WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Rainfall Prediction Using Machine Learning *Algorithms*

V. Sai Mahesh [a], R. Ramya [b], P. Anvesh Vardhan [c], T. Veera Aravinda Kumar [d], Mr. M. Bala Krishna [e]

[a,b,c,d] *UG Students, Department of Electronics and Communication Engineering, GMR Institute of Technology, Rajam, Vizianagaram District, 532127.*

[e] *Assistant Professor, Department of Electronics and Communication Engineering, GMR Institute of Technology, Rajam, Vizianagaram District,532127.*

ABSTRACT:

Rainfall prediction is a critical aspect of weather forecasting, with far-reaching implications for agriculture, disaster management, and water resource planning. This study explores the application of machine learning techniques to enhance the accuracy of rainfall prediction. Utilizing historical meteorological data, this research develops and evaluates predictive models that take into account various meteorological parameters such as temperature, humidity, wind speed, and atmospheric pressure. The proposed machine learning models, including regression and neural networks, are trained on extensive datasets to capture complex relationships between these meteorological variables and rainfall patterns. Through rigorous evaluation and validation procedures, the models demonstrate their ability to provide reliable rainfall predictions on both short and long time scales. This research advances weather forecasting through machine learning, highlighting its potential in refining rainfall prediction models to address climate challenges effectively.

Keywords: Machine learning, Logistic Regression, Decision Tree, Random Forest, Rainfall prediction

INTRODUCTION:

Precise and accessible precipitation estimation is projected to achieve another time of mediation for businesses adversely affected by extreme precipitation events. These sectors include, but are not limited to, energy and agriculture, which are heavily influenced by precipitation patterns. Various scholarly investigations have revealed that both the length and force of precipitation contribute to massive environmental disasters. The effects of these precipitation-related factors include, among other things, dry periods and floods. For example, in 2009, heavy rains impacted about 600,000 people in Senegal, Niger, Burkina Faso, and Ghana. Furthermore, floods in 2007 claimed the lives of about 1,000,000 people in Ethiopia, Uganda, Togo, Niger, Sudan, Mali, and Burkina Faso.

Furthermore, research suggests that the cost of hunger-related child deaths, which increased from 30,000 to 50,000 in Sub-Saharan Africa in 2009, may have been compounded by changes in precipitation patterns and extreme climatic events affecting the agricultural region. Aside from the tragic loss of life due to floods, several investigations have documented the extensive effects of precipitation on basic sections of the Ghanaian economy. It is highlighted that two big hydroelectric power plants, which account for more than 70% of Ghana's electricity interest, are heavily reliant on precipitation. This implies that any drop in precipitation might have severe consequences for the country's power age limit.

Farming, which employs around 44.7% of Ghana's labour population and plays an important role in the country's economy, is crucial for financial growth. Despite its current drop in execution, the horticulture sector remains an important component for poverty reduction and food security in Ghana. Regardless, Ghana's farming area is primarily rain-fed, with roughly 3% of cultivable land supported by a water system. In the meanwhile, precipitation is an important water source for farming, hydropower generation, and other uses in non-industrial countries such as Ghana.

For precipitation forecasting, several order operations such as Random Forest (RF), Decision Tree (DT), Neural Network (NN), K-Nearest Neighbor (KNN), and others have been investigated. The exhibition of these computations normally alters, allowing an opportunity for improvement by varying training and testing ratios or combining different methods. Nonetheless, forecasting precipitation continues to be a difficult issue. As a result, the careful selection of appropriate methodologies for describing precipitation patterns within a certain location is critical. AI computations have emerged as a viable technique to improve the accuracy of precipitation

forecasts. As a result, there has been a proliferation of precipitation forecast research using various approaches in various countries, including Malaysia, India, and Egypt, among others.

In one assessment, for example, AI methodologies were used to create precipitation forecast models for major Australian cities. A number of calculations were examined, including Decision Tree, Random Forest, Logistic Regression, AdaBoost, Gradient Boosting, and K-Nearest Neighbor.

