Grenze International Journal of Engineering and Technology, June Issue

Improving Rainfall Prediction Accuracy with Ensemble Machine Learning Techniques

Santoshi Deshmukh[1], Prateek Verma[2] and D.B. Bhoyar[3]
[1]Faculty of Engineering and Technology, Datta Meghe Institute of Higher Education and Research, Wardha, Maharashtra, India 442001
Email: santoshideshmukh2003@gmail.com
[2,3]Deptt. Of Electronics and Telecommunication Engineering, Yeshwantrao Chavan College of Engineering, Nagpur
Email: prateekv.feat@dmiher.edu.in

***Abstract*—Climate Change (CC) is drastically changing weather patterns and making extreme weather events more intense. Among these, periods of intense rainfall have increased in frequency and intensity, causing catastrophic effects on infrastructure, agriculture, ecosystems, and human life. Intense precipitation events are made possible by rising global temperatures, and increasing atmospheric moisture. Due to the disruption of hydrological cycles, there is an increased danger of flooding, soil erosion, and imbalances in water resources. (CC) exacerbates the effects of excessive rainfall, making sophisticated forecast techniques necessary to reduce risks and guide disaster management plans. The intricate relationships between climate factors are frequently difficult for traditional forecasting models to consider. Recent developments in ensemble machine learning approaches merge various models to increase forecast accuracy and dependability. These approaches present encouraging answers. These methods improve our knowledge of rainfall patterns and how they relate to shifting climatic circumstances by allowing the integration of historical data, live-streaming weather conditions, and environmental elements. This study emphasizes the pressing need to use cutting-edge technologies to tackle the problems caused by severe rains brought on by climate change, stressing the vital role that precise prediction models play in promoting resilience and sustainable development.**

***Index Terms*— Machine learning, Deep learning, Algorithm, Model, Classification, Forecasting, climate changes, weather.**

I. INTRODUCTION

Climate change (CC) affects human communities, natural systems, and the global economy; these issues occur in a rather thin line balancing the climate on our planet, causing it to experience environmental difficulties following the rising greenhouse gas concentration elevation. One of the most troubling effects is increased Heavy rainfall particularly has increased its frequency and intensity, especially as a result of the atmosphere being warmed to an increased ability to retain moisture. This process of intensifying the precipitation events enhances devastating landslides, floods, and soil erosion. Important industries that get affected include agriculture, where waterlogging causes harm to crops, and urban areas, where an overloaded drainage system worsens the situation. There is a need for a more profound understanding of the link between heavy rainfall and climate change as well as new approaches toward the precise forecasting and mitigation of it [1].

Grenze ID: 01.GIJET.11.2.288_20

Advanced Machine Learning (ML), especially the ensemble models, could facilitate significant power in developing more accurate rainfall predictions towards better adaptation plans and proactive disaster management. Rainfall datasets are an important application of meteorological data analysis in rainfall prediction. High-resolution datasets like GPM are ideal for machine learning models such as Random Forest and Gradient Boosting as they enable accurate analysis of historical and near-real-time rainfall data with their 0.1 spatial resolution and half-hourly temporal resolution. Machine learning is an important factor in improving rainfall prediction as it helps in analyzing vast, diverse, and complicated datasets.

