RESEARCH ARTICLE • OPEN ACCESS

Hyperparameter-Optimized Gradient Boosting for Daily Rainfall Prediction Using BMKG Meteorological Data in Malang Regency, Indonesia

Mohamad Arif Abdul Syukur, Suhartono, and Mochamad Imamudin

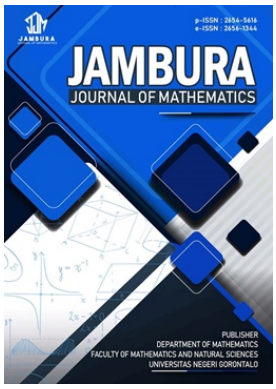
JAMBURA JOURNAL OF MATHEMATICS

Volume 8, Issue 2, Pages 188–201, August 2026

Received 30 April 2026, Revised 13 June 2026, Accepted 18 June 2026, Published 20 July 2026

To Cite this Article : M. A. A. Syukur, S. Suhartono, and M. Imamudin,"Hyperparameter-Optimized Gradient Boosting for Daily Rainfall Prediction Using BMKG Meteorological Data in Malang Regency, Indonesia", *Jambura J. Math*, vol. 8, no. 2, pp. 188–201, 2026, *https://doi.org/10.37905/jjom.v8i2.38554*

JOURNAL INFO • JAMBURA JOURNAL OF MATHEMATICS



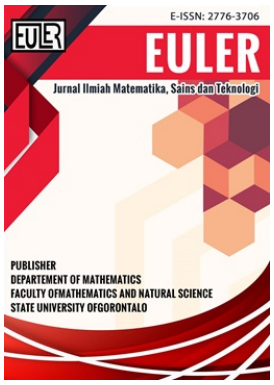
Homepage	:	http://ejurnal.ung.ac.id/index.php/jjom/index
Journal Abbreviation	:	Jambura J. Math.
Frequency	:	Biannual (February and August)
Publication Language	:	English (preferable), Indonesia
DOI	:	https://doi.org/10.37905/jjom
Online ISSN	:	2656-1344
Editor-in-Chief	:	Hasan S. Panigoro
Publisher	:	Department of Mathematics, Universitas Negeri Gorontalo
Country	:	Indonesia
OAI Address	:	http://ejurnal.ung.ac.id/index.php/jjom/oai
Google Scholar ID	:	iWLjgaUAAAAJ
Email	:	info.jjom@ung.ac.id

JAMBURA JOURNAL • FIND OUR OTHER JOURNALS

Jambura Journal of Biomathematics

Jambura Journal of Mathematics Education

Jambura Journal of Probability and Statistics



EULER : Jurnal Ilmiah Matematika, Sains, dan Teknologi

Jambura Journal of Mathematics, Volume 8, Issue 2, Pages 188–201, August 2026 https://doi.org/10.37905/jjom.v8i2.38554

Research Article

Check for updates

Hyperparameter-Optimized Gradient Boosting for Daily Rainfall Prediction Using BMKG Meteorological Data in Malang Regency, Indonesia

Mohamad Arif Abdul Syukur[1,*], **Suhartono**[1], **Mochamad Imamudin**[1]

[1]*Master's Program in Informatics, Universitas Islam Negeri Maulana Malik Ibrahim, Malang 65144, Indonesia*

ARTICLE HISTORY
Received 30 April 2026
Revised 13 June 2026
Accepted 18 June 2026
Published 20 July 2026

KEYWORDS
Prediction
Weather
Rainfall
Gradient Boosting
BMKG
RMSE

ABSTRACT. *Weather conditions significantly affect many aspects of modern life, including transportation, tourism, agriculture, and disaster risk management, particularly in relation to rainfall. Consequently, reliable meteorological information is essential for supporting daily decision-making, making rainfall prediction increasingly important. This study develops a daily rainfall prediction model using Gradient Boosting based on daily meteorological data obtained from the Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG). The dataset includes date, minimum, maximum, and average temperatures, relative humidity, sunshine duration, maximum and average wind speeds, and wind direction, with daily rainfall as the target variable. Four chronological train–test split scenarios were evaluated. The first scenario produced an RMSE of 13.97, an MAE of 7.96, and an R^2 value of 0.14. The second scenario yielded an RMSE of 12.81, an MAE of 8.72, and an R^2 value of 0.17. The third scenario achieved an RMSE of 12.21, an MAE of 7.70, and an R^2 value of 0.20, whereas the fourth scenario obtained an RMSE of 10.31, an MAE of 7.11, and an R^2 value of -0.27. Considering both prediction error and generalization capability, the third scenario was selected as the best-performing model. The main contribution of this study lies in demonstrating the effectiveness of hyperparameter optimization in improving the stability of rainfall prediction under complex tropical climatic conditions. Practically, the proposed model may support BMKG and regional policymakers in Malang Regency in hydrometeorological disaster mitigation and agricultural planning.*

1. Introduction

Daily rainfall prediction is a critical forecasting problem because it directly affects public safety, agricultural activities, water-resource management, and flood-risk mitigation [1]. Extreme weather events, particularly high-intensity rainfall, may also generate substantial economic consequences. Uncertain weather conditions can disrupt transportation operations through flight delays and cancellations, potentially reducing operator revenues because of the time required to restore normal operations [2]. The integration of daily observations from official meteorological agencies, such as the Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG), with machine-learning algorithms offers considerable potential for improving rainfall prediction in regions characterized by high climatic variability. Reliable rainfall forecasts have therefore become increasingly important because many modern activities depend on meteorological information for operational and strategic decision-making.

This study focuses on the application of Gradient Boosting to predict daily rainfall using BMKG meteorological data. This approach is motivated by empirical evidence showing that ensemble-based and boosted decision-tree models frequently outperform conventional linear models and individual learners in rainfall forecasting across different geographical scales and climatic conditions [3]. Daily BMKG records provide long-term, station-based, and location-specific meteorological information that is valuable for understanding local hydrometeorological variability and constructing operational daily forecasting models [4]. Meteorological time series also exhibit temporal trends, seasonal patterns, nonlinear relationships, and complex interactions among variables, making rainfall prediction a challenging data-driven problem [5].

Although highly detailed meteorological data may improve predictive performance, such data are often limited in spatial and temporal coverage. Gradient Boosting provides a flexible modeling framework capable of handling heterogeneous predictors and complex nonlinear relationships. When combined with robust nonparametric learning techniques, an enriched feature space can substantially improve predictive performance [6]. Previous studies employing ensemble-based methods and Gradient Boosting have demonstrated promising results for rainfall forecasting across various geographical regions and temporal resolutions, including daily and seasonal prediction [7].

Modern Gradient Boosting algorithms perform effectively on tabular meteorological datasets in which predictors may include lagged variables, daily statistics, and engineered features

*Corresponding Author.

Email : *mohamadarifabdulsyukur@mail.com* (M. A. A. Syukur)
Homepage : http://ejurnal.ung.ac.id/index.php/jjom/index / E-ISSN : 2656-1344

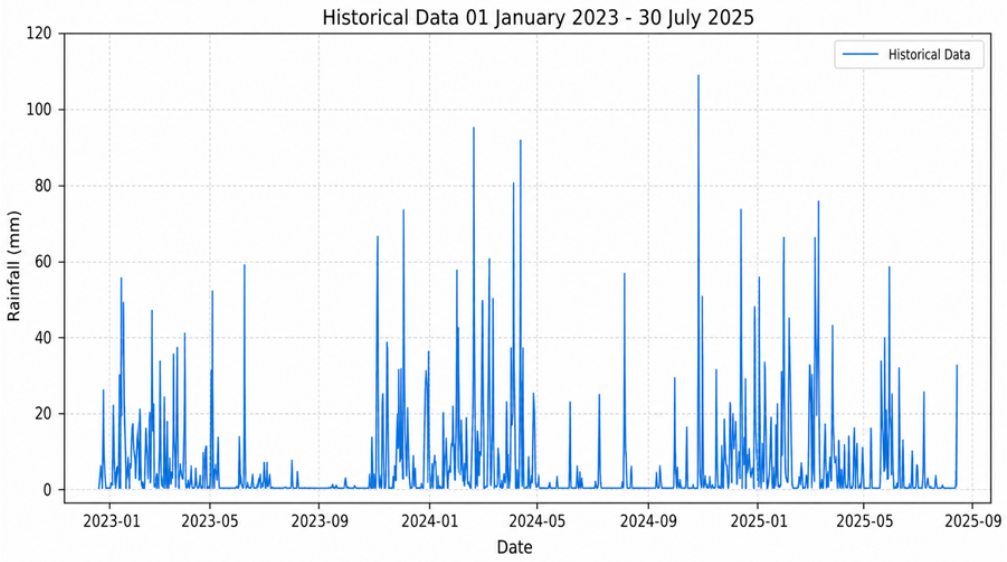


Figure 1. Historical daily rainfall data used in this study.

representing relevant atmospheric processes [8]. In rainfall prediction, Gradient Boosting variants have consistently achieved performance comparable to or better than that of conventional machine-learning models, particularly when the data exhibit nonlinear effects and strong interactions among predictors [9]. These characteristics provide a strong justification for selecting Gradient Boosting as the primary learning algorithm for daily BMKG data. Furthermore, the reliability of time-series prediction depends not only on the selected algorithm but also on appropriate validation procedures and careful hyperparameter adjustment. Hyperparameter optimization can identify configurations that remain robust across different weather regimes and forecasting horizons [6].

