Global NEST Journal, Vol 28, No 1, 07707

Forecasting machine learning decision tree, random forest, and Naïve Bayes in predicting hydrometeorological disasters in South Sumatra, Indonesia

Melly Ariska[1]*, Iin Seprina[2], Sardianto Markos Siahaan[3], Yenny Anwar[4] and Suhadi[5]

[1,3]Department of Phsycis Education, Faculty of Teaching and Education, Universitas Sriwijaya, Palembang, South Sumatera, Indonesia

[2] Information Systems Study Program, Faculty of Computer Science, Universitas Sriwijaya, Palembang, South Sumatera, Indonesia

[4]Department of Biology Education, Faculty of Teaching and Education, Universitas Sriwijaya, Palembang, South Sumatera, Indonesia

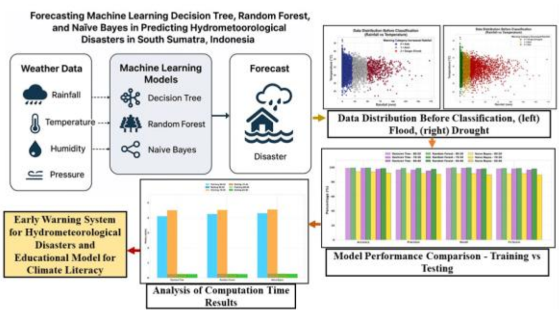
[5]Department of Phsycis Education, Universitas Islam Negeri Raden Fatah Palembang, South Sumatra, Indonesia

Received: 30/05/2025, Accepted: 04/11/2025, Available online: 11/12/2025

*to whom all correspondence should be addressed: e-mail: mellyariska@fkip.unsri.ac.id

https://doi.org/10.30955/gnj.07707

Graphical abstract



Abstract

Hydrometeorological disasters that still occur in cities or areas in South Sumatra, especially along the banks of the Musi River, are floods and peatland fires that trigger haze to cover all areas of South Sumatra, especially the capital city of Palembang. The cause of flooding is generally due to the increasing volume of water in the Musi River and high rainfall intensity, while peatland fires trigger prolonged thick haze disasters. Prevention of hydrometeorological disasters is difficult to do because of the inaccuracy of data in flood and land fire predictions provided by the local government to the community. Therefore, this study was conducted as a more accurate anticipation with better performance and accuracy. This study uses a dataset obtained from the South Sumatra Climatology Station and its surroundings with parameters of river water level and rainfall intensity from 1981 to 2024. The method used to detect the occurrence of hydrometeorological disasters, especially floods and droughts, is the decision tree, random forest, and Naïve Bayes machine learning algorithms. Model performance was assessed using stratified 10-fold cross-validation; Random Forest achieved the best average performance across folds. The experimental results show that the method with the best performance is Random Forest compared to other methods, with an average value of accuracy, precision, recall, and F1-score of 99.05%, 97.91%, 99.18%, and 98%, respectively, and an average computation time of 0.2561 seconds from 3 tests conducted based on different data sharing ratios. The results of this study provide a significant contribution to the use of machine learning methods for more accurate prediction of hydrometeorological disasters in the South Sumatra region. These findings are expected to support disaster risk mitigation efforts through a more effective early warning system, as well as being a strategic reference for policymakers and related parties in data-based disaster management planning.

Keywords: Decision Tree Algorithm; drought; Flood; Haze pollution; Machine Learning Model; Random forest.

1. Introduction

Hydrometeorological disasters in South Sumatra, Indonesia, include various events that are closely related to weather and climate dynamics, such as floods, strong winds, droughts, and haze (Lee 2015). Floods are the most frequent disasters, especially during the rainy season, caused by high rainfall, inadequate drainage systems, and reduced river capacity due to sedimentation and garbage accumulation. (Irfan *et al.* 2022). On the other hand, during the dry season, Palembang City often experiences prolonged droughts. This drought not only causes a decrease in the availability of clean water and disrupts agricultural activities but also triggers forest and land fires in the surrounding areas (Ariska *et al.* 2023). As a result, a haze disaster appears that covers the city and has a serious impact on air quality and public health. This haze usually occurs due to the burning of peatlands that dry out during the dry season, exacerbated by environmentally unfriendly land-clearing practices. In addition, global climate change and human activities such

Melly Ariska, Iin Seprina, Sardianto Markos Siahaan, Yenny Anwar and Suhadi (2025), Forecasting machine learning decision tree, random forest, and Naïve Bayes in predicting hydrometeorological disasters in South Sumatra, Indonesia, *Global NEST Journal*, **28**(1), 07707.

as land conversion, rapid urbanization, and minimal green open spaces also exacerbate the risks and impacts of these hydrometeorological disasters. (Byaruhanga *et al.* 2024).

Hydrometeorological disasters that are prone to occur are a strategic issue for the local government. This disaster does not only occur due to increased rainfall, but also extreme decreases in rainfall trigger drought and peatland fires (Field *et al.* 2016; Iskandar *et al.* 2022). Long dry seasons are the main basis for this drought problem. As happened in 2015 and 2019, rain fires created thick smoke all day long (Ariska *et al.* 2022). In fact, this disaster has become a national disaster globally because its impacts cover several vital aspects of the community environment (Putra *et al.* 2019). The total geographical area of South Sumatra is 91,592.43 km², with an average slope morphology of 0-8%, 8-15%, and above 45% (Ghiffari *et al.* 2023). The average rainfall in South Sumatra is between 2000 millimeters and 3000 millimeters per year. Considering the fairly strong current of the Musi River and its tributaries that carry garbage and mud as the cause of shallowing, flooding in the river flow that passes through the city of Palembang has great potential to cause losses, both material and fatalities (Koplitz *et al.* 2016). Several flood incidents that occurred in Palembang show that this area is still vulnerable to the disaster, especially during the rainy season with high intensity (Ariska *et al.* 2024). Based on the results of an interview with Mr. DM, Head of BNPB Palembang City, flood prevention efforts have been carried out by disseminating information to residents via WhatsApp messages and using sirens as an early warning before flooding occurs (Haylock & McBride 2001). However, there are still obstacles in the form of inaccurate or untimely information, so that people often do not have enough time to take anticipatory steps. In addition to flooding, Palembang City also often faces hydrometeorological disasters in the form of thick smoke from forest and land fires, especially during the dry season. This smoke not only disrupts community activities but also has a serious impact on health, especially respiratory problems in children and the elderly (Field *et al.* 2016; Koplitz *et al.* 2016).