ML works behind patterns, trends, and interactions within weather data to give accurate and useful forecasting. In this regard, machine learning has changed the scenario of rainfall forecasting whereby huge and complex datasets, which are otherwise hard to handle with the usual statistical methods, can be analyzed. These datasets are usually acquired from satellite imaging, IoT sensors, and historical weather records, and machine learning models are good at identifying such patterns and nonlinear correlations within such datasets. Among these include decision trees, support vector machines (SVM), and neural networks. Some studies have put forward weather parameters such as temperature, humidity, wind speed, and air pressure as inputs of neural networks for predicting frequent rainfall instances. Time series models, such as Long Short-Term Memory (LSTM) networks, have taken this further to provide accurate predictions for short-term and long-term allowances, which are useful for establishing relationships at a specific time. Moreover, for the preprocessing of the data, different ensemble techniques used are random forest and gradient boosting, combining two or more models [3-4]. These developments are very important in agricultural applications, disaster management, and urban planning since they make better use of resources, prepare them for extreme weather events, and design flood-resistant infrastructure. Prediction models of rainfall are often applied to specific regions and localized data are used to provide the highest level of accuracy and relevance. For example, machine learning models are commonly used within farming regions to predict patterns in rainfall that are essential for the management of irrigation processes as well as crop planning. Models like Random Forests, Gradient Boosting, and Long Short-Term Memory (LSTM) networks have been vital for studies conducted in places like India, where there is a need for monsoon rainfall. In producing forecasts relevant to a certain place, these models of work are premised on a whole variety of meteorological factors such as temperature, humidity, wind speed, and historical rainfall data [5]. Forecast models of rainfall can be used in enhancing the water draining systems and ensuring that the urban flooding dangers within metropolitan areas, especially flood-prone cities are reduced. Thus, such highly localized rainfall predictions assimilate information from the ensemble models into complex systems to ingest real-time feeds from meteorological stations and IoT-based weather stations [6]. In recent years, there has been much interest in applying the power of ML to overcome the limitations imposed by traditional statistical models for rainfall prediction. In the present study, it is attempted to design models that can process and evaluate large complex datasets including historical weather records, IoT-based weather sensor data, and satellite imagery. Most machine learning algorithms, including Random Forests, Gradient Boosting, and Long Short-Term Memory (LSTM) networks, tend to capture nonlinear relationships and temporal autocorrelations of weather patterns [7].

II. Literature Review

TABLE I: IMPROVING A LITERATURE REVIEW OF RAINFALL PREDICTION

Ref. No	Objective / Module Use Data Set Used	Output	Limitation
[8]	Objective: To better predict rainfall by utilizing a wealth of meteorological data. Module Use: selecting features and prepping data for ensemble learning. Data set Use: Global Precipitation Measurement (GPM) dataset	Rainfall forecasts with high resolution utilizing Random Forest and Gradient Boosting	Restricted to satellite-based observations; microclimatic fluctuations might be overlooked.
[9]	Objective: Boost forecast precision while maximizing resource utilization. Module Use: Random Forest, Gradient Boosting for efficiency. Data Set Use: ERA5 Reanalysis Dataset	Detailed projections using multivariable meteorological data	computationally demanding because of the large-scale and high-dimensional data
[10]	Objective: To develop reliable models that perform well in various climates and geographical areas. Module Use: Cross-validation, Hyperparameter tuning. Data set Use: Indian Meteorological Department (IMD) Gridded Data	precise regional forecasts for the various climate zones in India	limited to the Indian subcontinent; less appropriate for use worldwide

[11]	Objective: Build flexible models that require less input and can be used in several locations. Module Use: Hybrid models combining traditional statistics and ML. Data set use: NOAA Climate Data Online (CDO)	Information about past patterns of precipitation in particular areas	slower updates and lesser resolution than contemporary satellite datasets
[12]	Objective: To combine several data sources (historical, satellite, and Internet of Things) to improve prediction accuracy. Module Use: Data fusion techniques for combining diverse datasets. Data set Use: Tropical Rainfall Measuring Mission (TRMM)	Enhanced comprehension of tropical precipitation	For current data, GPM replaces temporal coverage, which ends in 2015.

The Various datasets are used in Predicting rainfall is essential in fields including agriculture, disaster relief, and water resource management. The accuracy of predictive models is enhanced when they utilize advanced meteorological datasets such as Global Precipitation Measurement (GPM), ERA5 Reanalysis, and regional sources. For instance, GPM data is suitable for machine learning models such as Random Forest because it provides high spatial and temporal resolution and near real-time global precipitation information. Similarly, the ERA5 Reanalysis dataset provides hourly global weather data, which is useful for historical modeling and long-term trend analysis using Gradient Boosting. These datasets capture intricate correlations and patterns in rainfall-related variables, allowing for reliable forecasting[13-14]. These datasets are useful, yet they have drawbacks. For instance, because GPM and TRMM databases are satellite-based, they necessitate a great deal of preprocessing and computer power. IMD and other regional datasets are quite dependable for a given area, but they are not globally applicable[15]. Although the ERA5 and NOAA's Climate Data Online (CDO) datasets offer extensive historical data, they encounter difficulties in areas with little instrumentation. The accuracy of rainfall predictions could be further increased while optimizing the use of these datasets by overcoming these constraints through sophisticated preprocessing methods and hybrid modeling methodologies.