Rainfall is influenced by multiple interdependent meteorological variables that interact nonlinearly. Studies conducted in different climatic regions have shown that Gradient Boosting can produce strong predictive performance for daily and short-term rainfall forecasting, often outperforming alternative machine-learning techniques and, under suitable data conditions, competing with physically based or hybrid approaches. Appiah-Badu et al. [7], for example, used 40 years of historical climate data from the Ghana Meteorological Agency and reported an accuracy of 83.16%. Monego et al. [6] employed NCEP/NCAR reanalysis data and obtained RMSE values of 1.53 for summer and 0.98 for spring. More recently, Alvines et al. [9] analyzed 2024 BMKG data from South Sumatra and reported an RMSE of 12.54. These findings support the use of Gradient Boosting as an efficient data-driven method for daily rainfall prediction [10].

This study applies Gradient Boosting to predict daily rainfall in Malang Regency using historical BMKG data from 2023 to 2025. To strengthen the methodological contribution, the analysis does not rely solely on a default model configuration. Instead, it compares a baseline Gradient Boosting model with an optimized model obtained through `RandomizedSearchCV` under several chronological train–test split scenarios. This comparative design is intended to identify the most stable configuration for capturing the complex variability of local rainfall. The methodological contribution of this study lies in evaluating a hyperparameter-optimization strategy tailored to the climatic characteristics of Malang Regency and in examining model stability under different proportions of training and testing data. The temporal distribution and variability of the historical daily rainfall observations used in this study are presented in Figure 1.

The rainfall prediction framework uses daily meteorological observations collected by BMKG in Malang Regency from 2023 to 2025. Gradient Boosting is combined with hyperparameter optimization and multiple chronological data-splitting scenarios to assess the sensitivity of predictive performance to the proportion of training data. The `RandomizedSearchCV` procedure is employed to identify parameter combinations that improve the model's ability to capture temporal rainfall patterns while maintaining generalization performance. The resulting model is expected to provide insight into both the potential and the limitations of machine learning as an early decision-support tool for BMKG and other stakeholders dealing with regional weather uncertainty.

Accordingly, the objective of this study is to develop and evaluate a daily rainfall prediction model using BMKG meteorological data from Malang Regency for the 2023–2025 period. Gradient Boosting is selected because of its ability to represent nonlinear, dynamic, and noisy weather patterns that may not be adequately captured by conventional statistical models. To assess the contribution of hyperparameter optimization, the predictive performance of the default Gradient Boosting model is compared with that of the optimized model using MAE, RMSE, and R^2 under four chronological train–test split scenarios. Through this framework, the study aims to determine a model configuration that balances prediction error and generalization capability and that may support decision-making in meteorology, hydrometeorological disaster mitigation, and agricultural planning.

2. Methods

This study implements a machine-learning approach using the Gradient Boosting Regressor in Python 3.13.5 to predict daily rainfall. The dataset was obtained from BMKG in Malang Regency and consists of 942 daily observations with 11 recorded vari-

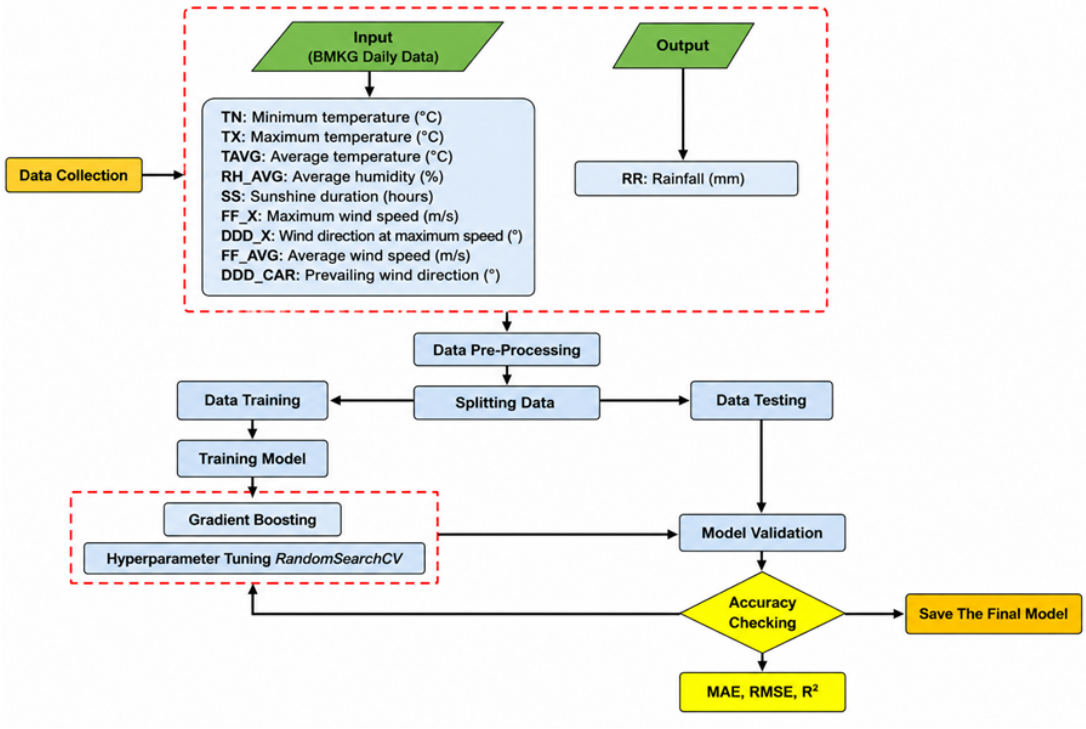


Figure 2. System design of the proposed daily rainfall prediction framework.

ables: date, minimum temperature, maximum temperature, average temperature, average relative humidity, rainfall, sunshine duration, maximum wind speed, average wind speed, wind direction at maximum speed, and prevailing wind direction. Each observation represents one day of meteorological measurement.

The data were processed using K-Nearest Neighbors imputation with $k = 5$, followed by feature standardization using `StandardScaler`. To preserve temporal ordering and prevent data leakage, the dataset was divided chronologically into training and testing sets using four split ratios: 50:50, 70:30, 80:20, and 90:10. Hyperparameter optimization was performed using `RandomizedSearchCV` with 50 iterations over the `n_estimators`, `learning_rate`, `subsample`, and `max_depth` parameter spaces.

Figure 2 illustrates the overall research framework, beginning with the collection of daily meteorological data from BMKG. The predictor variables include minimum, maximum, and average temperatures; average relative humidity; sunshine duration; maximum and average wind speeds; wind direction at maximum speed; and prevailing wind direction. Daily rainfall is defined as the target variable.

The next stage involves data preprocessing to improve dataset quality through data-type conversion and missing-value treatment. The cleaned dataset is then divided chronologically into training and testing sets using four split scenarios: 50:50, 70:30, 80:20, and 90:10. During the modeling stage, the Gradient Boosting algorithm is used to construct the rainfall prediction model, while `RandomizedSearchCV` is applied to optimize its hyperparameters. Finally, model performance is quantitatively evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). The detailed sequence of the research procedures is presented in Figure 3.

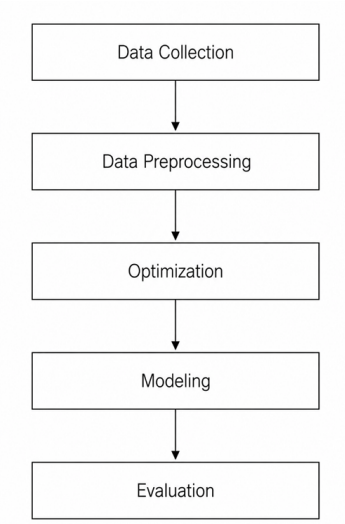


Figure 3. Stages of the daily rainfall prediction study.

2.1. Data Collection

Data collection involves identifying and obtaining information that is relevant to the research objectives and can be used as input for model development. Historical meteorological records constitute the primary data source for rainfall prediction. When combined with an appropriate predictive model, historical weather patterns provide temporal and multivariate relationships that can be used to estimate future rainfall. Consistent with the supervised-learning regression framework, in which rainfall is predicted from historical meteorological variables, previous studies have emphasized the importance of multivariate inputs obtained from weather stations or remote-sensing products [11].

Although machine-learning models commonly rely on historical weather measurements collected at the target location, rainfall prediction may also benefit from multisource data to strengthen the predictive signal [12]. Nevertheless, data quality, temporal coverage, consistency, and stationarity during the training period are important considerations when selecting a dataset. These considerations support the use of a sufficiently long and consistent single-source dataset for chronological model training and testing [13].

This study uses daily meteorological observations obtained from BMKG for the period from January 1, 2023, to July 15, 2025. The dataset was selected because its attributes represent the actual meteorological conditions required for daily rainfall modeling. Data collection was followed by an initial examination of the dataset structure, variable definitions, temporal coverage, and data completeness to ensure its suitability for subsequent preprocessing and model development.

The data were recorded at the East Java Climatological Station operated by the Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG). The observation station has World Meteorological Organization identification number WMO 96943 and is located at latitude $-7.90080°$, longitude $112.59790°$, and an elevation of 590 m above sea level. The data were obtained through the official BMKG online data portal at https://dataonline.bmkg.go.id/. To replicate the data-acquisition procedure, users may register an account, select the "Climate Data" category, specify the required observation period, and choose the station according to the technical specifications described above. Before public distribution, the meteorological records provided by BMKG undergo an internal quality-control procedure.

