Rainfall Prediction Models Based on Machine Learning: A Survey

Maneesh Kumar, Dr. Jawahar Thakur
Research Scholar, Computer Science Department, Himachal Pradesh Technical University, Shimla
Professor, Computer Science Department, Himachal Pradesh Technical University, Shimla

Abstract - The rainfall is a significant part of water resource ecosystem and acts efficiently in the field of hydrology and meteorology. In particular, rainfall is the outcomes of multi-scale air system interference and various natural factors such as thermal power, flow field, and terrain have influence on it. The task to predict the rainfall becomes complex due to these complex physical mechanisms. In the process of forecasting rainfall, the probability of precipitation present in a certain region is predicted and rainfall in future is foreseen along with the estimation of the amount of rainfall in particular regions. The various schemes of rainfall prediction are reviewed in this paper. The techniques for the rainfall prediction are based on machine learning and are compared in terms of various parameters.

Keywords - *Rainfall, Time series, Machine Learning, Analysis*

I. INTRODUCTION

The rainfall is a significant part of water resource ecosystem and acts efficiently in the field of hydrology and meteorology. In particular, rainfall is the outcome of multi-scale air system interference and various natural factors such as thermal power, flow field, and terrain have influence on it. The task to predict the rainfall becomes complex due to these complex physical mechanisms. In the process of forecasting rainfall, the probability of precipitation present in a certain region is predicted and rainfall in future is foreseen along with the estimation of the amount of rainfall in particular regions [1]. One way to predict rainfall from data is to build a predictive model with DM (data-mining) technology. DM algorithms help in building models from massive amounts of data. The science of DM evolves from AI (Artificial Intelligence), ML (Machine Learning), Statistics and Database Systems.

1.1 Rain Fall Prediction Process

Data mining techniques have brought a major change in the traditional way of weather prediction. In the last few years, researchers have developed and practiced many weather forecasting models using data mining methodologies. These forecasting models have shown great accuracy in weather prediction. Precipitation prediction with data mining technology, which is different from classic techniques of weather forecasting, has drawn a lot of attention from the research community. Based on machine learning theory, historical observational data can be used to predict future rainfall. Contrary to other types of models, this computing operation is clearly more suitable. There have been many worthwhile studies applying historical data to predict rainfall. Figure 1 illustrates a generic framework of rainfall prediction [2].

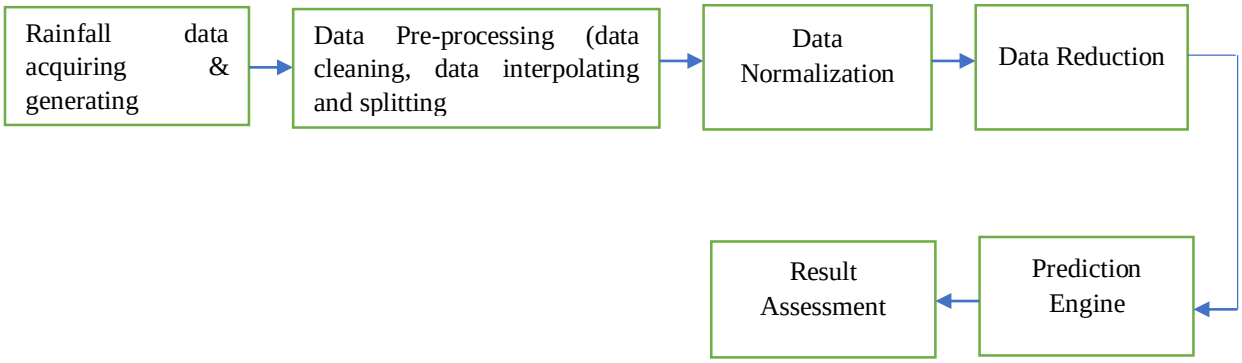

Fig. 1. Pipeline of Rainfall Prediction

All steps involved in the designing of a generic rainfall prediction framework have been described below:

i. Data Acquisition: This phase is focused on gathering the primary data for this study across India. Diverse metrics such as Average Monthly Temperature, Relative Humidity and annual rainfall are some metrics which are contained in the data. These metrics are taken from CES (Centre for Environmental Studies), from the Department of Forest and Environment, Government of India [3]. The non-monsoon

regression is analyzed monthly on the basis of weather metrics such as average temperature, cloud cover, potential evapotranspiration and rainfall.

