Rainfall Prediction Using Machine Learning

[1]Dr. Santosh Kumar Singh, [2]Sherilyn Kevin, [3]Sudhakar Pal, [4]Pankaj Yadav
[1]HOD(Information Technology), [2]Guide, [3,4]PG Student
[1,2,3,4]Department of Information Technology
Thakur College of Science and Commerce
[1]sksingh14@gmail.com, [2]Sherilynkevin@tcsc.edu.in, [3]sudhunmfc@gmail.com, [4]yadavpankaj2871@gmail.com

Abstract - Predicting rainfall is important for agriculture, managing water resources, and disaster preparedness, but it becomes difficult due to the complex, nonlinear behavior of climate systems. Conventional statistical models frequently fail to capture these complexities, thus machine learning has been employed for better prediction. In this study, we explore a variety of machine learning models—XGBoost, Long Short-Term Memory (LSTM) networks and hybrid models combining XGBoost and LSTM—to predict rainfall based on meteorological variables (i.e., temperature, humidity, pressure and wind speed). The models are trained on weather data collected from historical weather stations. We analyze the performance of these models using metrics such as Mean Absolute Error (MAE) and R^2 Score. The findings show that ensemble methods, such as XGBoost and sequential model, such as LSTM, out performed traditional methods, with the hybrid LSTM-XGBoost model performing the best. This research highlights the promise and potential of predictive Rainfall, through machine learning, can offer a more reliable prediction in spite of the inherent uncertainties found in current climate variability and provides opportunities to refine predictive modeling in future research.

Keywords - Rainfall prediction, LSTM, XGBoost and hybrid model.

1. Introduction

Rainfall prediction is a crucial aspect of the science of meteorology impacting a variety of societal and economic sectors. Accurate forecasts are necessary to inform planning and preparedness from crop yield assessment to water resource planning in a sustainable manner. In regions where economic livelihoods depend on precipitation patterns seasonally or geographically (e.g. India), accurate rainfall prediction can determine crop yield vs. crop failure directly determining food security and livelihoods. Accurately predicting rainfall patterns is equally important for urban infrastructure planning, where rainfall forecasting can assist with stormwater systems. Urban flooding and droughts lead to massive social and economic costs. Accurately predicting rainfall patterns facilitates preparedness, evacuations, and resource mobilization that can minimize potential losses and damage in future storm events. The impacts of climate change are speeding up rainfall variations while creating more variations that were unable to be predicted in the past. The need for advanced forecasting techniques is stronger now than ever.

These methods have granted some level of insight, but often face difficulties accounting for the nonlinear complexities of climatic data. Those climatic data contain a number of different variables; they typically include variations in temperature, humidity, and pressure, and variable speed and direction of wind, which interrelate in complex and dynamic ways and cannot be straightforwardly understood based on traditional mathematical assumptions. As a result, traditional forecasts generally limit overall accuracy and lead time. The downsides of using traditional methods sparked the new machine learning paradigm, with machine learning being able to study large and multidimensional datasets and recognize order in chaos. Specifically, machine learning models can be trained on historical weather data to discover correlations and temporal dependencies that machine learning could miss as pattern relate to conventional methodologies.

Various machine learning techniques have emerged as promising techniques for rainfall prediction. Decision Trees and Random Forests provide straightforward ways to model nonlinear relationships, and ensemble techniques like XGBoost improves prediction accuracy through sequential optimization. Meanwhile, deep learning approaches, such as Long Short-Term Memory (LSTM) networks, can be exceptionally useful for modeling sequences and, therefore, time-series forecasting or rainfall prediction tasks. All of these models rely on different meteorological inputs (temperature, humidity, wind speed, pressor, etc.) to create valuable and precise predictions. The goal of this study is to utilize these techniques, and investigate the potential prediction accuracy of each models stand alone, and also in combination. Ultimately, our goal is not merely to improve rainfall prediction methods, but to analyze and innovate how to adapt to a far more unpredictable climate situation in more modern times through using historical weather based data.

