Machine Learning-Based Rainfall Prediction for Various Regions

L. Meenachi [1], S. Ramakrishnan [1], S. Vigashini [1], A. Preethi [1] and R. Kavya [1]

[1] *Dr. Mahalingam College of Engineering and Technology, Coimbatore, Tamilnadu, India*

Abstract

As rainfall intensity varies randomly, rainfall in urban or rural areas can inflict significant harm. Rainfall is one of the most catastrophic natural disasters, yet modelling it is extremely difficult. Additionally, they are non-stationary occurrences that are challenging to describe. As a consequence of combining a parameter estimation model based on a machine learning method, a regression and classification-based actual rainfall forecast. The rainfall dataset is produced ahead of time for a variety of rainfall scenarios using different models. The rainfall depth data for each state and region are divided by year based on the monthly rainfall prediction based on ground truth using the Random Forest (RF) method, which is a classification technique for more recall and precision rainfall prediction. If the measured rainfall is entered, a model is built to predict the appropriate accumulated volume. A proposed system capable of generating a real-time rainfall map is constructed for the temporal widening of the rainfall depth with the predicted reflective quantity by linking each grid's cumulative volume with the representative amount of substance using linear regression. In the proposed Random Forest (RF)algorithm, rainfall accuracy is based on the ground truth. If the measured rainfall is entered, a model is built to predict the original volume. A new methodology able to generate a real-time rainfall map is built for the rainfall depth with the expected typical flow rate by applying a combination of linear and nonlinear regression to link each system's total volume with the representative cumulative volume. Rainfall accuracy based on ground truth is used in the proposed Random Forest (RF) technique.

Keywords

Linear Regression, Random Forest, Classification, Regression, Rainfall, Accuracy, Prediction, Machine Learning.

1. Introduction

The extent and frequency of extreme precipitation are rising as a result of the recent regional effects of climate change, producing regular rainfall destruction to metropolitan areas. In current history, the occurrence of torrential downpours with levels more than 30 mm/h has increased dramatically as compared to the early 80s. Impermeable regions are becoming moreprevalent in both urban and rural settings, particularly in low-cost urban neighbourhoods and locations with extensive underground infrastructure. Nonetheless, the likelihood of damage from urban rain falling is significantly growing because the potential rainfall control adjustments to existing urban districts are restricted. As a result, there is an increasing demandfor forecasting/warning systems to aid in successful decisions with respect to rainfall prediction in urban regions. The accuracy is used to predict the rainfall using the suggested algorithm random forest.

Models for predicting rainfall are very important for determining risk and managingunpredictable situations. The management of water resources benefits greatly from reliable andaccurate prediction, which is usually correlated to a subset of machine learning. But statisticallearning is only one type of machine learning. The study of mathematical optimization provides machine learning with theory,

WINS 2023: Workshop on Intelligent Systems, May 20 – 21, 2023, Chennai, India.
EMAIL: ram_f77@yahoo.com (S. Ramakrishnan)
ORCID: 0000-0002-8224-4812 (S. Ramakrishnan)

methods, and application areas. Unsupervised data analysis through exploratory data mining is the focus of a related field of study. Data and neural networks are used in some machine learning applications in a way that is similar to how a human brain works. The dataset includes average rainfall for each state and sub-region from 1901 to 2015. This model will make predictions of monthly and annual rainfall (in mm). Predictive analytics is another name for machine learning, which is used to solve problems inbusinesses. The dataset is taken from the Kaggle website. After contrasting multiple predictive models, our rainfall forecast system uses the top model. Our programme will forecast rainfall for several states in real-time.

2. Related Work

It is commonly known that rainfall has a significant impact on agriculture and the workers based on the work. As a result, various studies and research are being undertaken globally and locally in every country after that. By testing remotely sensed satellite algorithms, WF, James. A, Smith et al. [1] predicted intense precipitation frequency and prediction of a flood using Radar and several innovative methods. Kumar Abhishek et al. suggested a neural network-based rainfall forecast system. The suggested model in [2] forecasts rainfall in the district of Udupi in Karnataka of India. The use of Back-propagation neural networks in conjunction with cascade neural network feed-forward networks is being investigated. The suggested model uses 70 per cent of the overall data for training and 30 per cent of the total for testing. When compared to Back- propagation neural networks, the recurrent network performs better in terms of accuracy. Back- propagation neural networks have a high MSE.