ii. Pre-processing: The initial stage to pre-process the data is executed to attain the rainfall data and to execute the generating procedure. An automatic tool is constructed with the objective of obtaining and generating the data from the online data source. When the data is achieved and produced, the row data is cleaned, interpolated and divided. The raw datasets are consisted of some empty items and some duplicates. These duplicated and empty items are detected. Various empty values transmission around the whole dataset is handled in the data interpolating stage [4]. The fundamental goal of this phase is to normalize the min-max and mitigate the data. The average value is calculated from its chronicle neighbours and inserted as the estimated value. As an essential step for training and testing in machine learning models, data splitting is applied to divide the dataset into an appropriate proportion (generally in 70:30 ratio)

iii. Normalization: Z-score and Minmax schemes are adopted for computing the effect of normalization and non-normalization in the framework. The computation of Z-score value denoted with z_i is done as [5]:

$$z_i = \frac{x_i - \mu}{\delta}$$

In which, x_i represents $i - th$ observed value, the mean of all values is illustrated by μ in the variable and δ denotes the standard deviation in the variable.

iv. Dimensionality reduction: After normalization, dimensionality of the input data is reduced. PCA is a common theory executed to pre-process the data and mitigate the dimension and extensively utilized in multivariate statistics. The processing becomes more complicated with the maximization of number of variables. Principal Component Analysis is capable of displaying more information with fewer variables [6]. In case, the correlation coefficient among the initial variables is not 0, this implies a certain overlap among these variables. For all primary variables, PCA is effective for deleting the frequent relations of variables and generating novel variables in least amount. These new variables are independent of each other. However, the original information is stored. The vector space composed of input samples is converted to create the direction of the largest error as the base vector of the new linear space. Actually, PCA has potential to provide more information with relatively fewer factors. The complexity of the problem is simplified and the usage of hardware resources is alleviated using this algorithm.

v. Forecasting Engines: The computing model provides the choices of machine learning models and algorithms for either classification or regression. A set of algorithms such as SVM, KNN, ANN, and LSTM, are used for rainfall prediction. These models and algorithms are provided as forecasting engines in the prediction system.

vi. Result evaluation: Various evaluation metrics are used to evaluate classification and regression respectively. In classification, accuracy is most commonly used to measure the performance of classifiers, which can reveal a lot of predictive models' potential [7]. In regression, the coefficients of determination (R^2), mean square error (MSE), root mean square error (RMSE), and Pearson correlation coefficient (Pcc) are all adopted to measure the fitness of the regression model in the dataset. Once experimental results are received and saved to local storage, comparisons are made based on these evaluation metrics.

1.1.1 Rainfall Prediction Techniques

In the present times, the rainfall is predicted using two techniques namely atmospheric models and ML (machine learning) techniques.

a. Atmospheric models: These models are utilized for simulating the atmospheric operation. The atmospheric equation is present in the form of a closed system in which atmospheric motion is defined. These equations are assisted in predicting the atmospheric physical quantities and weather components, such as rainfall. The division of atmospheric models is done into three classes in accordance with their functions. Atmospheric circulation models are adopted for forecasting the atmospheric physical quantities. Different climatic factors are predicted using the Climate model that leads to forecast long-term rainfall [8]. Recently, the numerical models are extensively implemented to forecast the rainfall of medium and short-term in China. The atmospheric equations contain partial differential equations; thus, the overall accuracy of numerical computation is relied on the reasonable approximation. Moreover, there is necessity of utilizing more computations and resources in the atmospheric models. Due to these strict requirements for hardware, the implementation of atmospheric predictive models becomes complex.

b. Machine learning: ML (Machine learning) techniques are planned on the basis of the theory of several ML classifiers. In particular, some ML techniques are designed on the basis of statistics that are capable of learning only time series variation attributes from historical rainfall series. Other techniques often focus on analyzing the factors due to which the rainfall and the internal relation of these factors with rainfall is affected. A suitable ML algorithm is assisted in learning the relation amid diverse factors and rainfall due to the capacity of ML of learning the internal relation of data [9].

