Predicting Annual Rainfall for the Indian State of Punjab Using Machine Learning Techniques

Sunil Kaushik,Akashdeep Bhardwaj
University of Petroleum and Energy Studies
Dehradun. India
sunil.kaushik@rediffmail.com,
abhardwaj@ddn.upes.ac.in

Luxmi Sapra
Chitkara University
Rajpura. India
luxmi.verma@gmail.com

2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) | 978-1-7281-8337-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICACCCN51052.2020.9362742

***Abstract*—Excess and deficit of rain has always been a concern for an agricultural state like Punjab. Measure to save crop can be taken if the rainfall values are predicted in advance. The current study attempts to solve rainfall prediction problem using machine learning techniques. Current study evaluates 3 machine learning algorithms – KNN, ELM and SVM applied on the rainfall data and other parameters – humidity, wind speed, max temperature and min temperature for the data from 1973 to 2019. These algorithms were compared on the universal performance parameters – MAE, RMSE, SD, PP and time to predict. We found out that SVM predicts the values nearest to the observed values in the trainer data set and test dataset. The SVM predicted values were not only close to predicted values but also had the least RMSE, MAE and ET. SVM predicted the results with 95% and 92% accuracy for trainer data set and test data set. SVM could have shown better results if number of data points had been more.**

Keywords—Rainafall, Punjab, Machine Learning, SVM,ELM, KNN.

I. INTRODUCTION

Punjab, a northern state in India (between 29’30’’ N to 32’32’’ N latitude and 73’55 E to 76’50 E longitude) ,shown in Figure 1, is known for producing 40% of Rice and 50-70% of wheat of total India’s production [22]. It is often called as considered to be food basket of India. The agriculture is not only dependent on the fertile land of India but also it depends heavily on irrigation by rain or tubewell. Water table of the tubewell is replenished by the rain. Hence, it can be easily said that rain fall is very important to sustain the agriculture of the Punjab state of India, Accurately forecasting the rain is very important in Punjab where not only agriculture is dependent on rain but also the danger of flood which creates problems in the livelihood of the residents and to maintain the growth of the industry such as textile and steel which are water intensive and affect Punjab’s economy.

The knowledge of prediction of rainfall can help in choosing the crop and deciding the contingencies or alternatives that will help in absorbing the impact on GDP growth.

Previous studies [8.9] on the rainfall of Punjab is done through statistical analysis such as Log Normal, Gumbel and Log Pearson Type-III, non parametric and spatial analysis. Previous studies have found a decreasing trend and the inconsistency in the annual rain fall in Punjab. We started taking interest in the rainfall of the Punjab when floods were reported in Punjab in 2019. More than 300 villages were drown killing apx 1300 people and crops over 4000 hectares were reported to be damaged by floods in 2019.

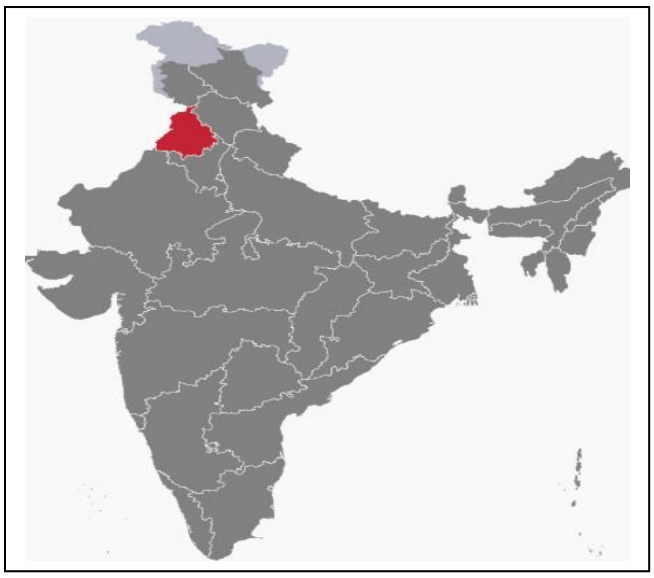

Fig. 1. Location of Punjab in India (Marked in Red Color)

Predicting rainfall is a challenging and data intensive task[1,2,3,4]. It not only depends on the so many variables that may be local or far environmental or it may be depending on the non environment variables [7,8,9,10]. However, recently machine learning approaches have been seen to be used in predicting the outcome of various phenomena in different areas ranging from science, economics and games[11,12]. Machine learning approaches analyze data which can be liner or non linear and come out with the predicted value with good accuracy. Thus, machine learning approaches are used as “universal approximators”. Hence, we deliberated on applying the machine learning approaches to predict the rainfall in Punjab where climate is complex and impacts are high. Previous researches on similar area has shown that phenomena of rain is dependent on many variables and approximates the rain falls based on the values which makes it eligible for using classification algorithms [15,16].

