INNSPUB

Journal of Biodiversity and Environmental Sciences (JBES)
ISSN: 2220-6663 (Print) 2222-3045 (Online)
Vol. 23, No. 5, p. 23-30, 2023
http://www.innspub.net

RESEARCH PAPER

OPEN ACCESS

A novel rainfall prediction model for North-East region of India using stacked LSTM model

Satyajit Sarmah*[1], Rajdeep Kr. Dutta[1], Chandrikka Pathak[1], Rubul Kumar Bania[2]

[1] *Department of Information Technology, Gauhati University, Guwahati, Assam, India*

[2] *Department of Computer Science, Birangana Sati Sadhani Rajyik Viswavidyalaya, Golaghat, Assam, India*

Article published on November 10, 2023

Key words: Rainfall prediction, Multivariate linear regression, Decision tree regressor, Random forest regressor, XGBoost regressor

Abstract

The main objective of this work is to analyze different atmospheric factors and their correlation to daily rainfall, and then further compare them using various machine learning models such as Multivariate Linear Regression (MLR), Decision Tree Regressor, Random Forest Regressor, XGBoost Regressor and our proposed Stacked LSTM RNN model for forecasting rainfall. Daily rainfall data for almost 40 years are taken as input for this study and comparison of the above mentioned models have been carried out with respect to the Mean Absolute Error (MAE), Coefficient of Regression (R^2) and the Mean Squared Error (MSE). The LSTM model emerged as the top performer, displaying the lowest MAE score. Hence, it was chosen to forecast daily rainfall for 7 days, weekly rainfall for 4 weeks and monthly rainfall for 12 consecutive months. Notably, the model's accuracy improves as the duration of recorded observations are increased. The key novelty presented in this work is the proposed stacked LSTM Model and the comparative study of the model to predict rainfall in daily, weekly and monthly format. Our study prominently underscores the effectiveness of the proposed stacked LSTM Model in accurately forecasting rainfall.

***Corresponding Author:** Satyajit Sarmah ✉ ss@gauhati.ac.in

Introduction

The North-East region of India receives abundant rainfall due to its close proximity to Bay of Bengal, its topography and its rich biodiversity. There are four seasons based on rainfall patterns: Southwest Monsoon (Jun-Sep), Post Monsoon Season (Oct-Nov), Winter Season (Dec-Feb) and the Pre Monsoon Season (Mar-May). Certain areas in this region experience rainfall throughout the year due to its local topography and lush greenery. Therefore, accurate prediction of rainfall can be beneficial from the different aspects including the socio-economic condition of the region. Rainfall prediction tasks can not only benefit farmers to plan their cropping patterns, but also benefit the government to make water management strategies, flood and landslide prevention policies, etc. Hydroelectric plants can optimize their performance with accurate rainfall predictions and many businesses and people in general can plan their day to day activities more effectively with accurate predictions.

In (Mujumdar *et al.*, 2023), Sounak Majumdar and their team, mentioned clearly in their study that the accurate prediction of rainfall is one of the most challenging tasks in meteorology. Here, the stacked LSTM model is employed for rainfall prediction of Silchar city of North-East India where Boruta feature selection is used to increase the efficiency. The proposed model when compared to RF, MLR and XGBoost, LSTM models achieve better results. RMSE and R-squared values are used for performance measurement in the study.

Chalachew Muluken Liyew and their team in their study (Liyew *et al.*, 2021), predicted the intensity of rainfall using MLR, RF, and XGBoost Algorithms. The Pearson correlation strategy was used to identify relevant environmental parameters used on the dataset which was gathered from the Meteorological office of Bahir Dar City, Ethiopia. Based on RMSE and MAE scores, XGBoost Algorithm outperformed others.

In (Salehin *et al.*, 2020), Imrus Salehin and their team proposed a method for rainfall prediction where it was determined using artificial intelligence and LSTM techniques. Here, they have considered 6 parameters for their model, namely - temperature, dew point, humidity, wind pressure, wind speed and wind direction. The dataset was collected from the Bangladesh Meteorological Department (BMD). Here they have taken only one month's data for the year 2020 from 1st August to 31st August and used it for training the model. 76% accuracy was obtained in their work.

