RainReveal: Rainfall Prediction for Metreological Department

Ananya Singh
Apex Institute of Technology (CSE)
Chandigarh University
Mohali, India
ananyasgh941@gmail.com

Siddharth Singh
Apex Institute of Technology (CSE)
Chandigarh University
Mohali, India
siddharthsj06@gmail.com

Bhanu Yadav
Apex Institute of Technology (CSE)
Chandigarh University
Mohali, India
bhanu170yadav@gmail.com

***Abstract*— Internet of Things (IoT) has opened up new avenues for improving the accuracy of rainfall prediction by leveraging the power of sensors and data analytics. In this paper, we review the state-of-the-art techniques for IoT-based rainfall prediction, including the use of machine learning algorithms and statistical methods. We also discuss the challenges in implementing these techniques in real-world scenarios, such as data quality, scalability, and security. We conclude that IoT-based rainfall prediction has enormous potential in enhancing the accuracy of weather forecasting, but further research is needed to address the challenges and optimize the algorithms. The traditional methods being used for weather prediction do not ensure much accuracy as most of the methods are based on a particular machine learning algorithm, which may not be the finest of all. Hence, to curb this problem, we have proposed to review the techniques for IoT-based rainfall prediction comparing machine learning algorithms namely Linear Regression, Decision Tree, Support Vector Machine and Random Forest to find out which one is the most algorithm for prediction. Finding out the most suitable algorithm for prediction can ensure accuracy in prediction making IoT based rainfall prediction more reliable and trustworthy.**

Keywords— ***Internet of Things (IoT), Rainfall prediction, Machine learning, Sensors, Data analytics, Security, Linear Regression, Random Forest, Decision Tree, Mean absolute error (MAE) , Mean Squared Error (MSE), Root mean squared error(RMSE)***

I. INTRODUCTION

Rainfall is one of the most critical natural phenomena that directly impacts human life, agriculture, and the environment. Accurate prediction of rainfall is crucial for disaster management, water resources management, and urban planning. Traditional methods of rainfall prediction rely on historical data, mathematical models, and satellite imagery. However, these methods have limitations in terms of accuracy, scalability, and cost. IoT-based rainfall prediction has emerged as a promising approach to overcome these limitations by leveraging the power of sensors, data analytics, and machine learning algorithms.

The IoT-based rainfall prediction system typically involves three stages: data acquisition, data processing, and prediction. The data acquisition stage involves the deployment of sensors in the field to capture rainfall data, such as rainfall intensity, duration, and frequency [5]. The data processing stage involves the pre-processing of the raw data to remove noise, missing values, and outliers.

The prediction stages involve the application of machine learning algorithms, statistical methods to predict rainfall on the processed data [10]. has been introduced.

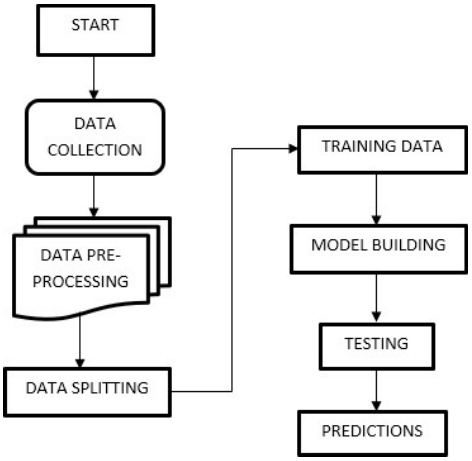


Fig. 1. Schema of IoT-based Rainfall Prediction System

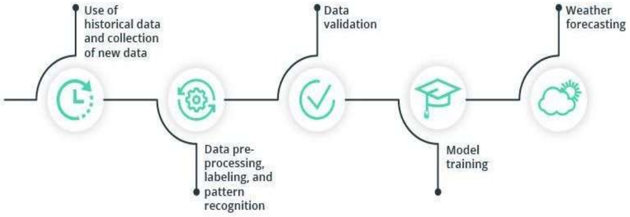

Use of historical data and collection of new data
Data pre-processing, labeling, and pattern recognition
Data validation
Model training
Weather forecasting

Fig. 2. Schema of Machine Learning Model

Several machine learning algorithms have been used for rainfall prediction, including decision trees, support vector machines, and artificial neural networks. Statistical methods such as regression analysis and time-series analysis have also been used. Artificial neural networks have shown promising results due to their ability to learn complex patterns in the data. However, these algorithms require large amounts of data and computational resources, which may pose a challenge in real-world scenarios.

