International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume: 3 | Issue: 3 | Mar-Apr 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 - 6470

Comparative Study of Machine Learning Algorithms for Rainfall Prediction

Mylapalle Yeshwanth[1], Palla Ratna Sai Kumar[1], Dr. G. Mathivanan[2] M.E., Ph.D

[1]Student IT Department, [2]Professor
[1,2]Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India

***How to cite this paper**:* Mylapalle Yeshwanth | Palla Ratna Sai Kumar | Dr. G. Mathivanan M.E., Ph.D "Comparative Study of Machine Learning Algorithms for Rainfall Prediction" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-3, April 2019, pp.677-681, URL: http://www.ijtsrd.com/papers/ijtsrd22961.pdf

IJTSRD22961

ABSTRACT

Majority of Indian framers depend on rainfall for agriculture. Thus, in an agricultural country like India, rainfall prediction becomes very important. Rainfall causes natural disasters like flood and drought, which are encountered by people across the globe every year. Rainfall prediction over drought regions has a great importance for countries like India whose economy is largely dependent on agriculture. A sufficient data length can play an important role in a proper estimation drought, leading to a better appraisal for drought risk reduction. Due to dynamic nature of atmosphere statistical techniques fail to provide good accuracy for rainfall prediction. So, we are going to use Machine Learning algorithms like Multiple Linear Regression, Random Forest Regressor and AdaBoost Regressor, where different models are going to be trained using training data set and tested using testing data set. The dataset which we have collected has the rainfall data from 1901-2015, where across the various drought affected states. Nonlinearity of rainfall data makes Machine Learning algorithms a better technique. Comparison of different approaches and algorithms will increase an accuracy rate of predicting rainfall over drought regions. We are going to use Python to code for algorithms. Intention of this project is to say, which algorithm can be used to predict rainfall, in order to increase the countries socioeconomic status.

KEYWORDS: *rainfall prediction, machine learning, multiple linear regression, Random Forest Regressor, adaboost Regressor*

I. INTRODUCTION

Precipitation expectation remains a genuine concern and has pulled in the consideration of governments, businesses, chance administration elements, just as mainstream researchers. Precipitation is a climatic factor that influences numerous human exercises like rural creation, development, control age, ranger service and the travel industry, among others. To this degree, precipitation forecast is fundamental since this variable is the one with the most astounding relationship with unfavorable normal occasions, for example, avalanches, flooding, mass developments and torrential slides. These episodes have influenced society for a considerable length of time. In this way, having a proper methodology for precipitation forecast makes it conceivable to take preventive and relief measures for these normal wonders.

Likewise these forecasts encourage the supervision of farming exercises, development, the travel industry, transport, and wellbeing, among others. For organizations in charge of catastrophe anticipation, giving exact meteorological forecasts can help basic leadership notwithstanding conceivable event of normal occasions.

Throughout the most recent couple of years, machine learning has been utilized as a fruitful instrument in regression for taking care of complex issues. Profound Learning is a general term used to allude to a progression of multilayer designs that are prepared utilizing unsupervised calculations. The primary improvement is learning a smaller, substantial, and non-straight portrayal of information by means ofunsupervised techniques, with the expectation that the new information portrayal adds to the forecast job needing to be done. This methodology has been effectively connected to fields like PC vision, picture acknowledgment, regular language handling, and bioinformatics. Profound learning has appeared for displaying time-arrangement information through systems like Restricted Boltzmann Machine (RBM), Conditional RBM, Autoencoder, Recurrent neural system, Convolution and pooling, Hidden Markov Model.

Precipitation forecast is useful to evade flood which spare lives and properties of people. In addition, it helps in overseeing assets of water. Data of precipitation in earlier causes ranchers to deal with their harvests better which result in development of nation's economy. Variance in precipitation timing and its amount makesprecipitation expectation a testing undertaking for meteorological researchers. In every one of the administrations given by meteorological office, Weather estimating emerges on top for every one of the nations over the globe. The assignment is intricate as it requires quantities of particular and furthermore all calls are made with no sureness. The diverse strategies utilized for precipitation expectation for climate determining with their restrictions. Different regression calculations which are utilized for expectation are talked about with their means in detail.