2.2. Data Preprocessing

Data preprocessing includes missing-value treatment, data-type conversion, feature engineering, standardization, and chronological separation of the training and testing sets. To preserve the temporal structure of the observations, all preprocessing parameters were estimated exclusively from the training data. In particular, the transformation parameters, including the mean and standard deviation used for standardization, were fitted to the training set and subsequently applied to the testing set. This procedure prevents information from future observations from influencing model training, thereby minimizing data leakage and reproducing a realistic forecasting setting [14].

Two principal procedures were initially applied to ensure the integrity and quality of the dataset. First, data-type conversion was performed to transform the meteorological variables into numerical formats compatible with machine-learning algorithms [15]. Second, missing observations and non-numerical entries were converted into `NaN` values and treated using the K-Nearest Neighbors imputation method. The `KNNImputer` was configured with $k = 5$ to preserve the number of observations and maintain the continuity of the meteorological time series.

The KNN-based imputation method was selected because meteorological variables exhibit temporal and multivariate relationships. Consequently, a missing value on a particular day may be estimated using observations with similar temperature, humidity, wind, and other meteorological characteristics. In contrast to mean or median imputation, which may reduce the natural variability of the data, KNN imputation estimates missing values based on local similarity among observations and therefore better preserves the underlying data distribution. This approach is also consistent with the procedure adopted in previous rainfall-prediction research [9].

To prevent data leakage, the imputer was fitted exclusively to the training data for each chronological split scenario. The learned neighborhood structure was then used to transform both the training and testing sets. The quality of the imputed data was assessed using two procedures. First, statistical characteristics before and after imputation were compared to ensure that the procedure did not substantially alter the original data distribution. Second, a comprehensive null-value inspection and graphical examination were conducted to verify that the imputed values remained consistent with the patterns of the corresponding meteorological variables [16].

Feature engineering was subsequently performed to capture the nonlinear and temporal dependence of daily rainfall. Previous meteorological observations can provide relevant predictors through the construction of lagged variables, particularly because weather data frequently contain nonlinear relationships and interactions that cannot be adequately represented by simple linear regression models [17]. Accordingly, seven lagged rainfall variables were constructed from the target variable RR:

$$RR_{t-1},\ RR_{t-2},\ RR_{t-3},\ RR_{t-4},\ RR_{t-5},\ RR_{t-6},\ \text{and}\ RR_{t-7}.$$

These variables represent rainfall observations from the preceding seven days and enable the model to identify short-term persistence in daily rainfall patterns. After data-type conversion, missing-value imputation, and lag-feature construction, the predictor variables were standardized using `StandardScaler` before being supplied to the Gradient Boosting model.

2.3. Hyperparameter Optimization

Gradient Boosting and its variants, including XGBoost and LightGBM, are widely used for meteorological prediction because of their ability to model nonlinear relationships and complex interactions in tabular datasets [18, 19]. The predictive performance of these methods depends strongly on their hyperparameter configuration. Consequently, optimization procedures are commonly employed to adjust parameters such as the learning rate, maximum tree depth, number of estimators, subsampling ratio, and regularization settings to improve predictive accuracy while reducing the risk of overfitting [6].

Regularization strategies can stabilize the learning process and improve model generalization. In stochastic Gradient Boosting, subsampling introduces randomness by training each weak learner on only a fraction of the available observations. This mechanism may reduce variance and limit overfitting while maintaining the ability of the ensemble to capture complex nonlinear patterns [18].

Hyperparameter optimization in this study was conducted using `RandomizedSearchCV`. The search space consisted of a `learning_rate` ranging from 0.01 to 0.20, `max_depth` ranging from 2 to 6, `n_estimators` ranging from 100 to 500, and `subsample` ranging from 0.80 to 1.00. A total of 50 parameter combinations were randomly sampled from the predefined

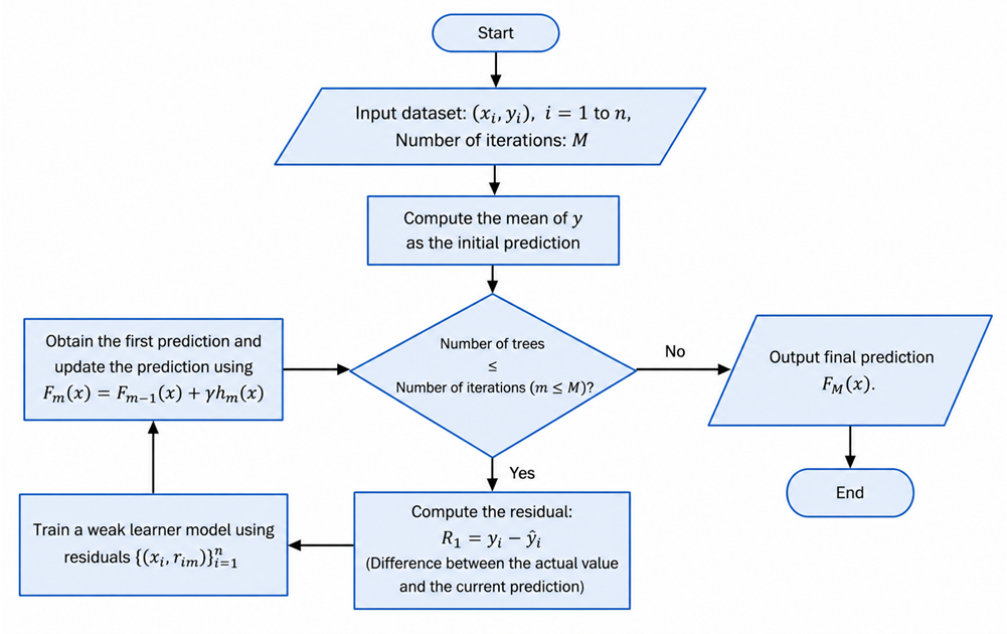


Figure 4. General procedure of the Gradient Boosting algorithm.

search space. Each candidate configuration was evaluated using time-series-based validation, and the optimal configuration was selected according to the minimum Root Mean Squared Error (RMSE).

2.4. Model Development

The regression model aims to predict continuous daily rainfall values using historical meteorological predictors. Accordingly, the problem is formulated within a supervised-learning framework in which the model learns a mapping from the predictor vector $\mathbf{x}$ to the rainfall target RR based on observations with known outcomes [13]. After completing the training process, the resulting ensemble model can be expressed as

$$F_M(\mathbf{x}) = \widehat{\mathrm{RR}},$$

where $F_M(\mathbf{x})$ denotes the final Gradient Boosting prediction after M boosting iterations.

Gradient Boosting was selected because it can represent nonlinear predictor effects and complex interactions among meteorological variables without requiring strong assumptions regarding the functional relationship between the inputs and rainfall. These characteristics make it suitable for rainfall prediction problems involving dynamic, nonlinear, and noisy data [20]. The selection of the modeling method was also determined by the continuous nature of the response variable, which requires a regression rather than a classification, clustering, or association-learning framework [21].

Alternative machine-learning methods, including Support Vector Regression, Random Forest, Artificial Neural Networks, and Long Short-Term Memory networks, have also been applied to rainfall forecasting. However, this study focuses on Gradient Boosting because of its predictive capability, flexibility in handling tabular meteorological data, and potential interpretability when combined with feature-importance analysis [22].

Gradient Boosting is an ensemble-learning method applicable to both regression and classification problems. For regression, the algorithm sequentially constructs weak learners, typically decision trees, in which each new learner is trained to correct the residual errors generated by the preceding ensemble. The objective is to minimize a differentiable loss function by iteratively adding models in the direction of the negative gradient [23]. The general Gradient Boosting procedure adopted in this study is illustrated in Figure 4.

Let the training dataset be

$$\{(\mathbf{x}_i, y_i)\}_{i=1}^n,$$

where $\mathbf{x}_i$ is the predictor vector, y_i is the observed rainfall value, $L(y, F(\mathbf{x}))$ is a differentiable loss function, and M is the total number of boosting iterations. For the squared-error loss used in regression, the initial model is defined as the average observed rainfall:

$$F_0(\mathbf{x}) = \bar{y}, \tag{1}$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

For each iteration $m = 1, 2, \ldots, M$, the negative gradient or pseudo-residual is calculated as

$$r_{im} = -\left[\frac{\partial L\left(y_i, F(\mathbf{x}_i)\right)}{\partial F(\mathbf{x}_i)}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}, \qquad i = 1, 2, \ldots, n. \tag{2}$$

A weak learner $h_m(\mathbf{x})$ is subsequently fitted to the residual training set

$$\{(\mathbf{x}_i, r_{im})\}_{i=1}^n.$$

The optimal multiplier γ_m is determined by solving the one-dimensional optimization problem

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(\mathbf{x}_i) + \gamma h_m(\mathbf{x}_i)\right). \tag{3}$$

The ensemble model is then updated according to

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \eta\gamma_m h_m(\mathbf{x}), \tag{4}$$

where η denotes the learning rate. A smaller value of η reduces the contribution of each weak learner and generally requires a larger number of estimators, whereas a larger value accelerates learning but may increase the risk of overfitting.

The notation used in Eq. (1) to (4) is defined as follows:

- $\mathbf{x}_i$: predictor vector for the ith observation;
- y_i: observed rainfall value for the ith observation;
- $\widehat{y}_i$: predicted rainfall value for the ith observation;
- $F_0(\mathbf{x})$: initial prediction model;
- $F_m(\mathbf{x})$: ensemble prediction after the mth iteration;
- n: number of observations;
- m: boosting iteration index;
- M: total number of boosting iterations;
- $\bar{y}$: average observed rainfall;
- r_{im}: pseudo-residual of the ith observation at the mth iteration;
- $\partial L/\partial F(\mathbf{x}_i)$: derivative of the loss function;
- γ_m: optimal multiplier for the mth weak learner;
- η: learning rate; and
- $h_m(\mathbf{x})$: weak learner fitted at the mth iteration.

