VNU Journal of Science: Earth and Environmental Sciences
Journal homepage: https://js.vnu.edu.vn/EES

Original Article

Utilizing Multi-source Data and Machine Learning Models for Multi-class Rainfall Estimation in Central Vietnam

Vu Duy Dong[1], Nguyen Hung An[1,*], Nguyen Tien Phat[1], Nguyen Thi Huyen[1], Nguyen Thi Nhat Thanh[2]

[1]*Le Quy Don Technical University, 236 Hoang Quoc Viet, Nghia Do, Hanoi, Vietnam*
[2]*VNU University of Engineering and Technology, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam*

Received 12th June 2025
Revised 26th August 2025; Accepted 22nd October 2025

Abstract: This study proposes a multi-layer machine learning architecture for multi-class rainfall estimation in Central Vietnam. The input data includes Himawari-8 satellite imagery, ERA5 reanalysis data, ASTER DEM, and rain gauge observations. Four regional satellite-based rainfall products, including IMERG Final Run V07, IMERG Early Run V07, GSMaP_MVK_Gauge V08, and PERSIANN_CCS, were used as comparative datasets. Three machine learning algorithms, including Light Gradient Boosting Machine (LGBM), Extreme Gradient Boosting (XGB), and Random Forest (RF), were employed within the proposed architecture. Performance evaluation based on rain gauge observations showed that the LGBM-based rainfall product achieved the highest classification performance among the three surveyed products, with a Probability of Detection (POD) of 0.80, a Critical Success Index (CSI) of 0.54, a Matthews Correlation Coefficient (MCC) of 0.59, and a Symmetric Extremal Dependence Index (SEDI) of 0.58. Compared to the best-performing rainfall product (GSMaP_MVK_Gauge V08), the LGBM-based product demonstrated significant improvements in classification performance, with increases of 6.67% in POD, 8.00% in CSI, 11.32% in MCC, and 20.83% in SEDI. In terms of rainfall regression performance, the LGBM-based product also outperformed the other evaluated products, exhibiting the lowest errors, with a Mean Absolute Error (MAE) of 2.91 mm/h, Root Mean Square Error (RMSE) of 5.81 mm/h, and Mean Logarithmic Squared Error (MLSE) of 0.47.

Keywords: Rainfall estimation, Machine learning, LGBM, Random forest, Himawari-8, ERA5.

* Corresponding author.
E-mail address: hungan@lqdtu.edu.vn

https://doi.org/10.25073/2588-1094/vnuees.5325

1. Introduction

The central region of Vietnam frequently experiences extreme weather events, including prolonged heavy rainfall events that cause serious consequences. Therefore, developing a highly accurate rainfall dataset for this region is of great significance, not only to support economic development but also to enable the government to formulate effective response strategies to rainfall-induced extreme weather events [1, 2].

Machine learning (ML) has emerged as an effective approach for rainfall estimation, capable of processing large-scale, multisource datasets and modeling complex nonlinear relationships [3]. Min et al., (2018) employed the Random Forest (RF) algorithm to estimate summer rainfall across East Asia, using infrared (IR) band data from the Himawari-8 satellite combined with digital elevation model (DEM) data and numerical weather prediction (NWP) data as input features for model training. Rainfall data from the IMERG product were used as labels during the training process. Their proposed model outperformed the IMERG rainfall product, achieving a rain/no-rain classification accuracy (Acc) of 0.87, an MAE of 0.51 mm/h, and an RMSE of 2.0 mm/h [4]. Similarly, Putra et al. (2024) conducted rainfall estimation for six different regions in Indonesia—Bandar Lampung, Banjarmasin, Pontianak, Deli Serdang, Gorontalo, and Biak—using the XGB model. The input data consisted of brightness temperature (BT) from the Himawari-8 satellite's IR band 13 (10.4 μm), along with weather radar data, which were used as training features. Rainfall data from the IMERG Early Run product served as the training label, while measurements from Automated Weather Observing System (AWOS) rain gauge stations were used for validation. The results demonstrated that the XGB model outperformed the IMERG Early Run product in rainfall classification, with Acc values of 0.89, 0.91, 0.89, 0.90, 0.92, and 0.90 for the respective regions, and corresponding RMSE values of 2.75 mm/h, 2.57 mm/h, 3.08 mm/h, 2.64 mm/h, 1.85 mm/h, and 2.48 mm/h [5]. In the study by Giang et al., (2023), the LGBM model was employed to estimate daily-scale rainfall over South Korea using three satellite precipitation products: the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS), the Global Satellite Mapping of Precipitation (GSMaP), and the Tropical Rainfall Measuring Mission (TRMM). Rainfall observations from Automatic Weather Stations (AWS) were used as training labels, while DEM data and Euclidean Distance (ED) between stations served as supplementary inputs. An independent dataset from the Automated Synoptic Observation System (ASOS) was used for validation. The results indicated that the rainfall estimates produced by their proposed model outperformed the original satellite precipitation products, achieving a correlation coefficient (CC) of 0.944, an MAE of 1.18 mm/day, and an RMSE of 4.55 mm/day [6]. Nevertheless, accurately estimating heavy rain events in mountainous regions remains a significant challenge and limitation in these studies.

In Vietnam, achieving high-accuracy rainfall estimation from satellite data remains a significant challenge, particularly for heavy and extreme rainfall events in mountainous regions [1, 7, 8]. To improve the accuracy of multi-class rainfall estimation, especially for heavy rainfall events in Central Vietnam, we propose to apply the multi-layer ML technique in this study. The input dataset comprises satellite imagery from Himawari-8, ground-based rainfall observations from meteorological stations, and auxiliary data sources including ERA5 reanalysis and the ASTER Digital Elevation Model (ASTER DEM). The proposed model categorizes rainfall intensity into four distinct classes: weak rain, moderate rain, heavy rain, and very heavy rain [9, 10]. Such stratification is expected to improve the accuracy of both classification and overall rainfall estimation in detail. Furthermore, to address the issue of class imbalance in the input data—particularly the underrepresentation of very heavy rainfall samples compared to weak rain—two data augmentation techniques (the

Randomized Value-based Rainfall Augmentation (RVR) and Class Weighting (CW)) were applied for the two-class classification model and the four-class classification model, respectively.

The remainder of this paper is structured as follows. Section 2 describes the datasets and research methodology. Section 3 presents the results and evaluation. Section 4 provides the conclusions and future research directions.

2. Case study and Datasets

2.1. Case Study

The study area extends from Quang Binh to Da Nang. This region experiences a very high average annual rainfall, with the majority occurring from August to December. The terrain is predominantly mountainous, with elevation increasing significantly from east to west. As a result, rainfall in this area tends to be locally distributed, primarily concentrated in mountainous regions [11]. The climate is influenced by both the southwest and northeast monsoons [12]. Monthly rainfall statistics for the years 2019–2020 in the study area are presented in Figure 1.

2.2. Datasets

The data used in this study were collected from various sources. The datasets used for model training include satellite data from Himawari-8, provided by the Japan Meteorological Agency (JMA), with a spatial resolution ranging from 0.5 to 2 km and a temporal resolution of 10 minutes [4]. In this study, BT data extracted from 10 infrared bands of Himawari-8 served as the primary input features for model training. In addition, the ERA5 reanalysis dataset, developed by the European Centre for Medium-Range Weather Forecasts (ECMWF) [13], provides climate, meteorological, and geophysical variables for the region and was used as supplementary input to improve model accuracy [14] The ASTER DEM, developed by NASA with a spatial resolution of 30 meters, was also integrated as an additional input to enhance the model's performance [15].

