জার্ণাল অব ইএলটি এ্যান্ড এডুকেশন | JOURNAL OF ELT AND EDUCATION
ISSN: 2618-1290 (Print), 2663-1482 (Online), Volume 5, Issue 4, December 2022, Page: 86-93
Published by, Hello-Teen Society & Center for Academic Research and Development (CARD)

[**Citation:** Shafiq, S. I. (2022). Machine Learning Techniques for the Investigation of Rainfall in Bangladesh. *Journal* of *ELT and Education*, *5*(4), 86-93.]

Research Paper

MACHINE LEARNING TECHNIQUES FOR THE INVESTIGATION OF RAINFALL IN BANGLADESH

Shafiqul Islam Shafiq*

Abstract

Rainfall significantly affects human lives in various contexts, including natural calamities like landslides, droughts, and floods. In these industries, artificial intelligence technologies have a strong probability of success. The development of a machine learning-based rain prediction model is the aim of this study. This model is based on daily and monthly rainfall data (millimeters) over the period of 1970 to 2020 from the 32 weather stations of Bangladesh obtained from the Bangladesh Meteorological Department (BMD). Significant rainfall since the projection is so directly related to the economy and human lifespan, it could be a significant setback for the earth science division. For nations like Bangladesh, where agriculture is the mainstay of the economy, the accuracy of the rainfall statement is very important. Python programming language is used to investigate and predict precipitation using machine learning techniques and to compare different machine learning models. There was a total of six models used during the experiment. Each model was trained using five input characteristics, and then the precipitation projections were validated. All of the model's performance results were positive. However, Random Forest successfully predicts the dataset, making it the best option for this investigation. Higher accuracy and low Mean Absolute Error were produced using the Random Forest classifier. The amount of rainfall is observed to rise throughout the months of July, August, and September. The year 2017 saw an overall rainfall maximum of 99790.0 mm while the year 1972 saw an overall rainfall minimum of 32020.0 mm. And 71827.59mm is the average rainfall over all 32 weather stations. The month of July saw the highest monthly average rainfall of 539.10mm, and the month of January saw the lowest monthly average rainfall of 7.45mm. Additionally, 202.96mm of rain falls on average per month. The seasonal amount of rainfall in the 32 meteorological stations that are represented reveals that Cox's Bazar, Sylhet, and Teknaf have good volumes of rainfall.

Keywords: Rainfall, Weather, Python, Machine Learning, Prediction

1. Introduction

The most significant part of the hydrological sequence is rainfall, which is also the main basis of seasonal rainfall that might have been exploited and controlled specifically to fulfill ingesting needs. Water is the most priceless natural resource that has a big impact on both social and economic growth. In many parts of the world, there is general stress on water resources or an imbalance between supply and demand. This is brought on by a surge in demand, a shortage of readily

*Corresponding Email: sishafiq19@gmail.com
Data Analyst, Bangladesh Oceanographic Research Institute, Cox's Bazar-4730, Bangladesh

accessible water, or inadequate management of the water resources currently in existence. Actually, the best ways to reduce water stress or close the supply-demand gap are through efficient water planning and management. Numerous attempts have been made on a global scale to forecast rainfall and comprehend the relationship between precipitation and runoff in order to modify water management techniques and guarantee that water is accessible whenever it is needed (Tikhamarine et al., 2020).