2.5. Model Evaluation

Regression-model performance is commonly evaluated using the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Root Mean Squared Percentage Error, and coefficient of determination (R^2). Lower MAE and RMSE values indicate smaller prediction errors, whereas a higher R^2 value indicates that a larger proportion of the observed variation is explained by the model. These metrics are widely used in machine-learning-based rainfall forecasting [24].

In this study, the Gradient Boosting Regressor was evaluated using MAE, RMSE, and R^2. The three metrics provide complementary assessments of model performance. MAE measures the average absolute difference between observed and predicted rainfall values in millimeters and therefore provides a direct interpretation of the typical prediction error. Because positive and negative errors are treated equally, MAE is not affected by the direction of overestimation or underestimation.

RMSE assigns a greater penalty to large prediction errors because the residuals are squared before averaging. This property is particularly important in rainfall forecasting because substantial errors during high-intensity rainfall events may have serious implications for hydrometeorological disaster preparedness. In contrast, R^2 measures the proportion of rainfall variability that can be explained by the meteorological predictors and provides an indication of the model's performance relative to a mean-based baseline. The combined use of MAE, RMSE, and R^2 therefore provides a more comprehensive evaluation than the use of a single metric [25].

Temporal validation, time-series cross-validation, and chronological hold-out testing are important for assessing model generalization under realistic forecasting conditions [26]. Although rainfall models may be compared with alternative algorithms such as CatBoost, XGBoost, Random Forest, Support Vector Regression, and Multilayer Perceptron, ensemble methods combined with hyperparameter optimization have frequently demonstrated strong predictive performance [6].

The MAE is defined as

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y}_i\right|, \tag{5}$$

whereas the RMSE is calculated as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}. \tag{6}$$

The coefficient of determination is given by

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{7}$$

In Eq. (5) to (7), y_i denotes the observed rainfall value, $\widehat{y}_i$ denotes the predicted rainfall value, $\bar{y}$ denotes the mean observed rainfall, and n denotes the number of observations in the testing set.

To evaluate the effect of hyperparameter optimization and training-data availability, the optimized Gradient Boosting model was assessed under four chronological train–test split scenarios: 50:50, 70:30, 80:20, and 90:10. The comparison was designed to determine whether the model retained stable predictive performance across different proportions of training and testing data. Model selection was based on the joint interpretation of MAE, RMSE, and R^2, rather than on the smallest error metric alone.

Several methodological limitations were also considered when interpreting the evaluation results. These include the representativeness of the available observations, possible nonstationarity caused by long-term climatic changes, the limited occurrence of extreme rainfall events, and the risk of overfitting despite the use of regularization and temporal validation. Previous studies have therefore recommended the development of more interpretable models, the integration of additional atmospheric predictors, and the implementation of prediction systems within operational decision-support frameworks [24, 26–28].

3. Results and Discussion

3.1. Data Collection

The dataset consists of daily meteorological observations recorded from January 1, 2023, to July 15, 2025. The 2023–2025 observation period was selected based on the availability of continuous station records with minimal sensor anomalies. This restriction was intended to ensure that the model learned weather characteristics from observations with a high level of completeness and reliability, thereby reducing the risk of bias caused by missing or invalid measurements over a longer observation period.

Although rainfall events outside the selected period may also influence the general climatic characteristics of the region, the observation window was operationally defined to prioritize data quality for machine-learning model development. Therefore, the study does not assume that observations outside this

period are unimportant; rather, its scope is limited to the period with the most complete and reliable records available for the present analysis.

The secondary dataset was obtained from daily observations at the BMKG climatological station in Malang Regency and contains 942 observations. The selected period covers multiple seasonal cycles, enabling the model to learn seasonal rainfall variability and its relationships with other meteorological variables. Each row represents one day of observation and contains 11 attributes: date, minimum temperature, maximum temperature, average temperature, average relative humidity, rainfall, sunshine duration, maximum wind speed, average wind speed, wind direction at maximum speed, and prevailing wind direction.

Missing values were identified in three variables. The rainfall variable (RR) contained 94 missing observations, the minimum-temperature variable (TN) contained two missing observations, and the sunshine-duration variable (SS) contained one missing observation. Recorded rainfall ranged from 0 to 105 mm, with a standard deviation of 12.41 mm. The dataset contained 467 rainy days, representing 49.58% of the observations, and 475 non-rainy days, representing 50.42%. Based on the rainfall-intensity categories used in the dataset, 384 observations were classified as light rainfall, 65 as moderate rainfall, 17 as heavy rainfall, and one as extreme rainfall, corresponding to 0.11% of the total observations. The variables included in the dataset are summarized in Table 1.

Table 1. Meteorological variables included in the research dataset.

No.	Variable	Description
1	TANGGAL	Date of measurement
2	TN	Minimum temperature (°C)
3	TX	Maximum temperature (°C)
4	TAVG	Average temperature (°C)
5	RH_AVG	Average relative humidity (%)
6	RR	Daily rainfall (mm)
7	SS	Sunshine duration (hours)
8	FF_X	Maximum wind speed (m/s)
9	FF_AVG	Average wind speed (m/s)
10	DDD_X	Wind direction at maximum speed (°)
11	DDD_CAR	Prevailing wind direction

The TANGGAL variable records the date of each daily observation and was converted into a datetime format to preserve temporal ordering and facilitate time-series analysis. Minimum, maximum, and average temperatures are expressed in degrees Celsius and describe daily temperature variability. Average relative humidity is reported as a percentage and represents atmospheric moisture conditions associated with rainfall formation.

Rainfall is measured in millimeters and serves as the target variable. Sunshine duration is measured in hours and represents the duration of solar radiation received at the Earth's surface during each day. Maximum and average wind speeds are expressed in meters per second, while wind direction at maximum speed is represented in degrees from $0°$ to $360°$. The prevailing wind direction is represented using compass-direction codes, such as N, E, S, and W, which provide information regarding the dominant movement of air masses.

3.2. Data Preprocessing

The preprocessing stage was essential for transforming the raw meteorological records into a format compatible with the Gradient Boosting Regressor. The main procedures focused on standardizing the data format, converting variable types, and handling incomplete observations. The original daily weather dataset used commas as decimal separators in accordance with the local numerical format. Therefore, a string-replacement procedure was applied to convert commas into periods so that the values could be recognized as standard decimal numbers.

The meteorological variables TN, TX, TAVG, RH_AVG, RR, and SS, which were initially stored as object-type variables, were converted into floating-point numbers using the `pd.to_numeric` function. Numerical conversion was required because the Gradient Boosting algorithm performs mathematical operations on the predictor values when minimizing the loss function. The TANGGAL variable was converted from text into a datetime format to preserve the chronological order of the observations. This temporal ordering subsequently served as the basis for chronological data splitting and time-series validation. The data-type conversions performed during preprocessing are summarized in Table 2.

Table 2. Data-type conversion performed during preprocessing.

No.	Variable	Before	After
1	TANGGAL	Text	Datetime
2	TN	Object	Float
3	TX	Object	Float
4	TAVG	Object	Float
5	RH_AVG	Object	Float
6	RR	Object	Float
7	SS	Object	Float
8	FF_X	Integer	Integer
9	FF_AVG	Integer	Integer
10	DDD_X	Integer	Integer
11	DDD_CAR	Object	Object

Several observations contained incomplete entries or non-numerical symbols, such as a dash (-). During the cleaning process, non-numerical entries and empty cells were identified and converted into NaN values. This conversion enabled the missing observations to be detected systematically before the imputation procedure was applied.

Deleting incomplete observations was avoided because the dataset contained only 942 daily records, and removing rows could reduce the amount of information available for model training. Instead, missing values were estimated using the K-Nearest Neighbors imputation method with $k = 5$. To preserve the validity of the evaluation and prevent data leakage, the imputation procedure was conducted separately for each chronological train–test split scenario. The `KNNImputer` was fitted exclusively to the training data to learn the multivariate relationships among the meteorological variables. The fitted imputer was subsequently used to transform both the training and testing sets.

This procedure ensured that the testing data remained unseen during model development and were not influenced by statistical information extracted from future observations. The se-

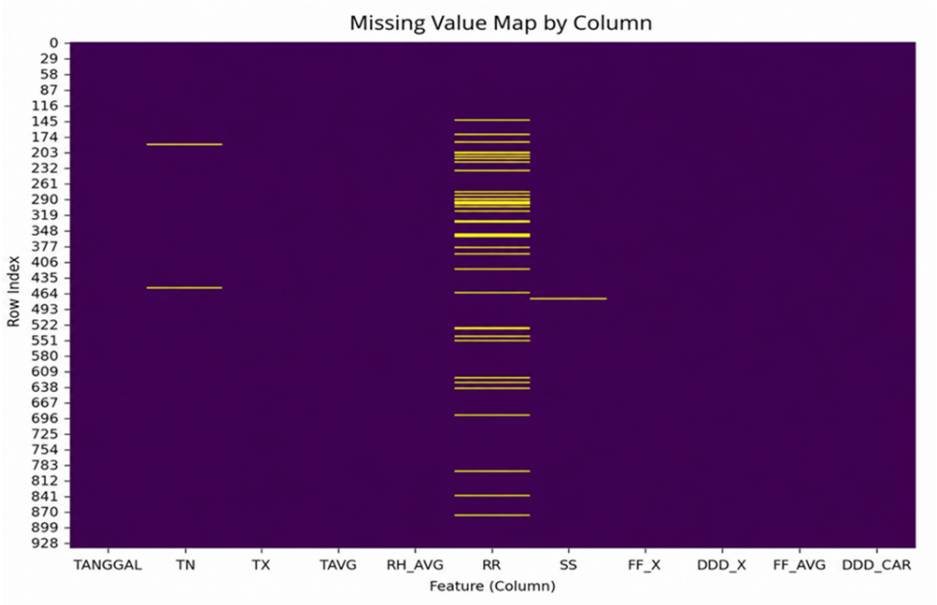


Figure 5. Distribution of missing values before KNN-based imputation.