This paper discusses the dataset , ML techniques used (SVM, ELM and KNN), the outcomes and comparison of the outcome in the subsequent sections.

II. METHODOLOGY

A. Dataset Used

The climatic data was collected from the Govt of India website (http://data.gov.in) for the state of Punjab from 1973 to 2019[5]. There was no single file of the data but data for the various variables ,such as below, was taken for said years and one consolidated file was prepared with all the values. All the data given followed SI system of units. No other processing of data was done.

B. Selection of Training Set and Testing Set

As given in previous section that time series data for Punjab state for the year 1973 to 2019 was considered. As all the ML techniques used were supervised, the data was divided in to two parts -

1. Training Set – Time series data of the annual rain fall (RAIN), Minimum Temperature (MINT), Maximum Temperature (Max T), Average Wind Speed (WIND) and Average Humidity (HUMD) for the year 1973 to 2008 were considered for the training of the ML Models. In the current study the we tried to adjust the training window in interval of 5 years to check the accuracy of the model.
2. Testing Set – The data of the year 2009-2019 was considered for testing the model.

C. Machine learning Techniques used

- KNN - K-nearest neighbor(KNN) is a machine learning technique that classifies the data objects in to various categories based on the few paramerters which are close its nearest neighbor. KNN uses the Euclidean distance and the data object which have the least Euclidean distance is qualified as the nearest neighbor [13,14]. The distance is calculated as

$$D_{x,y} = \sqrt{\sum_{n=1}^{N} \frac{1}{S_n^2} (x_n - y_n)^2} \quad (1)$$

Here S – standard deviation of the n^{th} data object.

In this study, we tested the data for k = 1.. 12 and optimum performance was seen at the k =8.

- ELM – Extreme learning machine (ELM) are performance driven single node feed forward neural network is a machine learning technique where all the hidden neurons are tuned in single go and help in solving the complex approximation problems faster. The output of the ELM algorithm is given as below in equation 2 [19,20,21]

$$f_L(x) = \sum_{i=1}^{L} \beta_i h_i(x) = h(x)\beta \quad (2)$$

Here

- β represents the weight of output for L nodes.
- h represents the non linear feature of hidden node

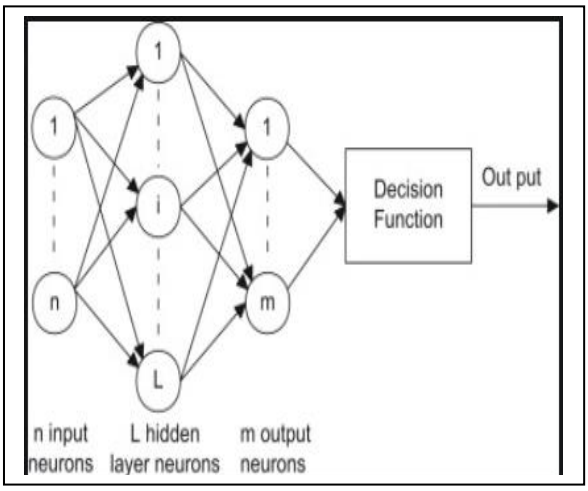

Fig. 2. ELM Schematic representation

- SVM – Support Vector Machine (SVM) is a supervised ML algorithm which is used for classification or regression and is known for solving complex problems [10,11,12]. To understand the SVM let us take a line represented by the equation

$$f(x) = (w.x)+b \quad (3)$$

where $w \in R^n$, $b \in r$. Here w is the adjustable weight of the vector , b scalar threshold and R is vector of dimension n.

Figure 3 provides the pictorial representation of the SVM.