1. RELATED WORKS:

The Paper [1] demonstrates that flood-related fatalities and economic losses have increased in Africa over the past 50 years, prompting the need to identify the causes for this increase. The study finds that intensive and unplanned human settlements in flood-prone areas are a major factor in increasing flood risk, and urgent actions such as discouraging settlements in these areas and implementing early warning systems are needed. Paper [2] focuses on the use of machine learning classifiers for rainfall prediction in Malaysia, comparing the performance of different techniques. The study concludes that the Neural Network (NN) classifier is the most effective for rainfall prediction in Malaysia. In paper [3] , it finds distinct differences in rainfall characteristics and length of the rainy season across different climatic zones in Ghana. The forest and coastal zones have their rainfall onset in March, while the transition zone has its onset from March to April, and the savannah zone has a late onset in April to May. The length of the rainy season varies across zones, with the forest zone having the longest rainy season. The paper [4] compares the onset, cessation, and duration of the rainy season in Ghana using simulated rainfall data from the Regional Climate Model (RegCM4) and rain gauge measurements from the Ghana Meteorological Agency (GMet). The paper [5] mentions the negative consequences of decreasing rainfall on agricultural practices, water resource management, and food security, which is a well-documented concern in the literature on climate change and its impacts on agriculture and food systems. The paper [6] proposes a deep-learning-based classification method for data pages used in holographic memory. The paper focuses on the classification of data pages used in holographic memory using deep learning techniques. The paper [7] presents a conjunction model for drought forecasting that combines dyadic wavelet transforms and neural networks. The paper [8] proposes a kTree method for kNN classification that assigns different optimal k values to different test samples, resulting in higher classification accuracy compared to traditional kNN methods. The paper [9] aims to derive optimal data-driven machine learning methods for forecasting rainfall in Odisha, India, by comparing three techniques: linear regression analysis, random forest method, and Artificial Neural Network (ANN) method. The study found that the maximum rainfall occurred in 1961 (385.3mm) and the minimum in 1974 (197.2mm). The paper [10] describes Understanding and quantifying long-term rainfall variability at regional scale is important for a country like India where economic growth is very much dependent on agricultural production which in turn is closely linked to rainfall distribution.

2. Methodology

2.1 Data Collection :

The initial stage in forecasting rainfall is to collect a dataset comprising numerous weather-related elements which include air pressure, wind speed, atmospheric humidity, temperature, and other important qualities.

2.2 Data Preprocessing:

Once the data has been acquired, it must be prepared. To maintain data quality and consistency, operations such as resolving missing values, correcting outliers, and normalizing feature values are performed. There are some stages in data pre-processing:

2.2.1 Formatting :

It's unlikely that the data we collected is accessible in a way that makes it usable. The information we obtained is presented as raw data. These data were transformed, and a CSV file was created. If you require it for a social database or content record, you might want it to be in a simple document or a restrictive record configuration.

2.2.2 Cleaning :

Information cleaning also includes the elimination of missing information or its completion. There may be instances where the data is lacking or partial, making it impossible for you to fix the issue. It could be necessary to eliminate these cases. Additionally, some of the characteristics could contain assistive data; as a result, it might be necessary to reveal these characteristics.

2.2.3 Sampling :

It can be essential to work with information that is far more scattered than what is readily accessible. Calculations that require more information may take longer to perform and consume more memory and computing resources. You may conduct a smaller agent test on the selected data while analyzing the entire dataset, which might help figuring out and testing the settings more easier.

2.2.4 Label Encoding :

We will need to convert the target variable and the category characteristics to a numerical representation in this. Label encoding transforms the label into a numerical form that is readable by machines. The functionality of these tags can be better understood using Ml. It is a critical supervised learning phase in the structured dataset preparation process.

2.2.5 Feature Scaling :

The process of "feature scaling" distributes the independent features in the data in a predictable way throughout a certain range. It happens while the data is being pre-processed. Additionally, the learning and generalization phases of the ML process are sped up by machine effort and data reduction to create variable combinations (features). Finally, we use the collected named dataset and the classifier approach to train our models using the classify module of the Python Natural Language Toolkit package. To assess the models, the remaining categorize data from our sample will be used. The already processed data was categorized using a few machine learning techniques. The classifiers chosen were Random Forests. These methods are frequently employed in text classification tasks.

2.3 Model Training :

After that, the dataset is divided into training and testing sets. Using the training data, a Random Forest model is trained. To produce predictions, this ensemble learning approach mixes many decision trees.

