RESEARCH
www.commprac.com
ISSN 1462 2815

MACHINE LEARNING TECHNIQUES IN RAINFALL PREDICTION – A COMPARISON STUDY

S. Anitha Rajathi [1] and Dr. J. Devagnanam [2]

[1]Assistant Professor, Department of CSBS/R.M.D. Engineering College, Kaveraipettai, Tiruvallur, TamilNadu.
[2]Associate Professor, Department of CS /St. George's Arts and Science College, Chennai, Tamilnadu.

Abstract

As a result of the potential for numerous calamities, rainfall prediction is crucial. In addition to assisting people in taking preventative action, the prediction should be accurate. Long-term rainfall prediction and short-term rainfall prediction are the two sorts of forecasts. Predictions, especially those made in the short future, can provide us with precise results. Creating a model for predicting long-term rainfall is the key challenge. Because it is so intimately related to the economy and the lifespan of humans, heavy precipitation forecasting could be a significant setback for the earth science department. Every year, people all around the world experience natural disasters like floods and droughts because of this. For nations like India, whose economy is primarily based on agriculture, the accuracy of the rainfall statement has considerable importance. The goal of this work is to provide a review of various methods for predicting rainfall early on. The review of rainfall forecasting using various methodologies is covered in the study, along with an analysis of their advantages.

Keywords: Rainfall, Prediction, Long-term, Short-term, Logistic Regression, LSTM

1. INTRODUCTION

Rainfall is one of the most difficult elements of the hydrologic cycle to forecast. This is due to the tremendous range of variability it displays over a wide range of scales both in space and time. Rainfall has a very important meaning for life, so the accurate rainfall prediction is needed. It can be used to support agricultural production, water resource management, and hydro meteorological disaster mitigation. Agricultural cropping patterns, yields, and the threat of disaster depend on the seasonal pattern of rainfall. The Accurate rainfall prediction can be used to anticipate the risks and minimize the losses. Predicting monthly rainfall is very important for flood cont control management and minimizing disaster risk. Heavy rainfall prediction is a major challenge for meteorological department as it is closely associated with the economy and life of human. It is a cause for natural disasters, like flood and drought which are encountered by people across the globe every year. Analyzing past rainfall data and using it to predict future rainfall is a complex task that only very robust and effective algorithms can achieve. The situation is aggravated by the fact that rainfall is also very difficult to measure at scales of interest to hydrology and climatology. Even rainfall measurement at a point by a single rain gauge is not an easy task. The methods available for rainfall forecasting depend on the scale of interest. In this article, we intend to classify various existing time series forecasting methods.

2. LITERATURE REVIEW

Researchers have been working to improve the accuracy of rainfall prediction by optimizing and integrating Machine Learning techniques. Some of the selected studies are discussed in this section.

In [1] S. Zhang, L. Lu, J. Yu, and H. Zhou performed a comparative analysis of Support

Vector Machine (SVM), Artificial Neural Networks (ANN), and Adaptive Neuro Fuzzy Inference System (ANFIS) on rainfall prediction. Four criteria have been used by the authors to compare the prediction models: I by using various lags as modeling inputs; (ii) by using training data of just heavy rainfall events; (iii) by performance analysis in peak values and all values; and (iv) by forecasting performance for 1 hour to 6 hours. The results show that ANNs that were trained using datasets of heavy rainfall performed better. The prior 2-hour input data was recommended for all three modeling strategies for forecasting 1 to 4 hours in advance (ANN, SVM and ANFIS). By utilizing various input lags, ANFIS demonstrated a higher capacity to avoid information noise. Finally, SVM showed itself to be more resilient during peak values under extreme typhoon events.

