ISSN: 2347-6540
International Journal of Latest Research In Engineering and Computing (IJLREC)
Volume 11, Issue 1, Page No.100-106, 2024
www.ijlrec.com

RAINFALL PREDICTION SYSTEM USING ML

Salim Akhter Ansari[1], Prince Kumar Raj[2], Sonu Kumar[3], KM Shaijal[4],Priyanka Garg[5]
Student, Computer Science and Engineering, Shivalik College of Engineering, Dehradun, Uttarakhand, India
ansarisalim786p@gmail.com[1]
Student, Computer Science and Engineering, Shivalik College of Engineering, Dehradun, Uttarakhand, India
princekumarcpr15@gmail.com[2]
Student, Computer Science and Engineering, Shivalik College of Engineering, Dehradun, Uttarakhand, India
thesonukumar357@gmail.com[3]
Student, Computer Science and Engineering, Shivalik College of Engineering, Dehradun, Uttarakhand, India
shejupathaniauk633@gmail.com[4]
Assistant Professor ,Computer Science and Engineering, Shivalik College of Engineering, Dehradun, Uttarakhand, India, priyanka.garg@sce.org.in

Abstract :- **Rainfall prediction is a crucial aspect of weather forecasting, with significant implications for various sectors. Machine learning techniques have been increasingly employed to enhance the accuracy of rainfall predictions. Studies have focused on identifying atmospheric features influencing rainfall intensity and utilizing various machine learning models such as Logistic Regression, Random Forest, and Multi-Layer Perceptron for rainfall density forecasting. Additionally, research has explored short-term rainfall prediction, extreme rainfall events detection, and the development of quantitative precipitation forecast correction techniques using machine learning. Novel machine learning models like TabNet have been proposed to improve rainfall prediction accuracy by combining satellite observations with machine learning methods. Furthermore, machine learning has been applied to predict rainfall- induced landslides and groundwater levels. Various studies have evaluated different machine learning methods for rainfall prediction, showcasing their effectiveness in enhancing predictive accuracy across different climates. Overall, machine learning has shown promise in improving rainfall prediction models, offering valuable insights for better preparedness and decision-making in response to changing weather patterns.**

Keywords: : Forecasting, short- and long-term data, geophysical, deep learning, sequence prediction, data leakage, baselining, error bars, shuffling, seasonality.

I. INTRODUCTION

Rainfall is a key geophysical parameter that is basic for numerous applications in water asset administration, particularly in the agriculture segment. Foreseeing rainfall can offer assistance supervisors in different divisions to make choices with respect to a range of vital activities such as crop planting, traffic control, the operation of sewer systems, and managing fiascos like dry seasons and surges. A number of nations such as Malaysia and India depend on the farming sector as a major contributor to the economy and as a source of nourishment security. Thus, a precise forecast of rainfall is required to make way better future choices to offer assistance manage activities such as the ones specified before.

Rainfall is one of the most complicated parameters to predict in the hydrological cycle. This is due to the energetic nature of environmental components and arbitrary varieties, both spatially and transiently, in these components. Subsequently, to address arbitrary varieties in rainfall, a few machine learning (ML) methods like artificial neural networks (ANN), k-nearest neighbours (KNNs), decision trees (DT), etc. are utilized in the literature to learn patterns in the dataset to predict rainfall. In this chapter, a survey of past work in the zone of rainfall forecast utilizing ML models is carried out.

A number of related audit papers exist as follows. The authors in centred on investigating studies that utilize ML for flood forecast, which closely takes after rainfall forecast. The authors in centred on the utilize of ML for generic spatiotemporal sequence forecasting. At last, the authors in conducted a study on the use of ML for rainfall forecast: however, the study was restricted to rainfall forecast in India.