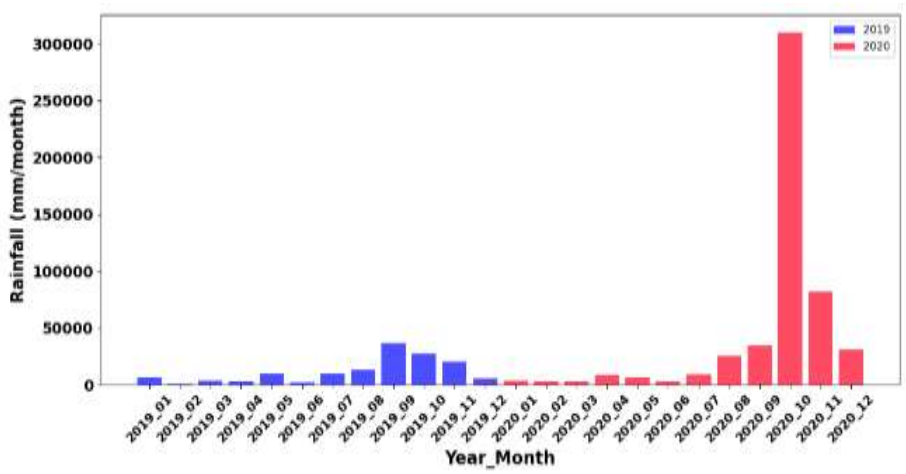

Figure 1. Monthly rainfall statistics for the study area in 2019–2020.

To evaluate the model's performance, four comparative precipitation products at the study area were used. Among them, the Integrated Multi-satellite Retrievals for GPM (IMERG) products, including IMERG Early Run (V07) and IMERG Final Run (V07), were developed by NASA and JAXA to monitor global precipitation. Both products offer data at a

spatial resolution of 0.1° × 0.1° and a temporal resolution of 30 minutes. IMERG Early Run V07 is a near-real-time product, available from 2023, with a latency of approximately six hours, whereas IMERG Final Run Version 07 is a post-processed product with a latency of about 3.5 months [16, 7]. The Global Satellite Mapping of Precipitation – Microwave–Infrared Combined Product with Kalman Filter (GSMaP_MVK_Gauge V08), developed by JAXA, provides hourly global precipitation estimates at a spatial resolution of 0.1° × 0.1° with a latency of approximately three days, covering the period from 1988 to 2022 [18]. Additionally, the Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks - Cloud Classification System (PERSIANN_CCS), developed by the Center for Hydrometeorology and Remote Sensing (CHRS), provides near-real-time precipitation estimates at a higher spatial resolution of 0.04° × 0.04° and a temporal resolution of one hour, with very low latency [19].

Rain gauge data provided by the National Centre for Hydro-Meteorological Network (NCN) during the period 2019–2020 were used as ground-truth labels for model evaluation and training. Specifically, rainfall data recorded at five time points of the day (00:00, 06:00, 12:00, 16:00, and 21:00) were used to construct the evaluation and reference dataset, while the remaining time points were used to build the training dataset.

3. Methodology

3.1. Proposed Architecture

The proposed architecture for rainfall estimation is illustrated in Figure 2. The input data, collected from multiple sources, were preprocessed to ensure consistency in both spatial and temporal resolutions while capturing the detailed variability of rainfall in the study area and optimizing computational resources. Specifically, the data were standardized to a spatial resolution of 4 km and a temporal resolution of 1 hour. For temporal resolutions, the Himawari-8 satellite data, originally available at 10-minute intervals, were aggregated by averaging six consecutive observations per hour to match the 1-hour temporal resolution of the ERA5 reanalysis data, ASTER DEM, and rain gauge measurements. For spatial resolutions, the ERA5 data, originally at a coarse resolution of 25 km, were resampled to 4 km using the nearest neighbor interpolation technique [20]. The Himawari-8 data, with an original resolution of 2 km, and the ASTER DEM data, originally at 30 m, were both resampled to 4 km using the average pooling technique [21].

The data input to model M1 is classified as either rain or no-rain events using a rainfall threshold of 0.1 mm/h. Model M2 then classifies the identified rain events into four intensity categories: weak (0.1–1.0 mm/h), moderate (1.0–5.0 mm/h), heavy (5.0–30.0 mm/h), and very heavy (> 30.0 mm/h). The division of rainfall into four classes with corresponding intensities is intended to reflect the characteristics of precipitation in the study area, where rapid changes in intensity and spatial extent occur due to variations in topography and complex, localized climatic conditions [22]. Models M3 to M6 are then used to perform regression on rainfall amounts for each corresponding rainfall category classified by M2. The outputs of these regression models are then combined based on the predicted rainfall classes to construct the final rainfall products. Subsequently, the proposed rainfall product was matching in both spatial and temporal resolution with the four comparative rainfall products, namely GSMaP_MVK_Gauge V08, IMERG Final Run V07, IMERG Early Run V07, and PERSIANN_CCS. The classification and regression performance of these rainfall products was then evaluated against rain gauge observations (ground truth) using the metrics detailed in Section 3.3.

To address data class imbalance in the classification models, the RVR technique is applied to the model M1, while a CW strategy is

used for the model M2. The RVR technique randomly augments rainfall samples based on intensity intervals. Specifically, the entire range of rainfall values recorded from rain gauge stations, ranging from 0.1 to 95.2 mm/h, was divided into 11 narrower sub-ranges, including: 0.1–1.0 mm/h, 1.0–2.0 mm/h, 2.0–3.5 mm/h, 3.5–5.0 mm/h, 5.0–8.0 mm/h, 8.0–12.0 mm/h, 12.0–20.0 mm/h, 20.0–30.0 mm/h, 30.0–40.0 mm/h, 40.0–50.0 mm/h, and > 50.0 mm/h. The objective of the RVR technique is to balance the distribution of samples within the moderate, heavy, and very heavy rain classes, thereby increasing the number of rainfall samples in these classes and reducing the disparity with the majority class (weak rain). For instance, the number of samples within each sub-range was randomly increased at different rates, with the criterion that sub-ranges with higher rainfall intensity values (fewer samples) were augmented at higher rates compared to those with lower rainfall intensity values (more samples). On the other hand, the CW strategy enhances the models' sensitivity to minority classes by assigning them higher weights during training. To enhance model performance and reduce computational complexity during training, feature selection and hyperparameter optimization techniques are employed for all ML models from M1 to M6 (refer to [9] for detail).