II. RELATED WORK

P. Goswami and Srividya have joined RNN and TDNN highlights and finish of their work was that composite models gives preferable exactness over the single model. C.Venkatesan et al. utilized Multilayer Feed Forward Neural Networks (MLFNN) for anticipating Indian summer storm precipitation. Blunder Back Propagation (EBP) calculation is prepared and connected to anticipate the precipitation. Three system models with two, three and ten info parameters have investigated. They additionally contrasted the yield result and the factual models. A.Sahai et al. utilized blunder back engendering calculation for Summer Monsoon Rainfall expectation of India on month to month and occasional time arrangement. They utilized information of past five years of month to month and regular mean precipitation esteems for precipitation expectation. N.Philip and K.Josheph utilized ABF neural system for yearly precipitation determining Kerala district. Their work recommends that ABFNN performs superior to the Fourier examination.

In these months, there is assurance that precipitation occasions will be available.

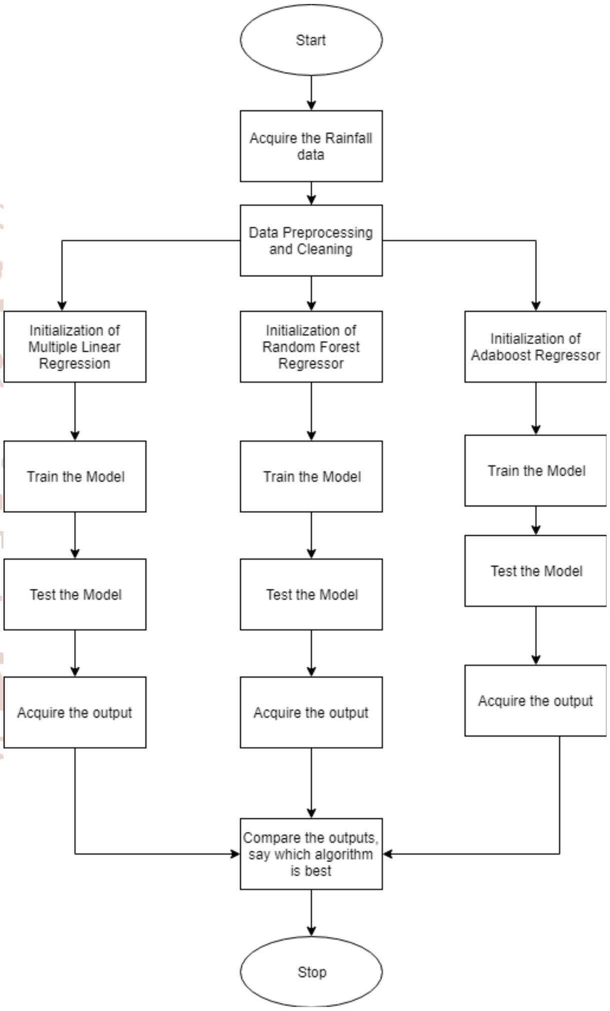
The creators utilize the normal mugginess and normal breeze speed as logical factors. The investigations were done with three kinds of various systems: Feed Forward Back Propagation, Layer Recurrent, and Cascaded Feed Forward Back Propagation. At that point, the outcomes acquired with each system are looked at, finding that the kind of system that got the best outcomes was Feed Forward Back Propagation. Liu et al. propose an option over the past model. They investigate the utilization of hereditary calculations as an element choice calculation, a then Naive Bayes as the prescient calculation. The issue is deteriorated into two expectation issues: precipitation occasion (i.e., a double forecast issue), and a classification of precipitation on the off chance that that precipitation is available (i.e., light, moderate and solid precipitation). The appropriation of hereditary calculations for the determination of information sources, demonstrates that it is conceivable to lessen the multifaceted nature of the dataset acquiring comparable or marginally better execution.

Liu et al. (2014) proposed a model dependent on Deep Learning (DL). In their examination they apply Deep Neural Network (DNN) to process enormous information including datasets of very nearly 30 years (1-1-1983 to 31-12-2012) of natural records given by the Hong Kong Observatory (HKO). The information is utilized to foresee the climate change in the following 24 hours, given four factors: temperature, dew point, Mean Sea Level Pressures (MSLP) and wind speed. The outcomes acquired for creators demonstrate that the DNN give a decent component space to climate datasets and a potential device for the element combination of time arrangement issues. Anyway they don't anticipate with their model progressively troublesome climate information, for example, precipitation dataset.

