Enhancing the accuracy of rainfall area classification in central Vietnam using machine learning methods

Nguyen Hung An*, Vu Duy Dong

Faculty of Radio and Electronics Engineering, Le Quy Don Technical University, 236 Hoang Quoc Viet, Nghia Do, Hanoi, Vietnam.
*Corresponding author: hungan@lqdtu.edu.vn
Received 8 Sep. 2025; Revised 2 Nov. 2025; Accepted 10 Nov. 2025; Published 28 Nov. 2025.
DOI: https://doi.org/10.54939/1859-1043.j.mst.107.2025.105-113

ABSTRACT

This study applies machine learning techniques, including Light Gradient Boosting Machine (LGBM), XGBoost (XGB), and Random Forest (RF), in conjunction with multi-source data comprising Himawari-8 satellite observations, ground-based rain gauge measurements, and auxiliary data such as ERA-5 reanalysis and the ASTER Digital Elevation Model (DEM), to enhance rainfall classification accuracy over Central Vietnam. Existing rainfall products in the region, including IMERG Final Run, IMERG Early, GSMaP_MVK_Gauge, PERSIANN_CCS, and FY-4A, are employed to evaluate the performance of the proposed classification approach. The results indicate that all proposed rainfall classification products exhibit high performance. Among them, the rainfall classification product based on LGBM achieved the highest performance across key evaluation metrics, including Probability of Detection (POD), Critical Success Index (CSI), Equitable Threat Score (ETS), and Heidke Skill Score (HSS). Compared to the investigated best-performing reference product, GSMaP_MVK_Gauge, the LGBM improves these metrics by 38.89%, 20.0%, 16.67%, and 13.04%, respectively. These findings highlight the potential of machine learning models, particularly LGBM, in enhancing the classification performance of meteorological models that utilize small but complex and high-dimensional datasets.

Keywords: Classification rainfall; Machine learning; LightGBM; Random forest; Himawari-8; ERA-5.

1. INTRODUCTION

Accurate rainfall datasets, especially for strong rainfall, are crucial for societal and economic planning. Rainfall classification can be done using traditional methods (rain gauges, radar, satellites) or modern approaches like machine learning (ML) and deep learning (DL). While DL models demand large datasets, high computational resources, and long training times, traditional ML models are better suited for small- to medium-sized datasets and limited computing capacity.

Numerous studies have demonstrated the effectiveness of ML models in rainfall classification tasks. Ouallouche et al. [1] employed three machine learning models, including RF, Support Vector Machine (SVM), and Artificial Neural Network (ANN), to classify daytime and nighttime rainfall over northern Algeria using MSG satellite infrared and radar data. Among them, the RF model achieved the highest performance, with a CSI of 0.78 during the day and a CSI of 0.77 at night. Liu et al. (2021) [2] applied a LGBM model to classify rainfall over regions in China using radar data, achieving a POD of 0.85 and a CSI of 0.65. Dong et al. (2024) [3] developed an XGB-based framework to classify rainfall over Vietnam's Central Coast based on data from Himawari-8, ECMWF Reanalysis V5 (ERA-5), and the Advanced Spaceborne Thermal Emission and Reflection Radiometer Digital Elevation Model (ASTER DEM). Incorporating ERA-5 improved performance, yielding a POD of 0.75, CSI of 0.45, and ETS of 0.32.

In this study, we propose a two-stage machine learning framework for rainfall classification in Central Vietnam that integrates multi-source data, including Himawari-8 satellite imagery, surface rain gauge observations, ERA-5 reanalysis, and ASTER DEM for the period 2019–2020. The main contributions of this work are: (1) a comprehensive investigation and optimization of the two-stage

classification framework across the three models (RF, XGB, and LGBM) to identify the most effective architecture (LGBM), incorporating feature selection and strategies for handling data imbalance; (2) the demonstration of the potential of machine learning models, particularly LGBM, in enhancing the classification performance of meteorological models that utilize small but complex and high-dimensional datasets; and (3) the development of a high-quality rainfall dataset that outperforms existing precipitation products, including IMERG Final Run, IMERG Early Run, GSMaP_MVK_Gauge, PERSIANN_CCS, and FY-4A, for the study region.