The traditional methods being used for weather prediction do not ensure much accuracy as most of the methods are based on a particular machine learning algorithm, which may not be the finest of all. Hence, to curb this problem, we have proposed to review the techniques for IoT-based rainfall prediction comparing machine learning algorithms namely Linear Regression, Decision Tree, Support Vector Machine and Random Forest to find out which one is the most algorithm for prediction. Finding out the most suitable algorithm for prediction can ensure accuracy in prediction making IoT based rainfall prediction more reliable and trustworthy.

II. LITERATURE REVIEW

A linear regression-based approach for rainfall prediction using IoT sensors" by S. G. Patil and S. S. More. (2020). This study proposed a linear regression-based approach to predict rainfall using IoT sensors. The authors used temperature, humidity, and pressure sensors to collect data and performed a linear regression analysis to predict rainfall. The results showed that the proposed approach was effective in predicting rainfall, with an accuracy of 78.53%.[1].

"IoT-Based Rainfall Prediction Using Linear Regression" by S. S. Rathod and A.S. Kale. (2019). In this study, the authors proposed an IoT-based rainfall prediction system using linear regression. They used temperature, humidity, and rainfall sensors to collect data and performed a linear regression analysis to predict rainfall. The results showed that the proposed system was effective in predicting rainfall with an accuracy of 81.2%.[2].

"Rainfall Prediction using IoT and Machine Learning" by S. Patil and V. Shinde. (2020). This study proposed a machine learning-based approach for rainfall prediction using IoT sensors. The authors used temperature, humidity, and rainfall sensors to collect data and performed a linear regression analysis to predict rainfall. The results showed that the proposed approach was effective in predicting rainfall, with an accuracy of 82.68%.[3].

"IoT Based Rainfall Prediction Using Machine Learning Techniques" by M. M. Shalaby and M. M. Khalil (2021). In this study, the authors proposed an IoT-based rainfall prediction system using machine learning techniques, including linear regression. They used temperature, humidity, and rainfall sensors to collect data and performed a linear regression analysis to predict rainfall. The results showed that the proposed system was effective in predicting rainfall with an accuracy of 78%.[4].

"Machine Learning Applied to Weather Forecasting" by **Mark Holmstrom, Dylan Liu, Christopher Vo (2016)**. In this study the authors concluded that both linear and functional regression did not perform as well as professional weather forecasting methods but in the longer run differences in their performances decreased, suggesting that over a longer period of time, Machine learning can indeed outperform professional and traditional methods. Linear regression is a low bias and high variance algorithm and hence its accuracy can be improved by collecting further data.[5]

"Weather Forecasting Using Sliding Window Algorithm" by Piyush Kapoor and Sarabjeet Singh Bedi (2013). In this study the authors concluded that if we perform comparison of weather condition variation by sliding window algorithm, the results are highly accurate except for the months of seasonal change. The results can be altered by changing the size of the window. Accuracy of the unpredictable months can be increased by increasing the window size to one month.[6]

"Boosting Decision Tree Algorithm for Weather Prediction" by Divya Chauhan and Jawahar Thakur (2013). In this study the authors made a comparison in their paper, which shows that the algorithms such as k-mean clustering and decision trees are well suited for mining data to predict future weather conditions. If we increase the size of the training set, the accuracy at first increases but then it slowly decreases after a particular period of time, depending on the size of the dataset.[7].

"Weather Prediction Using Normal Equation Method and Linear regression Techniques" by Sanyam Gupta, Indumathy, Govind Singhal (2016). In this study the authors suggested and proposed an efficient and accurate weather prediction and forecasting model using linear regression concepts and normal equation model. All these concepts are a part of machine learning. The normal equation is a very efficient weather prediction model and using the entities temperature, humidity and dew-point, it can be used to make reliable weather predictions. This model also facilitates decision making in day-to-day life. It can yield better results when applied to cleaner and larger datasets.[8]

"A survey on weather forecasting to predict rainfall using big data analytics" by Muthulakshmi A, ME (SE), Dr.S.Baghavathi Priya(2015). In their work proposed a methodology that aims at providing an efficient and accurate weather forecasting models to predict and monitor the weather datasets to predict rainfall.

