



NEXT-DAY RAINFALL CATEGORY PREDICTION USING ENSEMBLE MACHINE LEARNING AND SHAP

Ulfa Emi Rahmawati¹, Muhammad Galang Satrio Wicaksono*²

¹ Program Studi Teknik Informatika, Jurusan Teknologi Informasi, Politeknik Negeri Jember

² Program Studi Sains Data, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional “Veteran” Jawa Timur

*Corresponding author: muhammad_galang.sada@upnjatim.ac.id

ABSTRACT

Reliable day-ahead rainfall-category classification is challenging when the observation period is short and rainfall classes are highly imbalanced. This study re-evaluates a 2025 daily meteorological dataset using a chronological holdout and nested time-series hyperparameter selection. A source file contained 366 rows, but one row had no date; after excluding that malformed record, 365 dated daily records remained and 364 day-ahead supervised examples were available. Rainfall was grouped into five operational categories: no rain/trace (<0.5 mm), light ($0.5\text{--}<20$ mm), moderate ($20\text{--}<50$ mm), heavy ($50\text{--}<100$ mm), and very heavy/extreme (≥ 100 mm). Random Forest, XGBoost, LightGBM, logistic regression, a majority-class baseline, and a persistence baseline were compared using macro-F1 as the common tuning criterion. The final 20% chronological test set contained 73 examples. Random Forest achieved the highest macro-F1 (0.327; bootstrap 95% CI 0.202–0.441) and balanced accuracy (0.309), whereas XGBoost achieved the highest accuracy (0.466) but a lower macro-F1 (0.251). No model correctly classified the single very-heavy/extreme test example, and no model detected the four heavy-rain examples consistently. Therefore, the earlier claim that XGBoost was the best overall model is not supported under the revised evaluation. The results should be interpreted as exploratory because the series covers only one year and the station metadata are absent from the supplied file.

Keywords: Rainfall Prediction; BMKG; Ensemble Learning; Feature Engineering; SHAP

I. INTRODUCTION

Daily rainfall classification supports weather communication, agricultural planning, and hydrometeorological risk assessment. However, rainfall observations commonly contain many dry or low-rainfall days and relatively few heavy-rainfall events. This zero-inflated distribution makes overall accuracy an incomplete measure of model quality: a classifier can appear successful by predicting common categories while missing the rarer events that may be operationally important. BMKG daily rainfall guidance distinguishes rainfall levels using intensity thresholds, including 0.5–20 mm/day for light rain, 20–50 mm/day for moderate rain, 50–100 mm/day for heavy rain, and 100–150 mm/day for very heavy rain; higher values may be classified as extreme [1].

Machine-learning methods such as Random Forest, XGBoost, and LightGBM can represent nonlinear relationships in tabular meteorological data [2]–[4]. Indonesian studies have applied data-mining and machine-learning techniques to BMKG rainfall or weather data, but reported results are not directly comparable when targets, temporal partitions, and evaluation metrics differ [5], [6]. Explainability methods such as SHAP can describe how fitted models use predictors, but feature attribution should not be interpreted as evidence of physical causation [7].

A key methodological concern is validation design. Randomly shuffled train/test partitions can allow future observations to influence model development when the data are temporally ordered. Time-series validation instead trains on earlier observations and evaluates on later ones [8]. This is especially important for a one-year series, where rare categories may be absent or represented by only a few examples. Model tuning should use the same scoring metric across candidate algorithms, and the final comparison should include simple baselines rather than only sophisticated ensembles.

This study therefore asks: (RQ1) how do ensemble models compare with logistic regression, majority-class, and persistence baselines

* Corresponding author.

E-mail address: muhammad_galang.sada@upnjatim.ac.id

<http://doi.org/10.33005/jasid.v2i2.65>

under a chronological holdout? (RQ2) how do macro-F1, balanced accuracy, class-wise metrics, and uncertainty change the interpretation of model performance? (RQ3) which predictors contribute most to the selected model's outputs according to SHAP? The contribution is a transparent re-evaluation of the available dataset rather than a claim of operationally reliable forecasting.

II. RELATED WORK

Breiman introduced Random Forest as a bagging-based ensemble of decision trees [2]. XGBoost combines gradient boosting with regularization and efficient tree construction [3], while LightGBM uses histogram-based learning and leaf-wise growth to improve efficiency [4]. These algorithms are commonly applied to structured data, but algorithm choice alone does not ensure reliable performance on rare classes. A recent Indonesian study used XGBoost to estimate rainfall through the integration of rain-gauge, radar, and satellite data, illustrating the value of multi-source observations compared with a short single-station series [6]. Other Indonesian work has explored classification of BMKG rainfall and daily weather categories [5], [9].