The remainder of this paper is organized as follows. Section 2 describes the study area and the characteristics of the data sources used. Section 3 presents the methodology. Section 4 discusses the obtained results and their evaluation. Finally, section 5 concludes the study and outlines potential directions for future research.

2. CASE STUDY AND DATASETS

2.1. Case study

This study focuses on four coastal provinces of Central Vietnam (Quang Binh, Quang Tri, Thua Thien Hue, and Da Nang) located between 15.6°–18.4°N and 104.4°–108.8°E. The region has a high annual rainfall, mostly concentrated from August to December. Its steep terrain rises from east to west, leading to localized precipitation mainly in mountainous areas. The climate is complex, influenced by both the southwest and northeast monsoons.

2.2. Datasets

2.2.1. Data characteristics

The data used in this study, specifically collected for the four provinces of Central Vietnam mentioned above during the period 2019–2020, include training data (Himawari-8, ERA-5 reanalysis, ASTER DEM, and rain gauge observations), comparative data (IMERG Final Run, IMERG Early Run, GSMaP_MVK_Gauge, FY-4A, and PERSIANN_CCS), and ground-truth data (175 rain gauge stations) used to evaluate the classification performance of rainfall products. The characteristics and roles of these data sources are described in detail in table 1.

Table 1. *Characteristics of the data sources.*

Datasets	Role	Resolution	
		Temporal	Spatial
Himawari-8 [3]	Training	10 minutes	02 km
ERA-5 [4]	Training	60 minutes	25 km
ASTER DEM [5]	Training		30 m
Rain gauge stations	Labeling and ground truth	60 minutes	
IMERG Final Run [6]	Comparison	30 minutes	0.1° × 0.1°
IMER Early Run [6]	Comparison	30 minutes	0.1° × 0.1°
GSMap_MVK_Gauge [7]	Comparison	60 minutes	0.1° × 0.1°
FY-4A [8]	Comparison	15 minutes	0.04° × 0.04°
PERSIANN-CCS [9]	Comparison	30 minutes	0.04° × 0.04°

2.2.2. Data preprocessing

After collection, the datasets were spatially and temporally matched before model training, with all data standardized to a 1-hour temporal resolution and a 4 km spatial resolution. For temporal alignment, the Himawari-8 satellite data, originally recorded every 10 minutes, were averaged over six consecutive observations to obtain a 1-hour resolution, matching that of the ERA-5 reanalysis data, ASTER DEM, and rain gauge measurements. For spatial alignment, the ERA-5 data (originally 25 km resolution) were resampled to 4 km using the nearest-neighbor interpolation method. The Himawari-8 data (2 km) and ASTER DEM data (30 m) were also resampled to 4 km

using an average pooling technique. Similarly, all five global rainfall products were spatially and temporally adjusted to match the proposed rainfall dataset before comparison.

After matching, the input data were sampled in both space and time to ensure the highest possible independence between the training and testing datasets. Specifically, the initial dataset was divided into a training set of 512,844 samples (80%) and a testing set of 149,362 samples (20%). The evaluation dataset includes rainfall observations from April and June 2020 (dry season) and September and November 2020 (rainy season).

3. METHODOLOGY

3.1. Machine learning algorithm

This study investigates three different machine learning models, including RF, XGB, and LGBM. They are briefly described as follows. The RF algorithm is a machine learning method that builds an ensemble of decision trees (DTs), each trained in parallel on random subsets of the original dataset. The final classification is determined by majority voting among the trees.

The XGB is another high-performance, tree-based algorithm designed to combine multiple weak learners to minimize a specified loss function. XGB parallelizes certain steps in the training process, such as identifying optimal split points for DTs, which accelerates model training and reduces overall computation time.

The LGBM is a gradient boosting algorithm based on decision trees (DTs) that builds trees sequentially to minimize the loss of previous iterations. Unlike conventional tree models, LGBM uses a histogram-based and leaf-wise growth strategy instead of the level-wise approach. It also employs two key sampling techniques: EFB, which reduces dimensionality by merging mutually exclusive features, and GOSS, which focuses on samples with large gradients to retain essential information while reducing training data.

