## Title Page:

## A prominent technique for enhancing rainfall prediction involves utilizing a XG Boost classifier over Random forest algorithm for better accuracy of rainfall prediction

Shaik Towhid Hussain1, Dr.C.Sivasankar2

Shaik Towhid Hussain1, Research Scholar,

Department of Computer Science and Engineering, Saveetha School of Engineering,

Saveetha Institute of Medical And Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India – 602105

192111049.sse@saveetha.com

Dr.C.Sivasankar2  
professor,  
Department of Computer Science and Engineering,  
Saveetha School of Engineering,  
Saveetha Institute of Medical and Technical Sciences,   
Saveetha University,  Chennai, Tamil Nadu, India, Pincode: 602105  
E-Mail:sivasankarc.sse@saveetha.com

**Keywords:** XG Boost algorithm , Random forest , Agriculture, Rainfall, Prediction, Machine Learning Algorithm.

# ABSTRACT

**Aim**: This study attempts to assess the precision and accuracy of rainfall prediction using random forest (RF) and XG Boost (XGB) machine learning algorithms. **Materials and Methods:** 145461 records make up the weatherAUS dataset, on which the XG Boost method is deployed. This study presents and develops a standard module for rainfall prediction that contrasts the machine learning techniques Random Forest and XG Boost. 1010 sample records from each group were collected for assessment. Clinical analysis was used to measure the sample records; the evaluation's enrollment ratio is 1, the percentage of confidence is 95, the percentage of pretest power is 80, and the values of alpha and beta are 0.05 and 0.5, respectively. The acquired significance value (p), which is less than 0.05, is 0.001 for accuracy. Both methods' final accuracy were computed and reported.. **Results:** The rainfall on the used dataset is predicted by the machine learning algorithm XG Boost classifier with 93.00% accuracy, whereas the Random Forest classifier predicts the same event with 85.00% accuracy**. Conclusion:** According to the study, the XG Boost algorithm predicts rainfall more accurately than the Random Forest method.

**Keywords:** XG Boost algorithm , Random forest , Agriculture, Rainfall, Prediction, Machine Learning Algorithm.

# INTRODUCTION

Since agriculture has been the main industry in practically every state, it plays a major role in the economy of our nation. Predicting rainfall is therefore more important for the nation's economic development. Over the years, predicting the amount of rain on Earth has proven to be the most hard and difficult task. Over 60% of the people in India, a country of over a billion people, depend on agriculture as their main source of income (Kharwal 2020). In the modern world, resources and technological advancements are growing daily. Applications of science and technology can be used to forecast the weather in a specific area. In the past, several attempts were made to anticipate rainfall in order to assist farmers with their agricultural operations. These initiatives were first made informally in the eighteenth century and formally in the nineteenth (Oswal 2019). In essence, weather forecasting is the analysis and prediction of future climate conditions for a certain area. Predicting rainfall is a type of weather forecasting (Yen et al. 2019). Since the beginning of time, there has been interest in weather prediction, and numerous strategies for predicting rainfall have been developed. All of those methods have developed with time, and they are all unique in terms of accuracy and effectiveness (Shardoor and Rao 2018). The suggested study aids in accurately forecasting the amount of rainfall, which benefits the nation's overall development.

There are 105 research articles published on the rainfall prediction in the IEEE xplore, 165 research papers on google scholar and 34 articles were found in sciencedirect. [(Rudrappa](https://paperpile.com/c/1kWtcS/g9Fr) [2021)](https://paperpile.com/c/1kWtcS/g9Fr) presents a new algorithm that predicts the long term rainfall with the help of time series approach using Artificial Neural Networks and Bayesian Enhanced Approach. The Bayes technique offers an additional method for incorporating past knowledge into forecasting models because it is frequently challenging to outperform the best representation of weather forecast information. predicting circumstances in cases where prior knowledge is either unavailable or unhelpful. It seems that things are stable as a result. (2018, Shardoor and Rao) examined various rainfall prediction classifiers that aid in determining the frequency of annual precipitation, its total amount, and the combination of the frequency and total amount of annual precipitation. [(Narejo et al. 2021)](https://paperpile.com/c/1kWtcS/dooI) Proposed a multiple linear regression design for predicting the participation rate. The design proposed considers various weather parameters, few of them are temperature, pressure, wind speed and wind direction. National Climatic Data Centre is the website from where the data is collected for the research. [(Reddy 2021)](https://paperpile.com/c/1kWtcS/KBCZ) Proposed an algorithm for rainfall prediction through the Random Forest. It achieves an accuracy of 80.56% and outperforms all other feature representations, including k-nearest neighbours and decision tree algorithms.

