Developing a rainfall estimation model using XGBoost with Himawari-8/9 satellite and atmospheric data in East Java

Gede Gangga Wisnawa[1,2*], *Fajar Setiawan*[1,2]

[1]Department of Geomatics Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

[2]Marine Meteorological Station of Tanjung Perak, Badan Meteorologi Klimatologi dan Geofisika, Surabaya, Indonesia

Abstract. Accurate rainfall estimation in tropical regions is often hindered by non-linear atmospheric interactions and extreme data imbalance. This study develops a multi-stage precipitation estimation framework—comprising binary classification, multi-class classification, and regression—using an optimized Extreme Gradient Boosting (XGBoost) architecture. Applied to East Java, Indonesia, the model integrates Himawari-8/9 satellite brightness temperatures, global atmospheric indices, and high-resolution topography. To mitigate the dominance of non-rain events (91.6% of the dataset), Stratified Random Under-sampling (RUS) was employed. Hyperparameters were tuned using Bayesian Optimization and evaluated via 10-fold site-based cross-validation to prevent spatial data leakage. Results show that the optimized model significantly outperforms the baseline. In the regression stage, MAE and RMSE decreased by 21.5% and 23.0%, respectively, while the Pearson correlation coefficient improved by 43.1%. In classification, the Critical Success Index (CSI) rose by 16.2% for binary and 34.5% for multi-class stages, indicating an enhanced capability to detect rare rainfall events. Performance gains were most pronounced in mountainous regions, suggesting improved representation of orographic effects. The proposed hierarchical framework demonstrates potential as an effective approach for satellite-based rainfall estimation in topographically diverse tropical regions.

1 Introduction

East Java Province, Indonesia, is a region characterized by high vulnerability to disasters triggered by extreme weather conditions. Data from the National Disaster Management Agency of Indonesia (BNPB) reveals that floods accounted for approximately 48.8% of all natural disaster events in this region, culminating in material losses of IDR 275.5 billion in 2024 [1]. This situation highlights the urgent need for an accurate precipitation estimation system with high spatial and temporal resolution to support more effective mitigation and early warning efforts in East Java.

Numerical Weather Prediction (NWP) models are commonly used to generate weather forecasts. Although physically robust, NWP demands intensive computational resources and

*Corresponding author: gede.wisnawa@bmkg.go.id

is operationally expensive. Furthermore, NWP models face challenges in the parameterization of sub-grid scale processes, like cloud microphysics and turbulence, which are fundamental to understanding convective precipitation in tropical regions. These high costs and uncertainties at the physical process level create a distinct opportunity gap. This gap can be filled by data-driven empirical models leveraging machine learning (ML), which offer a potentially more efficient and accurate alternative [2].

The availability of high-resolution geostationary satellite data from Himawari-8/9, provides a strong foundation for developing precipitation estimation models. The Advanced Himawari Imager (AHI) sensor provides 16 observation channels with 10-minute temporal and 2-km spatial resolution, enabling the continuous capture of atmospheric dynamics. The utilization of predictor variables from this satellite is not merely based on arbitrary numerical values, they represent physically relevant atmospheric processes. For example, low Brightness Temperature (BT) values in Band 13 (10.4 μm) indicate high and cold cloud tops, which are associated with strong convective activity. Similarly, Brightness Temperature Difference (BTD) between channels offers additional information regarding cloud phase, particle size, and convective growth rates [3]. This combination of information establishes satellite infrared data as a valid predictor for precipitation, even though the relationship between satellite signals and precipitation is indirect and highly non-linear [4].

Algorithm selection is crucial given the complexity of remote sensing data and the variability of precipitation in tropical regions. Conventional ML models, such as Decision Trees and Support Vector Machines, generally struggle to accommodate high-dimensional, non-linear data characteristics influenced by complex feature interactions [5] In this context, ensemble methods offer more stable and accurate performance [6]. XGBoost was selected due to its numerous advantages, including high scalability, the ability to handle complex non-linearities, and built-in regularization (L1/L2) that effectively reduces overfitting [7]. Moreover, various meteorological studies have demonstrated that XGBoost consistently outperforms classical ML models in satellite-based precipitation estimation tasks [4,8]. Nevertheless, studies indicate that XGBoost's performance is highly sensitive to its hyperparameter configuration [4]. Therefore, the model's success depends not only on the algorithm itself but also on an appropriate hyperparameter optimization strategy. Accordingly, this study aims to develop a 3-hourly precipitation estimation model for the East Java region using XGBoost by integrating Himawari-8/9 data with atmospheric and topographic variables. The study will also evaluate the result of hyperparameter optimization on the model's accuracy, error patterns, and its capability to detect operationally significant heavy rainfall events.

2. Materials and methods

This section details the data sources, pre-processing stages, and analytical methods employed in this study. The process commences with the determination of the study area and the collection of primary data, followed by missing data handling, temporal and spatial resolution alignment, and data resampling to balance the occurrences of rain and no-rain events. Subsequently, model design, hyperparameter tuning, and evaluation of the results were conducted to obtain an accurate and reliable analysis.

2.1 Study area

The study area is East Java Province, Indonesia, geographically bounded by 110°–115° E longitude and 5°–9° S latitude. This region features complex topography, extending from coastal lowlands to highland areas, and is highly vulnerable to hydrometeorological hazards. The ground-truth data consist of 3-hourly precipitation accumulation, aggregated from 10-

minute observations from a network of 79 Automatic Weather Stations and Automatic Rain Gauges (AWS/ARG). This network is operated by the Meteorology, Climatology, and Geophysics Agency of Indonesia (BMKG), with provides data covering the period from 2020 to 2024

2.2 Data

This study utilizes a multi-scale and multi-source dataset compiled to capture patterns for generating accurate precipitation predictions. The predictor variables consist of Himawari-8/9 satellite observations (Brightness Temperature/BT and Brightness Temperature Difference /BTD), global and regional atmospheric indices including large-scale indices such as Sea Surface Temperature Anomaly (SSTA) Nino 3.4 and Dipole Mode Index (DMI) from The Bureau of Meteorology (BoM), and regional indices such as Western North Pacific Monsoon Index (WNPMI) and Australian Summer Monsoon Index (AUSMI) from European Centre for Medium-Range Weather Forecasts (ECMWF), and a Digital Elevation Model (DEM) from United States Geological Survey (USGS).