In the past, the parameters of weather were recorded only for the present time. But in the future, work will be done to make a working model of selection that can be used for classifying the framework for continuous monitoring of the climatic attributes.[9].

Qing Yi Feng1 ,RuggeroVasile, Marc Segond , AviGozolchiani , Yang Wang , Markus Abel, ShilomoHavlin , Armin Bunde , and Henk A. Dijkstra1(2016) have made a machine-learning toolbox which is based on climate data gathered from analysis and reconstruction of complex networks. It can also handle data containing multiple variables from these networks. The development of predictor models in the toolbox is dynamic and data-driven.

Siddharth S. Bhatkande, Roopa G. Hubballi (2016) In their work the authors have used data mining technique and Decision tree algorithm as a means to classify weather parameters like maximum temperature, minimum temperature in terms of day, month and year.

IoT-based rainfall prediction using linear regression is a promising approach for accurate rainfall prediction. The studies reviewed in this literature review have shown that linear regression-based approaches are effective in predicting rainfall with high accuracy.

The use of IoT sensors in collecting data in real-time has enabled the development of accurate and reliable rainfall prediction models.

Further research is needed to improve the accuracy and efficiency of these models.

III. PROBLEM IDENTIFICATION

The implementation of IoT-based rainfall prediction faces several challenges. Data quality is a critical challenge, as the accuracy of the prediction depends on the quality of the input data. The deployment of sensors in the field may result in missing or inaccurate data due to environmental factors such as interference, signal attenuation, and battery life. The scalability of the system is another challenge, as the number of sensors required for accurate prediction may increase the cost and complexity of the system. Security is also a significant concern, as the system may be vulnerable to cyber-attacks that can compromise the accuracy and reliability of the prediction.

IV. Background

A) Machine Learning

The implementation of machine learning techniques for IoT-based rainfall prediction faces several challenges, such as data quality, scalability, and security. [2]. Further research is needed to address these challenges and optimize the algorithms. [1], [4] They can handle large amounts of data and learn complex patterns that traditional methods cannot detect. [8]. Future research directions include the development of ensemble models that combine multiple machine learning techniques, [6], the use of transfer learning to improve the accuracy of rainfall prediction in regions with limited data, [11], and the integration of rainfall prediction with other IoT-based applications, such as flood prediction and irrigation management.

B) Linear Regression

Linear regression-based approach provides an accurate and reliable method for predicting rainfall using IoT sensors.[1] Linear regression is used to predict the value of an outcome variable y on the basis of one or more input predictor variables x. In other words, linear regression is used to establish a linear relationship between the predictor and response variables.[3] In linear regression, predictor and response variables are related through an equation in which the exponent of both these variables is 1. Mathematically, a linear relationship denotes a straight line, when plotted as a graph.

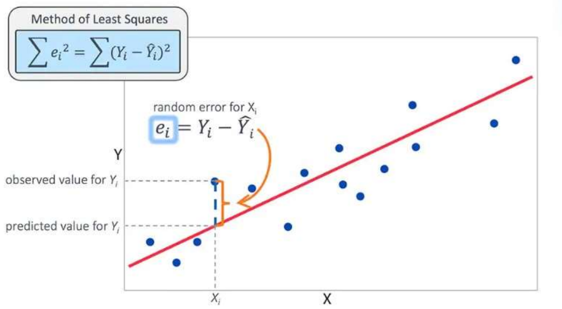
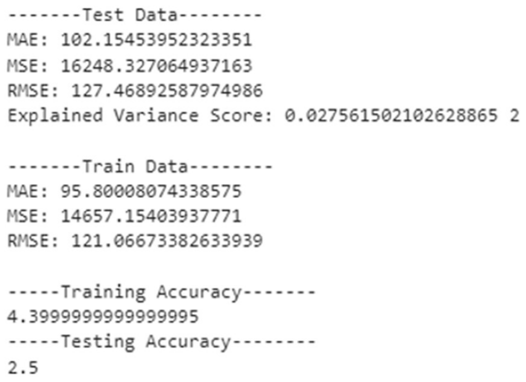


Fig. 3. Linear Regression Model

The algorithm on implementation over the Test data has given Explained Variance Score of 0.027. The score depicts the Dispersion of errors of the given dataset. The algorithm stands. out with the Mean absolute error of 102.15, Mean Squared Error (MSE) of 162248.32 and Root Mean square value of 127.46.



