IMPROVING PRECIPITATION ESTIMATION ACCURACY FOR THE CENTRAL VIETNAM REGION USING THE XGBOOST MODEL WITH MULTI-SOURCE DATA

Vu Duy Dong[1], Nguyen Hung An[1*], Nguyen Tien Phat [1], Nguyen Thi Nhat Thanh[2], Nguyen Thi Huyen[1]
[1]Le Quy Don Technical University
[2]University of Engineering and Technology - Vietnam National University, Hanoi

ARTICLE INFO

Received:	**17/10/2024**
Revised:	**22/11/2024**
Published:	**22/11/2024**

KEYWORDS

Rainfall estimation
Machine learning
XGBoost model
Himawari-8 satellite data
Digital elevation model

ABSTRACT

This paper presents a novel approach to enhancing the accuracy of precipitation estimation in Central Vietnam using the Extreme Gradient Boosting (XGBoost) machine learning model. The proposed method integrates multi-source data, combining satellite imagery from Himawari-8, atmospheric reanalysis from ERA-5, and digital elevation models from ASTER DEM to train the model. Rain gauge data from 175 stations across the region are used as target labels for validation. The proposed model achieved a CSI of 0.45, a POD of 0.75, and an RMSE of 4.53, with improvements of 11.11% to 86.67%, 28% to 93.33%, and 16.99% to 51.87%, respectively, compared to other precipitation products such as IMERG-Final Run, GSMaP_MVK, FengYun 4A, and PERSIANN-CCS. Detailed rainfall maps generated by the proposed model were compared with radar imagery during rainfall events, demonstrating a high degree of similarity. Furthermore, this approach serves as the basis for running near-real-time rainfall estimation models for the region of Vietnam.

TĂNG CƯỜNG ĐỘ CHÍNH XÁC TRONG DỰ BÁO LƯỢNG MƯA Ở KHU VỰC MIỀN TRUNG VIỆT NAM SỬ DỤNG MÔ HÌNH XGBOOST CHO DỮ LIỆU ĐA NGUỒN

Vũ Duy Đông[1], Nguyễn Hùng An[1*], Nguyễn Tiến Phát [1], Nguyễn Thị Nhật Thanh[2], Nguyễn Thị Huyền[1]
[1]Trường Đại học Kỹ thuật Lê Quý Đôn
[2]Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội

THÔNG TIN BÀI BÁO

Ngày nhận bài:	**17/10/2024**
Ngày hoàn thiện:	**22/11/2024**
Ngày đăng:	**22/11/2024**

TỪ KHÓA

Ước tính lượng mưa
Học máy
Mô hình học máy XGBoost
Dữ liệu vệ tinh Himawari-8
Mô hình độ cao số

TÓM TẮT

Bài báo này trình bày một cách tiếp cận mới để nâng cao độ chính xác trong ước tính lượng mưa tại miền Trung Việt Nam bằng cách sử dụng mô hình học máy Extreme Gradient Boosting (XGBoost). Phương pháp đề xuất tích hợp dữ liệu đa nguồn, kết hợp hình ảnh vệ tinh từ Himawari-8, phân tích lại khí quyển từ ERA-5 và mô hình độ cao kỹ thuật số từ ASTER DEM để đào tạo mô hình. Dữ liệu đo mưa từ 175 trạm trên khắp khu vực được sử dụng làm nhãn mục tiêu để xác thực. Mô hình đề xuất đạt được CSI 0,45, POD 0,75 và RMSE 4,53, với mức cải thiện lần lượt từ 11,11% tới 86,67%, 28% tới 93,33% và 16,99% tới 51,87%, so với các sản phẩm lượng mưa khác như IMERG-Final Run, GSMaP_MVK, FengYun 4A và PERSIANN-CCS. Bản đồ lượng mưa chi tiết do mô hình đề xuất tạo ra đã được so sánh với ảnh radar trong các sự kiện mưa, chứng minh mức độ tương đồng cao. Hơn nữa, phương pháp này tạo cơ sở để chạy các mô hình ước tính lượng mưa gần thời gian thực cho khu vực Việt Nam.

