Rainfall Prediction using Machine Learning Models from Meteorological Data

Krittiya Sunahu, Chanyaphas Chaikongcha, Suphawan Srimakorn, Kunnaree Wongsa and Thitikarn Techa
Associate Professor Dr. Thaned Rojsiraphisa and Assistant Professor Dr. Nawinda Chutsagulprom
Data Science Research Center, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand

Abstract

Rainfall prediction plays a crucial role in water management and agricultural planning in the northeastern region, that is influenced by various climatic factors. This study aims to develop rainfall prediction models using machine learning techniques and compare the performance of different models, namely XGBoost, long short-term memory (LSTM), and transformer. The meteorological data used in this prediction include rainfall, min temperature, max surface temperature, wind north, max wind direction, max wind speed, humidity, dry bulb temperature, and evaporation spanning from 1993 to 2018 obtained from Thailand Meteorological Department. The evaluation metrics for model performance include mean absolute error (MAE), root mean square error (RMSE) and R² score. By varying under specific parameter settings, the experimental findings showed that the transformer model outperformed the other two models, whereas the LSTM approach yielded unsatisfactory results. It is widely recognized that hyperparameter tuning significantly impacts analysis performance, often in a detrimental way. Future work could explore optimization techniques to determine optimal hyperparameters.

Introduction

Northeastern Thailand covers one third of the country with diverse geographical features. This region relies heavily on agriculture [1]. Rainfall significantly impacts agriculture, water resources, and agricultural planning. Understanding both expected rainfall amounts and the climatic variability influencing precipitation is crucial, especially when given the challenges of effective rainfall planning.

Methodology

This study employs a structured methodology for rainfall prediction in the northeastern region. Data understanding and preprocessing involve analyzing relationships, handling missing values using interpolation methods (last observation carried forward: LOCF, linear interpolation), and ensuring data consistency. Linear regression analysis aids in feature selection by identifying key variables using linear regression. Three models—XGBoost, LSTM, and transformer—are trained using Python (Google Colab) with hyperparameter tuning. Finally, model performance is evaluated using MAE, RMSE, and R² score as shown in Fig. 1.