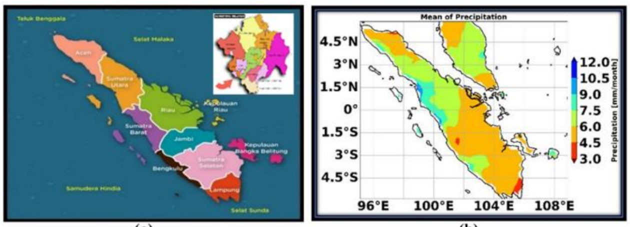
Based on data from the National Board for Disaster Management of South Sumatra, Indonesia, the level of haze occurrences increases significantly in the period from July to October each year, along with decreasing rainfall (Ariska *et al.* 2024a; Putra *et al.* 2019). Although various mitigation efforts have been carried out, such as community outreach, routine patrols in areas prone to forest and land fires, and the use of weather modification technology, the challenges are still great, especially in terms of law enforcement against illegal land burning (Ward *et al.* 2021). Coordination between related agencies and public awareness are key factors in efforts to overcome haze in this area. Based on field conditions related to flood warnings and the potential for major flood disasters in the Palembang City area, the presence of an information system that can predict flood disasters based on rainfall is very important to realize (Gordon *et al.* 2000; Katsumata *et al.* 2018). Moreover, this is very possible to realize considering that the dataset related to rainfall in Palembang City can be easily accessed by the agency in charge of river and rainfall monitoring. Due to the fairly dynamic natural conditions, a machine learning model is needed to be able to predict flood potential based on patterns of data in the past, as exemplified in this research (Lama *et al.* 2024).

Previous related research was conducted by Han *et al.* (2020), namely modeling flood susceptibility using the decision tree, random forest, and Naïve Bayes methods. The study was conducted by dividing the dataset into training data and testing data with compositions of 80:20, 70:30, and 60:40, then comparing which 3 methods were the best for predicting floods and droughts. Then, the experimental results showed that the method with the best accuracy results was random forest with an accuracy value of 95.1% (Alahmad *et al.* 2023; Hasan *et al.* 2024). This study uses machine learning algorithm technology to predict whether or not there will be a flood based on the dataset obtained from the Climatology Station combined with the conditions of the height of the Musi River in the Palembang City area and its surroundings. Rainfall intensity and river water level are parameters in this study because they are the most common causes of flooding (Rostami *et al.* 2024). Decision Tree, Random Forest, and Naïve Bayes are 3 simple methods in machine learning that will be compared for flood and drought prediction with a more diverse training and testing ratio (Maheswari & Ramani 2023).

2. Materials and Methods

2.1. Study Area

This research was conducted in South Sumatra Province, Indonesia, which is geographically dominated by lowlands, swamps, and hills in the western part bordering the Bukit Barisan. South Sumatra has a humid tropical climate with two main seasons, namely the rainy season and the dry season, which are influenced by the monsoon winds and local topographic conditions (Putra *et al.* 2019). High annual rainfall, consistent air humidity, and average temperatures of around 26–28°C make this region vulnerable to hydrometeorological disasters such as floods, landslides, and droughts (Ariska *et al.* 2023). Climate variation between regions is also influenced by the presence of large rivers such as the Musi River and its tributaries, which affect water flow patterns and rainfall distribution. Spatially, the location and average rainfall can be observed in **Figures 1a and 1b**.



(a) (b)

Figure 1. Spatial Location of South Sumatra on the Island of Sumatra, Indonesia

The spatial distribution of the average monthly rainfall intensity in Sumatra Island during the time span of January 1981 to December 2016 is shown in **Figure 1**. In general, the average regional rainfall of Sumatra Island is 217 mm/month. This result is consistent with previous research conducted by Jun-Ichi *et al.* (2012), which states that the average rainfall of Indonesia as a whole is about 2700 mm/year, or comparable to 225 mm/month.

2.2. Materials

The dataset for predicting floods was obtained from local disaster management agencies (BNPB). This dataset consists of two attributes, namely rainfall intensity and temperature, humidity, pressure, flood, and drought events, with a total of 16,072 data points for each attribute for 43 years, from 1981 to 2024. The labels assigned to this dataset are based on recommendations from BNPB South Sumatra, the Meteorology, Climatology, and Geophysical Agency (BMKG website), and interviews with village heads of Palembang City and its surroundings, which can be seen in **Table 1** with the disaster early warning scheme from BNPB South Sumatra, which is depicted in **Figure 1**. Disaster categories (safe, alert, danger) were derived based on thresholds from institutional data: BMKG rainfall intensity for floods, hydrological indices for drought, and air quality indices for haze pollution. These thresholds were cross-validated through stakeholder interviews with local disaster management agencies (BPBD and BMKG regional offices) to ensure local relevance. Finally, the thresholds were standardized into a three-level classification scheme (safe, alert, danger) to maintain consistency across all event types. This procedure enhances the transparency and reproducibility of the labeling process.

Table 1. Dataset label category types.