Table 1. Summary of Target and Predictor Variables.

Variable Name	Source	Spatial Resolution	Temporal Resolution	Purpose
Precipitation	BMKG (AWS/ARG)	Point (79 station)	10 min	(Target)
BT (Band 7-16)	JMA (via BMKG)	2 km x 2 km	10 min	Predictor
BTD	Derived features from BT	2 km x 2 km	10 min	Predictor
SSTA Nino 3.4	BoM	Global	7 days	Predictor
Dipole Mode Index (DMI)	BoM	Global	7 Hari	Predictor
WNPMI	ECMWF ERA5	Regional	3 Jam	Predictor
AUSMI	ECMWF ERA5	Regional	3 Jam	Predictor
Digital Elevation Model (DEM)	GMTED2010	500 m	Statis	Predictor
Julian Date	calculated	-	day	Predictor

Brightness Temperature Difference (BTD) data, which represent the difference in BT values between two infrared (IR) channels. These BTD features, as shown in Table 2, can provide additional information regarding cloud microphysical characteristics, which helps to differentiate cloud types and identify clouds with a higher precipitation potential [3, 8]. Specifically, this study using several key BTD combinations to capture distinct atmospheric

and cloud properties. For example, the difference value between Band-8 and Band-14 is used to assess upper tropospheric water vapor content. Other combinations, such as Band-7 – Band-14 and Band-7 – Band-10, provide information on cloud phase and cloud thickness, respectively. Critically for precipitation analysis, Band-11 and Band-14 difference helps detect low cloud growth, while Band-14 and Band-15 difference is used to identify the development of convective clouds.

Temporal and spatial resolution alignment was performed to ensure all input variables possessed a uniform data structure prior to model training. As the estimation target is on a 3-hourly scale, data with a finer temporal resolution, such as the 10-minute BMKG observed rainfall and Himawari-8/9 BT/BTD, were harmonized. Specifically, rainfall was accumulated into 3-hour totals, while the BT/BTD values were selected at the timestamp nearest to the target interval (e.g., 00:00, 03:00, 06:00 UTC, up to 21:00 UTC). Conversely, data with a coarser duration, such as the weekly Sea Surface Temperature Anomaly (SSTA) of Nino 3.4 and Dipole Mode Index (DMI), were applied constantly across all 3-hour slots within that specific week. Time variables, like the Julian Date, were calculated directly from each sample's timestamp. Spatially, gridded data, including Himawari-8/9 and the DEM, were extracted at the station coordinate points using the nearest neighbor method to ensure an accurate local representation.

2.3 Missing value handling

Data quality control was performed by applying a data completeness threshold. Observation stations with missing data rates exceeding 25% were excluded from the model development process to prevent potential bias caused by excessive data gaps. Incomplete data points for the remaining stations were omitted to ensure the training set consisted solely of reliable and valid observations. This rigorous filtering process maintains the overall integrity of the analysis while limiting performance degradation associated with input data inconsistencies.

2.4 Data resampling

The distribution of rainfall data as the target variable in this study exhibits a significant imbalance, where the frequency of no-rain events vastly outnumbers that of rain events. This condition poses a risk of biasing the model toward the majority class, thereby compromising its accuracy in detecting lower-frequency events. To handle this disparity, the study utilizes Random Under-sampling (RUS) with stratification specifically targeting the majority class. Unlike oversampling techniques such as SMOTE, which generate synthetic data and risk introducing noise or non-physical weather patterns, RUS operates by randomly reducing the number of samples in the majority class (CR) until its proportion is balanced with the minority class. This approach aims to proportionally reduce the dominance of the majority data without discarding critical patterns associated with global climate variability and monsoon systems. By utilizing stratification, the majority data is down-sampled evenly across each climate phase, ensuring that the representative characteristics of each phase are preserved [9]. Consequently, the RUS process yields a more proportional inter-class distribution prior to the application of augmentation techniques on the minority class.

2.5 Model design

The methodology incorporates Extreme Gradient Boosting (XGBoost), a sequential ensemble algorithm that improves predictive accuracy by iteratively generating decision trees to address the deficiencies of prior iterations. Model performance is optimized by

minimizing an objective function that integrates a loss function with a regularization term, ensuring high accuracy while effectively mitigating overfitting [10]. Regularization is achieved by penalizing the number of leaves and leaf weights within each tree; this enhances XGBoost's generalization capabilities, particularly for large-scale and sparse datasets through its sparsity-aware learning mechanism. In the context of spatiotemporal rainfall modeling, computational efficiency is further augmented by histogram-based algorithms and parallel processing support, which accelerate the identification of optimal splits for each feature.

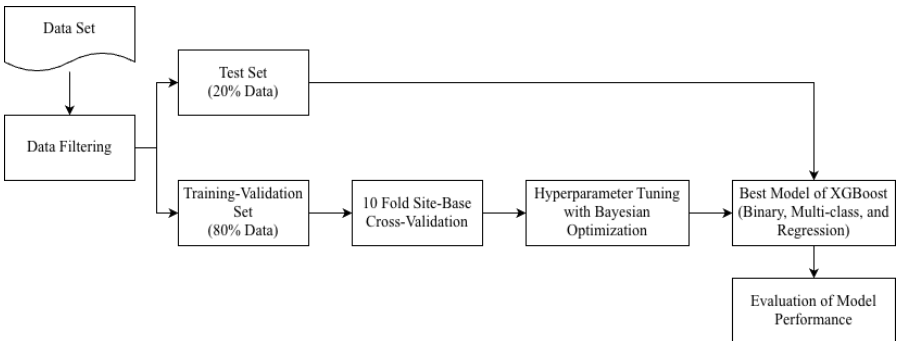

Fig. 1. Schematic diagram of the model development framework.