The paper of Michael F et al. [3] shows the effect of rainfall on multiple aspects such as soil water content, air temperature in the soil, and plant growth, with the result that reducing rainfall while retaining current monthly precipitation averages can minimize CO_2 release from the soil surface, plant biomass yield, and soil moistness. Aswin et al. [4] suggested a model for predicting rainy precipitation through Deep Learning Architectures. During 468 months, Convolution Network Frameworks predicts and model the average rainfall of the globe. The system developed has an RMSE of 2.55 while Convolution Network has an RMSE of 2.44. Problems can still be reduced by raising the convolution layer.

Neural network backpropagation networks for rainfall prediction have been explored by Xianggen Gan et al. [5] The model developed was examined using a dataset with 16 meteorological characteristics from 1970 to 2000. While Model training and the learning rate are set to 0.01. The developed model used Mat lab using the platform of neural network. Mohapatra et al. [6] presented rainfall forecast for Bangalore, India data from 1901 to 2002. Linear Regression was applied to investigate the collected data. Pandas and sk-learn Learn were applied to validate and obtain computational results. K fold was used to forecast rainfall for different seasons. Predictions for the rainy season were more realistic than forecasts for the summer season. Folorunsho Olaiya [7] used Random Forests, Artificial networks and Decision Tree methods to estimate temperature, rainfall, and wind speed for the city of Nigeria from 2001 to 2009.In their work, Brath, Montanari, et al. [8] investigated rainfall prognosis using historical data analysis approaches. They developed the K-Nearest Network technique and Artificial neural networks. This technique greatly improved flood prognosis and several scholars based their timely models. Lim et al. presented the neural network model as well as bridging [9].

L Meenachi, S Ramakrishnan In this paper t is to develop the classification model from reduced features and instances [10]. L Meenachi, S. Ramakrishnan, This paper is to predict cancer from microarray gene expression data by proposing two feature selection algorithms, namely (1) differential evolution with fuzzy rough set feature selection and (2) ant colony optimization with fuzzy rough set feature selection algorithms [11].

Loganathan Meenachi, Srinivasan Ramakrishnan, In this Letter, the authors propose a genetic search fuzzy rough (GSFR) feature selection algorithm, which is hybridized using the evolutionary sequential genetic search technique and fuzzy rough set to select features[12].

L. Meenachi, S. Ramakrishnan, This paper aims to discover cancer prediction from the microarray gene expression data using the selected features. The meta-heuristic search algorithms select the global and local optimal features using population and neighbourhood- based algorithms [13].

Rainfall forecasting model built around sensor network using the technique of data mining. The system built utilizing rainfall prediction algorithms helped to reduce the risk, recommendations of policy, the decreases the losses of human life, and the disasters caused by rainfalls. The existing Processes of rainfalls, for the past two decades using SVM and Linear Regression. These models are important for disaster prediction and planning events for various processes like agriculture. A reliable and precise forecast has a crucial benefit in saving the water resources methods, schemes and policy recommendations and analyses. To mitigate the disaster of properties, the need for technologies using modern algorithms for immediate and long processes and planning. Thus the forecast of rainfall and other hydrological phenomena is underlined.

The results suggest that Boosted Decision Tree Regression is the highest regression designed for rainfall prediction, with the largest coefficient for determination, while the overall performance of the model of the Random Forest is good and it is used for rainfall prediction. The Random Forests algorithm examines the outcome based on the predictions of various testing and training ratios. Random Forests predictions by averaging the results of several trees.

3. Materials and Methods

3.1. Dataset Description

From 1901 to 2015, the dataset contains rainfall by states of India on a monthly basis until its end. The dataset covers rainfall figures for the months of October, November, and so on in India. The information also includes rainfall totals for particular months such as January, February, and April. The rainfall data in the collection are from various subdivisions of the country where weather trends are observed and data is gathered. As seen in Fig. 1 the data indicated the distribution of rainfall in both seasons and monthly durations.