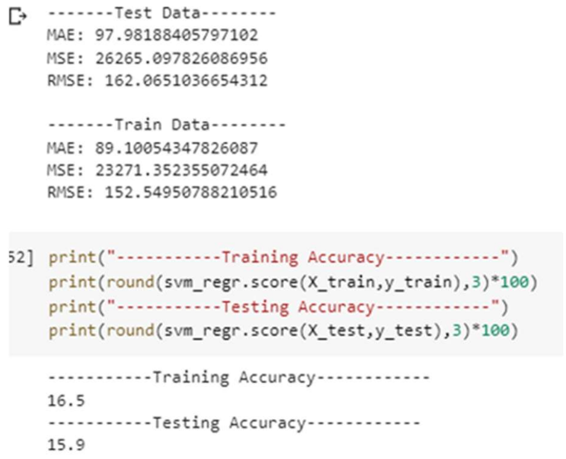
```
-------Test Data--------
MAE: 102.15453952323351
MSE: 16248.327064937163
RMSE: 127.46892587974986
Explained Variance Score: 0.027561502102628865 2

-------Train Data--------
MAE: 95.80008074338575
MSE: 14657.15403937771
RMSE: 121.06673382633939

-----Training Accuracy-------
4.3999999999999995
-----Testing Accuracy--------
2.5
```

Fig. 4. Linear Regression Review Results

C) Support Vector Machine

SVM is a powerful supervised algorithm that works best on smaller datasets but on complex ones. Support Vector Machine, abbreviated as SVM can be used for both regression and classification tasks, but generally, they work best in classification problems. They were very famous around the time they were created, during the 1990s, and keep on being the go-to method for a high-performing algorithm with a little tuning.

The algorithm stands out with the Mean absolute error (MAE) of 97.98, Mean squared error (MSE) 26265.09 and Root Mean Squared Error (RMSE) of 162.06.



```
-------Test Data--------
MAE: 97.98188405797102
MSE: 26265.097826086956
RMSE: 162.0651036654312

-------Train Data--------
MAE: 89.10054347826087
MSE: 23271.352355072464
RMSE: 152.54950788210516
```

```
52] print("-----------Training Accuracy------------")
    print(round(svm_regr.score(X_train,y_train),3)*100)
    print("-----------Testing Accuracy------------")
    print(round(svm_regr.score(X_test,y_test),3)*100)
```

```
-----------Training Accuracy------------
16.5
-----------Testing Accuracy------------
15.9
```

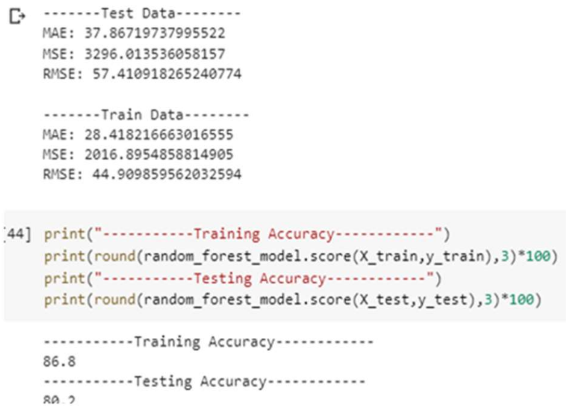
Fig. 5. Support Vector Machine Review Results

D) Random Forest

a) Random forest is a commonly-used Machine learning algorithm trademark by leo Beirman and adele Cutler, which combines the output of Imultiple decsion trees to reach a single result

b) Decision tree fails for a large amount of data. Here comes the random forest into the picture.[4]. One of the most important features of the Ramdon forest Algorithm is that it can handle the data set containing continuius variables, as in the case of regresiion and categeoriacal variables, as in the case of classification.