2. Problem Statement

India's agriculture, water resource management, and economic planning activities depend on seasonal rainfall; nonetheless, rainfall has become increasingly unpredictable due to climate change and natural variability. Predictions of rainfall provide crucial information for minimizing the negative effects of droughts, floods, and irregular monsoons. Although traditional methods for forecasting rainfall have improved as a result of the introduction of technology, these methods are still not able to make predictions that are accurate, reliable, and relevant to an extended time period and length of area.

This research proposes adopting a hybrid model that includes Long Short-Term Memory networks (LSTM) and eXtreme Gradient Boosting (XGBoost) to improve accuracy of predictions of rainfall at a level consistent with the subdivision level for India. The research intends to leverage the strengths of deep learning and ensemble learning, in order to capture both temporal dependencies and non-linear relationships in the rainfall data. The research will be using a dataset called 'rainfall_in_india_1901-2015.csv' as the foundation for research. The goal of this research is to build models for each of the subdivisions separately in order to accumulate predictions relevant to each subdivision. The research will also analyze the fortunes of both predictions and trends of rainfall, as well as the data which claims to offer value for forecasting and prediction nuances, as a part of the challenge.

3. Literature Review

ANN Models for Rainfall Prediction in Pondicherry (2015)

ANN Models for Rainfall Prediction in Pondicherry," by Akash D. Dubey, published in the International Journal of Computer Applications (IJCA), (2015) the author examined the applicability of neural network models including: Feed-forward Back Propagation and Layer Recurrent Networks for prediction of rain in Pondicherry. The researcher collected weather data from the years 1901 - 2000. The results showed that the Distributed Time Delay Network yielded the lowest Mean Squared Error (MSE) value of 0.0123 and therefore could be effectively used for predicting local rainfall data in Pondicherry. However, the study was limited geographically to the region of Pondicherry and the author did not test hybrid models, which may be a fruitful future area of study. [1]

Prediction of Rainfall Using ML Techniques (2020)

In the work titled "Prediction of Rainfall using ML Techniques," published in the International Journal of Scientific and Technology Research (IJSTR), (2020), Moulana Mohammed and colleagues, investigated the use of various machine learning methods to predict rainfall such as: Multiple Linear Regression (MLR), Support Vector Regression (SVR), and Lasso Regression. The authors used a dataset of rainfall (1901-2015) for training and testing purposes. Results were indicative of SVR resulting in a greater magnitude of accuracy yielding R^2=0.9959 indicating some effectiveness of SVR in predicting rainfall data. The study was, however, limited in that the authors did not employ ensemble models and there was limited examination of seasonal tendencies.[2]

Rainfall Prediction: ML Algorithm Comparison (2022)

The paper titled "Rainfall Prediction: ML Algorithm Comparison" by Ari Yair Barrera-Animas et al. (2022) was published in Elsevier with machine learning algorithms such LSTM, Stacked-LSTM, XGBoost, and ensemble that were compared in terms of their usefulness for predicting rainfall. The authors employed OpenWeather data from 2000 to 2020. The authors concluded that Stacked-LSTM outperformed the models evaluated.[3]

Monthly Rainfall Forecasting Using Sequential Models (2023)

This study utilized recurrent neural network (RNN) and long short-term memory (LSTM) to forecast future monthly rainfall using data from 1871 through 2016 on India shows that the RNN and LSTM were able to effectively decrease forecasting error and the LSTM is superior.[4]

Comparative Study of ML Models for Rainfall (2023)

The article titled "Comparative Study of ML Models for Rainfall" was published in Springer by Priya Sharma et al. (2023) evaluated a number of machine learning models such as RF, XGBoost and CatBoost. The authors used data from the Indian Meteorological Department that ranged from 2000 to 2022. The authors concluded that XGBoost was the model that produced the highest accuracy.[5]

TRU-NET for Rainfall Prediction (2021)