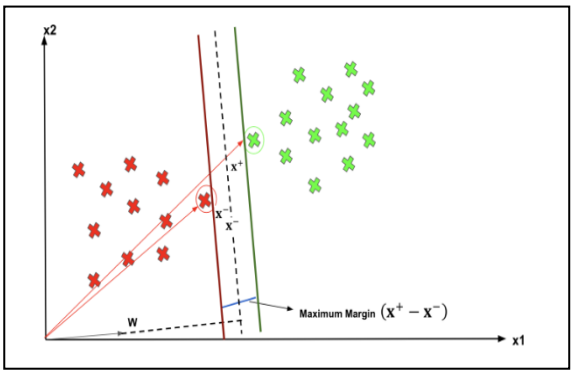

Fig. 3. SVM Categorization

D. *Performancy Parameters*

Standard statistical indicators were used to measure the performance of the ML models. These indicators are given below as –

- Mean Absolute Error (MAE)– This can be defined as ratio of absolute difference in the observed value and predicted values to predicted value. It can be calculated with the following formula [18,19].

$$MAE\ (\%) = \frac{100}{x}\sum_{i=1}^{x}\left|\frac{Obs_i - Prd_i}{Obs_i}\right| \quad (4)$$

- Standard Deviation (SD) – This can be described as amount of dispersion or difference from the mean value. Also, it can be defined as the root of the variance for a data set [18,19]. Mathematically it can be called as

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}, \quad (5)$$

where x_i is the value of data ordinate at i^{th} position $\bar{x}$ *is mean of data set* $\{x_i \ldots x_n\}$

- Root Mean Square (RMSE) -

$$RMSE\ (\%) = \sqrt{\frac{100}{x}\sum_{i=1}^{x}\left(\frac{Obs_i - Prd_i}{Obs_i}\right)^2}$$

- Elapsed Time (ET) – Time taken to predict the value of entire set. It is measured in Seconds [18,19].

E. *Experimental Process*

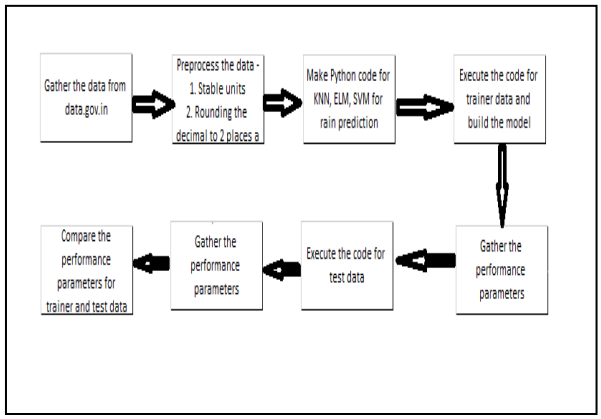

Fig. 4. Process flow of experiments

III. Results and Discussion

The annual rainfall of Punjab is wide spread. There had been years when the annual rainfall was very much more than the average rainfall and much lower than the annual rain falls. Mostly it falls in range of 400 mm to 600mm. The figure below represents the annual rain fall for the years from 1973 to 2019. We can easily see that observed value of rainfall is scattered and observe no standard pattern and don't follow any model.

