## Title Page:

## An efficient prediction of rainfall using XGBoost algorithm over LightGBM Classifier for improvement of accuracy.

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 , LightGBM , Agriculture, Rainfall, Prediction, Machine Learning Algorithm.

# ABSTRACT

**Aim**: This study uses the machine learning algorithms LightGBM (LGBM) and XG Boost (XGB) to evaluate the precision and accuracy of rainfall prediction. **Materials and Methods:** The XG Boost method is applied to the weatherAUS dataset, which consists of 145461 records. This work compares the machine learning methods LightGBM and XG Boost for rainfall prediction and presents and develops a standard module for the purpose. For evaluation, 1010 sample records were gathered from each group. The sample records were measured using clinical analysis; the evaluation's enrollment ratio is 1, its confidence percentage is 95%, its pretest power percentage is 80%, and its alpha and beta values are 0.05 and 0.5, respectively. For accuracy, the obtained significance value (p), which is less than 0.05, is 0.001. 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 88.50% accuracy, whereas the LightGBM predicts the same event with 83.60% accuracy**. Conclusion:** According to the study, the XG Boost algorithm predicts rainfall more accurately than LightGBM method.

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

# INTRODUCTION

Agriculture has been the primary industry in almost every state, making it a significant contributor to the national economy. Therefore, rainfall prediction is more crucial for the country's economic growth. Predicting how much rain will fall on Earth has proven to be the most challenging task over the years. In India, a nation of over a billion people, agriculture provides the primary means of subsistence for more than 60% of the population (Kharwal 2020). Resources and technological innovations are expanding on a daily basis in the modern world. The weather in a particular location can be predicted using scientific and technological applications. There have been multiple attempts in the past to predict rainfall to help farmers with their farming operations. First proposed informally in the eighteenth century, these efforts were formalised in the nineteenth (Oswal 2019). Weather forecasting is essentially the study and forecasting of future climate conditions for a specific region. One kind of weather forecasting is rainfall prediction (Yen et al. 2019). Weather prediction has piqued interest since the beginning of time, and various methods for forecasting rainfall have been established. Since they have all evolved over time, those techniques are all distinct in terms of precision and potency (Shardoor and Rao 2018). The proposed study contributes to the nation's overall development by helping to accurately forecast the amount of rainfall.

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. A few of the weather parameters taken into account in the proposed design are temperature, pressure, wind direction, and speed. The website where the data is gathered for the study is the National Climatic Data Centre. (Reddy, 2021) suggested an algorithm to forecast rainfall using LightGBM. It surpasses all other feature representations and achieves an accuracy of 83.60%.

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 LightGBM 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 the accuracy of LightGBM and XGBoost 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.

## LightGBM (LGBM) - Group 2

Input: weatherAUS dataset Output: Accuracy

1. Load and read weatherAus dataset
2. Choose properties randomly from dataset
3. Develop LGBM 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 LightGBM 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 LightGBM Classifier (LGBM).

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

In addition to experimental analysis, the work is statistically evaluated using the SPSS tool. The research objectives were the mean, standard deviation, accuracy, and standard error mean. Accuracy is the dependent variable, and the independent research variables are temperature, humidity, pressure, and wind speed (Dattalo 2013). An independent sample T-Test was used to compare the XGB and LGBM algorithms.

# RESULTS

Table 1 shows the comparison between the accuracy obtained while evaluating the XG Boost and the LightGBM 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 LGBM. The analysis of two groups presents XGB as having better accuracy (88.50%) and compared to LightGBM.

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

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 LightGBM.

Figure. 3 inferred the mean accuracy of XG Boost (XGB) and LightGBM (LGBM). The results showed that the XGB has better accuracy (88.50%) compared to the accuracy of LightGBM .

# DISCUSSION

### India's economy has historically relied heavily on agriculture, and forecasting rainfall is a big problem for the industry. By altering the test size, two groups—XG Boost (XGB) and LightGBM (LGBM)—performed experimental work. Based on the SPSS experimental results (Figure 3), XGB has an accuracy of 88.50%, while LightGBM offers an accuracy of 83.60%. This illustrates why XG Boost performs superior to LightGBM. According to the SPSS, the developed XG Boost algorithmic classification model outperformed the LightGBM in terms of accuracy comparison (88.50%). One of the most important factors of rainfall prediction is effectiveness. In a Research by [Rainfall Prediction Using Catboost Machine Learning Algorithm](https://ieeexplore.ieee.org/document/10276336/)

[Vempaty Prashanthi](https://ieeexplore.ieee.org/author/37085565227);[Srinivas Kanakala](https://ieeexplore.ieee.org/author/37089254383);[Deepika Borgaonkar](https://ieeexplore.ieee.org/author/37090043935);[D. Suresh Babu](https://ieeexplore.ieee.org/author/37062109200)

[2023 International Conference on Network, Multimedia and Information Technology (NMITCON)](https://ieeexplore.ieee.org/xpl/conhome/10275783/proceeding)

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 XGBoost is better than the LightGBM 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 LGBM which is not a desired

outcome. Lowering the mean error greatly improves the current research. Using optimisation algorithmic techniques on the algorithms is one possible way to reduce mean error and increase accuracy. Another option is to use feature selection algorithms prior to dataset classification in order to increase classifier accuracy and produce more desirable results.

# CONCLUSION

XG Boost is a machine learning classifier which uses decision trees for producing improved accuracy . The research study proves that the accuracy for predicting rainfall using XG Boost (XGB) algorithm looks to be higher when compared with LightGBM. It is found that XGB performs significantly better than LGBM in predicting effective rainfall. Therefore, the report concluded that the XG Boost (XGB) algorithm produces better accuracy (88.50%) compared with LightGBM (LGBM) accuracy (83.60%).

# 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)

[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)

[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)

[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)

[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)

[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)

[*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)

[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)

# TABLES AND FIGURES

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

|  |  |
| --- | --- |
| **GROUP** | **ACCURACY** |
| XGB | 88 |
| XGB | 91 |
| XGB | 85 |
| XGB | 90 |
| XGB | 86 |
| XGB | 92 |
| XGB | 85 |
| XGB | 86 |
| XGB | 93 |
| XGB | 92 |
| LGBM | 82 |
| LGBM | 81 |
| LGBM | 83 |
| LGBM | 78 |
| LGBM | 81 |
| LGBM | 74 |
| LGBM | 79 |
| LGBM | 81 |
| LGBM | 73 |
| LGBM | 78 |

**Table 2**.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Group Statistics** | | | | | |
|  | GROUP | N | Mean | Std. Deviation | Std. Error Mean |
| ACCURACY | XGB | 20 | 88.6000 | 3.08477 | .68977 |
| LGBM | 20 | 75.7500 | 4.26584 | .95387 |

|  |
| --- |
|  |

**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 | 2.133 | .152 | 10.916 | 38 | .000 | 12.8500 | 1.17714 | 10.46701 | 15.23442 |
| Equal  Variances  Not assumed |  |  | 10.916 | 34.6 | .000 | 12.8500 | 1.17714 | 10.45930 | 15.24717 |

GRAPH :

