Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
DOI: https://doi.org/10.17762/msea.v70i2.2456

Machine Learning-Based Rainfall Prediction: A Comprehensive Review

Ashok Kumar Sahoo
Asst. Professor, Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Article Info
Page Number:1660 - 1669
Publication Issue:
Vol 70 No. 2 (2021)

Abstract
Several industries, including agriculture, water resource management, disaster preparedness, and urban planning, depend heavily on rainfall forecasting. Traditional approaches for predicting rainfall rely on numerical simulations and statistical models, which frequently have accuracy and computing efficiency issues. Growing interest has been shown in using machine learning (ML) algorithms to enhance rainfall prediction as a result of recent developments in ML approaches. In-depth analysis of the methodology, difficulties, and potential applications of machine learning in rainfall prediction is provided in this work.

Article History
Article Received*: 05 September 2021*
Revised*: 09 October 2021*
Accepted*: 22 November 2021*
Publication*: 26 December 2021*

Keywords. Rainfall prediction, Machine learning, Artificial neural networks, Regression models, Classification models, Time series forecasting, Data preprocessing, Feature selection, Performance evaluation

I. Introduction:

In addition to being an essential part of the Earth's climate system, rainfall is also crucial for managing water resources, planning for disasters, and many other facets of human life. For these industries to make informed judgements, accurate and dependable rainfall forecasting is essential. The accuracy and computing efficiency of traditional techniques to rainfall prediction, such as statistical models and numerical simulations, are constrained by their very nature [1].

However, the subject of rainfall prediction has significantly changed with the introduction of machine learning (ML) approaches. By using massive datasets and identifying intricate patterns and correlations, ML systems have demonstrated considerable promise in increasing the accuracy of rainfall forecasts [2]. Since ML models can automatically learn from data and generate predictions without having explicit rules put in, they are ideally equipped to handle the difficulties involved in predicting rainfall.

The goal of this research is to present a thorough analysis of machine learning's use in predicting rainfall. It will examine several ML algorithms, their operating principles, and their effectiveness in forecasting rainfall [3]. The paper will also go through the difficulties encountered when using ML to predict rainfall and identify potential future research and development initiatives.

The many ML methods utilised for rainfall prediction [4], such as regression-based approaches, classification-based approaches, and time series forecasting techniques, will be covered in more detail in the sections that follow. The critical elements of data collection, preprocessing, and feature

selection in rainfall prediction will also be covered in this work. It will also offer assessment criteria that are frequently used to rate how well ML models forecast rainfall [5].

The difficulties with ML-based rainfall prediction will then be discussed, including the lack of readily available data, the need to account for geographical and temporal correlations, and the need to make ML models interpretable. It will also look at possible future research avenues, such as ensemble techniques, hybrid methods integrating physics-based models with ML, and transfer learning.

The paper will present case studies in diverse fields to show how ML-based rainfall prediction is used in real-world scenarios. These fields include projecting agricultural yields, managing water resources, predicting floods, and urban planning. The objective is to demonstrate how ML can help these industries directly by making precise and timely rainfall forecasts.

II. Machine Learning Algorithms for Rainfall Prediction

Machine learning (ML) algorithms have emerged as powerful tools for rainfall prediction, offering the ability to capture complex patterns and relationships in large datasets. In this section [5][7], we will discuss several ML algorithms commonly employed in rainfall prediction tasks.

2.1. Regression-based Approaches:

Regression-based algorithms aim to predict the continuous value of rainfall based on input features. Some commonly used regression algorithms for rainfall prediction include:

Linear Regression: Linear regression models establish a linear relationship between input features and rainfall. They assume a linear correlation between the predictors and the target variable.

Polynomial Regression: Polynomial regression models capture nonlinear relationships by introducing polynomial terms into the regression equation, allowing for more flexible fitting of the data.

Support Vector Regression (SVR): SVR utilizes support vector machines to perform regression. It seeks to find the optimal hyperplane that maximizes the margin while minimizing the error between predicted and actual rainfall.

