

Machine Learning Approaches for Rainfall Trend Analysis: Insights from Precipitation and Meteorological Data

Mahmud Hasan
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
saif108889@gmail.com

Md. Jahadi Hasan Joy
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
mdjihadihasan2551@gmail.com

Md. Rokonujjaman
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
tarekrokonujjaman@gmail.com

Umme Rumman
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
chaiti@vu.edu.bd

Mst. Jannatul Ferdous
Computer Science and Engineering
Varendra University
Rajshahi, Bangladesh
jannat@vu.edu.bd

Abstract- Rainfall significantly impacts agriculture, water resources, and natural disasters like floods and droughts. Understanding rainfall trends is crucial for effective planning and mitigation. This study analyses rainfall trends using machine learning models trained on precipitation data and four meteorological features: temperature, specific humidity, relative humidity, and wind speed. Five models Linear Regression, KNN, SVR, Random Forest, and Gradient Boosting were evaluated using k-fold cross-validation and performance metrics, including Mean Squared Error (MSE) and R^2 Score. Among these, Random Forest outperformed the others with the lowest MSE (15.79) and the highest R^2 Score (71.69%), demonstrating its ability to capture seasonal trends. Gradient Boosting followed closely with an R^2 Score of 68.68%, while KNN achieved a moderate prediction accuracy with an R^2 Score of 67.97%. These findings highlight the potential of machine learning models for rainfall prediction, offering valuable insights for water resource management, disaster preparedness, and agricultural planning.

Keywords— Rainfall Trend Analysis, Precipitation, Machine Learning Models, Gradient Boosting, Regression Metrics, Climate Patterns, Bangladesh.

I. INTRODUCTION

Rainfall is a vital component of the hydrological cycle because it directly affects water availability, agriculture, and disaster management. In Bangladesh, rainfall patterns vary significantly across regions and seasons, influencing the country's economy and agricultural practices. Reliable rainfall prediction is crucial for effective water management, disaster preparedness, and agricultural planning, particularly in climate change. Proper forecasting can mitigate the impacts of floods, droughts, and water scarcity, which are increasingly prevalent challenges. Bangladesh experiences complex rainfall patterns, with most rainfall occurring during the monsoon season (June to October) and significantly less during the dry season (November to February). This variability, compounded by climate change, challenges resource management, making accurate rainfall prediction essential for flood control, water resource planning, and agricultural scheduling.

Machine learning (ML) models have gained popularity in climate prediction due to their ability to capture non-linear dependencies. This study utilizes five ML models—Linear Regression, Random Forest, K-Nearest Neighbors (KNN), Support Vector Regression (SVR), and Gradient Boosting—to analyze rainfall trends based on meteorological features like temperature, humidity, wind speed, and precipitation. These models are evaluated to determine the most effective approach for improving forecasting accuracy. This study contributes to the understanding of long-term rainfall trends in Bangladesh by:

- 1) Analyzing changes in rainfall patterns over time.
- 2) Employing machine learning models to predict future rainfall more reliably.
- 3) Comparing multiple models to identify the most effective approach for rainfall prediction.

The paper is structured as follows: Section I reviews related works on rainfall prediction and ML techniques. Section II describes the dataset, pre-processing steps, and methodology used to train and evaluate the models. Section III presents the results, including a comparison of model performances. Finally, Section IV summarizes the findings and suggests directions for future research in rainfall forecasting.

II. RELATED WORK

This section reviews several notable studies on rainfall prediction and machine learning techniques, particularly focusing on research relevant to Bangladesh or regions with similar climatic conditions. Several studies have explored rainfall prediction using machine learning. Rani et al. [4] developed a flood detection system using Linear Regression, SVR, and ANN, emphasizing ML's role in disaster management. Kader et al. [5] applied PSO with MLP, achieving improved accuracy by integrating optimization techniques. Misra et al. [6] used Random Forest for rainfall prediction in Odisha, India, highlighting its strength in handling climatic data. Zhang et al. [7] employed SVR and MLP for seasonal rainfall forecasting, demonstrating ML's adaptability. Adnan et al. [8] compared ANFIS-based models with traditional methods but used a limited dataset. Anwar et al. [9,10] applied Decision Trees and XGBoost, showing their

effectiveness with meteorological features. Singh et al. [11] proposed a real-time prediction system using sensors, while Park et al. [12] focused on flood estimation in South Korea. Dikshit et al. [13] analyzed drought patterns, illustrating ML’s long-term forecasting capabilities. However, many studies relied on fewer input features or region-specific datasets. Our research improves upon this by analyzing 12,053 records from Kaggle using five key meteorological features. We evaluated multiple ML models, identifying Gradient Boosting as the most effective for rainfall prediction, achieving superior accuracy and reliability.