Many studies have been conducted to predict rainfall by different researchers. Data mining techniques are used to extract patterns and relationships with the various environmental factors that affect rainfall. Different machine learning algorithms have been used by the researchers for prediction of rainfall using some standard dataset. Machine learning techniques like MLR, XGBoost, Random Forest regressor, SVM, etc. show good accuracy in predicting rainfall using historical data. MLR is applicable for predictive analysis which forecast or predict the rainfall or weather, predict trends in business, finance etc (Gomathy *et al.*, 2021). Decision Tree Regressor is a non-linear regression technique that works for both numerical as well as categorical output variables. It is generally employed when the relationship between the dependent and independent variables is very complex (https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/) (Angela and Shi, 2023). RF Regression Algorithm combines the basic principles of linear regression, with the addition of multiple decision trees and obtains the aggregation of all their predictions which it considers as the final prediction(Jonsson and Fredrikson, 2021). XGBoost is a specific implementation of the Gradient Boosting method which uses more accurate approximations to find the best tree model. It is implemented for the supervised machine learning problem that has data with multiple features. Most authors use XGBoost for different regression and classification problems due to the speed and prediction accuracy of the algorithm (Browniee, 2021).

Artificial Neural Networks, more specifically, RNN's are able to capture the temporal characteristics of rainfall and other features through models like ARIMA, SARIMA and LSTM (https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/). An LSTM model can be used to forecast time series predictions even when the features or independent variables for that particular period of prediction are not available (https://www.predicthq.com/events/lstm-time-series-forecasting). LSTM is capable of capturing both long term seasonality such as a yearly pattern and short term seasonality such as weekly patterns.

In the majority of the studies conducted, the environmental features considered were Temperature, Wind Direction, Wind Speed, Sunlight, Humidity and Pressure. In our work also, we have selected the same features. The main motivation of this work is to study the existing rainfall prediction method using different ML approach, analyze their performances with respect to some parameters and propose a model for accurate prediction of rainfall in the north-east region of India. Accuracy is one of the key parameters for performance evaluation of the models.

The main contributions of the paper are as follows.

First, propose and implement a stacked LSTM model to predict rainfall by analysing different atmospheric features and their relationship with rainfall.

Second, compare the performance using different ML techniques with our proposed Stacked LSTM Model.

Material and methods

Proposed method

To predict the rainfall of north east region of India, a standard data set of NASA Prediction of World Wide Energy Resources are taken and this dataset has been used for implementing our proposed model. The proposed method is discussed below.

To build the stacked LSTM Model, a sequential model with two LSTM layers with dropout regularisation is created. The first LSTM layer has 64 units and the second LSTM layer has 32 units, with each layer having a dropout rate set to 10 percent and uses ReLU activation function. The final layer is a Dense layer having the same number of units as the target values. The model is compiled using the Adam Optimizer which adaptively adjusts the learning rate during training and the loss during each epoch is calculated in MSE. The model is trained for 1000 epochs with early stopping criteria and the best weights are stored. The Fig. 1 shows the proposed model.

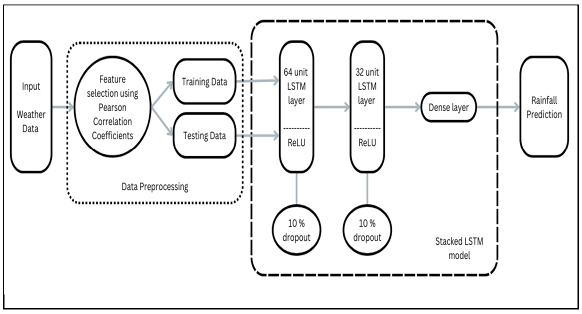

Fig. 1. Schematic Diagram for the proposed Stacked LSTM Model

Performance Measure: The metrics used for evaluating the models are Mean Squared Error, Mean Absolute Error, Root Mean Squared Error and R2 (coefficient of regression) (Chugh, 2020).

MSE- Mean Squared Error represents the average of the squared difference between the actual and predicted values of the dataset. It measures the variance of the residuals.

$$MSE = (1 \div N)\sum_{i=1}^{N} |y_i - \hat{y}|^2$$

Where, N → total number of observations

$\hat{y}$ → predicted value of y

MAE- The average of the absolute difference between the actual and the predicted values in the dataset is represented by the Mean Absolute Error which is commonly known as MAE. It measures the average of the residuals in the dataset.

$$MAE = (1 \div N)(\sum_{i=1}^{N} |y_i - \hat{y}|)$$

Where, N → total number of observations

$\hat{y}$ → predicted value of y

RMSE- It represents the square root of Mean Squared Error. It has the capacity to measure the standard deviation of residuals.

$$RMSE = \sqrt{MSE} = \sqrt{(1 \div N) \sum_{i=1}^{N} (y_i - \hat{y})^2}$$

Coefficient of determination R²- It measures how well the model is being able to predict the target variable. Its value ranges from 0 to 1, where 0 means that the target variable cannot be predicted and 1 means that the target variable can be exactly predicted by the model.

$$R^2 = 1 - \frac{\Sigma (y_i - \hat{y})^2}{\Sigma (y_i - \underline{y})^2}$$

Where $\underline{y}$ → mean value of y

$\hat{y}$ → predicted value of y

Results and discussion

After analyzing the data, the Pearson Correlation Coefficients |r| of all the features are found. The available features and the value of |r| in the dataset are as follows.