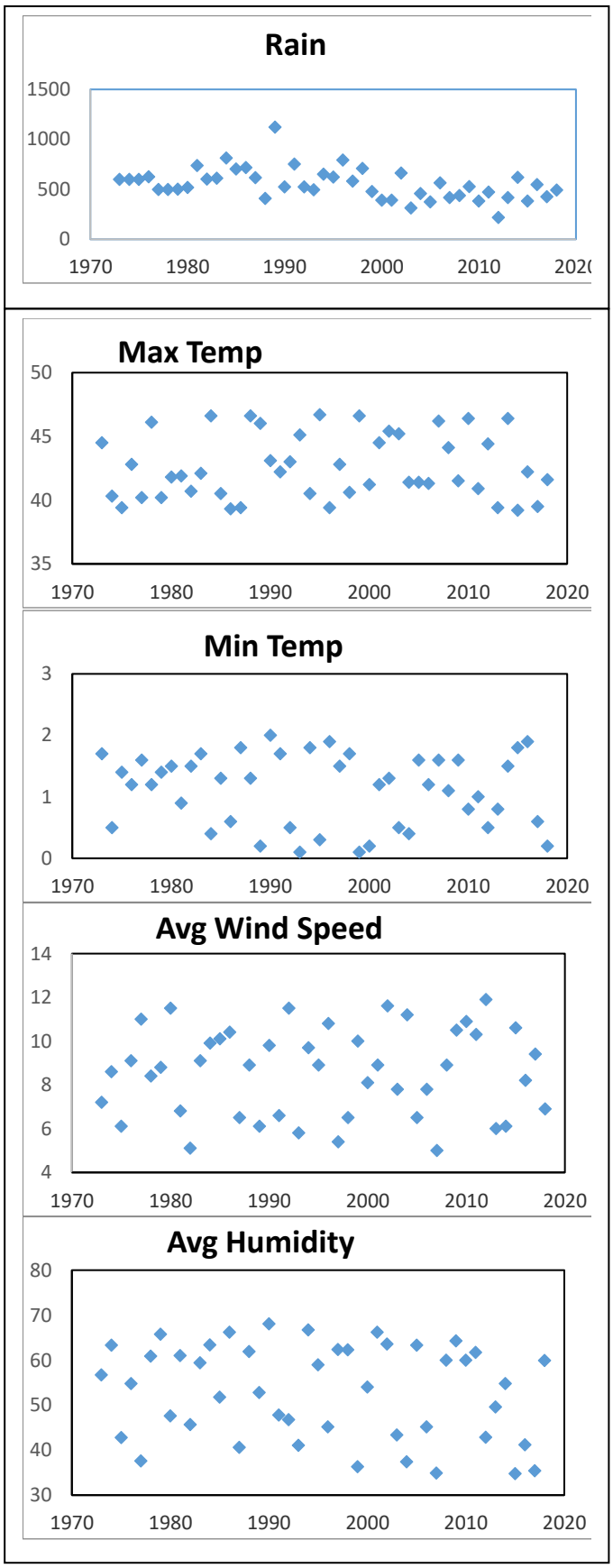

Fig. 5. Distribution of the parameters affecting rainfall 1973-2019

The rainfall is dependent on the parameters such as Average Humidity, Average Windspeed and Max Temperature and Minimum temperature for the year.

Looking at the fig 6, we concluded that there is no specific pattern of model that describes each of these parameters and no individual correlation between rain fall and these parameters can be seen. This is also validated using the heatmap of the pearson correlation coefficient (r) for rainfall and each of the parameters. This is depicted in the figure 7. The maximum correlation is seen between humidity and the rainfall , followed by minimum temperature of the year. Noteworthy to point that windspeed and maximum temperature showed a negative correlation with rain..

	Min Temp	Avg W S	Avg Hum	*Max Temp*	*Rain*
Min Temp	1.00	-0.13	0.10	-0.20	0.07
Avg W S	-0.13	1.00	0.04	-0.03	-0.18
Avg Hum	0.10	0.04	1.00	0.13	0.21
Max Temp	-0.20	-0.03	0.13	1.00	-0.04
Rain	0.07	-0.18	0.21	-0.04	1

Fig. 6. Heat Map of the correlation between rainfall and parameters

Figure 5,6, and 7 collectively suggest that approximation of the rainfall is a complex task and requires modern techniques such as ML techniques to predict the rainfall for the Indian state of Punjab.

Prediction performance of the trainer data for the annual rainfall approximation for Punjab is given in the below table (Table 1). The mean average error (MAE), root mean square error (RMSE), standard deviation (SD), performance parameter (PP) and elapsed time by each ML algorithm (KNN, ELM and SVM) were computed. The trainer data set consisted of the data for the year (1973-2008).

TABLE I. PREDICTION PERFORMANCE OF ML ALGORITHMS FOR TRAINER DATASET

	MAE	RMSE	SD	PP	ET
KNN	4.2	36.5	149.7	0.76	18
ELM	3.09	25.9	152.6	0.83	21
SVM	0.9	6.9	152.8	0.95	18

KNN showed the best efficiency on the trainer set at the k-value of 8 and ELM showed the best efficiency with 12 hidden nodes. The MAE score of KNN, ELM and SVM were found to be at 4.2, 3.09 and 0.9 with the RMSE of 36.5, 25.9 and 6.9 for KNN, ELM and SVM respectively. It can be easily concluded that SVM predicted values fit well the observed curve and low MAE of SVM also suggests that predicted values of the rainfall. Low MAE and low RMSE value of SVM predicted values of the rainfall mean that values are prediction error are low and provide a good fit with the observed values.

The fitness of predicted values and the observed values for part of the trainer dataset is shown in the figure 8. It can be seen that SVM predicted values fit the observed values. The SVM is shown to be an fitness of 95% while ELM and KNN have shown the fitness of 83% and 76%.

Fig. 7. Prediction performance for trainer data for KNN, ELM, SVM.

Prediction performance of the trainer data for the annual rainfall approximation for Punjab is given in the below table (Table II). The mean average error (MAE), root mean square error (RMSE), standard deviation (SD), performance parameter (PP) and elapsed time by each ML algorithm (KNN, ELM and SVM) were computed. The test data set consisted of the data for the year (2009-2019).

