# Bangladesh University of Engineering and Technology

---

# Rainfall Prediction in Bangladesh: A Deep Learning Approach

---

CSE-472 : Machine Learning Sessional Project


**AUTHORS**

Tanveer Rahman - 1905025
Mohaiminul Islam - 1905018

# Contents

# 1 Introduction

Bangladesh is an agricultural country where rainfall prediction plays a crucial role due to the heavy reliance of its population on agriculture. Accurate rainfall forecasting is a challenging and complex task, given the dynamic and unpredictable nature of rainfall patterns in the region.

In this study, we focus on monthly rainfall prediction across 35 weather stations in Bangladesh collected from 1948-2022. The rainfall data was collected from various online sources and carefully processed to create a comprehensive weather dataset. To perform the prediction, we implemented a range of machine learning models, including Linear Regression and Random Forest, as well as several deep learning models.

Additionally, this project introduces a novel GRU-Attention with XGBoost architecture, which demonstrates superior performance compared to the other models evaluated.

# 2 Readings

For this project, we reviewed several research papers to gain a deeper understanding of rainfall prediction techniques and the challenges associated with them. Usman et al. [1] provided a comparative analysis of various machine learning models for rainfall prediction in the Aligarh district of India. Similarly, De Nuno et al.[2] demonstrated the use of a hybrid model, M5P-SVR, for precipitation prediction in the Northern region of Bangladesh, showcasing its effectiveness in capturing localized rainfall patterns.

Among these, a particularly notable study focused on predicting monthly rainfall in two Australian cities - Darwin and Perth. This study utilized a dataset comprising 100 years of monthly rainfall data (1921 - 2020) to evaluate the efficacy of advanced deep learning approaches [3]. The authors proposed a deep learning model based on an **Encoder-Decoder with Attention** mechanism to address the inherent complexities and non-linear nature of rainfall patterns. The attention mechanism allowed the model to focus on significant temporal patterns in the data, thereby improving its prediction accuracy.

Inspired by this methodology, we explored its applicability to the context of Bangladesh, where rainfall prediction is both critical and challenging due to its dynamic, regionally varied climate. Leveraging these insights, we developed a model tailored to predict rainfall in Bangladesh, aiming to provide reliable and actionable forecasts.

# 3 Dataset Preparation

Preparing the dataset was one of the most challenging aspects of this project, as rainfall data for Bangladesh is not readily available in a centralized source. To address this, we adopted a multi-step process to compile a comprehensive dataset.

Initially, we collected a dataset from Kaggle, which contained weather data for various stations in Bangladesh spanning the years 1948 to 2013. To extend this dataset, we sourced additional data from the official website of the "Bangladesh Agricultural Research Council", covering rainfall records for various stations up to 2022.

After obtaining these datasets, we concatenated them to form a unified dataset. This required handling inconsistencies in data formats, resolving missing values, and ensuring that the records were accurately aligned across time and stations. The resulting dataset provided a robust foundation for the analysis and modeling phases of the project. The preview of the dataset is provided in fig: 1.

| | Station | Year | Month | Max Temp | Min Temp | Rainfall(mm) | Humidity (percent) | Wind Speed (m/s) | Cloud Coverage (Octs) | Sunshine (Hours) | Station Number |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Barisal | 1949 | 1 | 29.40 | 12.30 | 0.0 | 68.00 | 0.453704 | 0.60 | 7.831915 | 41950 |
| **1** | Barisal | 1949 | 2 | 33.90 | 15.20 | 9.0 | 63.00 | 0.659259 | 0.90 | 8.314894 | 41950 |
| **2** | Barisal | 1949 | 3 | 36.70 | 20.20 | 8.0 | 59.00 | 1.085185 | 1.50 | 8.131915 | 41950 |
| **3** | Barisal | 1949 | 4 | 33.90 | 23.90 | 140.0 | 71.00 | 1.772222 | 3.90 | 8.219149 | 41950 |
| **4** | Barisal | 1949 | 5 | 35.60 | 25.00 | 217.0 | 76.00 | 1.703704 | 4.10 | 7.046809 | 41950 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **24863** | Teknaf | 2022 | 8 | 31.75 | 25.49 | 529.0 | 86.16 | 1.580000 | 6.29 | 5.450000 | 41998 |
| **24864** | Teknaf | 2022 | 9 | 31.51 | 25.84 | 443.0 | 86.33 | 1.530000 | 6.33 | 4.400000 | 41998 |
| **24865** | Teknaf | 2022 | 10 | 32.40 | 25.55 | 174.0 | 83.23 | 1.970000 | 2.77 | 7.680000 | 41998 |
| **24866** | Teknaf | 2022 | 11 | 31.58 | 20.63 | 27.0 | 77.37 | 1.530000 | 0.33 | 9.290000 | 41998 |
| **24867** | Teknaf | 2022 | 12 | 30.21 | 18.13 | 0.0 | 75.32 | 1.980000 | 0.23 | 7.850000 | 41998 |