III. RESEARCH METHODOLOGY

This section outlines the data collection, preparation, and approach used for training and evaluating the machine-learning models in this study. The focus is on the collection and pre-processing of the dataset, followed by the application of various machine-learning techniques for predicting rainfall patterns in Bangladesh.

Fig. 1. illustrates the methodology followed in this study for rainfall trend analysis. The process begins with the Weather Bangladesh Dataset, which undergoes data pre-processing steps such as data quality checks, normalization, and outlier removal. Relevant features are then selected to create a modified dataset. Parameter tuning is applied to optimize the models, followed by splitting the data using Stratified K-Fold Cross-Validation ($n_splits=5$). Classification is performed using five machine learning models—Linear Regression, Random Forest, KNN, SVR, and Gradient Boosting. The models are evaluated using Regression Metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 Score and the results are compared to determine the most effective model for rainfall trend analysis. To enhance model performance and reduce complexity, feature selection was performed to identify the most relevant meteorological variables affecting rainfall prediction. After analyzing correlations and domain relevance, temperature, specific humidity, relative humidity, wind speed, and precipitation were selected as key input features.

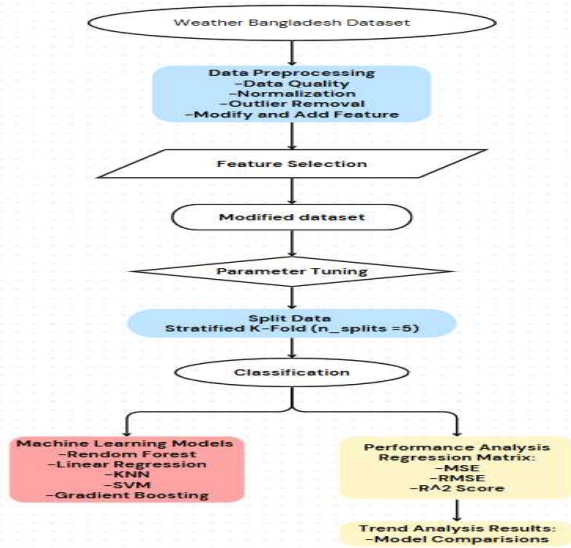


Fig. 1. Methodology Representation

A. Dataset

The dataset used in this study was collected from a weather data source, Kaggle's official website. The dataset contains weather records from Bangladesh spanning from the year 1990 to 2022. The dataset consists of 12,053 records, with daily measurements for the following meteorological attributes: temperature, specific humidity, relative humidity, wind speed, and precipitation data as shown in Table I. These records provide both input features (e.g., temperature, humidity) and the target variable (rainfall occurrence).

TABLE I. DESCRIPTION OF METEOROLOGICAL VARIABLES IN THE DATASET

Attribute	Description	Type
Temperature	Daily recorded temperature (°C)	Numeric
Specific Humidity	Amount of water vapor in the air (kg of water vapor per kg of air)	Numeric
Relative Humidity	Relative humidity (kg of water vapor per kg of air)	Numeric
Wind Speed	Wind speed (m/s or km/h)	Numeric
Precipitation	Precipitation on a given day (mm)	Numeric

Fig. 2 illustrates the distribution of rainfall based on precipitation data. The chart categorizes the dataset into two segments: "Rain," representing 34.4% of instances with measurable precipitation, and "No Rain," representing 65.6% of instances without measurable precipitation. This visualization highlights the imbalance in rainfall, with a larger proportion of no-rain events.



Fig. 2. Rain versus no-rain distribution based on precipitation.

Fig. 3 presents the correlation heatmap, highlighting key meteorological variables influencing precipitation. Specific humidity shows a strong positive correlation (0.77) with precipitation, indicating its significant role in rainfall formation. Relative humidity also exhibits a moderate correlation (0.59), reinforcing its impact. Wind speed and temperature display weaker correlations, suggesting indirect effects on precipitation patterns. These insights guided our feature selection, ensuring the inclusion of relevant variables—specific humidity, relative humidity, wind speed, and temperature—to enhance model accuracy in capturing rainfall trends, performance while minimizing noise from less significant factors.

