Rainfall Prediction using ARIMA and Linear Regression

Supriya S Surya P Swetha N Sangeetha K

*Department of Computer Department of Computer Department of Computer Department of Computer*

*Science Engineering, Science Engineering, Science Engineering, Science Engineering,*

*Panimalar Engineering Panimalar Engineering Panimalar Engineering Panimalar Engineering*

*College, Chennai, India. College, Chennai, India. College, Chennai, India. College, Chennai, India.*

[supriyapriya021@gmail.com](mailto:supriyapriya021@gmail.com) [suryapandian05@gmail.com](mailto:suryapandian05@gmail.com) [swetha3118@gmail.com](mailto:swetha3118@gmail.com) sangeethakalyaniraman@gmail.com

***Abstract*—Rainfall is the greatest of nature's gifts for our daily life, as well as the most important climate factor affecting human lives with farmers and agricultural complex systems. Rainfall forecasting is critical because excessive and irregular rainfall can have numerous consequences, such as crop destruction and property damage, so a better forecasting model is required for early warning that can limit risks to life and property while also better managing agricultural farms. Time series data have been used extensively in classical statistics. The proposed methodology predicts annual rainfall by time series ARIMA model and Linear Regression a machine learning algorithm. Time series data have been used extensively in classical statistics. The ARIMA has been trained to produce excellent outcomes. The ARIMA model demonstrated greater accuracy in all seasonal and yearly rains. To offer a solid prediction, this method, like time series ARIMA, requires a strict assumption of stationarity. We use real data from the Indian government website and Kaggle to compare model quality in ARIMA using different evaluation metrics. As a result, the ARIMA model accurately forecasts rainfall with less error, and the resultant model can be used to forecast rainfall for future years.**

***Keywords—*Rainfall, Forecasting, Machine Learning, Regression, Time series analysis, ARIMA.**

# INTRODUCTION

The rainfall analysis aids in the prevention and reduction of disasters such as droughts and floods. We can simply do it here by analysing historical rainfall data and forecasting rainfall for future seasons. According to the needs, we may apply a variety of approaches such as classification and regression, and we can also calculate the error between the real and predicted values as well as the accuracy. Because different methodologies yield varying degrees of accuracy, it is critical to select the appropriate algorithm and model in accordance with the requirements.

The machine learning technology was used to predict rainfall. Machine learning allows systems to learn and improve from experience without being designed by humans. The development of the machine learning idea has greatly simplified data analysis and prediction. Machine learning does not necessitate an understanding of the physical mechanisms that govern the atmosphere, but rather analyses past data to forecast future data. As a result, this procedure could be utilized to forecast weather. Rainfall data is time-series data that fluctuates as the climate and seasons change. The goal of this project is to predict rainfall using Linear Regression (LR) Auto-Regressive Integrated Moving Average (ARIMA) approaches. The parameters of the LR equation are obtained from the dataset, and the variables are derived from the dataset via correlation. ARIMA is an effective statistical technique for predicting rainfall modelling time series. The ARIMA model development technique includes iterative stages of discovery, estimation, diagnosis, and forecasting. Once chosen, the model is used to forecast monthly or seasonal rainfall series.

Timeseries forecasting is important for making predictions and informed strategic decisions, as well as predicting future rainfall using previously observed data. Time series analysis is a common approach to multivariate statistical analysis. It is used to forecast rainfall based on rainfall time series features. Exploration of time series offers numerous options for discovering, describing, and modelling climate inconsistencies and impacts. Modelling can be done using historical weather data collected by meteorological stations located throughout India.

# LITERATURE REVIEW

There are a lot of researchers who have carried out rainfall prediction using a different methodology.

Ashwini and Kalaivani [1] use time series machine learning model called SARIMA is used for forecasting rainfall in Tamilnadu. Forecasting data required for the analysis is available in the Indian meteorological department. Anosh and Mishra [2] used Seasonal ARIMA to predict the rainfall for the Allahabad region, using seasonal ARIMA provides a satisfactory and consistent result for prediction on month-wise but the Accuracy is reduced by some missing values.

Pengcheng and Yangyand [3] use DRCF and MLP methods to predict short-term rainfall it solves the clustered inference which is caused by the extension of the perception range, and DRCF is enhanced with several dynamic strategies. But the interval of DRCF is limited to 3 hours and cannot get a global optimal structure. Nitin Singh and Shamim [4] use raspberry pi and a random forest algorithm to predict using a low cost and portable solution and in this method, there are many false predictions of rainfall. Arnav and Pandey [5] use three types of methods SVR, AVM and KNN algorithm. Although all the algorithms are equally useful SVM and SVR are more accurate than KNN. Even though SVR perform better it performs a little lower due to nature anomalies.