3.2. Framework

A rainfall classification framework consisting of two models, X1 and X2, is proposed, as illustrated in figure 1. Each model is implemented using three algorithms: RF, XGB, and LGBM. The input features (55 BT features from Himawari-8, 13 meteorological features from ERA-5, and the ASTER DEM) are spatially and temporally matched through the data matching module described in section 2.2.2 and subsequently selected for input into models X1 and X2.

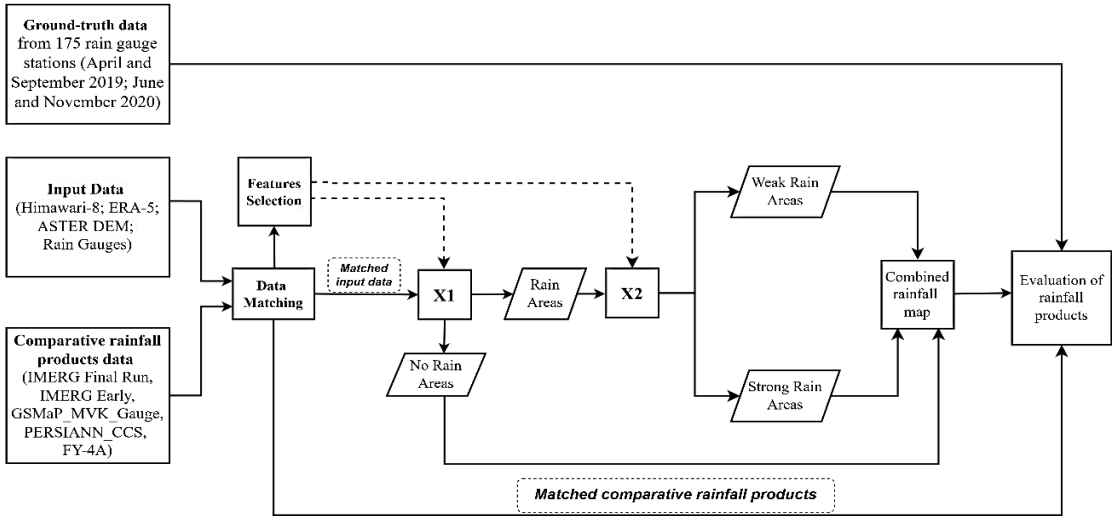
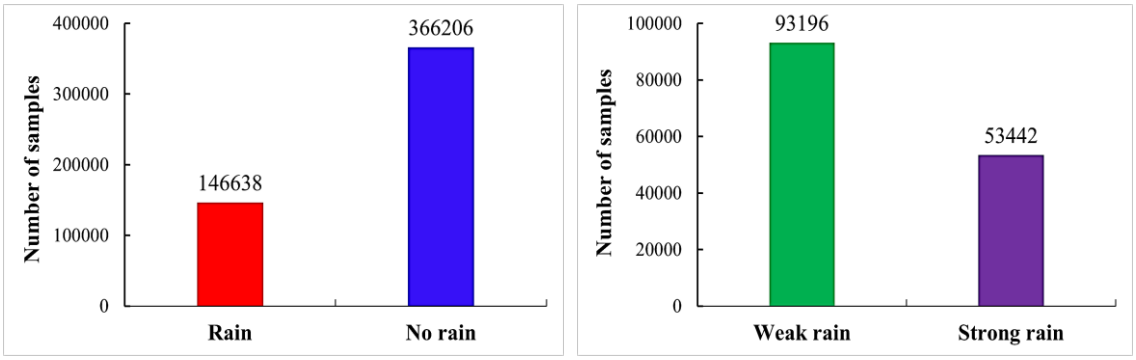


Figure 1. *The machine learning framework for rainfall classification.*

Rainfall was classified using two thresholds: 0.1 mm/h for rain/no-rain [10, 11] and 1.8 mm/h

for weak/strong rain [12]. Two models, X1 and X2, were trained independently: X1 identifies rain/no-rain areas, and X2 classifies rainfall intensity in the areas flagged as rainy.