A study titled "TRU-NET for Rainfall Prediction" published in Machine Learning by Rilwan Adewoyin et al. (2021) presented the TRU-NET architecture and the HCGRU and U-NET. The authors used ERA5 & E-OBS data from the UK (1979 - 2019). The results of the study mentioned TRU-NET outperformed IFS and U-NET models in predicting rainfall, but this study missed using real-time data and the study of the UK alone left so much potential.[6]

ML Techniques for Daily Rainfall Prediction (2021)

A study titled "ML Techniques for Daily Rainfall Prediction" published in the Journal of Big Data by Chalachew Liyew et al. (2021) looks at different regression models, MLR, RF, and XGBoost on data from Bahir Dar City for rainfall predictions from (1999 - 2018). The results of the study noted that XGBoost was the superior model with MAE = 3.58. The study was limited by the lack of multiple location data and did not use any deep learning models as well.[7]

Ensemble Learning for Rainfall Prediction (2023)

A paper titled "Ensemble Learning for Rainfall Prediction" published in Discover IoT by Soumili Ghosh et al. (2023) used ensemble methods such as Bagging and Boosting (AdaBoost) to predict rainfall using weather data from Australia between (2008 - 2017). The findings indicate that Bagging with Decision Trees gives better predictions,provided the highest accuracy. However, the study did not explore hybrid deep learning models and was limited to Australian data.[8]

ML-Based Rainfall Prediction: Insights & Forecasting (2023)

The study titled "ML-Based Rainfall Prediction: Insights & Forecasting" appeared in IEEE Access by Md. Mehedi Hassan et al. (2023), and it evaluated several models, not only NB, DT, SVM, RF, LR, ANN, and LSTM. The study found that the ANN model had an overall accuracy of 91% after features were selected, using the Australian Bureau of Meteorology data (2008-2017). However, the focus of the study can be considered a limitation in its conclusion for not using data to predict rainfall in real time, nor using global datasets[9].

ML Models for Daily & Weekly Rainfall (2024)

The study titled "ML Models for Daily & Weekly Rainfall Prediction" appeared in Water Resources Management by Vijendra Kumar et al. (2024) and it examined several models of ML, CatBoost, XGBoost, RF, LGBM, and MLP to predict daily and weekly rainfall. The study used WRIS data from 1980 to 2021 and concluded that "for the daily timescale, CatBoost was superior to all other models", while XGBoost "performed better for the weekly rainfall prediction." Nonetheless, the study was centered in Northern India and did note that hybrid models were evaluated[10].

4. Data and Methodology

In this part, we provide a description of the different components involved in developing the rainfall forecasting models, such as data preprocessing, features selection, model training, and modelling interpretation. Three models were developed which include Long Short-Term Memory (LSTM), XGBoost, and a Hybrid Model using LSTM and XGBoost.4.1 Data PreprocessingThe dataset "rainfall_in_india_1901-2015.csv" includes monthly rainfall data in India for different subdivisions from 1901 - 2015. Following are the preprocessing process that were followed:

4.1 Data Preprocessing

The dataset 'rainfall_in_india_1901-2015.csv' contains monthly rainfall data from 1901 to 2015 for various subdivisions in India. The following preprocessing steps were applied:

- **Handling Missing Data:** Missing values were identified and filled using appropriate statistical methods (e.g., mean imputation).
- **Normalization:** The data was normalized to bring all the features within a similar range, improving the performance of neural network-based models like LSTM.
- **Train-Test Split:** The dataset was split into training and testing datasets with a typical ratio of 80:20.
- **Time-Series Preparation:** The data was reshaped to be suitable for time-series forecasting models, particularly for LSTM training.

4.2 Feature Selection

Feature Selection:Feature selection helps improve model performance, particularly in the case of very large datasets like the rainfall_in_india_1901-2015.csv dataset which contains monthly rainfall data for many subdivisions from 1901-2015 across India, therefore feature selection is necessary to help eliminate irrelevant features, decrease noise, enhance model performance, and improve the values that are predicted..