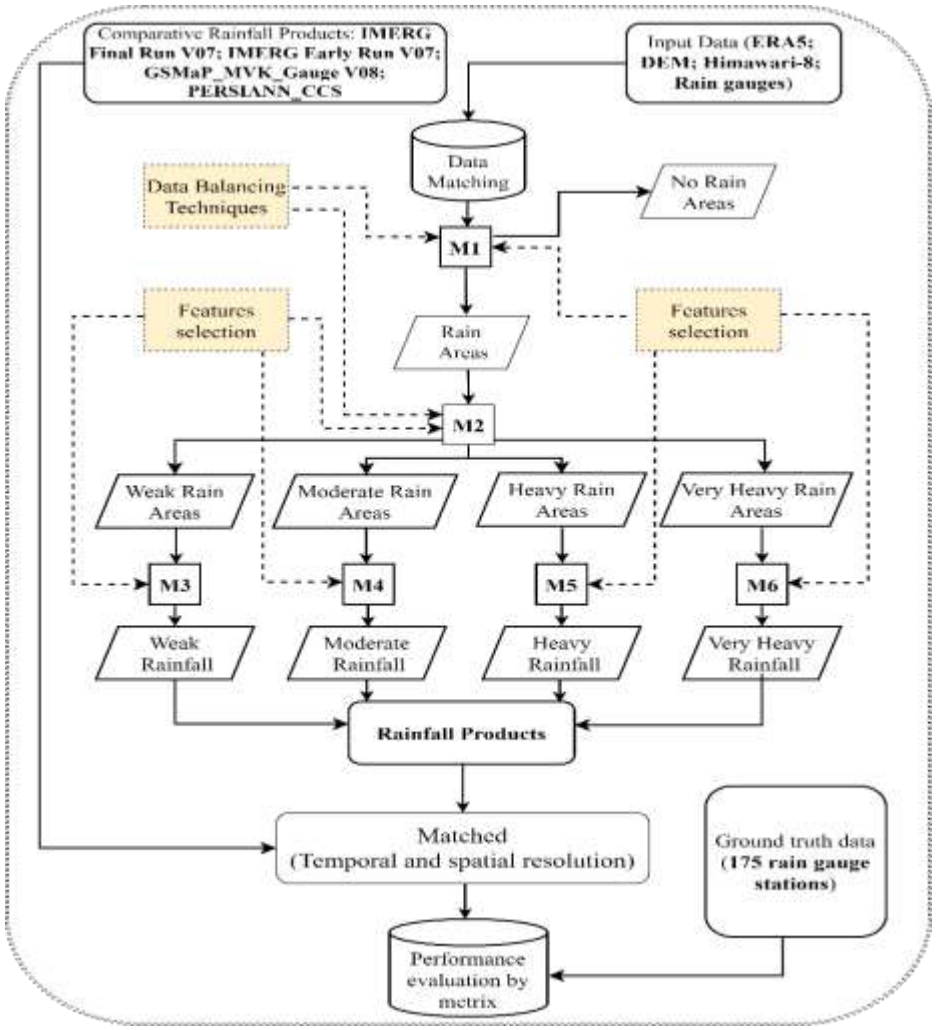

Figure 2. Proposed model architecture for rainfall estimation.

3.2. Machine Learning Algorithms

The RF algorithm, introduced by Breiman in 2001, builds an ensemble of decision trees using bootstrap aggregation [23]. The trees are trained in parallel on random subsets of the input data. The final predictions are made by majority voting for classification or averaging for regression [4]. During the training process, the decision trees are grown simultaneously in a level-wise manner. Deeper trees tend to provide more detailed learning and higher accuracy; however, they are also more prone to overfitting and require longer training time [24].

The XGB is an optimized implementation of the Gradient Boosting framework, which constructs an ensemble of weak learners, typically decision trees, to minimize a predefined loss function and enhance predictive performance [25]. XGB employs a level-wise tree growth strategy (breadth-first expansion), where each subsequent tree is built to correct the residuals (errors) from previous trees, thereby iteratively enhancing model accuracy [5].

The LGBM is an optimized Gradient Boosting framework that enhances efficiency and scalability. It grows trees leaf-wise, uses histogram-based learning to reduce memory usage, and applies Exclusive Feature Bundling (EFB) to lower feature dimensionality. Additionally, it employs Gradient-based One-Side Sampling (GOSS) to prioritize samples with large gradients, preserving key training information while reducing sample size [6].

3.3. Training and Evaluation

The input dataset is divided into training (80%) and testing (20%) subsets. This study applied 5-fold cross-validation on the training set using the Scikit-learn library to optimize the training process. Subsequently, the classification performance was evaluated using the F1-score, CSI, POD, FAR (False Alarm Ratio), BIAS (Bias Score), MCC [26], and SEDI [27], and the results are presented in Table 1. Meanwhile, the regression performance was assessed using metrics of the CC, MAE, RMSE, and MLSE [28], and is presented in Table 2. In these two tables, *TP*- denotes the number of rainfall samples correctly classified as rain; *FP* - represents the number of non-rain samples incorrectly classified as rain; *TN* - indicates the number of non-rain samples correctly classified as non-rain; and *FN* - refers to the number of rainfall samples incorrectly classified as non-rain; and *N* - the total number of samples; p_i, p_j represent the estimated and observed rainfall values, respectively.

Table 1. Basic classification metrics

Name	Equation	Range	Optimal
F1-score	$F1-score = (2TP)/(2TP+FP+FN)$	$[0,1]$	1
CSI	$CSI = TP/(TP+FP+FN)$	$[0,1]$	1
POD	$POD = TP/(TP+FN)$	$[0,1]$	1
MCC	$MCC = \frac{TP \times TN - FP \times FN}{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}$	$[-1,1]$	1
SEDI	$SEDI = \frac{\ln(F)-\ln(H)+\ln(1-H)-\ln(1-F)}{\ln(F)+\ln(H)+\ln(1-H)+\ln(1-F)}$ Here: $H = \frac{TP}{TP+FN}; F = \frac{FP}{FP+TN}$	[-1,1]	1
FAR	$FAR = FP/(FP+TP)$	$[0,\infty)$	0
BIAS	$BIAS = (TP+FP)/(TP+FN)$	$(-\infty,\infty)$	1

Table 2. Basic regression metrics

Name	Equation	Range	Optimal
MAE	$MAE = \frac{\sum \left\| p_i - p_j \right\|}{N}$	[0: inf]	0
RMSE	$RMSE = \sqrt{\frac{\sum \left(p_i - p_j \right)^2}{N}}$	[0: inf]	0
MLSE	$MLSE = \frac{\sum (\log(p_i + 1) - \log(p_j + 1))^2}{N}$	[0: inf]	0
CC	$CC = \frac{\sum (p_j - \bar{p_j})(p_i - \hat{p_i})}{\sqrt{\sum (p_j - \bar{p_j})^2} \sqrt{\sum (p_i - \hat{p_i})^2}}$	[-1: 1]	1

4. Result

4.1. Results of Features Selection

The original input data of the proposed architecture consists of 73 features: 55 derived from Himawari-8 BT data including (10 single IR bands and 45 band differences between them [29, 30], and 17 meteorological features from the ERA5 dataset, including K-Index (KX), Total Column Water (TCW), Total Column Water Vapor (TCWV), Convective Inhibition (CIN), Instantaneous Moisture Flux (IMF), Convective Available Potential Energy (CAPE), Slope of sub-gridscale orography (SLOR), Anisotropy of sub-gridscale orography (ISOR), Relative Humidity at 850 hPa, 500 hPa, and 250 hPa (R850, R500, R250), and zonal and meridional wind components at the same pressure levels (UWIND850, UWIND500, UWIND250, VWIND850, VWIND500, VWIND250), along with the ASTER DEM feature [31]. These features are ranked based on their importance using the RF Importance strategy [30], with a threshold of 0.02 to select the most relevant features for each model (M1–M6) and shown in Table 3.

Table 3. Selected features for the models M1, M2, M3, M4, M5, and M6

M1	M2	M3	M4	M5	M6
irb_b16	i2b_b16	UWIND850	TCWV	UWIND850	b11_b16
b11_b16	irb_b16	b14_b16	UWIND850	VWIND500	VWIND250
b10_irb	b11_i2b	UWIND500	KX	UWIND250	KX
b14_i2b	wvb_b14	CAPE	R850	VWIND850	R850
b14_b16	b14_b16	VWIND500	UWIND500	TCW	UWIND250
i2b_b16	UWIND850	i2b_b16	CAPE	KX	CAPE
b10_b11	VWIND850	irb_b16	VWIND850	irb_b16	b11_b14
UWIND850	R850	VWIND250	b14_b16	i2b_b16	VWIND850
R850	ISOR	TCW	ISOR	b14_i2b	TCWV
DEM	DEM	KX	DEM	b10_wvb	VWIND500

From Table 3, it can be observed that the Himawari-8 features, namely the BT differences in IR channels, reflect key physical properties of clouds, including cloud-top temperature, optical thickness, cloud-top height, and cloud water content [29]. Additionally, the ERA5 features provide supplementary information directly related to rainfall activity in the study area. Among them, R850 represents atmospheric moisture in the lower troposphere. TCW and TCWV indicate the total column water and water vapor content in the atmosphere. KX and CAPE represent atmospheric instability and vertical air motion. The UWIND and VWIND features (at 850, 500, and 250 hPa) describe the direction and intensity of wind, while ISOR and SLOR capture topographic characteristics such as slope and elevation [32]. The ASTER DEM provides detailed information on the elevation of grid points, which is relevant to the distribution and intensity of surface rainfall [33]. These features are directly associated with the formation and movement of rainfall and also reflect its characteristics on the surface.