Macro-F1 weights each class equally, while balanced accuracy averages class-specific recall [10]. Matthews correlation coefficient (MCC) is a further summary for multiclass confusion [11]. Standard machine-learning workflows and preprocessing tools are documented by scikit-learn [12], while general predictive-modeling practice emphasizes separating model selection from final evaluation [13]. Class weighting and resampling are commonly discussed approaches for imbalanced data [14], [15], but they cannot compensate for a class represented by only one or two observations. Model comparison should be interpreted cautiously when data are limited [16]. Time-series validation is particularly important where observations are temporally dependent [8], [17], [18]. Further work on multiclass correlation metrics supports using MCC as a complementary summary [19], and systematic reviews of classification metrics emphasize matching metrics to the problem structure [20]. Cohen's kappa provides a chance-corrected agreement measure [21]. Classical decision-tree methods provide the broader methodological context for tree-based classification [22], and Liaw and Wiener describe Random Forest implementation details [23]. Bootstrap intervals can help communicate uncertainty in estimated metrics [24]. Indonesian rainfall classification studies provide local context, although their data sources and validation designs differ from this study [5], [9], [25]. For model interpretation, SHAP assigns feature contributions to predictions based on Shapley values [7]; global importance rankings may differ from built-in tree importance because the two methods quantify different quantities.

III. RESEARCH METHOD

A. Research Data

The supplied CSV contains daily columns for date, temperature, relative humidity, rainfall, wind speed, and wind direction. The file has 366 rows, but one row has a missing date and missing wind direction. Because 2025 contains 365 calendar days, the undated row was excluded rather than assigned an invented date. The resulting dataset contains 365 dated observations from 1 January to 31 December 2025. Shifting the target by one day yields 364 supervised examples. The file supplied for this revision does not identify the station name, station identifier, coordinates, elevation, or source URL; these metadata must be recovered from the original download before a journal submission.

The rainfall variable contains 208 exact zero values among the 365 dated observations and 223 values below 0.5 mm. Values below 0.5 mm were grouped into a "no rain/trace" class to align with BMKG daily products that distinguish values below 0.5 mm/day from measurable rainfall [1]. Categories were defined as: <0.5 mm, no rain/trace; 0.5 - <20 mm, light; 20 - <50 mm, moderate; 50 - <100 mm, heavy; and ≥ 100 mm, very heavy/extreme. The final class combines very heavy and extreme rainfall because the supplied year contains only two daily observations at or above 100 mm; separate evaluation of those two subclasses would not be statistically defensible. The classification is an operational grouping of the available series, not a claim that all BMKG products use identical boundaries.

No missing rainfall values were present in the supplied cleaned CSV, but the provenance of the cleaning and any missing-value codes in the upstream source could not be independently verified. The raw source contains one malformed date row; this record was excluded. Rolling features at the beginning of the series use available prior observations only (partial windows), and binary flags indicate whether a complete 3-, 5-, or 7-day window is available. Median imputation values were learned from the training period only. These choices must be revisited if the original uncleaned data or station documentation becomes available.

B. Target and Temporal Features

The target for day t is the rainfall category observed on day $t+1$ (H+1). Predictors use meteorological measurements available on day t and trailing summaries ending on day t . The feature set includes temperature, humidity, rainfall, wind speed, 3-, 5-, and 7-day rainfall accumulation, 3-day rolling means of humidity and wind speed, full-window availability indicators, and one-hot-encoded wind direction. No future observations are used to compute predictors. Numeric missing values, if any, are imputed using training-period medians only, following a standard predictive-modeling workflow [12], [13]. The target classes and class counts are reported explicitly because rare heavy-rainfall events strongly affect the attainable metrics [14], [15].

C. Chronological Evaluation and Tuning

The 364 supervised examples were sorted chronologically and divided into the first 80% for model development (291 examples) and the final 20% for a held-out test (73 examples; target dates 20 October–31 December 2025). Hyperparameter selection used expanding-window TimeSeriesSplit with two folds within the training period. Macro-F1 was the single scoring metric used for tuning all tuned models; the final test set was not used for hyperparameter selection. Candidate models were logistic regression with standardization and class weighting, Random Forest, XGBoost, and LightGBM. Baselines were (i) majority class, which always predicts the most frequent training class, and (ii) persistence, which predicts tomorrow's category as the category observed today. The time-ordered evaluation follows guidance that time-series observations should not be randomly shuffled [8].