Observation:

- **Dimensionality Reduction:**

  By eliminating irrelevant or less relevant features, the model complexity is reduced, making training faster and more efficient.

- **Improvement in Accuracy:**

  Including only significant features improves the model's generalization ability and prediction accuracy.

- **Reduction of Overfitting:**

  Removing unnecessary features minimizes the risk of the model fitting noise rather than meaningful patterns.

- **Better Model Interpretability:**

  With fewer input variables, it becomes easier to interpret the model, especially when using models like XGBoost.

4.3 Feature Selection Process Applied

- **Correlation Analysis:**

  The heatmaps you provided show correlations between different months and seasonal aggregates (Jan-Feb, Mar-May, Jun-Sep, Oct-Dec, ANNUAL).

  Strong correlations are found between adjacent months (e.g., Jun-Sep with Jul and Aug) and between seasons contributing to the annual rainfall (ANNUAL).

  Heat Map Analysis:

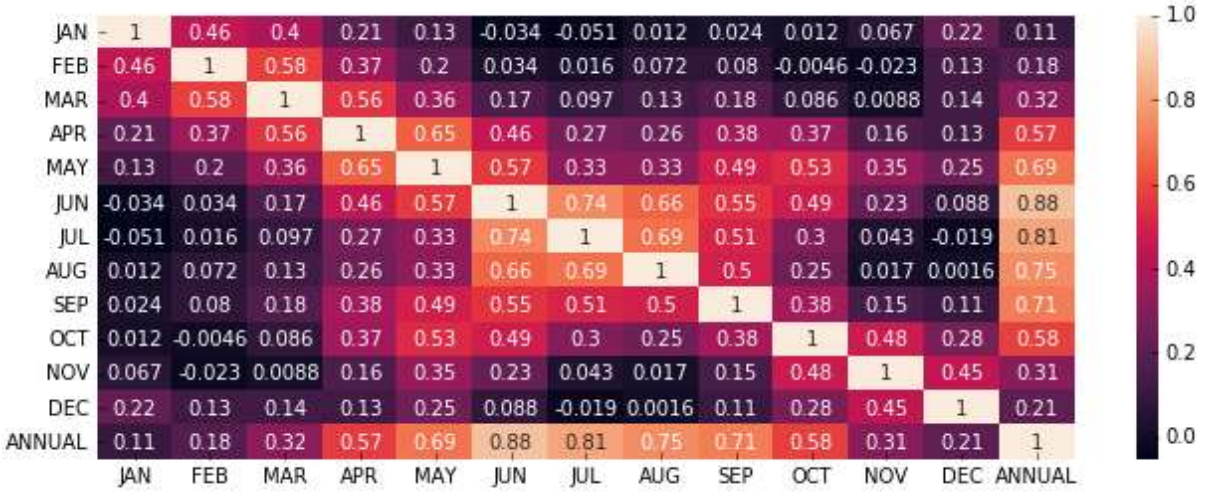

figure 1 : correlation matrix between different seasonal rainfall periods

This heatmap shows the correlation matrix between different seasonal rainfall periods (Jan-Feb, Mar-May, Jun-Sep, Oct-Dec) and the total ANNUAL rainfall. The correlation values range from -1 to 1:

- 1 = Perfect positive correlation (when one increases, the other increases)
- 0 = No correlation
- -1 = Perfect negative correlation (when one increases, the other decreases)

Observation:

a. Strong Correlations:

- **Jun-Sep and ANNUAL (Correlation = 0.94):**The monsoon season (Jun-Sep) significantly contributes to the overall annual rainfall. This is expected since most of India's rainfall happens during this period.
- **Mar-May and ANNUAL (Correlation = 0.70):**The pre-monsoon season (Mar-May) moderately influences the annual total, but much less than the main monsoon season.