The research gap identified from the survey is that there are many methods proposed for rainfall prediction but most of those methods have less accuracy rate. Several works have demonstrated that the performance of Random forest is poor and provides less accuracy in prediction of the rainfall. A study by [(Mohammed et al. 2020)](https://paperpile.com/c/1kWtcS/IaW3) compares the accuracy of various mining classification algorithms in predicting rainfall. The best way to determine which classification algorithm provides the best accuracy is to analyse and compare them. Therefore, this research paper compares the precision and the accuracy of XGBoost and Random Forest algorithm for accurate rainfall.

# MATERIALS AND METHODS

This research was performed at the Data Analytics lab, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, in which the lab provides extremely superior configured systems which help to get accurate results. Total no.of groups considered for the research were two, group1 consists of 10 sample sizes and group 2 consists of 10 sample sizes. The calculation is done with G-power 0.95, alpha value 0.005, beta value

0.95 and confidence interval 95%. The dataset used for the research was downloaded from Kaggle website [(kerneler 2019)](https://paperpile.com/c/1kWtcS/6ur6).

## XGBoost (XGB) - Group 1

Inputs: WeatherAUS dataset Output: Accuracy

1. Load the WeatherAUS.csv file
2. Divide the whole data records randomly into testing (20%) and training (80%)
3. Identify the output variable
4. Develop XG Boost classification design using training dataset
5. Train the developed classification
6. Project the testing dataset with respect to the training data file
7. Examine the developed classification model.
8. Show the accuracy value obtained.

In this study, sklearn.linear\_model library of XG Boost class was used for training the data. Read weatherAUS.csv file and load the same file. Data records of the dataset are distributed randomly, 80% of the dataset is taken for training and the remaining 20% of the dataset is taken for testing. The output variable is defined then, 3 of 10 XG Boost classifier are developed using the training dataset. Testing dataset is estimated using training dataset. The XG Boost classification design is tested and then accuracy is produced.

## Random Forest (RF) - Group 2

Input: weatherAUS dataset Output: Accuracy

1. Load and read weatherAus dataset
2. Choose properties randomly from dataset
3. Develop RF classification criteria as a variable for evaluation.
4. Gini was considered as an argument.
5. Design a classification model using RF classifiers and guess the output for each sample.
6. Random selection process was performed for each predicted output.
7. Most voted outputs were considered as final results.
8. Return the precision value obtained.

In this study, sklearn ensemble library of random forest class was used for training the data. Data records of the dataset are distributed randomly, 80% of the dataset is taken for training and the remaining 20% of the dataset is taken for testing. It randomly chooses samples from the dataset, and decision trees were gathered for predicting the outcome. Every anticipated outcome was voted on, and the most selected result was chosen as the final result. The algorithm uses a Random Forest Classifier (RF).

This study was implemented using Google collab and SPSS software, and hardware specifications needed in a system for evaluation is an intel i3 processor, 50GB Hard Disk Drive, 4GB and Random Access Memory (RAM) and software specifications needed is a windows operating system.

# STATISTICAL ANALYSIS

The SPSS tool is utilised to statistically assess the work in addition to experimental analysis. The mean, standard deviation, accuracy, and standard error mean were the research objectives. Temperature, humidity, pressure, and wind speed are the independent research variables, and accuracy is the dependent variable (Dattalo 2013). The XGB and RF algorithms were compared using an independent sample T-Test.

# RESULTS

Table 1 shows the comparison between the accuracy obtained while evaluating the XG Boost and the Random Forest model for analysis with various iterations.

Table 2 depicts the various parameters of both groups. The accuracy, Recall, F1 Score and support has been calculated for XGB and RF. The analysis of two groups presents XGB as having better accuracy (93.00%) and compared to Random Forest.

Table 3 shows the statistical analysis of XG Boost and Random Forest with different test datasets. The table depicts XG Boost algorithm produces better accuracy compared with Random Forest algorithm.