24868 rows × 11 columns

Figure 1: Preview of the created Dataset

# 4 Dataset Properties

The correlation matrix is given in figure 2

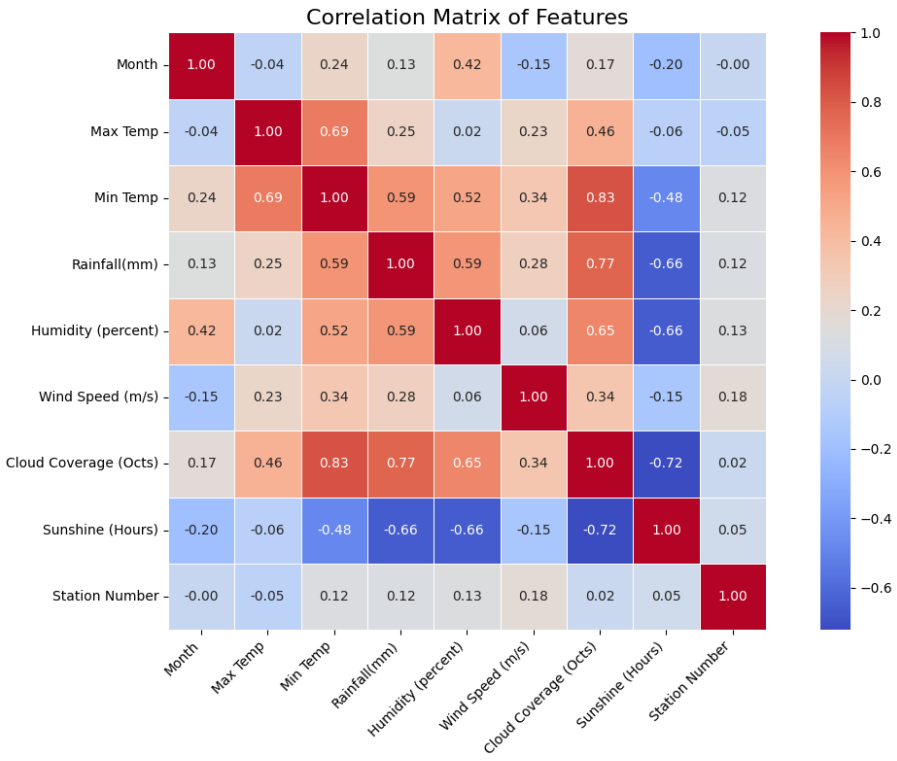


Figure 2: Correlation Matrix

Relationship of various features with Rainfall is shown in figure 3

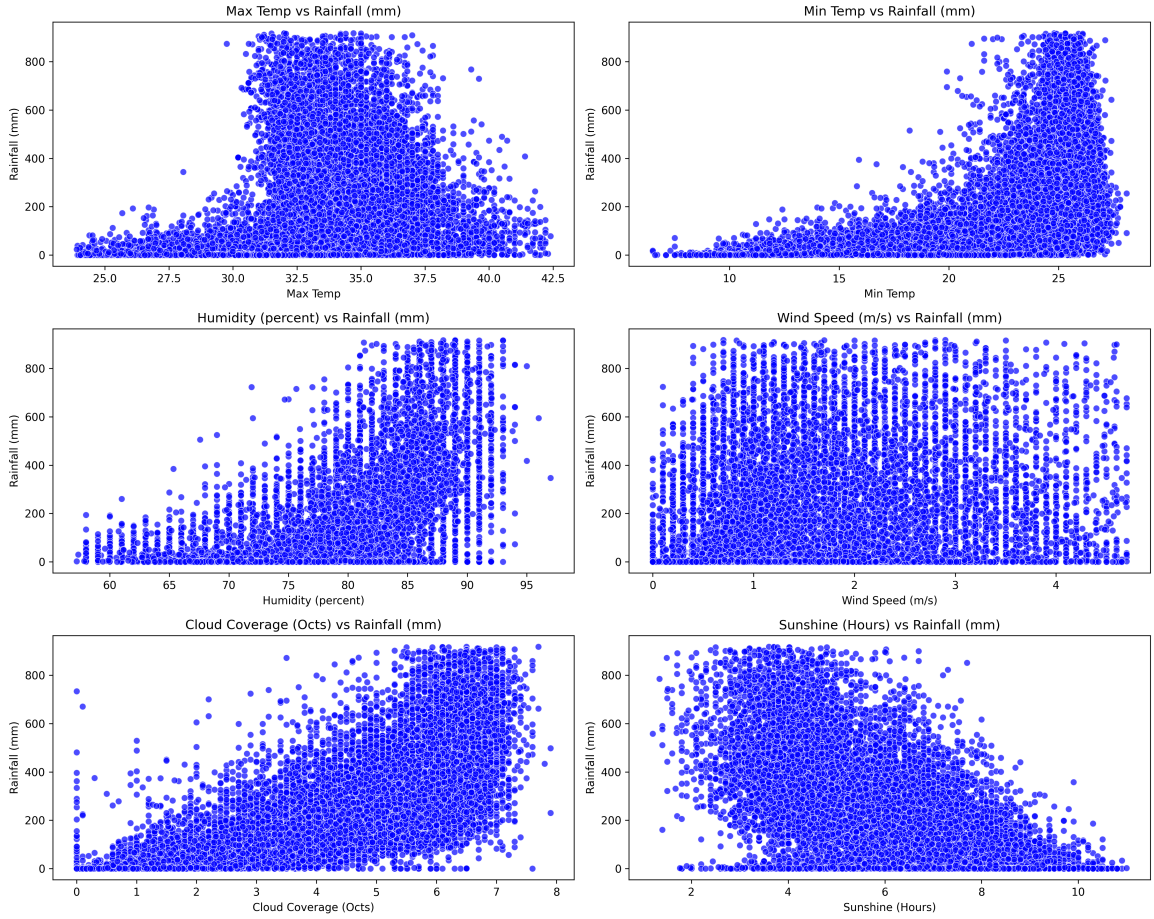


Figure 3: Relationship of rainfall with other features

# 5 Methodology

We employed several preprocessing, modeling, and evaluation steps to predict monthly rainfall accurately.

## 5.1 Data Preprocessing

To ensure all input features were on the same scale, we applied the *StandardScaler* to standardize the features. However, the target variable **Rainfall (mm)** was left unscaled to maintain interpretability.

The dataset was then split into training and testing sets in an 80:20 ratio, resulting in the shapes shown in Table 1.

Table 1: Test and Train Dataset Shape

| Type | Shape |
|---|---|
| Test Data | (18944, 8) |
| Train Data | (4736, 8) |

## 5.2 Modelling Approach

We experimented with multiple machine learning and deep learning models with our dataset. The output of such experiments are shown below:

- **Multiple Linear Regression Model** (fig-4)