DOI: https://doi.org/10.34238/tnu-jst.11346

[*] Corresponding author. *Email: hungan@lqdtu.edu.vn*

1. Introduction

Precipitation is an essential hydro-meteorological parameter for climate forecasting, disaster response, and water resource management, but accurately estimating it remains a significant challenge [1]. Rainfall estimation usually relies on three primary data sources: rain gauge, weather radar, and satellite data. Currently, a common trend in this field is the use of machine learning (ML) with multi-source data, including these primary rainfall data along with additional auxiliary information such as digital elevation model (DEM) and atmospheric reanalysis data [2], [3]. ML based methods for rainfall estimation can be categorized into two main groups: those employing Single Machine Learning (SML) models for only regression task, and those using Double Machine Learning (DML) models for classification and regression tasks [3].

Chen et al. [4] developed a SML model using an Artificial Neural Network (ANN) for estimating rainfall using satellite data and radar data over Dallas–Fort Worth metroplex in USA. The model achieved the Normalized Standard Error of 37% with a rain threshold of 1.0 mm/h. Putra et al. [5] proposed an XGBoost based SML model of rainfall estimation over six regions of Indonesia using data from the Himawari-8 satellite, radar, and rain gauge, achieving Probability of Detection (POD) values from 0.89 to 0.92 and Root Mean Square Errors (RMSEs) from 1.85 to 3.08 mm/h. Mohia et al. [6] used three SML models: K-Nearest Neighbors Regression (K-NNR), Support Vector Regression (SVR), and Random Forest Regression (RF) for rainfall estimation over the northern region of Algeria using Meteosat satellite and rain gauge data. The RF model achieved the best performance with a RMSE of 1.3 mm, and a Mean Absolute Error (MAE) of 3.0 mm for the daily estimates.

Besides the advantage of being simple and saving time and computational resources by not requiring a classification step, SML-based methods may be less accurate in non-rainfall areas due to the unclear differentiation between regions with and without rain. To overcome this drawback, the Dual Machine Learning (DML) method has been developed and is increasingly being applied. Ouallouche et al. [7] proposed a RF-based DML architecture for estimating rainfall on a 3-hour and 24-hour scale in Northern Algeria. This model was compared with SVM-based and ANN-based DML models, showing superiority with the best RMSE of 1.12 mm for daytime and 1.28 mm for nighttime. Zhang et al. [3] investigated four different DML models, including RF-RF, RF-ANN, RF-SVM, and RF-ELM, with data including rain gauge observations, three satellite precipitation products (IMERG, PERSIANN, and GSMap), Shuttle Radar Topography Mission (SRTM) DEM data, and ERA-5 atmospheric reanalysis data. The DML based products outperform the SML products, with the Median Kling-Gupta Efficiency (mKGE) values ranging from 0.67 to 0.71 for the former, compared to 0.47 to 0.65 for the latter. Lyu et al. [8] proposed a DML architecture for merging multi-source precipitation data from GSMaP-Gauge, IMERG Final Run, ERA-5, and STRM DEM over the Tibetan Plateau. This architecture combines different machine learning algorithms, including XGBoost, SVM, RF, and KNN for classification, alongside LSTM for regression. The best DML model, using XGBoost for classification and LSTM for regression, achieved a POD of 0.63, a Critical Success Index (CSI) of 0.59, and a RMSE of 3.73.

This paper proposes an XGBoost-based DML architecture for rainfall estimation using multi-source data, including Himawari-8 satellite, ground rain observations, ERA-5, and ASTER DEM data. The proposal includes four XGB models: the first classifies rain/no-rain, the second classifies weak/strong rain, and the third and fourth perform regression for weak and strong rainfall, respectively. Furthermore, a multi-method feature selection solution was proposed to improve performance and reduce model complexity. The results of the proposed model were compared with four common rainfall products over the study area - IMERG Final Run, GSMaP_MVK, FY4A, and PERSIANN_CCS - and demonstrated its superiority.

The remainder of the paper is organized as follows: Section 2 describes the methodology for the proposed approach. Section 3 presents the experimental results and the performance evaluation of the proposed model. Section 4 draws conclusions and outlines directions for future research.

2. Materials and methods

2.1. Materials

The study area includes four provinces: Quang Binh, Quang Tri, Thua Thien Hue, and Da Nang in Central Vietnam, located between 15.6° - 18.4° North latitude and 104.4° - 108.8° East longitude. The input data used in this study are hourly rain gauge data, Himawari-8 satelitte brightness temperature (BT) data, and auxiliary data including ERA-5 and DEM data.