Rainfall in Bangladesh varies depending on the time of year and geography. Only 1.54% of the annual rainfall falls during the very dry winter months of December to February. Due to westerly disturbances that originate in the northern areas of Bangladesh, precipitation varies throughout this season from 20 mm in the west and south to 40 mm in the northeast. In Bangladesh, the average precipitation varies from less than 1500 millimeters in the west and central areas to well over 3000 millimeters in the northeast and south. The Surma Valley and the neighboring hills get considerable precipitation. The average annual rainfall in Sylhet is 4180 mm, in Sunamganj it is 5430 mm at the base of the sloping Meghalaya Plateau, and in Lalakhal, the maximum height in Bangladesh is 6400 mm. Scientists from all across the world are working to increase the accuracy of rainfall predictions. Several statistical and mathematical models have been developed to anticipate precipitation. But as soft computing techniques continue to progress, numerical models' importance in the field of forecasting is waning. In the past few decades, artificial neural networks have emerged as an inductive technique for hydrological and clamatorial forecasting due to their ability to model both linear and nonlinear systems without previous knowledge of catchment behavior or flow processes. A rainfall data forecasting system that is based solely on the BPNN model can be used to provide early warning of the possibility of a landslide. The study offers four back-propagation neural networks with varying topologies that can forecast the severity of the following day's rainfall (Devi et al., 2014). Among the input characteristics were rainfall thresholds and previous rainfall histories. Support vector machine (SVM) is a cutting-edge machine learning approach that has been proven effective in handling difficult non-linear machine learning issues. Support vector regression (SVR), a variant of SVM, is used to address nonlinear regression problems. The accuracy of rainfall forecasting is improved by comparing several contemporary approaches and algorithms. The proposed study investigates the precision of monthly rainfall forecasts based on time-series analytic approaches (Srivastava et al., 2020), with historical precipitation intensity acting as the only input parameter. Various machine learning approaches, including linear regression, Support Vector Regression (SVR), lasso regression, Random Forest, XGBoost, and Light Gradient Boosting Machine (Light GBM) are used to create predictions. To calculate the possibility of a landslide, it is possible to compare rainfall projections for a certain site to the precipitation threshold for initiating a landslide. Based on daily climatic data, we have predicted rainfall in this study. The data was used to train and evaluate six machine learning algorithms. The article goes on like this from there: A thorough summary of the pertinent papers is included in the next section of the linked work. The dataset and the algorithm utilized in our work are then described in the methods section. The results and discussion section follows, where all of the models' findings and conclusions are discussed. Our effort is compiled in the last portion.

2. Related Research

The important investigations that other authors have carried out that are relevant to the study subject are summarized in this section. Rani et al. (2020) developed a flood detection method and an alerting system using machine learning techniques. To anticipate rainfall, they carried out research utilizing regression analysis and support vector machines. They employed the Particle Swarm (PSO) and Multi-Layered Perceptron (MLP) methods (As cited in Abdel-Kader et al., 2021). To forecast rainfall, four variables were tested: high and low temperatures, relative humidity, and wind speed. In order to forecast the typical rainfall during the rainy season in the Indian state of Odisha, Misra et al. (2020) utilized a machine learning technique. In their investigation, random forest regression was able to reach a maximum accuracy of 91%. Zhang et al. (2020) used Support Vector Regression and Multilayer Perception to estimate the maximum rainfall possible during monsoon and non-monsoon

seasons. Their study variables include monthly average temperature, wind speed, humidity, and cloud cover. The outcomes were contrasted with those of EBA4SUB, a physiologically event-based method. Utilizing rule-based machine learning, Anwar et al. (2020) built a rain prediction model using historical meteorological data. The decision tree model in their study had the best accuracy. The Extreme Gradient Boosting model was used by Anwar et al. (2021) to forecast rainfall. They used meteorological information gathered by meteorological stations over the past seven years for their dataset. The maximum temperature, minimum temperature, average humidity, time spent in the sun, and maximum wind speed, wind speed, and precipitation are the 8 attributes of their dataset. Using a Raspberry Pi 3 B, Singh et al. (2019) suggested an approach and created an application. This software forecasts whether it will rain today using actual information from humidity, temperature, and pressure sensors. Different machine learning techniques were employed by Park et al. (2020) to calculate the likelihood of flooding in South Korea's coastal areas. Using machine learning techniques, Dikshit et al. (2020) predicted temporal drought patterns in the southeast Australian state of New South Wales. To estimate the Standardized Streamflow Index for hydrologic dryness in this paper, machine learning approaches are once again applied (Shamshirband et al., 2020). Crop yield is predicted using machine learning methods (Klompenburg et al., 2020). They do research that is unique from other researchers in terms of scope, geological setting, and crop (Khan et al., 2020).