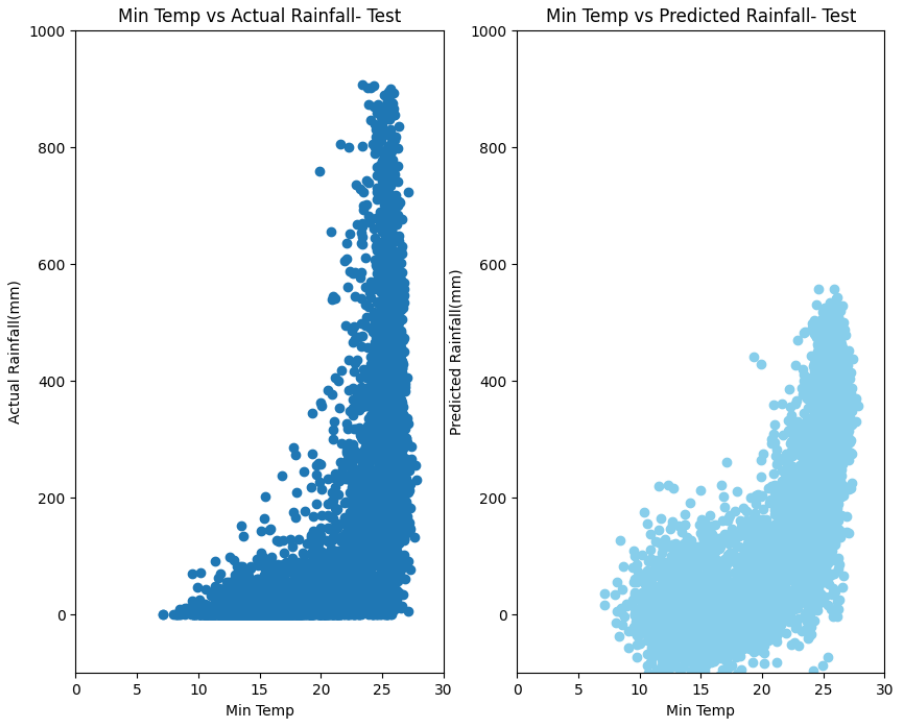


Figure 4: Prediction on the Test data using MLR

- **Polynomial Regression Model** (figure - 5)

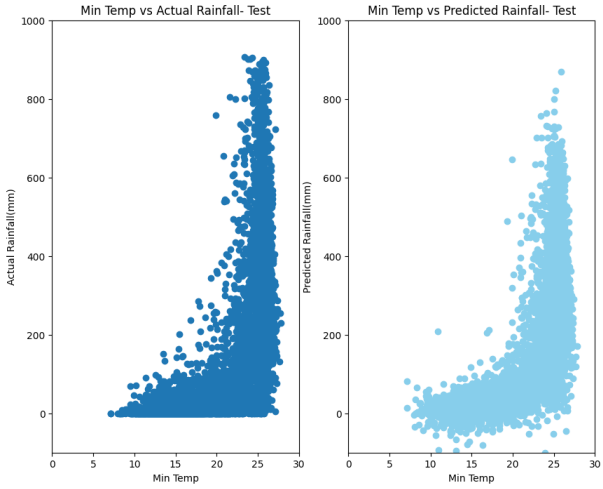


Figure 5: Prediction on the Test data using PLR

- **Decision Tree Model** (figure - 6)

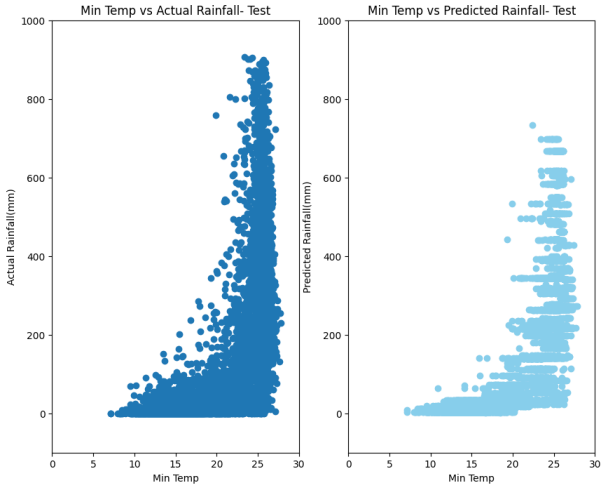


Figure 6: Prediction on the Test data using DT

- **k-nearest neighbors Model** (figure - 7)