4.2. Results of Data Augmentation

The number of rainfall samples across the 11 intervals within the four rainfall classes, before and after applying the RVR-based data augmentation technique, is presented in Table 4. As shown in Table 4, prior to data augmentation, the majority of rainfall samples were concentrated in the lower intensity intervals, and the number of samples gradually decreases as rainfall intensity increases. After performing RVR-based data augmentation, the distribution of samples within these three classes became more uniform across the entire class range, with a substantial improvement in the number of high-intensity rainfall samples. As a result, the disparity in sample counts between these three classes and the weak rain class was significantly reduced—particularly for the very heavy rain class, the ratio improved from approximately 1:43 (before augmentation) to around 1:3 (after augmentation).

Table 4. The number of rainfall samples class before and after augmentation using the RVR technique

Rainfall class	Value ranging (mm/h)	Rate of increase	Number of samples	
			Before	After
Weak	0.1 – 1.0	1.0	77,176	77,176
Moderate	1.0 – 2.0	1.0	21,285	21,285
	2.0 – 3.5	1.3	14,450	18,785
	3.5 – 5.0	1.8	9,765	17,577
Heavy	5.0 – 8.0	1.9	10,303	19,576
	8.0 – 12.0	2.5	6,462	16,155
	12.0 – 20.0	3.0	5,765	17,295
	20.0 – 30.0	5.0	2,655	13,275
Very heavy	30.0 – 40.0	7.0	1,030	7,210
	40.0 – 50.0	15.0	430	6,450
	> 50.0	20.0	338	6,760

4.3. Rain Classification Performance

In this section, models M1 and M2 are independently evaluated using the F1-score classification metric at first. After that, the classification performance of the final proposed rainfall products using the three algorithms (RF, XGB, and LGBM) is assessed and compared with the four comparative rainfall products, including IMERG Final Run V07, IMERG Early Run V07, GSMaP_MVK_Gauge V08, and PERSIANN_CCS, according to metrics of the POD, CSI, FAR, BIAS, SEDI, and MCC.

4.3.1. Two-class Rainfall Classification Results

The classification results of rain and no-rain events using model M1 with RF, XGB, and LGBM algorithms, in both cases without/with using the RVR-based balancing technique, are presented in Figure 3. As shown in Figure 3, the classification performance of model M1 improves after applying data balancing, particularly in terms of the F1-score for the rain class on the test set. Notably, the M1 model using LGBM achieves the highest F1-score of 0.76, slightly outperforming RF and XGB, both of which yield an F1-score of 0.75.

4.3.2. Four-classs Rainfall Classification Results

The four-class rainfall classification results of the M2 model are shown in Figure 4.

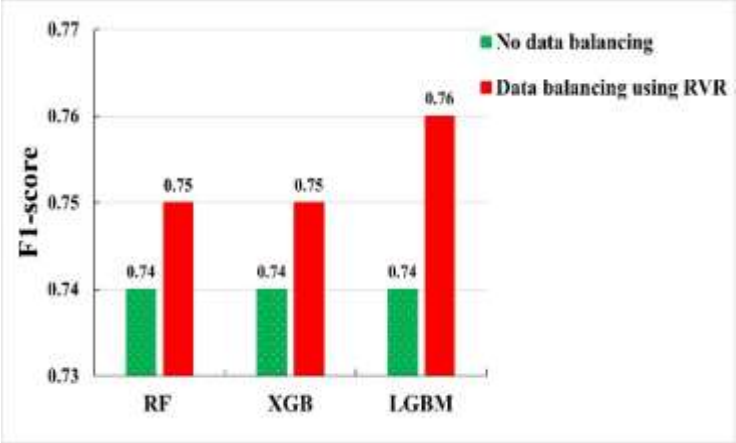

Figure 3. Two-class (rain and no-rain) classification results.

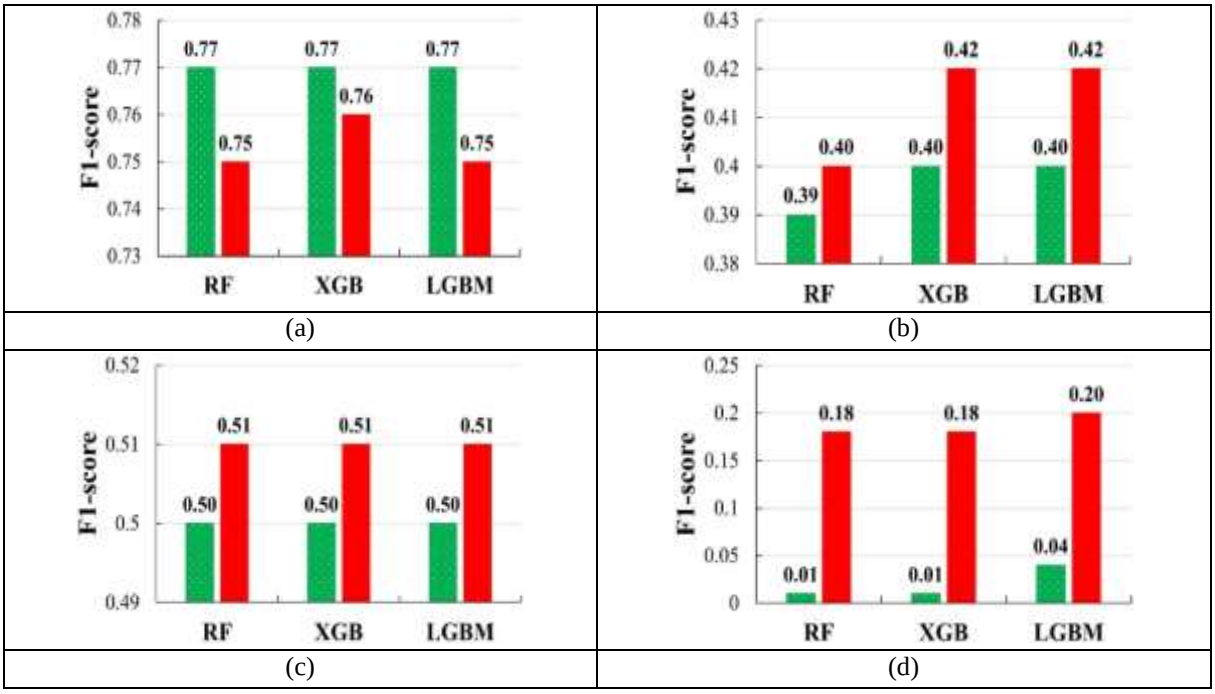

Figure 4. Four-class rain classification results: (a) Weak, (b) Moderate, (c) Heavy, (d) Very heavy.

The red and green colors represent the use of data balancing with CW and the absence of data balancing, respectively. As illustrated in this figure, the classification performance for the minority classes improved significantly after applying CW-based balancing techniques. In particular, the models were nearly incapable of classifying the very heavy rain events in the imbalanced scenario, with test F1-scores close to zero (Figure 4d). However, after balancing the data, performance in this class improved substantially, with test F1-scores reaching 0.20 for the LGBM model and 0.18 for both the XGB and RF models. Although data balancing helps significantly improve the classification performance for minority rain classes, it led to a slight decrease (~ 0.01) in the performance of the weak rain class (Figure 4a). However, this reduction is negligible, as the F1-score for the weak rain class remains relatively high, averaging 0.75.

4.3.3. Classification Performance of the Proposed Rain Classification Products

The performance of proposed rainfall products constructed from six models, M1 through M6, in which M1 applies the RVR technique and M2 utilizes the CW for data balancing, is evaluated. The proposed rainfall products based on LGBM, XGB, and RF are denoted as LGBM-RVR-CW, XGB-RVR-CW, and RF-RVR-CW, respectively, to highlight the data balancing techniques applied. The proposed rainfall products were matched with the comparative rainfall products to a common spatial and temporal resolution. A total of 3,332 rainfall maps were employed to evaluate their performance. The classification performance of these rainfall products was assessed using rainfall data from 175 independent rain gauge stations at five specific time intervals, based on metrics including POD, CSI, FAR, BIAS, SEDI, and MCC. The results are illustrated in Figure 5 and Figure 6.