3. Methodology

This section will go through the methods and procedures employed in this study's data preparation and gathering process. We trained and evaluated six of the most well-liked machine learning models using our dataset. Precipitation forecasting is done using the predictive model. The first step in conducting experiments is translating data into the proper format, followed by a thorough examination of the data and observation of variations in rainfall patterns (Fig-1).

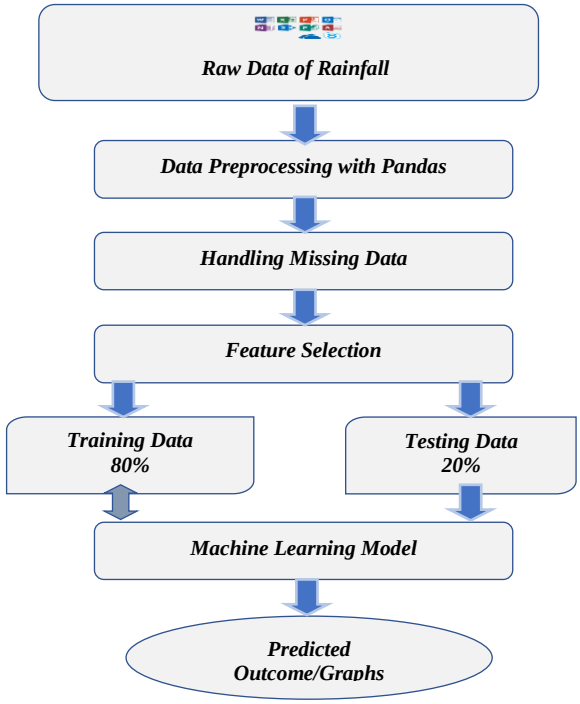


Fig-1: Rainfall Prediction Model

The amount of precipitation is forecasted by splitting the dataset into a training set and testing set, then using various statistical and machine learning techniques to compare and analyze the various methods. We try to reduce the error by using a variety of methods.

3.1. Dataset

The vital component of the research study is the dataset. At the Bangladesh Meteorological Department (BMD) in Dhaka, Bangladesh, all kinds of weather data are provided on a daily or monthly basis. In this study daily and monthly rainfall data (measured in millimeters) over the period of 1970 to 2020 of the 32 weather stations [Barishal, Bhola, Chandpur, Chittagong, Khulna, Kutubdia, M_Court, Madaripur, Mongla, Mymensingh, Patuakhali, Rajshahi, Rangamati, Rangpur, Sandwip, Satkhira, Sitakhunda, Srimangal, Sylhet, Tangail and Teknaf] of Bangladesh were obtained from the Bangladesh Meteorological Department (BMD). For looking a brief overview of the dataset utilized in the study, was created by the amazing Python tool "Pandas Profiling," which is mainly used for machine learning in form of dataframes. Initial calculations were used to determine total monthly precipitation from daily precipitation row data. Secondly, based on monthly total rainfall, annual and seasonal rainfalls were computed through the Python pandas package. For all ML models, the dataset was split for testing in 20% of cases and training in 80% of cases.

4. Machine Learning Algorithms

4.1. Linear Regression

The linear model is used in different ways according to the context. The most common occurrence is in connection with regression models and the term is often taken as synonymous with a linear regression model. However, the term is also used in time series analysis with a different meaning. In each case, the designation "linear" is used to identify a subclass of models for which a substantial reduction in the complexity of the related statistical theory is possible.