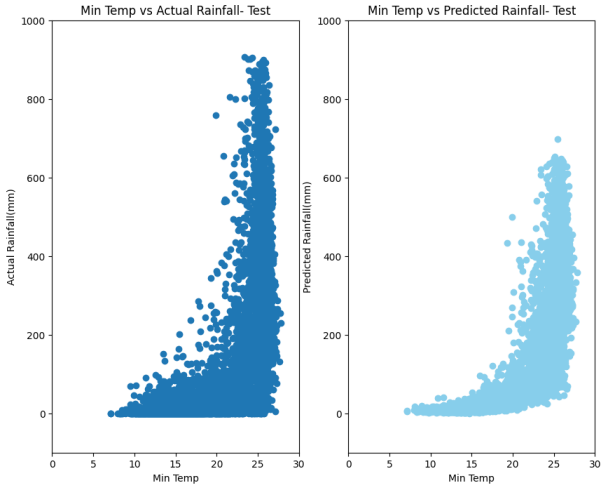


Figure 7: Prediction on the Test data using KNN

- **Support vector machine (SVM) Model** (figure - 8)

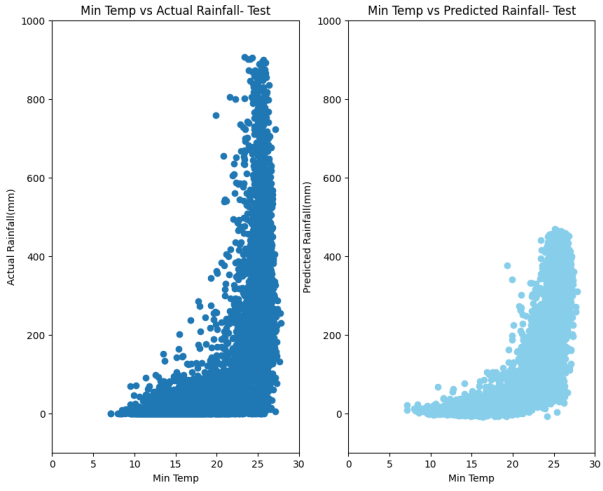


Figure 8: Prediction on the Test data using SVM

- **Random Forest Model** (figure - 9)

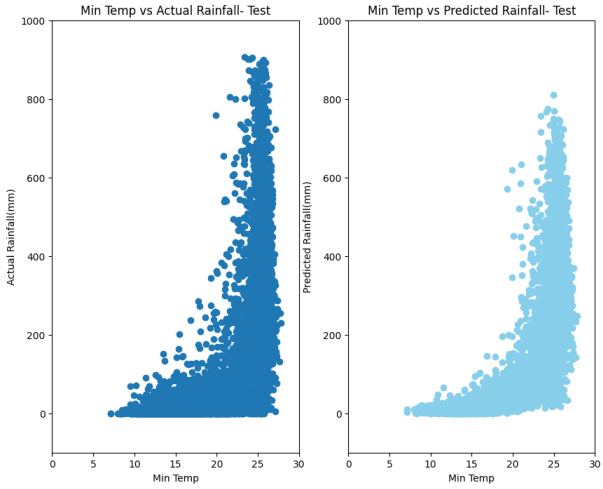


Figure 9: Prediction on the Test data using RF

- **AdaBoost Regressor Model** (figure - 10)

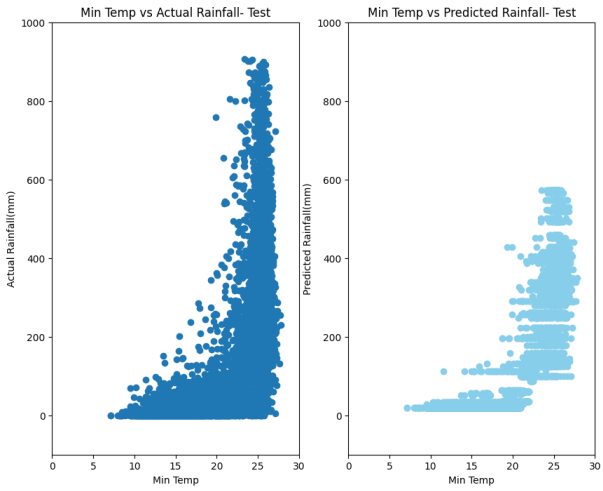


Figure 10: Prediction on the Test data using AdaBoost

- **Stacking Regressor Model** (figure - 11)