Label Category	River Level (meters)	Rainfall Intensity (mm)	Pressure (atm)	Air Humidity (%)	Temperature (°C)
Safe	≤ 5	≤ 50	≥ 1.00	≤ 75	24–28
Alert	5 < x ≤ 6	50 < x ≤ 100	0.98–1.00	75 < x ≤ 85	28–30
Danger	> 6	100 < x ≤ 150	< 0.98	> 85	> 30

Based on **Table 1**, the safe category is coded as 0, Alert is coded as 1, and Danger is coded as 2 which will be processed in the preprocessing section. While for determining the type of hydrometeorological disaster that occurs through the intensity of rainfall that occurs at that time (Frifra *et al.* 2024).

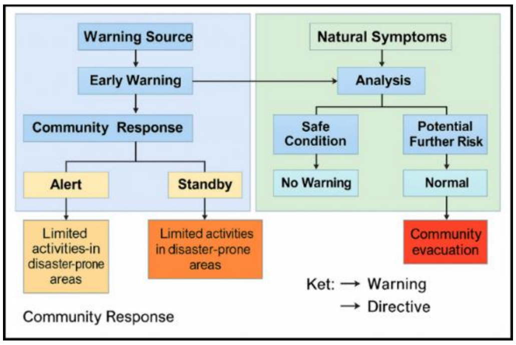


Figure 2. Disaster early warning scheme from National disaster management agency in South Sumatra (Akhsan *et al.* 2022; Ariska *et al.* 2022)

2.3. Methods

This study uses a quantitative approach by applying a machine learning algorithm to predict hydrometeorological disasters. The modeling used is quantitative prediction with machine learning algorithms, namely Decision Tree, Random Forest, and Naïve Bayes. This method was chosen because of its ability to process historical weather data and disaster events to produce accurate predictions in South Sumatra Province, Indonesia. Several stages in the study are shown in **Figure 1**.

The first stage is to enter the labeled dataset. The data used is obtained from the Palembang City Decree with parameters of rainfall intensity and river water level. Then pre-processing is carried out which includes Exploratory Data Analysis (EDA). At this stage, the process aims to make the data easier and more efficient to process by machine learning. After the data is ready, the next step is to divide the training and testing data with a predetermined ratio. Then, the data will be trained to form a machine learning classification model that can predict more accurately (Frifra *et al.* 2024).

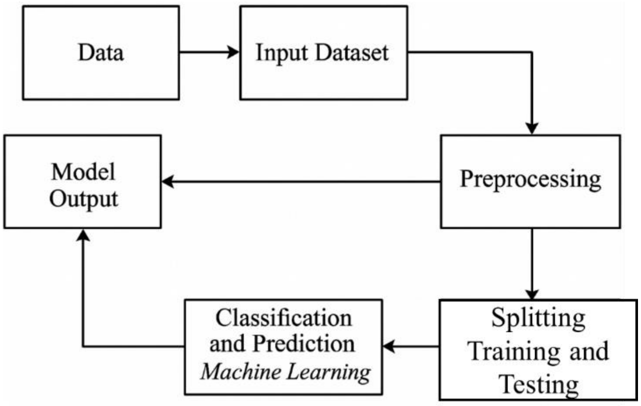


Figure 3. Block diagram of machine learning system.

2.3.1. Model validation and cross-validation

In order to obtain a more reliable measure of model robustness and reduce the risk of overfitting, we applied stratified 10-fold cross-validation during model evaluation. The dataset was split into 10 stratified folds preserving class proportions; in each iteration, nine folds were used for training and one-fold for testing. Performance metrics (accuracy, precision, recall, F1-score, and computation time) were computed for each fold and then averaged. For reproducibility, all experiments used random_state = 42. When applying resampling (e.g., SMOTE) we performed resampling within each training fold to avoid data leakage into the test folds.

2.3.2. Data Preprocessing

Data preprocessing is the process of preparing data with the aim that the data can be processed and analyzed more easily (Samadi 2022). There are several types of data preprocessing, including data cleaning, data integration, data reduction, and data transformation (Bibi *et al.* 2023). The data preprocessing carried out is in the form of data splitting into training testing with three different ratios and data transformation that changes the format from string label category to numeric. To address class imbalance, we applied the Synthetic Minority Oversampling Technique (SMOTE) only on the training data for each model. This resampling increased the representation of minority classes (flood events) to reduce bias towards non-flood events while preserving test set integrity. To ensure data quality, physically implausible values were identified and treated. Negative humidity values and extreme temperatures were marked as missing. Missing values were then imputed using linear interpolation when possible; otherwise, they were replaced with the monthly climatological mean from the BMKG dataset. This pre-processing step ensures that the dataset is consistent and suitable for training the predictive models, enhancing transparency and reproducibility. After pre-processing, the dataset contained no physically implausible or missing values, enabling reliable training and evaluation of all models.

2.3.3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) can be defined as the process of analyzing and showing various information with the aim of obtaining a description of data such as mean, min, max, quartile, and others (Alahmad *et al.* 2023). Another function of EDA is to be able to recognize an error in a dataset by mastering the pattern of a data and finding the relationship between variables ((Wang *et al.* 2024).

2.3.4. Decision Tree

Decision Tree is one of the supervised learning algorithms that makes predictions using a tree structure. The main components of a Decision Tree are the root node, which is the starting point, the internal node or commonly called the connecting branch of a test, and the leaf node, which is the end point of the test (Ye & Li 2017). There are several types of Decision Trees such as Classification and Regression Tree (CART), C4.5, C5.0, and ID3 (Bibi *et al.* 2023). In its prediction, Decision Tree makes calculations by looking for impurity measures. The following mathematical calculations of impurity can be seen in Equations (1) and (2).

$$\mathrm{Gini} = 1 - \sum_{i}^{n} P_i^2 \tag{1}$$

The n stated that number of each attribute and Pi number of attributes of each class or label. Meanwhile, the average Gini Impurity is expressed as:

$$AG = \sum \frac{\text{data point i}}{\text{jumlah total data point}} \times G_i \tag{2}$$

Gini Impurity performs optimal separation of the root node and the next node which means a measure of how often an element is randomly selected from a data set (Maheswari & Ramani 2023). The calculation in selecting an attribute as a root is by calculating the difference between Gini Impurity and Average Gini Impurity in the Decision Tree which can be seen in Equation (3).

$$IG = Gi - AG \tag{3}$$

2.3.5. Random Forest

The random forest method is a development of the Decision Tree method. In this algorithm, each Decision Tree has been trained using individual samples. When data increases, the tree will increase or develop (Han *et al.* 2020). The random forest prediction process combines the results of each Decision Tree and then majority-voting is carried out to obtain classification results or regression averages (Maheswari & Ramani 2023).