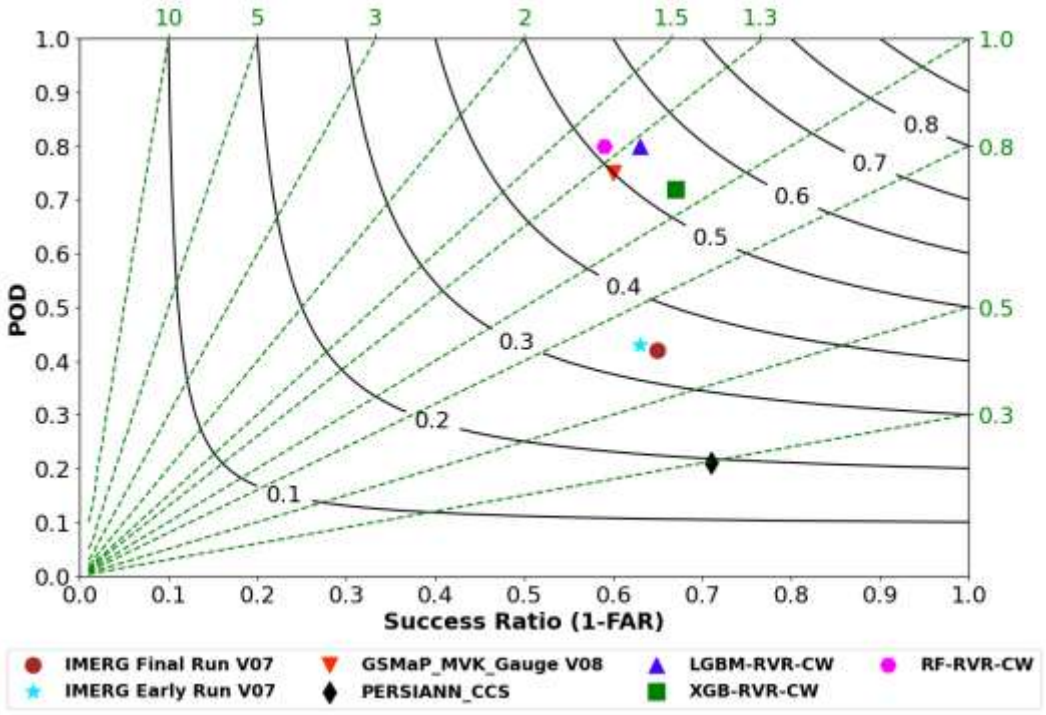

Figure 5. Overall rainfall classification performance of the proposed rainfall products, evaluated against ground-based rain gauge observations (ground truth). The green dashed line represents BIAS, while the black solid curve represents the CSI value.

Based on Figure 5, the performance of the rainfall products was evaluated using four metrics: POD, FAR, BIAS, and CSI. Specifically, for the POD, all three proposed rainfall products achieved higher values compared to the four comparative products. Among them, LGBM-RVR-CW and RF-RVR-CW obtained the highest POD value of 0.80, which is 11.11% higher than XGB-RVR-CW. Compared with the rainfall products, the POD of LGBM-RVR-CW and RF-RVR-CW was 6.67% higher than that of the best delayed product (GSMaP_MVK_Gauge V08) and 86.0% higher than that of the best near-real-time product (IMERG Early Run V07). These results show that LGBM-RVR-CW and RF-RVR-CW are the products with the highest rain/no-rain classification performance, and they positively affect the accuracy of subsequent rainfall intensity classification in the proposed rainfall products. In terms of the CSI metric, all three proposed products outperformed the four comparative rainfall products. Specifically, LGBM-RVR-CW achieved the highest CSI value of 0.54, followed by XGB-RVR-CW (0.53) and RF-RVR-CW (0.52). GSMaP_MVK_Gauge V08 had the highest CSI among the best delayed products but was still 8.0% lower than LGBM-RVR-CW, while IMERG Early Run V07, the best near-real-time rainfall product, was 54.29% lower than LGBM-RVR-CW.

Regarding the BIAS metric, XGB-RVR-CW achieved the best value of 1.07, followed by GSMaP_MVK_Gauge V08 (1.25). Although LGBM-RVR-CW and RF-RVR-CW showed slightly higher BIAS than the latter, they still performed better than the three other comparative products. In terms of the FAR, XGB-RVR-CW had the lowest value among the proposed models, at 0.33, followed by LGBM-RVR-CW (0.37) and RF-RVR-CW (0.41). Although PERSIANN_CCS had the lowest FAR value (0.29) among the products investigated, it exhibited the poorest CSI, POD, and BIAS values. Compared with the other comparative products, XGB-RVR-CW and LGBM-RVR-CW achieved FAR values of approximately 0.35, which are comparable to those of IMERG Final Run V07 and IMERG Early Run V07 but 12.5% better than GSMaP_MVK_Gauge V08.

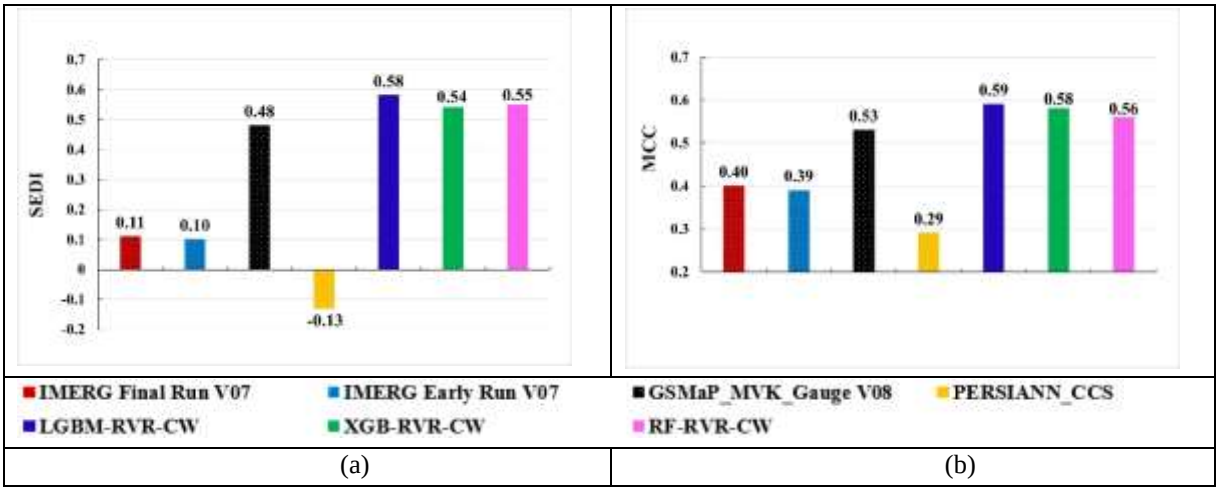

Figure 6. Rainfall classification performance of the proposed rainfall product: (a) SEDI, (b) MCC.

To evaluate the classification performance of the rainfall products in rare events (heavy and very heavy rainfall), the MCC and SEDI metrics were also used. The MCC reflects the overall accuracy of the model and is particularly effective for imbalanced datasets. In contrast, the SEDI focuses on assessing the ability to correctly classify rare rainfall samples within the

heavy and very heavy rain class. The results in Figure 6 indicate that the proposed rainfall products consistently outperform the comparative rainfall products in terms of these two classification metrics. Particularly, the LGBM-RVR-CW product achieves the highest SEDI and MCC values of 0.58 and 0.59, respectively. Compared with the comparative rainfall products, it is observed that LGBM-RVR-CW outperforms the best-performing reference product, GSMaP_MVK_Gauge V08, with improvements of 20.83% in SEDI and 11.32% in MCC.