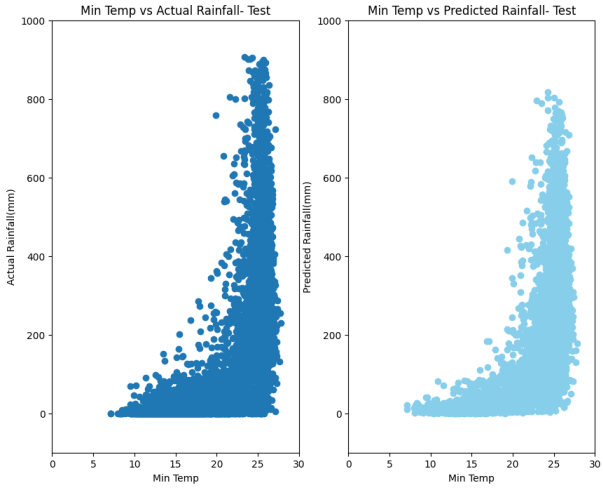


Figure 11: Prediction on the Test data using SR

- **Artificial Neural Network Model** (figure - 12)

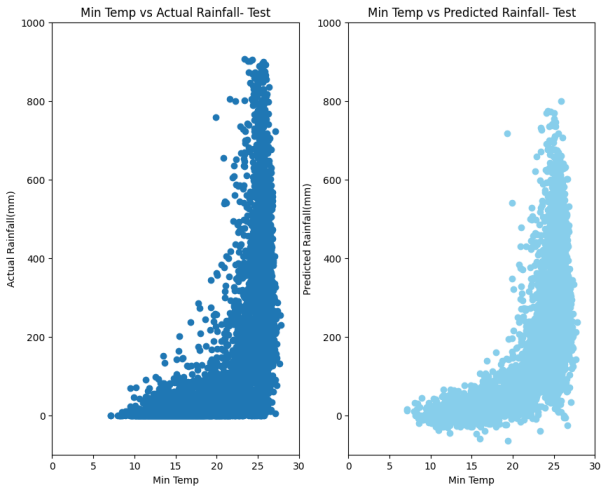


Figure 12: Prediction on the Test data using ANN

- **GRU with Attention + XGBoost Model** (figure - 13)

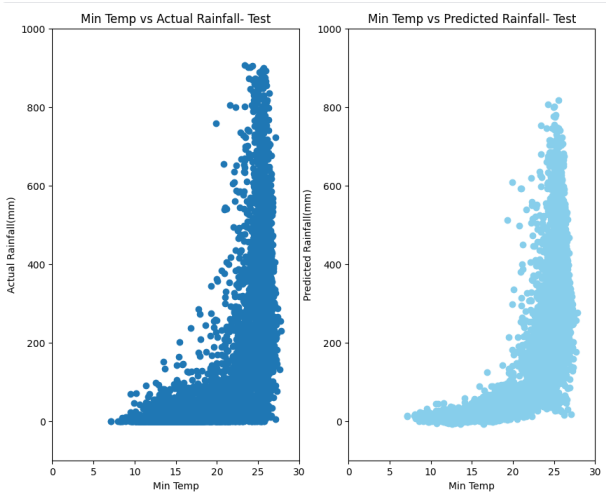


Figure 13: Prediction on the Test data using GRU and XGBoost

- **LSTM with Attention and XGBOOST Model** (figure - 14)