The proposed framework is designed as a multi-stage XGBoost architecture comprising three distinct phases: binary classification, multi-classification, and regression. This structure aims to generate more accurate and informative precipitation estimates. In the first phase, a binary classification model is utilized to discriminate between rain and no-rain events based on a specific precipitation threshold. The second phase subsequently focuses exclusively on samples identified as 'rain,' employing multi-classification to categorize rainfall intensity (e.g., light, moderate, and heavy). Finally, the third phase applies a regression model solely to the rain samples to estimate the precipitation quantity. By isolating the positive rain instances, this hierarchical strategy effectively mitigates the bias typically introduced by the zero-inflated nature of precipitation data, allowing the model to focus its learning capacity exclusively on the complex variability of rainfall intensities [11].

2.6 Hyperparameter optimization

Model performance is critically dependent on the selection of appropriate hyperparameters [4]. This study examines several pivotal hyperparameters, including: *learning_rate* (0.01–1) to control the weight update step size; *max_depth* (0–20) to define the maximum tree depth; *n_estimators* (100–1000) representing the total number of trees in the ensemble; *subsample* (0.1–1) to determine the fraction of data used per tree; *min_child_weight* (0.1–2) to set the minimum weight required for a node split; *gamma* was utilized (set between 0 and 1) to specify the minimum loss decrease needed to permit further partitioning , while *colsample_bytree* (spanning 0.1 to 1) was adjusted to manage the ratio of available features included in each sequential tree. These parameter ranges were utilized in the search for the optimal configuration across the XGBoost models (binary classification, multi-classification, and regression).

Hyperparameter optimization was conducted using Bayesian Optimization, which efficiently explores the parameter space by probabilistically modeling combinations likely to enhance model performance [7]. Performance evaluation was carried out using 10-Fold

Site-Based Cross-Validation (CVSI), a validation strategy that partitions data according to station locations. This approach prevents data leakage, which can occur when data from a single station appears simultaneously in both training and validation sets across different timestamps. Consequently, CVSI ensures that the model is rigorously evaluated based on its spatial generalization capability [8]. The research dataset was divided into 80% for training and validation, and 20% as an unseen test set. All reported evaluation results are derived from models using the optimal hyperparameters obtained during this validation process.

2.7 Evaluation of model result

Several statistical indicators were used to analyze the effectiveness of the regression stage. Specifically, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) served as the primary tools for quantifying the extent of prediction inaccuracies. Furthermore, the degree of linear agreement and the directional relationship between predicted outputs and observed data were evaluated through the Pearson Correlation Coefficient (R).

$$RMSE = \sqrt{\frac{1}{n}\sum_{n}^{n}(\hat{y}_i - y_i)^2} \tag{1}$$

$$MAE = \frac{1}{n}\sum_{n}^{n}|\hat{y}_i - y_i| \tag{2}$$

$$R = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \cdot \sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{3}$$

The developed XGBoost model also generates multi-classification predictions categorized into four rainfall intensity classes: No Rain (< 0.1 mm/3h), Light Rain (0.1–10 mm/3h), Moderate Rain (10–20 mm/3h), and Heavy Rain (> 20 mm/3h). Classification performance was evaluated via a confusion matrix, enabling the calculation of overall accuracy (ACC). However, as accuracy can be inflated by the dominance of True Negatives, this study also utilized the Critical Success Index (CSI). CSI provides a more objective assessment of rare events and is therefore considered the most representative metric for evaluating the model's capability to accurately detect rainfall occurrences.

$$Accuracy = \frac{Hits + True\ Negative}{Hits + False\ Alarm + Miss + True\ Negative} \tag{4}$$

$$CSI = \frac{Hits}{Hits + False\ Alarm + Miss} \tag{5}$$

3. Result and discussion

A comprehensive analysis of the data's inherent properties is a mandatory prerequisite to ensure that the subsequent model construction phase is built upon a valid understanding of the dataset's complexities. This analysis serves to validate methodological assumptions and

identify specific modeling challenges, particularly those arising from the data's inherent class imbalance and the non-linear relationships among variables.