From above analyses, it is concluded that the proposed product LGBM-RVR-CW has the highest rainfall classification performance among the seven investigateded rainfall products, especially effective for rare rainfall events with heavy and very heavy rainfall intensity.

4.4. Rainfall Regression Results

It can be seen from Figure 7a that the number of rainfall samples across the four rainfall classes is highly imbalanced, with the "very heavy rain" class being particularly underrepresented, accounting for only 1.2% (1798 samples). To evaluate the performance of the proposed rainfall products, the assessment process was carried out in two stages as follows.

In the first stage, the performance of the regression models M3 to M6, which use the LGBM, XGB, and RF algorithms, was independently evaluated for each rainfall class in Section 4.4.1.

In the second stage, the final proposed regression products were compared with the four regional rainfall products in terms of regression metrics in Section 4.4.2.

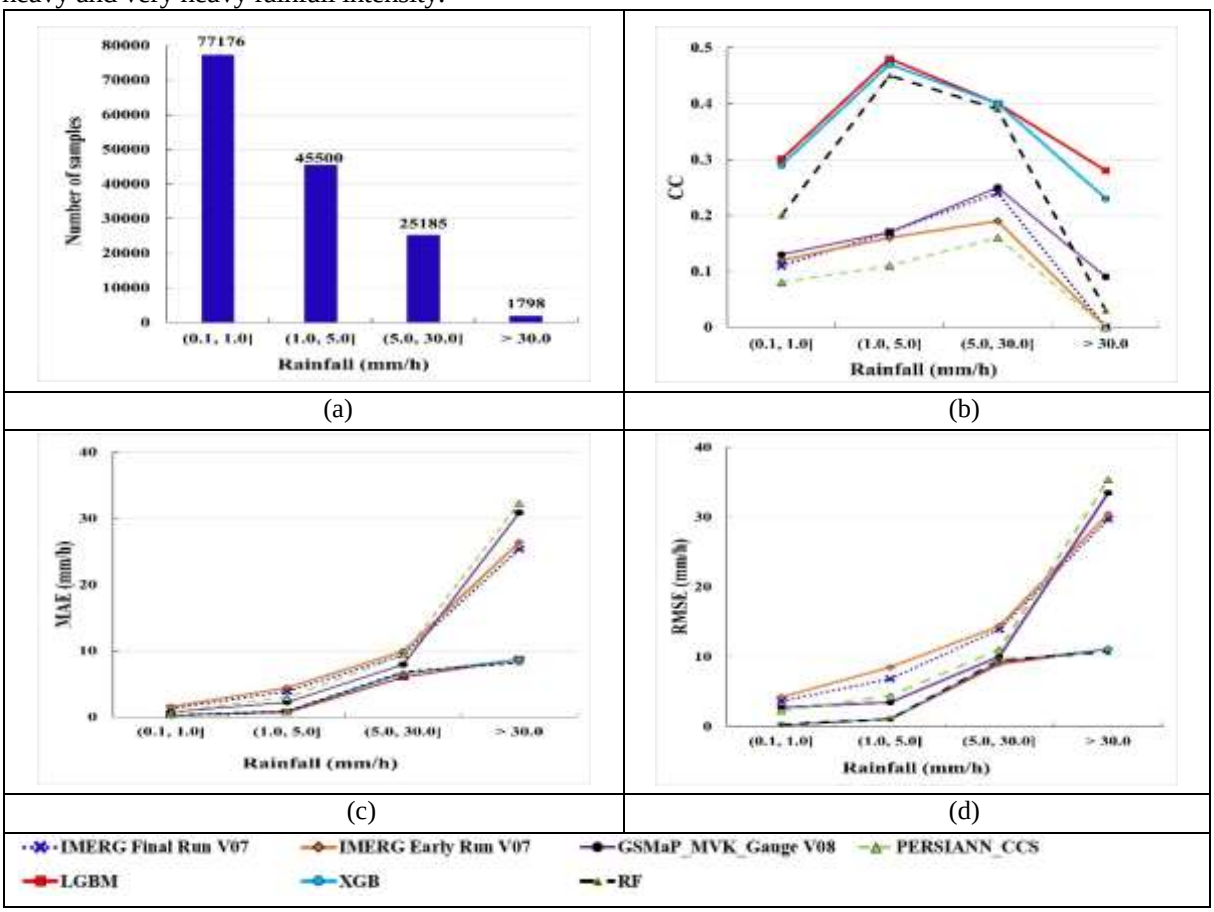

Figure 7. Rainfall regression performance for each rain class: (a) Number of samples per rain class, (b) CC, (c) MAE, (d) RMSE.

4.4.1. Regression Results for Each Rainfall Class

The regression performance for each rainfall class of the three proposed rainfall products was compared with the regression performance of the four regional comparative rainfall products using CC, MAE, and RMSE metrics, as illustrated in Figures 7b-7d. For the error metrics MAE (Figure 7c) and RMSE (Figure 7d), the three proposed rainfall products exhibit approximately similar values across all four-class of rainfall and consistently outperform the comparative rainfall products. Notably, for the very heavy rainfall class, these metrics show substantial improvements. Specifically, the MAE and RMSE of the three proposed products are approximately 8.70 mm/h and approximately 11.10 mm/h, respectively, representing improvements of approximately 65.69% and 62.60% compared to the best-performing delayed comparative product (IMERG Final Run V07). To the best-performing near real-time comparative product (IMERG Early Run V07), the improvements in MAE and RMSE are approximately 67.01% and 63.53%, respectively.

Regarding the CC values (Figure 7b), for the weak, moderate, and heavy rainfall classes, the three proposed rainfall products consistently outperform the comparative products. In the very heavy rainfall class, the proposed product LGBM-RVR-CW achieves the highest CC value among all evaluated products, reaching 0.28, followed by XGB-RVR-CW with a CC of 0.23. These results are higher than those of all comparative products. Specifically, the proposed products LGBM-RVR-CW and XGB-RVR-CW exhibit higher CC values compared to the best delayed comparative product, GSMaP_MVK_Gauge V08, by factors of 3.11 and 2.56, respectively. Meanwhile, the proposed RF-RVR-CW product records a CC of 0.03, which is lower than GSMaP_MVK_Gauge V08 but comparable to the other three comparative products.

4.4.2. Regression Performance of the Final Proposed Rainfall Products

The rainfall regression performance of the three proposed products, such as LGBM-RVR-CW, XGB-RVR-CW, and RF-RVR-CW, was compared against four comparative products using standard regression metrics: MAE, RMSE, MLSE, and CC, and are illustrated in Figure 8.

As shown in Figure 8, among the evaluated rainfall products, LGBM-RVR-CW demonstrates overall superior regression performance compared to the others. Specifically, among the three proposed products, LGBM-RVR-CW achieves an MAE of 2.91 mm/h (Figure 8a), an RMSE of 5.81 mm/h (Figure 8b), an MLSE of 0.47 (Figure 8c), and a CC of 0.38 (Figure 8d). LGBM-RVR-CW outperforms RF-RVR-CW and XGB-RVR-CW, with improvements of 24.81% and 42.38% in MAE, 10.62% and 32.99% in RMSE, and 29.85% and 43.37% in MLSE, respectively. Additionally, the CC of LGBM-RVR-CW also exceeds that of these two products by 5.56%.