Bhavya and Sowmya [6] used ARIMA, ANN, Logistic regression and SVM algorithms. Here ANN makes a superior solution to all algorithms but it will be difficult to predict even if small changes occur in climate. Moulana Mohammad [7] uses SVR methodology to predict rainfall in non-expert find very easy to access but the main challenge to building a model is for long term rainfall prediction.

Anwar and Windari [8] proposed a method called J48 to predict rainfall. By using this method it gives 86% but even if small changes in the data will result in a drastic change in the prediction result. Wanie [9] used BLR, DFR and NN for forecasting the rainfall it shows high accuracy but the accurate rainfall is predicted using hybrid machine learning models. Sun and Wu [10] implement GRU in their project, this method reduces the error obtained between the predicted and real values and improves the accuracy of rainfall prediction. Najim Ussiph And Emmanuel Ahene [11] use DT, RF, MLP, XGB and KNN algorithms to find the rainfall forecasting in Ghana. Here, the Decision tree is considered as fastest. But KNN performed worst in all zones on both train and test ratios.

# PROPOSED METHODOLOGY

1. **ARIMA:**

ARIMA stands for Auto-Regressive Integrated Moving Average which is used to perform time series forecasting. Time Series is nothing but a series of observations carried over a time period. There are two types of time series, univariant time series which contain values taken from a single variable and multivariant time series contain values taken from multiple variables.

ARIMA has p, d, q parameters these values are used to reduce the between the actual value and predicted values. Here we are choosing grid search here we can build a model by trying all the possible values and find the one which has a minimum error. This process is a time-consuming process when the number of values is large.

AR term refers to Auto-Regressive process it is defined by the current values is depends on its own p-previous value. P is the order of the AR process. MA term refers to Moving Average process here current deviation on q-previous deviation. Q is the order of the MA process.

The time Series has to be stationary, if the values are not stationary, we have to make it stationary. So that we have to set the parameter d=1. This is to remove the upward trend; therefore, it is called 1st order differencing.

ACF stands for Autocorrelation function. It is used to plot the correlation point against the log. PACF stands for Partial Auto Correlation Function which plots the correlation point between two variables.

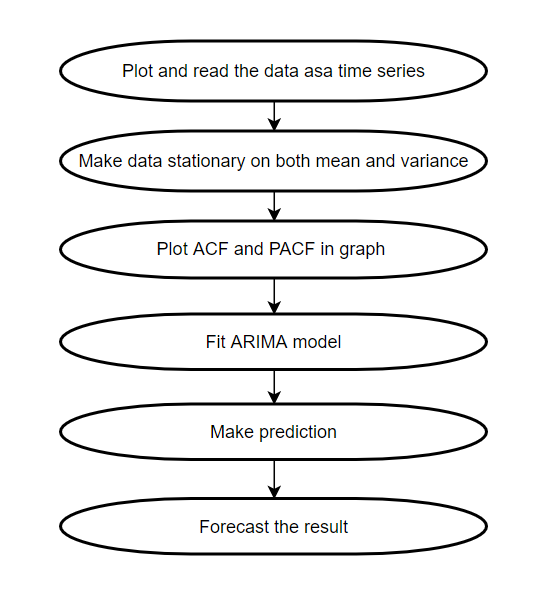


Fig 1. ARIMA Flow Diagram

**STEPS IN ARIMA:**

**Step 1:** Creating Models

This stage carefully selects a standard ARIMA formulation to model the rainfall data. This is done by carefully inspecting and selecting the most important characteristics of rainfall intensity and other weather parameters. Precipitation and rainfall are the data sources for this project.

**Step 2:** Identifying the Model

For the rainfall data, a trial model must be identified. To begin, the original rainfall data must be transformed in order to render the underlying mechanism stationary. The autocorrelation function (ACF) or unit root test can be used to check the data in this stage. Additional lag residual and lag dependent tests were performed using partial ACF.

**Step 3:** Estimation of Parameters

The parameters of the model's elements must be evaluated after the model's elements have been determined. Assuming the data are observations of a stationary time series, good parameter estimators can be derived (Step 2). If a Moving Average (MA) pattern is found, the next step is to use maximum likelihood or least square estimate to complete the optimization process.

