IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

Weather Forecasting System with Machine Learning based Rain Prediction using Random Forest Algorithm

Miss. Sayali R. Mhatre[1] and Nidhi HP Varma[2]
[1] Assistant Professor, Department of M. Sc.IT
[2]Student, M. Sc.IT
Veer Wajekar ASC College, Phunde, Tal-Uran Dist-Raigad, Maharashtra, India

Abstract: *Weather is defined as the state of the air in a place that varies over a specific or brief period and plays a significant role in influencing the patterns of life of individuals, organizations, and geographic areas. Unpredictable rainfall patterns can cause landslides and floods, among other tragedies. As a result, locals must have precise information in order to foresee and forecast heavy precipitation, which may result in flooding. The purpose of this research is to use time-series data to predict rain using machine learning (ML) algorithms. Additionally, ML algorithms will be compared in order to determine which one predicts rainfall with the highest accuracy. Two machine learning techniques, random forest and decision tree, were used in this research. Three features—relative humidity, temperature, and wind direction—are the data that are employed. The decision tree method yields an average accuracy value of 94.85%, but the random forest approach offers an average accuracy value of 95.64%. The random forest algorithm's accuracy results for rain prediction are superior to decision trees, according to test findings from the dataset used.*

Keywords: rainfall, machine learning, random forest, water sources, rain, weather forecasting

I. INTRODUCTION

The presence of water, which comes from rain, is essential to the existence of all living things. When other water sources, like rivers, lakes, or wells, are unreachable, rain is a critical supply of water. Rainwater offers a host of advantages, for instance to irrigate agricultural land, industries, and power plants. Depending on factors including topography, wind patterns, climate, and geographic position, rainfall characteristics might differ between places. On the other hand, heavy rains can trigger landslides and floods. Over the past ten years, Indonesia has seen an increase in the frequency of hydro-meteorological disasters [1]. Although the amount of rainfall in the future cannot be predicted with precision, it can be calculated by utilizing historical and current weather data.

Predicting when it will rain is crucial for many industries, including agriculture, urban planning, water resource management, and disaster mitigation. One promising way to improve rain forecasting accuracy and dependability is to employ machine learning algorithms. Using machine learning, an advanced computational technique, one can leverage the potential of large datasets to find complex relationships, trends, and patterns across different meteorological variables. Machine learning algorithms can better comprehend and anticipate rainfall patterns by using past data to generate precise predictions [2]. Basically, rain prediction is the process of estimating the probability of future rain events by examining current and historical weather data along with other environmental conditions. Prediction models can make more accurate forecasts by taking advantage of intricate patterns and correlations between weather variables using machine learning techniques.

This research focused on decision tree and random forest algorithms to create and evaluate machine learning-based rain prediction models. Decision tree is more likely to overfit, particularly when applied to complicated datasets, whereas

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 5, June 2026 Impact Factor: 8.2

random forests algorithm is a popular ensemble learning method used for classification. With this method, a vast number of decision trees are built, and the outcomes are combined to provide forecasts [2]. It is anticipated that this research will advance knowledge of the efficiency of machine learning algorithms in rain prediction and offer direction for the creation of more advanced and dependable rain prediction systems. Ultimately, it is hoped that the use of machine learning techniques in rain prediction can help society and the government in responding appropriately to the effects of extreme weather, lower financial losses, and boost resilience to natural disasters.

II.LITERATUREREVIEW

Knowing when it will rain is crucial, since it can affect many aspects of a person's life. Rain often occurs due to the geographic location of a region with a tropical climate. Tropical regions, particularly Indonesia, generally have varied climates, with each location having its own unique rainfall patterns. Rainfall has a variety of properties; it frequently affects not only food production but also natural disasters and economic activity. While "climate" refers to the typical conditions of the atmosphere over a longer time span, "weather" refers to daily variations in temperature and precipitation [3]. Rainfall is defined as the height at which precipitation gathers in a level area without evaporating, seeping, or flowing. In tropical regions like Indonesia, when the water level rises or is high, the water will overflow.