The final test metrics include accuracy, macro-F1, weighted-F1, balanced accuracy, MCC, Cohen's kappa, and per-class precision, recall, F1, and support [10], [11], [19]–[21]. Ninety-five-percent confidence intervals for macro-F1, balanced accuracy, and MCC were estimated by nonparametric bootstrap resampling of the 73 test examples (1,500 resamples; fixed random seed), a common approach to communicating uncertainty in performance estimates [24]. With such a small test set, these intervals are approximate and do not account for year-to-year variability. The same test set is used for all models, but no claim of statistically significant differences is made.

D. SHAP Interpretation

SHAP was computed for the selected Random Forest model using the held-out test feature matrix and a tree-based explainer. The dependence plot displays feature values against their SHAP contributions for the no-rain/trace class. SHAP contributions describe the fitted model's attribution of output differences; they do not establish that a predictor causes rainfall. Built-in tree impurity importance and SHAP values are not interchangeable: impurity importance reflects split-based reductions during training, whereas mean absolute SHAP summarizes contributions to predictions on a specified dataset. The dependence plot should therefore be interpreted as model-specific and test-set-specific.

E. Research Workflow

The research workflow consisted of data collection, preprocessing, exploratory data analysis (EDA), feature engineering, development of three ensemble models, evaluation and comparison, selection of the best-performing model, SHAP-based interpretation, and prediction simulation.

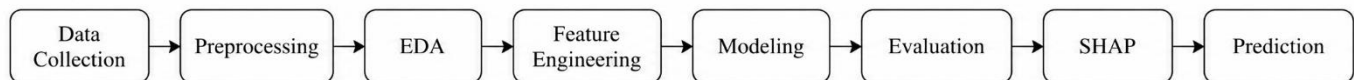


Fig. 1. Research workflow for next-day rainfall-category classification.

IV. RESULTS AND DISCUSSION

A. Data Distribution and Temporal Split

After excluding the undated row, the daily rainfall categories comprised 223 no-rain/trace observations, 90 light-rain observations, 31 moderate-rain observations, 19 heavy-rain observations, and 2 very-heavy/extreme observations. The 80:20 chronological split produced 291 training examples and 73 test examples. Test support was 29 no-rain/trace, 26 light, 13 moderate, 4 heavy, and 1 very-heavy/extreme example. The rarest class therefore represents only one observation in the final test period, making any estimate of its recall highly unstable.

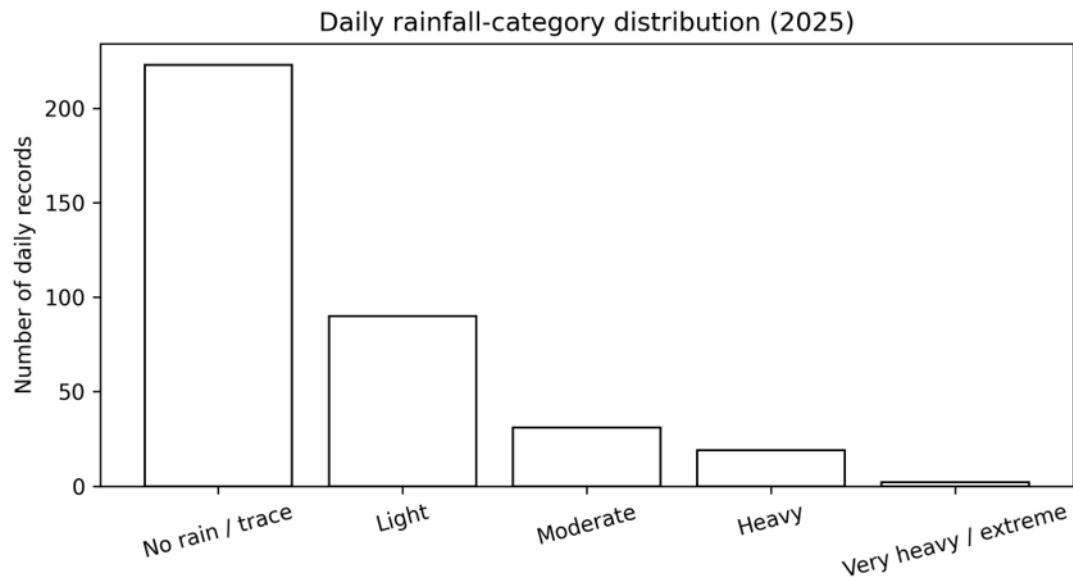


Fig. 2. Distribution of rainfall categories in the 2025 BMKG dataset.