**Step 4:** Validation of Hypotheses

If the model's assumptions are confirmed, proceed to Step 4, otherwise, return to Step 2 to enhance the model. The model assumptions made in Step 1 are validated using a diagnosis verification. This diagnosis checks whether the leftover hypotheses are correct.

**Step 5:** Prediction

Forecasting is now possible with the model. Predict future values of daily rainfall data using the model from Step 3.

1. **LINEAR REGRESSION:**

Regression is defined as an algorithm that is used to predict perpetual outcomes. It is a relation between features and outcomes. Algorithms are trained in such a way to understand the connection between independent variables and outcomes.

The Linear regression algorithm is a simple form of an algorithm. The main goal of the algorithm is to minimize the difference in elevation. The target of linear regression is to produce a quantitative type as output. **It is a method used to determine future outcomes.**

The line that provides the connection between the predictor variable and target variable is known as the linear regression line. This exhibits two types of relationships. The positive linear relationship is the target variable increases on the y-axis and the predictor variable increases on the x-axis, there exhibits a positive linear relationship. The negative linear relationship is the target variable decreases on the y-axis the and predictor variable decreases on the x-axis, there exhibits a positive linear relationship.

**The line of best fit is the one that has minimum error between predicted values and true values, it will have the least errors and is calculated by least squares regression and gradient descent.**

Linear regression is of two types: Simple linear regression is a single predictor variable is used to predict the value of a numerical target variable, this type is known as simple linear regression and it works with only one predictor variable. Multiple linear regression is more than one predictor variable is used to predict the value of a numerical target variable, this type is known as multiple linear regression and it works with more than one predictor variable.

# SYSTEM ARCHITECTURE

|  |  |  |
| --- | --- | --- |
| Features | R-Squared | R-Squared % |
| TemperatureMin | 0.999 | 99.9 |
| TemperatureMax | 0.996 | 99.6 |
| PrecipIntensity | 0.643 | 64.3 |

The rainfall dataset is collected in CSV file format. The dataset contains the month-wise aggregation. The dataset might contain empty values, negative values or errors. Dataset is cleaned in the pre-processing stage. The pre-processing methods involve removing records which is not complete. In this stage we have to make a dataset which has non-NAN (Not A Number) values, if the dataset has NAN values it should be replaced by 0 or succeeding values or by using the dropping option. Once the cleaned dataset is available, we have to prepare it to feed into the feature extraction process.

Feature Extraction is useful when you have a large dataset and need to reduce the number of resources without losing any important or relevant information. In this project, the original dataset contains unwanted information which is not required for training the model so, by using the feature extraction process only required information is extracted from the dataset.

After that, the dataset is given as input to the ARIMA model. ARIMA is a Statistical Analysis model that uses time-series data to predict the future value based on the past results. Linear regression is a machine learning algorithm for Supervised Learning. It performs regression tasks. The regression model is a value based on the target prediction independent variable.

After forecasting, the frequent itemset Searches for the frequent item in the dataset from the frequent itemset process frequent item can be retrieved. The frequent item is further divided into temperature, humidity and rainfall datasets. The overall future rainfall accuracy is predicted using parameters like temperature, humidity and rainfall dataset and validate the accuracy.

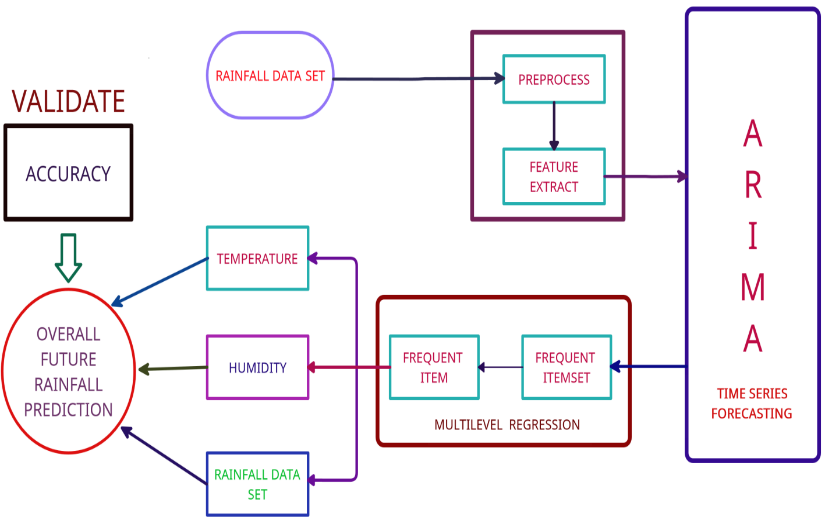


Fig 2. Architecture Diagram

# RESULT AND DISCUSSION

**A. LINEAR REGRESSION**

In Linear regression, the accuracy of the project is predicted and supported by the R-squared value. The R-squared value is often between 0 and 1. If the R-squared value is sort of capable 1 means the regression model fits the observation. R-Squared value of nearly 100% means the model is fit and performed well.