Random Forest Regression: Random forest regression builds an ensemble of decision trees and averages their predictions to obtain the final rainfall prediction. It can capture nonlinear relationships and handle high-dimensional feature spaces.

Artificial Neural Networks (ANN): ANNs are composed of interconnected nodes or neurons that mimic the structure of the human brain. Multilayer perceptron (MLP) neural networks have been widely used for rainfall prediction, where the input features are fed into the network, and the output layer predicts the rainfall.

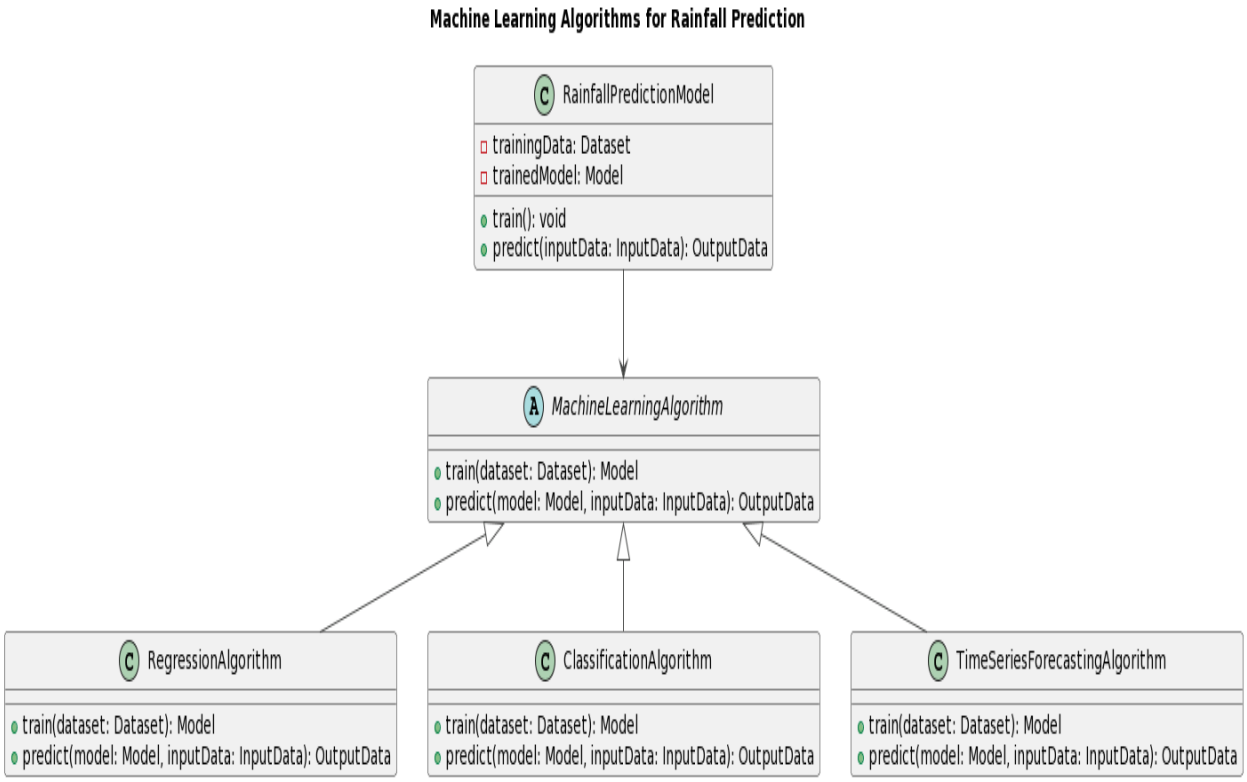

Figure.1 Machine Learning Algorithms for Rainfall Prediction

2.2. Classification-based Approaches:

Classification algorithms categorize rainfall into discrete classes (e.g., no rainfall, light rainfall, heavy rainfall) based on input features. Some commonly used classification algorithms for rainfall prediction include:

Decision Trees: Decision trees split the feature space based on specific conditions to create a tree structure for classification. Each leaf node represents a class label, indicating the predicted rainfall category.

