning techniques for forecasting extreme precipitation events. The model in question is built upon LSTM neural networks, drawing from data related to 8 atmospheric variables and 11 surface variables across 12 atmospheric pressure levels. The methodology amalgamates reanalysis data spanning 12 isobaric pressure levels, surface data, and information from meteorological stations in Brazil's southeastern region. The key takeaway from this methodology is that it exhibits potential in forecasting extreme rainfall volume, using multivariate time series data, presenting a fresh approach to predicting extreme rainfall events.

Exploring the promising potential of machine learning in weather forecasting, freely available meteorological data from various online sources has been utilized in research, highlighting the limitations of traditional meteorological models in terms of accuracy and adaptability [15]. The study emphasizes the pivotal role of ML in overcoming these challenges and introduces a specialized Python API tailored for meteorological data, facilitating the training and evaluation of neural network models. The study also emphasizes the advantages of neural networks over traditional models while shedding light on their applications in industries like hydropower and flood management. Singular-spectrum-analysis (SSA) has been integrated with supervised learning models, including least-squares support vector regression (LS-SVR) and Random Forest (RF), for rainfall prediction [16]. They have predicted monthly rainfall for Nellore region of India by utilizing weather parameters such as minimum and maximum temperature, cloud cover and relative humidity. They concluded that their model was better over MLR and ANN models. Bayesian Linear Regression, Boosted Decision Tree Regression, Decision Forest Regression, and Neural Network Regression have been explored for predicting rainfall in Terengganu, Malaysia. The models utilized data from 10 surrounding stations and were evaluated for different lead times. The results indicated that Boosted Decision Tree Regression and Decision Forest Regression performed better than the other models [17]. Random Forest regression, Support Vector regression, Neural Network regression, Least Absolute Shrinkage and Selection Operator (LASSO) regression, Gradient boosting regression, and Extreme Gradient boosting regression have been employed for short-term rainfall forecasting in the Lake Victoria Basin in Uganda. The study concluded that the Extreme Gradient boosting regression model performed the best in this region [18]. Machine learning techniques have been employed to classify rainfall into rainy

and no-rain days [19]. Various models, including logistic regression, K-NN, DT, and simple deep learning models, have been utilized to predict daily rainfall using multiple weather parameters. The results indicate that logistic regression and the simple deep learning model outperformed other models [20]. For short-term forecasting, feed-forward neural networks have been applied to nowcast rainfall using a classification approach every 10 minutes for lead times ranging from 30 minutes to 6 hours [21].

The literature review revealed that though various ML techniques have been employed by the researchers, because of the chaotic nature of weather, the rainfall pattern changes from place to place, making it difficult to predict. Hence, this study utilizes different ML techniques for a day-ahead rainfall prediction.

3. METHODOLOGY

3.1 Collection of data

The rainfall data used here is made freely available by IMD, Pune [22]. This is gridded daily rainfall data with a high spatial resolution of 1.0° latitude by 1.0° longitude. The dataset over India is available from the year 1901 to 2023. The data is available as a NetCDF file for every year. The data was downloaded for the years 2015 to 2023. The rainfall data provided by IMD, Pune, has three dimensions: Longitude, Latitude, and Time. The data is arranged in 35 × 33 grid points, i.e., longitudes and latitudes. The first record in the data is at 6.5 N and 66.5 E, and the last record in the data is at 38.5 N and 100.5E. The unit for the rainfall data is in millimeters (mm), and the time is 1st January to 31st December for each year. The data was downloaded for 9 years, starting from 2015 to 2023.

3.2 Data visualization and preprocessing

Panoply software, an open-source software for climate data visualization provided by NASA, is used to visualize rainfall data. Panoply helps extract data from NetCDF files in CSV format. The software facilitates visualization on a world map in various types of maps. The data was visualized by creating a georeferenced latitude-longitude plot with color contour. It was adjusted to the Equirectangular Regional projection to closely observe the Indian region. The sample view of rainfall on 27th July 2023 is shown in Figure 1.