1.1.2 Machine Learning Methods for Rainfall Prediction

Some ML algorithms utilized to predict the rainfall are Decision Tree (DT), SVM (Support Vector Machine), Logistic Regression (LR) and ANN (Artificial Neural Network) which are discussed as follow:

i. Decision Tree: Decision tree is one of the best-known predictive models. This model can extract valuable information and classify the futuristic episodes efficiently. The designing process of this classification model consists of two stages known as building and pruning. In the first stage of tree building, the training sets are divided recursively in accordance with the value of the features. The partitioning process continues until all partitions or most records in all partitions have the same value. In Tree pruning stage, the branches containing the highest projected error rate are selected and eradicated. This stage increases the prediction accuracy of the DT, and makes the prediction process simple [10]. The C5.0 is a well-known decision tree for used for classification. This classification model splits the instance space into smaller subsets in recursive manner to assemble classification trees unless only the instances of the identical class are identified to be as a pure node or a subset consisting of events from diverse classes remain recognized as impure nodes. This tree can grow to its full ability before it is pruned back for increasing its generalisation power on hidden data.

ii. Artificial Neural Networks: ANNs have been fruitfully employed for estimating complex non-linear functions. A neural network is a machine learning model, based on the biological brain. Artificial neural networks (ANNs) are a kind of information processing system. The most non-complex configuration of an ANN involves just one hidden layer. But this configuration highly corresponds to a brain neuron network as this structure includes huge interconnections amongst the neurons in successive layers [11]. The use of this model is quite common for different types of problems, such as prediction, control and classification. The major different between ANN and other classifiers is that ANN can interpret prediction as a probability.

iii. Logistic regression (LR): Logistic regression is one of the commonly used statistical modelling methods. In this approach, the probability of a dichotomous result relates to a group of possible independent variables. This technique has been successfully applied for predicting the value of a couple of class labels or sequence variables. The logistic regression models do not necessarily need to assume discriminant analysis; however, these models generate competent and truthful results as discriminant analysis.

iv. Support Vector Machine (SVM): Support Vector Machine model refers to a supervised machine learning model [12]. This model has been successfully used for classification and regression issues. However, it is most commonly used in a classification problem because of its ability of separating two classes using a hyperplane. The key idea of SVM is to discover a hyperplane that can clearly classify the data. Hyperplanes refer to as decision boundaries. These help in the classification of the data points. Support Vectors corresponds to data points that are closer to hyperplane and affect the location and orientation of the hyperplane.

II. LITERATURE REVIEW

F. Tang, et.al (2022) suggested a displacement prediction of rainfall actuated landslide, an AdaBoost BPNN (back propagation neural network) model [13]. This model integrated multiple weak BPNN into a strong predictor, for alleviating such defects as being not difficult to fall into the local optimal resolution and low accuracy. Accepting Xinpu landslide in Chongqing for instance the rainfall data, first and foremost, was dissected, the total rainfall, normal rainfall and rainfall days were extricated, and the historical sliding of the checking point was considered as the input attributes of the model. The analysis results demonstrated the way that AdaBoost BP neural network calculation able to develop the prediction accuracy and anticipate the day to day displacement successfully.

Y. Kim, et.al (2022) presents a CGAN (contingent generative adversarial network)-based radar algorithm of predicting rainfall for exceptionally short-range weather conjectures from 10 min to 4 h [14]. This algorithm was trained and tried utilizing KMA's CAPPI (constant altitude plan position indicator) data. The subjective correlation between the radar observation and the CGAN-predicted rain rates showed higher measurable scores, like the POD (probability of detection) of 0.8442, FAR (false alarm ratio) = 0.2913, and CSI (critical success index) = 0.6268, on account of a 1-h prediction for rainfall on September 5, 2019, 15:20 KST. This study demonstrated the ability of the CGAN model for momentary rainfall forecasting. Thus, the presented algorithm could supplement the KMA MAPLE framework and be valuable in different forecasting applications.

Ogochukwu Ejike, et.al (2021) established an application of LR (logistic regression) for predicting the rainfall the next day based on weather metrics from previous days [15]. The AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and LASSO (Least Absolute Shrinkage and Selection Operator) were utilized to select the suitable LR on the basis of selecting attributes to perform predictive modeling. Hosmer-Lemeshow test was executed for determining the accuracy of every model while predicting the rainfall occurrence the next day. The result demonstrated that the established application was effective to predict the rainfall of next day at accuracy of 87%.

Hanlin Yin, et.al (2021) introduced a new data-driven framework recognized as LSTM (Long Short-Term

Memory)-based MSV-S2S (multi-state-vector sequence-to-sequence) model in which m multiple state vectors were included in order to predict the m-step-ahead rainfall runoff [16]. Unlike the traditional algorithms, this framework was adaptable to predict the rainfall of multi-day-ahead. CAMELS (Catchment Attributes and Meteorology for Large-Sample Studies) data set having 673 basins was executed to test the introduced framework against other techniques and evaluate its efficiency. The results demonstrated that the introduced framework performed better in general and its multiple state vectors were helpful to predict the rainfall multi-day-ahead.