III. DATA AND METHODOLOGY

The yearly precipitation information was gathered from India Water Portal. The factual relapse system is connected over this information so as to build up a model to foresee the yearly precipitation esteems. Relapse is a measurable strategy that utilizes the connection between at least two factors on observational database so as to anticipate a result from different factors. There are numerous sorts of relapse examination out of which straight relapse is especially connected as it is straightforward. The quantitative factors are thought to be straightly identified with each other. There are essentially two sorts of straight relapse for example straightforward direct relapse and numerous direct relapse. The numerous straight relapse that is utilized in this investigation

SYSTEM ARCHITECTURE


Start
Acquire the Rainfall data
Data Preprocessing and Cleaning
Initialization of Multiple Linear Regression
Initialization of Random Forest Regressor
Initialization of Adaboost Regressor
Train the Model
Train the Model
Train the Model
Test the Model
Test the Model
Test the Model
Acquire the output
Acquire the output
Acquire the output
Compare the outputs, say which algorithm is best
Stop

IV. ALGORITHM:

MULTIPLE LINEAR REGRESSIONS

Massie and Rose (1997) connected a straightforward direct relapse technique to anticipate every day most extreme temperatures examined at Nashville, Tennessee.

Relapse endeavors, to decide the quality of the connection between one ward variable generally meant by Y and a progression of other changing factors known as autonomous factors. In straightforward relapse there are just two factors where one is the reliant variable and other is the free factor and the connection among them is of kind as beneath. This is known as the deterministic model

$$Y=A+BX$$

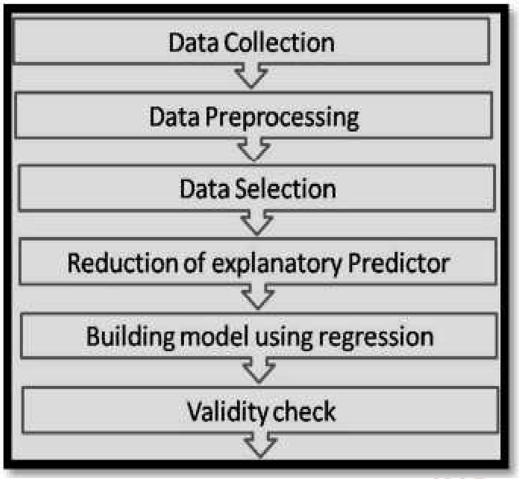
Here Y= Dependent variable X= independent variable A, B= Relapse parameters In Multiple relapses there are multiple factors among which one is reliant variable and all others are autonomous variable and the condition

$$Yi = \beta 0 + \beta 1xi1 + \beta 2\ xi\ 2 + \beta 3xi\ 3.....\beta pxip$$

To build up the different direct relapse condition the parameter is gotten from the preparation information and variable are extricated from the dataset utilizing relationship. The amount r, called the straight connection coefficient measure the quality and course of connection between the two factors. The direct connection coefficient is at some point called Pearson item minute relationship coefficient. The numerical formulae for r are given as

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2(\sum x)^2} - \sqrt{n(\sum y^2 - (\sum y^2)}}$$

The coefficient of assurance estimates how well the relapse line speaks to information, if the relapse line goes through each point on the dispersed plot it is ready to clarify the majority of the variety


Data Collection
Data Preprocessing
Data Selection
Reduction of explanatory Predictor
Building model using regression
Validity check

RANDOM FOREST

Random Forest is a procedure utilized for some, reasons including arrangement, relapse and expectation. Such method is a troupe of choice tree which goes for developing a huge number of choice trees inside the preparation and creating the class as a yield. Steps below demonstrates the pseudo code of such calculation

1. for straightforward Tree T
2. for every hub
3. Select m an arbitrary indicator variable
4. In the event that the target work accomplished (m=1)
5. Split the hub
6. End if
7. End for
8. Rehash for all hubs

The exploration procedure has been set to achieve the target of this examination, which spoken to by setting up another group model of numerous AI procedures for precipitation forecast. To do as such, a explore structure which contains a few stages. The primary stage which is the dataset stage is utilized to recognize the information inspected in this examination by representing its source, subtleties and amount. The second stage which is preprocessing readies the information for preparing. The stage incorporates two undertakings; cleaning which will deal with the missing qualities and standardization which plans to restrain the incentive into explicit range. Third is build up similar investigation among the five procedures so as to distinguish the best Regression Algorithm Multiple linear Regression, Random Forest (RF) and AdaBoost.