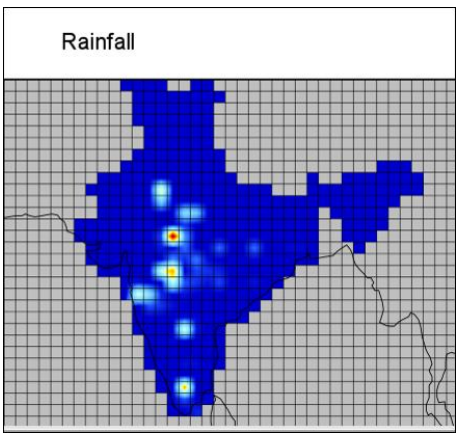


Figure 1. Visualization of rainfall on 27th July 2023 using Panoply

For rainfall prediction, data acquisition and preprocessing is highly important. Rainfall data can be very difficult to preprocess for ML models as, for a major part of the year, there is none to very minimal rainfall in India. Major rainfall occurs during the monsoon season, which usually starts in June and lasts until September. Therefore, the NaN values needed to be addressed properly so that they didn't negatively affect the models and hinder the prediction process. Another important aspect with the data was multidimensionality of the data.

After extracting the data from NetCDF files to CSV files using Panoply, Python libraries combined the CSV data files of all the years into a single data frame. Data from the years 2015 to 2023 were used. The data from all files was combined to form a single file with daily rainfall data over the above-mentioned latitude and longitude. This single data frame was then further preprocessed. For the latitudes and longitudes considered for this study, the NaN values were first identified, and number of such values was counted. The percentage of NaN values was found to be as negligible as 0.56% and hence they were changed to zero in the data frame. NaN values are not assumed or interpolated as it would have affected the training [19]. The sample data frame after preprocessing is shown in Table 1.

Table 1. Sample data frame after preprocessing

Latitude	Longitude	Time	Rainfall
15.5	72.5	2016-05-26	1.203877
		2016-05-29	1.547464
		2016-06-02	39.823242
		2016-06-03	2.956533
		2016-06-04	2.478646
		2016-06-05	19.569462
		2016-06-07	55.759205
		2016-06-08	7.372731
		2016-06-09	48.870586
		2016-06-10	35.458523

The data was further processed to remove the records of latitudes and longitudes other than the desired latitude and longitude. Two regions were considered for the study. The first was rainfall over 18.5 N and 73.5 E, and the second region was 15.5 N and 73.5 E. These regions are from Maharashtra and Gujarat states of India, respectively. Figure 2 shows the regions on the map.

Figure 2. Regions for rainfall prediction

After the preprocessing of data, various ML models were employed to predict rainfall of the next day. The methodology adopted for this study is as presented in Figure 3.

Figure 3. Methodology

The prediction models are described briefly in the following subsections. The performance of all the models deployed in the study was compared using various parameters such as mean absolute error, root mean squared error, correlation coefficient, and the maximum rainfall prediction capability of the model.

3.3 MLR model

Linear Regression is the most basic algorithm that provides a straightforward approach to modeling relationships between variables. The simplicity of linear regression provides us with the baseline model for predicting rainfall based on linear relationships.

Regression is a method used to examine the relationship between an independent variable or variables (X) and a dependent variable (y). The dependent variable is also called the response variable. Here, time series forecasting is done where an earlier day's rainfall is considered an independent variable to predict the next day's rainfall. The number of independent variables, if increased, is called MLP. Mathematically, it can be defined as shown in Eq. (1).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \quad (1)$$

where,

Y is the dependent variable (rainfall).

β_0, β_1, β_2 are the coefficients to be estimated.

X_1, X_2, X_3 are the values of the independent variables (previous 3 days' rainfall).

ε represents the error term.

The outcome, Y, represents the expected quantity of rainfall based on the linear relationship between the time variables provided. The goal is to reduce the residual sum of squares (RSS), which represents the sum of the squared variations between the rainfall's actual values and those predicted by the MLR model.

Mathematically, it is as given in Eq. (2) [23].

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \quad (2)$$

The MLR model looks for a linear equation that best fits the data to make predictions. Thus, the model's goal is to forecast the amount of rainfall likely to happen at a specific site the next day.

3.4 DT

The DT is a non-parametric supervised learning technique used in regression and classification applications. Its internal nodes, leaf nodes, branches, and root nodes make up its hierarchical tree structure. This arrangement enables a sequential decision-making process.

An important component of the DT algorithm is entropy. It is a measure used at each node in the tree (t_i) to decide how the data should be divided. It is used to quantify disorders or impurities in a dataset. Entropy quantifies the degree of uncertainty or unpredictability surrounding the target variable. The decision tree uses entropy analysis to produce splits that maximize information gain and, as a result, produce classification or regression results that are more exact and accurate.

In a regression problem, node t_i gives a prediction of the value of the dependent variable equal to the average value of this variable y for the elements of the training set assigned to t_i [24].