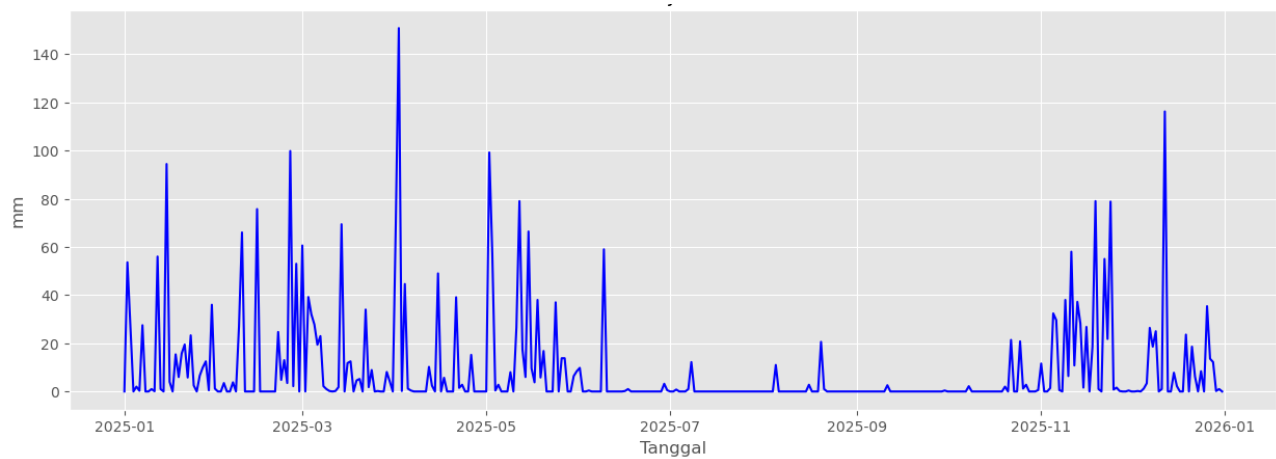


Fig. 3. Daily rainfall pattern during 2025.

