Volume 11 Issue 1 Year 2026 | Page 128-138 ISSN: 2527-9866
Received: 20-12-2025 | Revised: 30-12-2025 | Accepted: 13-01-2026

Ensemble-Based Machine Learning for Improving Local Weather Prediction Accuracy in Batam, Indonesia

Hendi Sama[1], Randy Heskyel Gumolung[2], Yefta Christian[3]
[1,2,3] Batam International University, Batam, Indonesia, 29431
e-mail: hendi@uib.ac.id[1], 2231149.randy@uib.edu[2], yefta@uib.ac.id[3]
*Correspondence: 2231149.randy@uib.edu

Abstract: *Accurate short-term rainfall prediction in tropical microclimates such as Batam remains challenging due to strong local atmospheric variability and the limited generalization capability of single-model classifiers. This study proposes an ensemble-based framework that integrates Naïve Bayes, C4.5, and Random Forest through a majority-voting mechanism for multi-class hourly rainfall prediction. The experiments were conducted using multi-year hourly meteorological data collected for Batam City from an open-source weather archive, covering key atmospheric variables and exhibiting an imbalanced rainfall-class distribution. Model performance was evaluated using ten-fold cross-validation with accuracy, precision, recall, and F1-score metrics. The proposed ensemble achieved an accuracy of 84.74%, consistently outperforming the corresponding base classifiers. The model demonstrated strong predictive capability for dominant rainfall classes (TidakHujan and HujanRingan), while reduced performance was observed for HujanSedang and HujanBerat due to class imbalance, a well-documented challenge in tropical rainfall modeling. Overall, the results indicate that combining probabilistic and tree-based learners yields a more stable and reliable prediction framework for localized tropical weather. This work contributes a practical and reproducible ensemble approach tailored to microclimate conditions, offering a foundation for improved data-driven rainfall forecasting in similar high-variability regions.*

Keywords: *Weather prediction, Naïve Bayes, C4.5, Random Forest, ensemble model.*

1. Introduction

Riau Islands Province in western Indonesia is characterized by a tropical maritime climate with strong ocean–atmosphere interactions that generate highly variable local weather. Its capital, Batam City, is surrounded by the Malacca Strait and exhibits pronounced microclimatic behavior, including rapid changes in rainfall, temperature, and wind that affect maritime transport, aviation, and coastal activities [1], [2], [3]. Operational forecasts often struggle to capture these fine scale variations. BMKG has reported that weather forecasts for Indonesia, particularly short-range numerical weather predictions, exhibit higher relative error compared with other regions due to the complex tropical dynamics of the archipelago, especially in areas with strong ocean–atmosphere interactions such as the Riau Islands, making high-resolution and locally adapted methods desirable [4]. In Batam, this limitation manifests as missed or mislocated rainfall events, which can disrupt daily activities and complicate risk management for local stakeholders. These conditions indicate a need for locally adapted forecasting methods that can better represent Batam’s microclimate.

At the same time, accurate short term weather prediction underpins early warning systems and operational planning in sectors such as transportation, agriculture, disaster preparedness, and coastal management. Recent Indonesian and international studies show that data driven approaches, including statistical post processing and machine learning, can enhance local forecast skill when models are adapted to station specific or region specific conditions [5], [6],

[7]. However, most existing work either focuses on broader regional scales or evaluates a limited set of algorithms, and relatively few studies target categorical rainfall prediction for a single tropical urban island with strong microclimatic effects such as Batam.

Machine learning offers a flexible framework for exploiting large meteorological datasets by learning non linear relationships between predictors and rainfall outcomes. Supervised learning methods are particularly relevant for categorical rainfall prediction, where models are trained on labeled data to distinguish between rain and no rain or among multiple rainfall intensity classes [8]. Within this context, several algorithm families have emerged as competitive baselines. Naïve Bayes provides a probabilistic formulation that can be extended to spatio temporal prediction by expanding feature representations across time, improving its suitability for forecasting tasks [9]. Decision tree based approaches such as C4.5 are valued for their interpretability and ability to handle mixed continuous and categorical inputs, and have shown good performance when combined with preprocessing or clustering techniques in precipitation classification [10]. Random Forest, as an ensemble of decision trees, has demonstrated robust skill in complex, nonlinear weather prediction settings and is frequently used as a benchmark model in rainfall studies [11].

Despite advances in machine learning–based weather prediction, a methodological gap remains for Batam's tropical island microclimate. Most prior studies evaluate Naïve Bayes, decision tree–based models, or Random Forest either individually or within ensembles designed for different climatological contexts, providing limited evidence on their behavior under Batam's conditions, where non-rain events dominate and rainfall classes are highly imbalanced. This gap is critical because moderate and heavy rainfall, though infrequent, have disproportionate impacts on maritime operations, aviation safety, and coastal risk management. Moreover, existing ensemble approaches rarely analyze how severe class imbalance affects the detection of higher rainfall intensities or whether combining models with distinct learning biases can alleviate this issue.