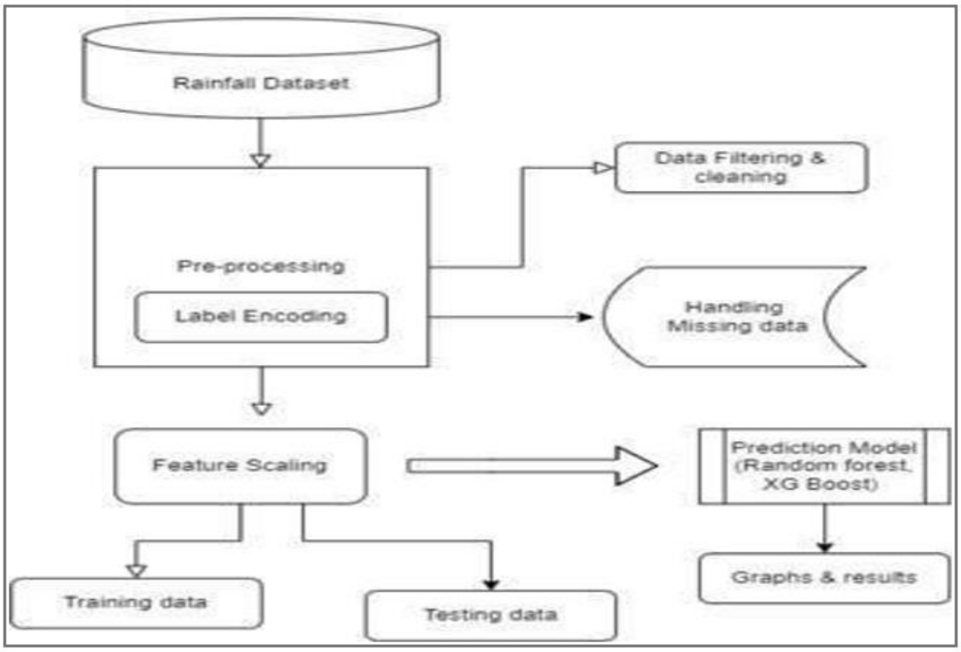
2.4 Model Evaluation :

The testing set is used to evaluate the model's performance. To quantify how successfully the model forecasts rainfall based on the testing data, several evaluation criteria, such as F1 score, Precision, Recall, and Accuracy, may be utilized.

2.5 Result :

When desired accuracy is not obtained, it is important to go back and carefully go over a specific method.

2.6 Architecture Model :



2.7 Applying algorithms :

2.7.1 Logistic Regression :

In order to represent the connection between a dependent variable and one or more independent variables, a statistical method known as logistic regression is used. It is used to determine the likelihood that an event will occur based on predictor factors. Logistic regression may be used to forecast rainfall using a variety of weather-related factors. These predictors are chosen based on their correlation with rainfall, including air pressure, wind speed, and humidity. The logistic regression model is then trained using the obtained data. By examining the values of these factors, a model that can forecast the possibility of rainfall must be developed. The associations between the predictor variables and the incidence of rainfall are assessed using statistical techniques.

2.7.2 Decision Tree :

An effective supervised machine learning technique used for both classification and regression applications is the decision tree. It operates by repeatedly dividing the dataset into subgroups according to the most important attribute, producing a decision tree-like structure. The interior nodes of the tree stand in for features, the branches for decision-making processes, and the leaf nodes for results or class labels.

Based on decision trees, ensemble approaches like Random Forest and Gradient Boosting improve their accuracy and resilience by mixing numerous trees. Decision trees are frequently employed in real-world applications for generating informed and data-driven decisions and serve a crucial role in the basis of more complex machine learning algorithms.

2.7.3 Neural Network :

By training a model to understand the complex correlations between the input factors, such as temperature, humidity, wind speed, and others, and the output variable, precipitation, a neural network may be used to forecast rainfall. By taking into account variables like temperature, humidity, wind speed, cloud cover, and geographical characteristics like elevation and proximity to water bodies, it collects information on rainfall patterns throughout a range of time periods and locales. By addressing missing values, outliers, and anomalies, preprocess the data. To verify that the input variables' value ranges are comparable, normalize them. Following that, the dataset is divided into training and testing sets. The training set will be used to build the neural network, and the testing set will be used to assess it.