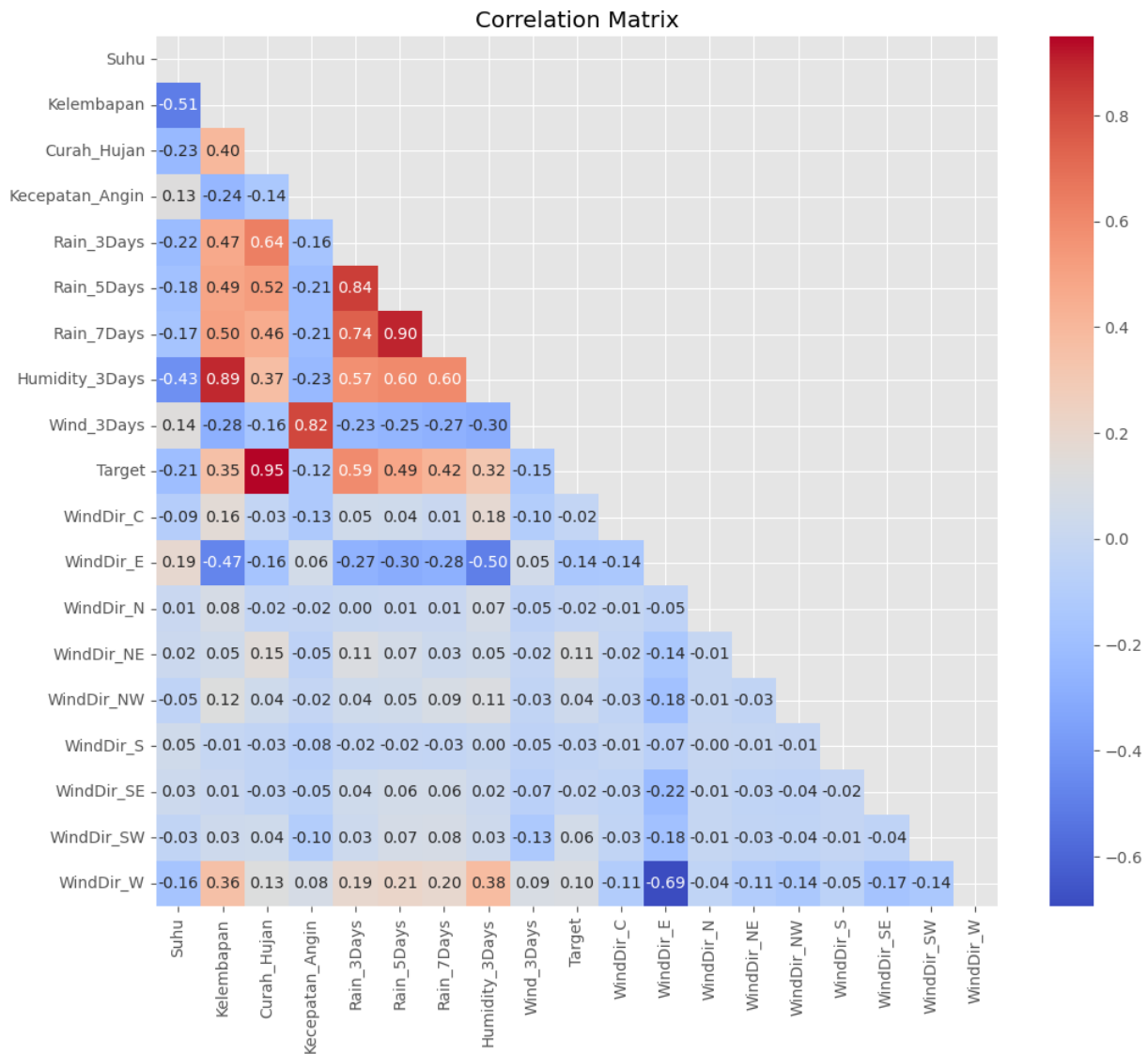


Fig. 4. Correlation matrix of meteorological and temporal features.

B. Overall Model Comparison

TABLE I. Chronological Holdout Performance; Macro-F1 Was the Common Tuning Metric.

Model	Accuracy	Macro-F1	Macro-F1 95% CI	Balanced acc.	MCC
Majority-class baseline	0.397	0.114	0.088–0.135	0.200	0.000
Persistence baseline	0.301	0.174	0.112–0.236	0.174	-0.027
Logistic regression	0.411	0.270	0.188–0.338	0.243	0.186
Random Forest	0.438	0.327	0.202–0.441	0.309	0.152
XGBoost	0.466	0.251	0.169–0.326	0.259	0.160
LightGBM	0.452	0.277	0.191–0.346	0.270	0.172

Random Forest achieved the highest macro-F1 (0.327; 95% bootstrap CI 0.202–0.441) and balanced accuracy (0.309). LightGBM reached macro-F1 0.277, logistic regression 0.270, XGBoost 0.251, persistence 0.174, and the majority-class baseline 0.114. XGBoost had the highest accuracy (0.466), but this did not translate into the best macro-F1 or balanced accuracy. Thus, the previous claim that XGBoost was the best overall model is withdrawn. Under the current protocol, Random Forest is the leading candidate by macro-F1, but its confidence interval is wide and its MCC confidence interval includes zero, so the evidence does not establish robust superiority.

TABLE II. Additional 95% Bootstrap Confidence Intervals for Balanced Accuracy and MCC.

Model	Balanced accuracy (95% CI)	MCC (95% CI)
Majority-class baseline	0.200 (0.200–0.250)	0.000 (0.000–0.000)
Persistence baseline	0.174 (0.118–0.287)	-0.027 (-0.171–0.119)
Logistic regression	0.243 (0.178–0.378)	0.186 (0.042–0.333)
Random Forest	0.309 (0.213–0.506)	0.152 (-0.025–0.321)
XGBoost	0.259 (0.202–0.394)	0.160 (-0.024–0.329)
LightGBM	0.270 (0.201–0.414)	0.172 (-0.003–0.337)