As shown in figure 2, the input data for models X1 and X2 are imbalanced: no-rain samples outnumber rain samples by 2.49 times for X1, while weak rain samples are 1.74 times fewer than strong rain samples for X2. To address this imbalance, the Class Weight (CW) technique is applied, assigning higher weights to minority classes (rain for X1, strong rain for X2) to improve classification performance (see section 4.2 for details).



a) X1 *b) X2*

Figure 2. *Statistics of the number of samples in the input data of the models. a) X1, b) X2.*

The proposed rainfall product is obtained by integrating the outputs of Models X1 and X2, resulting in a comprehensive rainfall map of the study area that distinguishes no rain, weak rain, and strong rain locations. Its performance is evaluated against five satellite-based precipitation datasets, IMERG Final Run, IMERG Early, GSMaP_MVK_Gauge, PERSIANN_CCS, and FY-4A, after preprocessing them to match its spatial and temporal resolution.

To simplify model training and reduce model complexity while maintaining performance, this study employs the combined correlation evaluation technique to identify the most relevant input features. Features are first ranked by their importance to the target variable, and then redundant features are removed based on correlation analysis.

3.3. Training and evaluation

During model training, the Scikit-learn library was used to perform 5-fold cross-validation (with k = 5) to optimize performance and reduce overfitting.

The performance of the proposed rainfall classification products was assessed using five commonly used metrics: POD, F1-Score, CSI, ETS, FAR, and HSS. They are summarized in table 2, where TP, FP, TN, FN, and N denote true positives, false positives, true negatives, false negatives, and total samples, respectively.

Table 2. *Basic classification metrics.*

Name	Equation	Range	Optimal
POD	$POD = TP/(TP + FN)$	$[0, 1]$	1
F1-score	$F1 = (2 \times PRE \times REC)/(PRE + REC)$	$[0, 1]$	1
CSI	$CSI = TP/(TP + FP + FN)$	$[0, 1]$	1
FAR	$FAR = FP/(TP + FP)$	$[0, 1]$	0
ETS	$ETS = (TP - X)/\ (TP + FP + FN - X)$ $X = [(TP + FN) \times (TP + FP)]/N$	$[-1/3, 1]$	1
HSS	$HSS = 2 \times \frac{TP \times TN - FP \times FN}{(TP + FN) \times (FN + TN) + (TP + FP) \times (FP + TN)}$	$[0, 1]$	1

4. RESULTS

4.1. Results of feature selection

This study compares the feature selection results obtained using two strategies: the combined correlation evaluation (CCE) technique and RF Importance (RI) (available in the scikit-learn library). The feature selection results for Models X1 and X2 are listed in table 3.

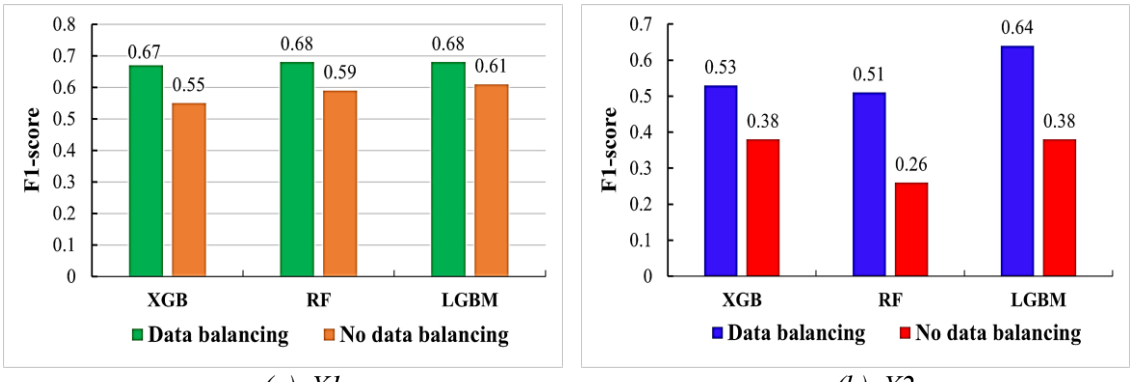
Table 3. *Features selected for model training.*