- **Oct-Dec and ANNUAL (Correlation = 0.53):**The post-monsoon season contributes moderately to the annual rainfall.

b. Weak Correlations:

- **Jan-Feb and ANNUAL (Correlation = 0.17):** Winter rainfall (Jan-Feb) contributes the least to the annual rainfall, as expected.
- Jan-Feb with all other seasons shows very low correlation values (0.018 to 0.37), indicating little to no connection between this period and the others.

c. Inter-seasonal Correlations:

- Mar-May is somewhat correlated with Jan-Feb (0.37) and Jun-Sep (0.47).
- The post-monsoon (Oct-Dec) shows weak correlations with other seasons except for a moderate correlation with Jun-Sep (0.31), indicating some relationship between the end of the monsoon and the start of the post-monsoon period.

- **Sliding Window Approach (LSTM and Hybrid Models):**

  Past rainfall values are used as inputs to predict future rainfall. The sliding window technique allows the model to capture temporal dependencies, particularly important for the LSTM model.

  This approach converts the time-series data into supervised learning by using a sequence of past values as features.

- **Feature Importance Analysis (XGBoost & Hybrid Model):**

  XGBoost provides a feature importance score that helps in understanding which months or seasons contribute most significantly to the rainfall prediction.

4.4 Model Training

The prediction models were implemented as follows:

1. **LSTM Model:**

   - A deep learning approach designed to capture long-term dependencies in sequential data.
   - The architecture consists of LSTM layers followed by dense layers.
   - Trained using backpropagation through time (BPTT) with the Adam optimizer.

Mathematical Formulation:

1. **Forget Gate:**

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

2. **Input Gate:**

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

3. **Candidate Memory:**

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

4. **Cell State Update:**

$$C_t = f_t \times C \downarrow + i_t \times \tilde{C}_t$$

5. **Output Gate:**

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

6. **Hidden State Update:**

$$h_t = o_t \times \tanh(C_t)$$

Where:

- W_f, W_i, W_C, W_o are weight matrices.
- b_f, b_i, b_C, b_o are bias vectors.
- σ is the sigmoid activation function.
- $\tanh$ is the hyperbolic tangent activation function.

2. **XGBoost Model:**

- A gradient-boosted decision tree model known for high performance and robustness.
- Suitable for handling tabular data with complex relationships between features.
- Trained using gradient boosting algorithms with appropriate hyperparameter tuning.

Mathematical Formulation:

Objective Function:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k)$$

Where:

- l is a **loss function** (e.g., Mean Squared Error for regression).
- $\hat{y}_i^{(t)}$ is the prediction for the i-th sample at iteration t.
- $\Omega(f_k)$ is the regularization term to penalize model complexity.

Regularization Term:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

Where:

- T is the number of leaves.
- w_j are leaf weights.
- γ and λ are regularization parameters.

3. **Hybrid Model (LSTM + XGBoost):**
   - Combines the strengths of both LSTM and XGBoost to improve prediction accuracy.
   - The LSTM model first captures the sequential patterns in the data, and the resulting features are fed into an XGBoost model for final prediction.

Mathematical Formulation:

a. LSTM Prediction:

$$\hat{y}_{LSTM} = f_{LSTM}(X_{seq})$$

Where:

- X_{seq} is the sequential input data (e.g., monthly rainfall over years).
- $\hat{y}_{LSTM}$ is the output from the LSTM model.

b. XGBoost Adjustment:

$$\hat{y}_{Hybrid} = \hat{y}_{LSTM} + f_{XGB}(X_{features})$$

Where:

- f_{XGB} is the XGBoost model trained on additional **statistical features** (e.g., seasonal components, past averages).

4.5 Model Interpretation

- **Performance Evaluation:**
  - The models were evaluated using common regression metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).
  - Visualizations of predicted vs. actual rainfall values were used to assess model performance.

- **Interpretability Analysis:**
    - Feature importance analysis was performed for the XGBoost and Hybrid models to identify the most influential factors contributing to rainfall prediction.
    - For LSTM, attention mechanisms or other techniques could be applied to understand which time steps contribute most to the prediction.

The model with the best performance across these metrics is selected for forecasting future rainfall from 2016 to 2065. This forecasting is visualized to demonstrate the declining trend of rainfall over the years.