Five rainfall categorization techniques—Naive Bayes, Decision Trees, Support Vector Machines (SVM), Neural Networks, and Random Forests—were compared in research [4] conducted in Malaysia. It was discovered that the Random Forest method detected rainfall accurately. Research by [5] used a Random Forest model in a rainfall prediction system to forecast weather. Model evaluation used training, testing, and k-fold cross-validation. The study indicated that the Random Forest model works better when all data are used as training data. While accuracy with total data was 99.45%, accuracy with 10-fold cross-validation was 71.09%.

Another research conducted by [6] implemented distribution patterns using the Ordinary Kriging geostatistical method from the prediction results of the SVM and Random Forest methods on air temperature. The research was carried out by identifying the accuracy of the SVM and Random Forest methods in estimating long-term air temperature through data transfer procedures.

Research [7] classified rainfall data using a Decision Tree model with the Classification and Regression Tree (CART) algorithm using the CRISP-DM data mining technique. The dataset used consisted of six attributes, such as average temperature, minimum temperature, maximum temperature, humidity, duration of sunlight, and rainfall. The rainfall dataset used contained 3,653 records, and only 123 records were categorized as heavy rain and very heavy rain. The results of the tests carried out obtained an accuracy value of 89.4%.

Another research is similar to previous research in comparing classification methods, but the focus is different. The aim of the research is to compare the performance of two machine learning models, Decision Tree and Random Forest, for modelling large landslide disasters triggered by rainfall.

Based on previous studies, considering both the advantages and disadvantages, rainfall prediction is developed in this research. Using a time-series dataset, this work employed a Random Forest model with a randomized search to determine the optimal hyperparameter values and a Decision Tree model.

III. RESEARCH METHOD

The most accurate machine learning method for rainfall prediction will be determined by comparing various techniques. This research included two machine learning methods: Decision Trees and Random Forests. The three features used are temperature, wind direction, and relative humidity. The collected data is then divided into training and test sets at a 70:30 ratio, and a confusion matrix is utilized to evaluate the model. The complete research method can be seen in Fig. 1.

After doing literature review and identifying problems, we collected time-series dataset from 2019 until 2021. Then, data preprocessing is done. The term data preprocessing is frequently used in the area of machine learning to refer to

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

the processes involved in organizing, cleaning, and preparing raw data before using it to build machine learning models. Preprocessing can be applied to remove certain anomalies while leaving others unaltered.

The next step is model training. Model training in machine learning refers to the process of teaching a machine learning algorithm to recognize patterns and make predictions or decisions based on input data. It involves presenting the algorithm with a dataset that contains examples of the problem the algorithm is trying to solve, along with the correct answers or outcomes for each example. During training, the algorithm learns from the data by adjusting its internal parameters or weights to minimize the difference between its predictions and the actual outcomes in the training data. This process is done through an optimization process, called hyperparameter tuning. Hyperparameter tuning involves searching for the combination of hyperparameters that result in the best performance of the model on a validation dataset. This research used grid search in decision tree model and random search in random forest model.

Once the training process is complete, the trained model can be used to make predictions on new data that has not been used during training in the model deployment step. It is essential to thoroughly test the model to ensure that it performs as expected, for example, by evaluating the model's accuracy.