2.7.4 Random Forest :

A Random Forest constructs a large number of decision trees, each using a random part of the training data and a random collection of features at each split. The final forecast is then calculated by averaging the predictions of all of these distinct trees. The use of randomization during data and feature selection reduces overfitting and improves the model's resilience. Within a Random Forest, decision trees are constructed independently of one another. Because of this capability, Random Forests are very scalable for huge datasets. Every tree is built with a distinct subset of the data and features, creating diversity and decreasing correlations between trees. This variety helps to increase model generalization.

Random Forest is versatile in that it may be used for both regression and classification applications. It handles categorical and numerical data equally well, giving it a versatile solution for a wide range of prediction applications. Estimation of Feature significance: Random Forest offers estimations of feature significance. This useful data may be used for feature selection and feature engineering, assisting in the identification of the most significant factors in the prediction process. Random Forest is a powerful and versatile machine learning algorithm that may be used for a wide range of applications. Its capacity to reduce overfitting, manage varied data formats, and provide insights on feature relevance makes it a dependable alternative for real-world problem solving across all disciplines.

2.7.5 LightGBM :

LightGBM is a gradient boosting framework developed by Microsoft which uses tree-based learning algorithms. Large datasets and high-dimensional feature spaces make good use of its speed and efficiency design. LightGBM is built on the gradient boosting framework, which develops a powerful predictive model by assembling a group of weak learners (often decision trees). Sequential tree construction is used, with each tree fixing the flaws of the preceding one. LightGBM has been enhanced for speed and effectiveness. It is quicker than many other gradient boosting implementations, especially when working with huge datasets. Histogram-based learning, which uses less memory and hastens training, is one strategy that may be used to accomplish this efficiency. LightGBM is extremely scalable and capable of handling enormous datasets, with millions of instances and features. Big data applications can benefit from its effective algorithms and parallel processing capability. There are several machine learning applications that make use of LightGBM, including click-through rate prediction, picture categorization, recommendation systems, and more. Large-scale and real-time applications can benefit from its speed and efficiency. A significant distinction between LightGBM and other gradient boosting implementations is its unique tree-building approach. Instead of constructing trees level by level, LightGBM builds them in a leaf-wise fashion, selecting split points that maximize loss reduction, resulting in efficient and fast model training.

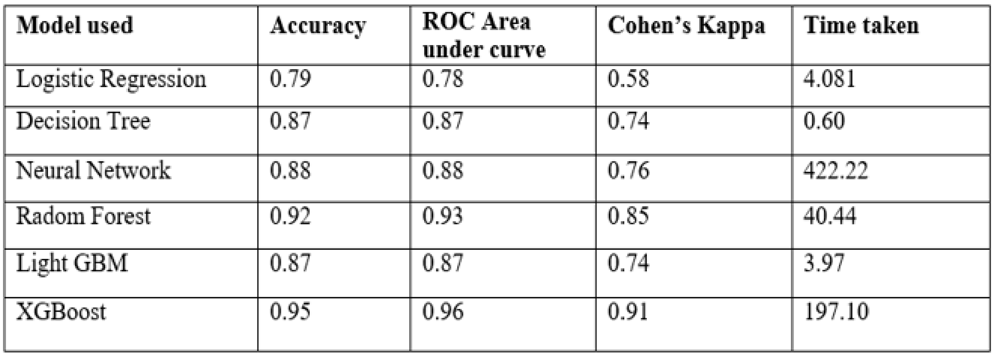
2.7.6 XG Boosting :

XG Boost is a well-liked method for rainfall predicting. In a variety of fields, including hydrology, agriculture, and water management, predicting rainfall is a critical issue. Accurate rainfall forecasting can help with planning and decision-making processes like irrigation scheduling and flood warning systems. The first step in utilizing XGBoost to predict rainfall is collecting historical rainfall data from diverse sources, such as meteorological stations or remote sensing data. This report should contain the amount of rainfall as well as other important elements like temperature, humidity, and wind speed. After it has been collected, the data can be cleaned up and preprocessed, including feature engineering, dealing with missing data, and removing anomalies. By modifying the data and creating new variables, feature engineering may include capturing the fundamental patterns in the data.