To address this gap, this study evaluates a majority-voting ensemble that integrates Naïve Bayes, C4.5, and Random Forest, selected for their complementary characteristics. Naïve Bayes provides probabilistic generalization on noisy meteorological data, C4.5 offers interpretable rule-based partitioning, and Random Forest captures nonlinear atmospheric interactions. By assessing this combination on Batam's historical weather data, the study quantifies its performance under strong class imbalance and clarifies its strengths and limitations across rainfall intensity categories. The contribution lies in providing empirical evidence and a practical, reproducible ensemble framework tailored to tropical island microclimates, with relevance for other coastal and island regions facing similar forecasting challenges.

2. Literature Review

Recent studies have highlighted the challenges of weather forecasting in Indonesia's tropical maritime climate, where strong ocean–atmosphere interactions produce highly variable local weather patterns. In Batam, persistent high temperatures, elevated humidity, and frequent rainfall create complex microclimatic conditions that often disrupt transportation, fisheries, and tourism activities [3]. Prior research has shown that ensemble learning can improve rainfall prediction under such conditions. For example, [12] demonstrated that combining multiple classifiers, including Naïve Bayes, Decision Tree, Support Vector Machine, and Random Forest, using majority voting significantly outperformed individual models. These findings suggest that ensemble approaches are particularly effective for capturing nonlinear rainfall behavior in tropical regions. Similarly, [13] reported that conventional weather forecasting methods often suffer from inefficiency and limited accuracy, particularly in data-scarce regions. By integrating multiple machine learning models, including Gradient Boosting Decision Tree, Random Forest, Naïve Bayes Bernoulli, and KNN, the study achieved high prediction accuracy with reduced computational cost, underscoring the practicality of ensemble approaches. In a different application domain, [14] proposed a hybrid model combining Random Forest and Naïve Bayes for student performance prediction. Although not meteorological, their ensemble design illustrates how probabilistic reasoning can be integrated with tree-based models, offering methodological insights relevant to weather forecasting tasks [15] highlights the application of machine learning techniques, particularly the Random Forest model, for weather prediction, addressing challenges posed by complex numerical climate models in tropical systems. Focusing on Tamil Nadu, India, it predicts Global Solar Radiation (GSR) and wind speed using data from IMD, Pune. The Random Forest model outperforms statistical regression and SVM models, achieving a minimum MSE of 0.750 and an R^2 score of 0.97. The findings demonstrate the Random Forest model's superior accuracy, reducing reliance on expensive measuring instruments for acquiring solar radiation and wind speed data [16] applies Random Forest models to probabilistic severe weather forecasting using long-term reforecast data from the Global Ensemble Forecast System. The study shows that Random Forest can capture meaningful statistical relationships and provide skillful forecasts in the medium range, particularly at days four and five. However, performance degradation for extreme events at longer lead times is also reported, indicating that while ensemble-based machine learning models improve forecast skill, their effectiveness remains constrained by data characteristics and event rarity. This finding aligns with challenges observed in tropical rainfall prediction, where extreme events are inherently difficult to detect [17] shows that rainfall classification plays a critical role in sectors such as aviation and agriculture, especially in data-scarce regions. Using a Naïve Bayes classifier with meteorological parameters such as humidity and precipitation, the study achieved high accuracy, demonstrating the effectiveness of probabilistic classifiers for rainfall categorization tasks[18] combines Naïve Bayes and C4.5 with ensemble techniques on ten years of BMKG weather data. Although ensemble variants slightly improved accuracy compared to individual models, the overall gains remained limited, highlighting the need for more robust ensemble configurations to achieve substantial performance improvements.

3. Methods

This study applies a systematic and replicable methodological framework designed to evaluate and enhance weather prediction performance in Batam. The approach consists of several sequential stages: data collection, data preprocessing, model development, ensemble construction, and performance evaluation. Each stage is structured to ensure transparency in the experimental process and to support reproducibility. A conceptual diagram is used to illustrate the end-to-end workflow, representing how raw meteorological data are transformed into predictive outputs through supervised machine learning techniques.

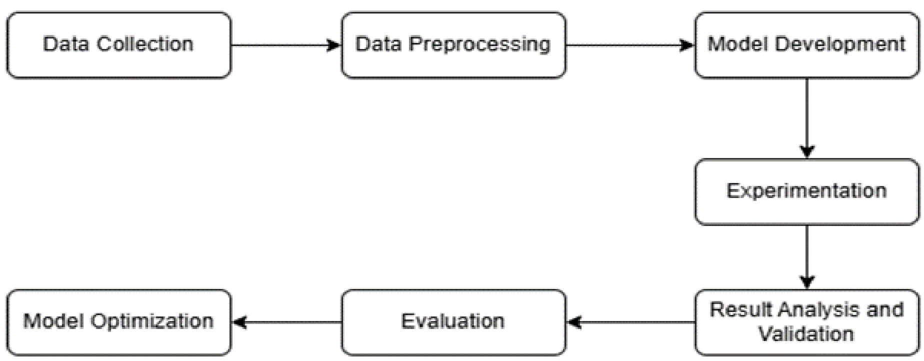

Figure 1. Research method

A. Data Collection

Meteorological data were collected from the Open-Meteo Historical Weather API for Batam City using fixed geographic coordinates (latitude 0.9, longitude 104.0), representing the central urban–coastal area of Batam. To comprehensively capture local atmospheric conditions, three complementary API requests were performed covering the period from January 2015 to January 2025.