ADABOOST

Boosting is a general outfit technique that makes a solid classifier from various frail classifiers. This is finished by structure a model from the preparation information, at that point making a second model that endeavors to address the mistakes from the main model. Models are included until the preparation set is anticipated flawlessly or a most extreme number of models are included.

AdaBoost was the first extremely effective boosting calculation created for paired arrangement. It is the best beginning stage for comprehension boosting.

PROPOSED ARCHITECTURE

In this area, we portray the general engineering of our proposed model. As referenced all through the paper, we utilize a profound learning design to foresee the gathered precipitation for the following day. The design is made out of two systems: an autoencoder organize and a multilayer perceptron arrange. The autoencoder arrange is dependable to highlight choice and as referenced. The autoencoder is a profound learning strategy guarantee for the element treatment in time arrangement. A multilayer perceptron organize is in charge of order, forecast undertaking. Next we will detail each system. The principal component in our engineering is the autoencoder. An autoencoder is an unsupervised system that expects to remove non-direct highlights for an information input. Being increasingly explicit, an autoencoder is made by three layers: the info layer, a concealed layer utilizing the sigmoid initiation work, and the yield layer. Diversely to great neural systems, auto encoders are prepared with the goal that the yield layer endeavors to be as comparative as conceivable to the info layer. Along these lines, the shrouded layer results in a non-straight reduced portrayal of the information layer, accomplished gratitude to the sigmoid enactment work. The method of reasoning behind this change is that information will be increasingly minimized (i.e., less inclined to over fitting) and ideally some fascinating non-direct connections that improve the clarification of the yield variable have been found. In our design, the kind of auto encoder that we utilized is a demising auto encoder given by Thaana, a Python GPU-based library for scientific improvement. The concealed layer of the auto encoder, the non-straight minimal portrayal of the first info, is specifically associated with a Multilayer observation. This system is the one in charge of making expectations in our concern, by accepting the new issue portrayal as info. The MLP comprises of one concealed layer and uses the sigmoid enactment work. Figure 1 displays our design; it appears in detail the information and yield layers and the manner by which the auto encoder associates with the MLP arrange.

EXPERIMENTS

So as to assess the execution of the proposed criteria we utilize the Mean Square Error (MSE) and the Root Mean Square Error (RMSE) as estimation blunders. Let Yˆ I be a vector of n forecasts and Yi be the vector of watched esteems comparing to the normal yield of the capacity which produces the expectation, at that point MSE and RMSE can be determined by the condition

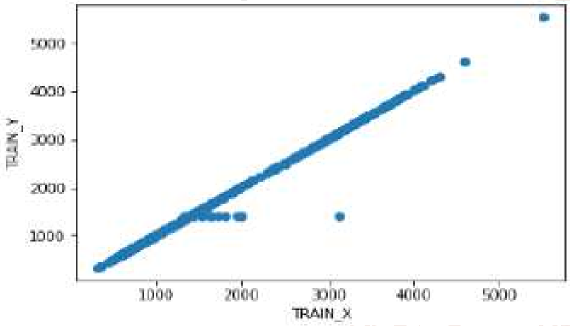
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$RMSE = \sqrt{MSE}$$

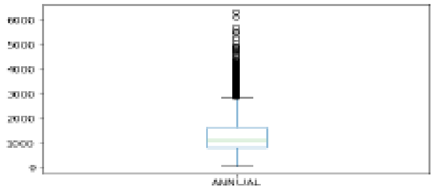
Comparisons of algorithms based on error rates:

Algorithms used:	% of trained data	% of tested data	Mean Squared Error	Root Mean Squared Error	r2_score Error
Multiple Linear regression	70%	30%	3326.4158	57.6751	0.9959
Random Forest Regressor	70%	30%	26617.2233	163.1479	0.9669
Adaboost Regressor	70%	30%	22743.9068	150.8108	0.9717

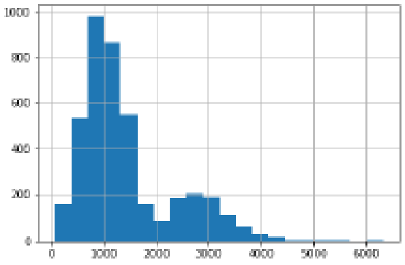
1. **Scatter plot of Multiple Linear regression between annual rainfall and periodic rainfall**