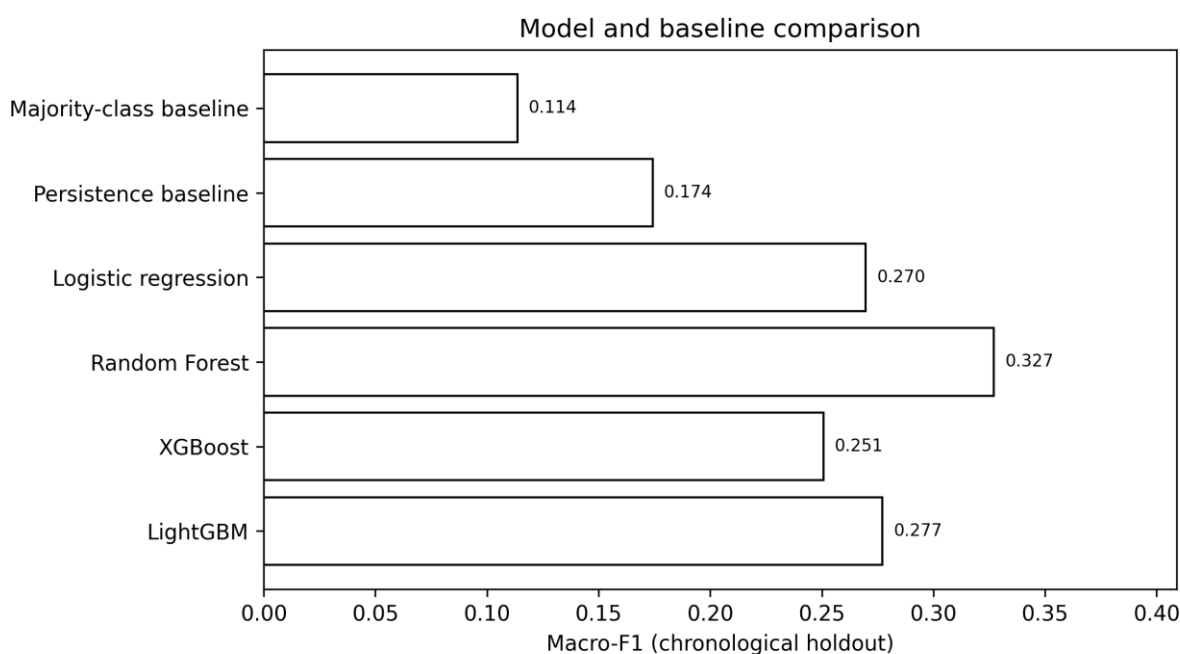


Fig. 5. Comparison of accuracy, precision, recall, and F1-score.

C. Per-class performance

The XGBoost confusion matrix on the test set contained 73 observations. Sixty Light observations were correctly classified, while three Light observations were predicted as Moderate. For the Moderate class, one observation was correctly classified and five were predicted as Light. All four Heavy observations in the test set were predicted as Light. This pattern explains why weighted accuracy and recall remained relatively high, while the model still had limited ability to detect higher rainfall categories.

TABLE II. Random Forest Class-Wise Metrics on the Test Set.

Target class	Support	Precision	Recall	F1
--------------	---------	-----------	--------	----

No rain/trace (<0.5)	29	0.542	0.448	0.491
Light (0.5–<20)	26	0.359	0.538	0.431
Moderate (20–<50)	13	0.500	0.308	0.381
Heavy (50–<100)	4	0.500	0.250	0.333
Very heavy/extreme (≥ 100)	1	0.000	0.000	0.000

Random Forest obtained F1 values of 0.491 for no-rain/trace, 0.431 for light, 0.381 for moderate, and 0.333 for heavy rainfall. It failed to identify the single very-heavy/extreme example. XGBoost and LightGBM also failed to identify that example, and all models had zero recall for at least some rare classes. This is a substantive limitation, not a cosmetic metric issue: the available data do not support a reliable claim about very heavy rainfall detection.