In terms of the correlation between variables, Fig. 2 shows a significant correlation betweensome months such as January to March and April to December, and a very low connection between January to July and February to October, as they are contrary in nature in terms of rainin the states.

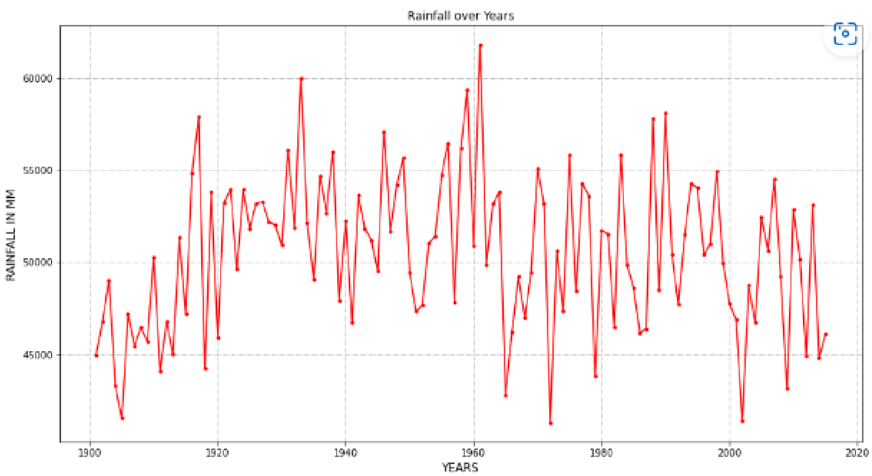

Figure 1: Rainfall distribution over months

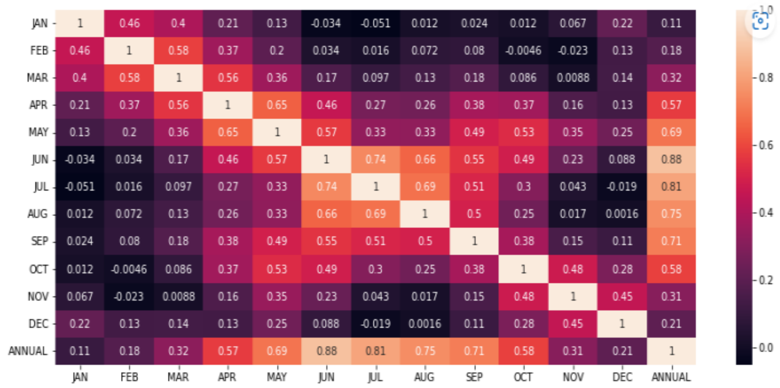

Figure 2: Monthly Correlation

3.2. Data Preprocessing

A data mining approach called data pre-processing includes organising the original data in a meaningful way. Actual data are frequently erroneous and missing in particular traits or patterns. It is also often inconsistent. Pre-processing data is an attempted method for resolvingto solve these issues. In the real world, data are generally inaccurate because they lack variables, or desirable qualities, or represent just collective knowledge. The proposed data pre-processing is used to check column data values and remove noisy data.

3.3. Data Feature Extraction

Object recognition is a step in the process of developing a predictive model that involves reducing the number of input variables. In certain scenarios, reducing the input parameters canimprove model performance while also lowering computational costs. Feature extraction is selecting the ones that will most certainly affect your predictor factor When there are fewer errors in the data, the modelling's accuracy improves. PCA (Principle Component Analysis) is used for Feature extraction. While applying PCA, we receive their initial data as a source and then target toward finding an arrangement of the characteristics of the input which most effectively describes the initial arrangement of data such that the initial parameters are reduced. PCA does that by increasing variants and reducing errors during reconstruction through investigating pairing wised similarities.

3.4. Modelling

Machine learning was used to forecast the amount of rain in this study. Using input variablesthat were marginally and significantly related to rainfall, three approaches to machine learning Support Vector Machine (SVM), Linear Regression, Random Forest and gradient descent boosting were investigated. The superior machine learning approach was found and presented based on the performance metrics RMSE and MAE. Random Forest (RF) is a simple approachin machine learning. The most frequently used metric to represent the RF method is median accuracy. The model's performance is significantly influenced by elements like the ground truthvalue, distance calculations, and the selection of suitable predictors. As input data, the classification and regression rainfall predictive algorithm collect information from regions, rainfall, and typical rainfall volume.