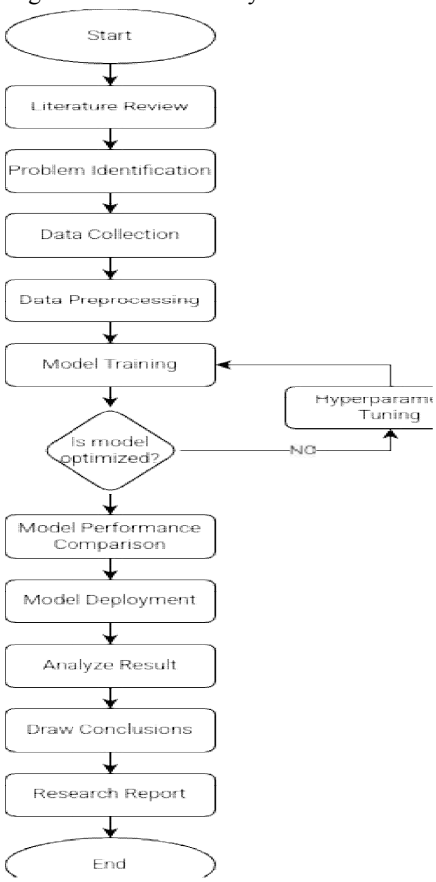


Fig. 1. Research flowchart

IV. RESULTS AND DISCUSSION

The relationships between the different variables in the dataset are shown visually in Fig. 2's correlation heatmap. Upon meticulous examination of this heatmap, numerous interesting positive correlations among various variables surface, providing valuable insights that enhance our comprehension of the dataset's fundamental dynamics.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

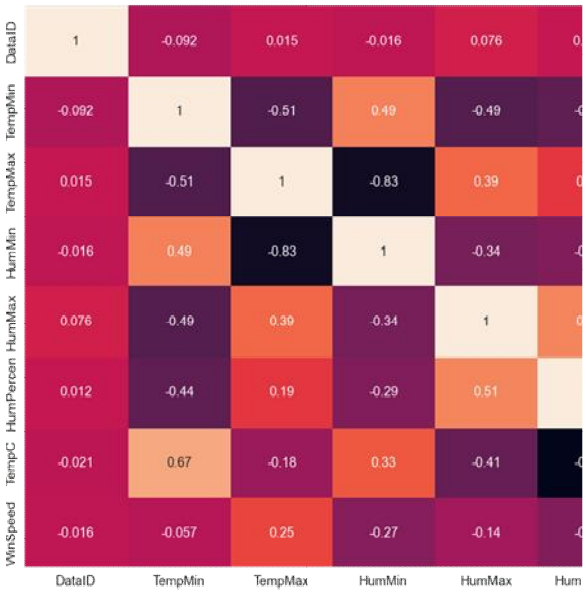


Fig. 2. Correlation heatmap of training dataset

DataID has no significant relationship with other features, which indicates that data numbering is not related to the values of other variables in the dataset. TempMin (minimum temperature) and TempMax (maximum temperature) are moderately negatively correlated, which means the lower the minimum temperature, the higher the maximum temperature. However, TempMin is moderately positively correlated with HumMin (minimum humidity) and strongly positively correlated with TempC (temperature in Celsius), which indicates that low minimum temperatures tend to be followed by low minimum humidity and low temperatures as well.

TempMax has a strong negative correlation with HumMin, which means that the higher the maximum temperature, the lower the minimum humidity. However, TempMax is moderately positive correlated with HumMax (maximum humidity), indicating that high maximum temperatures tend to be followed by high maximum humidity.

HumMin is moderately negatively correlated with HumMax, meaning the lower the minimum humidity, the higher the maximum humidity. In addition, HumMin also has a moderate positive correlation with HumPercen (humidity percentage), which indicates that high minimum humidity tends to be followed by a high average humidity percentage as well.

HumMax has a strong positive correlation with HumPercen, which means that the higher the maximum humidity, the higher the average humidity percentage. However, HumMax is moderately negatively correlated with TempC, indicating that high maximum humidity tends to be accompanied by low temperatures.