Mathematically, it is defined as given in Eq. (3).

$$\overline{y_i} = \frac{1}{N_i} \sum_{n=1}^{N_{train}} \mu_i(x_n) y_n \quad (3)$$

where,

$\mu_i(x)$ is a membership function by which node t_i is characterized.

N_i is the number of training samples

The tree is built using a divide-and-conquer approach, where the attribute space is divided into non-overlapping regions through a series of Boolean tests arranged in a hierarchy. Each region simplifies the decision problem, and each test in this hierarchy represents an internal node in the decision tree [24].

DT Regression aims to construct a tree structure for accurate rainfall forecasts using the previous day's rainfall. A parameter called max depth determines how complex the tree is. While deeper trees may prevent overfitting, they do catch finer rainfall distinctions. A suitable max depth is selected by calculating the Mean Squared Error (MSE) at various depths. To provide precise forecasts for rainfall based on geographic coordinates, this balance enables the model to identify important trends without overfitting. The depth of DT used in the model was 3 where minimum MSE was obtained. DT also helps us understand and visualize the decision-making process more easily, which helps us gain valuable insights into the contributing factors to rainfall conditions.

3.5 K-NN

K-NN is another supervised ML technique that is nonparametric [25]. K-NNs are also helpful in dealing with the complexities of rainfall patterns as this non-parametric algorithm doesn't make underlying assumptions about the data distribution. K-NN uses Euclidean, Manhattan, and Minkowski distance measurements to classify data. Studies reveal that when K-NN is used on datasets with a smaller feature set, it performs better. The variables x_{ij} and x_{io} are used to calculate the Euclidean distance. x_{ij} stands for the i^{th} data point in the j^{th} predictor, and x_{io} for the i^{th} predictand (Eq. (4)).

Mathematically,

$$d_j = \sum_{i=1}^{n} (x_{ij} - x_{io})^2 \quad (4)$$

$$Z_r = \sum_{k=1}^{K} f_k(d_j \times Z_k) \quad (5)$$

The idea of Euclidean distance is presented in these formulas, which measure the difference in similarity between data points. In addition, Z_r and Z_k represent the expected and neighboring data, respectively, while $f_k(d_j)$ is the kernel function, a key element of the K-NN method. These mathematical formulas are incorporated to highlight the complexity of K-NN (Eq. (5)), a flexible method that is frequently used for data classification. For regression using K-NN, prediction is the average of the K-NN outcome [26].

Every data point's location is determined by its previous day's rainfall amount, and the recorded precipitation is indicated by the related rainfall. By iterating through various neighbor counts (k), ranging from 1 to 20, the performance of the K-NN regression model is assessed. The model is trained and tested, and the MSE is calculated for each k. The MSE results are graphically displayed to help choose the best k value. The K-value used for the model was 9. This helps to shed light on the trade-off between bias and variance for accurate rainfall predictions.

3.6 RF

RF is a popular ensemble learning method that works well for regression and classification problems. To reduce overfitting and improve prediction accuracy, it combines several decision trees. Two important variables contribute to RF's "random" nature, which makes it unique. First, it uses random feature selection for every tree in the ensemble. Let's say there are M features in total. For every tree, a random subset of m features is chosen; m is usually significantly fewer than M. Only these m features are considered when building each tree in terms of node splits.

Secondly, RF introduces randomness by using random subsampling of the training data. Each tree in the ensemble is trained on a different subset of data points. Thirdly, the final prediction in an RF ensemble is calculated as the average of the predictions from all individual trees as given in Eq. (6).

$$\text{Final Prediction} = \frac{1}{N} \sum_{i=1}^{N} \text{Tree}_i \quad (6)$$

where,

Final Prediction: Represents the overall prediction made by the RF.

N: The total number of decision trees in the RF.

Tree_i: The prediction made by the i^{th} decision tree in the ensemble.

RF generates a more accurate and dependable overall prediction by averaging the guesses. Multiple trees' predictions work together to reduce individual errors and variations, creating a regression model that is both reliable and strong. An RF Regressor model is applied to forecast rainfall based on temporal data. The dataset is divided into training and testing sets for model evaluation. The initialization, training, and use of a RF Regressor to forecast rainfall values on the test set. Finally, the model's performance is assessed using various evaluation parameters.

3.7 Performance evaluation

In the last stage, the performance of all the models was evaluated on MAE, RMSE, R-value and the maximum rainfall prediction capability of the model for the test set. The formula for calculating MAE is given in Eq. (7) [27].