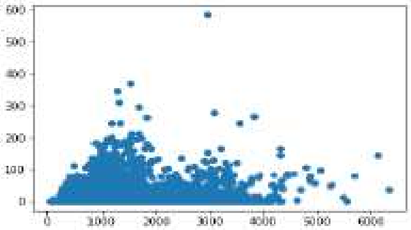
2. **Box plot of annual rainfall data in years 1901-2015**



3. **Histogram showing the annual rainfall of all states**



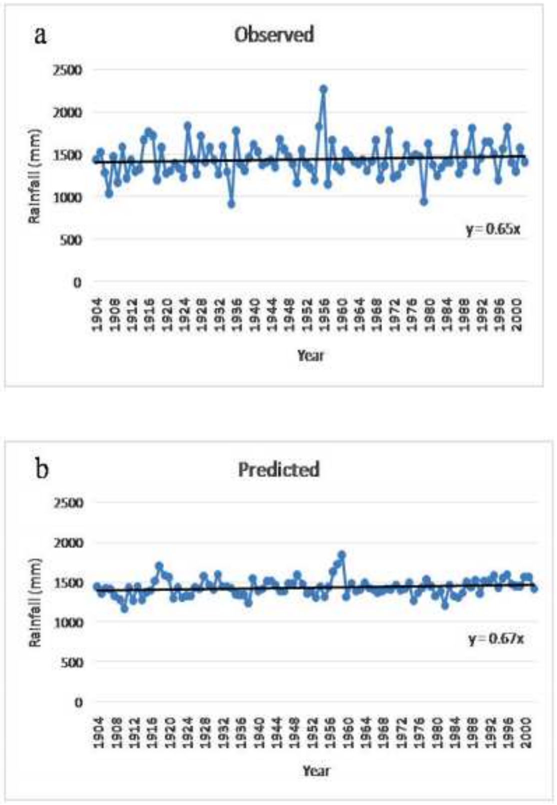
4. **Scatter plot of annual rainfall data in years 1901-2015**



V. RESULTS

The various straight relapse examinations created great results. The coefficients got for the principal, second and third earlier years are 0.492, 0.277 and 0.226 individually. The results got are very persuading as the predominance of the indicator factors are diminishing with that expanding the fleeting hole. The condition or various straight relapse display created is exhibited

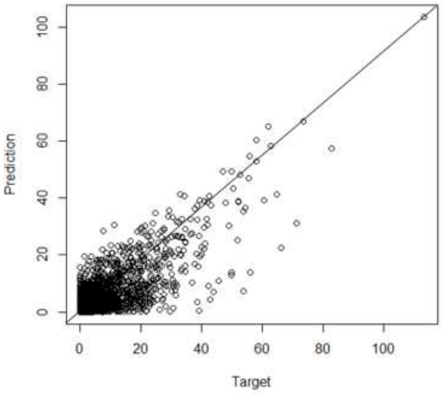
$$P = 0.492X_{-1} + 0.277X_{-2} + 0.226X_{-3}$$



From the approval of this model on yearly precipitation information for 1904-2002, the relapsed esteem has appeared extremely great coordinate with that of watched esteems. The connection coefficient (r) and coefficient of assurance (R2) are the viability measure utilized for approval of this model. The articulations for R2 furthermore, balanced R2 are displayed

$$R^2 = \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N\ \sigma_x \sigma_y}\right)^2$$

For the most part, in the different direct relapse models, the expectations may get one-sided because of the way that, progressively number of indicators causes over fitting of the model. Clearly whether the quantity of indicators expands, the precision of the model likewise increments. So each time another indicator will be presented, it might add to a superior understanding of model with watched esteems, regardless of whether because of chance alone. Besides, more indicators with higher arranged polynomials lead to formation of arbitrary clamors, which additionally creates high estimations of R2, which can be deluding. To battle such issues, the Adjusted R2 is considered which gets balanced as per the quantity of indicator factors. The estimation of Adjusted R2 increments when another indicator improves the model more than would be normal by possibility and diminishes if an indicator improves the model not as much as what is normal by chance. The intriguing thing to note is that the estimation of Adjusted R2 can even be negative, yet for the most part it doesn't happen so. Its esteem will be constantly lesser than R2. For a decent model, the distinction among R2 and Adjusted R2 is little.



