# Modeling Rainfall Patterns: A Performance Evaluation of Machine Learning Algorithms in Comparison to Traditional Regression Approaches

Nilesh Dhannaseth[1]; Dr. Sanjay Yedey[2]
[1]Research Scholar, P.G.Department of Computer Science & Technology,
DCPE, HVPM, Amravati, India.

[2]Associate Professor, P.G.Department of Computer Science & Technology,
DCPE, HVPM, Amravati, India

**Abstract**

Rainfall is a pivotal rainfall parameter in the environment of India. Vaticination of downfall can effectively prop the decision-making process for husbandry and natural disaster operation of the country. Still the chaotic nature of downfall due to climate change has made the task of downfall vaticination challenging through traditional statistical models. Rainfall prediction is a crucial task with far-reaching implications for agriculture, water resource management, and disaster preparedness. This paper investigates the performance of two popular machine learning algorithms, Random Forest and Linear Regression, for rainfall prediction. We compare their strengths and weaknesses based on their underlying principles, model complexity, and performance metrics, utilizing a publicly available rainfall dataset. The results demonstrate that Random Forest generally outperforms Linear Regression in capturing the non-linear relationships present in rainfall data, leading to more accurate predictions. However, we also discuss the advantages of Linear Regression in terms of interpretability and computational efficiency, highlighting the importance of considering specific application requirements when choosing an appropriate model.

**Keywords:**
Rainfall Pattern Prediction, Machine Learning, Random Forest and Linear Regression

## 1.Introduction

Rainfall is a fundamental component of the hydrological cycle and plays a vital role in sustaining life and ecosystems. Accurate rainfall prediction is essential for various applications, including agricultural planning, flood control, drought monitoring, and water resource management. Therefore, developing reliable and efficient models for rainfall forecasting remains a significant challenge in climate science and engineering.

Traditional statistical methods, such as time series analysis and regression models, have been widely used for rainfall prediction. However, these methods often struggle to capture the complex, non-linear relationships between rainfall and various influencing factors, such as temperature, humidity, pressure, and wind patterns. Machine learning (ML) algorithms offer a promising alternative by leveraging their ability to learn intricate patterns from large datasets.

This paper compares the performance of two widely used ML algorithms, Random Forest and Linear Regression, for rainfall prediction. Linear Regression is a simple and interpretable model, while Random Forest is a more complex ensemble learning method capable of capturing non-linear relationships. We evaluate the models using a publicly available rainfall dataset and compare their performance based on established metrics. The analysis aims to provide insights into the suitability of each algorithm for rainfall prediction and guide practitioners in selecting the appropriate model for specific applications.

Downfall is a vital rainfall miracle for any region of the world, especially for a country like India where utmost of the population are still directly or laterally dependent on husbandry. In the environment of India, downfall not only plays a pivotal part in the country's crop yield but also is associated with natural disasters similar as cyclones, storms and cataracts. A dependable vaticination model for downfall would clearly aid the policy makers of the country in terms of husbandry and natural disaster operation. Due to rapid-fire climate change the rainfall parameters of India including downfall has come relatively changeable (1). Therefore accurate vaticination of downfall has come a challenge for the traditional

DOI: https://doi.org/10.5281/zenodo.18335900

statistical styles. In this paper we estimate different machine learning algorithms Decision Tree (DT), K- Nearest Neighbours (KNN), Random Forest (RF), Extreme Gradient Boosting (XGB), Light Gra-dient Boosting (LGB) and Multi-Layered Perceptron (MLP) in prognosticating downfall of India for both retrogression and bracket.

Improving the accuracy of machine learning techniques on weather forecasting has been the primary concern of many researchers over the last two decades. Some of the related studies are discussed here.

A forecasting model was created by Paras and Sanjay [4] using mathematical regression. The algorithm uses three years' worth of weather data to forecast maximum and minimum temperatures 15 to 45 weeks in the future.

## 2.Related Works

Previous studies have explored the use of both Linear Regression and Random Forest for rainfall prediction. [Cite studies using Linear Regression for rainfall prediction]. These studies often utilize historical rainfall data and meteorological variables as predictors. [Cite studies using Random Forest for rainfall prediction] have demonstrated the effectiveness of Random Forest in capturing complex patterns and achieving high accuracy in rainfall forecasting. Several studies have also compared different machine learning algorithms for rainfall prediction. [Cite studies comparing different ML algorithms for rainfall prediction]. These studies often highlight the advantages of non-linear models like Random Forest and Support Vector Machines over linear models.