The first request retrieved hourly core atmospheric variables, including near-surface air temperature, relative humidity, dew point, surface pressure, wind speed, and total cloud cover. These variables form the primary predictors commonly associated with short-term rainfall processes in tropical environments. The second request focused on precipitation-related indicators, such as low, mid, and high cloud cover layers, hourly precipitation amount, and vapour pressure deficit, which directly influence rainfall occurrence and intensity. The third request provided daily extreme and radiative variables, including maximum and minimum temperature, apparent temperature, shortwave radiation sum, and wind gusts, supplying broader thermal and energy-balance context relevant to tropical weather variability.

All datasets were automatically downloaded through reproducible API calls, standardized to the same time zone (UTC+7), and subsequently temporally aligned and merged into a unified meteorological table. This integrated dataset served as the input for the subsequent preprocessing, feature engineering, and modeling stages.

	temperature_2m_c	relative_humidity_2m	dew_point_2m_c	surface_pressure_hpa	wind_speed_10m_kmh	cloud_cover	soil_moisture_0_7cm_m3m3	soil_temperature_0_7cm_c	soil_temperature_7_28cm_c	soil_temperature_28_100cm_c	soil_temperature_100_2
0:00	25.3	86	22.8	1011.5	24.6	100	325	24.9	26.0	27.1	28.3
0:00	25.5	86	23.0	1012.3	24.9	100	327	25.0	25.9	27.1	28.3
0:00	25.4	86	22.9	1012.8	24.4	100	335	25.2	25.9	27.1	28.3
0:00	25.5	86	23.1	1013.1	23.4	100	0.34	25.5	25.9	27.1	28.3
0:00	25.8	84	22.8	1012.8	23.5	100	337	25.6	25.9	27.1	28.3
0:00	26.0	83	22.8	1012.1	24.8	100	332	25.7	25.9	27.1	28.3
0:00	26.0	82	22.7	1011.2	24.9	100	331	25.9	25.9	27.1	28.3
0:00	25.6	84	22.8	1010.3	23.6	100	352	26.0	26.0	27.1	28.3
0:00	25.6	85	22.9	1009.7	22.3	100	351	25.9	26.0	27.1	28.3
0:00	25.6	86	23.0	1009.8	25.3	100	343	25.9	26.0	27.0	28.3
0:00	25.6	85	22.8	1010.3	24.4	100	339	25.7	26.0	27.0	28.3
0:00	25.8	83	22.6	1010.6	24.7	100	334	25.3	26.0	27.0	28.3
0:00	26.0	82	22.7	1011.2	25.4	99	328	25.1	25.9	27.0	28.3
0:00	26.0	82	22.6	1011.9	26.8	100	324	25.0	25.9	27.0	28.3
0:00	26.0	81	22.5	1012.7	28.1	100	322	24.9	25.8	27.0	28.2
0:00	25.9	84	22.9	1013.1	27.8	100	319	24.9	25.8	27.0	28.2
0:00	25.8	84	22.9	1012.8	26.3	100	316	24.9	25.7	27.0	28.2
0:00	25.7	84	22.8	1012.6	26.3	97	313	24.9	25.7	27.0	28.2
0:00	25.6	84	22.8	1011.6	27.4	95	311	24.8	25.7	27.0	28.2
0:00	25.4	86	22.8	1011.1	28.9	99	309	24.7	25.6	27.0	28.2
0:00	25.3	86	22.7	1010.6	29.3	99	307	24.6	25.6	27.0	28.2

Figure 2. Dataset

B. Data Preprocessing

The merged dataset initially consisted of hourly and daily observations spanning ten years. After temporal alignment, cleaning, and feature engineering, the final dataset comprised 87,668 hourly records with complete predictor variables. The target rainfall variable exhibited a highly imbalanced distribution, with non-rain observations dominating the dataset, reflecting the natural rainfall characteristics of Batam's tropical climate.