S. Zainudin, D. S. Jasim, and A. A. Bakar conducted a comparative examination of several data mining approaches, including Random Forest, Support Vector Machine, Naive Bayes, Neural Network, and Decision Tree, for predicting rainfall in Malaysia in [2]. Datasets for this experiment came from a number of weather stations in Selangor, Malaysia. Pre-processing activities were used to deal with the noise and missing values in the dataset before the classification procedure. As a result of successfully classifying a large number of examples with a minimal quantity of training data, the findings demonstrated Random Forest's significant performance.

D. Nayak, A. Mahapatra, and P. Mishra conducted a survey on several Neural Network architectures utilized for rainfall prediction over the previous 25 years in [3]. The authors noted that the majority of researchers had substantial success in predicting rainfall using propagation networks, and that SVM, MLP, BPN, RBFN, and SOM are better forecasting methods than other statistical and numerical ones. Moreover, some restrictions have been noted.

For the purpose of predicting rainfall in Thailand, B. K. Rani and A. Govardhan deployed artificial neural networks in [4]. For prediction, they employed a back propagation neural network, which indicated a respectable level of accuracy. It was suggested that a few further parameters, including sea surface temperature for the regions surrounding Andhra Pradesh and the southern half of India, be included to the input data for rainfall prediction in the future.

N. Tyagi and A. Kumar used back propagation, radial basis function, and neural networks to estimate monthly rainfall in [5]. The Coonoor region in the Nilgiri district served as the source of the dataset for prediction (Tamil Nadu). Mean Square Error was used to measure performance. Results showed that Radial Basis Function Neural Networks had higher accuracy and reduced Mean Square Error. Also, the researchers employed these methods for rainfall forecasting in the future.

In [6], N. Solanki and G. P. B proposed a Hybrid Intelligent System that combined Genetic Algorithm and Artificial Neural Network. In ANN, MLP functions as the Data Mining engine to produce predictions, whereas the Genetic Algorithm was used for inputs, the connection structure between the inputs, the output layers, and to improve the training of Neural Networks.

In [7], C. S. Thirumalai analysed rainfall rates in prior years in relation to several agricultural seasons, such as rabi, Kharif, and zaid, and then used the Linear Regression Approach to predict (rainfall) for subsequent seasons. The input dataset for prediction was chosen based on certain corps seasons from prior years.

In [8], N. Mishra, H. K. Soni, S. Sharma, and A. K. Upadhyay created one- and two-month forecasting models utilising artificial neural networks to predict rainfall (ANN). The input dataset, which covered the previous 141 years, was chosen from several stations in North India. These models made use of the Levenberg-Marquardt training function and Feed Forward Neural Network with Back Propagation. The effectiveness of both models was assessed using regression analysis, mean square error, and relative error magnitude. The outcomes demonstrated that a one-month forecasting model is more accurate in predicting rainfall than a two-month forecasting model.

H. Vathsala and S. G. Koolagudi presented a method in [9] that combined statistical and data mining techniques. The suggested method made rainfall predictions in five different categories, including flood, excess, normal, deficiency, and drought. The predictors were chosen based on association rules and obtained from the local and global environments with the maximum degree of confidence. Wind speed, sea level pressure, maximum temperature, and minimum temperature were recorded from the immediate area. Indian Ocean dipole conditions and the southern oscillation were drawn from the global environment.

R. Venkata Ramana, et al. used the Wavelet Neural Network Model (WNN), a fusion of the Wavelet Method and Artificial Neural Network, to predict the rainfall in [10]. (ANN). Using data from the Darjeeling rain gauge station in India, monthly rainfall prediction was carried out using both methodologies (WNN and ANN) to examine the performance. Performance evaluation was done using statistical methods, and the results showed that WNN outperformed ANN.

In [11], M. P. Darji et al. published a thorough examination and comparison of different neural networks for forecasting rainfall. In comparison to other statistical and numerical forecasting techniques, the survey finds that RNN, FFNN, and TDNN are suitable for predicting rainfall. Also, when forecasting annual, monthly, and weekly rainfall, respectively, TDNN, FFNN, and lag FFNN all performed well. This study also covered the numerous accuracy metrics employed by various academics to assess the effectiveness of ANNs.