In addition, when compared with the four comparative rainfall products, LGBM-RVR-CW also exhibits significant advantages. In terms of MAE (Figure 8a), it improves the delayed products IMERG Final Run V07 and GSMaP_MVK_Gauge V08 by 18.03% and 3.32%, respectively, and the near-real-time products PERSIANN_CCS and IMERG Early Run V07 by 4.90% and 26.32%, respectively. For RMSE (Figure 8b), LGBM-RVR-CW outperforms the two delayed products, IMERG Final Run V07 and GSMaP_MVK_Gauge V08, by 24.55% and 3.17%, respectively. Similarly, LGBM-RVR-CW outperforms the two near-real-time products, IMERG Early Run V07 and PERSIANN_CCS, by 31.73% and 7.93%, respectively. With respect to MSLE (Figure 8c), the proposed LGBM-RVR-CW product shows lower values, with improvements of 10.64% over IMERG Final Run V07 and 2.08% over GSMaP_MVK_Gauge V08, while also achieving improvements of 21.67% and 22.95% over the near-real-time products IMERG Early

Run V07 and PERSIANN_CCS, respectively. Regarding CC (Figure 8d), LGBM-RVR-CW shows a correlation coefficient 17.39% lower than IMERG Final Run V07 and 7.89% lower than IMERG Early Run V07, but equivalent to GSMaP_MVK_Gauge V08 and 15.15% higher than PERSIANN_CCS. When comparing the CC of the three proposed products, LGBM-RVR-CW outperforms XGB-RVR-CW and RF-RVR-CW by 5.56%. Accordingly, the proposed LGBM-RVR-CW rainfall product demonstrates the best rainfall regression performance among all evaluated rainfall products.

Based on the classification and regression performance of the evaluated rainfall products presented in Sections 4.3 and 4.4, the LGBM-RVR-CW product demonstrated the highest overall performance and was consequently selected as the final proposed rainfall product. This finding indicates the strong capability of the proposed model to accurately estimate rainfall distribution. However, accurately predicting extreme rainfall events (>30.0 mm/h) remains a challenging task.

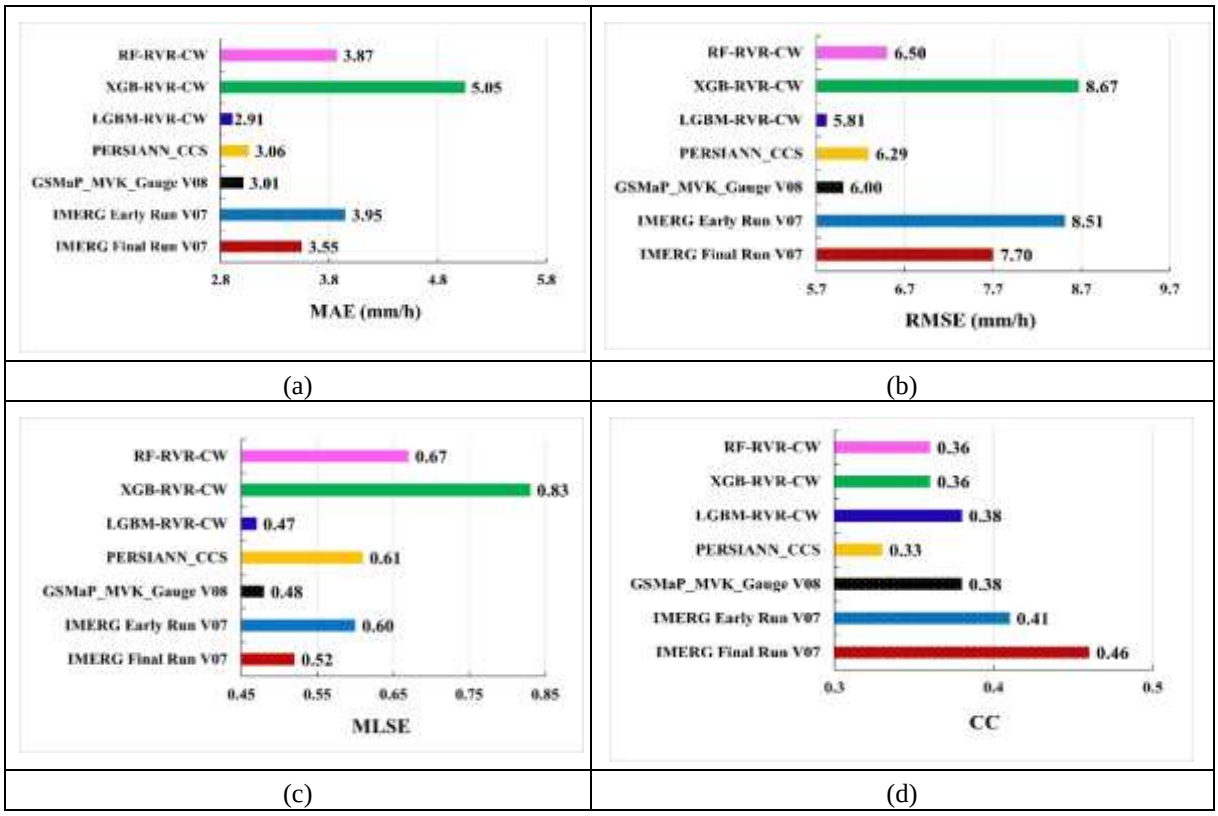

Figure 8. Rainfall estimation performance of rainfall products based on regression metrics: (a) MAE, (b) RMSE, (c) MLSE, (d) CC.

5. Conclusions

This study introduces a multi-layered machine learning architecture for enhanced rainfall estimation over Central Vietnam, leveraging three algorithms: RF, XGB, and LGBM. The model utilizes a combination of input data sources, including Himawari-8 satellite imagery, ERA5 reanalysis data, ASTER DEM, and ground-based rainfall observations. To address class imbalance, two data balancing techniques were employed: RVR for two

(rain/no-rain) classification and CW for four-class classification (weak/moderate/heavy/very heavy rain).

Three rainfall estimation products were proposed based on the integration of classification and regression models, which are RF-RVR-CW, XGB-RVR-CW, and LGBM-RVR-CW. Among these, the LGBM-RVR-CW product demonstrated the highest overall performance. When compared with four benchmark rainfall products, including IMERG Final Run V07, IMERG Early Run V07, GSMaP_MVK_Gauge V08, and PERSIANN_CCS, the LGBM-RVR-CW product consistently outperformed them. In terms of classification performance, the proposed product achieved substantial improvements compared to the best-performing comparative product, GSMaP_MVK_Gauge V08, with increases of 6.67% in POD, 8.00% in CSI, 11.32% in MCC, and 20.83% in SEDI. Regarding rainfall regression performance, LGBM-RVR-CW produced the lowest error values, with an MAE of 2.91 mm/h, an RMSE of 5.81 mm/h, and an MSLE of 0.47. For CC, the proposed product, LGBM-RVR-CW, achieved a value of 0.38, which is slightly lower than IMERG Final Run V07 (0.46) and IMERG Early Run V07 (0.41), but equal to GSMaP_MVK_Gauge V08 and higher than PERSIANN_CCS (0.33).

However, the prediction accuracy of the proposed products for extreme rainfall events (> 30.0 mm/h) remains a limitation. Future work may focus on incorporating advanced deep learning models to improve the prediction of very heavy rainfall.

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] L. Trinh, J. Matsumoto, T. N. Duc, M. Nodzu, T. Inoue, Evaluation of Satellite Precipitation Products Over Central Vietnam, Vol. 6, 2019, https://doi.org/10.1186/s40645-019-0297-7.

[2] T. N. Duc, T. Long, Future Rainfall Projections in Vietnam based on a CMIP6 Dynamical Downscaling Experiment, VNU Journal of Science: Earth and Environmental Sciences, Vol. 39, 2023, https://doi.org/10.25073/2588-1094/vnuees.4933.

[3] M. Kühnlein, T. Appelhans, B. Thies, T. Nauss, Improving The Accuracy of Rainfall Rates from Optical Satellite Sensors with Machine Learning — A Random Forests-Based Approach Applied to MSG Seviri, Remote Sens Environ, Vol. 141, 2014, pp. 129-143, https://doi.org/10.1016/j.rse.2013.10.026.

[4] M. Min et al., Estimating Summertime Precipitation from Himawari-8 and Global Forecast System Based on Machine Learning, IEEE Transactions on Geoscience and Remote Sensing, Vol. 57, 2019, pp. 2557–2570, https://doi.org/10.1109/TGRS.2018.2874950.