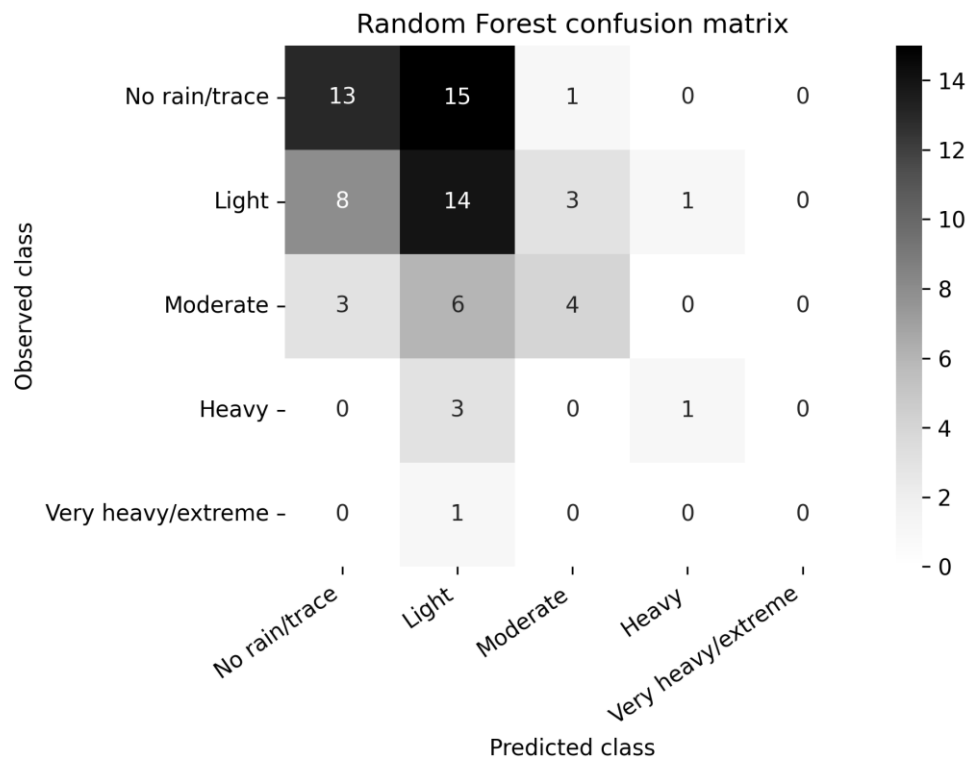


Fig. 6. Random Forest confusion matrix. Rows are observed categories and columns are predicted categories.

D. Feature Importance and SHAP Analysis

The dependence plot is class-specific to No rain/trace (<0.5); feature shown is Kelembapan. SHAP values describe model attribution, not causality.

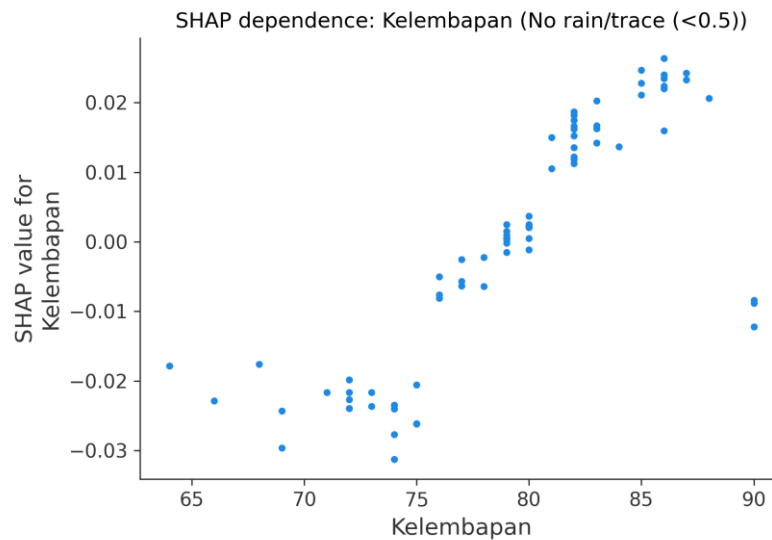


Fig. 7. SHAP dependence plot for the selected Random Forest model.

E. Discussion

The revised evaluation changes the model-selection conclusion. The earlier random stratified split and weighted metrics made the ensemble results appear substantially stronger, but the chronological holdout is closer to the intended forecasting use case. Random Forest provides the highest macro-F1 in this experiment, while XGBoost provides the highest accuracy. Because class-wise performance is uneven and rare-event support is very small, neither result warrants a broad claim that one algorithm is generally superior. The majority baseline's accuracy is lower than the leading models, but all models' ability to detect heavy and very-heavy/extreme rainfall remains inadequate.

The dataset's temporal coverage is a central constraint. A single calendar year cannot represent interannual climate variability, and a 73-example test period yields very wide uncertainty. The training set contains only one very-heavy/extreme target example; the model cannot learn a stable decision boundary for this class. Class weighting may help the optimization objective but cannot create information absent from the observations. Future work should obtain multiple years of data from a documented station or multiple stations, preserve an untouched final year for evaluation, compare operational BMKG forecasts where available, and consider a two-stage task (rain/no-rain followed by intensity conditional on rain) only if supported by enough events. Resampling must be performed within training folds only and should be evaluated against unresampled baselines.

The supplied data file does not document the station identity, coordinates, measurement instrument, download date, or the original missing-value codes. These details are essential for reproducibility and cannot be inferred safely from the numeric series. In addition, the source file's extra undated row and the cleaning history need to be reconciled against the original BMKG download. Consequently, the results are a reproducible analysis of the supplied CSV, not a verified station-level BMKG forecast product.