The rain gauge data were collected from 175 automatic rain gauge stations for the years 2019 and 2020 by the National Centre for Hydro-Meteorological Network for label assignment purposes. The satelitte BT data with a temporal resolution of 10 minutes and a spatial resolution of 2 km were extracted from 10 single infrared bands and 45 temperature brightness differences between these bands. The ERA-5 data of a spatial resolution of 25 km, developed by the European Centre for Medium-Range Weather Forecasts (ECMWF), and the ASTER DEM data from NASA with a spatial resolution of 30 m, were used for improving the accuracy of precipitation estimates. These input data were preprocessed to match the data sources to the same temporal resolution of 1 hour and the same spatial resolution of 4 km.

In addition, the proposed model was compared with the IMERG Final Run, GSMaP_MVK, FY4A, and PERSIANN_CCS precipitation products according to classification and regression metrics and reference radar rainfall maps for performance evaluation.

2.2. Methods

2.2.1. Proposed model

The proposed architecture for precipitation estimation includes four XGBoost models, M1 to M4, as described in Figure 1. Initially, the input data are classified into rainy and non-rainy events by M1 using a threshold of 0.1 mm/h. The rainy events are then further categorized into strong and weak rain by M2, utilizing a threshold of 1.8 mm/h. Subsequently, these categorized data are passed to the precipitation regression models: M3 for weak rain and M4 for strong rain. The output of the proposed product is a detailed precipitation map for the study area.

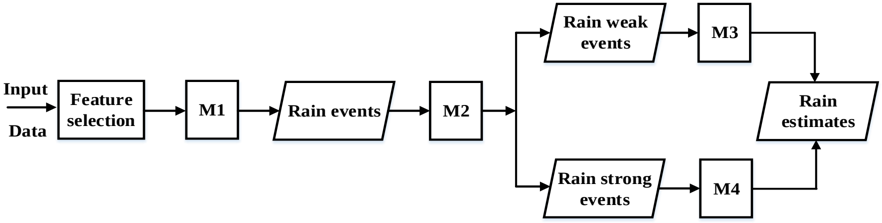

Figure 1. *Proposed architecture model for precipitation estimation*

To investigate the influence of auxiliary data (DEM, ERA-5) on rainfall estimation accuracy, two datasets were used: the first with 55 BT features of Himawari-8 (BT data), and the second combining these 55 features with the ASTER DEM and 17 ERA-5 features (BT+DEM+ERA data).

From the above features, only those with high importance and low correlation with other features will be selected as inputs of the model. This study proposed a feature selection strategy as follows: Firstly, original input features were ranked by five different methods: Mutual Information [9], Point-Biserial correlation [10], Sequential forward selection, Sequential backward belection [11], and Recursive feature elimination [12]. After that, the sum of the features' rankings by these five methods was sorted in ascending order, with the feature having the lowest cumulative score being deemed the most important. Finally, the most important features that correlate with other features below a certain threshold (0.6, according to the study) will be retained as actual inputs. As a result, we derived four reduced input feature sets for the four models (M1 to M4).

2.2.2. XGB method

XGBoost is an algorithm based on decision trees (DT), which combines multiple weak learners to minimize the loss function, thereby improving training performance. XGBoost parallelizes certain steps in the training process, such as finding the optimal split points for each DT, which accelerates the training process and reduces overall runtime. Detail information of XGBoost can be referred to in [13].

2.2.3. Training and evaluating the models

The original dataset is divided into a training set (80%) with 828,600 samples, of which 20% (165,720 samples) is used for validation, and a testing set (20%) with 170,261 samples. Specifically, the testing dataset includes data from April 2019 and June 2020, representing dry season months, and September 2019 and November 2020, representing rainy season months.

For each dataset, either BT or BT+DEM+ERA, feature selection (as described in the last paragraph of Section 2.2.1) and parameter tuning for each component of the proposed model, M1, M2, M3, and M4, were performed to achieve the best evaluation metric on the validation set. During the tuning process, their parameters, such as n_estimators, max_depth, subsample, colsample_bytree, min_child_weight, and learning_rate, are explored across different value ranges to identify the optimal set, which is evaluated on the validation set to determine the best-performing model.

Basic classification metrics for evaluating models M1, M2, and the proposed product are described in Table 1, where, TP - correctly predicted rainy samples; FP - non-rainy samples predicted as rainy; TN - correctly predicted non-rainy samples; FN - rainy samples predicted as non-rainy; N - total samples.