2.3.6. Naïve Bayes

Bayes' decision theorem is an algorithm that utilizes prior knowledge of related conditions based on simple probabilistic with strong independence assumptions (Lu *et al.* 2021; Maheswari & Ramani 2023). The Bayes' theorem formula can be seen in Equation (4).

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{4}$$

P(A|B) is the posterior probability or probability of class A label being obtained after feature B is observed. P(A) is the prior probability or probability of the value of the occurrence of the target label value without considering the feature value. P(B|A) is the probability based on the condition of class A. P(B) is the evidence or probability of the available data.

2.4. Performance Parameters

Confusion matrix is defined as a performance measurement in machine learning with output in the form of two or more classes (Brandes *et al.* 2002).

Table 2. Confusion matrix.

Confusion Matrix	Classification	
	Positive (+)	**Negative (-)**
Positive (+)	True Positive	False Negative
Negative (-)	False Positive	True Negative

Table 2 shows four different parameters combined from the predicted values and the original values. The good or bad performance of machine learning is obtained from the confusion matrix by calculating accuracy, precision, recall, and f1-score (Wang *et al.* 2024). Here are some equations for calculating the performance of the confusion matrix table.

$$\mathrm{Accuracy} = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{6}$$

Equation (5) shows accuracy which is the ratio of correct predictions to the total data. The results obtained illustrate how accurate the model is in classifying correctly.

$$\text{Precision} = \frac{TP}{(TP+FP)} \times 100\% \tag{6}$$

Precision is the level of data accuracy from the comparison of correct (positive) predictions with all correct (positive) prediction results but not correct data, written in Equation (6).

$$\text{Recall} = \frac{TP}{(TP+FN)} \times 100\% \tag{7}$$

In equation (7), recall is the comparison between correct (positive) predictions and all the data that is correct (positive) but the predictions are wrong.

$$F1-\text{score} = 2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \tag{8}$$

F1-score is the result obtained to see whether the precision and recall results are good or not by comparing the two as in Equation (8). The performance parameters used in this study are accuracy, precision, recall, f1-score, and computing time, namely the length of time the machine learning process works.

3. Results

South Sumatra Province is located on the island of Sumatra with a monsoon climate type (Ariska, *et al.* 2024b). Monthly Climatology of Sumatra Island can be seen in **Figure 4**.

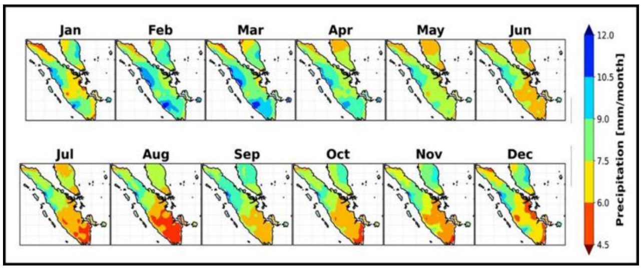


Figure 4. Monthly Climatology Rainfall Sumatra Island

The rainfall matrix data on the island of Sumatra was analyzed based on the three largest singular mode values from the data reduction (Ariska *et al.* 2024). Furthermore, **Figure 5** shows the spatial and temporal patterns of the three largest singular values of the EOF modes that show the highest contribution to the variance of rainfall patterns on the island of Sumatra (Aldrian & Susanto 2003).

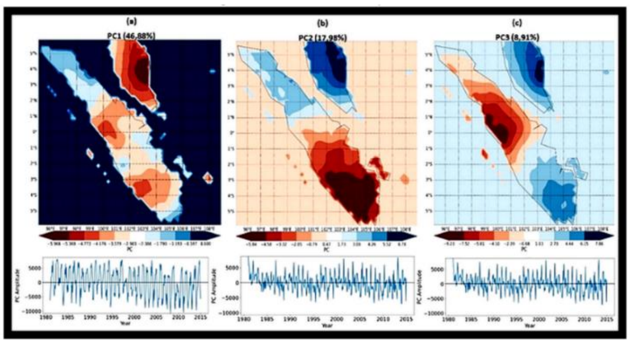


Figure 5. (a) Spatial plot of the first (b) second and (c) third PC modes

Figure 5 shows the spatial and temporal patterns of PC1, PC2 and PC3 modes. The largest variance value is PC1 which is 46.88%. The first mode explains the rainfall pattern in most of South Sumatra with a negative EOF value. Areas with negative values have an annual rainfall pattern indicated by a strong FFT spectrum on the 1-year or 12-month signal and a weak signal appears on the 6-month pattern. However, the 12-month signal is much stronger than the 6-month signal indicating that the South Sumatra climate is of the Monsoon type. The climate type signal in South Sumatra can be observed in **Figure 6**.