4.2. Support Vector Regression

A supervised machine learning approach called Support Vector Machine (SVM) can be applied to problems involving classification and regression. Each data point is represented as a point in n-dimensional space (where n is the number of features), and each feature's value is represented by a specific coordinate. Then, we carry out classification by identifying the hyper-plane that effectively distinguishes the two classes. A feature space, which is a finite-dimensional vector space with each dimension representing a "feature" of a given item, is created by a support vector machine to model the situation. The SVM's objective is to train a model that categorizes newly discovered items. It achieves this by creating a linear partition of the feature space into two categories.

4.3. Lasso Regression

Lasso regression is an approach to regularization. For a more precise prediction, it is utilized instead of regression techniques. It makes use of shrinking. Data values shrink towards a mean, which is a center point, in shrinkage. The lasso approach promotes straightforward, sparse models. When a model exhibits significant levels of multicollinearity or when certain steps in the model selection process, such as variable selection and parameter removal, need to be automated, this specific sort of regression is a good choice. L1 regularization technique is used in Lasso Regression. Because it automatically performs feature selection, it is used when there are more features.

4.4. Light GBM

The Light GBM framework uses gradient boosting and decision trees to improve model performance while using less memory. Since high-dimensional data are frequently quite sparse, we may create a practically lossless method of reducing the number of features. In particular, many features in sparse feature space are mutually incompatible, never taking simultaneous nonzero values. It is safe to combine the exclusive features into a single feature.

4.5. XGBoost

It is an implementation of gradient boosted decision trees designed for speed and performance that is dominative competitive machine learning. It is a rapid and efficient distributed gradient boosting library that has been tuned. XGBoost can give you the tools you need to use decision trees to create a stronger model. XGBoost examines the leaves of the existing decision tree and wonders whether the model would be improved by converting each leaf into a new "if" statement with distinct predictions.

4.6. Random Forest

Random Forest is a multipurpose, user-friendly machine learning technique that, in the majority of situations, produces good outcomes whether or not the hyper-parameters are changed. It is also one of the most used algorithms because it is easy to use and easy to change.

5. Experimental Results

To understand the rainfall in different regions, data on rainfall from 1970 to 2020 is gathered, analyzed, and plotted. The histograms for the monthly, yearly, and consecutive three-month rainfall data are shown below (Fig-2). In the months of July, August, and September, it is seen that there is an increase in the amount of rainfall (Y-axis). The below plot is the line graph (Fig-3) for the amount of

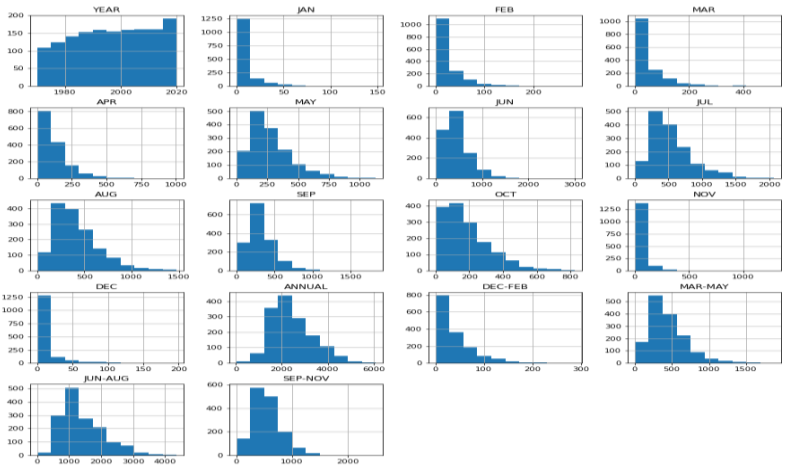


Fig-2: Histograms of the rainfall data monthly, annual and consecutive of three months