Table 1. *Basic classification metrics*

Name	Equation	Range	Optimal
Accuracy (*ACC*)	$ACC = (TP + TN)/N$	$[0, 1]$	1
Precision (*PRE*)	$PRE = TP/(TP + FP)$	$[0, 1]$	1
Recall (*POD*)	$REC(POD) = TP/(TP + FN)$	$[0, 1]$	1
F1-score (*F1*)	$F1 = (2 \times PRE \times RCL)/(PRE + RCL)$	$[0, 1]$	1
Critical Success Index (*CSI*)	$CSI = TP/(TP + FP + FN)$	$[0, 1]$	1
Equitable Threat Score (*ETS*)	$ETS = (TP - X)/\ (TP + FP + FN - X)$ $X = [(TP + FN) \times (TP + FP)]/N$	$[-{}^{1}/_{3}, 1]$	1

Basic regression metrics for evaluating the proposed and considered products are described in Table 2. Here, e_i, o_i, μ, and σ represent the values of estimation, observation, mean, and standard deviation, respectively.

Table 2. *Basic regession metrics*

Name	Equation	Range	Optimal
Root Mean Square Error (*RMSE*)	$RMSE = \sqrt{\sum (e_i - o_i)^2/N}$	$[0, +\infty]$	0
Mean Absolute Error (*MAE*)	$MAE = \sum \lvert e_i - o_i \rvert$	$[0, +\infty]$	0
Correlation Coefficient (*CC*)	$CC = \sigma_{o,e}/\sigma_o \sigma_e$	$[0, 1]$	1
Modified Kling-Gupta Efficiency (*mKGE*)	$mKGE = 1 - \sqrt{(CC - 1)^2 + (\beta - 1)^2 + (\gamma - 1)^2}$ $\beta = \mu_e/\mu_o\,; \gamma = \sigma_e \mu_o/\sigma_o \mu_e$	$[-\infty, 1]$	1

3. Results and discussion

The proposed model was evaluated in two steps. Step 1 involved independently assessing the performance of the classification models (M1, M2) and regression models (M3, M4). Step 2 focused on evaluating the classification performance of the integrated model (the proposed rainfall product), as follows.

3.1. Classification results

3.1.1. Rain/no rain model M1

The rain/no rain classification results of the model M1 for two datasets are shown in Table 3. This table demonstrates that the proposed model using BT+DEM+ERA data in the test set provided better performance in rain identification than the model using only BT data, with a Precision of 0.59, a Recall of 0.79, and an F1-score of 0.68, compared to 0.58, 0.76, and 0.66, respectively, for the BT-only model. It can be explained that the auxiliary data help achieve a better classification due to the additional information related to rainfall formation.

Table 3. *Rain/no-rain classification results of **M1***

Data	Class	Precision	Recall	F1_score	Accuracy
BT	No rain	0.89	0.78	0.83	0.77
	Rain	**0.58**	**0.76**	**0.66**	
BT+DEM+ERA	No rain	0.90	0.78	0.84	0.78
	Rain	**0.59**	**0.79**	**0.68**	

To provide a more visual evaluation, the rainfall maps produced by M1 for the BT+DEM+ERA dataset over two rain events were compared with those produced by the baseline model (M1 without feature selection) and those of the radar, as shown in Figure 2. Realizing that the baseline model incorrectly identified most of the no-rain points as rain (almost entirely red maps), while the maps produced by the proposed model show great similarity to the reference radar maps.