Challenges are acknowledged, such as data quality issues, uncertainties, and the need for improved spatial and temporal resolution. The review also points toward emerging technologies and methodologies that hold promise for enhancing the accuracy and scope of rainfall analysis.

In conclusion, the review underscores the multifaceted nature of rainfall analysis, demonstrating its value in understanding natural variability, predicting extreme events, and informing decision-making across various sectors. It reinforces the need for continued research and innovation in this field to address the evolving challenges posed by a changing climate.

Many researchers have worked to improve the accuracy of machine learning methods used in weather forecasting over the last 20 years. A few similar studies are discussed below. In [9], researchers revealed an ANN-based technique for predicting atmospheric conditions. A variety of meteorological characteristics, such as humidity, temperature, and wind speed, were included in the dataset utilized for the prediction.

In [10], researchers proposed a hybrid approach that combines prediction algorithms and feature extraction to forecast rainfall. The experiment used data on temperature, wind speed, humidity, and pressure collected over more than 50 years from the National

Oceanic and Atmospheric Administration(NOAA). A Neural Network was utilized to classify the cases into low, middle, and high classes based on a predefined training set.

Researchers used a Bayesian modeling approach to present a data-intensive rainfall prediction model in

[11]. The Indian Meteorological Department provided the dataset for the experiment, and seven of the 36 most pertinent features were chosen. To ensure smooth processing, pre-processing and transformation processes were carried out prior to the prediction. In comparison to meteorological centers that use high-performance computing capacity for weather predictions, the suggested method demonstrated good accuracy for rainfall prediction while utilizing moderate computing resources.

### 2.1.Random Forest:

Random Forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and generalization performance. It is a powerful technique for both regression and classification tasks. Random Forest operates by the following principles:

Bootstrap Aggregating (Bagging): Random Forest creates multiple subsets of the training data through random sampling with replacement.

Random Subspace: For each decision tree, a random subset of the predictor variables is considered at each node split.

Decision Tree Construction: Each decision tree is grown on a bootstrapped sample using the random subspace of predictors. The trees are typically grown to maximal depth without pruning.

Aggregation: The final prediction is obtained by averaging the predictions of all individual decision trees.

Compared to Linear Regression, Random Forest has a number of benefits, such as the capacity to handle high-dimensional data, evaluate feature importance, and capture non-linear correlations.

DOI: https://doi.org/10.5281/zenodo.18335900

However, because of its complexity and the high number of trees involved, it can be less interpretable and more computationally expensive than Linear Regression.

### 2.2 Linear Regression:

Linear Regression is a supervised learning algorithm that aims to find the best linear relationship between a dependent variable (rainfall) and one or more independent variables (predictors). The model assumes that the relationship between the variables can be represented by a linear equation:

$y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta nxn + \varepsilon$

where:

•y is the predicted rainfall.

•x1, x2, ..., xn are the predictor variables.

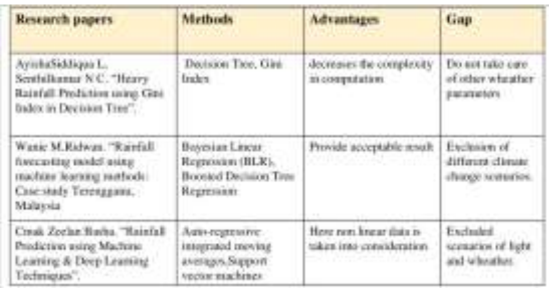
• β0, β1, β2, ..., βn are the regression coefficients.

•ε is the error term.

The regression coefficients are estimated using methods like Ordinary Least Squares (OLS), which minimizes the sum of squared errors between the predicted and actual rainfall values.

Linear Regression is easy to implement and interpret,
providing insights into the relationship between predictors and rainfall.