This chapter serves as an expansion to the field by looking over recent relevant thinks about centering on the utilize of ML in rainfall forecast in a variety of geographic areas from 2016–2020. After enumerating the strategies utilized to forecast rainfall, one of the imperative contributions of this chapter is to illustrate different pitfalls that lead to an overestimation in model execution of the ML models in different papers. This in turn leads to unlikely buildup and desires encompassing ML in the current writing. It too leads to an unreasonable understanding of the headways in, and picks up by, ML inquire about in this field. It is hence critical to clearly state and illustrate these pitfalls in arrange to offer assistance analysts dodge them.

The rest of this audit is organized as takes after: Area 2 talks about the strategy utilized to overview and audit the writing which characterizes the dialog system utilized in all consequent segments; Segment 3 portrays the information sets utilized; Segment 4 gives a portrayal of the yield objective in the different papers; Areas 5 – 7 portray the input highlights utilized, common strategies of preprocessing and the ML models utilized; Area 8 summarizes the comes about gotten in different considers; and Area 9 at that point gives a dialog of the methods utilized, particularly indicating out the pitfalls said some time recently towards getting over-estimated and unreasonable comes about. The segment that takes after concludes the paper.

METHODOLOGY AND MODEL

This chapter carries out an in-depth review of relevant literature to reveal the practices we took to predict rainfall. The review covers several aspects which relate to the input into, output from, and methods used in the various systems devised in the literature for this purpose. The review specifically focuses on studies that use supervised learning for both regression and classification problems.

The following figure shows the generic structure of supervised ML models.

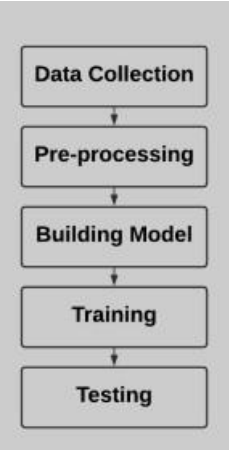

Fig1 . Basic flow for building machine learning (ML) models

Data collection:

For this study, the raw data were collected from the regional meteorological station at Bahir Dar City, Ethiopia. Ten data features such as year, month, date, evaporation, sunshine, maximum temperature, minimum temperature, humidity, wind speed, and rainfall were included. The meteorology station records the values of the environment tal variable every day for each year directly from the devices in the station. Ten, the data were recorded in the Microsoft Excel file tabular format. The year and the days of the month were arranged in the row of tables related to environmental variables in the column of the table. The raw data recorded at the station for 20 years (1999–2018) were used for the study.

Data preprocessing:

The data preprocessing step included the data conversion, manage missing values, categorical encoding, and splitting dataset for training and testing dataset. A total of 20 years (1999–2018) data were collected from the meteorology office. Since the data were raw, they contained missing values, and wrongly encoded values so that the missing values of the target variable were removed and the other features were filled using the mean of the data. Inthe meteorology office, the raw data were also arranged in a year based and the attributes in rows that need to combine and rearrange features in columns. Tus, data were converted from excel data to CSV data. Encoding the dataset was performed and then the dataset was prepared for the experiment. The important features for rainfall prediction were selected and the dataset splitting as 80% for training and 20% for testing were considered as an input for the model.

-Reformate the dataset
-Convert the excel file type to CSV
-Manage the Missing Values
-Relevant Feature Selection
-Splitting Training and Testing dataset
RMSE
-MAE

Model:

In this paper, the rainfall was anticipated utilizing a machine learning strategy. Tree machine learning algorithms such as Multivariate Linear Regression (MLR), Random Forest (RF), and gradient descent XGBoost were analysed

which took input variables having tolerably and strongly related environmental variables with rainfall. The way

better machine learning algorithm was recognized and detailed based on the performance measure utilizing RMSE and MAE (Fig. 2).

Measuring performance:

Pearson correlation was utilized to measure the quality of the relationship between two variables. The two variables can be positively or negatively related and no relationship between the two variables if the Pearson correlation coefficient is zero. The Pearson correlation coefficient model is mathematically depicted as:

$$r_{xy} = \frac{\sum_{i=n}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where rxy is the Pearson correlation coefficient, {(x1, y1), (x2, y2), …, (xn, yn)} are paired data comprising of n pairs and x and y are mean of x and y respectively. To appear the significant highlights of the natural factors to predict every day rainfall intensity, the taking after Pearson coefficient ranges and interpretations are utilized as appeared in Table 1.