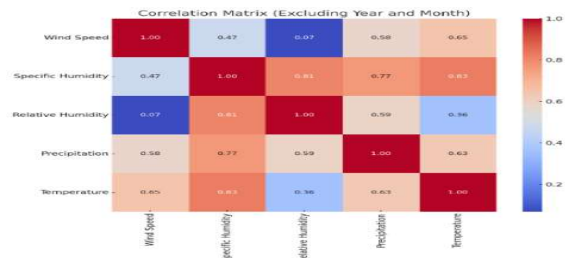


Fig. 3. Feature Correlation Analysis

Fig. 4 illustrates the monthly rainfall trends in Bangladesh from 1990 to 2022. The graph reveals a clear seasonal pattern, with peak rainfall observed during the monsoon months (June to September), characterized by significantly higher rainfall amounts than other months.

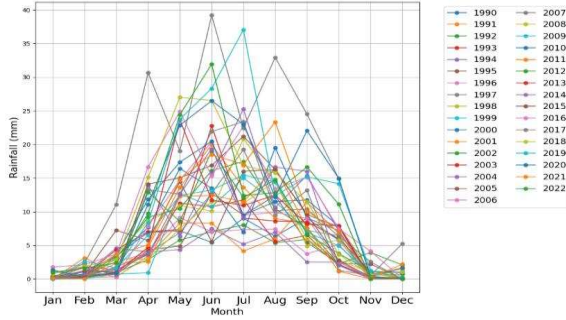


Fig. 4. Monthly rainfall representation (1990 to 2022)

Conversely, the dry season (November to February) consistently records minimal rainfall across all years. The inter-annual variability in rainfall during the monsoon months is also evident, reflecting fluctuations influenced by climatic and environmental factors. This analysis highlights the critical role of the monsoon in shaping rainfall patterns and its implications for water resource management and agriculture.

B. Regression algorithms

Linear Regression (1) models the relationship between the target variable (rainfall) and the input features using a linear combination of the predictors. The model was fitted using the least-squares method to minimize the residual sum of squares. K-Nearest Neighbors (KNN) (2), which predicts based on the average value among k nearest data points, was optimized with $n_neighbors = 9$ and uniform weights. Support Vector Regression (SVR) (3) constructs a hyperplane in high-dimensional space to minimize prediction errors within a margin, using a radial basis function kernel ($Kernel = 'rbf'$) and a regularization parameter $C = 10$. Random Forest Regressor (4), an ensemble of decision trees, was tuned with 50 trees ($n_estimators = 50$) a maximum tree depth of 10, and a minimum of 2 samples per split. Finally, Gradient Boosting Regressor (5) iteratively combines weak learners to minimize prediction error, using a learning rate of 0.1, 100 estimators, and a maximum depth of 3 for each tree.

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n \quad (1)$$

$$d(i, j) = \sqrt{\sum_{k=1}^n (x_{i,k} - x_{j,k})^2} \quad (2)$$

$$f(x) = Mean\{f_1(x), f_2(x), \dots, f_T(x)\} \quad (3)$$

$$f(x) = w \cdot x + b \quad (4)$$

$$y_t = y_{t-1} + \eta \cdot g(x; \theta_t) \quad (5)$$

The performance of the regression algorithms was assessed using three key metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 Score. MSE (6) measures the average squared difference between predicted and actual values, penalizing larger errors more heavily, and serves as a comprehensive indicator of prediction accuracy. RMSE (7), the square root of MSE, provides an interpretable measure of error in the same units as the target variable, making it more intuitive for understanding prediction deviations. Finally, the R^2 Score (8) quantifies the proportion of variance in the target variable that the model explains, with values closer to 1 indicating better performance. Together, these metrics provide a balanced evaluation of regression model performance, highlighting both the magnitude of prediction errors and the model's ability to capture underlying trends in the data.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (6)$$

$$RMSE = \sqrt{MSE} \quad (7)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n ((y_i - \hat{y}_i)^2)}{\sum_{i=1}^n ((y_i - \bar{y})^2)} \quad (8)$$

IV. RESULT ANALYSIS AND DISCUSSION

Table II presents a comparative analysis of supervised machine learning algorithms for predicting rainfall probability based on monthly climate data. The comparison is conducted using Residual Plots and three regression performance evaluation metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 Score.