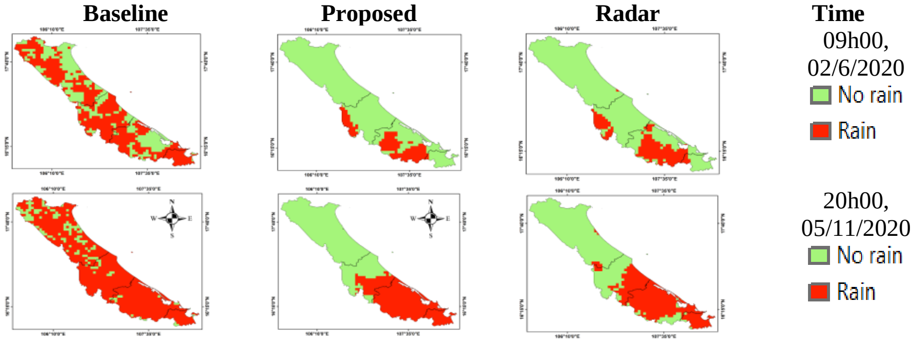

Figure 2. *Rain/no-rain classification maps of proposed model and radar maps*

3.1.2. Weak/strong rain model M2

Table 4. *Weak /strong rain classification results of **M2***

Data	Class	Precision	Recall	F1_score	Accuracy
BT	Weak rain	**0.77**	**0.64**	**0.70**	0.65
	Strong rain	0.52	0.66	0.58	
BT+DEM+ERA	Weak rain	**0.76**	**0.66**	**0.71**	0.65
	Strong rain	0.52	0.65	0.58	

The weak/strong rain classification results of the model M2 are shown in Table 4. As shown in Table 4, in the test set, model M2 provides nearly identical strong rain classification indices for the BT and BT+DEM+ERA data, with an F1 score of 0.58. However, the weak rain classification indices for the BT+DEM+ERA model are slightly better than those for the BT model, with F1 scores of 0.71 and 0.70, respectively. Overall, M2 with the BT+DEM+ERA dataset achieves higher accuracy compared to the BT model.

Similar to the evaluation of M1, weak/strong rain classification maps generated by M2 using the BT+DEM+ERA dataset for the same two rain events were compared with those produced by the baseline model and corresponding radar maps, as shown in Figure 3. There is still a great similarity between the maps produced by M2 and the reference radar maps, while the baseline model provides rain maps that are much different from the reference maps.

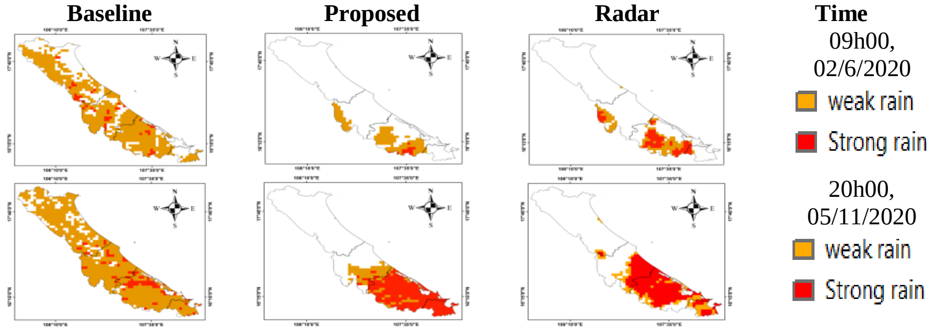

Figure 3. *Weak/strong rain classification maps of M2 and radar maps*

3.1.3. Classification performance of the proposed method product

To evaluate the classification performance of the proposed method, the classification results averaged from its 2,600 rainfall maps are compared with those of the rainfall products considered in the study area, as detailed in Table 5. As shown in this table, the proposed model using BT+ DEM+ ERA data achieves the best metrics, with a CSI of 0.45, an ETS of 0.32, and a POD of 0.75, representing an improvement of 11.11% and 40.00% in CSI, 6.25% and 40.63% in ETS, and 28% and 58.67% in POD compared to the second-best and third-best products, GSMaP_MVK and IMERG Final Run, respectively.

Table 5. *Comparison of classification performance of rain products*

Name	CSI	ETS	POD
IMERG Final Run	0.27	0.19	0.31
GSMaP_MVK	0.40	0.3	0.54
FY4A	0.05	0.03	0.05
PERSIANN_CCS	0.06	0.04	0.07
Proposed model _BT	0.3	0.21	0.70
Proposed model_BT+DEM+ERA	**0.45**	**0.32**	**0.75**
Optimal value	1	1	1

3.2. Regression model

3.2.1. Weak regression model M3

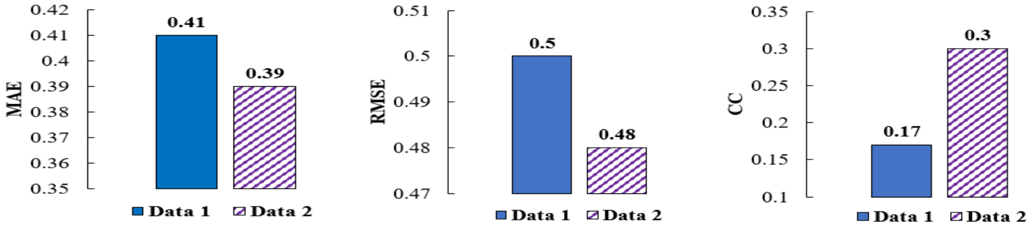

Figure 4. *Evaluation results of the weak rainfall regression model*

The weak rainfall regression results of the M3 model for both the BT (Data 1) and BT+DEM +ERA (Data 2) datasets are presented in Figure 4. This figure shows that the MAE and RMSE metrics for Data 2 are slightly better than those for Data 1. However, the CC value improved significantly by 76.47% (0.17 for Data 1 and 0.30 for Data 2).