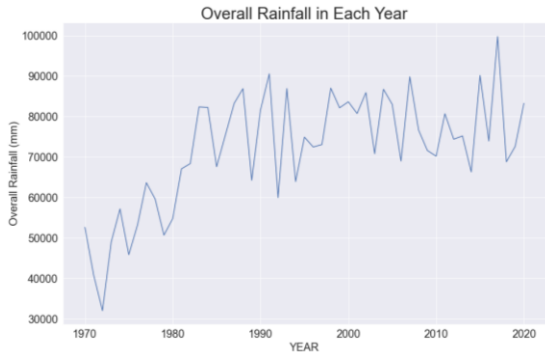


Fig-3: Line plot of the year wise overall rainfall

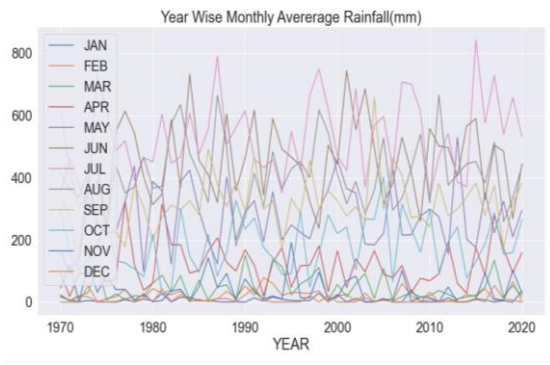


Fig-4: Line plot of year wise monthly average rainfall

Maximum overall rainfall (sum of all 32 weather stations) of 99790.0mm occurred in the year 2017, and the Minimum overall rainfall (sum of all 32 weather stations) of 32020.0 mm occurred in the year 1972. And the average rainfall of all 32 weather stations is 71827.59mm. The line graph (Fig-4) for the amount of maximum overall average rainfall is 539.10mm occurred in the month of July, and the minimum overall average rainfall is 7.45mm occurred in the month of January. And the average rainfall of all months is 202.96mm. In the 32 weather stations represented in the bar graph (Fig-5), the seasonal amount of rainfall is shown. It can be seen that during the rainy season, the volume of rainfall is sensibly good in the areas of Cox's Bazar, Sylhet, and Teknaf.