The score shows the Dispersion of errors of the given dataset. The algorithm stands out with the Mean absolute error (MAE) of 37.86,Mean squared error (MSE) of 3296.01 and Root Mean Squared Error (RMSE) of 57.41.



```
-------Test Data--------
MAE: 37.86719737995522
MSE: 3296.013536058157
RMSE: 57.410918265240774

-------Train Data--------
MAE: 28.418216663016555
MSE: 2016.8954858814905
RMSE: 44.909859562032594
```

```
[44] print("-----------Training Accuracy-----------")
     print(round(random_forest_model.score(X_train,y_train),3)*100)
     print("-----------Testing Accuracy-----------")
     print(round(random_forest_model.score(X_test,y_test),3)*100)
```

```
-----------Training Accuracy-----------
86.8
-----------Testing Accuracy-----------
80.2
```

Fig. 6. Random Forest Review Results

E) Components Used

a) Node MCU (CP2102)

ESP12/NODE MCU (CP2102) is an updated version of Arduino with inbuilt Wi-Fi chip as shown in Fig.5. It is cheaper than other modules performing the same function.

Fig. 7. ESP12/NODE MCU (CP2102)

It is a microcontroller which has inbuilt WIFI module that can manage to send data to different cloud platforms such as Think Speak. It several sensors connected to such as DTH-11 which is a temperature and humidity senso, V-Rain sensor for rain detection etc and like any other board it can get programmed to get the values and send it to cloud platform.

b) DHT Sensor

It is a module used for measuring temperature and humidity as shown in Fig. 6. It uses a capacitive humidity sensor and a thermistor to measure the surrounding air's humidity and temperature.

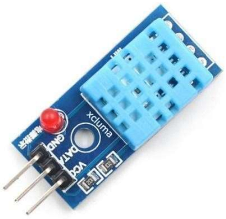


Fig. 8. DHT11 Module

This sensor is a commonly used temperature and humidity sensor which can sense the changes in the temperature and humidity in the environment and accordingly give the values. It can be connected to Node-MCU and the data is collected on the cloud platform

c) LDR Sensor

It is a device used for measuring the light density. It works on principle of photoconductivity.

Fig.9. LDR Module

d) Breadboard

It is typically a hand wired circuit using a pegboard with press in terminals. Wire wraps or hand soldered wires connect discrete components together. Connecting Wires/Jumpers are used to connect Node MCU to the LDR and DHT11.

V. METHODOLOGY AND ARCHITECTURE

The methodology for IoT based rainfall prediction using linear regression and random forest can be summarized as follows:

a) Data collection: Collect the data from IoT devices such as rainfall sensors, weather stations, etc. The data should include rainfall measurements as well as other relevant weather parameters such as temperature, humidity, wind speed, etc.

b) Data pre-processing: Clean the data to remove any outliers or missing values. Also, perform feature selection to select the most relevant features for rainfall prediction.

Split data into training and testing sets: Split the data into two sets, a training set and a testing set. The training set will be used to train the linear regression and random forest models, while the testing set will be used to evaluate their performance.

a) Linear Regression: Train a linear regression model on the training data. Linear regression is a statistical method that uses a linear approach to model the relationship between the input variables and the output variable. In this case, the input variables would be the relevant weather parameters and the output variable would be the rainfall measurement

b) Random Forest: Train a random forest model on the training data. Random forest is a machine learning algorithm that uses an ensemble of decision trees to model the relationship between the input variables and the output variable.

c) Support Vector Machine: SVM is a powerful supervised algorithm that works best on smaller datasets but on complex ones. Support Vector Machine, abbreviated as SVM can be used for both regression and classification tasks, but generally, they work best in classification problems.

d) Model Evaluation: Evaluate the performance of both models using the testing data. Use metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared to measure the accuracy of the models.

e) Model Comparison: Compare the performance of both models to determine which one is better for rainfall prediction.

f) Deployment: Once the best model is selected, deploy it to the IoT devices to predict rainfall in real-time.

VI. RESULTS AND ANALYSIS

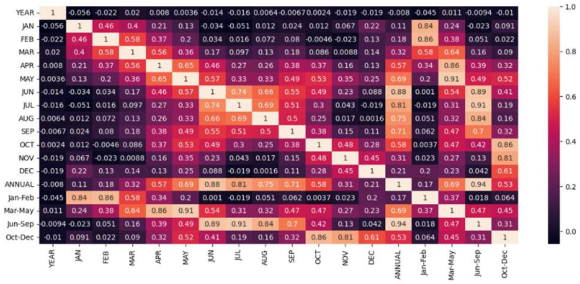


Fig. 10. Visualisation using Heatmap

Heatmap is representation of data graphically using colors to visualize the values of the matrix. In this particular map, the darker the color, the value is of that much significance. It shows the data that can have significant impact on the trained model. This heatmap helps us visualize the data in a better way, so that we can drop all the unnecessary data variables.