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |e_i| \quad (7)$$

where, *n* is the number of samples, *e* is the error calculated as the difference between observed and predicted values. MAE is an appropriate measure of performance evaluation over RMSE as RMSE increases with increased variance in error [27] The formula to calculate RMSE is given in Eq. (8).

$$RMSE = \left[\frac{1}{n} \sum_{i=1}^{n} |e_i|^2 \right]^{1/2} \quad (8)$$

Microsoft Excel was utilized to compute R-value. The formula for the correlation coefficient is given in Eq. (9).

$$R - Value = \frac{\sum (x - \overline{x})(y - \overline{y})}{\sqrt{\sum (x - \overline{x}^2)} \sqrt{\sum (y - \overline{y}^2)}} \quad (9)$$

where, *x* and y are two arrays of samples and $\overline{x}$ and $\overline{y}$ are the mean of all values of x and y, respectively. In each example, the maximum rainfall capacity of the network was estimated independently for the testing datasets. This indicates the observed maximum rainfall during the testing period and the maximum predicted rainfall during the same period. This parameter assisted in evaluating the network's ability to predict excessive rainfall.

4. RESULTS AND DISCUSSION

The preprocessed data was used to train four ML models. The data was organized from 1st January to 31st December for each year. Two separate files were created for the two regions, namely, Region 1 (Pune Region, Maharashtra State, India) and Region 2 (Panchmahal Region, Gujarat State, India). After testing for various input combinations, the number of previous days for predicting the next day's rainfall were decided to be five for each model. For each model, 80% data was used for training and 20% was used for testing. The ratio of training and testing can range from 50:50 to 90:10. For the dataset of the size in thousands, 80:20 ratio has been selected so as to provide enough samples for training purposes and sufficient samples to test the model's performance [28]. The total records available in the dataset were 3287 with no missing data for the two chosen regions. The test dataset in each case had 658 records. The training, testing and evaluation of these ML models were done using Python and its libraries. For each model, Actual Vs Predicted Rainfall was plotted. The details of the maximum rainfall, no-rain days, average rainfall during the years 2015 to 2023 are shown in Table 2.

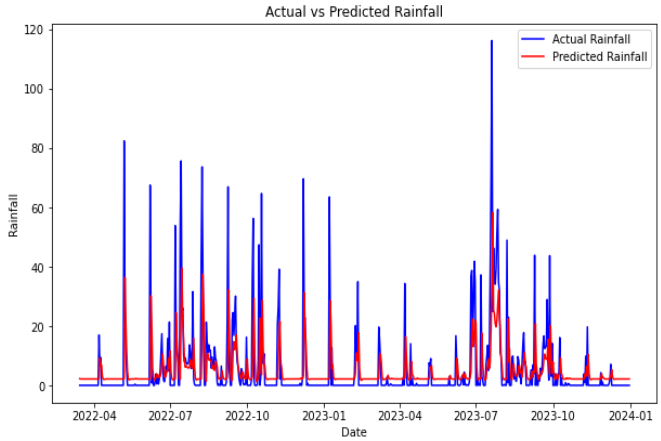
Table 2. Details of the rainfall at Pune and Panchmahal region

Region	No. of Rainy Days	No. of No-Rain Days
Region 1-Pune	918	2369
Region 2-Panchmahal	455	2832

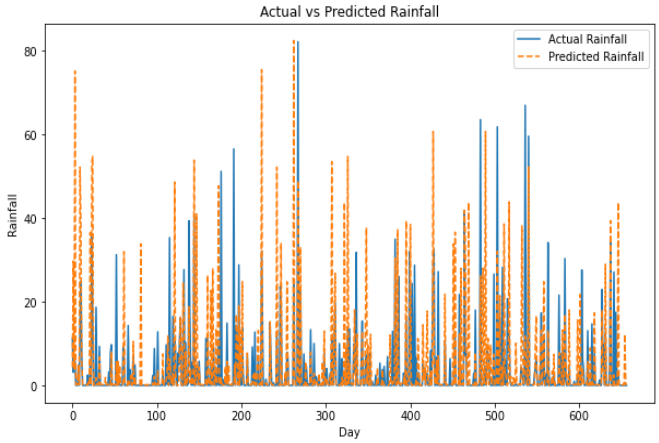
The number of rainy and no rainy days indicates that both the regions have different rainfall patterns where rainy days for Pune region are almost double than that of Panchmahal Region. The no-rain days considered here are the days where rainfall is less than 2.5 mm [29].

Figure 4(a)-(d) shows the plots of predicted Vs Actual rainfall for Maharashtra's Pune Region-Region 1.