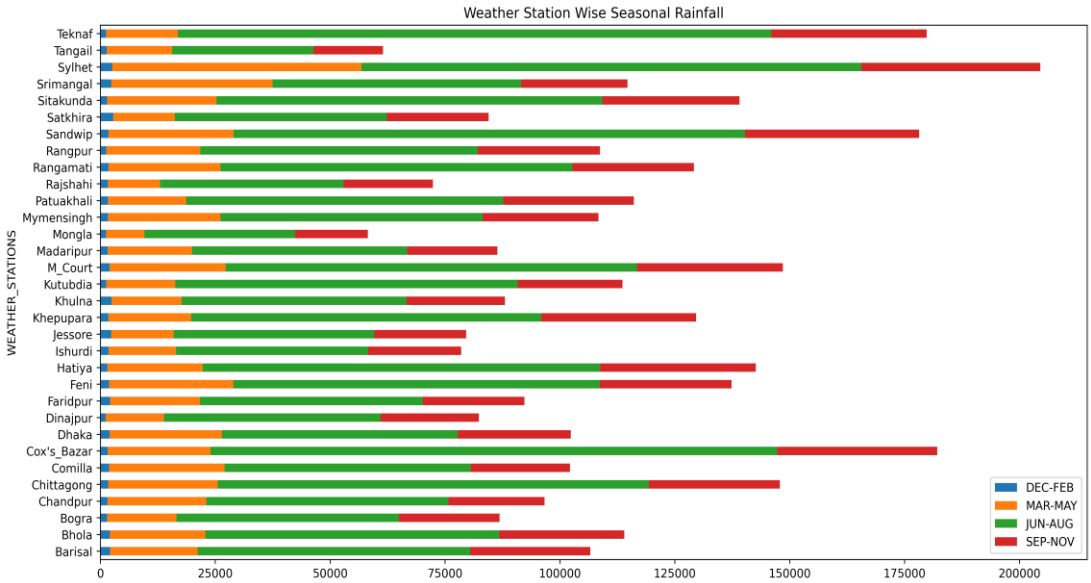


Fig-5: Bar plot of the seasonal rainfall in 32 stations

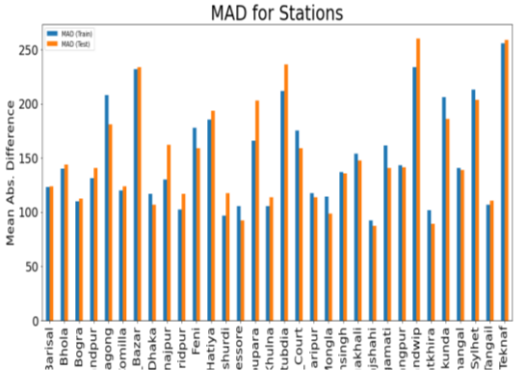


Fig-6: Bar plot of the mean absolute difference of train & test data in 32 stations

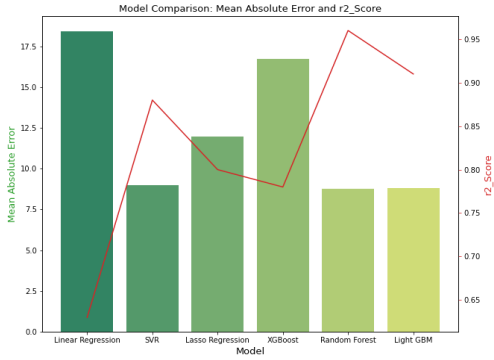


Fig-7: Comparison among applied 6 models

After data analysis, regression models (LRM, SVR, XGBoost, LightGBM, and Lasso) are applied together with pre-processing techniques. Then, for each model, the MAE and r2 scores are calculated and compared and a graph is plotted.

6. Conclusion

For humans, the rain forecast model is incredibly beneficial since it will make daily life simpler. This study suggests utilizing machine learning algorithms and meteorological databases to predict rainfall. Raw data is never complete; hence, it cannot be fed into a model. Therefore, we separated the dataset into two parts—train and test—and preprocessed it. In this work, we employed six machine learning models. The results have demonstrated the capability of the Random Forest to produce accurate forecasts for daily rainfall projections. In order to evaluate the effectiveness of the models, we also process Mean Absolute Error (MAE) for both MLR and SVR models. These models are verified using meteorological data provided by BMD. The current study demonstrates that categorization methods such as random forest regression perform more accurately than linear regression. The random forest regression and random forest classification algorithms outperform all others in terms of outcomes. The data amount has to be increased by at least ten years in order to boost accuracy. We will use the most recent ten years' worth of climate data from every meteorological station in Bangladesh in the future.

References

Abdul-Kader, H., Abd-Elsalam, M., & Mohamed, M. (2020). Hybrid Machine Learning Model for Rainfall Forecasting. *Journal of Intelligent Systems and Internet of Things*, 5–12. https://doi.org/10.54216/jisiot.010101

Adnan, R. M., Petroselli, A., Heddam, S., Santos, C. A. G., & Kisi, O. (2021). Short term rainfall-runoff modelling using several machine learning methods and a conceptual event-based model. *Stochastic Environmental Research and Risk Assessment*, *35*(3), 597–616. https://doi.org/10.1007/s00477-020-01910-0.