Sharma et al. introduced a Bayesian network model in [12] for the prediction of the mean monthly rainfall at 21 sites in Assam, India. This research may help with better water resource management. The study used monthly data from several sources over a 20-year period from 1981 to 2000 for all atmospheric parameters. For this model, rainfall at a station is used as a variable, and a Bayesian network is used to highlight the dependencies between rainfalls at various stations. In this paper, the conditional probability was determined using maximum likelihood approximations and the K2 algorithm.

There are five different atmospheric variables that are used: temperature, cloud cover, relative humidity, wind speed, and Southern Oscillation Index (SOI). The findings showed that temperature is most effective and wind speed is least effective. Moreover, SOI is thought to be crucial for enhancing outcomes. While some stations achieved efficiency levels above 95%, other stations received satisfactory results.

Akash D. Dubey suggested a strategy for predicting rainfall using artificial neural networks in [13]. (ANN). The author of this piece uses weather information from Pondicherry, India. To generate ANN models, three distinct training algorithms—feed-forward back propagation, layer recurrent, and feed-forward distributed time delay—were used, with a cap of 20 neurons for each model. The feed-forward distributed time

delay algorithm, out of all the algorithms, has the best accuracy and an MSE value as low as 0.0083, according to the data.

While employing artificial intelligence and LSTM techniques,[23] Iftakhar suggested a model for predicting how much rain will fall. This method of determining the rainfall is sophisticated. For this kind of method implementation and accuracy research, the deep learning methodology is most beneficial. When measuring memory sequence data, a long short-term memory algorithm is used to quickly calculate historical data and produce the best prediction. To do this, we gathered information from six different regions. We have used six parameters in order to predict (temperature, dew point, humidity, wind pressure, wind speed, and wind direction).

2.1. Data Collection

Indian weather information is used in the forecast model. For each district, for each month, the average rainfall is included in this dataset from 1951 to 2000. The nine measured attributes that make up the raw weather data are the date, temperature (high, low, average), Dew point (high, low, average), humidity (high, low, average), sea level pressure (high, low, average), visibility (high, low, average), wind (high, low, average), precipitation (high, low, average), and events (Rainfall snow, thunderstorm, fog). The average temperature, average humidity, average sea level pressure, average wind, and average events features as stated in table I were used for this work out of these nine features. For improved model computation and prediction, we have excluded less important features in the dataset.

Table 1: Weather Data Description

Attribute	Type	Description
Temperature	Numerical	Temp is in ºC
Humidity	Numerical	Humidity in Percentage
Sea Level Pressure	Numerical	Sea Level Pressure in hPa
Windy	Numerical	Wind Speed in km/h
Events	Numerical	Rainfall in mm

2.2 Data Preprocessing and Data Cleaning

The fundamental problem with weather forecasting is the selection and quality of the data. For this reason, extensive preprocessing of the data is performed in order to produce precise and reliable predictions. In this step, undesired information or noise is eliminated from the obtained data set by eliminating irrelevant information and maintaining information that is most important to better prediction. The missing values in the gathered data set are another significant problem that needs to be fixed. The data set's missing values are filled utilising a variety of methods. The modes and means calculated from the data are used to fill in the missing values for the attributes in the dataset. The dataset from which the classifiers can be trained is made more complete by include the missing values.

2.3 Research Methodology

Unsupervised learning and supervised learning are the two primary categories of machine learning methodologies. Predictive models are constructed through the application of supervised learning techniques. Experimental implementation and comparisons of the classification techniques ANN, Logistic Regression, Naive Bayes, Random Forest and LSTM are conducted. Artificial neural networks are first. A neural network is a sizable distributed processor that employs parallel processing techniques

and is made up of basic processing units that store empirical knowledge and are prepared to use it when needed. In that they can accept inputs and process the data utilising various computational nodes, neural networks function similarly to how the human brain does. Based on the requested application, a pertinent result is generated. The primary benefit of neural networks is their capacity to show the existence of non-linearity between input and output variables.