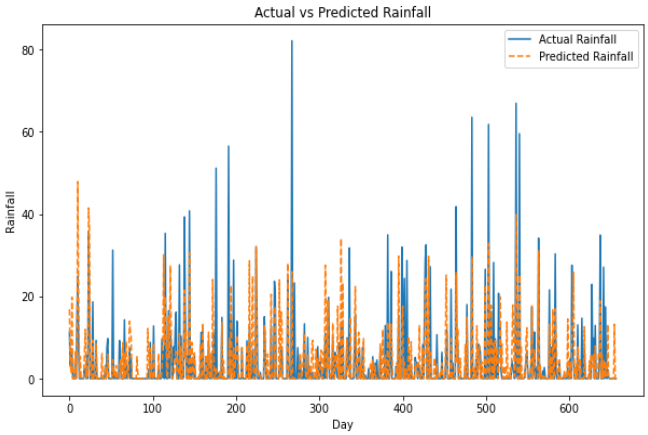
The figures show how the models performed. It can be observed that predicted rainfall by the MLR model closely followed the actual rainfall. However, it could not anticipate the increase in rainfall. As shown in Table 3, MAE obtained from MLR is the highest among all the models under study. The R-value is also less as compared to the other models. As the dataset consisted of many non-rainy days, the time series became highly nonlinear and hence, MLR could not perform well. Secondly, the number of inputs for each model were kept the same. Further experimentation with MLR parameters might provide better results. The results obtained from DT model show that it could anticipate heavy rain days better than the other models, however, the correlation values were less than that obtained by K-NN and RF models. If MAE, RMSE and maximum rainfall prediction capability of the model is compared with other models, DT was found to be better than MLR, however, K-NN outperformed in terms of MAE, RMSE and R-value. The results of K-NN and RF were found to be comparable, where K-NN was found better in terms of error and correlation and RF was found better in terms of maximum rainfall prediction capability of the model. The MAE, RMSE, R-value and Maximum Rainfall Capability for Region 1 is shown in Table 3.



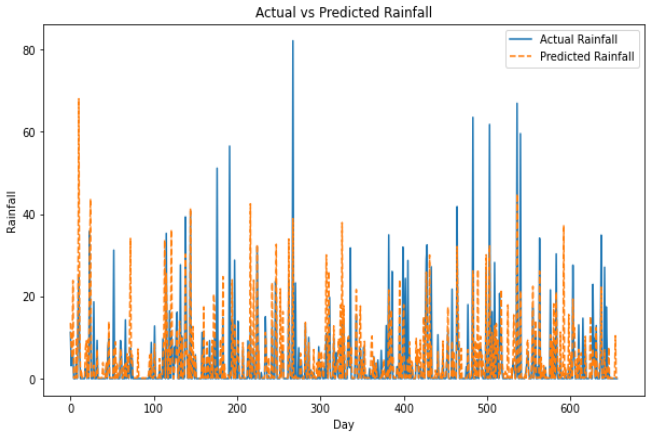
(a) Actual vs predicted values by MLR model for the test dataset



(b) Actual vs predicted values by DT model for the test dataset



(c) Actual vs predicted values by K-NN model for the test dataset

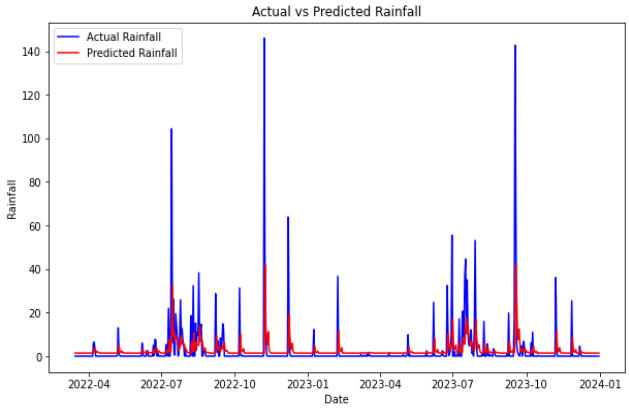


(d) Actual vs predicted values by RF model for the test dataset

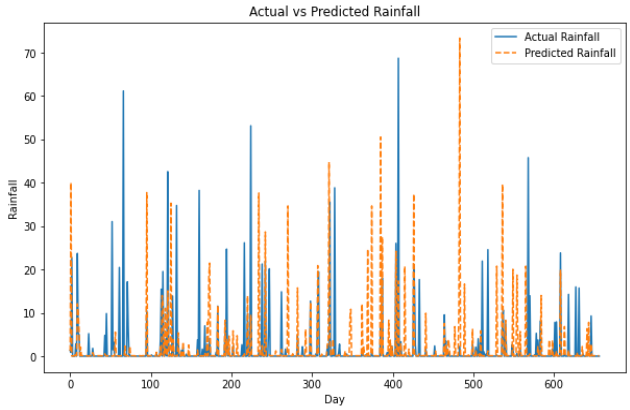
Figure 4. Rainfall prediction plots for Region 1