Anwar, M. T., Nugrohadi, S., Tantriyati, V., & Windarni, V. A. (2020). Rain Prediction Using Rule-Based Machine Learning Approach. *Advance Sustainable Science, Engineering and Technology*, *2*(1). https://doi.org/10.26877/asset.v2i1.6019

Anwar, M. T., Winarno, E., Hadikurniawati, W., & Novita, M. (2021). Rainfall prediction using Extreme Gradient Boosting. In *Journal of Physics: Conference Series* (Vol. 1869). IOP Publishing Ltd. https://doi.org/10.1088/1742-6596/1869/1/012078

Devi, S. R., Venkatesh, C., Agarwal, P., & Arulmozhivarman, P. (2014). Daily rainfall forecasting using artificial neural networks for early warning of landslides. In *Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2014* (pp. 2218–2224). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICACCI.2014.6968566

Dikshit, A., Pradhan, B., & Alamri, A. M. (2020). Temporal hydrological drought index forecasting for New South Wales, Australia using machine learning approaches. *Atmosphere*, *11*(6). https://doi.org/10.3390/atmos11060585

Khan, N., Sachindra, D. A., Shahid, S., Ahmed, K., Shiru, M. S., & Nawaz, N. (2020). Prediction of droughts over Pakistan using machine learning algorithms. *Advances in Water Resources*, *139*. https://doi.org/10.1016/j.advwatres.2020.103562

Misra, R. K., Panda, P. K., Sahu, A. K., Sahoo, S., & Behera, D. P. (2020). Rainfall Prediction using Machine Learning Approach: A Case Study for the State of Odisha. *Indian Journal of Natural Sciences*, *10*(60), June Issue, 20124-20133, Retrieved 19 June 2022 from https://www.researchgate.net/publication/342410970_Rainfall_Prediction_using_Machine_Learning_Approach_A_Case_Study_for_the_State_of_Odisha

Park, S. J., & Lee, D. K. (2020). Prediction of coastal flooding risk under climate change impacts in South Korea using machine learning algorithms. *Environmental Research Letters*, *15*(9). https://doi.org/10.1088/1748-9326/aba5b3

Rani, D. S., Jayalakshmi, G. N., & Baligar, V. P. (2020). Low Cost IoT based Flood Monitoring System Using Machine Learning and Neural Networks: Flood Alerting and Rainfall Prediction. In *2nd International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2020 - Conference Proceedings*, 261–267. https://doi.org/10.1109/ICIMIA48430.2020.9074928

Shamshirband, S., Hashemi, S., Salimi, H., Samadianfard, S., Asadi, E., Shadkani, S., … Chau, K. W. (2020). Predicting Standardized Streamflow index for hydrological drought using machine learning

models. *Engineering Applications of Computational Fluid Mechanics*, *14*(1), 339–350. https://doi.org/10.1080/19942060.2020.1715844

Singh, N., Chaturvedi, S., & Akhter, S. (2019). Weather Forecasting Using Machine Learning Algorithm. In *2019 International Conference on Signal Processing and Communication, ICSC 2019* (pp. 171–174). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICSC45622.2019.8938211

Srivastava, S., Anand, N., Sharma, S., Dhar, S., & Sinha, L. K. (2020). Monthly rainfall prediction using various machine learning algorithms for early warning of landslide occurrence. In *2020 International Conference for Emerging Technology, INCET 2020*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/INCET49848.2020.9154184

Tikhamarine, Y., Souag-Gamane, D., Ahmed, A. N., Sammen, S. S., Kisi, O., Huang, Y. F., & El-Shafie, A. (2020). Rainfall-runoff modeling using improved machine learning methods: Harris hawks optimizer vs. particle swarm optimization. *Journal of Hydrology*, 589, 125133, 01-18. https://doi.org/10.1016/j.jhydrol.2020.125133

van Klompenburg, T., Kassahun, A., & Catal, C. (01 October 2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture,* 177. Elsevier B.V. https://doi.org/10.1016/j.compag.2020.105709

Zhang, X., Mohanty, S. N., Parida, A. K., Pani, S. K., Dong, B., & Cheng, X. (2020). Annual and Non-Monsoon Rainfall Prediction Modelling Using SVR-MLP: An Empirical Study from Odisha. *IEEE Access*, *8*, 30223–30233. https://doi.org/10.1109/ACCESS.2020.2972435.