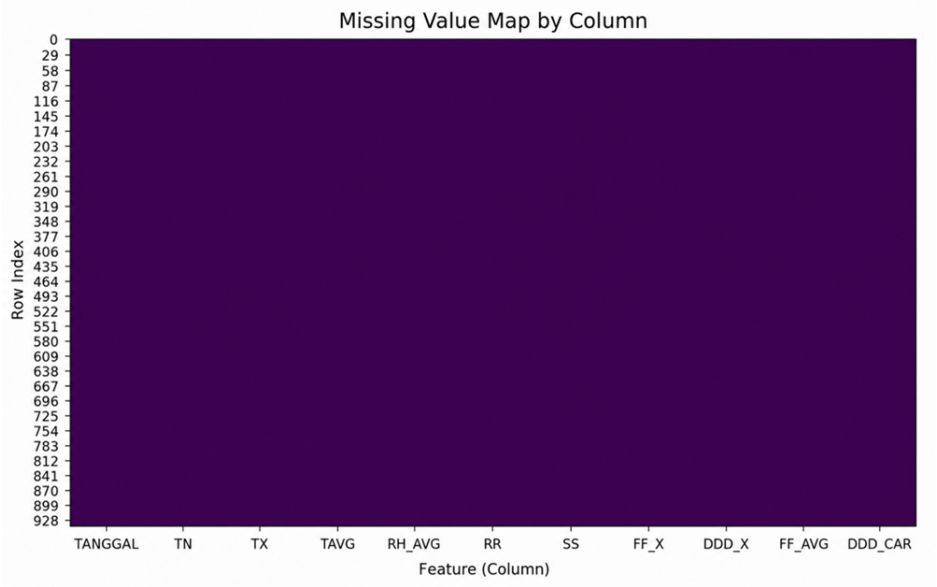


Figure 6. Distribution of missing values after KNN-based imputation.

lection of $k = 5$ was based on the multivariate relationships among meteorological variables. The KNN procedure estimates each missing value using five observations with the most similar meteorological characteristics, such as comparable temperature and humidity conditions. Compared with mean imputation, this approach is better able to preserve nonlinear relationships and the natural variability among weather variables. The distribution of missing observations before imputation is illustrated in Figure 5. Missing entries were identified primarily in the rainfall variable `RR`, with smaller numbers in `TN` and `SS`.

After the KNN-based imputation procedure, no missing values remained in the numerical variables, as shown in Figure 6. This result indicates that the preprocessing procedure successfully produced a complete dataset while retaining all 942 daily observations.

3.3. *Hyperparameter Optimization*

Following data preprocessing, hyperparameter optimization was conducted to identify the most effective Gradient Boosting Regressor configuration for daily rainfall prediction. The optimization procedure employed `RandomizedSearchCV` in combination with `TimeSeriesSplit`. A total of 50 hyperparameter combinations were sampled randomly from the predefined search space.

The use of `TimeSeriesSplit` ensured that the chronological order of the observations was preserved throughout the validation process. This procedure is particularly important for time-series forecasting because future observations must not be used to inform model training on earlier periods. The experiment used `n_splits=7`, dividing the training data into seven chronological folds. In each fold, the model was trained exclusively on earlier observations and validated on subsequent observations, thereby

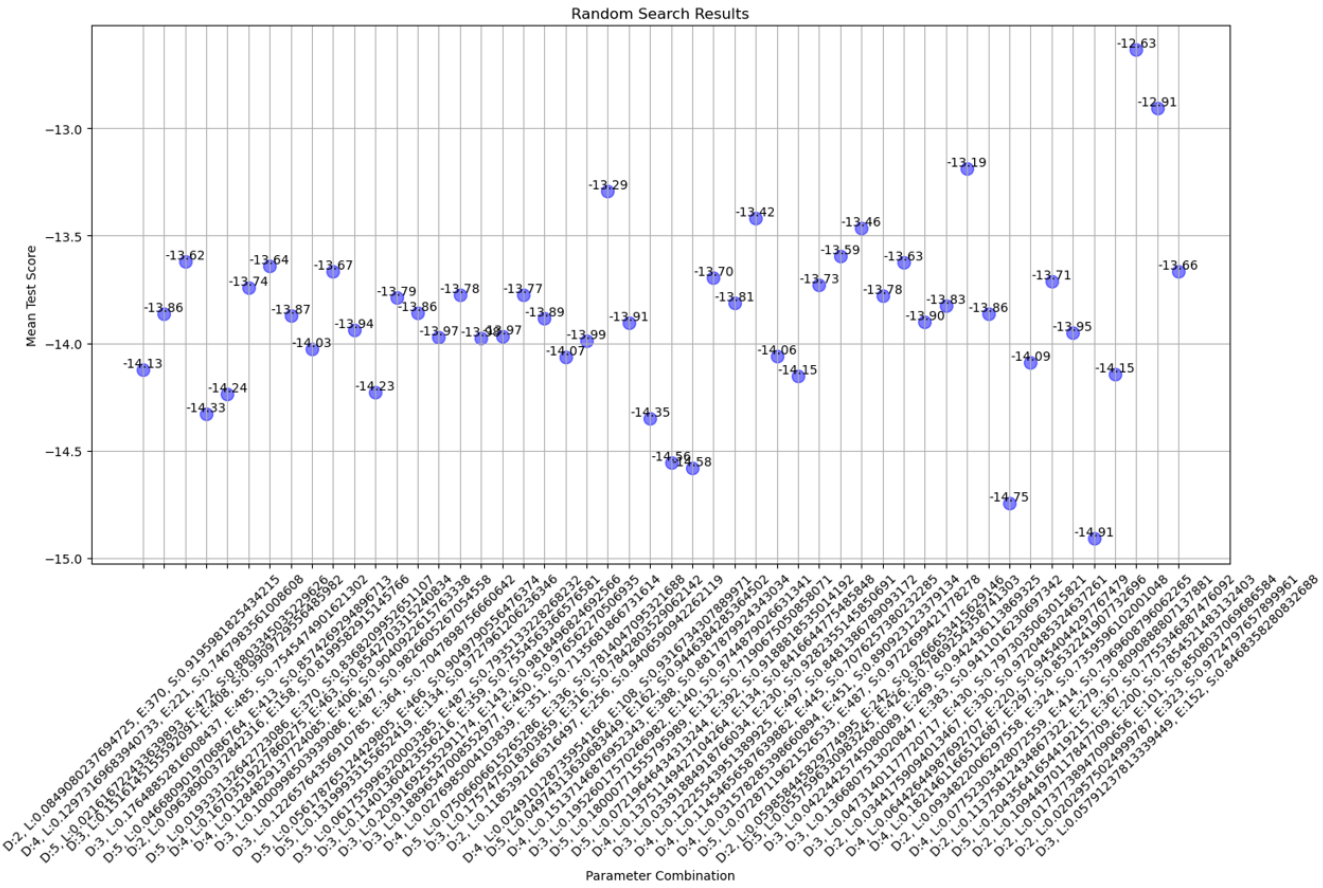


Figure 7. Hyperparameter optimization results obtained using `RandomizedSearchCV`.

reducing the risk of data leakage and providing a more realistic assessment of generalization to future data.

The optimization procedure used `neg_mean_absolute_error` as the scoring function, such that parameter configurations with lower MAE values were preferred. The `random_state` parameter was fixed at 42 to ensure reproducibility and consistency across repeated experiments. The optimal hyperparameter configuration obtained from the search procedure is presented in Table 3.

Table 3. Optimal hyperparameter configuration obtained using `RandomizedSearchCV`.

No.	Hyperparameter	Description	Selected value
1	`n_estimators`	Number of decision trees constructed	101
2	`learning_rate`	Rate at which the model corrects residual errors	0.017
3	`max_depth`	Maximum depth of each decision tree	2
4	`subsample`	Fraction of training observations used for each tree	0.850
5	RMSE score	Root mean squared prediction error	12.63

The optimized configuration consisted of 101 estimators, a learning rate of 0.017, a maximum tree depth of 2, and a subsampling ratio of 0.850. This relatively small learning rate, combined with shallow decision trees, indicates that the model applied gradual residual correction while limiting the complexity of individual weak learners. The subsampling ratio further introduced stochasticity into the training process, which may help reduce variance and overfitting. The best configuration produced an RMSE of 12.63 during the seven-fold time-series cross-validation procedure. The optimization results are visualized in Figure 7.