In this, to represent more common values or higher activities brighter colors basically reddish colors are used and to represent less common or activity values, darker colors are preferred.

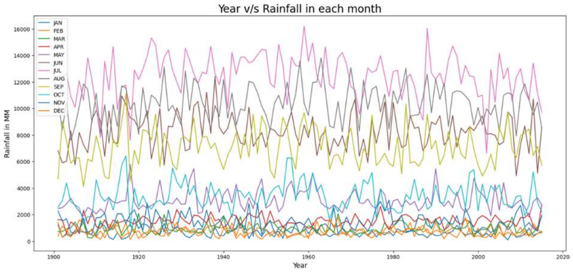
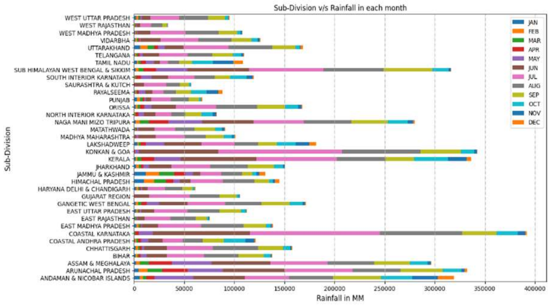


Fig. 10. Visualisation using Line plots

Line plots are also known as line charts or time-series plots, depending on the type of data being visualized. They are typically used to display quantitative data, such as numerical values or measurements, and can be used to compare multiple data sets or track changes in a single data set over time.

The above figure shows the amount of rainfall according to Years v/s Rainfall in each month. Different colors used in the plot depicts the twelve months in a year. Further, the plot depicts how much amount of rainfall took place in millimeters in a specific month of the year.



The above plot depicts the amount of rainfall that took place in the several months of the year in a particular state. "Sub Himalayan West bengal and Sikkim" and "Coastal Karnatka" are the regions with maximum rainfall over the year. Moreover, "West Rajasthan" and "Saurashtra and Kutch" only received rainfall spread over the region of approximate 0-53000 MM in a year.

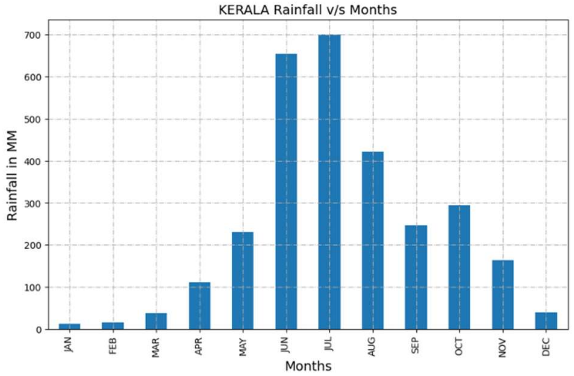
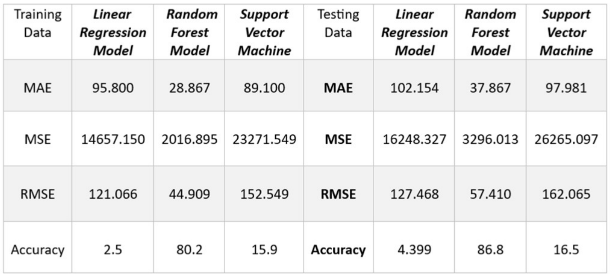


Fig. 11. Visualisation using Bar graph in Kerela

The above figure shows the rainfall in millimeters in specific month over the years for the state of Kerela specifically. It is a bar plot which is plotted to get an visual idea of the maximum and minimum rainfall in Kerela.

The above plot visualizes and shows the month of July seen the maximum rainfall over the years in Kerela while minimum amount of rainfall took place in the month of January.