3.1 Data distribution and correlation

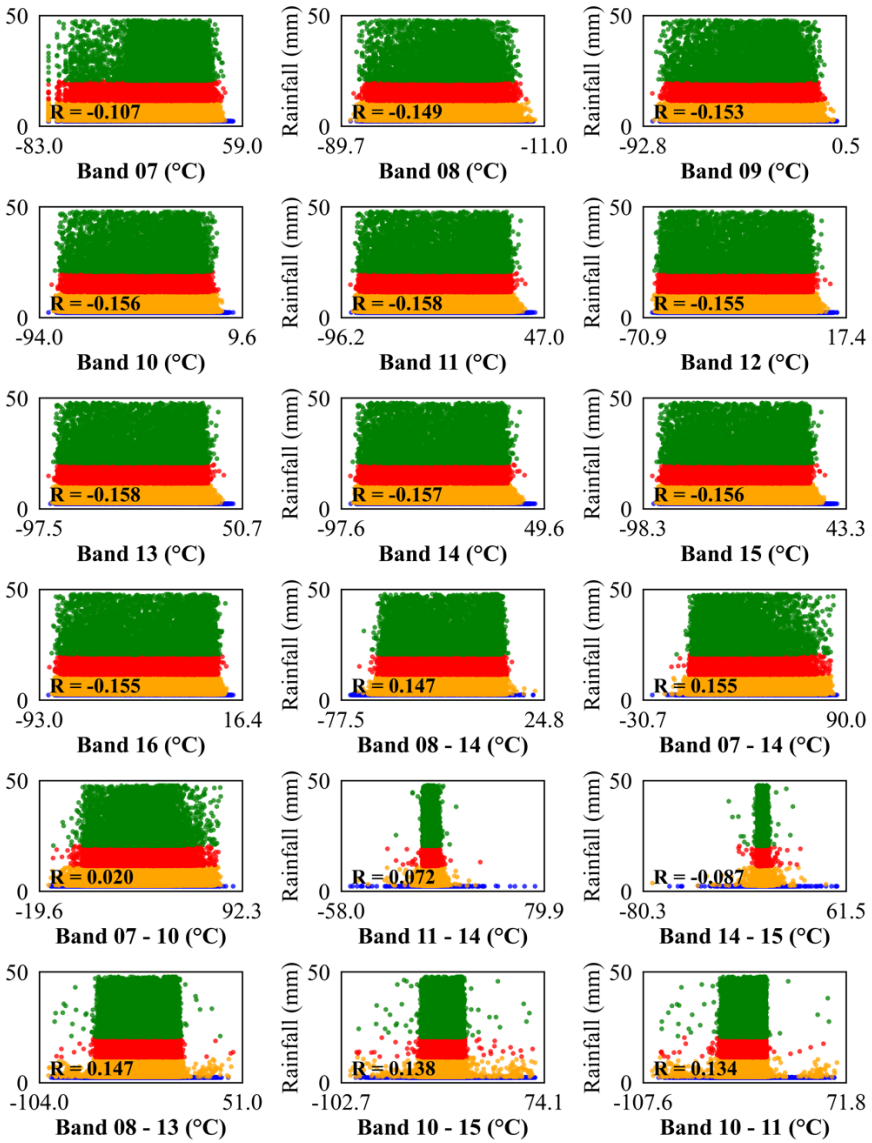

Fig. 2. Scatter plots illustrating the relationship between Himawari-8/9 predictors (BT and BTD) against observed 3-hourly rainfall intensity, overlayed with linear regression lines and Pearson correlation coefficients (*R*).

Scatterplots illustrating the relationship between various BT of Himawari channels and BTD against rainfall exhibit a highly dispersed pattern (Fig. 2). No discernible strong linear correlation (*R*) was identified. The Pearson correlation coefficients (*R*) for all BT channels

and BTD against surface rainfall range from -0.158 to 0.155. The majority of data points are clustered near zero rainfall, with the variance widening as predictor values increase. Furthermore, Fig. 3 reveals that similar patterns are observed in the auxiliary data, including climate indices and topographic parameters. Relationships between these predictors and the target precipitation also demonstrate substantial scatter and lack a distinct linear trend. These findings have important implications for model selection. The lack of linear correlation indicates that linear regression cannot effectively handle the complex dynamics of precipitation. Therefore, the use of XGBoost is validated. Unlike traditional models, XGBoost is designed to handle complex, non-linear, and multivariate relationships.

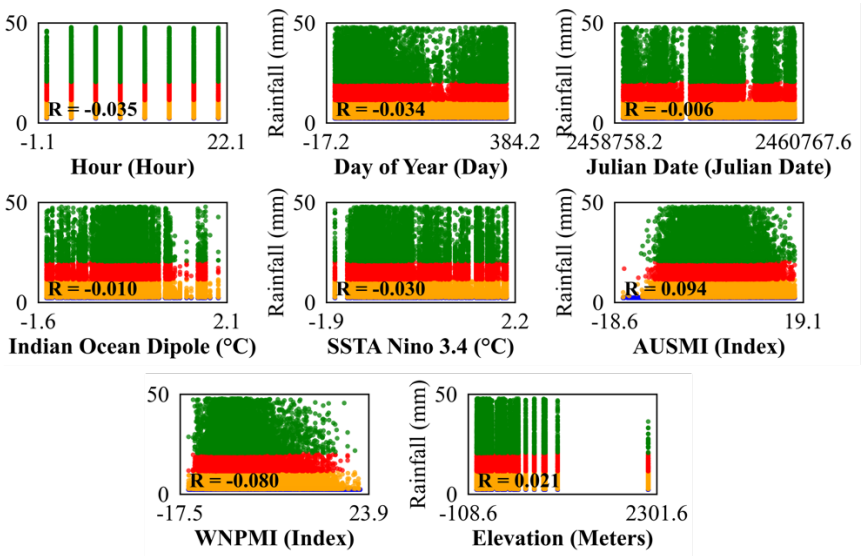

Fig. 3. Scatter plot matrix of auxiliary predictors (Time, Climate Indices, and Elevation) against observed 3-hourly rainfall intensity, overlayed with linear regression lines and correlation coefficients (*R*).

Beyond the challenge of non-linearity, the dataset also exhibits extreme class imbalance. Fig.4 illustrates the initial probability mass function (PMF) of the data from 71 observation stations following the filtering process. As evidenced in the figure, the Clear (CR) class overwhelmingly dominates the dataset, comprising 693,572 samples (91.6%). Conversely, rainfall events constitute the minority classes, with Light Rain (HR) accounting for 50,731 samples (6.7%), followed by Moderate Rain (HS) with 6,815 samples (0.9%), and Heavy Rain (HL) with a mere 6,057 samples (0.8%). Visually, the distribution exhibits a sharp leptokurtic pattern near 0 mm, confirming the prevalence of non-rainy conditions. In contrast, the probability curve for Heavy Rain (HL) appears nearly flat (platykurtic), reflecting the scarcity of extreme events. This severe imbalance, compounded by the distribution overlap between classes, poses a significant risk of model bias, where the algorithm may favor the majority class (CR) and fail to detect extreme precipitation events.

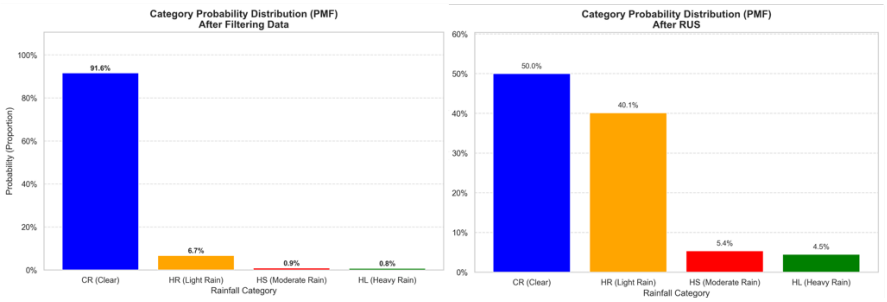

Fig. 4. Comparison of rainfall category probability distributions before (left) and after (right) applying the Stratified Random Under-sampling (RUS) technique.