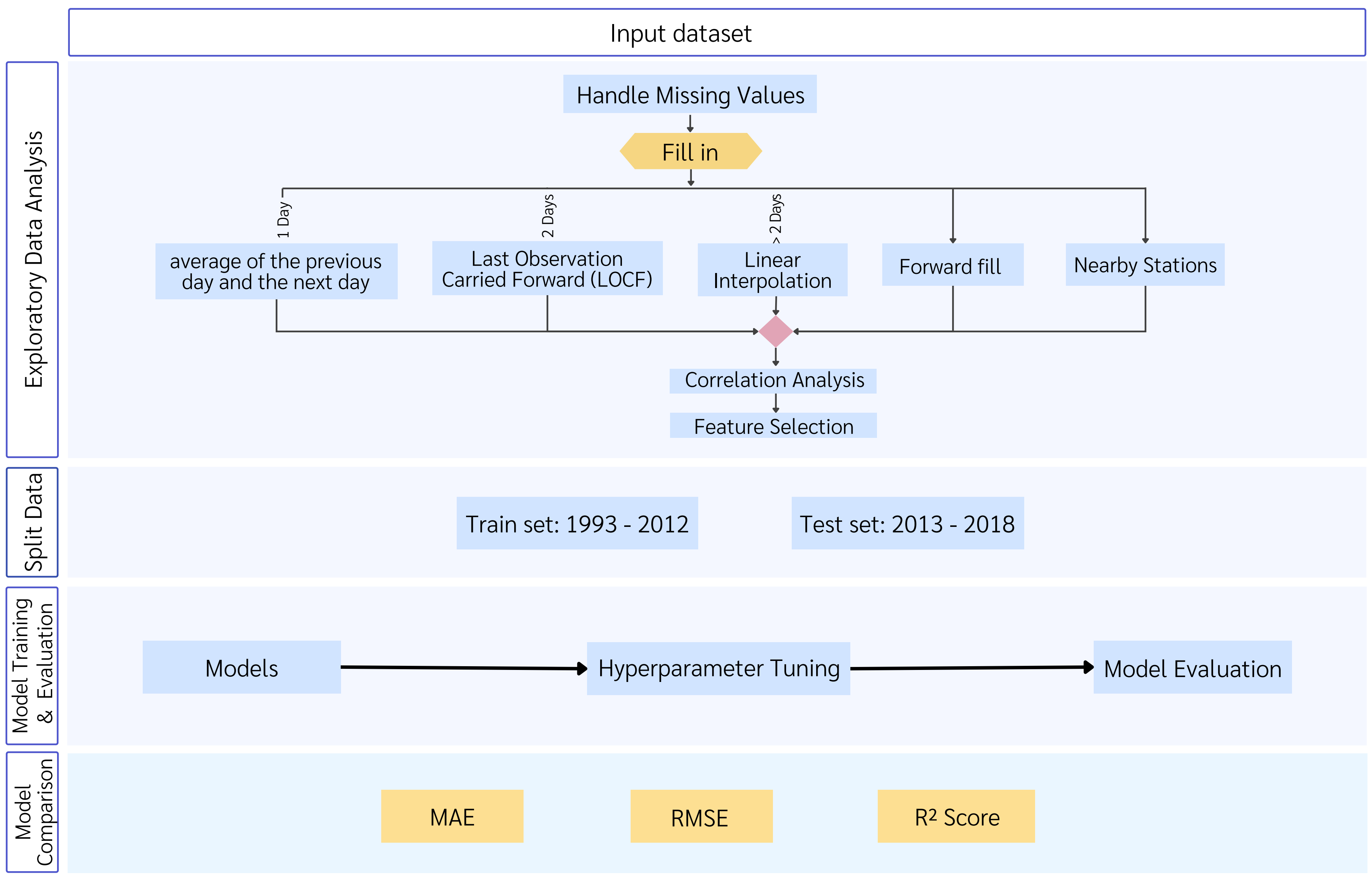

Fig. 1. Methodology Flowchart

Model Architecture

In this study, the models for rainfall prediction are developed using three approaches: the transformer model is autoregressive, generated new symbols based on previously ones at each stage. The encoder-decoder layers are stacked and interconnected for self-awareness [2] as displayed in Fig. 2. A simple LSTM network (as seen in Fig. 3, left) is commonly called vanilla LSTM. A stacked LSTM network consists of two or more simple LSTM networks connected as sequential hidden layers (as shown in Fig. 3, right). The stacked architecture enhances its ability to capture complex temporal dependencies, making it more effective for time-series prediction [3]. The XGBoost model is a supervised learning algorithm that utilizes the gradient boosting technique to combine multiple decision trees, creating a highly efficient model. It transforms features to reflect the patterns of past rainfall behavior [4] as shown in Fig. 4.