Table 4 depicts the statistical analysis of Significant levels for both groups. There is a negligible Significant difference 0.001 for accuracy among the two groups. Hence XG Boost is better than Random Forest.

Figure. 3 inferred the mean accuracy of XG Boost (XGB) and Random Forest (RF). The results showed that the XGB has better accuracy (93.00%) compared to the accuracy of Random Forest .

# DISCUSSION

Agriculture has been the backbone of India and prediction of rainfall is a major issue in the agriculture sector. Experimental work was done among two groups: XG Boost(XGB) and Random Forest (RF) by varying the test size. From the experimental results (Figure 3 ) done in SPSS, the accuracy of XGB is 93.00%, whereas Random Forest provides the accuracy of 85%. This depicts that XG Boost is better than Random Forest. From the SPSS, we can say that the developed XG Boost algorithmic classification model out performed in comparing accuracy (93.00%) with the Random Forest algorithm.

One of the most important factors of rainfall prediction is effectiveness. In a Research by Alshareet, O., Itradat, A., Doush, I.A., Quttoum, A.: Incorporation of ISO 25010 with machine learning to develop a novel quality in use prediction system (QIUPS). Int. J. Syst. Assur. Eng. Manag.Specifically, the XGBoost learning accuracy was found to be 100% during teams learning and production phase, while its prediction accuracy was found to be 95.60% and 93.08%, respectively for the same phases. Similarly, the learning accuracy of the DNN was found to be 89.26% and 81.23%, while its prediction accuracy was found to be 80.50% and 77.36%, during the two phases. In a research [Journal of Physics: Conference Series](https://iopscience.iop.org/journal/1742-6596), [Volume 1324](https://iopscience.iop.org/volume/1742-6596/1324), [The Second International Conference on Physics, Mathematics and Statistics 22–24 May 2019, Hangzhou, China](https://iopscience.iop.org/issue/1742-6596/1324/1). And its prediction accuracy was found to be 93.40% and 91.80%, respectively for the same phases. The research from [(Zhang et al. 2020)](https://paperpile.com/c/1kWtcS/omyJ) proves that the multilayer perceptron is better than the Random Forest algorithm in predicting the effective and accurate rainfall.

Although the proposed methodology attained satisfactory results, still there are some limitations in research. Accuracy evaluation may not produce satisfactory results on bulk datasets. Moreover in XGB, the mean error is higher compared with RF which is not a desired

outcome. Reducing the mean error is a great boost to the existing research work. One of the possible ways to increase accuracy and to reduce mean error is by using optimization algorithmic techniques on the algorithms. Another scope can be using feature selection algorithms before classifying the dataset for improved classification accuracy of the classifiers which helps to get predominant outcomes.

# CONCLUSION

XG Boost is a machine learning classifier which uses decision trees for producing improved accuracy and precision. The research study proves that the accuracy for predicting rainfall using XG Boost (XGB) algorithm looks to be higher when compared with Random Forest. It is found that XGB performs significantly better than RF in predicting effective rainfall. Therefore, the report concluded that the XG Boost (XGB) algorithm produces better accuracy (93.00%) compared with Random Forest (RF) accuracy (85.00%).

# DECLARATIONS

## Conflicts of Interests

No conflicts of interest in this manuscript.

## Author Contributions

Author Towhid played a key role in collecting and analysing data as well as writing the manuscript. Additionally, S.Towhid contributed significantly to conceptualization, data validation and providing critical feedback during manuscript reviews.

## Acknowledgements

The authors would like to thank the management, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (formerly known as Saveetha University) for providing the infrastructure needed to successfully complete this project.

**Funding:** We extend our gratitude to the following organisations for the financial support they have provided which helped us to conduct and complete the research.

* 1. Saveetha University.
  2. Saveetha Institute of Medical And Technical Sciences.
  3. Saveetha School of Engineering.