To mitigate this issue, Stratified Random Under-sampling (RUS) was implemented. This method downsamples the majority class (CR) while preserving the climatological structure associated with IOD, ENSO phases, and seasonal variability. The under-sampling process targeted a reduction of the CR class to approximately 63,603 samples, aligning it with the total accumulated count of rainfall samples (HR + HS + HL). The outcome of the RUS application is presented in Figure 4.4. As shown, the proportion of the CR class was successfully adjusted to 50.0%, achieving a balanced 1:1 ratio between rain and no-rain events. However, it is crucial to note that an internal imbalance within the rainfall categories persists. Light Rain (HR) continues to dominate the rain category at 40.1%, far exceeding HS (5.4%) and HL (4.5%). This residual imbalance underscores that model evaluation cannot rely solely on Accuracy. Consequently, metrics sensitive to minority classes, such as the Critical Success Index (CSI), are essential for a robust performance assessment [11].

3.2 Hyperparameter optimization result

The hyperparameter optimization results presented in Table 2 reveal that each modeling stage (binary, multi-classification, and regression) requires a distinct configuration tailored to its specific function within the multi-stage framework. For the binary classification stage, a notably high *max_depth* (17) and relatively high *min_child_weight* indicate the necessity for a complex tree structure to effectively separate the initial two classes (rain and no-rain). This complexity is primarily driven by the overwhelming dominance of the Clear (CR) class at 0 mm rainfall [8]. Outputs identified as no-rain are immediately assigned to the CR class with a value of 0 mm and undergo no further processing. Conversely, samples detected as "rain" are forwarded to the multi-classification stage.

In multi-calssification model, the optimal configuration is more moderate, characterized by a lower *max_depth* (7), a higher *colsample_bytree* (0.76), and a substantial *min_child_weight*. These settings suggest that the model must constrain its complexity to prevent overfitting, particularly given the significant distributional overlap between HR, HS, and HL rainfall classes. Subsequently, the regression stage designed to estimate the specific rainfall magnitude for samples labeled HR, HS, or HL yielded a lighter hyperparameter combination, such as significantly fewer *n_estimators* (117) and a higher *learning_rate* (0.33). This implies that the regression task operates more efficiently on continuous value patterns and does not require the extensive tree complexity demanded by the classification stages. Overall, these configurational variations are consistent with the hierarchical nature of the proposed modeling framework, where each stage addresses a distinct statistical challenge.

Table 2. Optimal hyperparameter configurations for binary classification, multiclass classification, and regression tasks.

Hyperparameter	Best Hyperparameter of XGBoost		
	Binary	**Multi-classification**	**Regression**
colsample_bytree	0.23	0.76	0.57
gamma	0.63	0.36	0.08
learning_rate	0.01	0.01	0.33
max_depth	17.00	7.00	6.00
min_child_weight	3.82	8.47	1.17
n_estimator	426.00	354.00	117.00
subsample	0.66	0.77	0.98

3.3 Regression model performance evaluation

Quantitative evidence from Table 3 illustrates how the optimized XGBoost architecture consistently outperforms the default configuration across all evaluation criteria. The Mean Absolute Error (MAE) decreased by 21.5%, dropping from 3.93 to 3.08 mm/3hours, indicating a reduction in the model's average absolute error. A more substantial improvement was observed in the Root Mean Square Error (RMSE), which fell by 23.0% from 8.66 to 6.67 mm/3hours. This performance boost signals that optimization effectively mitigates large prediction errors. The sharper decline in RMSE compared to MAE suggests that the optimization successfully suppressed errors in predicting rainfall events, for example cases where the model predicts 0 mm during heavy observed rainfall or the opposite condition. Furthermore, the Pearson correlation coefficient (R) increased by 43.1%, rising from 0.33 to 0.48. This represents an improvement in the linear relationship between predictions and observations, shifting from a weak category to a moderate one.

Table 3. Performance comparison between the optimized and non-optimized (default) XGBoost models across regression, binary, and multi-classification stages.

Scheme	Regression			Binary		Multi-classification	
	RMSE	**MAE**	**R**	**ACC**	**CSI**	**ACC**	**CSI**
Optimized	6.68	3.08	0.48	0.76	0.59	0.90	0.49
Non-Optimized	8.66	3.92	0.33	0.72	0.51	0.84	0.37
Improvement (%)	-23.0	-21.5	43.1	5.5	16.2	6.8	34.5

These improvements align with the spatial findings illustrated in Fig. 5, which reveal significant changes in the distribution of regression performance across observation stations in East Java. In the default model, the MAE map displays extensive areas of high error, particularly in the central and southern regions dominated by mountainous topography. Following optimization, the red areas on the map are visibly reduced and replaced by lower MAE values. A similar enhancement is evident in the correlation map. While the default model generally exhibited low correlation ($R \approx 0.3$), the optimized model demonstrates a

clear shift toward higher correlation ($R > 0.4$), especially in the coastal areas and the eastern part of East Java.

This spatial analysis provides direct insight into the influence of elevation on model performance, where RMSE and MAE values tend to increase, and correlation coefficients tend to decrease, as regional elevation rises [12]. The default model proved less capable of representing rainfall dynamics in topographically complex regions, as reflected by the high MAE and low correlation in highlands and mountainous areas. Conversely, the optimized model showed the most substantial performance gains specifically in these rugged terrains. These findings indicate that while the optimized model outperforms the default version, the challenge of orographic rainfall remains a significant barrier to accurate detection.