Table 3. Results obtained for the Region 1

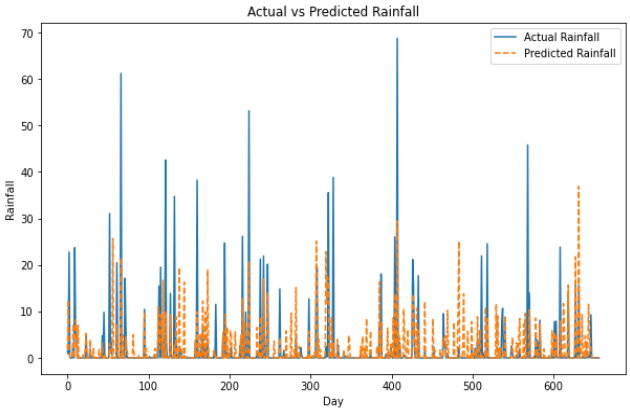
Model	MAE (in mm)	RMSE (in mm)	R-Value	Actual Maximum Rainfall —Test Set (in mm)	Predicted Maximum Rainfall —Test Set (in mm)
MLR	5.64	11.50	0.48	82.15	58.34
DT	4.82	11.11	0.49	82.15	82.4
K-NN	3.33	7.41	0.63	82.15	47.91
RF	3.50	7.70	0.59	82.15	68.02



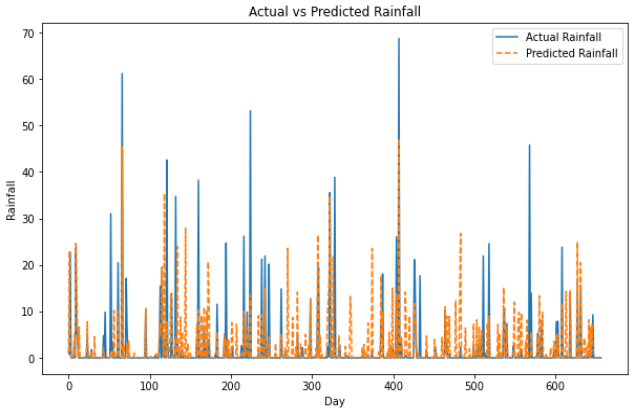
(a) Actual vs predicted values by MLR model for the test dataset



(b)Actual vs predicted values by DT model for the test dataset



(c) Actual vs predicted values by K-NN model for the test dataset



(d) Actual vs predicted values by RF model for the test dataset

Figure 5. Rainfall prediction plots for Region 2

Table 4. Results obtained for the Region 2

Model	MAE (in mm)	RMSE (in mm)	R-Value	Actual Maximum Rainfall —Test Set (in mm)	Predicted Maximum Rainfall —Test Set (in mm)
MLR	10.15	20.72	0.42	68.73	42.03
DT	8.05	20.68	0.22	68.73	73.34
K-NN	7.12	15.22	0.60	68.73	36.88
RF	6.66	15.26	0.60	68.73	47.22

Table 5. Assessing models' reliability

Parameter	Actual Rainfall	MLR	DT	K-NN	RF
Region 1					
Mean	4.15	4.99	5.12	4.11	4.52
Standard Deviation	9.58	6.28	11.99	7.27	8.22
Error margin	0.73	0.48	0.92	0.56	0.63
Region 2					
Mean	1.97	2.64	2.38	1.99	1.93
Standard Deviation	6.93	3.61	5.49	4.50	6.68
Error margin	0.53	0.28	0.42	0.34	0.51

The results on the test dataset obtained for the Gujarat region, Region 2 are shown in figures from Figure 5(a)-(d). From the results, it can be observed that RF and K-NN models were able to predict the rainfall better over the other two models. The DT model could predict the maximum rainfall amount beyond the actual rainfall value of the training set. However, the prediction of this value was not correct, and it affected R-value.