Table 1. Pearson correlation

Features	|r|
Year	0.032
Month	0.037
Day	0.027
Temperature	0.419
Dew/Frost	0.465
Specific Humidity	0.481
Relative Humidity	0.407
Precipitation	1.000
Surface Pressure	0.117
Wind Speed	0.176

The features selected after analyzing the correlation coefficients in Table 1 are Temperature, Dew/Frost Point Temperature, Surface Humidity, Relative Humidity, Surface Pressure, Precipitation and Wind Speed. Features having | r | < 0.1 are not considered further. We have implemented different ML models on the dataset. We also implemented our proposed stacked LSTM Model and calculated the MAE, MSE, RMSE and R2 values (Table 2). The obtained scores in the following table, gives us a comparison of the performance of all the ML and RNN models using present day's features to predict next day's rainfall using MAE, MSE, RMSE and R2 as performance measures.

Model selection for this work was done based on the MAE score and the R² score. Therefore, in this work, the proposed Stacked LSTM model has the best accuracy to predict next day's rainfall.

Table 2. Comparison of the models for next day's rainfall prediction using current day's feature values.

Algorithms	MAE	MSE	RMSE	R²
Linear	2.61	19.46	4.41	0.257
Decision tree	2.91	30.29	5.50	0.155
Random forest	2.32	17.57	4.19	0.329
Xgboost	2.34	18.38	4.28	0.29
Lstm	2.30	17.83	4.22	0.434

A demonstration of the LSTM model preparing to generate a one-day prediction by analysing the previous day is shown in Fig. 2.

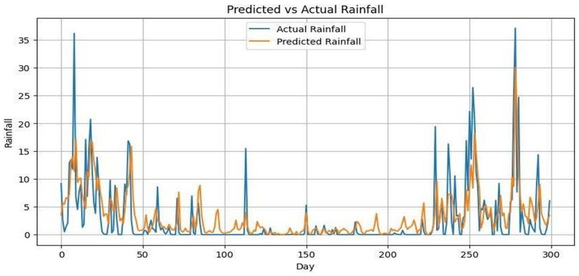

Fig. 2. Rainfall prediction of one-day vs actual rainfall of one day

In the above figure, the actual and predicted rainfall is plotted for the first 300 entries from the test set where rainfall is measured in millimeters. Daily rainfall data has a lot of outliers but the LSTM model captures the basic trend of daily rainfall in a fairly accurate manner.A demonstration of the LSTM model preparing to generate a 7-day prediction by analysing the 30 previous days is shown in Fig. 3.

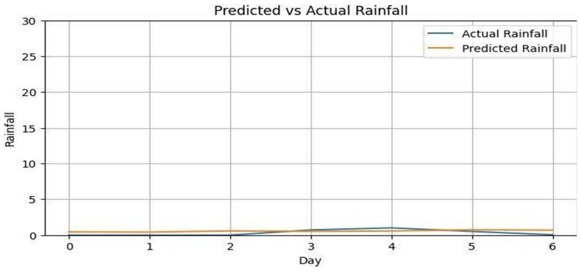

Fig. 3. Rainfall prediction of seven days vs actual rainfall of seven days.

Here, in the figure, the actual and the predicted rainfall are plotted for the first 7 days of the test set where rainfall is measured in millimeters. The LSTM model predictions are very close to the actual observations.

A demonstration of the LSTM model preparing to generate a 4-week prediction by analysing the 12 previous weeks is shown as Fig. 4.

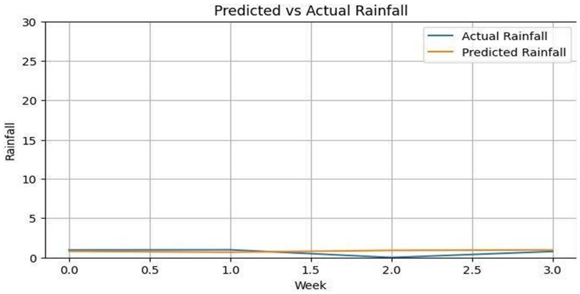

Fig. 4. Rainfall prediction of 4 weeks vs actual rainfall of 4 weeks

In the above figure, the actual and the predicted rainfall is plotted for the first 4 weeks of the test set where rainfall is measured in millimeters. The proposed LSTM model is able to capture the weekly trend of rainfall very accurately.