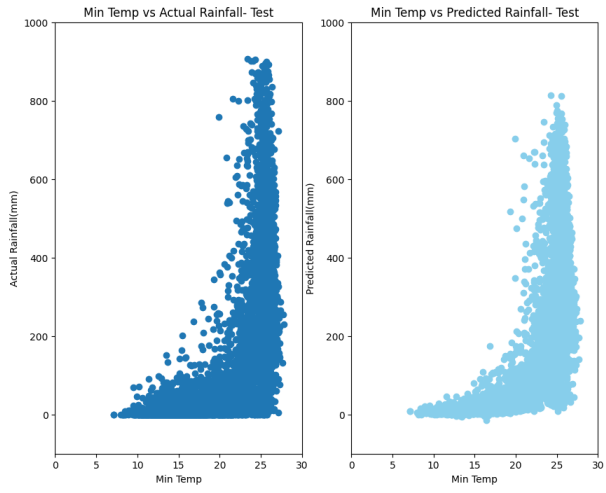


Figure 14: Prediction on the Test data using LSTM

# 6 Compilation of Results

For the above models, the compiled results are shown in the table 2

Table 2: Performance of Different Machine Learning Models

| ML Model | $R^2$ Score | RMSE Score |
|---|---|---|
| Multiple Linear regression | 0.676096 | 117.853209 |
| Polynomial regression | 0.779087 | 97.329476 |
| Decision Tree model | 0.737548 | 106.085994 |
| k-nearest neighbors | 0.751117 | 103.307324 |
| Support vector machine (SVM) | 0.670801 | 118.812651 |
| Random Forest model | 0.780543 | 97.008062 |
| AdaBoost Regressor | 0.718801 | 109.809502 |
| Stacking Regressor | 0.756089 | 102.270203 |
| Artificial Neural Network | 0.778553 | 97.446908 |
| GRU with Attention and XGBoost | 0.795308 | 93.687942 |
| LSTM with Attention and XGBoost | 0.8033 | 65.25 |

The **LSTM with Attention and XGBoost model** achieves the **highest $R^2$ score** of $0.8033$ and the **Lowest RMSE score** of $65.25$, indicating it performs slightly better than the GRU with Attention and XGBoost model, which also shows strong results with an $R^2$ score of $0.7953$ and an RMSE score of $93.6879$.
Deep Learning models (GRU/LSTM with Attention and XGBoost) consistently outperform the more traditional machine learning models in terms of both R2 score and RMSE.

# 7 Benchmark Comparison

In the paper under consideration [3], a different dataset is utilized. Specifically, two separate datasets are used, each containing 1200 data points, corresponding to two cities: Darwin and Perth.

Table 3: Performance Comparison of the Encoder-Decoder with Attention Mechanism

| Dataset | Model from Paper ($R^2$) | Our Model ($R^2$) |
|---|---|---|
| Darwin | 0.832 | 0.805 |
| Perth | 0.796 | 0.753 |
| Bangladesh | 0.7094 | 0.8033 |

### 7.1 Verdict:

The results clearly indicate that our proposed model is competitive with the Encoder-Decoder with Attention Mechanism presented in [3]. While the original model achieves slightly higher $R^2$ values on the Darwin ($0.832$) and Perth ($0.796$) datasets, our model achieves comparable results with $R^2$ values of $0.805$ and $0.753$, respectively.
Moreover, our model significantly outperforms the original model on the Bangladeshi dataset, achieving an $R^2$ value of $0.803$, compared to the original model's $0.7094$. This demonstrates that our model not only performs well on the original datasets but also generalizes better to unseen data, making it a robust and reliable alternative.

## 8 Conclusion

In this project, we proposed an **LSTM model** with an **Attention** Mechanism, combined with **XGBoost**, to predict rainfall at various stations in Bangladesh using the provided dataset. The primary objective was to evaluate the effectiveness of our proposed model in comparison to the benchmark Encoder-Decoder with Attention Mechanism model from the literature.

Our results demonstrate that the proposed model is competitive with the benchmark model, achieving superior performance on the Bangladeshi dataset. This highlights the adaptability and effectiveness of our approach in handling rainfall prediction tasks for Bangladesh.

Future work could focus on integrating additional features (e.g., external climate indices or satellite data), optimizing hyperparameters to further enhance performance, or extending the application of this framework to other datasets and regions.

# List of Figures

# List of Tables

# References

[1] M. Usman Saeed Khan, K. Mohammad Saifullah, A. Hussain, and H. Mohammad Azamathulla, "Comparative analysis of different rainfall prediction models: A case study of aligarh city, india," *Results in Engineering*, vol. 22, p. 102093, 2024.

[2] F. Di Nunno, F. Granata, Q. B. Pham, and G. de Marinis, "Precipitation forecasting in northern bangladesh using a hybrid machine learning model," *Sustainability*, vol. 14, no. 5, 2022.

[3] R. He, L. Zhang, and A. W. Z. Chew, "Data-driven multi-step prediction and analysis of monthly rainfall using explainable deep learning," *Expert Systems with Applications*, vol. 235, p. 121160, 2024.