The results obtained in terms of errors are as given in Table 4. These results indicate that RF model is better than K-NN in terms of MAE and K-NN performed better than RF in terms of RMSE. The R-value for both the models was the same for the test dataset. Considering the maximum rainfall prediction capability of the models under study, RF model outperformed other models. The MAE and RMSE values obtained by MLR and DT models were higher than RF and K-NN. A very less correlation was observed in the case of DT between predicted and actual rainfall values.

It must be noted that, for fair comparison between the models, the input was kept the same, i.e., the previous five

days were used to predict the next day's rainfall. The fine tuning of the model in each case might help to improve the results. If the performance of all the models is compared for both the regions, it can be seen that, MAE, RMSE are quite less for Region 1 than Region 2. The correlation coefficient was also obtained better for Region 1 than Region 2.

This might be because of a greater number of no-rain days present in the case of Region 2 causing high rainfall variability. Hence, a same model cannot be employed for every region.

To assess the reliability of predicted values obtained by different models, we calculated the 95% confidence interval (CI) for each predicted value. We focused on the mean, standard deviation and error margin of the actual rainfall values and compared it with the predicted values. The results are shown in Table 5.

It can be observed that the mean and standard deviation obtained from predictions by K-NN and RF models for Region 1 and Region 2 are close to the actual rainfall values. Hence, we can say that these model predictions are within the 95% CI.

5. CONCLUSION

For predicting a day-ahead rainfall, two regions from Indian states were studied. The ML models, MLR, K-NN, DT and RF were employed for the same. It was observed that for the Region 1 (Pune Region, Maharashtra State), K-NN performed better over the other models, whereas, for the Region 2 (Panchmahal Region, Gujarat State), RF performed better over the other models. Overall, K-NN and RF models have proven to be the most effective among all the four ML models with the lowest RMSE and MAE, better correlation coefficient. The results were obtained when the five previous days were used to predict the next day's rainfall. In future, the performance of these models can be further evaluated for 2 to 7 days lead time. This would need the parameter tuning for obtaining better results. By including other weather parameters and understanding the complex patterns that cause heavy rainfall, this approach might help to identify heavy rainfall days and issue appropriate warnings beforehand.

REFERENCES

[1] Fathi, M., Haghi Kashani, M., Jameii, S.M., Mahdipour, E. (2022). Big data analytics in weather forecasting: A systematic review. Archives of Computational Methods in Engineering, 29(2): 1247-1275. https://doi.org/10.1007/s11831-021-09616-4

[2] Maqsood, I., Khan, M.R., Abraham, A. (2004). An ensemble of neural networks for weather forecasting. Neural Computing & Applications, 13: 112-122. https://doi.org/10.1007/s00521-004-0413-4

[3] Smagorjnsky, J. (1983). The beginnings of numerical weather prediction and general circulation modeling: Early recollections. Advances in Geophysics, 25: 3-37. https://doi.org/10.1016/S0065-2687(08)60170-3

[4] Brotzge, J.A., Berchoff, D., Carlis, D.L., Carr, F.H., et al. (2023). Challenges and opportunities in numerical weather prediction. Bulletin of the American Meteorological Society, 104(3): E698-E705. https://doi.org/10.1175/BAMS-D-22-0172.1

[5] Mohammed, M., Kolapalli, R., Golla, N., Maturi, S. S. (2020). Prediction of rainfall using machine learning techniques. International Journal of Scientific and Technology Research, 9(1): 3236-3240.

[6] Subrahmanyam, K.V., Ramsenthil, C., Girach Imran, A., Chakravorty, A., et al. (2021). Prediction of heavy rainfall days over a peninsular Indian station using the machine learning algorithms. Journal of Earth System Science, 130: 240.

[7] Fang, W., Xue, Q., Shen, L., Sheng, V.S. (2021). Survey on the application of deep learning in extreme weather prediction. Atmosphere, 12(6): 661. https://doi.org/10.3390/atmos12060661

[8] Hu, M.J.C., Root, H.E. (1964). An adaptive data processing system for weather forecasting. Journal of Applied Meteorology, 3(5): 513-523.

[9] French, M.N., Krajewski, W.F., Cuykendall, R.R. (1992). Rainfall forecasting in space and time using a neural network. Journal of Hydrology, 137(1-4): 1-31. https://doi.org/10.1016/0022-1694(92)90046-X

[10] Michaelides, S.C., Tymvios, F.S., Michaelidou, T. (2009). Spatial and temporal characteristics of the annual rainfall frequency distribution in Cyprus. Atmospheric Research, 94(4): 606-615. https://doi.org/10.1016/j.atmosres.2009.04.008

[11] Petre, E.G. (2009). A decision tree for weather prediction. Buletinul UniversităŃii Petrol – Gaze din Ploieşti, 61(1): 77-82.