# REFERENCES

[Narejo, Sanam, Muhammad Moazzam Jawaid, Shahnawaz Talpur, Rizwan Baloch, and Eros](http://paperpile.com/b/1kWtcS/dooI) [Gian Alessandro Pasero. 2021. “Multi-Step Rainfall Forecasting Using Deep Learning](http://paperpile.com/b/1kWtcS/dooI) [Approach.” *PeerJ Computer Science* 7. https://doi.org/](http://paperpile.com/b/1kWtcS/dooI)[10.7717/peerj-cs.514](http://dx.doi.org/10.7717/peerj-cs.514)[.](http://paperpile.com/b/1kWtcS/dooI)

[Oswal, Nikhil. 2019. “Predicting Rainfall Using Machine Learning Techniques,” October.](http://paperpile.com/b/1kWtcS/8E1g) [https://doi.org/](http://paperpile.com/b/1kWtcS/8E1g)[10.36227/techrxiv.14398304.v1](http://dx.doi.org/10.36227/techrxiv.14398304.v1)[.](http://paperpile.com/b/1kWtcS/8E1g)

[Alexandridis, Antonios K., and Achilleas D. Zapranis. 2014. *Wavelet Neural Networks: With*](http://paperpile.com/b/1kWtcS/CIIe)[*Applications in Financial Engineering, Chaos, and Classification*. John Wiley & Sons.](http://paperpile.com/b/1kWtcS/CIIe)

[Dash, Yajnaseni, Saroj K. Mishra, and Bijaya K. Panigrahi. 2019. “Predictability Assessment](http://paperpile.com/b/1kWtcS/67qd) [of Northeast Monsoon Rainfall in India Using Sea Surface Temperature Anomaly](http://paperpile.com/b/1kWtcS/67qd) [through Statistical and Machine Learning Techniques.” *Environmetrics*.](http://paperpile.com/b/1kWtcS/67qd) [https://doi.org/](http://paperpile.com/b/1kWtcS/67qd)[10.1002/env.2533](http://dx.doi.org/10.1002/env.2533)[.](http://paperpile.com/b/1kWtcS/67qd)

[Dattalo, Patrick. 2013. *Analysis of Multiple Dependent Variables*. Oxford University Press.](http://paperpile.com/b/1kWtcS/nPq0) [Ingsrisawang, L., S. Ingsriswang, S. Somchit, Prasert Aungsuratana, and](http://paperpile.com/b/1kWtcS/MiTV)

[WarawutKhantiyanan. 2008. “Machine Learning Techniques for Short-Term Rain](http://paperpile.com/b/1kWtcS/MiTV) [Forecasting System in the Northeastern Part of Thailand.”](http://paperpile.com/b/1kWtcS/MiTV) [https://www.semanticscholar.org/paper/Machine-Learning-Techniques-for-Short-Term-R](https://www.semanticscholar.org/paper/Machine-Learning-Techniques-for-Short-Term-Rain-in-Ingsrisawang-Ingsriswang/559d0ebab46656f64c6eb5e347de891be2170b1d) [ain-in-Ingsrisawang-Ingsriswang/559d0ebab46656f64c6eb5e347de891be2170b1d](https://www.semanticscholar.org/paper/Machine-Learning-Techniques-for-Short-Term-Rain-in-Ingsrisawang-Ingsriswang/559d0ebab46656f64c6eb5e347de891be2170b1d)[.](http://paperpile.com/b/1kWtcS/MiTV)

[kerneler. 2019. “Starter: weatherAUS 536c1115-4.” Kaggle. November 29, 2019.](http://paperpile.com/b/1kWtcS/6ur6) <https://kaggle.com/kerneler/starter-weatheraus-536c1115-4>[.](http://paperpile.com/b/1kWtcS/6ur6)

[Shardoor, Nikhilkumar B., and Mandapati Venkateswar Rao. 2018. “Analysis of Rainfall](http://paperpile.com/b/1kWtcS/qfo6) [Prediction Using Machine Learning Data Mining and Satellite Techniques.”](http://paperpile.com/b/1kWtcS/qfo6) [*International Journal of Engineering & Technology* 7 (4): 4362–67.](http://paperpile.com/b/1kWtcS/qfo6)

[Kharwal, Aman. 2020. “Rainfall Prediction with Machine Learning.” September 11, 2020.](http://paperpile.com/b/1kWtcS/H1FD) <https://thecleverprogrammer.com/2020/09/11/rainfall-prediction-with-machine-learning/>[.](http://paperpile.com/b/1kWtcS/H1FD) [Krishnan, R., J. Sanjay, Chellappan Gnanaseelan, Milind Mujumdar, Ashwini Kulkarni, and](http://paperpile.com/b/1kWtcS/SKV9) [Supriyo Chakraborty. 2020. *Assessment of Climate Change over the Indian Region: A*](http://paperpile.com/b/1kWtcS/SKV9)