The figure presents the performance of the parameter combinations evaluated during the randomized search and highlights the configuration associated with the lowest validation error. The selected hyperparameter configuration was subsequently applied to the four chronological train–test split scenarios. Although the amount of training data differed across the scenarios, the optimized model exhibited relatively consistent prediction errors. This consistency suggests that the selected configuration was not excessively dependent on a particular training-set size and was able to capture recurring relationships among the meteorological predictors.

Nevertheless, the similarity of the validation errors across different split scenarios should not be interpreted as definitive evidence of complete robustness. Each scenario evaluates the model on a different testing period and therefore may contain different rainfall distributions, seasonal characteristics, and extreme events. Consequently, the final model selection was based on the joint interpretation of MAE, RMSE, and R^2 on the chronological hold-out sets rather than on the cross-validation RMSE alone.

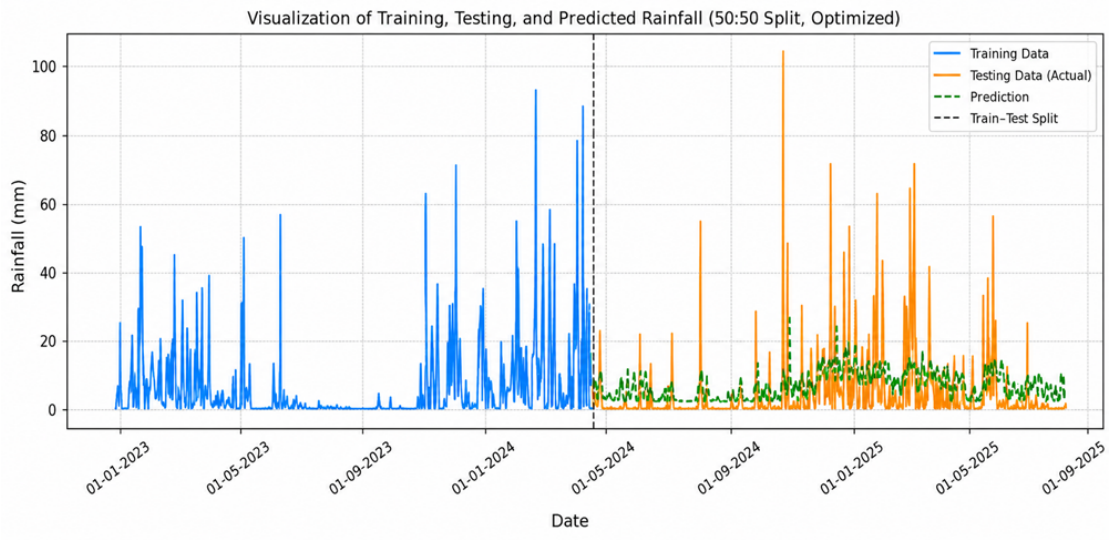


Figure 8. Time-series visualization of the optimized Gradient Boosting model under the 50:50 train–test split scenario.

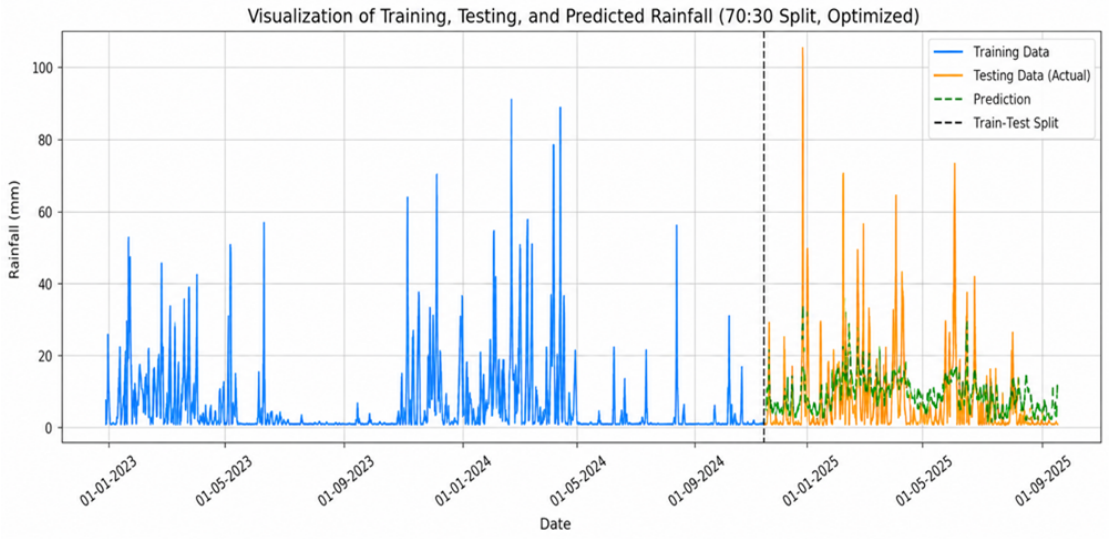


Figure 9. Time-series visualization of the optimized Gradient Boosting model under the 70:30 train–test split scenario.

3.4. Modeling Results

During the modeling stage, four chronological train–test split scenarios were evaluated. In each scenario, the Gradient Boosting Regressor was trained using the same optimized hyperparameter configuration obtained in the previous stage. The first scenario employed a 50:50 split, where 50% of the observations were used for training and the remaining 50% for testing. The second scenario used a 70:30 split, the third scenario used an 80:20 split, and the final scenario adopted a 90:10 split.

The purpose of evaluating multiple split ratios was to analyze the sensitivity of model performance to the amount of training data. Because rainfall exhibits complex stochastic behavior, it was necessary to examine whether the model required a large amount of historical data to converge or whether it could already learn meaningful rainfall patterns from a more limited training set. This strategy also avoided selecting the train–test ratio subjectively and instead allowed the most stable configuration to be determined empirically. By presenting all four scenarios, the study provides a transparent assessment of the minimum amount of historical data required for the model to maintain satisfactory generalization capability.

The 50:50 scenario provides the most conservative training configuration and the largest testing set, making it useful for evaluating whether the model can generalize without relying excessively on the training data. The 70:30 and 80:20 scenarios offer a more balanced allocation between training and testing data and are therefore expected to provide stable and representative performance estimates. In contrast, the 90:10 scenario allows the Gradient Boosting model to learn from the largest possible amount of historical data, which may be advantageous for recognizing rare or extreme rainfall events, but it also leaves a much smaller testing set for evaluation. The visualization of the predicted and observed rainfall series for the 50:50 scenario is presented in Figure 8.

The figure displays the training data in blue, the observed testing data in orange, and the predicted values in a dashed green line. Visually, the model captures the general low-to-moderate rainfall fluctuations, but it tends to underestimate extreme rain-

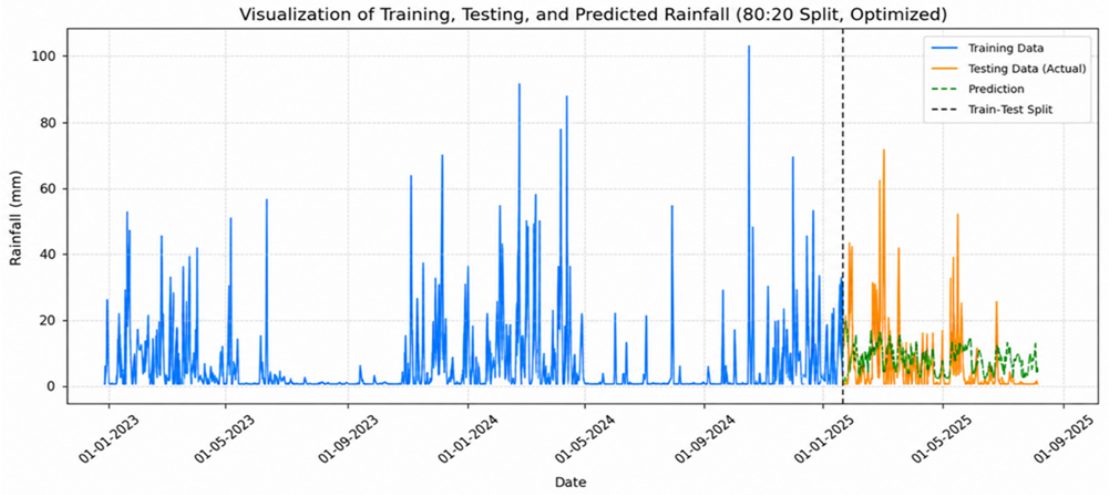


Figure 10. Time-series visualization of the optimized Gradient Boosting model under the 80:20 train–test split scenario.