1. Data Cleaning and Time Alignment

All timestamps were parsed into a uniform datetime format and aligned at the hourly resolution, as Open-Meteo data occasionally contains duplicate or irregular time entries. Duplicate timestamps were removed, and non-numeric fields were coerced into numeric types where appropriate. Rows with invalid timestamps were discarded. To ensure physical validity of meteorological measurements, range bounding was applied. Variables such as humidity (0–100%), cloud cover (0–100%), temperature (–20 to 60°C), soil moisture (0–0.6 m^3/m^3), and surface pressure (870–1085 hPa) were clipped to scientifically plausible ranges. Values falling outside these ranges were treated as missing.

2. Handling Missing Values and Outliers

Missing values were handled using a strategy tailored to the temporal nature of the dataset. Continuous meteorological variables were interpolated based on time to preserve hourly continuity, while variables less suitable for interpolation, such as cloud cover layers, were completed using forward and backward filling. Observations with excessive missing information were excluded to maintain data reliability. To reduce the influence of extreme values, numerical features were clipped using percentile-based thresholds, ensuring that outliers did not dominate model learning while retaining natural atmospheric variability.

3. Feature Engineering

Feature engineering was performed to capture temporal dependencies and periodic atmospheric patterns that influence rainfall formation. Time-based features such as hour, day of week, and month were derived from the timestamp, while cyclic transformations using sine and cosine functions were applied to model diurnal, weekly, and seasonal periodicity. To incorporate short-term temporal memory, lagged values at 1, 3, and 6 hours were generated for key meteorological variables, complemented by 3-hour rolling averages to smooth short-term fluctuations. These transformations enrich the temporal representation of the data, enabling the models to learn both instantaneous conditions and recent atmospheric trends.

4. Target Construction and Final Dataset Preparation

The target variable, next-hour rainfall, was derived by shifting the precipitation column one step forward to create `target_precip_next1h`. Rows containing undefined target values were removed. Finally, the dataset was exported into a clean analytical format and served as the input for model development. All preprocessing steps were conducted automatically using a reproducible Python pipeline, facilitating transparency and replicability.

C. Model Development

This study implemented three supervised machine learning algorithms: Naïve Bayes, C4.5, and Random Forest, to construct rainfall classification models. Each algorithm was selected based on its theoretical suitability for meteorological data and its complementary role within an ensemble framework. The Naïve Bayes classifier was implemented using the Gaussian Naïve Bayes formulation, which assumes continuous meteorological predictors follow a normal distribution within each rainfall class. This variant is appropriate for atmospheric variables such as temperature, humidity, pressure, and wind speed, which exhibit approximately continuous behavior. Gaussian Naïve Bayes estimates class posterior probabilities using Bayes' theorem under the conditional independence assumption, enabling efficient learning and robust performance on high-dimensional climate data with limited computational cost [19].

$$P(C\,|X) \propto P(C)\prod_{i=1}^{n} P(\,x_i|\,C)$$

The C4.5 algorithm was implemented using the J48 decision tree formulation, which constructs classification trees based on information gain ratio. This approach enables interpretable rule-based partitions of meteorological variables, allowing explicit analysis of how attributes such as humidity, cloud cover, and pressure contribute to rainfall classification. C4.5 is particularly suitable for mixed-type features and nonlinear decision boundaries commonly found in tropical weather data [18].

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInfo(S,A)}$$

Random Forest was implemented as an ensemble of decision trees trained using bootstrap sampling and random feature selection. This method improves generalization by reducing variance and capturing nonlinear interactions among atmospheric variables. In this study, the Random Forest model used 100 trees (n_estimators = 100), with maximum tree depth unconstrained, and class weighting enabled to partially address rainfall class imbalance. Random Forest has been shown to provide stable probabilistic forecasts in complex meteorological systems [20].

After training the individual classifiers, their predictions were combined using a majority-voting ensemble strategy. For each input instance, Naïve Bayes, C4.5, and Random Forest independently produced a rainfall class label, and the final ensemble output was determined by the class receiving the highest number of votes. This integration exploits complementary modeling characteristics: probabilistic inference from Naïve Bayes, rule-based decision boundaries from C4.5, and nonlinear pattern learning from Random Forest. Majority voting was chosen due to its simplicity, interpretability, and robustness under heterogeneous classifier behavior. All models were trained using the preprocessed dataset with stratified cross-validation to ensure fair evaluation under class imbalance conditions. Hyperparameters such as tree depth (C4.5), number of trees (Random Forest), and smoothing parameters (Naïve Bayes) were tuned empirically to optimize performance.