Once the data has undergone preprocessing, it may be separated into training, validation, and testing datasets. It is a well-liked option for many machine learning tasks, including as classification, regression, ranking, and recommendation systems, because to its adaptability and good performance. XGBoost, or Extreme Gradient Boosting, is a versatile machine learning algorithm used for both regression and classification tasks. It's known for its exceptional performance in various applications. XGBoost is an ensemble learning method that combines the predictions of multiple decision trees to create a strong predictive model. Each tree corrects the errors made by the previous ones, making the model more accurate. The algorithm optimizes an objective function, which combines a loss function (measuring prediction accuracy) and a regularization term (to prevent overfitting. XGBoost can calculate feature importance scores, helping with feature selection and understanding which features drive predictions. The algorithm is designed for efficiency, allowing for parallel and distributed computing, making it suitable for large datasets and distributed systems.

3. RESULTS AND DISCUSSIONS:

The dataset for rainfall in India, which contains factors like wind speed and atmospheric humidity, among others, is used to assess MI-based techniques including XGBoost, Decision trees, Neural Networks, LightGBM, Random Forest, and Logistic Regression. The primary goal of this project is to utilize current rainfall data to estimate future precipitation from the years 1901 to 2015, and from those predictions, select the best method with the highest level of accuracy. Because of the quick fluctuations in the climate, people cannot predict when it will rain. This erratic weather is killing farmers because it ruins their crops whether there is too much or not enough rain. As a result, the main purpose of the project is to help farmers in a particular region who significantly rely on rainfall. The forecast of rainfall using more suitable factors is a difficult undertaking to do. The accuracy for a decision tree is 85.92, whereas it is 79.514 for a logistic regression. Neural networks' accuracy is 88.40, Random forests' accuracy is 92.90, LightGBM's accuracy is 87.43, and XG Boost's accuracy is 95.79.



Model used	Accuracy	ROC Area under curve	Cohen's Kappa	Time taken
Logistic Regression	0.79	0.78	0.58	4.081
Decision Tree	0.87	0.87	0.74	0.60
Neural Network	0.88	0.88	0.76	422.22
Radom Forest	0.92	0.93	0.85	40.44
Light GBM	0.87	0.87	0.74	3.97
XGBoost	0.95	0.96	0.91	197.10

TABLE 1: COMPARISON TABLE OF USED ALGORITHMS

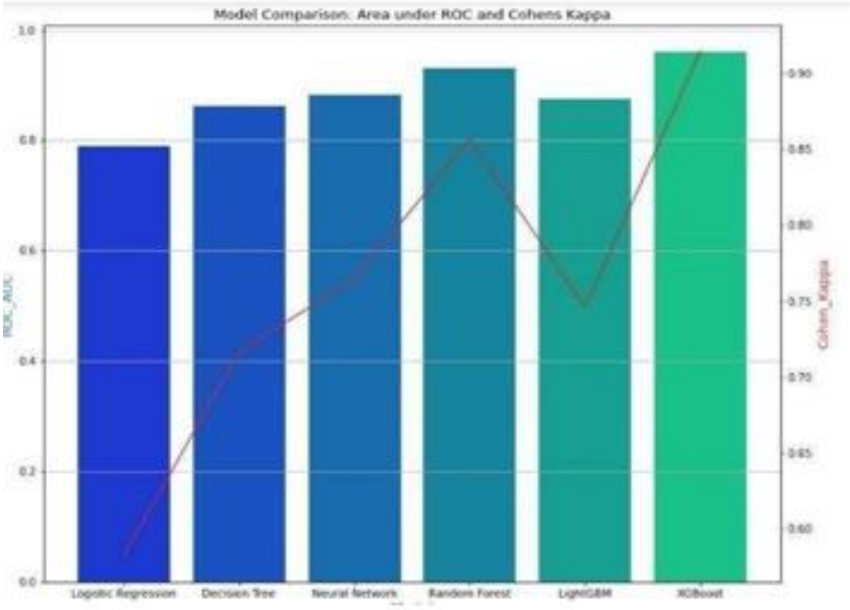


FIGURE 1: Area under ROC and Cohens Kappa

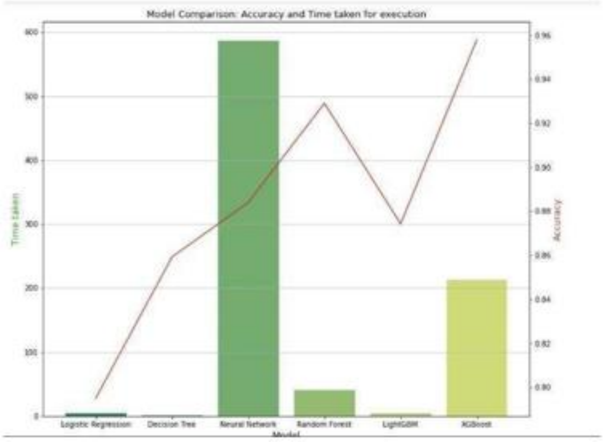


FIGURE 2: Accuracy and Time taken for execution

CONCLUSIONS AND FUTURE SCOPE :