V. CONCLUSION

This study re-evaluated next-day rainfall-category prediction using a chronological 80:20 split, time-series hyperparameter selection with macro-F1 as the common tuning metric, and explicit majority, persistence, and logistic-regression baselines. Random Forest achieved the highest test macro-F1 (0.327; 95% bootstrap CI 0.202–0.441), whereas XGBoost achieved the highest accuracy (0.466). The prior claim that XGBoost was the best overall model is therefore not supported by the revised evaluation. All models struggled with rare heavy-rainfall categories, and none detected the single very-heavy/extreme test example. The one-year record, small test set, and missing station metadata limit generalization. Additional years, verified station metadata, and comparison with operational forecasts are required before practical use or claims of operational readiness.

ACKNOWLEDGMENT

The authors would like to thank the Meteorology, Climatology, and Geophysical Agency (BMKG) for providing the meteorological data used in this study. The authors also acknowledge Universitas Pembangunan Nasional "Veteran" Jawa Timur for its academic support

during the research process.

REFERENCES

- [1] Badan Meteorologi, Klimatologi, dan Geofisika (BMKG), "Ikhtisar Cuaca Harian," BMKG Data Web. Available: <https://dataweb.bmkg.go.id/cuaca/dokumen/ikhtisar-cuaca-harian.pdf>. Accessed: 27-Oct-2026.
- [2] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [3] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [4] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems* 30, 2017, pp. 3146–3154.
- [5] A. Agung, A. Fauzi, A. Nur Risal, and F. Adiba, "Implementasi teknik data mining terhadap klasifikasi data prediksi curah hujan BMKG di Sulawesi Selatan," *Jurnal Tekno Insentif*, vol. 17, no. 1, pp. 22–23, 2023, doi: 10.36787/jti.v17i1.955.
- [6] M. Putra, M. S. Rosid, and D. Handoko, "High-resolution rainfall estimation using ensemble learning techniques and multisensor data integration," *Sensors*, vol. 24, no. 15, Art. no. 5030, 2024, doi: 10.3390/s24155030.
- [7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems* 30, 2017, pp. 4765–4774.
- [8] scikit-learn developers, "TimeSeriesSplit," scikit-learn documentation. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html. Accessed: 27-Oct-2026.
- [9] N. Wibowo, W. Hardyanto, A. Sepriando, D. Djuniadi, T. Nurmayati, and F. Alam, "Evaluasi kinerja algoritma Random Forest dan KNN untuk prediksi cuaca di Jakarta," *Jurnal Pendidikan Matematika: Judika Education*, 2026, doi: 10.31539/wzj0r898.
- [10] scikit-learn developers, "Model evaluation: balanced accuracy and classification metrics," scikit-learn documentation. Available: https://scikit-learn.org/stable/modules/model_evaluation.html. Accessed: 27-Oct-2026.
- [11] scikit-learn developers, "Matthews correlation coefficient," scikit-learn documentation. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.matthews_corrcoef.html. Accessed: 27-Oct-2026.
- [12] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [13] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY, USA: Springer, 2013.
- [14] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.
- [15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
- [16] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.
- [17] J. Bergmeir, R. J. Hyndman, and B. Koo, "A note on the validity of cross-validation for evaluating autoregressive time series prediction," *Computational Statistics & Data Analysis*, vol. 120, pp. 70–83, 2018, doi: 10.1016/j.csda.2017.11.003.
- [18] A. Roberts et al., "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," *Ecography*, vol. 40, no. 8, pp. 913–929, 2017, doi: 10.1111/ecog.02881.
- [19] G. Gorodkin, "Comparing two K-category assignments by a K-category correlation coefficient," *Computational Biology and Chemistry*, vol. 28, no. 5–6, pp. 367–374, 2004, doi: 10.1016/j.compbiolchem.2004.09.006.
- [20] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.
- [21] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960, doi: 10.1177/001316446002000104.
- [22] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1993.
- [23] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
- [24] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artificial Intelligence*, 1995, pp. 1137–1145.
- [25] A. F. Fahmi, A. Amali, and A. Badruzzaman, "Prediksi curah hujan Jawa Barat menggunakan algoritma machine learning: Analisis komparatif berbasis data BMKG 2024," *Journal of Information System Research*, 2026, doi: 10.47065/Josh.v7i2.9018.