TABLE II. PREDICTION PERFORMANCE OF ML ALGORITHMS FOR TEST DATASET

	MAE	RMSE	SD	PP	ET
KNN	7.9	38.3	101.9	0.62	8
ELM	1.15	24.5	96.7	0.75	19
SVM	1.7	7.6	100.4	0.92	13

The configurations which showed the best efficiency for trainer data set were considered. Here KNN was taken with the k-value of 8 and ELM was considered at 12 hidden nodes. The MAE score of KNN, ELM and SVM were found to be at 7.9, 1.15 and 1.7 with the RMSE of 38.3, 24.5 and 7.6 for KNN, ELM and SVM respectively. It can be easily concluded that SVM predicted values fit well the observed curve and low MAE of SVM also suggests that predicted values of the rainfall. Low MAE and low RMSE value of SVM predicted values of the rainfall mean that values are prediction error are low and provide a good fit with the observed values.

The fitness of predicted values and the observed values for part of the trainer dataset is shown in the figure 9. It can be seen that SVM predicted values fit the observed values. The

SVM is shown to be an fitness of 92% while ELM and KNN have shown the fitness of 76% and 62%.

On carefully observing the table I and table II , it is seen that all the ML algorithms showed deterioration in the performance when the number of data points decreased in a set. The difference in the SD of ELM and SVM was minimal for the trainer set but the SD increased for SVM for test set and it decreased significantly for the ELM. However, elapsed time did not reduce significantly on reducing the number of data points in a dataset. The SVM algorithm has always taken less time in predicting the values [11,12] and ELM algorithm [18,19,20,21] has always taken the highest time among the algorithms in the study.

Fig. 8. Prediction performance for test data for KNN, ELM, SVM.

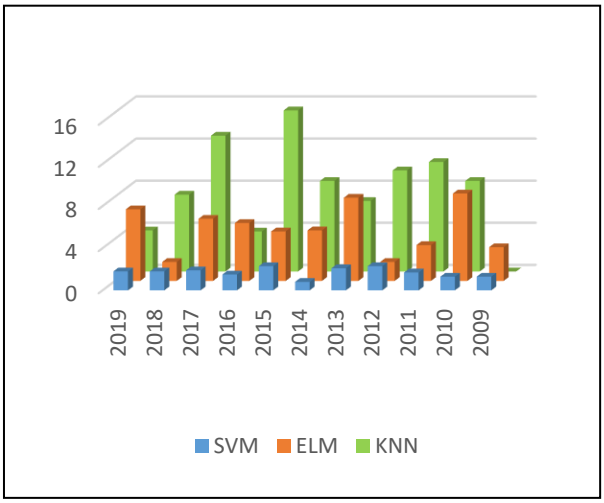

Fig. 9. Comparision of MAE of test data set in SVM, ELM and KNN

All the values predicted by SVM for the test data set were closed to the observed values of rainfall. Average MAE for SVM was 1.7% with a range of 0.8% to 2.1%. While the ELM and KNN showed the average MAE of approximately 5% and 8% respectively. Hence, though KNN is easy to use and takes reasonably lesser time compared to ELM and SVM but the average error in prediction is highest in KNN followed by ELM and SVM.

IV. CONCLUSION

The study compares three widely used ML algorithms- KNN, ELM and SVM and tries to find out the solution of the practical problems. It can be easily concluded that SVM performed best with least ET, MAE , RMSE and predicted the values those fit well the curve of the observed rainfall. ELM had taken the most time because of the complex calculation of weights in the hidden node. ELM were found to be CPU intensive and the CPU usage was seen to be increasing by 15-20% while performing ELM and CPU utilization hardly increased by 5% for KNN. The KNN and ELM could have performed better if the data set was bigger and number of parameters were more than the current study. We would like to expand this study for other states and include more parameters such as direction of wind, month of the year , nearness to see and height from sea level.

ACKNOWLEDGMENT

We would like to thank Govt of India for creating such a beautiful website – data.gov.in that helps in getting the data for the academic purpose.

REFERENCES

[1] Damodaran VK, Krishnan KP. Energy and Climate Change in Kerala. Proceedings of Kerala environment congress. 2015. p. 3–10. National workshop on climate change and sustainable development 2015.

[2] Guhathakurta P, Rajeevan M. Trends in the rainfall pattern over India. Int J Climatol 2008;28:1453–69.

[3] Nikhil Raj PP, Azeez PA. Trend analysis of rainfall in Bharathapuzha river basin, Kerala, India. Int J Climatol 2012;32:533–9.