1 Transformer model

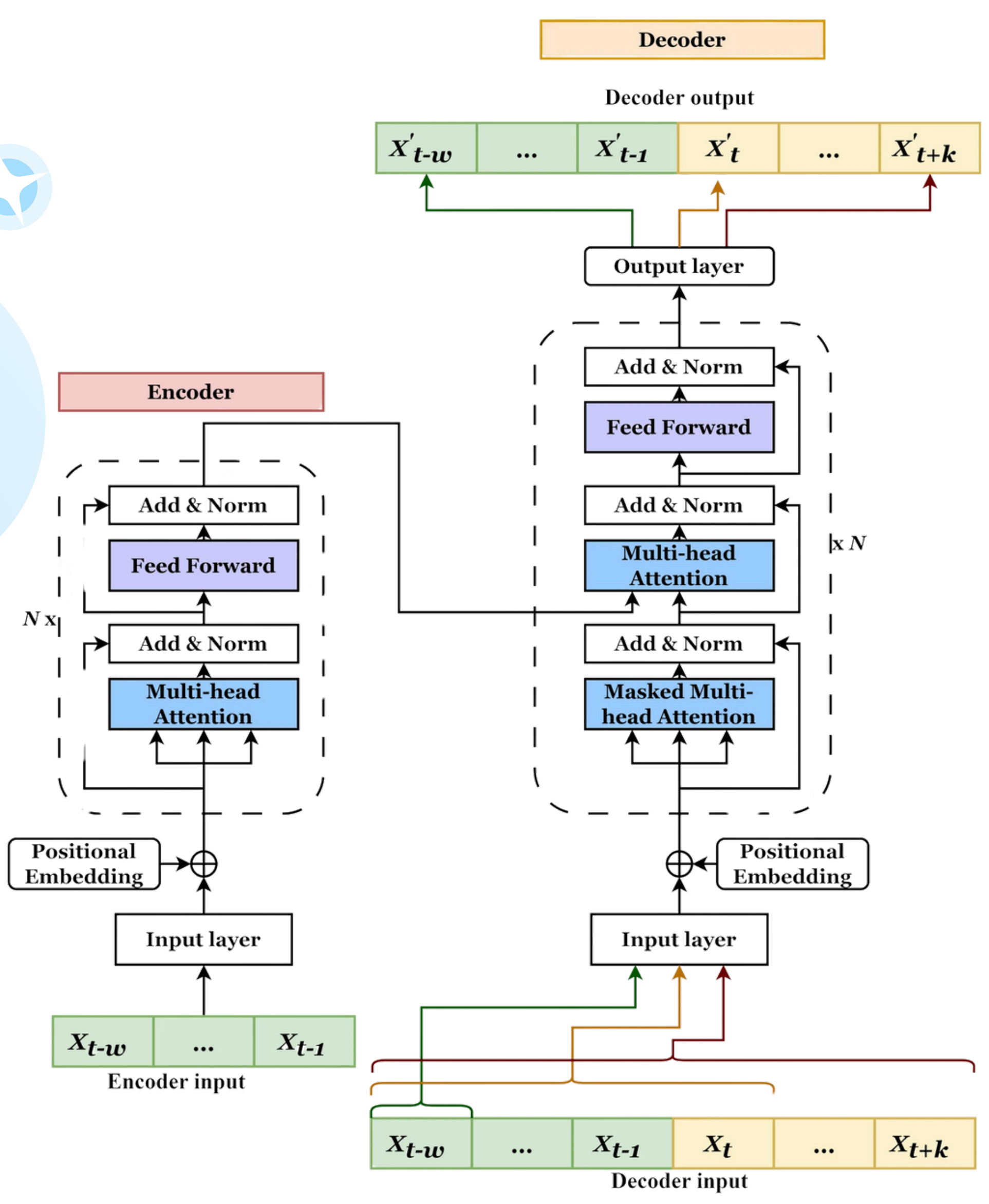

Fig. 2. Transformer architect [2]