One of the most significant natural occurrences that has an influence on both people and other living things is rain. Rainfall cycles are altering as a result of changing climatic conditions and increasing global temperatures. Keeping an eye on this natural event is crucial since climate change hurts trade, agriculture, and sporadically causes land slips and flooding. Therefore, predicting rainfall has benefits for agriculture, conservation, and public awareness of natural disasters like floods. A system to anticipate rainfall using neural artificial intelligence, which is common in modern technology, is needed to address these problems and provide basic needs.

Researchers should investigate ways to adapt rainfall prediction models to account for changing climatic factors and evaluate the models' efficacy in a range of climatic conditions. creating software to help decisions: Once a reliable prediction model has been developed, it may be applied to the development of decision-making applications for a number of sectors, such as agriculture and water management. Researchers may look into ways to develop user-friendly interfaces and decision-support tools to help stakeholders make good decisions based on the expected rainfall.

***REFERENCES*:**

1. G. Di Baldassarre, A. Montanari, H. Lins, D. Koutsoyiannis,L. Brandimarte, and G. Blöschl, ‘‘Flood fatalities in Africa: From diagnosis to mitigation,’’ Geophys. Res. Lett., vol. 37, no. 22, pp. 529–546, Nov. 2010.

2. N. SamsiahSani, I. Shlash, M. Hassan, A. Hadi, and M. Aliff, ‘‘Enhancing Malaysia rainfall prediction using classification techniques,’’ J. Appl. Environ. Biol. Sci., vol. 7, no. 2S, pp. 20–29, 2017.

3. L. Amekudzi, E. Yamba, K. Preko, E. Asare, J. Aryee, M. Baidu, and S. Codjoe, ‘‘Variabilities in rainfall onset, cessation and length of rainy season for the various agro-ecological zones of Ghana,’’ Climate, vol. 3, no. 2, pp. 416–434, Jun. 2015.

4. C. Mensah, L. Amekudzi, N. Klutse, J. Aryee, and K. Asare, ‘‘Comparison of rainy season onset, cessation and duration for Ghana from RegCM4 and GMet datasets,’’ Atmos. Climate Sci., vol. 6, no. 1, pp. 300–309, 2016, doi:10.4236/acs.2016.62025.

5. M. Baidu, L. K. Amekudzi, J. Aryee, and T. Annor, ‘‘Assessment of long-term spatio-temporal rainfall variability over Ghana using wavelet analysis,’’ Climate, vol. 5, no. 2, p. 30, Mar. 2017.

6. T. Shimobaba, N. Kuwata, M. Homma, T. Takahashi, Y. Nagahama, M. Sano, S. Hasegawa, R. Hirayama, T. Kakue, A. Shiraki, N. Takada, and T. Ito, ‘‘Deep-learning-based data page classification for holographic memory,’’ 2017, arXiv:1707.00684.

7. T.-W. Kim and J. B. Valdés, ‘‘Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks,’’ J. Hydrol. Eng., vol. 8, no. 6, pp. 319–328, Nov. 2003.

8. S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, ‘‘Efficient kNN classification with different numbers of nearest neighbors,’’ IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 5, pp. 1774–1785, May 2018.

9. Misra, R. K., Panda, P. K., Sahu, A. K., Sahoo, S., & Behera, D. P. (2021). Rainfall prediction using machine learning approach: A case study for the state of odisha. Indian Journal of Natural Sciences.

10. Mohapatra, G., Rakesh, V., Purwar, S., & Dimri, A. P. (2021). Spatio-temporal rainfall variability over different meteorological subdivisions in India: analysis using different machine learning techniques. Theoretical and Applied Climatology, 145(1-2), 673-686.