D. Experimentation

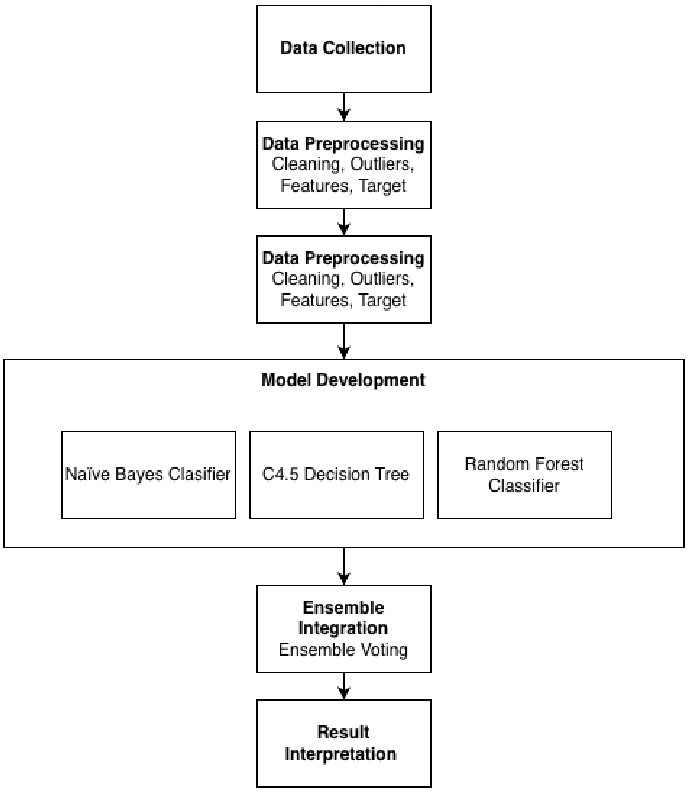

Figure 3. Ensemble voting

After the models were developed, an experimentation phase was carried out to evaluate the performance of each individual algorithm (Naïve Bayes, C4.5, and Random Forest) and to assess the effectiveness of the ensemble approach. This stage served as the empirical foundation of the study, since every model was tested under identical conditions to produce comparable and replicable results. The experiments were conducted using the fully preprocessed dataset, and all models were evaluated using the same validation procedure to ensure fairness and reduce variance.

During this phase, each classifier was executed on the dataset to generate predictions for rainfall categories. The predictive performance of every model was measured using standard classification metrics, which included accuracy, precision, recall, and F1 score. These metrics were selected because they provide a balanced evaluation of both overall correctness and class specific behavior, particularly in datasets affected by imbalance. Through this process, strengths and weaknesses of each algorithm became evident, such as the tendency of certain models to overpredict majority classes or to fail in identifying minority rainfall categories. The outcomes of the experimentation phase offered essential insight into how each model performed when confronted with Batam's highly variable meteorological patterns. They also provided evidence to justify the construction of the ensemble model and informed the subsequent optimization process.

3. Results and Discussion

Although this study is formulated as a multi-class classification task, additional error-based metrics are reported to provide complementary insight into model behavior. Classification metrics (accuracy, precision, recall, and F1-score) are computed on discretized rainfall classes (TidakHujan, HujanRingan, HujanSedang, and HujanBerat). Meanwhile, absolute error, RMSE, correlation, and squared correlation are calculated using an ordinal encoding of rainfall intensity (0–3), which preserves class ordering. These metrics do not aim to estimate exact rainfall amounts, but to assess the severity of misclassification across rainfall intensities. As such, error-based metrics function as auxiliary indicators that complement classification results, enabling a more nuanced interpretation of model performance under highly imbalanced rainfall conditions.

A. Model accuracy

accuracy: 84.74% +/- 0.31% (micro average: 84.74%)

	true TidakHujan	true HujanRingan	true HujanSedang	true HujanLebat	class precision
pred. TidakHujan	67795	7640	565	67	89.13%
pred. HujanRingan	3622	6507	1314	168	56.04%
pred. HujanSedang	1	2	2	2	28.57%
pred. HujanLebat	0	2	2	0	0.00%
class recall	94.93%	45.98%	0.11%	0.00%	

Figure 4. Confusion matrix

The ensemble model achieved an average accuracy of 84.74% ± 0.31%, demonstrating a relatively high predictive performance, especially in distinguishing between TidakHujan and Hujan conditions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

However, the confusion matrix reveals an uneven class performance. The TidakHujan class obtained the highest recall at 94.93% and precision at 89.13%, indicating that the model performed very well in recognizing dry conditions. In contrast, the recall for HujanRingan, HujanSedang, and HujanLebat classes were considerably lower at 45.98%, 0.11%, and 0.00%, respectively. This result suggests that the dataset is imbalanced, leading the model to favor the dominant class (TidakHujan) while underperforming on minority rainfall classes.