A demonstration of the LSTM model preparing to generate 12-month prediction by analyzing the 24 previous months are shown in Fig. 5.

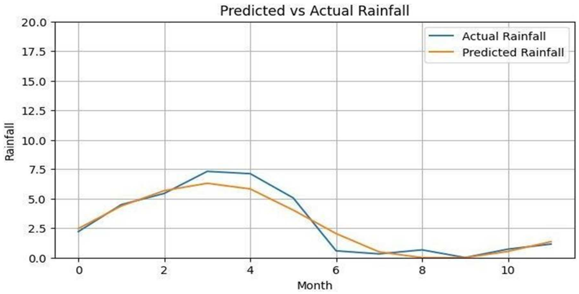

Fig. 5. Rainfall prediction of 12 months vs actual rainfall for those 12 months

Here, in the above figure, the actual and the predicted rainfall is plotted for the first 12 months from the test set. The model shows highest accuracy in predicting monthly rainfall.

Table 3. The evaluation metrics used for the model and their scores

SL	Models	MAE	MSE	RMSE	R^2
1.	LSTM (1 day prediction)	0.47	0.69	0.83	0.43
2.	LSTM (7 days prediction)	0.54	0.97	0.98	0.23
3.	LSTM (4 week prediction)	0.49	0.64	0.80	0.7
4.	LSTM (12 months prediction)	0.37	0.29	0.54	0.75

The performance of this proposed LSTM used to predict 7 future days, 4 future weeks and 12 future months is given in Table 3. The performance of the LSTM models improves as we take the weekly and monthly mean of the observations.

Table 4 gives a detailed comparison of the proposed stacked LSTM model with other rainfall prediction models in use. The comparison is made on the dataset, models, features and the performance measures used in the prediction process along with the findings of the existing work. From the comparison it is seen that for real-time rainfall prediction, the Stacked LSTM model has least MAE of 2.30 and highest R2 of 0.43. MAE of 7-day, 4-week and 12-month predictions was 0.54, 0.49 and 0.37 respectively.

Upon comparison of the findings of other research papers on rainfall prediction, it is seen that RNN models like various LSTM models including Stacked LSTM, Bidirectional LSTM and Intensified LSTM have outperformed ML Regression models. Utilizing the strengths of multiple models to create a hybrid model can capture the complex rainfall patterns more accurately. The proposed paper additionally provides a comparison between the performance of the Stacked LSTM models for 7-day, 4-week and 12-month predictions, where the model shows better accuracy as the corresponding time period of observations increases.

Table 4. Comparison of our proposed model with the existing models

Source	Dataset used	Models used	Features used	Performance measures	Findings
Liyew *et al.*, 2021	Data from the regional eteorological station at Bahir Dar City, Ethiopia for a period of 20 years from 1999 to 2018.	MLR, RF regressor and XGBoost gradient descent.	Evaporation, Relative Humidity, Sunshine, Maximum Daily Temperature, and Minimum Daily Temperature.	MAE and RMS E	MAE of RF, MLF and XGBoost were 4.49, 4.97 and 3.58 respectively. RMSE of RF, MLF and XGBoost were 8.82, 8.61 and 7.85 respectively.
Salehin *et al.*, 2020	Raw data of Dhaka City from 2000-2014 collected from Bangladesh Meteorological Department(B DM).	LSTM and Recurrent Neural Network (RNN)	Temperature, Dew Point, Humidity, Wind Pressure, Wind Speed, and Wind Direction.	R^2	After analysing all the data using LSTM and RNN , 76% accuracy was found in predicting monthly rainfall.
Gomathy *et al.*, 2021	Dataset consists of the measurement of rainfall from 1901-2015 for each state.	MLR, Support Vector Regression, Lasso Regression.	19 attributes (individual months, annual, and combinations of 3 consecutive months) for 36 sub divisions.	R^2 and MAE	While comparing the errors, it is found that the errors from the SVR model was the least and that of the Lasso model was the highest.
Poornima and Pushpalatha, 2019.	Rainfall data of Hyderabad region starting from 1980-2014 is the dataset considered for the prediction process.	Intensified LSTM, LSTM, Holt–Winter, ARIMA, ELM, and RNN	Maximum Temperature, Minimum Temperature, Minimum Relative Humidity, Wind Speed, Sunshine and Evapotranspiration.	RMSE	The RMSE of the proposed model that is intensified LSTM was found to be the RMSE of the proposed model that is intensified LSTM was found to be.
Chhetri *et al.*, 2020	Rainfall data was collected from National Center of Hydrology and Meteorology Department (NCHM) of Bhutan	Linear Regression, MLP, CNN, LSTM, Gated Recurrent Unit (GRU), BLSTM	Maximum Temperature, Minimum Temperature, Rainfall, Relative Humidity,	MSE and RMSE	MSE and RMSE were 0.0075 and 0.087 of the proposed BLSTM GRU model which outperformed all