[12] Sharif, M., Burn, D.H. (2006). Simulating climate change scenarios using an improved K-nearest neighbor model. Journal of Hydrology, 325(1-4): 179-196. https://doi.org/10.1016/j.jhydrol.2005.10.015

[13] KanimozhiSelvi, D., Sowmiya, G. (2019). Prediction of extreme weather events using machine learning technique. International Journal of Applied Engineering Research, 14(4): 924-929.

[14] de Sousa Araújo, A., Silva, A.R., Zárate, L.E. (2022). Extreme precipitation prediction based on neural network model–A case study for southeastern Brazil. Journal of Hydrology, 606: 127454. https://doi.org/10.1016/j.jhydrol.2022.127454

[15] Patkar, U., Maske, S., Ahmad, S., Mengade, R., Sadawarti, G. (2021). Machine learning for weather forecasting using freely available weather data in Python. GIS Science Journal, 8(12): 886-898.

[16] Reddy, P.C.S., Yadala, S., Goddumarri, S.N. (2022). Development of rainfall forecasting model using machine learning with singular spectrum analysis. IIUM Engineering Journal, 23(1): 172-186. https://doi.org/10.31436/IIUMEJ.V23I1.1822

[17] Ridwan, W.M., Sapitang, M., Aziz, A., Kushiar, K.F., Ahmed, A.N., El-Shafie, A. (2021). Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia. Ain Shams Engineering Journal, 12(2): 1651-1663. https://doi.org/10.1016/j.asej.2020.09.011

[18] Gahwera, T.A., Eyobu, O.S., Isaac, M. (2024). Analysis of machine learning algorithms for prediction of short-term rainfall amounts using Uganda's lake Victoria basin weather dataset. IEEE Access, 12: 63361-63380. https://doi.org/10.1109/ACCESS.2024.3396695

[19] Hudnurkar, S., Rayavarapu, N. (2022). Binary classification of rainfall time-series using machine learning algorithms. International Journal of Electrical & Computer Engineering, 12(2): 1945-1954.

https://doi.org/10.11591/ijece.v12i2.pp1945-1954
[20] Raval, M., Sivashanmugam, P., Pham, V., Gohel, H., Kaushik, A., Wan, Y. (2021). Automated predictive analytics tool for rainfall forecasting. Scientific Reports, 11(1): 17704. https://doi.org/10.1038/s41598-021-95735-8
[21] Pirone, D., Cimorelli, L., Del Giudice, G., Pianese, D. (2023). Short-term rainfall forecasting using cumulative precipitation fields from station data: A probabilistic machine learning approach. Journal of Hydrology, 617: 128949. https://doi.org/10.1016/j.jhydrol.2022.128949
[22] Rajeevan, M., Bhate, J., Jaswal, A.K. (2008). Analysis of variability and trends of extreme rainfall events over India using 104 years of gridded daily rainfall data. Geophysical Research Letters, 35(18): L18707. https://doi.org/10.1029/2008GL035143
[23] Weisberg, S. (2005). Applied Linear Regression. John Wiley & Sons.
[24] Suárez, A., Lutsko, J.F. (1999). Globally optimal fuzzy decision trees for classification and regression. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(12): 1297-1311. https://doi.org/10.1109/34.817409
[25] Souza, D.V., Nievola, J.C., Dalla Corte, A.P., Sanquetta, C.R. (2020). K-nearest neighbor and linear regression in the prediction of the artificial form factor. Floresta, 50(3): 1669-1678. https://doi.org/10.5380/rf.v50 i3. 65720
[26] Imandoust, S.B., Bolandraftar, M. (2013). Application of k-nearest neighbor (KNN) approach for predicting economic events: Theoretical background. International Journal of Engineering Research and Applications, 3(5): 605-610.
[27] Willmott, C.J., Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30(1): 79-82. https://doi.org/10.3354/cr030079
[28] Muraina, I. (2022). Ideal dataset splitting ratios in machine learning algorithms: General concerns for data scientists and data analysts. In 7th International Mardin Artuklu Scientific Research Conference, Mardin, Turkey, pp. 496-504.
[29] Bhatla, R., Verma, S., Pandey, R., Tripathi, A. (2019). Evolution of extreme rainfall events over Indo-Gangetic plain in changing climate during 1901–2010. Journal of Earth System Science, 128, 120. https://doi.org/10.1007/s12040-019-1162-1