3.2.2. Strong regression model M4

The strong rainfall regression results of the M4 model for both the BT (Data 1) and BT+DEM +ERA (Data 2) datasets are shown in Figure 5. Similar to the weak regression model results, the metrics for Data 2—MAE of 4.46, RMSE of 6.61, and CC of 0.29—are slightly better than those for Data 1, which are 4.51, 6.88, and 0.25, respectively.

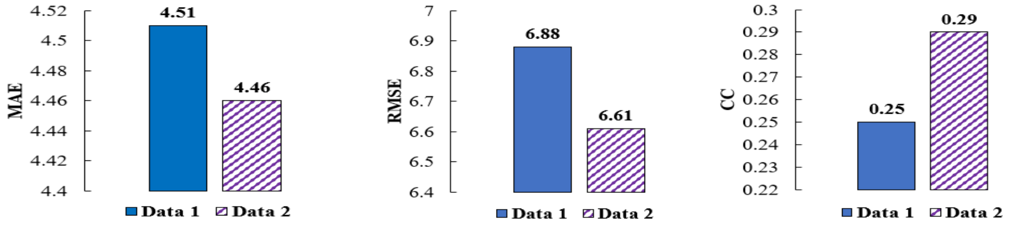

Figure 5. *Evaluation results of the strong rainfall regression model*

3.2.3. Regression performance of the proposed product

Similar to the classification performance evaluation, the regression performance of the proposed product was evaluated based on the average regression metrics from the 2600 rainfall maps and compared with those of the considered rainfall products, as detailed in Table 6. It can be seen from Table 6 that the proposed method yields the best values of RMSE (4.53) and mKGE (0.25) while providing the second-best value of MAE at 2.77, compared to the best MAE value of 2.65 for the GSMaP_MVK product.

Table 6. *Comparison of regression results of rain products*

Name	MAE	RMSE	mKGE
IMERG_Final	3.33	6.88	0.25
GSMaP_MVK	2.65	5.3	0.22
FY4A	3.22	6.33	-1.84
PERSIANN_CCS	2.81	5.45	-0.6
Proposed model _BT	3.22	5.97	0.18
Proposed model_BT+DEM+ERA	**2.77**	**4.53**	**0.25**
Optimal value	0	1	1

Additionally, the quality of the proposed rainfall product is also demonstrated by the high similarity of the rainfall maps it generates compared to the reference radar rainfall maps, as shown in Figure 6 with four rainfall events as examples.

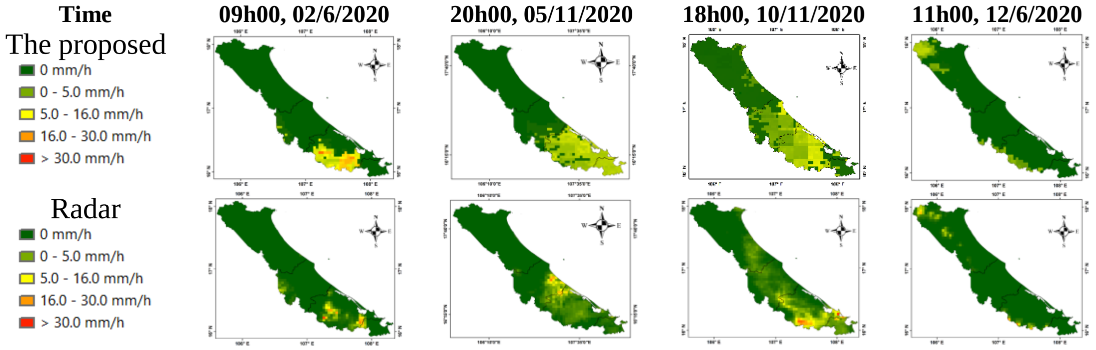

Figure 6. *Rainfall maps produced by the proposed product and reference radar maps*

In summary, the classification and regression performance of the proposed model with feature selection and auxiliary data is better than that of the model using only satellite BT data and the baseline model without feature selection.