Table 1. R-Squared %

Fig 3. Statistical summary of R-Squared %

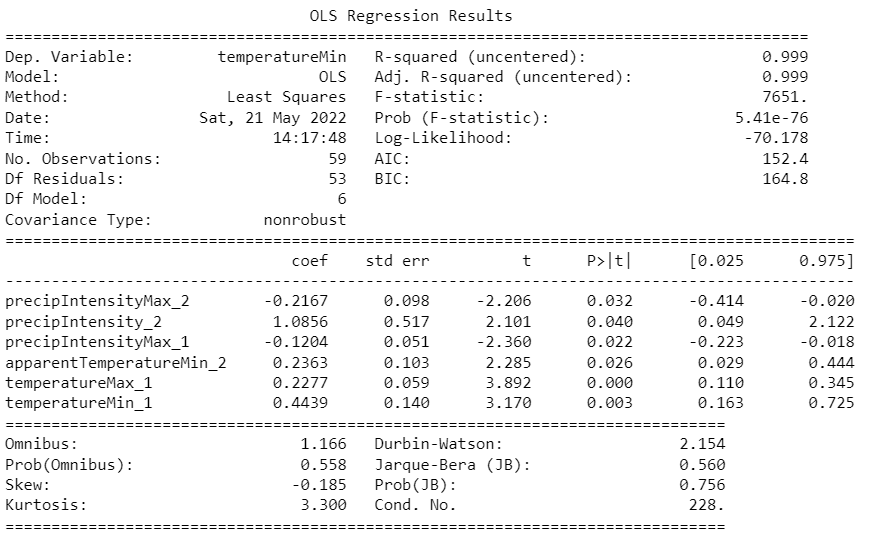


Fig.4 Linear Regression Summary

**B. ARIMA**

First, calculate the mean squared error of this model. And the root means squared error for this particular model. Also, the value of the **root mean squared should be very smaller than the mean squared error.** In this case, we can see the average error is going to be roughly 5.26/27.68 \*100=19% of the actual value.

Table 2. Error percentage

|  |  |  |  |
| --- | --- | --- | --- |
|  | MSE | RMSE | Error % |
| Observed | 27.68 | 5.26 | 19% |
| Forecasted | 1058.85 | 32.54 | 3% |

Here in our project, the forecasted error is very low compared to the observed error so this model is the best fit for rainfall prediction.