CCE technique		RFI technique	
X1	**X2**	**X1**	**X2**
I2B_IRB; B10_B11	B10_I4B; I2B_IRB	B10_B14; B16_IRB	B14_B16; B12_IRB
B10_B09; B11_B14	B11_B14; B11_IRB	B11_B16; B14_I2B	B10_B14; B14_WVB
B11_IRB; B12_B16	B12_B09; B12_WVB	B14_B16; B11_B09	B14_B09; WVB_IRB
B12_I4B; B14_I2B	B12_I4B; B14_I2B	B10_B11; B10_IRB	B11_WVB; TCWV
B14_IRB; B16_I4B	B14_IRB; B09_I4B	R850; CAPE; KX	KX; R850
I4B_I2B; KX; CIN	WVB_I4B; R850	UWIND850; UWIND500	TCW;VWIN250
CAPE; TCW	DEM	UWIND250; VWIND500	UWIND500
SLOR; DEM	SLOR	VWIND850; TCW	UWIND850

4.2. Evaluation of the rainfall classification performance of individual models

In these evaluations, the F1-score of the rain class is used to evaluate model X1, while that of the strong rain class is used to assess model X2. Additionally, the classification results of models X1 and X2 using balanced data are compared with those obtained from imbalanced data, employing the RF, XGB, and LGBM algorithms.

The detailed results are presented in figure 3. From figure 3, it can be observed that applying the CW technique to balance the data significantly improved the F1-scores for both the rain and strong rain classes across all three algorithms. For the rain class (figure 3a), the LGBM and RF algorithms achieved the highest F1-score of 0.68, while the XGB algorithm obtained a slightly lower value of 0.67. Compared to the unbalanced case, the F1-scores of LGBM, XGB, and RF improved by 11.47%, 21.82%, and 15.25%, respectively, after applying the balancing technique. Similarly, for the strong rain class (figure 3b), the LGBM algorithm achieved the highest F1-score of 0.64, followed by XGB with 0.53 and RF with 0.51. Compared to the unbalanced data case, applying the CW technique led to improvements in the F1-score for the strong rain class of 68.42%, 39.47%, and 96.15% for LGBM, XGB, and RF, respectively.



(a) X1 *(b) X2*

Figure 3. *Rainfall classification results of the individual models. (a) X1, (b) X2.*

The final rainfall classification product is derived from the combined outputs of models X1 and X2, enabling the identification of areas with weak rain, strong rain, and no rain. Its classification

performance is then compared with the other rainfall products in the study area, using rain gauge observations as reference, as presented in the following sections.

4.3. Evaluation of the classification performance of the proposed rainfall product

To assess the overall rainfall classification performance of the three proposed products based on RF, XGB, and LGBM, this study analyzed 2,599 rainfall classification maps generated by these models and compared them with the corresponding maps from five comparative rainfall products: IMERG Final Run, IMERG Early, GSMaP_MVK_Gauge, PERSIANN_CCS, and FY-4A. The classification performance was assessed using the metrics CSI, ETS, POD, and HSS, as described in table 2.

The detailed results are illustrated in figure 4. Figure 4 illustrates the performance of the three proposed rainfall classification products and five comparative products. Among the proposed products, LGBM achieved the highest scores, with a CSI of 0.48, a POD of 0.75, an ETS of 0.35, and an HSS of 0.52. XGB ranked second with all metrics higher than RF, except for the POD, which was 0.63 compared to 0.71 for RF. RF ranked last among the three proposed products, with the lowest CSI (0.37) and significantly lower ETS and HSS scores, 66.67% and 48.57% of those achieved by LGBM, respectively.