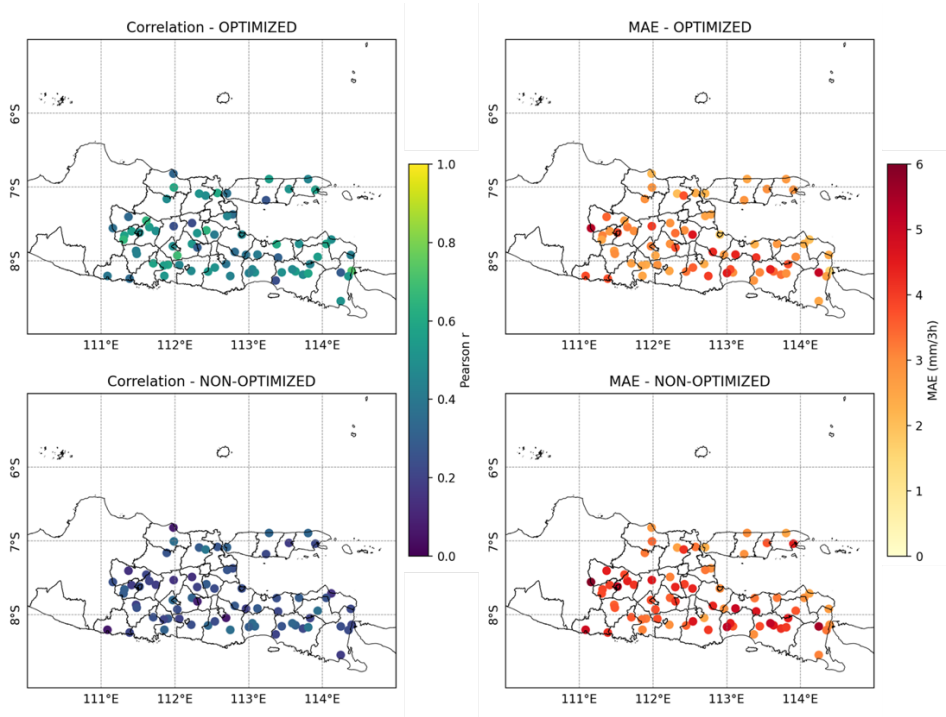

Fig. 5. Spatial distribution of regression performance metrics comparing the Optimized (top row) and Non-Optimized (bottom row) XGBoost models across observation stations in East Java. The left panels display Pearson Correlation coefficients (R), while the right panels display Mean Absolute Error (MAE in mm/3h).

The quantitative evaluation of model performance across varying topographic gradients reveals a distinct degradation in predictive performance at higher elevations, as evidenced by the systematic shift in error metrics and correlation coefficients. This trend is substantiated by the positive linear regression slopes observed for both Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) relative to station elevation, alongside a corresponding decline in the Pearson correlation coefficient (R) (Fig. 6). To facilitate this assessment, observation stations were partitioned into three distinct elevation groups: lowlands (stations below 60 m above sea level [asl]), mid-elevation (61–200 m asl), and high-altitude regions (stations exceeding 200 m asl). Specifically, categorical analysis shows that the mean RMSE increases progressively from 6.23 in the lowland group to 7.28 in the high-altitude areas,

while the mean MAE rises from 2.67 to 3.65, indicating that the model's predictive uncertainty intensifies in complex mountainous terrains. Simultaneously, the mean correlation drops from 0.50 to 0.45 at higher elevations, suggesting a weakened linear agreement between the satellite-derived predictors and surface observations. Such performance discrepancies likely stem from the intricate orographic effects and localized convective processes inherent to high-altitude topography, which present significant challenges for infrared-based brightness temperature (BT) and difference (BTD) channels to resolve accurately compared to more uniform lowland conditions. Consequently, these findings highlight the necessity for enhanced topographic integration or localized recalibration to mitigate elevation-dependent biases in satellite-based rainfall estimation models.

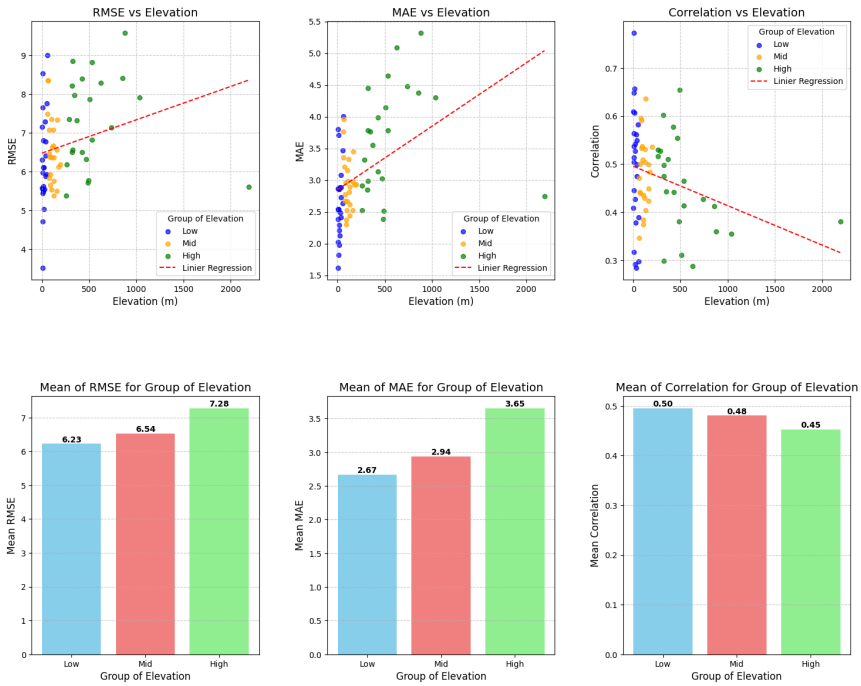

Fig. 6. Analysis of model performance metrics (RMSE, MAE, and Correlation) relative to station elevation. The upper panels show scatter plots with linear regression trends, while the lower panels display the mean error values for low, mid, and high-elevation groups.

3.4 Classification model performance evaluation

The evaluation results presented in Table 3 also demonstrate that hyperparameter optimization yields consistent performance improvements across both classification modeling stages. In binary classification, ACC increased by 5.5%; however, the most significant improvement was observed in CSI, which rose by 16.2%, signaling a tangible enhancement in the model's ability to detect rainfall events. A similar pattern is evident in multi-classification, where ACC increased by 6.8% to 0.90, while CSI experienced a much larger surge of 34.5%, rising from 0.38 to 0.49. The dominance of CSI improvement over ACC indicates that optimization was particularly effective in reducing misses and false alarms, thereby improving the model's precision in identifying rainfall classes that are difficult to distinguish due to class imbalance. Thus, hyperparameter optimization is proven

to not only enhance general accuracy but, more importantly, to improve the model's sensitivity toward rare yet critical rainfall events [11].

The spatial evaluation illustrated in Fig. 7 and Fig. 8 provides essential context regarding the multi-classification model's performance. The Accuracy map in Fig. 7 displays very high values for all rainfall classes across nearly all stations in East Java; however, this pattern reflects the Accuracy Paradox inherent in imbalanced datasets. Accuracy appears high not because the model is genuinely superior, but because the 'Clear' class is the dominant event, meaning a prediction of 'no rain' is almost always correct.