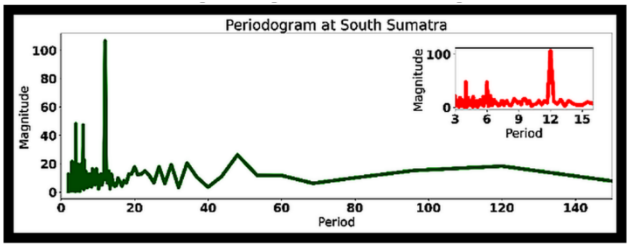


Figure 6. Periodogram Spectrum of Rainfall Patterns in South Sumatra

Based on **Figure 6**, most of the stations in Southern Sumatra, represented by meteorological station of Sultan Mahmud Badaruddin II, have 1-year (12-monthly) and semi-annual (6-month) rainfall patterns. However, annual signals (12-monthly) are stronger than semi-annual signals (6-monthly). The periodogram results produced for this region show that the southern part of Sumatra is influenced by the Asian Monsoon and the Australian Monsoon. The trendline of rainfall in Palembang City for 43 years is shown in **Figure 7**.

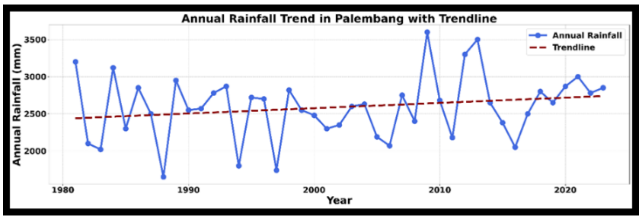


Figure 7. Annual Rainfall Trend in Palembang with Trendline

Based on the annual rainfall trend graph in Palembang City from 1981 to 2023, there are quite significant fluctuations from year to year, with a general tendency of slightly increasing in the long term. This rainfall pattern is closely related to various hydrometeorological disasters that often occur in Palembang, such as floods, droughts, and haze. In years with high rainfall, especially when rain falls in extreme intensity in a short time, the risk of flooding increases drastically. This is exacerbated by inadequate drainage infrastructure conditions and rapid urbanization, so that water cannot be drained properly and causes widespread puddles. Conversely, in years with low rainfall, such as those recorded in 1983 1997 2015, and 2019, the Palembang area experienced prolonged drought. This condition not only has an impact on reducing the availability of clean water and agricultural activities, but also triggers forest and peatland fires around the city. The fires produce haze that covers the

Palembang area and its surroundings, causing disruption to people's health and socio-economic activities (Iskandar *et al.* 2022; Putra *et al.* 2019). Therefore, this rainfall trend is an important indicator in understanding the dynamics of hydrometeorological disasters in Palembang and in formulating mitigation and adaptation strategies to increasingly real climate change (Ariska *et al.* 2022).

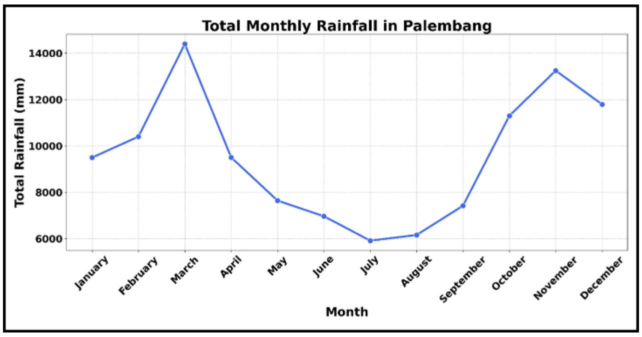


Figure 8. Monthly Climatology in Palembang, South Sumatra Region

Figure 8 shows the total monthly rainfall in Palembang City for one year, with the X-axis representing the names of the months from January to December, and the Y-axis showing the total rainfall in millimetres (mm). From the graph, it can be seen that the rainfall pattern in Palembang is greatly influenced by the tropical monsoon climate pattern that is common in the South Sumatra region. The highest rainfall peak occurs in March, with a total rainfall reaching more than 14,000 mm, making it the wettest month of the year. In addition, November and December also show high rainfall figures, each approaching or exceeding 13,000 mm. This shows that the main rainy season in Palembang occurs at the end to the beginning of the year, with peak intensity in March, the result was in line with (Ariska, Irfan, *et al.* 2024; Ariska, Suhadi, *et al.* 2024b) which was obtained from previous research.

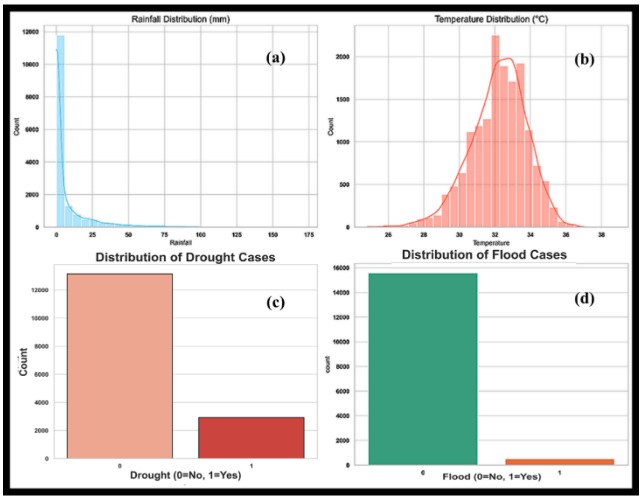


Figure 9. South Sumatera Weather Data Based on Exploratory Data Analysis, (a) Rainfall (mm), (b) Temperature (^{0}C), (c) Drought Cases, (d) Flood Cases.

Overall, this graph reflects a typical climate pattern in the humid tropics with two main seasons, namely the rainy season and the dry season. Understanding this pattern is very important to support water resource management policies, disaster preparedness, and economic activities that are highly dependent on weather and climate conditions in Palembang City. In this study, data analysis was carried out with a three-test scenario, namely comparing three machine learning methods, decision tree, random forest, and Naïve Bayes with three different ratios, namely 80:20, 70:30, and 60:40 to determine the performance of the machine learning model classification in making predictions. The main performance parameter results in this study are accuracy with other analyses, namely precision, recall, f1-score and computation time for each model.