Carlos H. R. Lima, et.al (2021) intended a hierarchical Bayesian mixture model to predict daily rainfall by the means of endogenous and external information [17]. The predictors were included in this model for mitigating the bias and variance of the forecasts and analyzing the climate indices generated through the low-level wind over the region. For this, PCA (Principal Component Analysis) algorithm was implemented. The intended model utilized these indices for boosting its forecast skills. The model structure was planned on the basis of comprehensive data analysis in which the application of SOM (Self-Organizing Maps) was adopted for examining the spatio-temporal patterns of rainfall. The results demonstrated that the intended model had improved skills as compared to the reference models.

R. Kingsy Grace, et.al (2020) discussed that the major method for predict the climatic conditions in any country was to predict the rainfall [18]. A model was presented to predict the rainfall with the implementation of MLR (Multiple Linear Regression) for Indian dataset. The multiple meteorological metrics were comprised in the input data and the rainfall was predicted more precisely using these metrics. Diverse factors such as MSE (Mean Square Error), accuracy and correlation were considered in the evaluation of the presented model. The results indicated that the presented model had generated higher results in comparison with other algorithms.

Arief Bramanto Wicaksono Putra, et.al (2020) projected a new DNN framework recognized as DAESCNN (Deep Auto-Encoder Semi Convolutional Neural Network) for enhancing the efficacy of the traditional algorithm [19]. This framework was tested using the annual rainfall data. This data was gathered from weather stations of Indonesia in the period 2006-2016. A program developed via MATLAB was utilized to train the projected framework so that its efficiency was evaluated with respect to diverse metrics. The results exhibited that the accuracy achieved from the projected framework was calculated 99%.

T. Dananjali, et.al (2020) presented 3 DM (data mining) models namely LR (linear regression), SMO (Sequential Minimal Optimization) regression, and M5P model for predicting the rainfall [20]. The data related to rainfall was gathered from Badulla district, Sri Lanka for training these models so that weekly rainfall was predicted for the following five months lead-time. Various metrics such as MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), RRSE (Root Relative Squared Error) and residual analysis were utilized to quantify every model. The outcomes revealed the supremacy of the M5P model tree due to its lower error value, higher direction accuracy and superior randomness of the error values in comparison with other models.

2.1 Table

Author	Year	Technique Used	Findings
F. Tang, et.al	2022	AdaBoost BPNN (back propagation neural network) model	The analysis results demonstrated the way that AdaBoost BP neural network calculation able to develop the prediction accuracy and anticipate the day to day displacement successfully.
Y. Kim, et.al	2022	CGAN (contingent generative adversarial network)-based radar algorithm	This study demonstrated the ability of the CGAN model for momentary rainfall forecasting. Thus, the presented algorithm could supplement the KMA MAPLE framework and be valuable in different forecasting applications.
Ogochukwu Ejike, et.al	2021	LR (logistic regression)	The result demonstrated that the established application was effective to predict the rainfall of next day at accuracy of 87%.
Hanlin Yin, et.al	2021	LSTM (Long Short-Term Memory)-based MSV-S2S (multi-state-vector sequence-to-sequence) model	The results demonstrated that the introduced framework performed better in general and its multiple state vectors were helpful to predict the rainfall multi-day-ahead.
Carlos H. R. Lima, et.al	2021	A hierarchical Bayesian mixture model	The results demonstrated that the intended model had improved skills as compared to the reference models.

R. Kingsy Grace, et.al	2020	MLR (Multiple Linear Regression)	The results indicated that the presented model had generated higher results in comparison with other algorithms.
Arief Bramanto Wicaksono Putra, et.al	2020	DAESCNN (Deep Auto-Encoder Semi Convolutional Neural Network)	The results exhibited that the accuracy achieved from the projected framework was calculated 99%.
T. Dananjali, et.al	2020	3 DM (data mining) models namely LR (linear regression), SMO (Sequential Minimal Optimization) regression, and M5P model	The outcomes revealed the supremacy of the M5P model tree due to its lower error value, higher direction accuracy and superior randomness of the error values in comparison with other models.

III. CONCLUSION

Data mining techniques have brought a major change in the traditional way of weather prediction. In the last few years, researchers have developed and practiced many weather forecasting models using data mining methodologies. These forecasting models have shown great accuracy in weather prediction. Precipitation prediction with data mining technology, which is different from classic techniques of weather forecasting, has drawn a lot of attention from the research community. Based on machine learning theory, historical observational data can be used to predict future rainfall. The machine learning model are best performing models for the rainfall prediction. In future deep learning models can be proposed for the rainfall prediction.