TABLE II: EVALUATION METRICS OF MACHINE LEARNING MODELS FOR RAINFALL PREDICTION (MSE, RMSE, R^2)

Model	MSE	RMSE	R^2 Score
Random Forest	15.7870	3.9499	0.7169
KNN	17.9872	4.1961	0.6796
Linear Regression	18.1854	4.2552	0.6693
SVR	29.4636	5.4179	0.4727
Gradient Boosting	17.4423	4.1697	0.6860

Fig. 5 presents the MSE, RMSE and R^2 Scores for five machine learning models from— Linear Regression, Random Forest, KNN, SVR, and Gradient Boosting—evaluated using k-fold cross-validation for precipitation prediction. Among the models, Random Forest achieves the best performance with the lowest MSE (15.79) and the highest R^2 Score (0.7169), indicating strong prediction accuracy and reliability. Gradient Boosting follows closely with an R^2 Score of 0.6868 and an MSE of 17.44, demonstrating competitive performance. KNN also performs well, achieving an R^2 Score of 0.6797 and an MSE of 17.99. In contrast, Linear Regression and SVR show

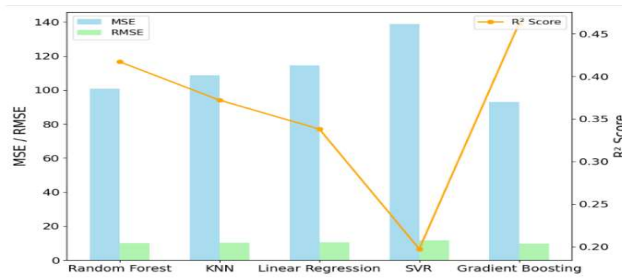


Fig. 5. Model Performance Metrics: MSE, RMSE, and R^2 Score for Rainfall Trend Analysis

comparatively lower R^2 Score (0.6693 and 0.4727, respectively) and higher MSE values, reflecting reduced prediction effectiveness. The visualization underscores the models' comparative strengths and stability in capturing precipitation trends. The residual plots in Fig. 6 illustrate the difference between observed and predicted values for each model, providing insights into their performance.

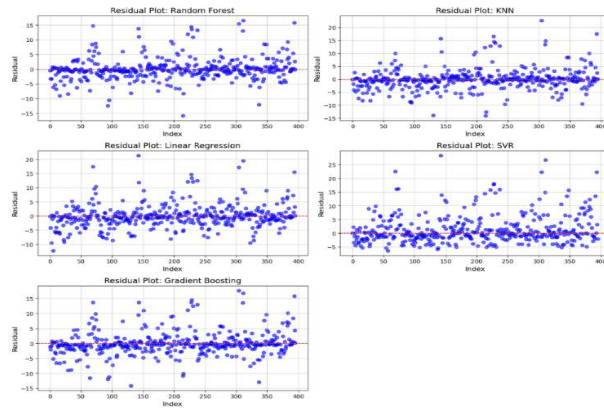


Fig. 6. Residual Plots for Model Performance Evaluation (5-Fold Cross-Validation)

V. CONCLUSION

This study analyzed rainfall trends in Bangladesh using five machine learning models—Linear Regression, KNN, SVR, Random Forest, and Gradient Boosting—based on meteorological features like temperature, humidity, wind speed, and precipitation. KNN achieved the highest accuracy ($R^2 = 0.6797$, $MSE = 17.99$), followed by Random Forest and Gradient Boosting, demonstrating their effectiveness in capturing rainfall patterns. While Linear Regression and SVR showed lower performance, ensemble techniques proved more reliable. The study highlights the potential of ML for rainfall prediction but acknowledges limitations such as the exclusion of additional climatic factors and dataset generalizability. Future research should incorporate more diverse features, larger datasets, and advanced models like neural networks to enhance accuracy and scalability. These findings provide a foundation for improved rainfall forecasting, benefiting water management, disaster preparedness, and agriculture in Bangladesh.