III. METHODOLOGY

Input Data: The data [1901-2015] shows the rainfall at the subdivision level area of Andaman & Nicobar Islands also and Assam as well as how each month and season deviates from the norm[16].
Pre-Processing of Data: In this stage, the data must be cleaned and ready for analysis. This might involve activities like dealing with missing numbers, eliminating anomalies, normalizing data as needed, and guaranteeing data consistency.
Spitting of Data: A training set and a testing set to be derived from the preprocessed data. This is where, after training on the training set, the classification algorithm is put to test by the testing set.

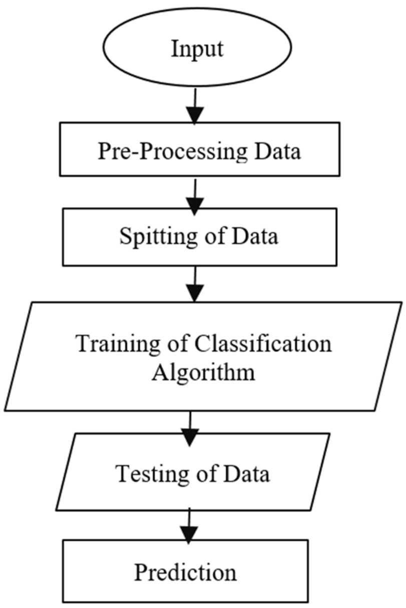

Fig 1: Methodology on Rainfall Prediction Workflow

Training of classification Algorithm: Choose a suitable classification method for the task at hand. Neural networks, support vector machines (SVM), decision trees, and random forests are common techniques used in this kind of study. Utilize the training data to teach the algorithm so that it may discover patterns and correlations in the data.
Data testing: Use the testing data set to check the accuracy of the trained classification model. This is the final step in identifying the strength of how well the algorithm applies to the previously unseen data.
Forecast : Use the model to predict previously unknown or discovered data after you have trained and tested it. In this step, you would use the trained model to make predictions about future rainfall patterns.
Result: Examine the algorithm's predictions compared to the data that was viewed. Analyze accuracy as well as other relevant metrics to view how well the classification system is performing. Look at the data and make conclusions based on your findings. Be sure to note each step in this process with any parameter settings, algorithm choice, and metrics used for assessments. The above documentation ensures the transparency and repeat-ability of your analysis. And, you must also improve your process based on the learnings you acquire through your results, through iteration in the new knowledge.

IV. PROPOSED SOLUTION

- Linear Regression: This is a basic machine-learning technique that can be applied to predict the relationship between a dependent variable and one or more independent variables. It assumes that the target variable and the input characteristics are linearly related. Future results may be predicted since the model learns the coefficients that best suit the data. The equation for linear regression can be given as:
- Lasso Regression: a regularization approach named Lasso regression, or Least Absolute Shrinkage and Selection Operator, which reduces the complexity of a model by penalizing the absolute size of coefficients. Feature selection is achieved using a penalty parameter that shrinks the coefficients of less important characteristics down to zero.
- Ridge Regression Ridge regression is another type of regularization, which eliminates multicollinearity by including a penalty term in the cost function. Unlike Lasso regression, ridge regression penalizes the squared magnitude of coefficients and tends to minimize the contribution of each feature while letting it contribute towards prediction.
- Support vector machine: SVM classifies by seeking the best separation hyperplane of the classes in which to lie. The amount of confidence about the forecast relates directly to the distance from the hyperplane. It therefore must be as large as possible for that gap separating the nearest data point from the hyperplane. To find a best-fit line or decision boundaries, SVM uses a kernel approach. Essentially it maps the low-dimensional input data onto the high-dimensional feature space. Though there isn't any hyperplane between the classes, the hyperplane, which divides the two classes with the maximum gap between them, is chosen as the best one. The optimization problem when SVM occurs during the training time of the Support Vector Machine is achieved by using this iterative approach. The SVM optimization problem for a binary classification problem:
- Random forest: The supervised learning method random forest does produce a classification tree. In this method, an input data vector is inserted into each tree in the forest to classify a new item based on an input feature vector. Each tree in the forest "votes" for other trees; the trees that have the highest number of "votes" are rated the highest. pseudo-code. Randomly select "k" features out of a total of "m" features. To generate some elements, divide the element with the best split. Algorithms can handle classification and regression problems. There is a forest of decision trees, and the more trees there are, the more accurate and reliable the results are.