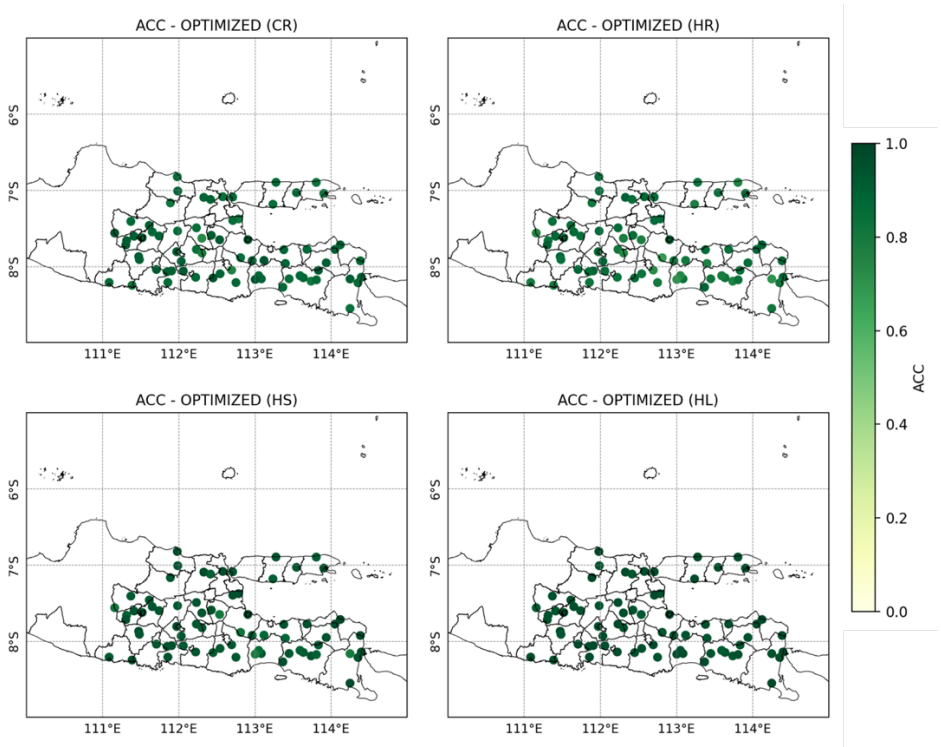

Fig. 7. Spatial distribution of Accuracy (ACC) scores for the optimized XGBoost model across four rainfall categories: Clear (CR), Light Rain (HR), Moderate Rain (HS), and Heavy Rain (HL).

The CSI maps provide a clear spatial overview of model performance, revealing distinct behavioral patterns across each rainfall class (Fig. 8). For the Clear class, CSI values remain high and uniform across the entire region, indicating that the binary classification stage excels at identifying no-rain conditions. In contrast, the CSI distribution for Light and Moderate Rain classes exhibits significantly greater variability. Coastal areas and the northern lowlands demonstrate superior CSI values, whereas the central mountainous regions consistently display a decline in performance. This pattern reinforces the understanding that topographic complexity, particularly orographic rainfall mechanisms and local convective processes, serves as a primary factor diminishing prediction accuracy in highland areas. Regarding the Heavy Rain class, CSI values are low across nearly all stations, confirming that these rare high-intensity events remain the model's most significant challenge. This finding aligns with the underestimation observed in the PDF analysis and illustrates that the scarcity of extreme events lies at the root of the data imbalance problem in precipitation modeling. However, the difficulty in detecting heavy rainfall is not attributable solely to data

imbalance, but also stems from the inherent physical limitations of infrared sensors. In tropical regions, heavy precipitation is frequently driven by 'warm rain' processes where cloud tops are not sufficiently cold to be flagged by IR sensors [13], while cold cirrus clouds often trigger false alarms [14]. Overall, these spatial patterns confirm that CSI provides a more representative measure of the model's detection capability than accuracy.

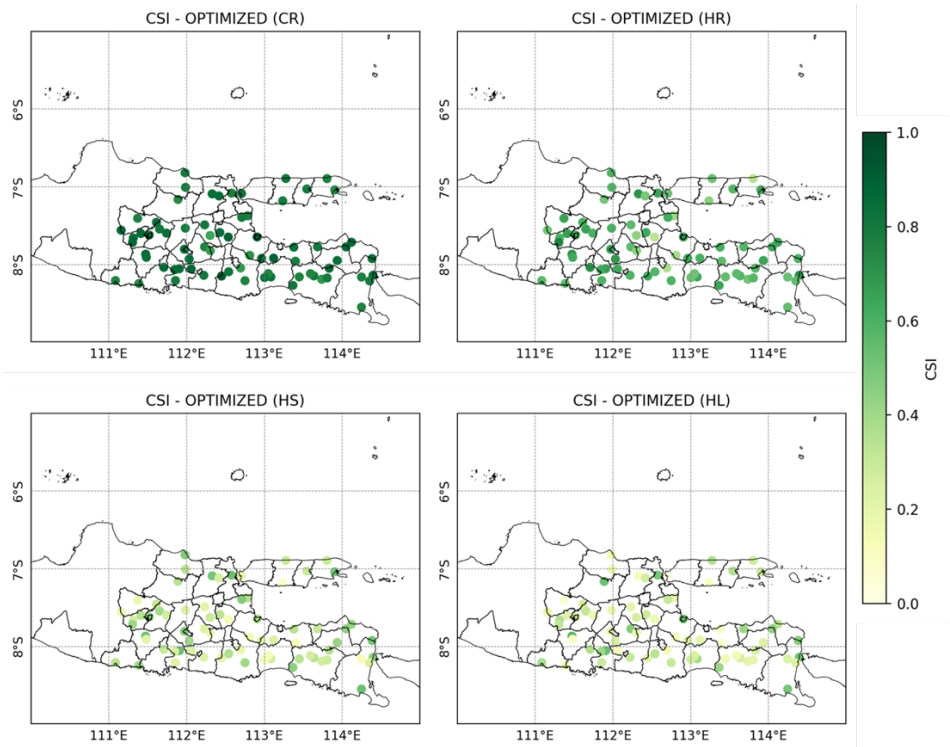

Fig. 8. Spatial distribution of Critical Success Index (CSI) scores for the optimized XGBoost model across four rainfall categories: Clear (CR), Light Rain (HR), Moderate Rain (HS), and Heavy Rain (HL).