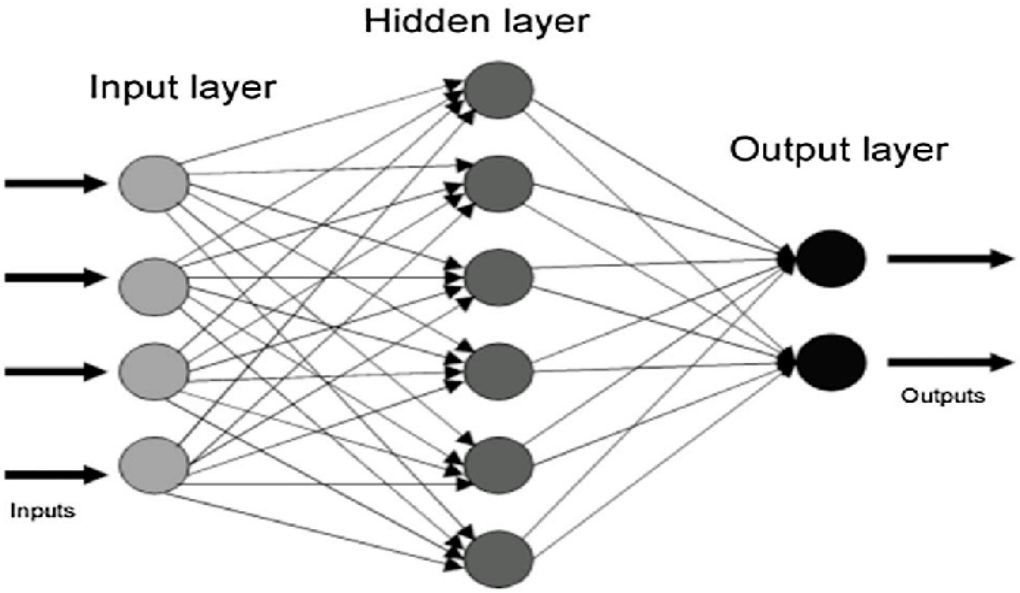

Fig.1: Neural Network

3. SIGNIFICANT ISSUES WITH NEURAL NETWORKS

1. Number of Hidden Nodes and Layers

The Input, Hidden, and Output Layers are the three standard layers that make up the generic Artificial Neural Network shown in Fig. 1. The Hidden Layer is the present-day layer with which we are most concerned. The majority of neural network calculations take place in this layer, which is also where the actual nonlinear mapping between input and output occurs. The entire neural network's output would diverge if any mistake were made in this layer, which might have disastrous consequences for prediction. As a result, we need to think about how many hidden layers we require and their nodes. These parameters can improve the network's overall accuracy.

2. Logistic Regression

One of the methods used to do a regression analysis on a set of binary input parameters is logistic regression. The sigmoidal function, which is the foundation of the entire approach, is primarily the focus of the logistic regression function shown in Fig. 2. It creates an S-shaped map of the features of any supplied data between the values of 0 and 1. It was initially created by statisticians to examine the rise of the human population in a controlled setting. Since then, it has been improved to fit new domains and precisely map their properties.