B. Absolute error and RMSE

The obtained absolute error was 0.256 ± 0.028, and the RMSE was 0.390 ± 0.008. These relatively low values indicate that the average deviation between predicted and actual outcomes is modest. Nevertheless, the remaining errors primarily arise from misclassification in less frequent rainfall categories, especially moderate and heavy rain events.

absolute_error

absolute_error: 0.256 +/- 0.028 (micro average: 0.256 +/- 0.294)

Figure 5. Absolute error

root_mean_squared_error

root_mean_squared_error: 0.390 +/- 0.008 (micro average: 0.390 +/- 0.000)

Figure 6. RMSE

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

C. Correlation

The correlation between predicted and actual values was 0.509 ± 0.014, while the squared correlation was 0.259 ± 0.015. These results suggest a moderate linear relationship between model predictions and observed data. Although the ensemble captures general weather patterns, its ability to model more complex rainfall behaviors remains limited.

correlation

correlation: 0.509 +/- 0.014 (micro average: 0.508)

Figure 7. Correlation

$$r = \frac{\Sigma(x_i - \underline{x})(y_i - \underline{y})}{\sqrt{\Sigma(x_i - \underline{x})^2 \, \Sigma(y_i - \underline{y})^2}}$$

Table 1. Performance trends of individual and ensemble classifiers in rainfall prediction studies

	Study	Models Compared	Key Findings
1	A review on rainfall forecasting using ensemble learning techniques [7]	Individual ML models vs ensemble methods	Ensemble learning consistently outperformed single classifiers in rainfall prediction accuracy across multiple studies.
2	Weather Forecasting Using Machine Learning Techniques: Rainfall and Temperature Analysis [21]	SVC, LR, RF vs voting ensemble	Ensemble model achieved higher accuracy and recall than individual classifiers, particularly for rainfall occurrence detection.
3	Perbandingan Model Machine Learning pada Klasifikasi Curah Hujan di Bogor [22]	Naïve Bayes, Decision Tree, Random Forest	Random Forest outperformed Naïve Bayes and Decision Tree, indicating complementary strengths among base learners.
4	Comparison Analysis of Random Forest and Naïve Bayes Algorithms for Rainfall Classification Based on Climate in Indonesia [23]	Naïve Bayes vs Random Forest	Random Forest showed significantly better precision and recall than Naïve Bayes on climatic datasets.
5	Rainfall Prediction based on Historical Weather Data using Naive Bayes Classification Model in Southeast Sulawesi [24]	Naïve Bayes (single model)	Naïve Bayes achieved moderate accuracy (~74%), serving as a baseline classifier for rainfall prediction.

Table 1 summarizes performance trends reported in recent rainfall prediction studies comparing individual classifiers and ensemble-based methods. Prior research consistently indicates that ensemble models generally achieve higher overall accuracy and more stable performance than single classifiers, particularly when probabilistic and tree-based learners are combined. This pattern is consistent with the results of this study, where the Naïve Bayes–C4.5–Random Forest ensemble outperformed its base models in terms of overall accuracy. At the same time, several studies report that ensemble performance degrades under severe class imbalance, especially for moderate and heavy rainfall categories. This limitation aligns with the behavior observed in the Batam dataset, indicating that the findings of this study reflect broader trends in rainfall prediction literature rather than isolated results.

The experimental results further show that the accuracy improvement achieved by the proposed ensemble is primarily driven by its strong ability to identify TidakHujan conditions, which dominate the data distribution. However, the pronounced imbalance substantially reduces sensitivity to rarer rainfall events, with recall values dropping to 0.11% for HujanSedang and 0.00% for HujanBerat. These results indicate that, while effective for dominant-condition classification, the ensemble is not reliable for detecting moderate and heavy rainfall events without additional strategies to address class imbalance. This limitation has important implications for model applicability. While the ensemble achieves high overall accuracy, its effectiveness is restricted when forecasting rare but high impact rainfall events. As a result, the proposed model should not be interpreted as a reliable solution for disaster early warning systems that depend on accurate detection of extreme rainfall. Instead, its reliability is confined to operational scenarios focused on dominant weather conditions, such as distinguishing dry periods from light rainfall, which remain relevant for maritime operations, aviation scheduling, and routine outdoor planning in Batam.

Despite these constraints, the ensemble voting strategy still demonstrates methodological value. The integration of probabilistic reasoning from Naïve Bayes, rule-based partitioning from C4.5, and nonlinear modeling from Random Forest contributes to improved stability and generalization relative to single classifier approaches. This confirms that ensemble learning remains a viable framework for short term weather classification in tropical microclimates, provided its limitations are clearly acknowledged. The primary source of performance degradation lies in the severe imbalance of rainfall classes, where moderate and heavy rainfall events are sparsely represented. Addressing this issue is critical for future research. Potential improvements include the application of resampling techniques such as SMOTE, cost sensitive learning strategies, or the incorporation of additional historical rainfall records from BMKG. These enhancements are necessary to improve fairness across classes and to extend the applicability of ensemble-based models to extreme rainfall prediction.

4. Conclusions

This study proposed and evaluated an ensemble-based approach for short-term rainfall prediction in Batam by integrating Naïve Bayes, C4.5, and Random Forest algorithms. The ensemble was designed to address the high variability and complex microclimatic characteristics of Batam's tropical environment. Experimental results showed that the ensemble achieved an average accuracy of 84.74%, outperforming the corresponding individual classifiers. The model demonstrated strong predictive performance for dominant rainfall categories, particularly TidakHujan and HujanRingan, supported by relatively high precision and recall values. These findings are consistent with previous studies reporting that ensemble machine learning methods generally achieve higher predictive accuracy for rainfall forecasting than single-model approaches [25].