4. Conclusion

This study demonstrates that integrating multi-source data (comprising Himawari-8/9 Brightness Temperature (BT) and Brightness Temperature Difference (BTD), atmospheric indices, and topography) with the XGBoost algorithm yields an accurate and operationally relevant 3-hourly rainfall estimation system for East Java. Initial analysis confirmed that the relationship between satellite predictors and precipitation is highly non-linear, further complicated by extreme class imbalance and distributional overlap among rainfall categories.

The optimization of XGBoost hyperparameters proved effective, resulting in significant performance enhancements. In the regression stage, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) decreased by 21.5% and 23.0%, respectively, while correlation increased by 43.1%. This indicates that optimization effectively suppresses large errors and improves prediction precision. Spatial evaluation indicates that the most substantial improvements occurred in mountainous regions, signaling the optimized model's capability to capture complex orographic rainfall dynamics. Regarding classification, the Critical Success Index (CSI) improved by 16.2% for the binary stage and 34.5% for the multi-

classification stage, confirming the model's enhanced sensitivity in detecting rare rainfall events. Nevertheless, performance for the heavy rainfall class remains limited, reflecting challenges stemming from the scarcity of observational data.

Overall, this study confirms that the optimized XGBoost based model delivers significant performance gains and successfully mitigates the spatial heterogeneity of predictions. The model demonstrates strong potential as an operational component within a satellite-based early warning system for the East Java region. However, the detection of rare heavy rainfall remains a critical limitation. This opens avenues for future research utilizing image-based deep learning approaches, radar data integration, or hybrid physical-statistical models to further enhance sensitivity toward extreme events.

The Acknowledgements: The authors would like to express their sincere appreciation to the Meteorology, Climatology, and Geophysics Agency (BMKG) for providing the data and necessary facilities that supported this research.,
Fundings: This research was financially supported by the Indonesia Endowment Fund for Education (LPDP),
Data availability: This study has been conducted using BMKG Data.

References

1. BNPB, Jumlah kejadian bencana menurut jenis bencana. (2025). Diakses dari https://data.bnpb.go.id/dataset/data-bencana-indonesia/resource/9b41007e-c998-456b-8cbc-385b17986e46
2. S. Berkhahn, L. Fuchs, I. Neuweiler, An ensemble neural network model for real-time prediction of urban floods, *J. Hydrol.* **575**, 743 (2019).https://doi.org/10.1016/J.JHYDROL.2019.05.066
3. M. Min, C. Bai, J. Guo, F. Sun, C. Liu, F. Wang, H. Xu, S. Tang, B. Li, D. Di, L. Dong, J. Li, Estimating Summertime Precipitation from Himawari-8 and Global Forecast System Based on Machine Learning, *IEEE Trans. Geosci. Remote Sens.* **57**, 2557 (2019).https://doi.org/10.1109/TGRS.2018.2874950
4. M. Putra, M.S. Rosid, D. Handoko, High-Resolution Rainfall Estimation Using Ensemble Learning Techniques and Multisensor Data Integration, *Sensors* **24**, 5030 (2024). https://doi.org/10.3390/s24155030
5. S. Kundu, S.K. Biswas, D. Tripathi, R. Karmakar, S. Majumdar, S. Mandal, A review on rainfall forecasting using ensemble learning techniques, *e-Prime* **6**, 100296 (2023). https://doi.org/10.1016/j.prime.2023.100296
6. H. Hang, J. Mallick, S. Alqadhi, A.A. Bindajam, H.G. Abdo, Exploring forest fire susceptibility and management strategies in Western Himalaya: Integrating ensemble machine learning and explainable AI for accurate prediction and comprehensive analysis, *Environ. Technol. Innov.* **35**, 103655 (2024). https://doi.org/10.1016/j.eti.2024.103655
7. B. Wu, P. Chen, M. Wei, Bayesian optimization-based XGBoost for performance Prediction of Carbon Nanotube Membranes. (2024).https://doi.org/10.21203/RS.3.RS-4562640/V1
8. S. Zhou, Y. Wang, Q. Yuan, L. Yue, L. Zhang, Spatiotemporal estimation of 6-hour high-resolution precipitation across China based on Himawari-8 using a stacking ensemble machine learning model, *J. Hydrol.* **609**, 127718 (2022). https://doi.org/10.1016/j.jhydrol.2022.127718

9. G.E.A.P.A. Batista, R.C. Prati, M.C. Monard, A study of the behavior of several methods for balancing machine learning training data, *ACM SIGKDD Explor. Newsl.* **6**, 20 (2004). https://doi.org/10.1145/1007730.1007735
10. T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), pp. 785–794. https://doi.org/10.1145/2939672.2939785
11. A.U.G. Senocak, M.T. Yilmaz, S. Kalkan, I. Yucel, M. Amjad, An explainable two-stage machine learning approach for precipitation forecast, *Journal of Hydrology.* **627**, 130375 (2023). https://doi.org/10.1016/j.jhydrol.2023.130375
12. J. Ko, K. Lee, H. Hwang, S. G. Oh, S. W. Son, K. Shin, Effective training strategies for deep-learning-based precipitation nowcasting and estimation, *Computers & Geosciences* **161**, 105072 (2022). https://doi.org/10.1016/J.CAGEO.2022.105072
13. B. Sohn, G. Ryu, H. Song, Observational Characteristics of Warm-Type Heavy Rainfall, *Advances in Global Change Research* **69**, pp. 745-759 (2020). https://doi.org/10.1007/978-3-030-35798-6_15
14. C. Kidd, V. Levizzani, Status of satellite precipitation retrievals, *Hydrology and Earth System Sciences* **15**, pp. 1109-1116 (2011). https://doi.org/10.5194/hess-15-1109-2011