[4] Nair A, Joseph KA, Nair KS. Spatio-temporal analysis of rainfall trends over maritime state (Kerala) of India during the last 100 years. Atmos Environ 2014;88:123–32.

[5] Data on Punjab- https://data.gov.in. (accessed 15.05.2029).

[6] Dash Y, Mishra SK, Panigrahi BK. Rainfall prediction of a maritime state (Kerala), India using SLFN and ELM techniques. International conference on intelligent computing, instrumentation and control technologies (ICICICT). 2017. p. 1714–8.

[7] Dash Y, Mishra SK, Panigrahi BK. NEMR predictability assessment over Indian peninsula using ELM. In: Lyubchich V, Oza NC, Rhines A, Szekely E, editors. Proceedings of the 7th international workshop on climate informatics: CI 2017, NCAR technical note NCAR/TN-536+PROC 2017. p. 77–80. http://dx.doi.org/ 10.5065/D6222SH7.

[8] Dhaliwal, L. K. "Variability in meteorological parameters during kharif season and its impact on rice crop at Ludhiana, Punjab, India." Journal of Agrometeorology 22, no. 1 (2020): 92-96..

[9] Kaur, J., Kingra, P. K., Setia, R., & Singh, S. P. (2019). Relationships among wheat yield, climate and technology variables in different agro-climatic zones of Punjab. Agricultural Research Journal, 56(3), 436-443.

[10] Karunakaran, V., Joseph, S. I., Teja, R., Suganthi, M., & Rajasekar, V. (2019). A wrapper based feature selection approach using bees algorithm for extreme rainfall prediction via weather pattern recognition through svm classifier. International Journal of Civil Engineering and Technology (IJCIET), 10, 1745-1750.

[11] Samantaray, S., Tripathy, O., Sahoo, A., & Ghose, D. K. (2020). Rainfall Forecasting Through ANN and SVM in Bolangir Watershed, India. In Smart Intelligent Computing and Applications (pp. 767-774). Springer, Singapore.

[12] Zhang, Y., Yang, H., Cui, H., & Chen, Q. (2019). Comparison of the Ability of ARIMA, WNN and SVM Models for Drought Forecasting in the Sanjiang Plain, China. Natural Resources Research, 1-18.

[13] Pham, B. T., Le, L. M., Le, T. T., Bui, K. T. T., Le, V. M., Ly, H. B., & Prakash, I. (2020). Development of advanced artificial intelligence models for daily rainfall prediction. Atmospheric Research, 104845.

[14] Amale, O., & Patil, R. (2019, March). IOT Based Rainfall Monitoring System Using WSN Enabled Architecture. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 789-791). IEEE.

[15] Singh, G., & Kumar, D. (2019, January). Hybrid Prediction Models for Rainfall Forecasting. In 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 392-396). IEEE.

[16] Murad, S. H., Mohammed, Y., & Salih, M. (2020). COMPARABLE INVESTIGATION FOR RAINFALL FORECASTING USING DIFFERENT DATA MINING APPROACHES IN SULAYMANIYAH CITY IN IRAQ. International Journal, 4(1), 11-18.

[17] Jimenez, S., Aviles, A., Galán, L., Flores, A., Matovelle, C., & Vintimilla, C. (2019, November). Support Vector Regression to Downscaling Climate Big Data: An Application for Precipitation and Temperature Future Projection Assessment. In Conference on Information Technologies and Communication of Ecuador (pp. 182-193). Springer, Cham.

[18] Dash, Y., Mishra, S. K., & Panigrahi, B. K. (2019, July). Neural Network Based Approaches for Prediction of the Indian Summer Monsoon Rainfall. In 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT) (Vol. 1, pp. 550-554). IEEE.

[19] Dash, Y., Mishra, S. K., & Panigrahi, B. K. (2019). Predictability assessment of northeast monsoon rainfall in India using sea surface temperature anomaly through statistical and machine learning techniques. Environmetrics, 30(4), e2533.

[20] Hu, J., Liu, B., & Peng, S. (2019). Forecasting salinity time series using RF and ELM approaches coupled with decomposition techniques. Stochastic Environmental Research and Risk Assessment, 33(4-6), 1117-1135.

[21] Maimaitiyiming, M., Sagan, V., Sidike, P., & Kwasniewski, M. T. (2019). Dual activation function-based Extreme Learning Machine (ELM) for estimating grapevine berry yield and quality. Remote Sensing, 11(7), 740.

[22] *Brass, Paul R. (2005). Language, Religion and Politics in North India. iUniverse. p. 326.* ISBN 978-0-595-34394-2.