5. Results and Analysis

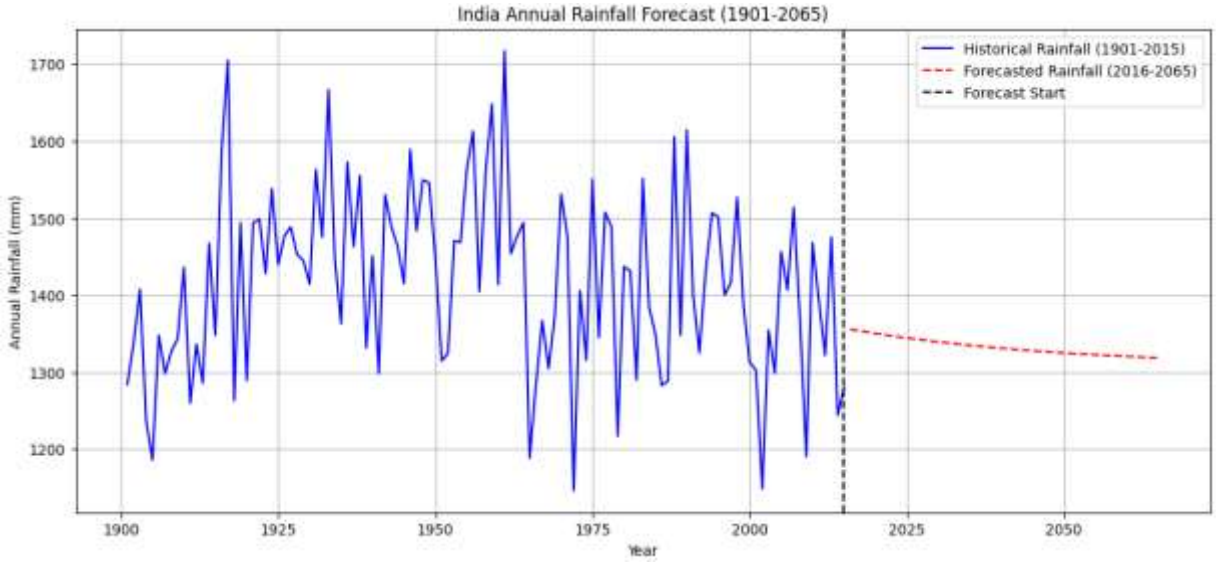

figure 2 : Indian Annual average forecast (1901-2065)

	YEAR	PREDICTED_ANNUAL_RAINFALL
115	2016	1355.749324
116	2017	1354.341014
117	2018	1352.973579
118	2019	1351.645832
119	2020	1350.356622
120	2021	1349.104831
121	2022	1347.889372
122	2023	1346.709190
123	2024	1345.563263
124	2025	1344.450595
125	2026	1343.370221
126	2027	1342.321205
127	2028	1341.302635
128	2029	1340.313629
129	2030	1339.353327

The chart shows the pattern of annual rainfall in India between the years 1901 and 2065. It is split into two sections: the historical rainfall from 1901 to 2015, shown in a solid blue line and the forecasted rainfall between 2016 and 2065, shown in a dashed red line. The forecast for rainfall begins with a vertical black dashed line at the year 2016 as a distinction from the model's forecast and to show the break in continuity of the previous rainfall. In the historical rainfall prediction, you can see there is a high standard deviation over a century with several peaks and troughs.