REFERENCES

- [1] J.K. Basak and R.A.L. Mahmud Titumir, "Climate Change in Bangladesh: A Historical Analysis of Temperature and Rainfall Data," [Online].
- [2] K. Prathibha, G. Rithvik Reddy, Harsh Kosre, and K. S. Prasad, "Rainfall Prediction Using Machine Learning," in *Machine Intelligence Techniques for Data Analysis and Signal Processing*, Springer, 2023, pp. 637–645. [Online].
- [3] S. Belayneh and J. Adamowski, "Machine Learning Techniques to Predict Daily Rainfall Amount," *Journal of Big Data*, vol. 8, no. 1, Article 45, 2021. [Online].
- [4] A. Krasnopolsky and D. V. Chalikov, "Machine Learning to Improve Numerical Weather Forecasting," in *Proceedings of the 2020 IEEE Conference on Big Data and Analytics (ICBDA)*, IEEE, 2020, pp. 124–134. [Online].
- [5] S. S. R. Depuru, "Rainfall Prediction: A Comparative Analysis of Modern Machine Learning Techniques," *Machine Learning with Applications*, vol. 4, Article 100204, 2021. [Online]. Available: <https://doi.org/10.1016/j.mlwa.2021.100204>.
- [6] Md. Sarwar Jahan and S.M. Abdullah, "A Trend Analysis of Rainfall in Khulna District of Bangladesh," [Online]. Available: https://www.researchgate.net/publication/384562899_A_Trend_Analysis_of_Rainfall_in_Khulna_District_of_Bangladesh
- [7] R.K. Sutradhar, S. Dey, Md. Tauhid Ur, and B.C. Mondal, "Spatial and Temporal Analysis of Rainfall and Temperature Trends of Bangladesh during 1989 to 2019 and the Possible Impacts of Rainfall and Temperature Changes," DOI: <https://doi.org/10.55248/gengpi.4.823.50981>
- [8] S.R. Devi, C. Venkatesh, P. Agarwal, and P. Arulmozhivarman, "Daily Rainfall Forecasting Using Landslides," 2014 *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2014.
- [9] S. Srivastava, N. Anand, S. Sharma, S. Dhar, and L.K. Sinha, "Monthly Rainfall Prediction Using Various Machine Learning Algorithms for Early Warning of Landslide Occurrence," 2020 *International Conference for Emerging Technology (INCET)*, 2020.
- [10] H. Abdel-Kader, M. A.-E. Salam, and M. Mohamed, "Hybrid Machine Learning Model for Rainfall Forecasting," *Journal of Intelligent Systems and Internet of Things*, 2021.
- [11] R.K. Misra, P.K. Panda, A.K. Sahu, S. Sahoo, and D.P. Behera, "Rainfall Prediction Using Machine Learning Approach: A Case Study for the State of Odisha," *Indian Journal of Natural Sciences*, 2020.
- [12] X. Zhang, S.N. Mohanty, A.K. Parida, S.K. Pani, B. Dong, and X. Cheng, "Annual and Non-Monsoon Rainfall Prediction Modelling Using SVR-MLP: An Empirical Study from Odisha," *IEEE Access*, vol. 8, pp. 30223–30233, 2020.
- [13] M.M. Hossain and K. Ahmed, "Climate Change Impacts and Adaptation Strategies in Bangladesh: A Review," *Cogent Food & Agriculture*, vol. 5, no. 1, 1615717, 2019.
- [14] M.S. Rahman and M.M. Rahman, "Recent Trend and Variability of Monsoon Rainfall Over Bangladesh: Implications for Sustainable Agricultural Production," *Agriculture, Ecosystems & Environment*, vol. 288, 106717, 2020.
- [15] M.H. Khan and M.R. Chowdhury, "Future Changes in Rainfall Variability Over Bangladesh Using High-Resolution Regional Climate Models," *Climate Dynamics*, vol. 51, no. 1–2, pp. 79–98, 2018.
- [16] M.R. Islam and H. Uyeda, "Analysis of Historical Rainfall Variability and Trends in Bangladesh," *Theoretical and Applied Climatology*, vol. 137, no. 1–2, pp. 281–294, 2019.
- [17] F.M. Chowdhury and M.S. Alam, "Predicting Rainfall Patterns in Bangladesh Using Machine Learning Algorithms," *IEEE Access*, vol. 9, pp. 102632–102644, 2021.
- [18] D. Das and M.R. Khan, "Impact of Climate Change on Rainfall Patterns and Its Implications for Water Resources Management in Bangladesh," *Journal of Water and Climate Change*, vol. 8, no. 4, pp. 761–774, 2017.