DOI: 10.48175/IJARSCT-35636

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

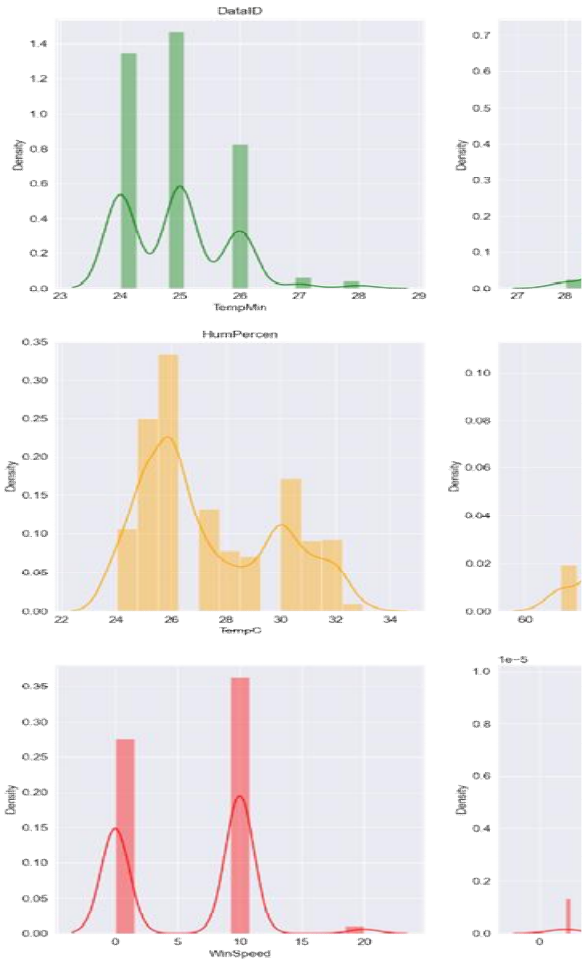


Fig. 3. Dataset features distribution

HumPercen has a strong negative correlation with TempC, which means the higher the humidity percentage, the lower the temperature. TempC has a moderate positive correlation with WinSpeed (wind speed), which indicates that high temperatures tend to be followed by high wind speeds as well. Meanwhile, WinSpeed does not have a strong correlation with other features in this dataset.

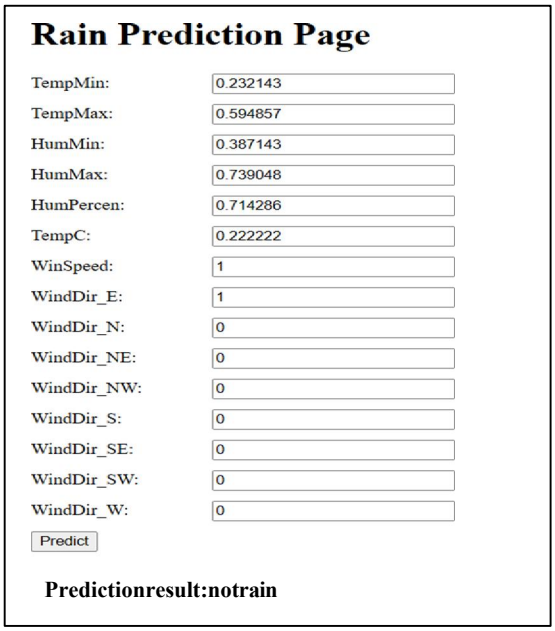
A thorough visual representation of the distribution of all numerical features is shown in Fig. 3. The nature of these features is better understood by looking at this visualization. For DataID, the distribution shows that the data ID values tend to be concentrated around certain values. There are several peaks or groups of values that stand out, indicating the possibility of certain groupings or categories in the data numbering. TempMin has a distribution with two main peaks that are quite prominent. This suggests that there are two dominant groups of minimum temperatures in the data, perhaps representing different weather conditions or seasons. Meanwhile, the TempMax distribution also displays two main peaks, with one peak being slightly higher than the other peaks. This indicates that there are two dominant groups of maximum temperatures, but with one group having slightly more data.