Random Forest: Random forest is an ensemble of decision trees. It combines multiple decision trees to make predictions and utilizes the majority voting or averaging of the individual tree predictions to determine the final rainfall category.

Gradient Boosting Methods: Gradient boosting algorithms, such as Gradient Boosting Machines (GBM) and XGBoost, iteratively build a sequence of weak learners (decision trees) to improve the prediction accuracy. Each subsequent tree focuses on correcting the errors made by the previous trees.

K-Nearest Neighbors (KNN): KNN assigns a new instance to a class based on the majority class of its k nearest neighbors in the feature space. It relies on the assumption that instances with similar features tend to have similar rainfall categories.

Naive Bayes: Naive Bayes is a probabilistic classification algorithm that assumes independence among the input features. It calculates the probability of each class given the feature values and selects the class with the highest probability as the predicted rainfall category.

2.3. Time Series Forecasting Techniques:

Time series forecasting techniques are specifically designed to handle sequential data, where the order of observations is important. In rainfall prediction, time series models are used to forecast rainfall values at future time steps. Some commonly used time series forecasting techniques for rainfall prediction include:

Autoregressive Integrated Moving Average (ARIMA): ARIMA models capture the linear dependencies and trends in the historical rainfall data. They incorporate autoregressive (AR) and moving average (MA) components, along with differencing to handle non-stationarity in the time series.

Long Short-Term Memory (LSTM): LSTM is a type of recurrent neural network (RNN) that can capture long-term dependencies in sequential data. It is well-suited for rainfall prediction as it can retain information from past observations over extended periods.

Recurrent Neural Networks (RNN): RNNs are designed to process sequential data by maintaining a hidden state that captures past information. They can model temporal dependencies in rainfall data and make predictions based on historical observations.

Convolutional Neural Networks (CNN): CNNs, commonly used in image processing, can also be applied to rainfall prediction. They treat rainfall data as 2D spatial data and use convolutional layers to capture spatial patterns and relationships.

I. Data Acquisition and Preprocessing:

Accurate rainfall prediction relies on the availability of reliable and representative data. In this section [9][10], we discuss the key aspects of data acquisition and preprocessing for machine learning-based rainfall prediction.

3.1. Data Sources:

There are various sources of data that can be used for rainfall prediction:

Weather Stations: Ground-based weather stations provide direct measurements of rainfall at specific locations. They are often equipped with rain gauges or pluviometers to collect precipitation data.

Satellites: Satellite-based remote sensing platforms, such as those equipped with passive microwave sensors, can provide wide-area coverage of rainfall data. These sensors estimate rainfall by measuring the microwave radiation emitted or scattered by precipitation particles.

Radars: Weather radars use electromagnetic waves to detect and track precipitation in real-time. They provide high-resolution rainfall data, capturing spatial and temporal variations.

3.2. Data Quality Issues:

Rainfall data can be affected by various quality issues that need to be addressed during preprocessing:

Missing Data: Rainfall data often have missing values due to sensor failures, data transmission issues, or other factors. Techniques such as data imputation or interpolation can be used to fill in the missing values.

Outliers: Outliers can occur due to measurement errors or extreme weather events. Outliers need to be identified and handled appropriately to avoid their influence on the model's training and predictions.

Temporal and Spatial Resolution: Rainfall data may have different temporal and spatial resolutions depending on the data source. It is essential to harmonize the data to a consistent resolution for effective analysis and modeling.

3.3. Feature Selection and Extraction:

In rainfall prediction, selecting relevant features and extracting meaningful information from the data is crucial. Some common techniques for feature selection and extraction include:

Statistical Measures: Various statistical measures such as mean, maximum, minimum, and standard deviation can be calculated from the rainfall data to capture the data's central tendencies and variability.

Time-Based Features: Time-based features such as day of the week, month, or season can provide additional temporal context to the model.

Spatial Features: If multiple weather stations or radar data are available, spatial features such as distance to the nearest weather station or average rainfall in the surrounding area can be extracted to capture spatial correlations.