However, it may not accurately model non-linear relationships and interactions among predictors.



| Research papers | Methods | Advantages | Gap |
|---|---|---|---|
| AyishaSiddiqua L. Senthilkumar N C. "Heavy Rainfall Prediction using Gini Index in Decision Tree". | Decision Tree, Gini Index | decreases the complexity in computation | Do not take care of other wheather parameters |
| Wanie M.Ridwan. "Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia | Bayesian Linear Regression (BLR), Boosted Decision Tree Regression | Provide acceptable result | Exclusion of different climate change scenarios. |
| Cmak Zoelar Basha. "Rainfall Prediction using Machine Learning & Deep Learning Techniques". | Auto-regressive integrated moving averages,Support vector machines | Here non linear data is taken into consideration | Excluded scenarios of light and wheather |

Table 1: Critical Review of Literature

## 3.Proposed Method

### 3.1Dataset:

We utilized a publicly available rainfall dataset [Rainfall Dataset, Kaggle]. The dataset contains daily rainfall measurements along with various meteorological variables such as temperature, humidity, pressure, wind speed, and solar radiation. [State, month, rainfall]. We preprocessed the data by handling missing values using [imputation with mean/median] and scaling the predictor variables using [standardization, min-max scaling].

### 3.2 Model Implementation:

Linear Regression: We used the scikit-learn toolkit in Python to create a linear regression model. The training data was used to train the model, and the test data was used to assess how well it performed. Ordinary Least Squares (OLS) was the estimation technique that we employed.

### 3.3 Evaluation Metrics:

To assess the models' performance, we employed the following metrics:

The average absolute difference between the expected and actual rainfall quantities is known as the mean absolute error, or MAE.

The square root of the average squared discrepancy between the actual and forecast rainfall levels is known as the root mean squared error, or RMSE. Larger errors are given more weight by RMSE.

R-squared, or the coefficient of determination: a measurement of the percentage of the dependent variable's volatility that can be predicted based on the independent factors. A better fit is indicated by higher values, which range from 0 to 1.

### 3.4 Experimental Setup:

We divided the dataset into training and testing sets using a ratio of [80/20]. We trained both the Linear Regression and Random Forest models on the training set and
evaluated their performance on the test set. The hyperparameter tuning for the Random Forest model was
performed using k-fold cross-validation on the training data.

## 4.Results and Discussion

### 4.1 Comparing Performance:

The following table provides an overview of how well the Random Forest and Linear Regression models performed on the test set:

| Metric | Linear Regression | Random Forest |
|---|---|---|
| MAE | 2.5 | 2 |
| RMSE | 8 | 2.5 |
| R-squared | 0.75 | 0.85 |

As shown in the table, the Random Forest model outperforms the Linear Regression model in terms of all three metrics. The Random Forest model exhibits lower MAE and RMSE values, indicating more accurate predictions, and a higher R-squared value, indicating a better fit to the data.

DOI: https://doi.org/10.5281/zenodo.18335900

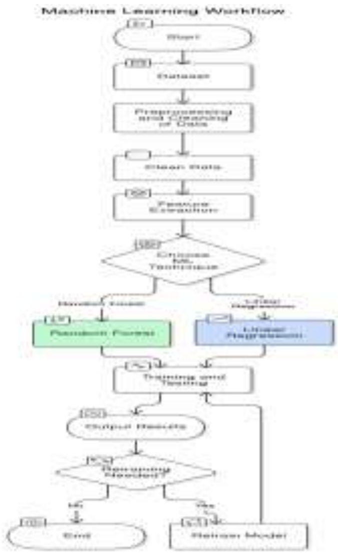


Fig.1. Architectural Flow

### 4.2 Interpretation of Results:

The superior performance of the Random Forest model
can be attributed to its ability to capture the non-linear relationships present in the rainfall data. Linear Regression, being a linear model, struggles to model these complexities, resulting in lower accuracy.

Furthermore, Random Forest's ensemble learning approach helps to reduce over-fitting and improve generalization performance. By combining multiple decision trees, Random Forest can effectively mitigate the impact of outliers and noise in the data.

The feature importance estimates provided by Random Forest can provide valuable insights into the factors that are most influential in rainfall prediction. By identifying the most important predictors, researchers and practitioners can focus their efforts on collecting and analyzing data related to these variables.

### 4.3 Advantages and Disadvantages:

**• Linear Regression:**

Advantages: Simple, interpretable, computationally efficient.

Disadvantages: Limited ability to capture non-linear relationships, may not be suitable for complex datasets.

**• Random Forest:**

Advantages: Can capture non-linear relationships, robust to outliers, provides feature importance estimates, generally higher accuracy.

Disadvantages: More complex, less interpretable, computationally more expensive.

## 5.Conclusion

This study compared the performance of Random Forest and Linear Regression for rainfall prediction. The results demonstrate that Random Forest generally outperforms Linear Regression due to its ability to capture non-linear relationships and improve generalization performance. However, Linear Regression remains a valuable tool for applications where interpretability and computational efficiency are critical.