V. RESULT AND DISCUSSION

The entire globe is always concerned about climate change, and making predictions about it these days is challenging and unexpected. The present trend of global warming and human growth are to blame for climate change. As a result, the sea level is rising, the air and oceans are warming, there is more flooding and drought, etc. Today predicting the amount of rainfall is a difficult problem that most major international bodies take into consideration. Understanding the temporal patterns, trends, and features of the historical rainfall data is essential for performing machine learning-based rainfall prediction for the years [1901–2015]. This is done through the use of data visualization. The historical background meant that, in comparison to modern methods, the visualization approaches and tools accessible at this time were restricted. The following are some streamlined techniques for data visualization that may have been used at that time: To visualize bar charts and line charts that can be applied to a machine learning model. To represent discrete data points such as monthly or yearly rainfall averages so, in the bar chart, there is rainfall in mm & annual rainfall v/s subdivision using the data visualization technique.

TABLE II. COMPARISON OF MACHINE LEARNING ALGORITHMS BASED SUBDIVISIONS

SR.NO	Algorithm	Accuracy Score	Accuracy (%)
1	Linear Regression	0.79	79%
2	Lasso Regression	0.58	68%
3	Ridge Regression	0.62	62%
4	Support Vector Machine	0.51	51%
5	Random Forest	0.65	65%

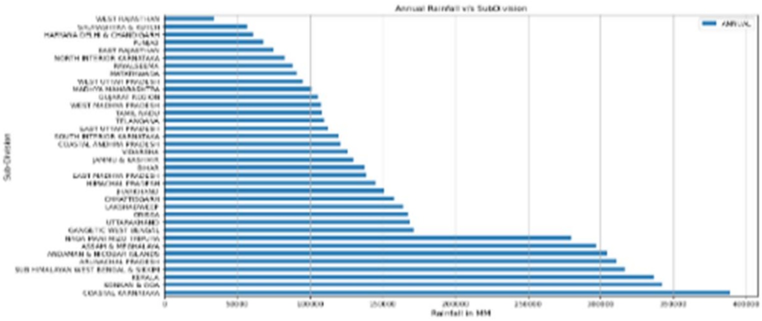

Fig 2. Monthly Rainfall Distribution Across Different

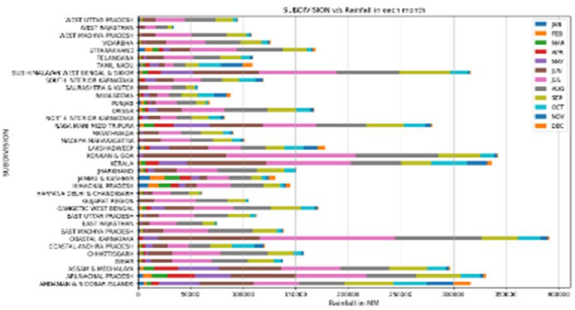

Fig 3. Bar Chart Representing Rainfall Data Across Indian States for Different Months

Fig 4. Correlation Matrix of Rainfall Features

VI. CONCLUSION

In this paper, a supervised rainfall learning model that classified rainfall data using machine learning methods is provided. The accuracy of precipitation prediction was verified using machine learning algorithms such as SVM, Random Forest, Naive Bayes, and MLP, crucial for water resource management and environmental sustainability. Due to the influence of regional and local changes and characteristics on rainfall estimation, it may encounter inaccurate or partial estimation difficulties. In this research, various techniques for rainfall forecasting were reviewed, along with potential issues that may arise when using these approaches. Artificial Neural Networks are the most effective method due to their ability to learn from past nonlinear correlations in rainfall data. There are currently no industries using machine learning. The complexity of the data will rise along with it, which is why we are leveraging machines to help humans interpret the data better. It provides reasonably accurate weather forecasts with a decent accuracy score, and it also provides reasonably accurate rainfall forecasts. We intend to intensify our efforts in storm and crop prediction in the future along with rainfall prediction.