VI. CONCLUSION

It is essential to evaluate precipitation appropriately for an improved water assets arranging, advancement and the executives. A different direct relapse demonstrate was created to appraise the yearly precipitation over India, utilizing the yearly precipitation estimations of three earlier years. The model can create great outcome and conveyed a magnificent coordinating with the genuine information in this manner acquiring a high coefficient of assurance (R2) equivalent to0.974 and a balanced R2 of 0.963. Such a high R2 esteem is sufficient to legitimize the ability of the model to assess yearly precipitation over the territory that may help for further hydro meteorological examinations in future.0.974 and a balanced R2 of 0.963. Such a high R2 esteem is sufficient to legitimize the ability of the model to assess yearly precipitation over the zone that may help for further hydro meteorological examinations in future. This paper introduced survey of various techniques utilized for precipitation expectation and issues one may experience while applying distinctive methodologies for precipitation estimating. Because of nonlinear connections in precipitation information and capacity of gaining from the past makes multiple linear regression is the best methodology from every single accessible methodologies. The findings from this study offer a few commitments to the present writing. First, it was shown that the use of machine learning techniques shows a good to significant improvement in rainfall prediction models study area. It is important to note that in general, the multiple linear regression method consistently outperforms the Random Forest (RF) and Adaboost algorithm. Hopefully, the outcomes from this study may help on addressing a suitable machine learning technique that has a significant impact on improving the performance of rainfall forecasting prediction.

VII. REFERENCES

[1] P. Goswami and Srividya, "An epic Neural Network structure for long range gauge of precipitation plan," CurrentSci.(Bangalore), vol. 70, no. 6, pp. 447-457, 1996.

[2] C. Venkatesanet, S. D. Raskar , S. S. Tambe , B. D. Kulkarni , and R. N. Keshavamurty , "Conjecture of all India summer rainstorm precipitation using Error-Back-Propagation Neural Networks," Meteorology and Atmospheric Physics, pp. 225-240, 1997.

[3] A. K. Sahai, M. K. Soman, and V. Satyan, "All India summer storm precipitation expectation utilizing an Artificial Neural Network," Climate elements, vol. 16, no. 4, pp. 291-302, 2000.

[4] N. S. Philip and K. B. Joseph, "On the consistency of precipitation in Kerala-A use of ABF neural system," Computational ScienceICCS, Springer Berlin Heidelberg, pp. 1-12, 2001.

[5] N. S. Philip and K. B. Joseph, "A Neural Network contraption for looking at examples in precipitation," Compute. furthermore, Geosci.,vol. 29, no. 2, pp. 215-223, 2003.

[6] N. Chantasut, C. Charoenjit, and C. Tanprasert, "Farsighted mining of precipitation desires using fake neural frameworks for Chao Phraya River," fourth Int Conf. of the Asian Federation of Inform. Development in Agriculture and the second World Congr. on Comput. in Agriculture and Natural Resources, Bangkok, Thailand, pp. 117-122, 2004.

[7] V. K. Somvanshi, O. P. Pandey, P. K. Agrawal, N.V.Kalanker1, M.Ravi Prakash, and Ramesh Chand, "Showing and estimate of precipitation using Artificial Neural Network and ARIMA techniques," J. Ind. Geophys. Affiliation, vol. 10, no. 2, pp. 141-151, 2006.

[8] S. Chattopadhyay, "Desire for summer rainstorm precipitation over India by Artificial Neural Network with Conjugate Gradient Descent Learning," arXiv preprint nlin/0611010, pp. 2-14, 2006.

[9] S. Chattopadhyay and M. Chattopadhyay, "A Soft Computing system in precipitation gauging," Int. Conf. on IT, HIT, pp. 19-21, 2007. S. Chattopadhyay and G.

[10] Chattopadhyay, "Similar examination among various neural net learning calculations connected to precipitation time arrangement, "Meteorological application., vol. 15, no. 2, pp. 273-280, 2008.

[11] P.Guhathakurta, "Long lead storm precipitation expectation utilizing deterministic Artificial Neural Network demonstrates," Meteorology and Atmospheric Physics 101, pp. 93-108, 2008.

[12] C. L. Wu, K. W. Chau, and C. Fan, "Forecast of precipitation time arrangement utilizing Modular Artificial Neural Networks combined with data preprocessing systems," J. of hydrology, vol. 389, no. 1, pp. 146-167, 2010.

[13] K. K. Htike and O. O. Khalifa, "Precipitation estimating models utilizing Focused Time-Delay Neural Networks," Comput. also, Commun. Eng. (ICCCE), Int. Conf. on IEEE, 2010.