	from 1997-2017.		Sunshine and Wind Speed.		the other models.
Our proposed Method	Atmospheric Dataset from The Power Project, NASA for the entire NE region of India from Feb, 1981 -Feb 2021	MLR, DT, RF regressor, XGBoost regressor, and Stacked LSTM model.	Temperature, Dew/Frost Point Temp, Surface Humidity, Relative Humidity, Surface Pressure, Precipitation and Wind Speed.	MAE and R^2	For Real-time rainfall prediction, the Stacked LSTM model had least MAE of 2.30 and highest R^2 of 0.43. MAE of 7-day, 4-week and 12-month predictions were 0.54, 0.49 and 0.37 respectively.

Conclusion

Rainfall prediction is a difficult task which depends on many atmospheric conditions like temperature, pressure systems, topography, forest cover, humidity, wind patterns. Accurate prediction of rainfall can provide various benefits to the people of a particular region to plan their socio-economic activities, disaster management, agriculture and many other related activities. The region of study we have taken here is the North East region of India for which 40 years of daily historical data was collected from NASA Prediction of World Wide Energy Resources. ML algorithms like MLR, RF and XGBoost, and a Stacked LSTM RNN model were analyzed on their performance to predict rainfall using real time temperature, dew/frost point temperature, relative humidity, specific humidity, surface pressure, wind speed data as features.

The LSTM model showed better performance with the lowest MAE score, thus it was chosen to forecast daily rainfall for 7 days, weekly rainfall for 4 weeks and monthly rainfall for 12 consecutive months. Notably, the model's accuracy improves as the duration of recorded observations are extended. While considering daily observations, the high fluctuation of rainfall increases the complexity of the prediction task. But when we consider monthly observations, the result reflects higher accuracy compared to weekly and daily observations. In the proposed LSTM models, the difference of the mean of predictions and the mean of actual observations was less than 1 mm.

References

Majumdar S, Biswas SK, Purkayastha B, Sanyal S. 2023. Rainfall Forecasting for Silchar City using Stacked- LSTM. 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON), Jaipur, India. pp. 1-**5**.
DOI: 10.1109/IEMECON56962.2023.10092355.

Liyew CM, Melese HA. 2021. Machine Learning Techniques to Predict Daily Rainfall Amount. Journal of big Data **8**, 153.
https://doi.org/10.1186/s40537-021-00545-4

Salehin I, Talha IM, Hasan MM, Dip ST, Saifuzzaman M, Moon NN. 2020. An Artificial Intelligence Based Rainfall Prediction Using LSTM and Neural Network. IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), Bhubaneswar, India. pp. **5-8**.
DOI: 10.1109/WIECON-ECE52138.2020.9398022

Gomathy CK., Reddy ABN, Kumar AP, Lokesh A. 2021. A Study Of Rainfall Prediction Techniques. International Journal of Scientific Research in Engineering and Management **5(10)**, 1-15.

Python Decision Tree Regression using sklearn. Available from: https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/

Jonsson E., Fredrikson S. 2021. An Investigation Of How Well Random Forest Regression Can Predict Demand.

Browniee J. 2021. XGBoost for Regression. Available from: https://machinelearningmastery.com/xgboost-for-regression

The Value of LSTM in Time Series Forecasting. Available from: https://www.predicthq.com/events/lstm-time-series-forecasting

Angela, Shi K. 2023. Decision Tree Regressor- A Visual Guide with Scikit Lear. Towards Data Science.

Chugh A. 2020. MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared-Which metric is better. Analytics Vidhya.

Introduction to Recurrent Neural Networks. Available from: https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network

Poornima S, Pushpalatha M. 2019. Prediction of Rainfall Using Intensified LSTM Based Recurrent Neural Network with Weighted Linear Units. Atmosphere **10(11)**, 668. https://doi.org/10.3390/atmos10110668

Chhetri M., Kumar S, Roy PP, Kim B-G. 2020. Deep BLSTM-GRU Model for Monthly Rainfall Prediction: A Case Study of Simtokha, Bhutan. Remote Sensing **12(9)**, 3174. https://doi.org/10.3390/rs12193174