4. Conclusions

This paper proposes a DML model of precipitation estimation over the Central Vietnam region using the XGBoost method and input feature selection. The proposed architecture was investigated on two datasets: (i) only Himawari-8 BT data and (ii) the BT data combined with ASTER DEM and ERA-5 data. The experimental evaluations indicate that incorporating additional auxiliary data, such as ERA-5 and DEM, contributes to improving accuracy, while the multi-method feature selection enhances the quality of the model and reduces its complexity.

The proposed product, which integrates auxiliary data, demonstrates its superiority with a CSI of 0.45, a POD of 0.75, and an RMSE of 4.53, showing improvements of 11.11% to 86.67%, 28% to 93.33%, and 16.99% to 51.87%, respectively, compared to the precipitation products considered, including IMERG-Final Run, GSMaP_MVK, FengYun 4A, and PERSIANN-CCS.

However, for extreme rainfall events (exceeding 50 mm/h), which constitute only about 0.2% of the dataset, the scarcity of data limits the XGBoost model's training, reducing its accuracy for these events. Therefore, applying advanced deep learning models, which can effectively capture hidden spatial and temporal features in the data, along with data augmentation techniques, could address this issue effectively in the future.

Acknowledgment

Our research is funded by a bilateral project between Vietnam and Italy. The project is titled "Research and Development of precipitation Estimation Methods Using Integrated precipitation Measurement Platforms for Agriculture" with project code NDT/IT/22/07.

REFERENCES

[1] Q. Sun, C. Miao, Q. Duan, H. Ashouri, S. Sorooshian, and K. L. Hsu, "A review of global precipitation data sets: Data sources, estimation, and intercomparisons," *Reviews of Geophysics*, vol. 56, no. 1, pp. 79-107, 2018.

[2] M. Guarascio, G. Folino, F. Chiaravalloti, S. Gabriele, A. Procopio, and P. Sabatino, "A machine learning approach for rainfall estimation integrating heterogeneous data sources," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-11, 2020.

[3] L. Zhang, X. Li, D. Zheng, K. Zhang, Q. Ma, Y. Zhao, and Y. Ge, "Merging multiple satellite-based precipitation products and gauge observations using a novel double machine learning approach," *Journal of Hydrology*, vol. 594, 2020, Art. no. 125969.

[4] H. Chen, V. Chandrasekar, R. Cifelli, and P. Xie, "A machine learning system for precipitation estimation using satellite and ground radar network observations," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 2, pp. 982-994, 2019.

[5] M. Putra, M. S. Rosid, and D. Handoko, "High-Resolution Rainfall Estimation Using Ensemble Learning Techniques and Multisensor Data Integration," *Sensors*, vol. 24, no. 15, 2024, Art. no. 5030.

[6] Y. Mohia, R. Absi, M. Lazri, K. Labadi, F. Ouallouche, and S. Ameur, "Quantitative Estimation of Rainfall from Remote Sensing Data Using Machine Learning Regression Models," *Hydrology*, vol. 10, no. 2, 2023, Art. no. 52.

[7] F. Ouallouche, M. Lazri, and S. Ameur, "Improvement of rainfall estimation from MSG data using Random Forests classification and regression," *Atmospheric Research*, vol. 211, pp. 62-72, 2018.

[8] Y. Lyu and B. Yong, "A novel Double Machine Learning strategy for producing high-precision multi-source merging precipitation estimates over the Tibetan Plateau," *Water Resources Research*, vol. 60, no. 4 , 2024, Art. no. e2023WR035643.

[9] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Computing and Applications*, vol. 24, pp. 175-186, 2014.

[10] J. D. Brown, "Point-biserial correlation coefficients," *Statistics*, vol. 5, no. 3, pp. 6-12, 2001.

[11] L. Čehovin and Z. Bosnić, "Empirical evaluation of feature selection methods in classification," *Intelligent Data Analysis*, vol. 14, no. 3, pp. 265-281, 2010.

[12] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, "Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products," *Chemometrics and Intelligent Laboratory Systems*, vol. 83, no. 2, pp. 83-90, 2006.

[13] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, August 2016, pp. 785-794.