The machine learning algorithms take the input data highlights which are chosen utilizing the Pearson correlation coefficient as relevant features. the rainfall prediction execution of each machine learning algorithm that was utilized in this study was measured utilizing Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to compare which machine learning algorithms outflank way better than others. RMSE and MAE were two of the most common metrics utilized to measure precision for continuous variables. the MAE measures the normal magnitude of the errors in a set of forecasts and the comparing perception, without considering their direction.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Table 1. Pearson coefficient ranges and interpretations

Pearson coefficient |r|	Interpretation
0.00 < 0.10	Negligible
0.10 < 0.20	Weak
0.20 < 0.40	Moderate

Relatively strong
Strong
Very strong

The RMSE is a quadratic scoring rule which measures the average magnitude of the error. It is the square root of the average of squared differences between forecast and real observation.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

RMSE gives a generally high weight to huge errors. This implies the RMSE is most valuable when huge errors are especially undesirable. the MAE and the RMSE can be utilized together to diagnose the variety in the errors in a set of forecasts. The RMSE will continuously be larger or equal to the MAE; the greater contrast between them, the more prominent the variance in the individual errors in the test. If the RMSE=MAE, at that point all the errors are of the same magnitude.

III. RESULTS AND DISCUSSION

The fundamental objective of this study was to recognize the important atmospheric features that cause rainfall and foresee the intensity of day-by-day rainfall utilizing machine learning strategies. Subsequently, the research findings are summarized below. To select the environmental variables that connect with the rainfall, the Pearson correlation was analysed on the environmental variables displayed in Table 1 previously. Since the dataset is huge, the variables that relate more noteworthy than 0.20 with rainfall were considered as the member environmental features to the test for rainfall prediction. Subsequently, to foresee the amount of every day rainfall, the comes about of environmental properties relevant to every day rainfall forecast like Evaporation, Relative Humidity, Sunshine, Maximum Every day Temperature, and Least Every day Temperature are appeared in Table 2. The Pearson Correlation coefficient exploratory results on the given data appeared that the attributes such as year, month, day, and wind speed had no critical affect on the forecast of rainfall. This paper took environmental values which had a correlation coefficient more prominent than 0.2 and examined the rainfall prediction. The exceedingly correlated environmental features for rainfall prediction were relative humidity and the everyday daylight which measured the Pearson coefficient of 0.401 and 0.351 separately.

Table no. 2 ENVIRONMENTAL FEATURE AND THEIR PEARSON COEFFICIENT VALUE

Features	|r|
Year	0.012
Month	0.101
Day	0.017
Evaporation	0.279
Relative	0.401
humidity	0.296
Max daily temperature	0.204
Min daily temperature	0.351
Sunshine	0.046

Table 3. Performance Measurements

Algorithms	MAE	RMSE
Random forest	4.49	8.82
MLR	4.97	8.61
XGBoost	3.58	7.85

The machine learning model utilized the chosen environmental features as an input for the algorithms. The regression models were enforced in python and the performances of the MLR, RF, and XGBoost were scaled applying MAE and RMSE. In Table 3 before, the comparison of results of the three algorithms similar as the MLR, RF, and XGBoost was produced. The performance results denoted that XGBoost Gradient worthy outperformed MLR and RF. The MAE and RMSE values of the XGBoost gradient descent algorithms were3.58 and7.85 independently so that The XGBoost algorithm foretold the rainfall utilizing applicable selected environmental features better than the RF and the MLR.