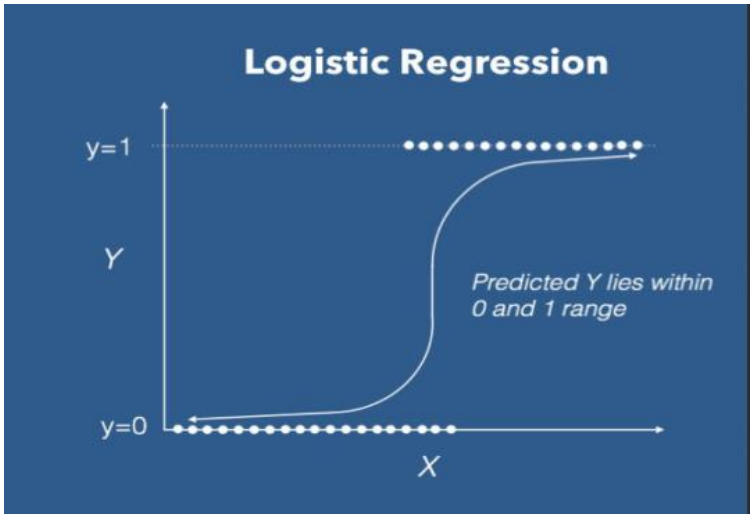

Fig.2: Logistic regression

3. Naive Bayes

In 1973, [14] and 1992, [15] respectively, described the Naive Bayesian classifier. Statistical classifiers include Bayesian classifiers. One of the most reliable machine learning techniques for predicting rainfall is the Naive Bayes algorithm [11]. The Bayes rule of conditional probability serves as the foundation for the Nave Bayes classifier [16]. It examines each quality in isolation and makes the assumption that they are all significant and independent. For example, in [17], naive Bayes classifiers were heavily utilized in fault-proneness prediction. The naive Bayes classifier has the benefit of just needing a modest quantity of training data to estimate the classification-related parameters.

4. Random Forest

Another technique used as part of an ensemble classifier is Random Forest [18]. Random Forest is a decision tree-based classifier that performs admirably in computer engineering experiments by Guo et al., [19]. One significant benefit of random forest is its speed and ability to handle a high number of input attributes. There are dozens or hundreds of trees there. A random selection of attributes is used while building a decision tree. The approach used to generate the trees is as follows [20]:

1) The sample bootstrap data at each tree's root node is identical to the real data. Each tree has a distinct bootstrap sample
2) A subset of the input variables is randomly chosen using the best split approach.
3) Each tree is then allowed to develop to its full potential without being pruned.
4) After the forest's trees have all been constructed, new instances are added to each tree, and the classification that receives the most votes is chosen as the forecast for the new instance(s).

5. LSTM

The deep learning strategy is the most beneficial for implementing this kind of method and determining its accuracy. A long short-term memory technique is used to measure memory sequence data, quickly calculate historical data, and produce the most accurate prediction. We have developed a model to assist us in estimating the amount of rainfall. To do this, we have data from six different regions. We have used 6

parameters to forecast (temperature, dew point, humidity, wind pressure, wind speed, and wind direction). We can precisely forecast the rainfall after examining all of our data.

4. DISCUSSION & CONCLUSION

To analyse widely used machine learning techniques for predicting rainfall using multiple performance metrics over Indian weather data. The various measuring characteristics are essential for providing accurate rainfall forecasts. In this instance, the long-short term memory strategy for rainfall prediction is effective and appropriate. The data being utilised as input for classification and prediction has a significant impact on the level of accuracy and prediction. Every algorithm has benefits and drawbacks, making it challenging to select the best algorithm. By creating a hybrid prediction model that combines different machine learning techniques, the model's prediction accuracy can be improved. We forecast the quantity of rainfall in a year or a month using data on wind speed, wind direction, temperature, and pressure. Based on this methodology, it may be suggested to use LSTM to forecast the Rainfall forever.

References

1) Zhang S., Lu L., Yu J., and Zhou H. (2016). “Short-term water level prediction using different artificial intelligent models,” in 5th International Conference on Agro Geoinformatics, Agro-Geoinformatics.

2) Zainudin S., Jasim D. S., and Bakar A. A. (2016). “Comparative Analysis of Data Mining Techniques for Malaysian Rainfall Prediction,” Int. J. Adv. Sci. Eng. Inf. Technol., vol. 6, no. 6, (pp. 1148–1153).