For TempC, the distribution is bell-shaped or a normal distribution with one peak in the middle. This shows that most temperature values are centered around a certain average value, with symmetric variations around it. The HumPercen has a shape like a beta distribution, with one dominant peak in the middle and descending tails on either side. This indicates that most moisture percentage values are centered around a certain value, with less variation in both tails of

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 5, June 2026 Impact Factor: 8.2

the distribution. Finally, the WinSpeed distribution looks multimodal with several prominent peaks. This suggests that there are some dominant wind speed groups or categories in the data, perhaps related to certain weather conditions or seasons.

The vertical axis represents the number of data around it. The HumPercen has a shape like a beta distribution, with one dominant peak in the middle and descending tails on either side. This indicates that most moisture percentage values are centered around a certain value, with less variation in both tails of the distribution. Finally, the WinSpeed distribution looks multimodal with several prominent peaks. This suggests that there are some dominant wind speed groups or categories in the data, perhaps related to certain weather conditions or seasons.



Rain prediction page

The results of the deployment carried out in this research used a random forest model, the results of no rain by entering the parameters minimum temperature, maximum temperature, maximum humidity, minimum humidity, humidity percentage, temperature, and wind direction. The parameters entered in the deployment are in the form of normalized data, to show the prediction process carried out in the IoT system.

V. CONCLUSION

Random forest and decision tree algorithms can be used to predict rain based on historical weather data. From the research, it was found that the decision tree algorithm is better than random forest. The difference between training and test was very small, that is 0.69%, while the difference between training and test for the random forest algorithm was 0.97%. However, in terms of accuracy, random forest is superior with a value of 94.67% (95.64% training), while decision tree accuracy is 94.17% (94.86% training).

REFERENCES

1. L. Q. Avia, "Change in rainfall per-decades over Java Island, Indonesia," IOP Conference Series: Earth and Environmental Science, vol. 374, no. 1, p. 012037, Nov. 2019, doi: https://doi.org/10.1088/1755-1315/374/1/012037.
2. Hassan, Md.Mehedi & Rony, Abu & Khan, Md & Hassan, Md & Yasmin, Farhana & Nag, Anindya & Zarin, Tazria & Bairagi, Anupam & Alshathri, Samah & El-Shafai, Walid. (2023). Machine Learning-Based Rainfall

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

Prediction: Unveiling Insights and Forecasting for Improved Preparedness. IEEE Access. 11. 132196-132222. 10.1109/ACCESS.2023.3333876.

3. P. Mishra, A. Biancolillo, J. M. Roger, F. Marini, and D. N. Rutledge,‘‘New data preprocessing trends based on ensemble of multiple prepro-cessing techniques,’’ TrAC Trends Anal. Chem., vol. 132, Nov. 2020,Art. no. 116045.
4. K. Gao, T. Liu, B. Hu, M. Hao, Y. Zhang, and X. Ning, “Establishment of Economic Forecasting Model of High-Tech Industry Based on Genetic Optimization Neural Network,” Intell. Neurosci., vol. 2022, Jan. 2022, doi: 10.1155/2022/2128370.
5. S. Zainudin, D. Jasim, and A. Abu Bakar, “Comparative Analysis of Data Mining Techniques for Malaysian Rainfall Prediction,” Int. J. Adv. Sci. Eng. Inf. Technol., vol. 6, p. 1148, 2016, doi: 10.18517/ijaseit.6.6.1487.
6. A. Salam, S. Prasetiyowati, and Y. Sibaroni, “Prediction Vulnerability Level of Dengue Fever Using KNN and Random Forest,” J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 4, pp. 531–536, 2020, doi: 10.29207/resti.v4i3.1926.
7. S. Mohsenzadeh Karimi, O. Kisi, M. Porrajabali, F. Rouhani-Nia, and J. Shiri, “Evaluation of the support vector machine, random forest and geo-statistical methodologies for predicting long-term air temperature,” ISH J. Hydraul. Eng., vol. 26, pp. 1–11, 2018, doi: 10.1080/09715010.2018.1495583.