The environmental features utilized in this study taken from the meteorological station re-collected by scaling devices are analysed their applicability on the impact of rainfall and selected the applicable features grounded on experiment outcome of Pearson correlation values as displayed in Table 2 for the day-to-day rainfall forecast. This paper took environ internal features which had a correlation measure higher than0.2 and analysed the rainfall forecast. also, Manandhar et al. (7) identifies the five important environmental features like as Temperature, Relative Humidity, Dew Point, Solar Radiation, precipitable water vapor utilizing a point of correlation among each point. According to the experiment result of the study, a high negative correlation coefficient of around −0.9 is seen between Temperature and Relative Humidity. The researcher Prabakaran et al. (15) utilized the year, temperature, cloud cover and year attribute for the experiment without analysing the relation between environmental features, and Gnanasankaran and Ramaraj, (14) didn't display the impact of environmental features on rainfall rather utilized the monthly and yearly rainfall data to forecast the average yearly rainfall.
This study utilized the applicable environmental feature to train and test the three machine learning Algorithms similar as RF, MLR, and XGBoost for the day-to-day rainfall amount forecasting. The performance of these machine learning Algorithms was scaled applying MAE and RMSE. The RAM of RF, MLR, XGBoost are 4.49,
4.97, and 3.58, and the RMSE is 8.82, 8.61, and 7.85 independently. also, the researcher Manandhar et al. (7) utilized data- driven machine learning algorithms to forecast the yearly rainfall utilizing the selected relevant environmental features and jotted an overall accuracy of 79.6 percent. The researcher viewed the attributes to forecast the measure of annual rainfall amount by taking the average value of temperature, cloud cover, and rainfall for a year as an input. The correlation analysis between attributes was not assessed. The average error chance of the annual rainfall forecast utilizing modified linear regression was 7 percent. The researcher Gnanasankaran and Ramaraj(14), didn't display the impact of environmental features on rainfall. The research held the monthly and yearly rainfall for the forecast of rainfall and measures the performance utilizing RMSE which was 0.1069 and MAE which was 0.0833 utilizing multiple linear regression.
Hence, this study imposed the impact of environmental features on the everyday rain fall intensity utilizing the Pearson correlation and selected the applicable environmental variables. The relevant features are used as an input for the everyday rainfall amount forecast machine learning models and the performance of the models are scaled utilizing MAE and RMSE.

IV. CONCLUSION

Rainfall forecast is the operation area of data science and machine learning to forecast the state of the atmosphere. It's important to forecast the rainfall intensity for productive use of water resources and crop production to reduce mortality due to food and any complaint caused by rain. This paper analysed various machine learning algorithms for rainfall forecast. Tree machine learning algorithms similar as MLR, FR, and XGBoost were presented and tested utilizing the data collected from the meteorological station at Bahir Dar City, Ethiopia.

The applicable environmental features for rainfall forecast were named utilizing the Pearson correlation coefficient. The chosen features were utilized as the input variables for the machine learning model utilized in this paper. A comparison of results among the three algorithms(MLR, RF, and XGBoost) was formed and the results

displayed that the XGBoost was a more- suited machine learning algorithm for everyday rainfall amount forecast utilizing selected environmental features. The accuracy of the rainfall amount forecast may up if the sensor data is incorporated for the study. But the sensor data wasn't considered in this paper.

The Rainfall forecast accuracy can be bettered using sensor and meteorological datasets with additional different environmental features. Hence, in future work, big data analysis can be utilized for rainfall forecast if the sensor and meteorological datasets are utilized for the everyday rainfall amount forecast study.

Abbreviations:

XGBoost: Extreme Gradient Boosting; MLR: Multivariate Linear Regression; RF: Random Forest; RMSE: Root Mean Squared

Error; MAE: Mean Absolute Error; SVM: Support Vector Machine; DT: Decision Tree.

Acknowledgements:

We gratefully acknowledge the North West of Ethiopia Meteorology Agency for providing meteorological data, valuable information, and kind help for the completion of this study.

Authors' contributions

CML designed and coordinated this research, drafted the manuscript, and experiment. CML and HAM carried out the data collection and data analysis. Both the authors read and approved the final manuscript.