Lagged Features: Lagged features involve including past rainfall values as predictors in the model. This allows the model to capture temporal dependencies and trends in the data.

3.4. Data Preprocessing Techniques:

Data preprocessing is essential for ML algorithms to effectively learn from the data. Some common preprocessing techniques for rainfall prediction include:

Normalization: Normalizing the input features to a common scale can help prevent features with larger magnitudes from dominating the model's training process.

Data Splitting: The rainfall dataset is typically divided into training, validation, and testing sets. The training set is used to train the model, the validation set helps in model selection and hyperparameter tuning, and the testing set evaluates the model's performance on unseen data.

Feature Scaling: Scaling the input features, such as using techniques like min-max scaling or standardization, can ensure that all features contribute equally to the model's learning process.

Handling Imbalanced Data: In classification problems where rainfall categories are imbalanced, techniques like oversampling or undersampling can be used to address the class imbalance and prevent bias towards the majority class.

Data acquisition and preprocessing play a crucial role in ensuring the quality and suitability of data for machine learning-based rainfall prediction. Careful consideration of data sources, addressing data quality issues, and extracting relevant features are essential steps in preparing the data for training ML models. Proper preprocessing techniques enhance the performance and accuracy of the rainfall prediction models.

II. Performance Evaluation Metrics:

Evaluating the performance of machine learning models for rainfall prediction is crucial to assess their accuracy and reliability. In this section [11][12], we discuss commonly used evaluation metrics for assessing the performance of rainfall prediction models.

4.1. Mean Absolute Error (MAE):

The Mean Absolute Error measures the average absolute difference between the predicted rainfall values and the actual rainfall values. It provides a measure of the average magnitude of errors in the predictions, without considering the direction of the errors. A lower MAE indicates better accuracy.

4.2. Root Mean Square Error (RMSE):

The Root Mean Square Error calculates the square root of the average squared differences between the predicted and actual rainfall values. RMSE penalizes larger errors more than MAE, as it squares the differences. Like MAE, a lower RMSE value indicates better accuracy.

4.3. Correlation Coefficient (R-squared):

The Correlation Coefficient, also known as R-squared, measures the linear relationship between the predicted and actual rainfall values. It indicates the proportion of variance in the actual rainfall values that can be explained by the predicted values. R-squared ranges from 0 to 1, with a higher value indicating a stronger correlation between predictions and actual values.

4.4. Probability of Detection (POD):

POD measures the fraction of correctly predicted rainfall events out of all observed rainfall events. It assesses the model's ability to detect rainfall when it occurs. A higher POD value indicates better detection accuracy.

4.5. False Alarm Rate (FAR):

FAR determines the percentage of all projected rainfall occurrences that were incorrectly forecasted. It displays the frequency of erroneous warnings or incorrectly optimistic predictions. A lower FAR number denotes more accurate rainfall forecasting.

It is essential to remember that different assessment measures put varying emphasis on the performance of the model. While R-squared shows the model's quality of fit, MAE and RMSE shed light on the precision and size of errors. When predicting rainfall using classification, POD and FAR are very useful.

Domain-specific assessment criteria may also be taken into account in addition to these indicators. To assess the practical usability and efficacy of the rainfall prediction models, for instance, agricultural applications may add indicators relating to crop output prediction or water resource management, such as water availability or reservoir levels.

It is important to employ a variety of criteria when assessing rainfall prediction models in order to fully comprehend their performance. A mix of evaluation metrics offers a more thorough review because no one evaluation indicator can fully capture all facets of model performance.

The assessment criteria that researchers and practitioners choose should take into account the precise goals and specifications of the job at hand—predicting rainfall—and they should interpret the findings in light of the application domain.

III. Challenges and Future Directions:

While machine learning-based rainfall prediction has shown promising results, there are several challenges that researchers and practitioners face in this field. In this section, we discuss some of these challenges and potential future directions for advancing the field of machine learning-based rainfall prediction.