However, the results also reveal a clear limitation of the proposed model. Performance degraded substantially for the HujanSedang and HujanBerat categories, with recall values of only 0.11% and 0.00%, respectively. This indicates that the current ensemble configuration is not suitable for reliable detection of rare but high-impact rainfall events. Such degradation is primarily caused by severe class imbalance, where non-rain and light-rain observations dominate the dataset and bias the learning process toward majority classes. Similar behavior has been reported in prior rainfall prediction studies, which show that imbalanced datasets can reduce the effectiveness of conventional ensemble models and that alternative approaches may outperform ensembles when minority rainfall classes are severely underrepresented [26].

Therefore, the applicability of the proposed ensemble must be interpreted within a clearly defined scope. In its current form, the model is most suitable for forecasting dominant weather conditions, particularly distinguishing dry periods from light rainfall, which is relevant for maritime operations, aviation scheduling, and routine activity planning in Batam. However, due to very low recall for heavy rainfall events, the model is not yet appropriate for early warning applications that require reliable detection of extreme precipitation. Future research should focus on improving minority-class detection through targeted strategies, including resampling methods such as SMOTE or ADASYN, cost-sensitive learning, and the incorporation of additional large-scale climate features such as sea surface temperature and ENSO indices. Evaluation should prioritize class-aware metrics, including macro-F1, class-specific recall, and precision–recall AUC, to ensure measurable improvement in heavy rainfall detection.

References

[1] D. C. Siregar, V. P. Ardah, and A. M. Navitri, “Analisis Kondisi Atmosfer Terkait Siklon Tropis Pabuk Serta Pengaruhnya Terhadap Tinggi Gelombang di Perairan Kepulauan Riau,” *Tunas Geografi*, vol. 8, no. 2, p. 111, Mar. 2020, doi: 10.24114/tgeo.v8i2.17049.

[2] M. T. Anwar, W. Hadikurniawati, E. Winarno, and W. Widiyatmoko, “Performance Comparison of Data Mining Techniques for Rain Prediction Models in Indonesia,” in *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, Dec. 2020, pp. 83–88. doi: 10.1109/ISRITI51436.2020.9315460.

[3] D. R. Pratama, I. Jaya, and M. Iqbal, “Algorithm design for wind prediction in Berakit Bay, Bintan Island using Long Short-Term Memory (LSTM) method,” *IOP Conf Ser Earth Environ Sci*, vol. 944, no. 1, p. 012006, Dec. 2021, doi: 10.1088/1755-1315/944/1/012006.

[4] Kompas, “Tantangan Akurasi Prediksi Cuaca Indonesia,” https://www.kompas.com/sains/read/2023/01/02/104238523/tantangan-akurasi-prediksi-cuaca-indonesia.

[5] X. Ren *et al.*, “Deep Learning-Based Weather Prediction: A Survey,” *Big Data Research*, vol. 23, p. 100178, Feb. 2021, doi: 10.1016/j.bdr.2020.100178.

[6] A. V. Wulandari, N. J. Trilaksono, and M. Ryan, “Improving Short-Term Weather Forecasting using Support Vector Machine Method in North Barito,” *Jurnal Meteorologi dan Geofisika*, vol. 25, no. 2, pp. 113–121, Jun. 2025, doi: 10.31172/jmg.v25i2.1096.

[7] Tedyyana, A., Ahmad, A. A., Idrus, M. R., Mohd Shabli, A. H., Abu Seman, M. A., Ghazali, O., & Abd Razak, A. H. (2024). Enhance Telecommunication Security Through the Integration of Support Vector Machines. *International Journal of Advanced Computer Science & Applications*, *15*(3).

[8] T. Tan, H. Sama, G. Wijaya, and O. E. Aboagye, “Studi Perbandingan Deteksi Intrusi Jaringan Menggunakan Machine Learning: (Metode SVM dan ANN),” *Jurnal Teknologi dan Informasi*, vol. 13, no. 2, pp. 152–164, Aug. 2023, doi: 10.34010/jati.v13i2.10484.

[9] S. S. Prasetiyowati and Y. Sibaroni, “Unlocking the potential of Naive Bayes for spatio temporal classification: a novel approach to feature expansion,” *J Big Data*, vol. 11, no. 1, p. 106, Aug. 2024, doi: 10.1186/s40537-024-00958-x.

[10] D. Lou *et al.*, “K-Means and C4.5 Decision Tree Based Prediction of Long-Term Precipitation Variability in the Poyang Lake Basin, China,” *Atmosphere (Basel)*, vol. 12, no. 7, p. 834, Jun. 2021, doi: 10.3390/atmos12070834.