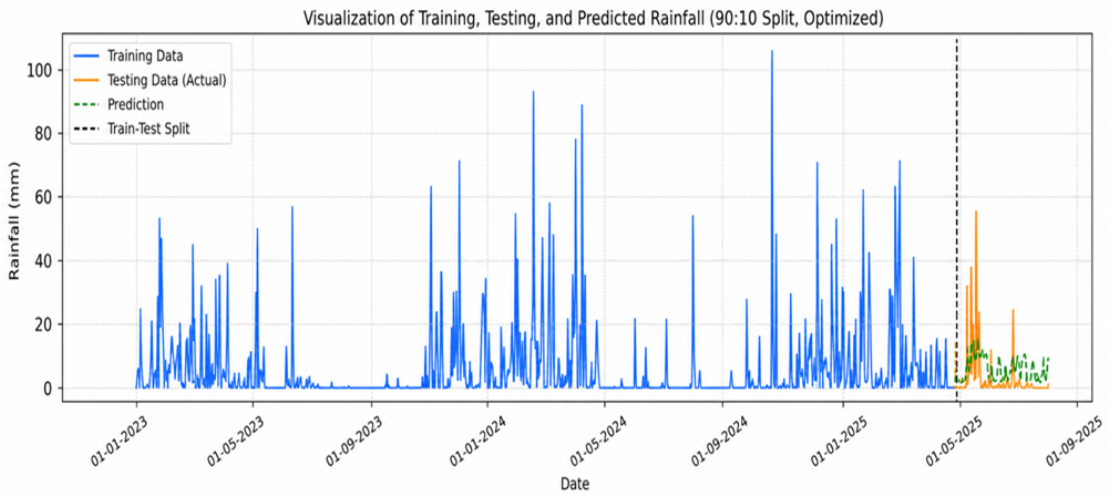


Figure 11. Time-series visualization of the optimized Gradient Boosting model under the 90:10 train–test split scenario.

fall events. The predicted series appears smoother than the observed series and fails to reproduce several substantial rainfall peaks above 40 mm. This pattern suggests that, under the 50:50 split, the model still exhibits relatively high bias and tends to produce predictions closer to the average rainfall level than to the extreme observations.

The 70:30 scenario is shown in Figure 9. Compared with the previous configuration, the model predictions follow the trend of the observed rainfall series more closely. The predicted line becomes more responsive to daily rainfall fluctuations and exhibits smaller deviations from the actual observations. A particularly important improvement is observed in the model's ability to respond to higher rainfall intensities, including several events above 40 mm. This result indicates that the increased amount of training data, combined with the optimized hyperparameter configuration, improved the model's flexibility and reduced the underfitting tendency observed in the 50:50 scenario.

The 80:20 scenario, illustrated in Figure 10, provides the most visually consistent agreement between the predicted and observed rainfall values among all scenarios. The predicted series shows a sharper and more accurate response to daily rainfall variability, including periods with relatively high rainfall intensity. With 80% of the observations allocated to training, the model appears to have learned the seasonal and short-term rainfall structure more effectively, while the optimized hyperparameters prevent the model from becoming excessively smooth. Visually, this scenario achieves the best balance between learning capacity and generalization.

The final 90:10 scenario is shown in Figure 11. Although the model had access to the largest portion of training data, its performance on the final 10% testing set still reveals a tendency to underestimate several rainfall peaks. The predicted values appear relatively stable, yet they do not fully follow the abrupt increases observed in the actual rainfall series. This behavior suggests that, despite hyperparameter optimization, the model may have become overly adapted to the dominant patterns present in the earlier 90% of the data, thereby reducing its sensitivity to the specific characteristics of the final testing segment. From an analytical perspective, the 90:10 configuration exhibits indications of reduced flexibility in generalizing to newly observed patterns and is therefore less favorable than the more balanced 70:30 and 80:20 scenarios.

Overall, the comparative visual analysis indicates that the model performance improves as the proportion of training data

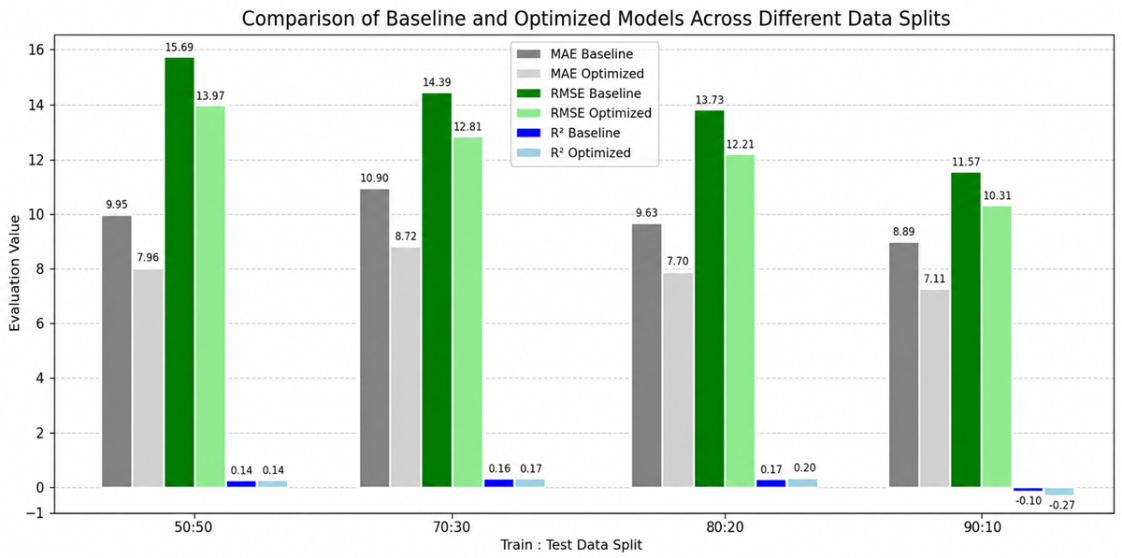


Figure 12. Comparison of MAE, RMSE, and R^2 across the four chronological train–test split scenarios.

increases from 50% to 80%, but this improvement does not continue monotonically when the training proportion reaches 90%. The 80:20 split appears to provide the most balanced and stable configuration, suggesting that an adequate amount of historical information is required for rainfall prediction, but an excessively small testing set may reduce the reliability of the evaluation and may expose the model to reduced sensitivity to new rainfall patterns.

3.5. Model Evaluation

Model evaluation was conducted using the chronological testing sets that were not observed during model training or hyperparameter optimization. This stage assessed the ability of the Gradient Boosting Regressor to generalize the learned meteorological patterns to future rainfall observations. The comparative evaluation results for the four train–test split scenarios are presented in Figure 12.

As shown in Figure 12, the four scenarios produced different prediction errors and coefficients of determination. The 50:50 scenario yielded an MAE of 7.96 mm, an RMSE of 13.97 mm, and an R^2 value of 0.14. The 70:30 scenario produced an MAE of 8.72 mm, an RMSE of 12.81 mm, and an R^2 value of 0.17. The 80:20 scenario achieved an MAE of 7.70 mm, an RMSE of 12.21 mm, and an R^2 value of 0.20, whereas the 90:10 scenario obtained the lowest MAE and RMSE values, namely 7.11 mm and 10.31 mm, respectively, but generated a negative R^2 value of -0.27.

Differences in RMSE indicate that the models varied in their sensitivity to large prediction errors. A lower RMSE generally indicates better performance in estimating high-intensity rainfall because this metric assigns greater weight to large residuals. The MAE values represent the average absolute daily prediction errors in millimeters and therefore provide a direct interpretation of the typical deviation between observed and predicted rainfall. Meanwhile, R^2 measures the proportion of rainfall variability explained by the meteorological predictors.

The 80:20 scenario was selected as the best-performing configuration because it provided the most balanced combination of prediction error and generalization capability. Its MAE of 7.70 mm and RMSE of 12.21 mm were accompanied by the highest positive R^2 value of 0.20. This result indicates that the model retained relatively low prediction errors while explaining a greater proportion of rainfall variability than the other scenarios.

Although the 90:10 scenario produced numerically lower MAE and RMSE values, its negative R^2 value indicates that the model performed worse than a simple baseline that predicts the mean rainfall value of the testing set. The final 10% testing segment may not have represented the full range of rainfall variability, particularly rare high-intensity events. Because daily rainfall has a highly skewed distribution, a model that produces predictions near the average may obtain relatively small absolute errors when the testing set contains predominantly low rainfall values, while still failing to reproduce the temporal variability of the observations.

In contrast, the larger testing set in the 80:20 scenario provided a more representative evaluation of the observed rainfall distribution. Consequently, selecting this scenario constitutes a more conservative and reliable decision because its performance reflects genuine generalization rather than an apparently low prediction error obtained from a limited testing segment. A negative R^2 value in the 90:10 scenario also suggests that allocating more observations to training does not necessarily improve model performance when the remaining testing period contains patterns that differ from those dominant in the historical training data.

The relatively low R^2 values obtained in this study reflect the difficulty of predicting daily rainfall using standard surface meteorological variables alone. Rainfall is highly stochastic and may be influenced by atmospheric pressure, local convection, topographic interactions, cloud dynamics, and large-scale climate variability that were not explicitly represented in the dataset. Therefore, the model should not be interpreted as a stand-alone system capable of producing precise deterministic forecasts. Instead, it is more appropriately positioned as a decision-support tool that provides preliminary rainfall estimates for meteorological monitoring, hydrometeorological risk mitigation, and agricultural planning.

4. Conclusion

This study evaluated the performance of a hyperparameter-optimized Gradient Boosting Regressor for daily rainfall prediction in Malang Regency using BMKG meteorological data from 2023 to 2025. Hyperparameter optimization was performed using `RandomizedSearchCV` with time-series cross-validation, and model performance was assessed under four chronological train–test split scenarios.

The results showed that the 80:20 scenario provided the most balanced and consistent predictive performance, with an MAE of 7.70 mm, an RMSE of 12.21 mm, and an R^2 value of 0.20. This scenario was selected because it combined relatively low prediction errors with the highest positive explanatory performance. Although the 90:10 scenario produced lower MAE and RMSE values, its negative R^2 value indicated limited generalization and an inability to adequately reproduce rainfall variability in the testing period.

The findings demonstrate that hyperparameter optimization and chronological validation can improve the stability of Gradient Boosting for rainfall prediction. Nevertheless, the R^2 value of 0.20 indicates that a substantial proportion of daily rainfall variability remains unexplained. This limitation is associated with the stochastic nature of rainfall and the absence of several atmospheric and physical processes that cannot be represented solely by surface-station variables.