5.1. Limited Data Availability:

One of the significant challenges in rainfall prediction is the limited availability of high-quality and long-term rainfall data. Obtaining reliable and extensive datasets can be challenging, especially in regions with sparse weather station coverage. Addressing this challenge requires efforts to improve data collection infrastructure, promote data sharing and collaboration among institutions, and explore alternative data sources such as remote sensing data and citizen science initiatives.

5.2. Incorporating Spatial and Temporal Correlations:

Rainfall exhibits spatial and temporal correlations, where rainfall patterns in one location can be influenced by neighboring regions and previous time steps. Capturing these correlations accurately in machine learning models remains a challenge. Future research can focus on developing advanced techniques, such as spatio-temporal modeling approaches, that effectively incorporate the spatial and temporal dependencies in rainfall data.

5.3. Model Interpretability:

While machine learning models often achieve high accuracy in rainfall prediction, their interpretability can be limited. Understanding the factors and features driving the predictions is crucial for gaining insights into the underlying processes and building trust in the models. Future research should explore techniques for improving the interpretability of machine learning models for rainfall prediction, such as model explainability methods and feature importance analysis.

5.4. Ensemble Methods:

Ensemble methods, which combine multiple models to make predictions, have shown promise in improving the accuracy and robustness of rainfall predictions. Future research can focus on developing ensemble techniques tailored specifically for rainfall prediction, such as combining regression models, classification models, or different time series forecasting methods, to harness the strengths of individual models and mitigate their weaknesses.

5.5. Hybrid Approaches:

Integrating physics-based models with machine learning techniques can potentially leverage the strengths of both approaches. Hybrid models can incorporate physical laws and domain knowledge into machine learning algorithms, improving the accuracy and interpretability of predictions. Exploring hybrid approaches that combine the physics of atmospheric processes with data-driven machine learning methods is an exciting direction for future research.

5.6. Transfer Learning:

Transfer learning, which involves transferring knowledge learned from one task or dataset to another, holds promise in rainfall prediction. Pretrained models or feature representations from related domains, such as weather forecasting or climate modeling, can be used as a starting point to improve the performance of rainfall prediction models. Investigating transfer learning techniques for rainfall prediction can enhance model generalization and reduce the need for large amounts of data.

5.7. Integration with Decision-Making Systems:

To maximize the practical utility of rainfall prediction, there is a need to integrate machine learning-based rainfall prediction models with decision-making systems. This integration can provide timely and accurate rainfall forecasts to stakeholders in sectors such as agriculture, water resource management, and disaster preparedness. Future research should focus on developing frameworks and tools that facilitate the seamless integration of rainfall prediction models into decision support systems.

IV. Conclusion:

By allowing for precise and timely forecasts, machine learning algorithms have revolutionised the field of rainfall prediction. The use of machine learning algorithms in the prediction of rainfall was covered in this work, with a focus on regression-based approaches, classification-based approaches,

and time series forecasting methods. To make sure the data were of high quality and appropriate for using in ML model training, we investigated data gathering and preprocessing strategies, including feature selection and extraction. In order to evaluate the precision and dependability of rainfall prediction models, we also spoke about performance assessment indicators. Despite the advancements, issues including the scarcity of data, the difficulty of capturing spatial and temporal correlations, the interpretability of models, and integration with decision-making systems continue. The use of hybrid strategies, transfer learning, and ensemble algorithms, among other interesting future avenues, present prospects for additional improvements in rainfall prediction. Predicting rainfall using machine learning has the potential to have a big influence on industries including agriculture, water resource management, and disaster preparedness. Informed planning and decision-making may result in better resource allocation, crop management, and flood mitigation techniques. Accurate rainfall forecasts help make this happen. However, to improve the precision, understandability, and usefulness of rainfall forecast models, research and development in this area must be continued. Machine learning-based rainfall prediction will continue to progress and help lessen the effects of rainfall-related disasters as technology advances, with improvements in data collecting, processing capability, and algorithmic methodologies. It is a fascinating topic that is quickly developing and has great potential to improve society and the environment. In conclusion, machine learning has expanded the possibilities for predicting rainfall, and more research and innovation in this area will continue to propel advancements, enhancing our knowledge of and capacity for predicting rainfall patterns and enabling better resource management.