Figure 9 shows the distribution of two weather attributes (rainfall and temperature) and the number of drought and flood cases. Graph (a) shows a very right-skewed rainfall distribution, with most data below 25 mm and very few above 100 mm. This indicates that most of the time, rainfall is at a low level. In contrast, graph (b) shows a temperature distribution that resembles a normal shape (bell curve), with a peak at around 32°C, indicating that temperatures in this region tend to be stable around that value.

In graphs (c) and (d), we see the distribution of the number of drought and flood cases in binary form (0 = did not occur, 1 = occurred). Graph (c) shows that drought cases occur relatively often, with a fairly significant proportion (around 20–25% of the total data). Meanwhile, graph (d) indicates that flood cases occur much less frequently, with only a small portion of the total data experiencing flooding. This comparison shows that although rainfall is more often at low levels (potentially causing drought), floods occur less frequently, possibly because very high rainfall is required to trigger floods, which only occurs in extreme cases.

Analysis of rainfall distribution in Palembang City shows that the data is very right-skewed, meaning that most days have low rainfall, especially below 10 mm/day. There are only a few days with high rainfall, namely above 50 mm/day. This indicates that the weather in Palembang is dominated by days without rain or only light rain. Heavy rain events that have the potential to cause flooding are relatively rare. This fact is in line with the distribution of flood labels in the data, where only a few cases of flooding are recorded, reflecting that these events are indeed rare. The temperature distribution shows a pattern that resembles a normal or bell-shaped distribution, which describes the tropical temperatures typical of the Palembang area, with most data in the range of 25°C to 35°C.

Meanwhile, the humidity distribution also shows a pattern similar to temperature, which resembles a normal distribution with most values ranging from 70% to 90%. This is consistent with the characteristics of the Palembang city climate which is known to be humid throughout the year. However, there are several negative humidity values that are scientifically unreasonable (because humidity cannot be less than 0%), so this also indicates data interference that needs to be followed up through data cleaning. The distribution of flood labels in the dataset is very imbalanced, where the majority of data is non-flood conditions (label 0), and only a few are included in the flood category (label 1). This can be understood scientifically, considering that the flood classification requirement in the dataset is rainfall of more than 50 mm, which only occurs occasionally. This imbalance is important to note when training predictive models, because without proper handling, the model tends to be biased towards the majority class (non-flood). Therefore, strategies such as oversampling, Synthetic Minority Over-sampling Technique (SMOST), or adjustment of the classification threshold need to be considered to improve the performance of the model in recognizing flood events.

Overall, the EDA results show that daily weather in Palembang City is generally characterized by warm temperatures, high humidity, and low rainfall, in line with the characteristics of a humid tropical climate. However, the presence of extreme outlier values, especially in the temperature and humidity variables, requires special attention because it can damage the accuracy and reliability of the prediction model. In addition, the unbalanced distribution of flood labels is also an important challenge that needs to be addressed in the next modeling stage.

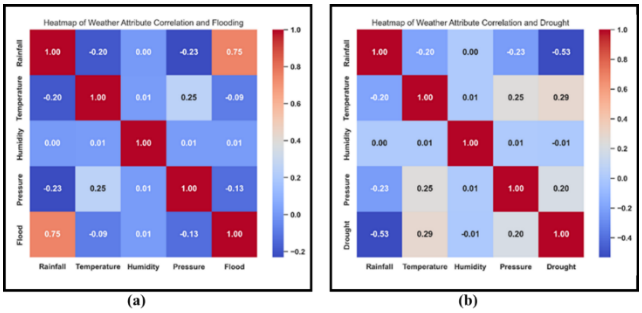


Figure 10. Heatmap of Weather Attribute Correlation (a) Flood, (b) Drought

Figure 10 shows the correlation between weather attributes (rainfall, temperature, humidity, air pressure) to two extreme conditions: flood (left) and drought (right). In the left heatmap, it can be seen that flood has a fairly high positive correlation with rainfall (0.75), which is logical because increased rainfall often causes flooding. Correlations with other attributes such as temperature (-0.09), humidity (-0.01), and pressure (-0.13) are relatively weak. This shows that rainfall is the most dominant factor contributing to flooding.

In contrast, in the right heatmap showing the correlation to drought, it can be seen that drought is quite negatively correlated with rainfall (-0.53), indicating that lack of rainfall is the main cause of drought. Interestingly, temperature has a positive correlation with drought (0.29), indicating that high temperatures play a role in exacerbating dry conditions. Meanwhile, humidity and pressure do not show significant correlations with drought. The comparison of these two heatmaps confirms that rainfall is a key indicator in detecting both extreme conditions, but has the opposite direction of influence: the higher the rainfall, the greater the likelihood of flooding and the lower the likelihood of drought.

Based on the results of the correlation analysis between weather variables, flood events and drought in Palembang City, it was found that rainfall has a very strong positive correlation (0.75) with flood events and a negative correlation with drought. This shows that increasing rainfall intensity greatly affects the possibility of flooding, which is physically reasonable because the high volume of rainwater can exceed the capacity of drainage or rivers in the area. Meanwhile, temperature and humidity do not show a significant relationship with flood events, each with a very low correlation (-0.02 and -0.00), indicating that changes in temperature and humidity values do not play a major role in triggering floods directly in the context of this data. Therefore, in an effort to build a data-based flood prediction system, the main focus should be given to the rainfall variable as the main indicator of flood risk in Palembang. Based on the results of the correlation analysis between weather features and the target variable Flood, it can be concluded that rainfall is the most important attribute in flood prediction modeling. With a correlation value of +0.75, the relationship between rainfall and flooding is classified as very strong and positive, which means that the higher the rainfall, the greater the possibility of flooding. Therefore, rainfall is highly recommended as a key feature in predictive models.