Funding

There are no funding organizations or individuals.

Availability of data and materials

The raw data collected from the North West of Ethiopia Meteorology Agency is available for researchers if it is requested and the materials that the authors used are available at the authors' hands

REFERENCES

[1] Ehsan MA. Seasonal predictability of Ethiopian Kiremt rainfall and forecast skill of ECMWF's SEAS5 model. Climate Dynamics. 2021; 1–17.

[2] Kusiak A, Verma AP, Roz E. Modeling and prediction of rainfall using radar reflectivity data: a data-mining approach. IEEE Trans Geosci Remote Sens. 2013; 51:2337–42.

[3] Chowdari KK, Girisha R, Gouda KC. A study of rainfall over India using data mining. In 2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT). IEEE: New York. 2015; pp. 44–47.

[4] Namitha K, Jayapriya A, SanthoshKumar G. Rainfall prediction using artificial neural network on map-reduce frame work. ACM. 2015. https://doi.org/10.1145/2791405.2791468.

[5] Tharun VP, Prakash R, Devi SR. Prediction of Rainfall Using Data Mining Techniques. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). IEEE Xplore. 2018; pp. 1507–1512.

[6] Zainudin S, Jasim DS, Bakar AA. Comparative analysis of data mining techniques for malaysian rainfall prediction. Int J Adv Sci Eng Inform Technol. 2016;6(6):1148–53.

[7] Manandhar S, Dev S, Lee YH, Meng YS, Winkler S. A data-driven approach for accurate rainfall prediction. IEEE Trans Geosci Remote Sens. 2019;5(11):9323–31.

[8] Arnav G, Kanchipuram Tamil Nadu. Rainfall prediction using machine learning. Int J Innovative Sci Res Technol. 2019. 56–58.

[9] Aswin S, Geetha P, Vinayakumar R. Deep learning models for the prediction of rainfall. In 2018 International Conference on Communication and Signal Processing (ICCSP). IEEE: New York. 2018; pp. 0657–0661.

[10] Zeelan BCMAK, Bhavana N, Bhavya P, Sowmya V. Rainfall prediction using machine learning & deep learning techniques. Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020). Middlesex University: IEEE Xplore. 2020; pp. 92–97.

[11] Vijayan R, Mareeswari V, Mohankumar P, Gunasekaran G, Srikar K, (JUNE, Estimating rainfall prediction using machine learning techniques on a dataset. Int J Sci Technol Res. 2020;9(06):440–5.

[12] Chaudhari MM, Choudhari DN. Study of various rainfall estimation & prediction techniques using data mining. Am J Eng Res. 2017;6(7):137–9.

[13] Thirumalai C, Harsha KS, Deepak ML, Krishna KC. Heuristic prediction of rainfall using machine learning techniques. In 2017 International Conference on Trends in Electronics and Informatics (ICEI). IEEE: New York. 2017; pp. 1114–1117.
[14] Gnanasankaran N, Ramaraj E. A multiple linear regression model to predict rainfall using Indian meteorological data. Int J Adv Sci Technol. 2020;29(8):746–58.
[15] Prabakaran S, Kumar PN, Tarun PSM. Rainfall prediction using modified linear regression. ARPN J Eng Appl Sci. 2017;12(12):3715–8.
[16] Balan MS, Selvan JP, Bisht HR, Gadgil YA, Khaladkar IR, Lomte VM. Rainfall prediction using deep learning on highly non-linear data. Int J Res Eng Sci Manage. 2019;2(3):590–2.
[17] Sarker IH. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. SN Comput Sci. 2021;2(6):1–20.
[18] Sarker IH. Machine learning: algorithms, real-world applications and research directions. SN Comput Sci. 2021;2(3):1–21. 19. Srinivas AST, Somula R, Govinda K, Saxena A, Reddy PA. Estimating rainfall using machine learning strategies based on weather radar data. Int J Commun Syst. 2020;33(13):1–11.