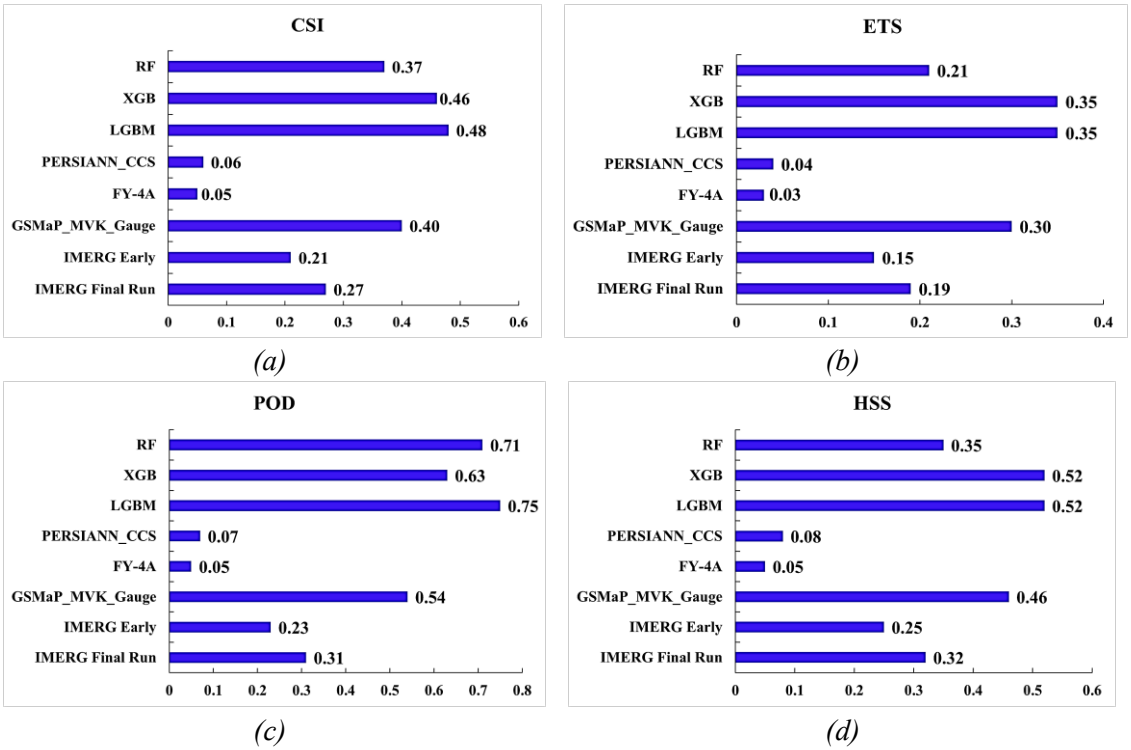


Figure 4. *Comparison of classification performance among different rainfall products: (a) CSI, (b) ETS, (c) POD, (d) HSS.*

In addition, when comparing the three proposed products with the five comparative rainfall products, it can be seen that the LGBM and XGB rainfall products outperformed all comparative rainfall products across all four evaluation metrics. Particularly when compared to the GSMaP_MVK_Gauge product - the best-performing among the four reference products - LGBM achieved improvements of 20.0% in CSI, 38.99% in POD, 13.04% in HSS, and 16.67% in ETS. Although the RF rainfall product performed worse than GSMaP-MKV-Gauges in three metrics, including CSI (0.37 vs 0.40), ETS (0.21 vs 0.30), and HSS (0.35 vs 0.46), it still outperformed all the other reference rainfall products.

Overall, the three proposed rainfall products outperformed all comparative rainfall products, except for GSMaP_MVK_Gauge, which performed better than the RF-based product. Also, the LGBM-based product with the best evaluation scores was selected as our final rainfall product.

4.4. Discussion

4.4.1. Impact of the feature selection techniques

To evaluate the effectiveness of the feature selection techniques, we compared the rainfall classification performance of the LGBM rainfall product using the feature set selected from the Combined Correlation Evaluation technique (referred to as LGBM-CCE) with that of the LGBM rainfall product using the feature set selected by the RF Importance technique (referred to as LGBM-RFI). Four evaluation metrics, namely CSI, ETS, POD, and FAR, were employed for the assessment. The comparative results are presented in table 4.

Table 4. *Impact of different feature selection techniques.*

Rainfall products	CSI	ETS	POD	FAR
LGBM-CCE	**0.48**	0.35	**0.75**	**0.43**
LGBM-RFI	0.42	0.35	0.74	0.50

From table 3, it can be observed that the proposed LGBM-CCE product achieves higher classification performance compared to the LGBM-RFI product, which uses the RF importance technique for feature selection. LGBM-CCE shows improvements in CSI and POD over LGBM-RFI by 14.29% and 1.35%, respectively. In terms of FAR, LGBM-CCE achieves a 14.0% improvement compared to LGBM-RFI, while the ETS values of the two products remain identical.

4.4.2. Comparison of rainfall classification performance with previously studied rainfall products

The rain classification performance of the proposed product (LGBM-CCE) was compared with the results reported in the previously published study [3]. Both studies employed the same dataset but utilized different algorithms, specifically LGBM in this study and XGB in the published study, referred to as XGB-CCE. Three evaluation metrics, including the CSI, ETS, and POD, were used for assessment. The comparison results are presented in figure 5.