[11] S. S. Sofi and I. Oseledets, “A case study of spatiotemporal forecasting techniques for weather forecasting,” *Geoinformatica*, Sep. 2024, doi: 10.1007/s10707-024-00530-y.
[12] N. S. Sani, A. Hadi, A. Adam, I. Shlash, and M. Aliff, “Ensemble Learning for Rainfall Prediction,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, 2020, doi: 10.14569/IJACSA.2020.0111120.
[13] H. Shaiba *et al.*, “Weather Forecasting Prediction Using Ensemble Machine Learning for Big Data Applications,” *Computers, Materials & Continua*, vol. 73, no. 2, pp. 3367–3382, 2022, doi: 10.32604/cmc.2022.030067.
[14] Y. Manzali, Y. Akhiat, K. Abdoulaye Barry, E. Akachar, and M. El Far, “Prediction of Student Performance Using Random Forest Combined With Naïve Bayes,” *Comput J*, vol. 67, no. 8, pp. 2677–2689, Aug. 2024, doi: 10.1093/comjnl/bxae036.
[15] R. Meenal, P. A. Michael, D. Pamela, and E. Rajasekaran, “Weather prediction using random forest machine learning model,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 2, p. 1208, May 2021, doi: 10.11591/ijeecs.v22.i2.pp1208-1215.
[16] A. J. Hill, R. S. Schumacher, and I. L. Jirak, “A New Paradigm for Medium-Range Severe Weather Forecasts: Probabilistic Random Forest–Based Predictions,” *Weather Forecast*, vol. 38, no. 2, pp. 251–272, Feb. 2023, doi: 10.1175/WAF-D-22-0143.1.
[17] A. U. Azmi, A. F. Hadi, D. Anggraeni, and A. Riski, “Naive bayes methods for rainfall prediction classification in Banyuwangi,” *J Phys Conf Ser*, vol. 1872, no. 1, p. 012028, May 2021, doi: 10.1088/1742-6596/1872/1/012028.
[18] V. I. Yani, A. Aradea, and H. Mubarok, “Optimasi Prakiraan Cuaca Menggunakan Metode Ensemble pada Naïve Bayes dan C4.5,” *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, no. 3, Dec. 2022, doi: 10.28932/jutisi.v8i3.5455.
[19] C. Li, L. Chen, S. Wu, Y. Mo, and L. Chen, “Application Research of Naive Bayes Algorithm Based on DIKW in Weather Website,” in *2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, IEEE, Dec. 2021, pp. 2253–2257. doi: 10.1109/HPCC-DSS-SmartCity-DependSys53884.2021.00337.
[20] A. J. Hill, G. R. Herman, and R. S. Schumacher, “Forecasting Severe Weather with Random Forests,” *Mon Weather Rev*, vol. 148, no. 5, pp. 2135–2161, May 2020, doi: 10.1175/MWR-D-19-0344.1.
[21] A. Hussain, A. Aslam, S. Tripura, Vi. Dhanawat, and V. Shinde, “Weather Forecasting Using Machine Learning Techniques: Rainfall and Temperature Analysis,” Sep. 13, 2024. doi: 10.20944/preprints202402.1566.v2.
[22] I Dewa Gede Loka Maheswara and Ahmad Hanif Al’aziz, “PERBANDINGAN MODEL MACHINE LEARNING PADA KLASIFIKASI CURAH HUJAN DI BOGOR,” *INTI Nusa Mandiri*, vol. 19, no. 2, pp. 202–210, Feb. 2025, doi: 10.33480/inti.v19i2.6296.
[23] N. Indaryono and R. Saedudin, “COMPARISON ANALYSIS OF RANDOM FOREST AND NAÏVE BAYES ALGORITHMS FORRAINFALL CLASSIFICATION BASED ON CLIMATE IN INDONESIA,” *JOURNAL OF SITEKNIK: Sistem Informasi, Teknik dan Teknologi Terapan*, vol. 1, no. 2, Jul. 2025.
[24] A. Samudin, R. A. Saputra, F. Agsaria, N. H. Judanto, and A. S. Badarudin, “Rainfall Prediction based on Historical Weather Data using Naive Bayes Classification Model in Southeast Sulawesi,” *SISTEMASI*, vol. 14, no. 5, p. 2280, Sep. 2025, doi: 10.32520/stmsi.v14i5.3882.
[25] G. Tuysuzoglu, K. U. Birant, and D. Birant, “Rainfall Prediction Using an Ensemble Machine Learning Model Based on K-Stars,” *Sustainability*, vol. 15, no. 7, p. 5889, Mar. 2023, doi: 10.3390/su15075889.
[26] M. Usman Saeed Khan, K. Mohammad Saifullah, A. Hussain, and H. Mohammad Azamathulla, “Comparative analysis of different rainfall prediction models: A case study of Aligarh City, India,” *Results in Engineering*, vol. 22, p. 102093, Jun. 2024, doi: 10.1016/j.rineng.2024.102093.