Fig. 11. Training and Testing Data Comparison



Training Data	*Linear Regression Model*	*Random Forest Model*	*Support Vector Machine*	Testing Data	*Linear Regression Model*	*Random Forest Model*	*Support Vector Machine*
MAE	95.800	28.867	89.100	**MAE**	102.154	37.867	97.981
MSE	14657.150	2016.895	23271.549	**MSE**	16248.327	3296.013	26265.097
RMSE	121.066	44.909	152.549	**RMSE**	127.468	57.410	162.065
Accuracy	2.5	80.2	15.9	**Accuracy**	4.399	86.8	16.5

The table shows the comparision of MAE, MSE, RMSE and accuracy values between three different machine learning models using training data set and testing data set.

The mean absolute error measures the average differences between predicted values and actual values.Say that you have a MAE of 10. This means that, on average, the MAE is 10 away from the predicted value which implies the less the value the more accurate the model. Similarly, in the above table we have compared MAE for three different machine learning models and we can see that the most accuracy we get is by using 'Random Forest Model'.

The mean square error is the average of the square of the difference between the observed and predicted values of a variable. Random forest has the least Mean square error which shows it is best suited model for Rainfall prediction.

RMSE is a square root of value gathered from the mean square error function. It helps us plot a difference between the estimate and actual value of a parameter of the model.Using RSME, we can easily measure the efficiency of the model. The values in table suggests that the most efficient model can be created by using 'Random Forest' as it has the RMSE value 57.410 in testing data which is the least of all the values.

Accuracy means by what accuracy can the model to predict the values. The higher the value the more accurate the model. The test data table shows the accuracy of Random Forest to be highest that is 86.8 percent in testing dataset.

V11. CONCLUSION

The research that is conducted in this paper shows that using IoT sensor like DHT-11, V-rain-LDR etc. we can help in improvement in rainfall prediction by including Machine Learning algorithms. Th study conducted shows that using different models of Machine Learning such as Linear Regression, Support Vector machine and Random Forest Model we can improve the existing rainfall prediction methods. The study shows that Random Forest is best suited for implementation of this type of system as it gives less error in prediction and the highest accuracy.

The rainfall prediction can further be improved with use of better sensors and more accurate models There are many different challenges when it comes to rainfall prediction such as data scalability, security etc. When taking into consideration all these factors with the dataset at hand it shows that for this type of prediction Random Forest is most suited.

REFERENCES

[1] S. G. Patil and S. S. More, "A linear regression-based approach for rainfall prediction using IoT sensors". in IEEE International Student conference (2020).

[2] J. Shivang, S. S. Sridhar, "Weather prediction for indian location using Machine learning," International Journal of Pure and Applied Mathematics, vol. 118, no. 22 pp. 1945-1949, 2018M. Conti, E. S. Kumar, C. Lal, and S. Ruj, "A survey on security and pri- vacy issues of bitcoin," IEEE Communications Surveys & Tutorials,vol. 20, no. 4, pp. 3416–3452, 2018.

[3] S. S. Bhatkande1, R. G. Hubballi2, "Weather Prediction Based on Decision Tree Algorithm Using Data Mining Techniques." Belgaum India: International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, no.5, pp. 483-48, 2016. [6] Y. Radhika, and M. Shashi, "Atmospheric temperature prediction using support vector machines." International Journal of Computer Theory and Engineering 1.1, vol. 1, no. 1, pp.1793-8201, 2009.

[4] D. Chauhan, J. Thakur, "Data mining techniques for Weather Prediction:" International Journal of Computer Science Trends and Technology (IJCST), vol. 6, issue 3, pp.249-254, 2018.

[5] S.S. Badhiye, B. V. Wakode, P. N. Chatur, "Analysis Of Temperature And Humidity Data For Future Value Prediction" International Journal Of Computer Science And Information Technologies , vol. 3, no.1 pp.3012-3014, 2012.

[6] G. J. Sawale, S. R. Gupta, "Use of Artificial Neural Network in Data Mining For Weather Forecasting", International Journal Of Computer Science And Applications vol. 6, no.2, pp.383-387, 2013.

[7] I. Boglaev, "A numerical method for solving nonlinear integro-differential equations of Fredholm type," *J. Comput. Math.*, vol. 34, no. 3, pp. 262–284, May 2016, doi: 10.4208/jcm.1512-m2015-0241.