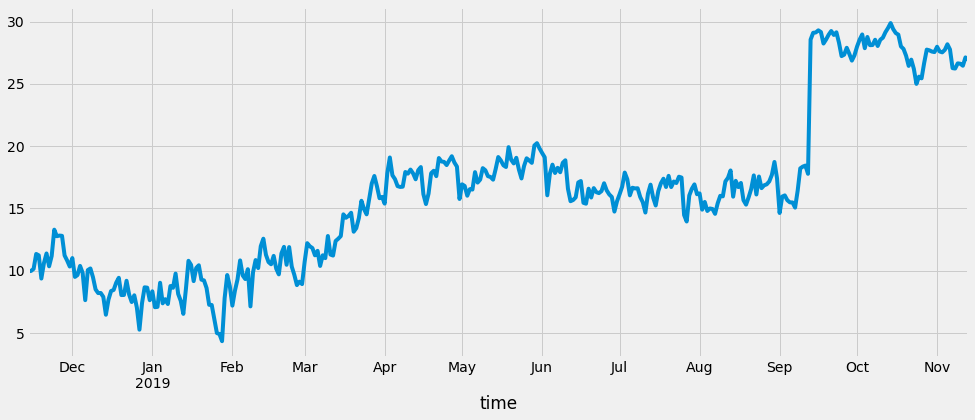


Fig.5. Observed Rainfall Series of Kerala

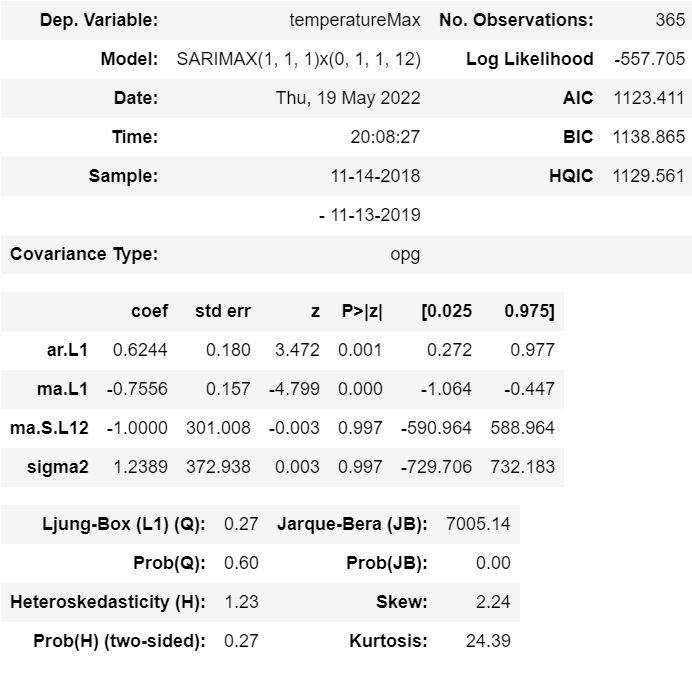


Fig.6 ARIMA Summary

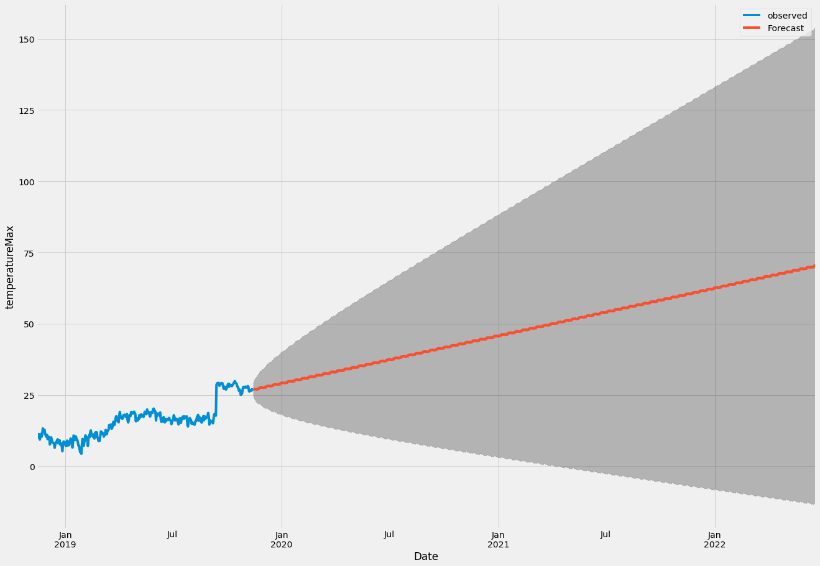


Fig.7. Temperature Minimum prediction

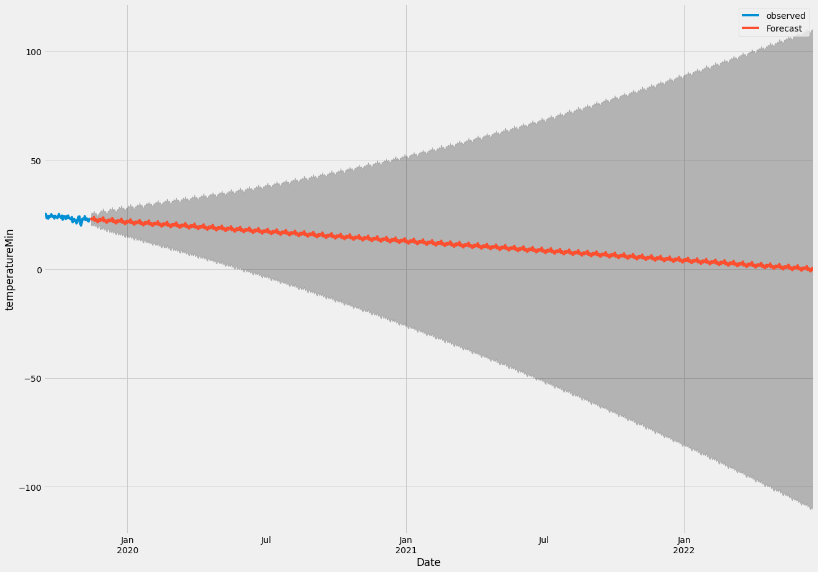
****

Fig.8. Temperature Maximum prediction

# FUTURE SCOPE AND CONCLUSION

The proposed study introduces a rainfall recommendations system that uses ARIMA and Linear Regression to provide computer-assisted results. Depending on the data set used in a particular area, the model focuses on the variety of rainfall and its production in each area, as well as the weather and seasonal patterns. The algorithms used for the recommendation function with ARIMA can be achieved when it rains during the rainy season, so relationships between parameters (such as optimal temperature, rainfall, wind speed, humidity, soil availability, and seed varieties required), rainfall, and region have been established. investigated and demonstrated.

The response is measurable and may be used to calculate the recommended rainfall in additional provinces in the same way as the method. This work can be continuously improved to avoid the problem of inequality in production and demand by adding humidity and airspeed to all regions, which will result in a more accurate recommendation. Rainfall, irrigation, and other features may be added to the system to improve its output. In addition, the recommendation can be modified to warn of timely rainfall data over the course of a particular season and to suggest the types of fertilizers or nutrients needed in the soil for the crop to thrive and produce its superior ARIMA accuracy.

# REFERENCE

[1]. U.Ashwini, K.Kalaivani\* , K.Ulagapriya and A.Saritha, year 2021, “*Time Series Analysis based Tamilnadu Monsoon Rainfall Prediction using Seasonal ARIMA*” ,IEEE, Chennai, India.

[2]. Anosh Graham , Ekta Pathak Mishra, year 2017, “*Time series analysis model to forecast rainfall for Allahabad region”,* Journal of Pharmacognosy and Phytochemistry, Uttar Pradesh, India.