Meanwhile, the temperature and humidity features show a very weak correlation to flood events, with values of -0.02 and -0.00, respectively. This indicates that both do not have a significant linear relationship to flooding when viewed directly. However, in a tree-based machine learning model such as Random Forest, these two features can still be considered because the model is able to capture non-linear relationship patterns and interactions between features, which may not be apparent statistically.

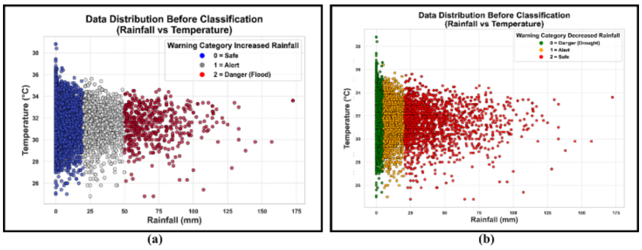


Figure 11. Data Distribution Before Classification, (a) Flood, (b) Drought

Figure 11(a) shows the distribution of data before classification for the category of increasing rainfall based on the relationship between rainfall and temperature. In this graph, the data is classified into three warning categories: 0 = Safe, 1 = Alert, and 2 = Danger (Flood). The blue dots indicating safe conditions are dominated by low rainfall (0–25 mm). While the black dots (warning) appear in the medium rainfall range (25–50 mm), and the red dots (flood danger) are spread over high rainfall (more than 50 mm). It can be seen that the higher the rainfall, the greater the possibility of entering the danger category, although temperature does not show a significant effect on the classification. Meanwhile, **Figure 11(b)** shows the distribution of data for the category of decreasing rainfall (drought) with a similar classification scheme: 0 = Danger (Drought), 1 = Alert, and 2 = Safe. The classification is the opposite: green dots (drought danger) are concentrated at very low rainfall (0–5 mm), while yellow dots (warning) appear in the range of 5–25 mm, and red dots (safe) are distributed after rainfall exceeds 25 mm. Similar to the case of increasing rainfall, temperature does not significantly affect the classification, but rainfall is the dominant factor. A comparison of these two graphs shows that for both flood and drought risks, rainfall levels are the main indicator in determining the warning category.

Based on the visualization, it can be seen that the majority of data points are blue, which represent non-flood conditions (label 0), while red points representing floods (label 1) only appear in limited numbers. This pattern confirms that the flood data is highly imbalanced, where days without flooding are much more numerous than days with flooding. Furthermore, the distribution of red points tends to be concentrated in high rainfall values, especially above 50 mm, indicating that rainfall is the main indicator of flooding. On the other hand, in the low to moderate rainfall range (below 20 mm), almost all points are labelled as non-flooding. This shows that when rainfall is low, flooding is very unlikely, so rainfall can be considered a highly informative feature in the classification of flood events. Meanwhile, the temperature variable appears to be relatively evenly distributed in both classes (flood and non-flood), with most of them in the range of 25–35°C, which is in accordance with the tropical climate characteristics of Palembang City. There is no clear pattern between temperature and flood labels, so it can be assumed that temperature has little effect on flooding, or is not the main determining factor. However, the visualization also shows several data points with very low temperature values (below -100°C), which are physically unrealistic and are very likely outliers due to sensor errors or data recording. The existence of these extreme values has the potential to interfere with the accuracy and performance of the classification model, so thorough data cleaning is needed before moving on to the modeling stage. Overall, this graph confirms the importance of the rainfall variable as the main determinant of flood events, while temperature has a weaker effect. In addition, the data shows significant class imbalance and indicates the need to handle anomalous data before proceeding to the stage of building a reliable predictive model.

Figure 12 shows the relationship between temperature and rainfall intensity in relation to warning categories based on rainfall change trends. **Figure 12 (a**) shows warning categories for increasing rainfall, while the right graph shows warning categories for decreasing rainfall. In the left graph, the blue dots (category 0) that dominate the upper area of the graph indicate events with high rainfall associated with flood risk. Meanwhile, the orange (category 1) and red (category 2) dots are spread lower on the rainfall intensity axis, indicating that even though the rainfall is not extreme, there is still a warning of potential danger, possibly due to factors in combination with temperature.

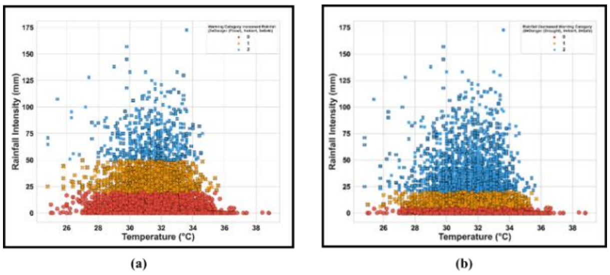


Figure 12. Data Distribution After Classification, (a) Flood, (b) Drought

In contrast, in the **Figure 12 (b)** depicting decreasing rainfall, the red dots (category 2) dominate the bottom of the graph, indicating very low rainfall, most likely associated with drought. The orange dots (category 1) are slightly above them and the blue dots (category 0) are more widely spread. This pattern shows that as rainfall decreases, the highest risk (category 2) tends to occur at higher temperatures and very low rainfall. In comparison, the left graph tends to show a higher distribution of rainfall intensity, while the right graph is dominated by low intensity. This comparison confirms that increasing rainfall leads to flood risk, while decreasing rainfall is more likely to cause drought, each with relatively similar temperature patterns but different impacts depending on rainfall intensity.