3.5. Performance Measuring

The accuracy of the machine learning algorithm introduced in this study's rainfall prediction was examined through Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), in order to determine which methods outperformed others (MAE). The RMSE and MAE statistics remain some of the most regularly used statistics to assess the correctness of time series or variables. The Mean Absolute Error calculates the average degree of error over a series of estimates and observations while accounting for them.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right| \qquad (1)$$

The nonlinear rating approach known as root means square error is utilized to calculate the average amount of mistakes (RMSE). It's the average of the squared variance among the actual and expected values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (2)$$

4. Proposed Methodology

The efficient installation of a rainfall surveillance and forecasting system is difficult since it necessitates dependability as well as the accessibility of pertinent information. The recommended machine learning algorithm of Random Forest is used to effectively and precisely estimate rain falling and river basin circumstances. A rainfall dataset was put togetherbased on rainfall occurrences, it has data on the maximum rainfall range, rainfall regularity, and average rate of rainfall events. Data on rainfall were generated using the given input development approach in order for the Artificial neural network to forecast the rainfall areas in the target areas. The continuous downpour amount was produced as time-analysis data to express the features of rainfall throughout time.

Showers and levels of water are recorded by ground weather stations or relatively new technology such as rockets, and multichannel systems in the preceding year. The workflow of the current proposal is seen in Fig.3. The presented Random Forest (RF) techniqueis utilized to determine the performance of rainfall accuracy.

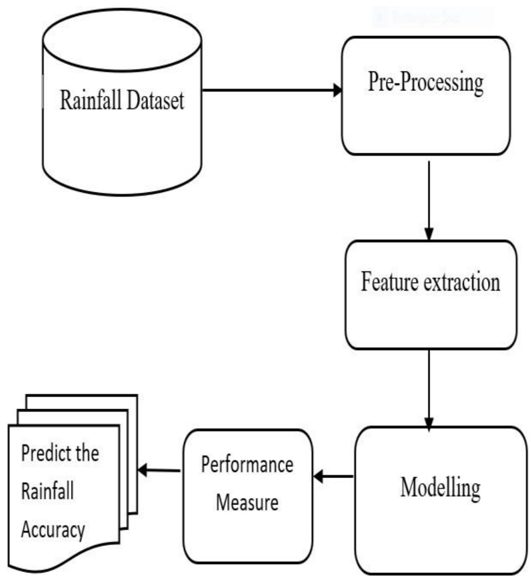

Figure 3: Flow Diagram

5. Result Analysis

5.1. Data Analysis

Because rainfall is the expected output of the models generated, we examined the entire dataset using plots of average monthly rainfall events. The element "Date" serves as an identifier for the prediction models. The month-to-month rain gauge station plots for Tamil Nadu, Rajasthan, and Telangana are given on the y-axis in relation to every season from 1901to 2015 on the x-axis. shown in Figures 4, 5, and 6, respectively. The month-to-month averagerainfall figures are