3) Nayak D., Mahapatra A., and Mishra P. (2013). “A Survey on Rainfall Prediction using Artificial Neural Network,” Int. J. Comput…., vol. 72, no. 16, (pp. 32–40).

4) Rani B. K., and Govardhan A. (2013). “Rainfall Prediction Using Data Mining Techniques - A Survey,” (pp. 23–30).

5) Tyagi N., and Kumar A. (2017). "Comparative analysis of back propagation and RBF neural network on monthly rainfall prediction," Proc. Int. Conf. Inven. Comput. Technol. ICICT 2016, vol.1.

6) Solanki N. and G. P. B. (2018). “A Novel Machine Learning Based Approach for Rainfall Prediction,” Inf. Commun. Technol. Intell. Syst. (ICTIS 2017) - Vol. 1, vol. 83, no. Ictis 2017.

7) Thirumalai C. S. (2017). “Heuristic Prediction of Rainfall Using Machine Learning Techniques”.

8) Mishra N., Soni H. K., Sharma S., and Upadhyay A. K. (2018). “Development and Analysis of Artificial Neural Network Models for Rainfall Prediction by Using Time-Series Data,” Int. J. Intell. Syst. Appl., vol. 10, no. 1, (pp. 16–23).

9) Vathsala H., and Koolagudi S. G. (2017). “Prediction model for peninsular Indian summer monsoon rainfall using data mining and statistical approaches,” Comput. Geosci., vol. 98, (pp. 55–63).

10) R. VenkataRamana, B. Krishna, S. R. Kumar, and N. G. Pandey. (2013). “Monthly Rainfall Prediction Using Wavelet Neural Network Analysis,” Water Resour. Manag., vol. 27, no. 10, (pp. 3697–3711).

11) Darji M. P., Dabhi V. K., and Prajapati H. B. (2015). “Rainfall forecasting using neural network: A survey,” 2015 Int. Conf. Adv. Comput. Eng. Appl., no. March, (pp. 706–713).

12) Sharma, Ashutosh and Manish Kumar Goyal. (2015). “Bayesian network model for monthly rainfall forecast”, Research in Computational Intelligence and Communication Networks (ICRCICN), IEEE International Conference.

13) Dubey and Akash D. (2015). “Artificial neural network models for rainfall prediction in Pondicherry”, International Journal of Computer Applications, Vol. 120, No. 3.

14) Duda R. O.,and Hart P. E.(1973). Pattern classification and scene analysis, John Wiley and Sons.

15) Langley P., Iba W., and Thompson K. (1992). “An analysis of Bayesian Classifiers”, in Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, CA.

16) Mccallum A., and Nigam K. (1998). “A Comparison of Event Models for Naive Bayes Text Classification”, Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98)-Workshop on Learning for Text Categorization, (pp. 41-48).

17) Ibrahim Raaed K., Kadhim Roula A.J.. (2016). “Incorporating SHA-2 256 with OFB to realize a novel encryption”, IEEE paper on image encryption.

18) T. Menzies, J. Greenwald and A. Frank. (2007). “Data Mining Static Code Attributes to Learn Defect Predictors”, IEEE Transactions on Software Engineering, Vol. 33, No. 1, 2-13.

19) L. Breiman.(2001). “Random forests”, Machine Learning, Vol. 45, No. 1,(pp. 5-32).

20) Guo L., Ma Y., Cukic B. and Singh H. (2004). Robust prediction of fault-proneness by random forests, In Proc. of the 15th International Symposium on Software Reliability Engineering ISSRE’04, (pp. 417-428).

21) Jiang Y., Cukic B., Menzies T., and Bartlow N. (2008). “Comparing design and code metrics for software quality prediction”, Proc. Fourth Int. Workshop on Predictor Models in Software Engineering, PROMISE’08, New York, USA, (pp. 11-18).

22) Pradeep Nijalingappa and Sandeep B (2015). “Machine learning approach for the identification of diabetes retinopathy and its stages”. International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), IEEE International Conference

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