[3]. Pengcheng Zhang, Yangyang Jia, Jerry Gao, Wei Song, Hareton Leung, year 2018, “*Short-term Rainfall Forecasting Using Multi-layer Perceptron*”, IEEE, China.

[4]. Nitin Singh, Saurabh Chaturvedi, Shamim Akhter, year 2019, “*Weather Forecasting Using Machine Learning Algorithm*”, IEEE, noida, India.

[5]. Arnav Garg, Himanshu Pandey, year 2019, “*Rainfall Prediction Using Machine Learning”,* International Journal of Innovative Science and Research Technology, Kanchipuram Tamil Nadu, Volume 4, Issue 5.

[6]. CMAK Zeelan Basha, Nagulla Bhavana, Ponduru Bhavya, Sowmya V, year 2020, “*Rainfall Prediction Using Machine Learning & Deep Learning Techniques* “, International Conference on Electronics and Sustainable Communication Systems, Guntur, India.

[7]. Moulana Mohammed, Roshitha Kolapalli, Niharika Golla, Siva Sai Maturi, Year 2020, "*Prediction Of Rainfall Using Machine Learning Techniques*", international journal of scientific & technology research, volume 9, issue 01.

[8]. Uchamad Taufiq Anwar, Saptono Nugrohadi, Vita Tantriyati, Vikky Aprelia Windarni, year 2020, “*Rain Prediction Using Rule-Based Machine Learning Approach*”, Advance Sustainable Science, Engineering and Technology, Indonesia.

[9]. Wanie M. Ridwan, Michelle Sapitang, Awatif Aziz, Khairul Faizal Kushiar, Ali Najah Ahmed, Ahmed El-Shafie, year 2020, “*Rainfall forecasting model using machine learning methods*”, Ain Shams Engineering Journal, Terengganu, Malaysia.

[10]. Dechao Sun, Jiali Wu, Hong Huang, Renfang Wang , Feng Liang, and Hong Xinhua, year 2021, “*Prediction of Short-Time Rainfall Based on Deep Learning*”, Hindawi, Ningbo, China, Volume 2021, 8 pages.

[11]. Nana kofi ahoi appiah-badu, yaw marfo missah, leonard k. amekudzi, najim ussiph, twum frimpong, and emmanuel ahene, year 2021, “*Rainfall Prediction Using Machine Learning Algorithms for the Various Ecological Zones of Ghana*”, IEEE, volume 10.