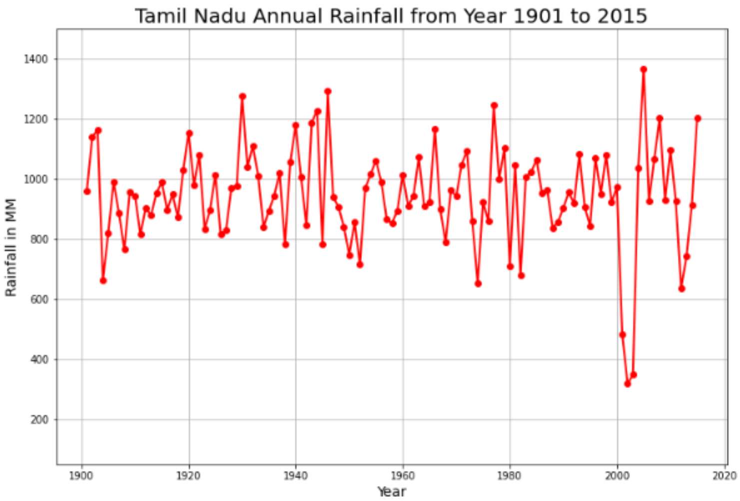

Figure 4: Tamil Nadu Rainfall

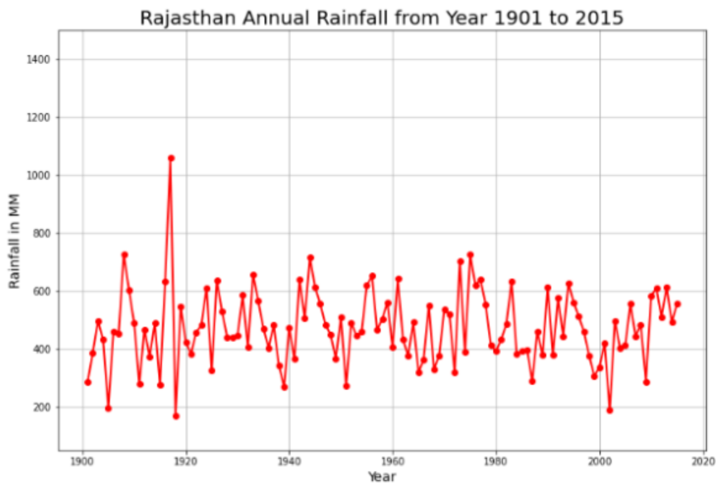

Figure 5: Rajasthan Rainfall

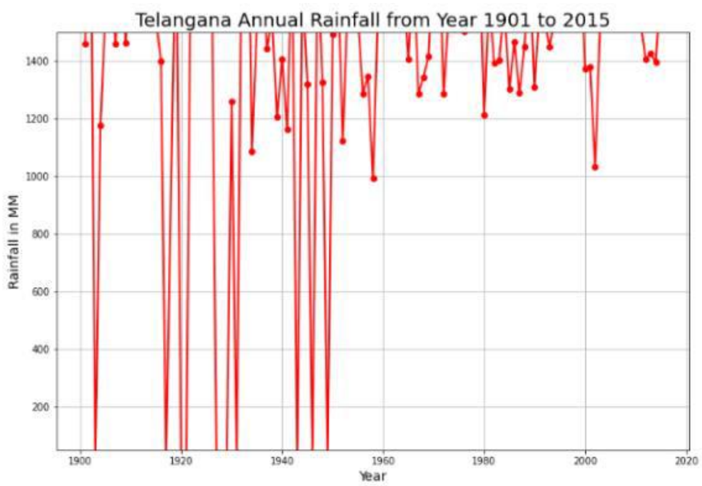

Figure 6: Telangana Rainfall

5.2. Model Accuracy

The testing performance metrics of the Linear Regression model is 33.1 and Training Accuracy is 41.699. The testing Accuracy of the Random Forest model is 42.3 and Training Accuracy is 72.6. Ensemble machine learning approach that employs a large number of classification trees, hence the term random forest. It integrates various decision trees for regression and classification applications during regression calculation. Despite its preference for values with high rates across categorical variables with different levels, the RF algorithm has recently been rated as a very demanding machine learning algorithm for usage.

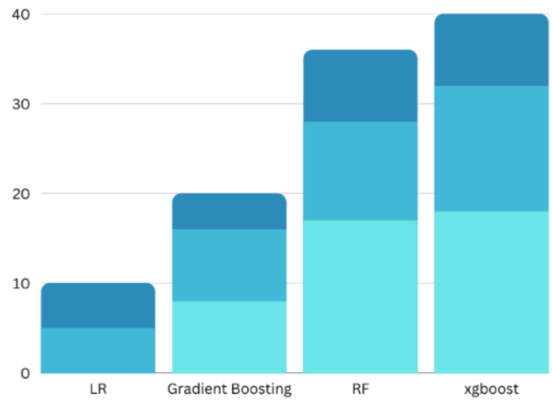

Figure 7: Comparison of algorithms

At the initial stage of the Random Forest algorithm's work, it gathers samples at random from the given data. Following that, the result predicted are decided on, and the classification that received the most points is chosen.