2 LSTM model

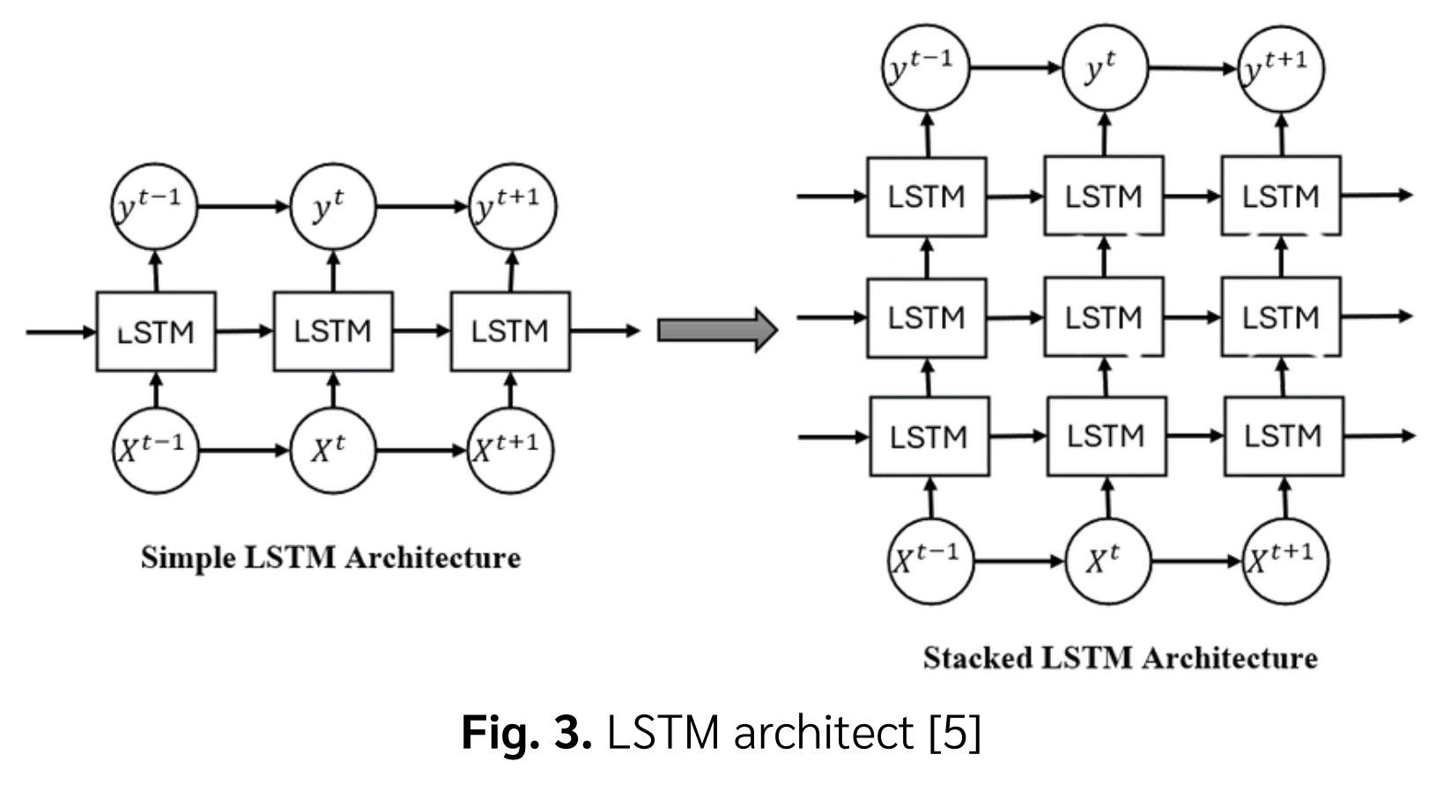

Fig. 3. LSTM architect [5]