References

[1] Hsu, K., Gupta, H. V., & Sorooshian, S. (1995). Artificial neural network modeling of the rainfall-runoff process. Water resources research, 31(10), 2517-2530.

[2] Chatterjee, C., & Maharatna, K. (2019). Review of machine learning techniques for rainfall-runoff modelling. Journal of Hydroinformatics, 21(2), 228-250.

[3] Elshorbagy, A., & Simonovic, S. P. (2007). A review of artificial intelligence techniques for runoff prediction in ungauged basins. Journal of hydrology, 347(3-4), 299-309.

[4] Prakash, A., Gupta, M., & Jangid, N. K. (2020). Machine learning-based models for rainfall prediction: A review. International Journal of Intelligent Systems and Applications in Engineering, 8(1), 50-58.

[5] Moradkhani, H. (2005). Hydrologic prediction in ungauged basins: synthesis across processes, places, and scales. Water Resources Research, 41(12).

[6] Shukla, S., & Pandey, P. C. (2020). Machine learning models for rainfall prediction: A comprehensive review. Journal of Hydrology, 589, 125091.

[7] Ratnadass, A., Paturel, J. E., & Mahé, G. (2019). Rainfall prediction in the Sahel region using machine learning approaches. Water, 11(10), 2103.

[8] Khandelwal, A., & Verma, A. K. (2019). Rainfall prediction using machine learning algorithms: A review. International Journal of Emerging Trends in Engineering Research, 7(3), 274-282.

[9] Akram, A., Hossain, M. S., & Rahman, M. M. (2020). Rainfall prediction using machine learning algorithms: A comprehensive review. Water Resources Management, 34(14), 4311-4333.

[10] Zare, A., & Samadianfard, S. (2017). Application of artificial neural networks in rainfall prediction: A review. Environmental Processes, 4(3), 605-632.
[11] Jia, H., Jiang, S., Yu, M., & Tian, Y. (2018). A review of machine learning methods for precipitation prediction. Journal of Hydrology, 561, 573-586.
[12] Rakibuzzaman, M., & Buytaert, W. (2019). Machine learning for precipitation downscaling: A review. Wiley Interdisciplinary Reviews: Water, 6(4), e1356.
[13] Saha, S., & Akanda, M. A. R. (2021). A comprehensive review of machine learning-based precipitation prediction models. SN Applied Sciences, 3(5), 1-21.
[14] Khoi, D. D., & Trung, T. N. (2020). Rainfall prediction using hybrid machine learning models: A comprehensive review. Journal of Hydrology, 590, 125304.
[15] Nguyen, V. T., Wang, L. P., & Jin, J. H. (2018). A review of machine learning methods for rainfall–runoff modelling. Environmental Modelling & Software, 109, 232-246.
[16] Yu, L., & Liong, S. Y. (2005). Rainfall forecasting using least square support vector machines and relevance vector machines. Journal of hydrology, 301(1-4), 47-66.
[17] Wang, L., Huang, D., Zhang, C., & Li, Z. (2020). Deep learning for rainfall prediction using radar and satellite images: A review. Remote Sensing, 12(14), 2321.
[18] Anh, N. K., Pradhan, B., & Tien Bui, D. (2021). A comprehensive review of machine learning models for rainfall-induced landslides prediction. Landslides, 18(2), 303-325.
[19] Chen, S., Zhang, Z., Chen, Q., Wang, H., & Shu, L. (2020). Deep learning for short-term rainfall prediction: A review and new perspectives. Journal of Hydrology, 590, 125391.
[20] Sharif, H. O., Ibrahim, M. T., & Hossain, M. S. (2019). A comprehensive review on machine learning-based rainfall prediction models: Past, present, and future. Geosciences, 9(4), 162.
[21] Remember to consult the references to find specific information related to your research interests and to ensure that you include the most relevant and up-to-date sources in your paper.