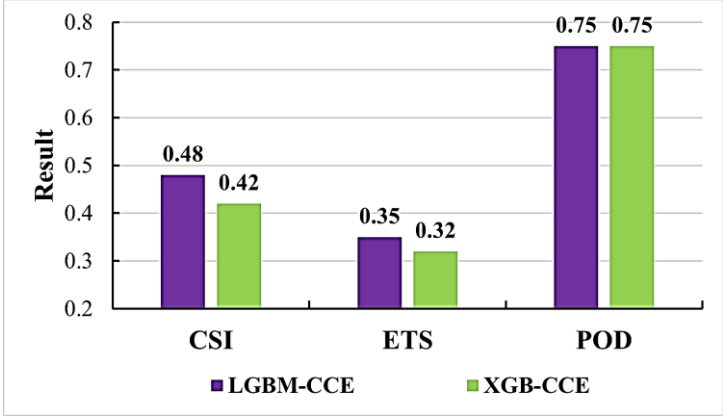


Figure 5. *Comparison of classification performance between the proposed LGBM-CCE product and the published XGB-CCE product.*

From figure 5, it can be observed that, compared with the XGB-CCE rainfall product presented in the study [3], the proposed LGBM-CCE rainfall product in this study achieved higher classification performance, with improvements of 6.67% in CSI and 9.38% in ETS, while the POD remained unchanged.

5. CONCLUSIONS

This study proposes a two-stage rainfall classification framework using RF, XGB, and LGBM with a multisource dataset of rain gauge, Himawari-8, ASTER DEM, and ERA-5 data from 2019 to 2020. The proposed rainfall classification products outperform five comparative datasets (IMERG Final Run, IMERG Early, GSMaP_MVK_Gauge, PERSIANN_CCS, and FY-4A), with the LGBM-based product achieving the highest performance. Compared to the best dataset (GSMaP_MVK_Gauge), it shows improvements of 38.89% in POD, 20.0% in CSI, 16.67% in ETS, and 13.04% in HSS.

The proposed approach is effective for developing rainfall estimation models under limited data availability and computational resources. However, the classification performance for the strong rain class remains relatively low and could be improved by employing advanced deep learning models with larger datasets and greater computational capacity.

Acknowledgments: *The Himawari-8 satellite, weather radar images, and surface rain gauge data used in this study were provided by the National Centre for Hydro-Meteorological Network.*

REFERENCES

[1]. F. Ouallouche, M. Lazri, and S. Ameur, "*Improvement of rainfall estimation from MSG data using Random Forests classification and regression*", Atmos Res, vol. 211, pp. 62–72, (2018), doi: 10.1016/J.ATMOSRES.2018.05.001.

[2]. X. Liu, H. Duan, W. Huang, R. Guo, and B. Duan, "*Classified Early Warning and Forecast of Severe Convective Weather Based on LightGBM Algorithm*", Atmospheric and Climate Sciences, vol. 11, pp. 284–301, (2021), doi: 10.4236/acs.2021.112017.

[3]. V. Dong, A. Nguyen, N. Phat, N. Thanh, N. Huyen, "*Improving precipitation estimation accuracy for the Central Vietnam region using the XGBoost model with multi-source data*", TNU Journal of Science and Technology, vol. 229, pp. 69–77, (2024), doi: 10.34238/tnu-jst.11346.

[4]. D. Lavers, A. Simmons, F. Vamborg, and M. Rodwell, "*An evaluation of ERA5 precipitation for climate monitoring*", Quarterly Journal of the Royal Meteorological Society, vol. 148, (2022), doi: 10.1002/qj.4351.

[5]. A. Mohammadi et al., "*A Multi-Sensor Comparative Analysis on the Suitability of Generated DEM from Sentinel-1 SAR Interferometry Using Statistical and Hydrological Models*", Sensors, vol. 20, p. 7214, (2020), doi: 10.3390/s20247214.

[6]. L. Xuegang et al., "*Comparative evaluation of GPM IMERG V07 early, late and final run products compared to IMERG V06 in Sichuan Province, China*", Theor Appl Climatol, vol. 156, (2025), doi: 10.1007/s00704-025-05569-x.