The proposed model is entirely data-driven and does not yet incorporate complex atmospheric-dynamics parameters. Consequently, its ability to capture anomalous and extreme rainfall events remains limited. Future studies should therefore incorporate additional predictors, including Outgoing Longwave Radiation, sea-surface temperature anomalies, atmospheric-pressure variables, and satellite imagery, to improve sensitivity to regional and large-scale weather patterns. Further investigation of ensemble and deep-learning approaches, such as Long Short-Term Memory networks, may also improve the representation of longer and more complex temporal dependencies in meteorological data.

Author Contributions. **Mohamad Arif Abdul Syukur**: Software, validation, formal analysis, investigation, visualization, writing–original draft preparation. **Suhartono**: Conceptualization, methodology, supervision. **Mochamad Imamudin**: Resources, data curation, supervision.

Acknowledgement. The authors would like to express their gratitude to the editors and reviewers for their valuable comments, constructive suggestions, and support in improving the quality of this work.

Funding. This research received no external funding.

Conflict of interest. The authors declare that there is no conflict of interest related to this article.

Data availability. The weather dataset used in this study was obtained from the Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG). The dataset generated and analyzed during this study is publicly available at https://dataonline.bmkg.go.id.

References

[1] D. Oettl and G. Veratti, "A comparative study of mesoscale flow-field modelling in an Eastern Alpine region using WRF and GRAMM-SCI," *Atmospheric Research*, vol. 249, p. 105288, 2021. doi: 10.1016/j.atmosres.2020.105288.

[2] A. C. Travieso Bello, O. F. Martínez, M. L. Hernández Aguilar, and J. C. Morales Hernández, "Comprehensive risk management of hydrometeorological disaster: A participatory approach in the metropolitan area of Puerto Vallarta, Mexico," *International Journal of Disaster Risk Reduction*, vol. 87, p. 103578, 2023. doi: 10.1016/j.ijdrr.2023.103578.

[3] R. Meenal, K. Kailash, P. A. Michael, J. J. Joseph, F. T. Josh, and E. Rajasekaran, "Machine learning based smart weather prediction," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 1, pp. 508–515, 2022. doi: 10.11591/ijeecs.v28.i1.pp508-515.

[4] R. Prasad and P. Kumar Shukla, "Weather forecasting using machine learning for smart farming," in *Future Farming: Advancing Agriculture with Artificial Intelligence*. Bentham Science Publishers, 2023, pp. 97–113. doi: 10.2174/9789815124729123010009.

[5] Suhartono and Subanar, "The effect of decomposition method as data preprocessing on neural networks model for forecasting trend and seasonal time series," *Jurnal Teknik Industri*, vol. 8, no. 2, pp. 156–164, 2006.

[6] V. S. Monego, J. A. Anochi, and H. F. de Campos Velho, "South America seasonal precipitation prediction by gradient-boosting machine-learning approach," *Atmosphere*, vol. 13, no. 2, 2022. doi: 10.3390/atmos13020243.

[7] N. K. A. Appiah-Badu, Y. A. W. M. Missah, L. K. Amekudzi, N. Ussiph, T. Frimpong, and E. Ahene, "Rainfall prediction using machine learning algorithms for the various ecological zones of Ghana," *IEEE Access*, vol. 10, pp. 5069–5082, 2022. doi: 10.1109/ACCESS.2021.3139312.

[8] C. D. Usman, A. P. Widodo, K. Adi, and R. Gernowo, "Rainfall prediction model in Semarang City using machine learning," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 30, no. 2, pp. 1224–1231, 2023. doi: 10.11591/ijeecs.v30.i2.pp1224-1231.

[9] M. Alvines et al., "Komparasi Ridge Regression, Random Forest, dan Gradient Boosting untuk prediksi curah hujan harian di Sumatra Selatan berbasis time series cross-validation," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 4, pp. 5742–5748, 2025. doi: 10.36040/jati.v9i4.13915.

[10] M. T. Anwar, E. Winarno, W. Hadikurniawati, and M. Novita, "Rainfall prediction using Extreme Gradient Boosting," *Journal of Physics: Conference Series*, vol. 1869, no. 1, 2021. doi: 10.1088/1742-6596/1869/1/012078.

[11] M. Hassan et al., "Machine learning-based rainfall prediction: Unveiling insights and forecasting for improved preparedness," *IEEE Access*, vol. 11, pp. 132196–132222, 2023. doi: 10.1109/ACCESS.2023.3333876.

[12] H. Chen, V. Chandrasekar, H. Tan, and R. Cifelli, "Rainfall estimation from ground radar and TRMM precipitation radar using hybrid deep neural networks," *Geophysical Research Letters*, vol. 46, nos. 17–18, pp. 10669–10678, 2019. doi: 10.1029/2019GL084771.

[13] H. Xu, G. Zhai, and X. Li, "Convective-stratiform rainfall separation of Typhoon Fitow (2013): A 3D WRF modeling study," *Terrestrial, Atmospheric and Oceanic Sciences*, vol. 29, no. 3, pp. 315–329, 2018. doi: 10.3319/TAO.2017.10.11.01.

[14] W. Sun et al., "Improved prediction of extreme rainfall using a machine learning approach," *Advances in Atmospheric Sciences*, vol. 42, no. 8, pp. 1661–1674, 2025. doi: 10.1007/s00376-024-4269-5.

[15] M. S. Islam et al., "Explainable deep learning for rainfall prediction: A CNN-XGBoost hybrid approach in the northern region of Bangladesh," *Neural Computing and Applications*, vol. 37, no. 33, pp. 28125–28160, 2025. doi: 10.1007/s00521-025-11646-z.

[16] A. Y. Barrera-Animas, L. O. Oyedele, M. Bilal, T. D. Akinosho, J. M. D. Delgado, and L. A. Akanbi, "Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting," *Machine Learning with Applications*, vol. 7, p. 100204, 2022. doi: 10.1016/j.mlwa.2021.100204.

[17] X. Zhang, S. N. Mohanty, A. K. Parida, S. K. Pani, B. Dong, and X. Cheng, "Annual and non-monsoon rainfall prediction modelling using SVR-MLP: An empirical study from Odisha," *IEEE Access*, vol. 8, pp. 30223–30233, 2020. doi: 10.1109/ACCESS.2020.2972435.

[18] V. Svetnik, T. Wang, C. Tong, A. Liaw, R. P. Sheridan, and Q. Song, "Boosting: An ensemble learning tool for compound classification and QSAR modeling," *Journal of Chemical Information and Modeling*, vol. 45, no. 3, pp. 786–799, 2005. doi: 10.1021/ci0500379.

[19] T. Talan, "Machine learning-based rainfall prediction across temporal scales: Model benchmarking and explainability analysis," *Stochastic Environmental Research and Risk Assessment*, vol. 40, no. 5, pp. 1–17, 2026. doi: 10.1007/s00477-026-03245-8.

[20] I. U. Hassan, Z. A. Lone, S. Swati, and A. Gamal, "Forecasting weather and water management through machine learning," pp. 71–93, 2023. doi: 10.4018/979-8-3693-1194-3.ch004.

[21] A. Pambudi, "Penerapan CRISP-DM menggunakan MLR K-Fold pada data saham PT Telkom Indonesia (Persero) Tbk (TLKM): Studi kasus Bursa Efek Indonesia tahun 2015–2022," *Jurnal Data Mining dan Sistem Informasi*, vol. 4, no. 1, p. 1, 2023. doi: 10.33365/jdmsi.v4i1.2462.

[22] C. Zoremsanga and J. Hussain, "Particle swarm optimized deep learning models for rainfall prediction: A case study in Aizawl, Mizoram," *IEEE Access*, vol. 12, pp. 57172–57184, 2024. doi: 10.1109/ACCESS.2024.3390781.

[23] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation," *Expert Systems with Applications*, vol. 244, p. 122778, 2024. doi: 10.1016/j.eswa.2023.122778.

[24] M. A. Rodriguez-Ramirez and Ó. A. Fuentes-Mariles, "Daily rainfall assimilation based on satellite and weather radar precipitation products along with rain gauge networks," *Journal of Hydroinformatics*, vol. 25, no. 6, pp. 2354–2368, 2023. doi: 10.2166/hydro.2023.104.

[25] Z. Cui, X. Qing, H. Chai, S. Yang, Y. Zhu, and F. Wang, "Real-time rainfall-runoff prediction using light gradient boosting machine coupled with singular spectrum analysis," *Journal of Hydrology*, vol. 603, p. 127124, 2021. doi: 10.1016/j.jhydrol.2021.127124.

[26] V. Kumar, N. Kedam, K. V. Sharma, K. M. Khedher, and A. E. Alluqmani, "A comparison of machine learning models for predicting rainfall in urban metropolitan cities," *Sustainability*, vol. 15, no. 18, 2023. doi: 10.3390/su151813724.

[27] M. I. Hastuti, K. H. Min, and J. W. Lee, "Improving radar data assimilation forecast using advanced remote sensing data," *Remote Sensing*, vol. 15, no. 11, 2023. doi: 10.3390/rs15112760.

[28] N. Gill, P. Hall, K. Montgomery, and N. Schmidt, "A responsible machine learning workflow with focus on interpretable models, post-hoc explanation, and discrimination testing," *Information*, vol. 11, no. 3, pp. 1–32, 2020. doi: 10.3390/info11030137.