Table 1: Accuracy Table

Algorithm	Testing Accuracy	Training Accuracy
Linear Regression	33.1	41.69
Random Forest	42.1	72.6
XgBoost	9.4	96.3
Gradient Boosting	39.4	71.3

6. Conclusion

Real-time rainfall prediction through classification for urban or rural rainfall analysis, a Random Forest (RF) algorithm was used. This approach was then implemented realistically to accurately anticipate rainfall regions in advance. To do this, a classification-based real-time rainfall prediction model was developed using the best input data, which were determined by a correlation and uncertainty analysis of the precipitation and rainfall data. The suggested machine learning-based RF method makes it possible to reduce harm to people's health and property by forecasting rainfall-induced rain falling based on the built-in database.

7. References

[1] T. A. Khan, M. M. Alam, Z. Shahid and M. M. Su'ud, "Investigation of Flash Rainfallson Early

Basis: A Factual Comprehensive Review," in IEEE Access, vol. 8, pp. 19364-19380, 2020, doi: 10.1109/ACCESS.2020.2967496.

[2] S. Miau and W. -H. Hung, "River Rainfalling Forecasting and Anomaly Detection Based on Deep Learning," in IEEE Access, vol. 8, pp. 198384-198402, 2020, doi:10.1109/ACCESS.2020.3034875

[3] L. Ling, Z. Yusop and M. F. Chow, "Urban Rainfall Depth Estimate with a New Calibrated Curve Number Runoff Prediction Model," in IEEE Access, vol. 8, pp. 10915- 10923, 2020, doi: 10.1109/ACCESS.2020.2964898. 70375-70386, 2020, doi: 10.1109/ACCESS.2020.2986090.

[4] M. Khalaf et al., "I oT-Enabled Rainfall Severity Prediction via Ensemble Machine Learning Models," in IEEE Access, vol. 8, pp.

[5] Y. Xu, J. Wang and Y. Yu, "Alarm Event Prediction from Historical Alarm Rainfall Sequences Based on Bayesian Estimators," in IEEE Transactions on Automation Science and Engineering, vol. 17, no. 2, pp. 1070-1075, April 2020, doi: 10.1109/TASE.2019.2935629.

[6] Machine learning techniques for rainfall prediction: A review," by Aakash, Parmar, Kinjal Mistree, and Mithila Sompura. 2017 International Conference on Information, Embedded, and Communication Systems.

[7] Dash, Yajnaseni, Saroj K. Mishra, and Bijaya K. Panigrahi are the authors. "Using artificial intelligence methodologies, rainfall forecast for the Indian state of Kerala." 70 Computers and Electrical Engineering (2018).

[8] D. Karthika, and K. Karthikeyan are the authors. “Analysis of mathematical models for rainfall prediction using seasonal rainfall data” in IEEE Access, doi: 10.1109/ICEEICT53079.2022.9768602.

[9] Veera Ankalu, Vuyyuru, S.Anuradha are the authors.” Rainfall Prediction Using Machine Learning Based Ensemble Model” doi: 10.1109/ISCON52037.2021.9702409

[10] L Meenachi, S Ramakrishnan, Random global and local optimal search algorithm based subset generation for diagnosis of cancer, Curr Med Imaging 16 (2020) 249–261 2020

[11] Loganathan Meenachi, Srinivasan Ramakrishnan, Evolutionary sequential genetic search technique-based cancer classification using fuzzy rough nearest neighbor classifier, Healthcare Technology Letters 5 (2018) 130–135

[12] L. Meenachi, S. Ramakrishnan, Metaheuristic Search Based Feature Selection Methods for Classification of Cancer, Pattern Recognition 19(2021), 108079.doi.org/10.1016/j.patcog. 2021.108079

[13] L Meenachi, S. Ramakrishnan, Differential evolution and ACO based global optimal feature selection with fuzzy rough set for cancer data classification, Soft Computing 24 (2020) 18463–18475

[14] Sunil Kaushik, Akashdeep Bhardawaj, Luxmi Sapra are the authors. “Predicting Annual Rainfall for the Indian State of Punjab Using Machine Learning Techniques” in IEEE Access doi:10.1109/ICACCCN51052.2020.9362742

[15] Kala, Ganesh Vaidyanathan are the authors. “Prediction of Rainfall Using Artificial Neural Network” in IEEE Access doi:10.1109/ICIRCA.2018.8597421