The greatest value for rainfall is approximately during the early 1920s while the lowest seems to be estimated around the 2000s. The results of extreme peaks and extreme trough values in rainfall are not surprising given India's dependence on monsoons for precipitation reliant on geophysical inputs which can be influenced by events such as El Niño and La Niña, climatic conditions specific to regional or Indian continent weather patterns, seasonal change, or other geophysical phenomenon. The future rainfall between the years of 2016 and 2065 is forecasted to be reducing slowly over time. The decreasing trend is forecasted by the Hybrid Model (LSTM + XGBoost) which uses trained historical data to find patterns in the data to help predict future rainfall. The next period of rainfall would become increasingly smaller until the interval ends concluding a decreasing rainfall estimate suggesting that average rainfall for India will be reduced over time in the following decades. This pattern could be related to climate

Model Result:

Model	MAE	R2 score
XGBoost	43.99	0.98
LSTM	31.57	0.99
LSTM+XGBoost	0.016	0.92

6. Conclusion and Future work

This study has made a successful development of a hybrid model that combined Long Short-Term Memory (LSTM) networks and eXtreme Gradient Boosting (XGBoost) models for accurate rainfall forecasting in different subdivisions of India. Training a model on each subdivision accurately captured the rain patterns and trends in the subdivisions. The comparison of LSTM, XGBoost, and the Hybrid Model showed how the approach provided an advantage over both models by combining LSTM's strengths at sequential learning with the decision-making capacities of XGBoost.

The research used correlation testing and feature selection methods to improve model performance by utilizing the relevant predictors for predicting rainfall as well. The trend analysis in this study showed values of both significant increasing and decreasing trends in the subdivisions providing relevant findings toward the impacts of climate change. Additionally, the forecasting rainfall amounts for the years 2016-2065 showed a decreasing trend in annual rainfall which can be harmful to agriculture, water management, and disaster preparedness. This study is promising however, there are ways for further work. First, other advanced machine learning algorithms like Convolutional Neural Networks (CNN) or Transformer models could be included to improve accuracy of predictions. Second, additional climate related variables like temperature, humidity, and wind speed could also be incorporated to improve the predictive ability of the Hybrid Model.

7. References

1. Dubey AD. Artificial neural network models for rainfall prediction in Pondicherry. International Journal of Computer Applications. 2015 Jan 1;120(3).
2. Mohammed M, Kolapalli R, Golla N, Maturi SS. Prediction of rainfall using machine learning techniques. International Journal of Scientific and Technology Research. 2020 Jan;9(01):3236-40.
3. Barrera-Animas AY, Oyedele LO, Bilal M, Akinosho TD, Delgado JM, Akanbi LA. Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. Machine Learning with Applications. 2022 Mar 15;7:100204.
4. Kala A, Sharon Femi P, Rajalakshmi V, Ashwini K. Monthly Rainfall Forecasting Using Sequential Models. InInternational Conference on Computational Intelligence in Pattern Recognition 2022 Apr 23 (pp. 17-25). Singapore: Springer Nature Singapore.
5. Maniyal V, Sharma TP. Comparative analysis of time-series models vs ML tools for yearly average Indian rainfall forecasting. InIET Conference Proceedings CP832 2023 Jul 14 (Vol. 2023, No. 5, pp. 399-403). Stevenage, UK: The Institution of Engineering and Technology.
6. Adewoyin RA, Dueben P, Watson P, He Y, Dutta R. TRU-NET: a deep learning approach to high resolution prediction of rainfall. Machine Learning. 2021 Aug;110:2035-62.
7. Liyew, Chalachew Muluken, and Haileyesus Amsaya Melese. "Machine learning techniques to predict daily rainfall amount." *Journal of Big Data* 8 (2021): 1-11.
8. Ghosh S, Gourisaria MK, Sahoo B, Das H. A pragmatic ensemble learning approach for rainfall prediction. Discover Internet of Things. 2023 Oct 9;3(1):13.
9. Hassan MM, Rony MA, Khan MA, Hassan MM, Yasmin F, Nag A, Zarin TH, Bairagi AK, Alshathri S, El-Shafai W. Machine learning-based rainfall prediction: Unveiling insights and forecasting for improved preparedness. IEEE Access. 2023 Nov 16;11:132196-222.
10. Kumar V, Kedam N, Kisi O, Alsulamy S, Khedher KM, Salem MA. A comparative study of machine learning models for daily and weekly rainfall forecasting. Water Resources Management. 2024 Sep 9:1-20.