[5] M. Putra, M. Rosid, D. Handoko, High-Resolution Rainfall Estimation Using Ensemble Learning Techniques and Multisensor Data Integration, Sensors, Vol. 24, 2024, pp. 5030, https://doi.org/10.3390/s24155030.

[6] G. V. Nguyen, X. Le, L. N. Van, S. Jung, C. Choi, G. Lee, Evaluating the Performance of Light Gradient Boosting Machine in Merging Multiple Satellite Precipitation Products Over South Korea, 2023.

[7] M. I. Nodzu, J. Matsumoto, L. T. Tuan, T. N. Duc, Precipitation Estimation Performance by Global Satellite Mapping and Its Dependence on Wind Over Northern Vietnam, Prog Earth Planet Sci, Vol. 6, No. 1, 2019, pp. 58, https://doi.org/10.1186/s40645-019-0296-8.

[8] T. Cong, L. Quyen, G. N. Minh, L. Quyet, The Application of Himawari Satellite Data in Forecast and Warning of Rain and Thunderstorm, Vietnam Journal of Hydrometeorology, Vol. 719, 2020, pp. 1-13, https://doi.org/10.36335/VNJHM.2020(719).1-13.

[9] Y. Huang, Y. Bao, G. P. Petropoulos, Q. Lu, Y. Huo, F. Wang, Precipitation Estimation Using FY-4B/AGRI Satellite Data Based on Random Forest," Remote Sens (Basel), Vol. 16, No. 7, 2024, https://doi.org/ 10.3390/rs16071267.

[10] Y. Kim, S. Hong, Very Short-term Prediction of Weather Radar-Based Rainfall Distribution and Intensity Over the Korean Peninsula Using Convolutional Long Short-Term Memory Network, Asia Pac J Atmos Sci, Vol. 58, 2022, https://doi.org/10.1007/s13143-022-00269-2.

[11] T. C. Chen, J. D. Tsay, M. C. Yen, J. Matsumoto, Interannual Variation of the Late Fall Rainfall in Central Vietnam, J Clim, Vol. 25, 2012, pp. 392-413, https://doi.org/10.1175/JCLI-D-11-00068.1.

[12] V. Hang, N. Pham, H. P. Thanh, Evaluation of GSMaP Satellite Precipitation Over Central Vietnam in 2000-2010 Period and Correction Ability, VNU Journal of Science: Earth and Environmental Sciences, Vol. 34, 2018, https://doi.org/10.25073/2588-1094/vnuees.4341.

[13] B. M. Tuan, P. Yen, T. Mai, C. Ta Huu, N. Hoa, Distinct Characteristics of Early Summer Rainfall Over the Red River Delta and Southern Floodplain, VNU Journal of Science: Earth and Environmental Sciences, 2025, https://doi.org/10.25073/2588-1094/vnuees.5255.

[14] D. Lavers, A. Simmons, F. Vamborg, M. Rodwell, An Evaluation of ERA5 Precipitation for Climate Monitoring, Quarterly Journal of the Royal Meteorological Society, Vol. 148, 2022, https://doi.org/10.1002/qj.4351.

[15] 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, 2020, pp. 7214, https://doi.org/10.3390/s20247214.

[16] 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, https://doi.org/ 10.1007/s00704-025-05569-x.

[17] F. Gan, X. Cai, Y. Gao, X. Zhang, A Performance-Enhancement-oriented Evaluation System to Scrutinize the Changes From IMERG V06 Updated to V07 in Capturing and Presenting Typhoon Process, Atmos Res, vol. 326, 2025, pp. 108292, https://doi.org/10.1016/j.atmosres.2025.108292.

[18] C. Zhou, L. Zhou, J. Du, J. Yue, 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, 2024, pp. 102017, 2024, https://doi.org/10.1016/j.ejrh.2024.102017.

[19] P. Nguyen et al., The Persiann Family of Global Satellite Precipitation Data: A Review and Evaluation of Products, Hydrol Earth Syst Sci, Vol. 22, 2018, pp. 5801–5816, https://doi.org/10.5194/hess-22-5801-2018.

[20] A. Giordani, I. Cerenzia, T. Paccagnella, S. Di Sabatino, Sphera, A New Convection-Permitting Regional Reanalysis Over Italy: Improving the Description of Heavy Rainfall, 2022.

[21] D. Piyush, A. Varma, P. K. Pal, G. Liu, An Analysis of Rainfall Measurements over Different Spatio-Temporal Scales and Potential Implications for Uncertainty in Satellite Data Validation, Journal of the Meteorological Society of Japan, Vol. 90, 2012, https://doi.org/10.2151/jmsj.2012-401.

[22] R. K. Sumesh et al., Microphysical Aspects of Tropical Rainfall During Bright Band Events at Mid and High-altitude Regions Over Southern Western Ghats, India, Atmos Res, Vol. 227, 2019, pp. 178-197, https://doi.org/10.1016/j.atmosres.2019.05.002.

[23] L. Breiman, Random Forests, Mach Learn, Vol. 45, 2021, pp. 5-32, https://doi.org/10.1023/A:1010950718922.

[24] N. H. Pham, Q. Pham, T. Tran, Apply Machine Learning to Predict Saltwater Intrusion in the Ham Luong River, Ben Tre Province, VNU Journal of Science Earth and Environmental Sciences, Vol. 38, 2022, pp. 79-92, https://doi.org/10.25073/2588-1094/vnuees.4852.

[25] F. Baig, L. Ali, F. Ma, H. Chen, M. Sherif, How Accurate are the Machine Learning Models in Improving Monthly Rainfall Prediction in Hyper Arid Environment?, J Hydrol (Amst), Vol. 633, 2024, pp. 131040, https://doi.org/10.1016/j.jhydrol.2024.131040.

[26] S. Panigrahi, V. Vidyarthi, Assessing the Suitability of McKee et al., Drought Severity Classification Across India, Natural Hazards, Vol. 120, 2024, pp. 13543-13572, https://doi.org/10.1007/s11069-024-06762-3.

[27] K. Sharma, R. Ashrit, S. Kumar, A. Mitra, E. Rajagopal, Unified Model Rainfall Forecasts Over India During 2007–2018: Evaluating Extreme Rains Over Hilly Regions, Journal of Earth System Science, Vol. 130, 2021, https://doi.org/10.1007/s12040-021-01595-1.

[28] P. Krithik, M. Raghavan, K. Vasanth, Real Time Rainfall Prediction for Hyderabad Region using Machine learning Approach, 2021, pp. 1-6, https://doi.org/10.1109/iPACT52855.2021.9697017.

[29] H. Hirose, S. Shige, M. Yamamoto, A. Higuchi, High Temporal Rainfall Estimations from Himawari-8 Multiband Observations Using the

Random-Forest Machine-Learning MethodHigh 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, https://doi.org/10.2151/jmsj.2019-040.

[30] 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, 2024, pp. 69-77, https://doi.org/10.34238/tnu-jst.11346.

[31] Q. Jiang et al., Evaluation of the ERA5 Reanalysis Precipitation Dataset Over Chinese Mainland, J Hydrol (Amst), Vol. 595, 2021, pp. 125660, https://doi.org/10.1016/j.jhydrol.2020.125660.

[32] X. Li et al., Leveraging Machine Learning for Quantitative Precipitation Estimation from Fengyun-4 Geostationary Observations and Ground Meteorological Measurements, Atmos Meas Tech, Vol. 14, No. 11, 2021, pp. 7007-7023, 2021, https://doi.org/10.5194/amt-14-7007-2021.

[33] S. Ullah et al., GPM-Based Multitemporal Weighted Precipitation Analysis Using Gpm_Imergdf Product and Aster Dem in EDBF Algorithm, Remote Sens (Basel), Vol. 12, 3020, pp. 3162, https://doi.org/10.3390/rs12193162.