3 XGBoost model

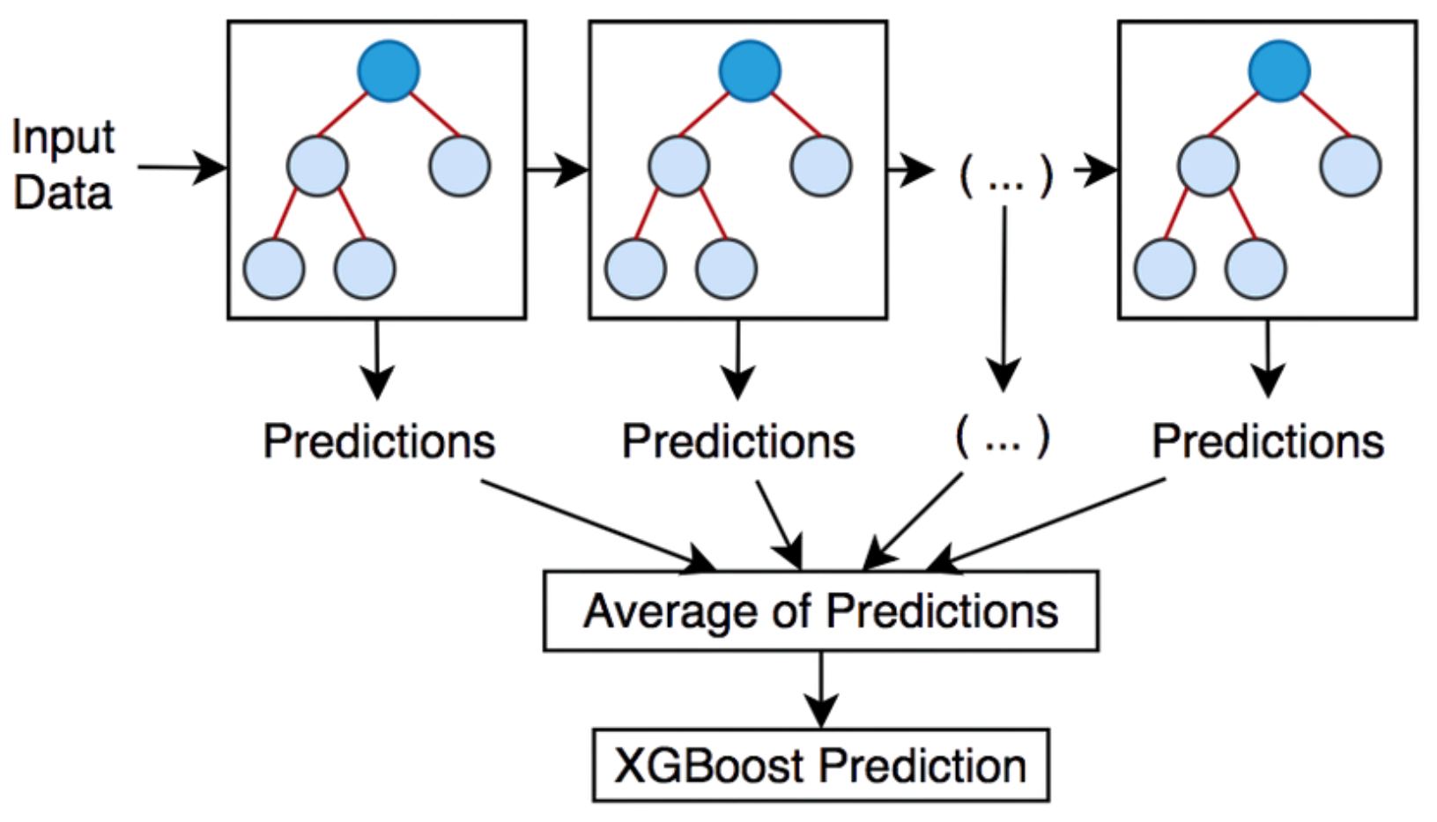

Fig. 4. XGBoost architect [6]

Objectives

1. To identify the climatic variables that impact rainfall prediction in the northeastern region of Thailand.
2. To compare the performance of rainfall prediction models in the northeastern region of Thailand using machine learning techniques.

Data Description

The meteorological data used include rainfall, max surface temperature, wind north, max wind direction, max wind speed, humidity, dry bulb temperature, and evaporation, sourced from 28 meteorological stations across northeastern Thailand. These data, collected by the Northeastern Meteorological Center of the Thai Meteorological Department [7], span the period from January 1993 to December 2018.

- Training set: January 1993 to December 2012
- Test set: January 2013 to December 2018