REFERENCES

[1] A. Gupta, H. K. Mall, and S. Janarthanan, Rainfall Prediction Using Machine Learning, no. January. 2022. doi: 10.1109/ICAITPR51569.2022.9844203.

[2] H. Interfaces and M. Systems, Proceedings 2012, no. July. 2012.

[3] S. Manandhar, S. Dev, Y. H. Lee, Y. S. Meng, and S. Winkler, “A Data-Driven Approach for Accurate Rainfall Prediction,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 11, pp. 9323–9330, 2019, doi: 10.1109/TGRS.2019.2926110.
[4] M. Mohammed, R. Kolapalli, N. Golla, and S. S. Maturi, “Prediction of rainfall using machine learning techniques,” Int. J. Sci. Technol. Res., vol. 9, no. 1, pp. 3236–3240, 2020, doi: 10.46647/ijetms.2023.v07i05.006.
[5] A. Parmar, K. Mistree, and M. Sompura, “Machine Learning Techniques For Rainfall Prediction: A Review Machine Learning View project Rainfall prediction using ANN View project Machine Learning Techniques For Rainfall Prediction: A Review,” Int. Conf. Innov. Inf. Embed. Commun. Syst., no. 3, p. 6, 2017.
[6] C. Z. Basha, N. Bhavana, P. Bhavya, and V. Sowmya, “Rainfall Prediction using Machine Learning Deep Learning Techniques,” Proc. Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2020, pp. 92–97, 2020, doi: 10.1109/ICESC48915.2020.9155896.
[7] W. M. Ridwan, M. Sapitang, A. Aziz, K. F. Kushiar, A. N. Ahmed, and A. El-Shafie, “Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia,” Ain Shams Eng. J., vol. 12, no. 2, pp. 1651–1663, 2021, doi: 10.1016/j.asej.2020.09.011.
[8] G. Huffman, D. Bolvin, D. Braithwaite, K. Hsu, R. Joyce, and P. Xie, "Integrated Multi-satellite Retrievals for GPM (IMERG), version 4.4," NASA's Precipitation Processing Center, 2014. [Online]. Available: https://disc.gsfc.nasa.gov/. [Accessed: Mar. 31, 2015].
[9] Copernicus Climate Data Store, "ERA5 Reanalysis Data." [Online]. Available: https://cds.climate.copernicus.eu/.
[10] Indian Meteorological Department, "IMD Gridded Rainfall Data." [Online]. Available: https://imdpune.gov.in/.
[11] National Oceanic and Atmospheric Administration (NOAA), "Climate Data Online (CDO)." [Online]. Available: https://www.ncdc.noaa.gov/cdo-web/.
[12] NASA Earth Science Data, "Tropical Rainfall Measuring Mission (TRMM) Dataset." [Online]. Available: https://disc.gsfc.nasa.gov/.
[13] P. Jiang, S. Ye, D. Chen, Y. Liu, and P. Xia, “Retrieving precipitable water vapor data using GPS zenith delays and global reanalysis data in China,” Remote Sens., vol. 8, no. 5, pp. 2072–4292, 2016.
[14] J. Liu, X. Chen, J. Sun, and Q. Liu, “An analysis of GPT2/GPT2w+Saastamoinen models for estimating zenith tropospheric delay over Asian area,” Adv. Space Res., vol. 59, no. 3, pp. 824–832, 2017.
[15] Kar, Kaveri, Neelima Thakur, and Prerika Sanghvi. "Prediction of Rainfall Using Fuzzy Dataset." (2019).
[16] Business Bliss Consultants FZE, "Rainfall Prediction Using Machine Learning Algorithms," Nov. 2018. [Online]. Available: https://ukdiss.com/examples/rainfall-prediction-machine-learning.php?vref=1