[*Report of the Ministry of Earth Sciences (MoES), Government of India*. Springer Nature.](http://paperpile.com/b/1kWtcS/SKV9)

[Reddy, Ajay Vardhan. 2021. “Rain Prediction in Australia.” June 29, 2021.](http://paperpile.com/b/1kWtcS/KBCZ) [https://www.analyticsvidhya.com/blog/2021/06/predictive-modelling-rain-prediction-in-](https://www.analyticsvidhya.com/blog/2021/06/predictive-modelling-rain-prediction-in-australia-with-python/) [australia-with-python/](https://www.analyticsvidhya.com/blog/2021/06/predictive-modelling-rain-prediction-in-australia-with-python/)[.](http://paperpile.com/b/1kWtcS/KBCZ)

[Rudrappa, Gujanatti. 2021. “Machine Learning Models Applied for Rainfall Prediction.”](http://paperpile.com/b/1kWtcS/g9Fr) [*Revista Gestão Inovação E Tecnologias*.](http://paperpile.com/b/1kWtcS/g9Fr) [https://doi.org/](http://paperpile.com/b/1kWtcS/g9Fr)[10.47059/revistageintec.v11i3.1926](http://dx.doi.org/10.47059/revistageintec.v11i3.1926)[.](http://paperpile.com/b/1kWtcS/g9Fr)

[Yen, Meng-Hua, Ding-Wei Liu, Yi-Chia Hsin, Chu-En Lin, and Chii-Chang Chen. 2019.](http://paperpile.com/b/1kWtcS/Rtuw) [“Application of the Deep Learning for the Prediction of Rainfall in Southern Taiwan.”](http://paperpile.com/b/1kWtcS/Rtuw)

[*Scientific Reports* 9 (1): 1–9.](http://paperpile.com/b/1kWtcS/Rtuw)

[Zhang, Pengcheng, Yangyang Jia, J. Gao, Wei Song, and H. Leung. 2020. “Short-Term](http://paperpile.com/b/1kWtcS/omyJ) [Rainfall Forecasting Using Multi-Layer Perceptron.”](http://paperpile.com/b/1kWtcS/omyJ) [http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8468083](http://ieeexplore.ieee.org/stamp/stamp.jsp?tp&arnumber=8468083)[.](http://paperpile.com/b/1kWtcS/omyJ)

# TABLES AND FIGURES

**Table 1.** Accuracy values of XG Boost and Random Forest algorithms obtained for each iteration while evaluating the dataset for various test sizes.

|  |  |
| --- | --- |
| **GROUP** | **ACCURACY** |
| XGB | 93 |
| XGB | 89 |
| XGB | 91 |
| XGB | 87 |
| XGB | 94 |
| XGB | 92 |
| XGB | 93 |
| XGB | 90 |
| XGB | 88 |
| XGB | 92 |
| RF | 85 |
| RF | 75 |
| RF | 79 |
| RF | 84 |
| RF | 82 |
| RF | 76 |
| RF | 81 |
| RF | 79 |
| RF | 71 |
| RF | 83 |

**Table 2**.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Group Statistics** | | | | | |
|  | GROUP | N | Mean | Std. Deviation | Std. Error Mean |
| ACCURACY | XGB | 20 | 90.0500 | 2.64525 | .59150 |
| RF | 20 | 77.3000 | 4.75837 | 1.06400 |

|  |
| --- |
|  |

**Table 3**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Levene’s Test for Equality of Variances | | T-test for Equality of means | | | | | | |
| F | Sig. | t | df | Sig( 2-ta iled  ) | Mean Difference | Std. Error Difference | 95% confidence interval of the Difference | |
| Lower | Upper |
| Accuracy Equal  variances  Assumed | 9.618 | .004 | 10.473 | 38 | .000 | 12.7500 | 1.21736 | 10.28558 | 15.21442 |
| Equal  Variances  Not assumed |  |  | 10.473 | 29.7 | .000 | 12.7500 | 1.21736 | 10.26283 | 15.23717 |