Results

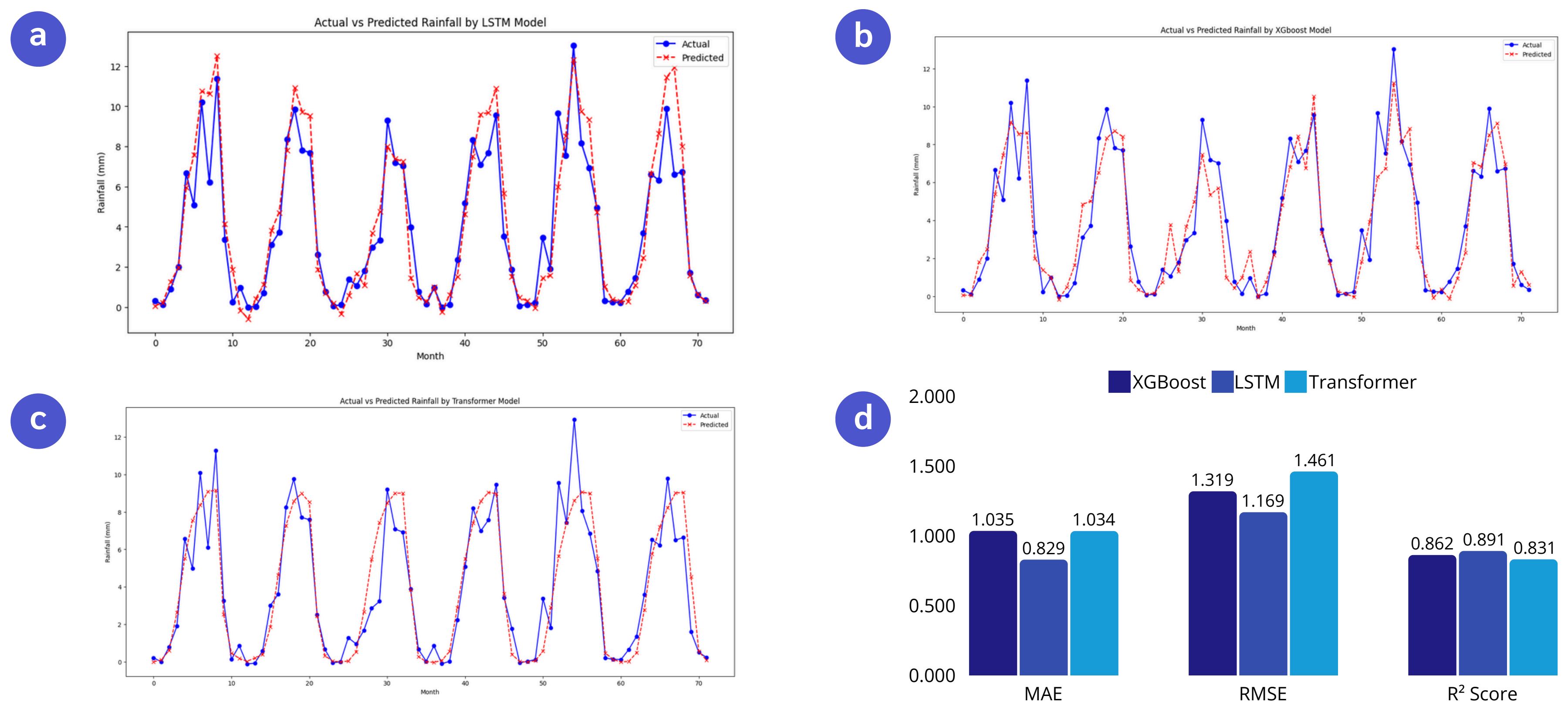

Fig. 5. Prediction Results of Transformer, LSTM, and XGBoost Models

We first perform linear regression analysis to identify climatic variables that affect rainfall prediction. There are eight features including min temperature, max surface temperature, wind north, max wind direction, max wind speed, humidity, dry bulb temperature, and evaporation that are mostly influenced rainfall. Next, we predict rainfall using LSTM, XGBoost and transformer. Results are showed in Fig 5. The experiment showed that the LSTM model (Fig. 5a) produced the lowest error, with a mean absolute error (MAE) of 0.829, a root mean square error (RMSE) of 1.169, and an R² score of 0.891, indicating the highest prediction accuracy among the models tested. Meanwhile, the XGBoost model (Fig. 5b) had an MAE of 1.035, an RMSE of 1.319, and an R² score of 0.862. Although the XGBoost model had higher errors than those of the LSTM, its results were still fairly close. The transformer model (Fig. 5c), on the other hand, had an MAE of 1.034, an RMSE of 1.461, and an R² score of 0.831, which signifies higher prediction errors compared to both LSTM and XGBoost. These results highlight the differences in model performance, with the LSTM and XGBoost models yielding similar results, while transformer showed higher errors than the other models tested.

Conclusion & Discussion

In conclusion, by comparison of model performance, LSTM model gave the lowest MAE value of 0.829, RMSE value of 1.169 and the highest R² score value of 0.891, representing the highest accurate prediction ability in all models used in the experiment. On the other hand, the relatively lower performance of the transformer model may be attributed to the limitations in tuning its parameters. Due to insufficient computational resources available in the notebook environment, we were unable to optimize the hyperparameters effectively, which likely impacted its predictive accuracy.

References

[1] วราฤทธิ์ พานิชกิจโกศลกุล., วารสารวิจัยและพัฒนา มจธ. ปีที่ 28., การพยากรณ์ปริมาณน้ำฝนของจังหวัดนครราชสีมา., https://digital.lib.kmutt.ac.th/journal/kmuttv28n2_3.pdf
[2] Nayak, G. H., Alam, W., Singh, K. N., Avinash, G., Ray, M., & Kumar, R. R. (2024). Modelling monthly rainfall of India through transformer-based deep learning architecture. Modeling Earth Systems and Environment, 1–18. https://doi.org/10.1007/s40808-023-01944-7
[3] Barrera-Animas, A. Y., Oyedele, L. O., Bilal, M., Akinosho, T. D., Davila Delgado, J. M., & Akanbi, L. A. (2021)., Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. Environmental Challenges, 4, 100102., https://doi.org/10.1016/j.mlwa.2021.100204
[4] Jian Rong Ban; Qi Gou; Ya Shi Li., Study on Rainfall Prediction of Yibin City Based on GRU and XGBoost, 10.1109/CTISC54888.2022.9849730
[5] Ayesha Sahar, Dongsoo Han., An LSTM-based Indoor Positioning Method Using Wi-Fi Signals., the 2nd International Conference., An LSTM-based Indoor Positioning Method Using Wi-Fi Signals., https://dl.acm.org/doi/10.1145/3271553.3271566
[6] Lara Marie Demajo., XGBoost model., https://www.researchgate.net/figure/GBoost-model-Source-Self_fig2_350874464
[7] the Thai Meteorological Department., https://www.tmd.go.th/en