[7]. C. Zhou, L. Zhou, J. Du, J. Yue, and T. Ao, "*Accuracy evaluation and comparison of GSMaP series for retrieving precipitation on the eastern edge of the Qinghai-Tibet Plateau*", J Hydrol Reg Stud, vol. 56, p. 102017, (2024), doi: 10.1016/j.ejrh.2024.102017.

[8]. Z. Wang, H. Chai, C. Zhu, H. Ma, N. Zheng, and P. Chen, "*Reconstruction of High-Resolution Precipitable Water Vapor of FY-4A Based on GNSS and Remote Sensing Data*", (2025), doi: 10.2139/ssrn.5243923.

[9]. P. Nguyen et al., "*The PERSIANN family of global satellite precipitation data: a review and evaluation of products*", Hydrol Earth Syst Sci, vol. 22, pp. 5801–5816, (2018), doi: 10.5194/hess-22-5801-2018.

[10]. C. Gianoglio, S. Zani, M. Colli, and D. Caviglia, "*Rainfall Classification in Genoa: Machine Learning vs. Adaptive Statistical Models Using Satellite Microwave Links*", IEEE Access, vol. PP, p. 1, (2024), doi: 10.1109/ACCESS.2024.3458407.

[11]. S. Kolios, N. Hatzianastassiou, C. J. Lolis, and A. Bartzokas, "*Accuracy Assessment of a Satellite-Based Rain Estimation Algorithm Using a Network of Meteorological Stations over Epirus Region, Greece*", Atmosphere (Basel), vol. 13, no. 8, (2022), doi: 10.3390/atmos13081286.

[12]. H. Hirose, S. Shige, M. Yamamoto, and A. Higuchi, "*High Temporal Rainfall Estimations from Himawari-8 Multiband Observations Using the Random-Forest Machine-Learning Method*", Journal of the Meteorological Society of Japan. Ser. II, vol. 97, (2019), doi: 10.2151/jmsj.2019-040.

TÓM TẮT

Nâng cao độ chính xác trong việc phát hiện các khu vực có mưa ở miền Trung Việt Nam bằng phương pháp học máy

Nghiên cứu này ứng dụng các phương pháp học máy, bao gồm Light Gradient Boosting Machine (LGBM), XGBoost (XGB), và Random Forest (RF), kết hợp với dữ liệu đa nguồn dữ liệu vệ tinh Himawari-8, dữ liệu quan trắc mưa từ trạm đo mưa, và dữ liệu phụ trợ bao gồm tái phân tích ERA-5 và mô hình độ cao số ASTER (DEM), nhằm nâng cao độ chính xác trong phân loại mưa tại khu vực miền Trung Việt Nam. Các sản phẩm hiện có trong khu vực, bao gồm IMERG Final Run, IMERG Early, GSMaP_MVK_Gauge, PERSIANN_CCS, FY-4A và ảnh rađa, được sử dụng để đánh giá hiệu quả của phương pháp phân loại được đề xuất. Kết quả cho thấy mô hình phân loại mưa đề xuất sử dụng phương pháp LGBM vượt trội hơn so với các phương pháp khác. Kết quả cho thấy, sản phẩm phân loại mưa đề xuất đạt giá trị cao nhất đối theo các chỉ số đánh giá bao gồm: Xác suất phát hiện (POD), Chỉ số thành công (CSI), Chỉ số đe dọa công bằng (ETS) và Chỉ số kỹ năng Heidke (HSS). So với sản phẩm tham chiếu có hiệu suất cao nhất (GSMaP_MVK_Gauge), sản phẩm phân loại mưa đề xuất cho thấy mức cải thiện đáng kể khi tăng 20.0% về CSI, 16,67% về ETS, 38,89% về POD và 13,04% về HSS. Những kết quả này cho thấy tiềm năng của các mô hình học máy, đặc biệt là LGBM, trong việc nâng cao độ chính xác phát hiện mưa tại các khu vực có địa hình phức tạp, biến động khí hậu lớn và điều kiện thời tiết khắc nghiệt, qua đó cung cấp cơ sở quan trọng để cải thiện các hệ thống dự báo thời tiết.

Từ khoá: Phân loại lượng mưa; Học máy; LightGBM; Random Forest; Himawari-8; ERA-5.