IV. REFERENCES

[1] S. Manandhar, S. Dev, Y. H. Lee, Y. S. Meng and S. Winkler, "A Data-Driven Approach for Accurate Rainfall Prediction," in IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 11, pp. 9323-9331, Nov. 2019

[2] B. Abishek, R. Priyatharshini, M. A. Eswar and P. Deepika, "Prediction of effective rainfall and crop water needs using data mining techniques," 2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), 2017, pp. 231-235

[3] A. Sharma and M. K. Goyal, "Bayesian network model for monthly rainfall forecast," 2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), 2015, pp. 241-246

[4] A. Pranolo, Y. Mao, Y. Tang, Haviluddin and A. P. Wibawa, "A Long Short-Term Memory Implemented for Rainfall Forecasting," 2020 6th International Conference on Science in Information Technology (ICSITech), 2020, pp. 194-197

[5] Yajnaseni Dash, S.K. Mishra, B.K. Panigrahi, "Rainfall prediction of a maritime state (Kerala), India using SLFN and ELM techniques", 2017, International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)

[6] A. Kala, S.Ganesh Vaidyanathan, "Prediction of Rainfall Using Artificial Neural Network", 2018, International Conference on Inventive Research in Computing Applications (ICIRCA)

[7] Mary N. Ahuna, Thomas J. Afullo, Akintunde A. Alonge, "Rainfall rate prediction based on artificial neural networks for rain fade mitigation over earth-satellite link", 2017, IEEE AFRICON

[8] Hiyam Abobaker Yousif Ahmed, Sondos W. A. Mohamed, "Rainfall Prediction using Multiple Linear Regressions Model", 2020, International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)

[9] Imrus Salehin, Iftakhar Mohammad Talha, Md. Mehedi Hasan, Sadia Tamim Dip, Mohd. Saifuzzaman, Nazmun Nessa Moon, "An Artificial Intelligence Based Rainfall Prediction Using LSTM and Neural Network", 2020, IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)

[10] Oswalt Manoj S, Ananth J P, "MapReduce and Optimized Deep Network for Rainfall Prediction in Agriculture", 2020, The Computer Journal

[11] Gunawansyah, Thee Houw Liong, Adiwijaya, "Prediction and anomaly detection of rainfall using evolving neural network to support planting calender in soreang (Bandung)", 2017, 5th International Conference on Information and Communication Technology (ICoIC7)

[12] Sankhadeep Chatterjee, Bimal Datta, Soumya Sen, Nilanjan Dey, Narayan C. Debnath, "Rainfall prediction using hybrid neural network approach", 2018, 2nd International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom)

[13] F. Tang, X. Wang, Z. Wang, M. Xiao, Y. Ma and Y. Wan, "Displacement prediction of rainfall induced landslide based on AdaBoost BP neural network," 2022 3rd International Conference on Geology, Mapping and Remote Sensing (ICGMRS), 2022, pp. 794-799

[14] Y. Kim and S. Hong, "Very Short-Term Rainfall Prediction Using Ground Radar Observations and Conditional Generative Adversarial Networks," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-8, 2022, Art no. 4104308

[15] Ogochukwu Ejike, David L. Ndzi, Abdul-Hadi Al-Hassani, "Logistic Regression Based Next-Day Rain Prediction Model", 2021, International Conference on Communication & Information Technology (ICICT)

[16] Hanlin Yin, Xiuwei Zhang, Jin Jin, "Rainfall-runoff modeling using LSTM-based multi-state-vector sequence-to-sequence model", 2021, Journal of Hydrology

[17] Carlos H. R. Lima, Hyun-Han Kwon, Yong-Tak Kim, "A Bernoulli-Gamma hierarchical Bayesian model for daily rainfall forecasts", 2021, Journal of Hydrology

[18] R. Kingsy Grace, B. Suganya, "Machine Learning based Rainfall Prediction", 2020, 6th International Conference on Advanced Computing and Communication Systems (ICACCS)

[19] Arief Bramanto Wicaksono Putra, Rheo Malani, Bedi Suprapty, Achmad Fanany Onnilita Gaffar, "A Deep Auto Encoder Semi Convolution Neural Network for Yearly Rainfall Prediction", 2020, International Seminar on Intelligent Technology and Its Applications (ISITIA)

[20] T. Dananjali, S. Wijesinghe, J. Ekanayake, "Forecasting Weekly Rainfall Using Data Mining Technologies", 2020, From Innovation to Impact (FITI)