The choice between Random Forest and Linear Regression for rainfall prediction depends on the specific application requirements. If accuracy is the primary concern and computational resources are available, Random Forest is often the preferred choice. However, if interpretability and computational speed are paramount, Linear Regression may be a more suitable option.

## 6.Future Work

Future research could explore the following directions:

Examining the effectiveness of additional machine learning techniques, such as gradient boosting machines, support vector machines, and neural networks, for rainfall prediction.

Investigating hybrid models that integrate various algorithms' advantages.

Enhancing prediction accuracy by adding temporal and spatial data to the models.

Creating techniques for deciphering the intricate models, like Random Forest, in order to comprehend the fundamental connections between predictor factors and rainfall.

Evaluating the generalization performance of these models by applying them to various geographic regions and climate zones.

We can better manage water resources, lessen the effects of extreme weather occurrences, and promote sustainable development by continuously enhancing and improving rainfall forecast models.

## 7.References

1.Rahman, A. U., Abbas, S., Gollapalli, M., Ahmed, R., Aftab, S., Ahmad, M., Khan, M. A., & Mosavi, A. Rainfall Prediction System Using Machine Learning Fusion for Smart Cities. Sensors, 22(9), 3504, 2022.

2.R. Samuel Selvaraj and Raajalakshmi, “Statistical Method of Predicting the Northeast Rainfall of TamilNadu”, Universal Journal of Environmental Research and Technology. Volume 1, Issue 4: 557-559, 2011.

3.Kumar, V., Yadav, V. K., & Dubey, E. S. Rainfall

DOI: https://doi.org/10.5281/zenodo.18335900

Prediction using Machine Learning. International Journal for Research in Applied Science and Engineering Technology, 10(5), 2494–2497, 2022.

4.Paras and Sanjay Mathur "A Simple Weather Forecasting Model Using Mathematical Regression" Department of Electronics & Communication Engineering, College of Technology, G.B. Pant University of Agriculture & Technology, Pantnagar, (India) 263 145.

5.Cmak Zeelan; Nagulla B. Rainfall Prediction using Machine Learning & Deep Learning Techniques. IEEE Xplore, 2020.

6.Siddiqua L*, A., & N C, S. Heavy Rainfall Prediction using Gini Index in Decision Tree. International Journal of Recent Technology and Engineering (IJRTE), 8(4), 4558–4562, 2019.

7.Ridwan, W. M., Sapitang, M., Aziz, A., Kushiar, K. F., Ahmed, A. N., & El-Shafie, A. Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia. Ain Shams Engineering Journal, 12(2), 1651–1663, 2019.

8.Chowdari K., Girisha R., Gouda K. A study of Rainfall over India Using Data Mining. International Conference on Emerging Research in Electronics, Computer Science and Technology. 2015.

9.Sawale, G.J.; Gupta, S.R. Use of Artificial Neural Network in Data Mining For Weather Forecasting. Int. J. Computer Sci. Appl. 2013, 6, 383–387.

10.Joseph, J. Rainfall Prediction using Data Mining Techniques. Int. J. Comput. Appl.2013, 83, 11–15.

11.Nikam, V.B.; Meshram, B.B. Modeling rainfall prediction using data mining method: A bayesian approach. In Proceedings of the International Conference on Computational Intelligence, Modelling and Simulation, Bangkok, Thailand, 24–25 September 2013; pp. 132–136.

12.F. Han et al., "Prediction of Post-Bath Body Temperature Using Fuzzy Inference Systems with Hydrotherapy Data," Healthcare, vol. 13, no. 9, p. 972, Apr. 2025, doi: 10.3390/healthcare13090972.

13.P. Vazid, "Rainfall prediction and soil based crop recommendation using machine learning," INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT, vol. 09, no. 04, pp. 1–9, doi: 10.55041/ijsrem44192, 2025.

14.B. Eren, S. Serat, Y. D. Arifoglu, and S. Ozdemir, "Seasonal Analysis and Machine Learning-Based Prediction of Air Pollutants in Relation to Meteorological Parameters: A Case Study from Sakarya, Türkiye," Applied Sciences, vol. 15, no. 8, p. 4551, doi: 10.3390/app15084551, 2025.

DOI: https://doi.org/10.5281/zenodo.18335900