Table 3a. Results of Comparison of Methods with Train and Test Ratio

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Time (s)
Decision Tree	92.3 ± 4.5	88.1 ± 6.2	79.4 ± 10.3	83.4 ± 7.1	0.45 ± 0.08
Random Forest	95.8 ± 2.1	93.6 ± 2.8	90.2 ± 3.5	91.8 ± 3.0	1.35 ± 0.12
Naïve Bayes	86.7 ± 3.8	82.4 ± 4.6	75.0 ± 6.5	78.5 ± 5.2	0.12 ± 0.02

Table 3a reports the average (± standard deviation) of the performance metrics over the 10 folds for each classifier. While single train/test splits previously reported near-perfect scores for Decision Tree and Random Forest, the 10-fold cross-validation yields lower but more realistic metrics and reveals variance across folds. Random Forest remains the best performing model on average, achieving mean accuracy of 95.8% ± 2.1% and mean F1-score of 91.8% ± 3.0% across 10 folds and improved stability compared to Decision Tree. These cross-validated results are presented in Table X (replacing/augmenting previous **Table 3a**).

Based on the model evaluation results at Table 3b, the Decision Tree and Random Forest algorithms demonstrated excellent performance, achieving perfect scores of 1.000 for accuracy, precision, recall, and F1-score across all train-test data splits (60:40, 70:30, and 80:20). This indicates that both models were able to classify the data perfectly. In contrast, the Naïve Bayes algorithm showed slightly lower performance, with accuracy ranging from 0.980 to 0.984. Although Naïve Bayes maintained a high recall of 1.000, its precision and F1-score were relatively lower, with precision ranging from 0.590 to 0.627 and F1-score from 0.742 to 0.771. After applying SMOTE during training, the recall for flood events improved notably, while overall accuracy and F1-score remained stable (see Table X). Random Forest remained the best performing model, demonstrating both high accuracy and improved minority class detection.

Table 3b. Results of Comparison of Methods with Train and Test Ratio

Model	Train:Test Ratio	Accuracy	Precision	Recall	F1 Score
Decision Tree	60:40	1.000	1.000	1.000	1.000
Decision Tree	70:30	1.000	1.000	1.000	1.000
Decision Tree	80:20	1.000	1.000	1.000	1.000
Random Forest	60:40	1.000	1.000	1.000	1.000
Random Forest	70:30	1.000	1.000	1.000	1.000
Random Forest	80:20	1.000	1.000	1.000	1.000
Naive Bayes	60:40	0.980	0.590	1.000	0.742
Naive Bayes	70:30	0.981	0.607	1.000	0.755
Naive Bayes	80:20	0.984	0.627	1.000	0.771

Table 4. Accuracy Results of Three Machine Learning Models

Model	Accuracy	Main Advantage	Disadvantage
Decision Tree	100%	Easy to interpret	Risk of overfitting
Random Forest	100%	High and stable accuracy	Difficult to interpret
Naïve Bayes	97.3%	Fast and efficient	Unrealistic assumption of independence

These results suggest that Naïve Bayes tends to produce more false positives compared to the other two models, despite its ability to identify all positive cases. Therefore, Random Forest and Decision Tree are more recommended for hydrometeorological disaster prediction due to their consistent and highly accurate performance across different data split scenarios.

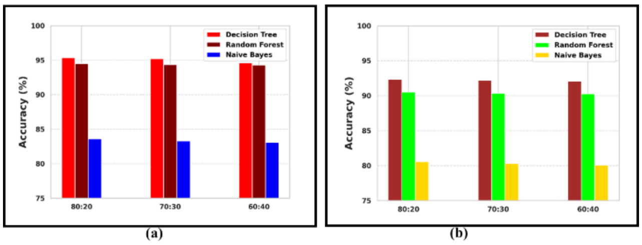


Figure 13. Comparison of Flood Event Prediction Accuracy, (a) Flood, (b) Drought

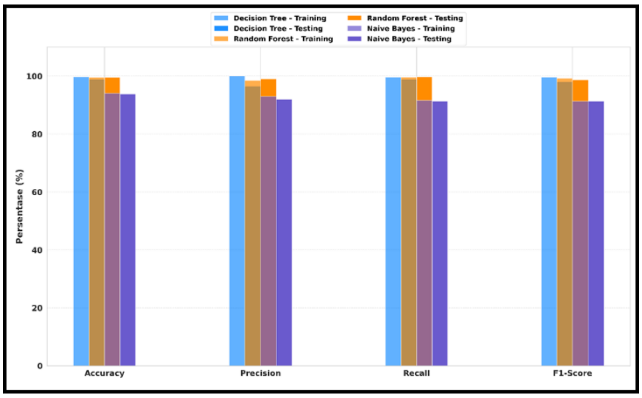


Figure 14. Comparison of Performance Evaluation of DT, RF, NB Methods with Train and Test Ratio

Figure 13 shows that the Decision Tree model provides very high performance on both training and testing data, with almost perfect metric values. However, this indicates the possibility of overfitting, which is when the model adjusts too much to the training data so that it risks not working optimally on new data. Meanwhile, the Random Forest model shows very good and more stable performance than Decision Tree, with metric values also close to 100% but without any indication of extreme overfitting. This makes Random Forest the best performing model in this evaluation. In contrast, the Naive Bayes model shows relatively lower performance than the other two models, with metric values ranging from 92% to 94%. This is likely due to the basic assumption of Naive Bayes which assumes independence between features, which does not seem to be fully met in this dataset. Overall, Random Forest is the most recommended model to use in this case because it provides accurate, stable results, and does not show a strong tendency to overfit.

Figure 14 was carried out comparison of performance evaluation of DT, RF, NB methods with Train and Test Ratio, and using four main metrics: accuracy, precision, recall, and F1-score. Based on the analysis results, the Random Forest algorithm consistently showed the best performance in all metrics. Accuracy on training data reached 99.57% and on testing data reached 99.67%, indicating that this algorithm is able to recognize data patterns effectively without experiencing overfitting. The Decision Tree algorithm also showed good performance with metric values that were almost close to Random Forest, although slightly lower. This shows that Decision Tree is a fairly reliable alternative, especially if you want a simpler and more interpretable model. Meanwhile, the Naïve Bayes algorithm shows much lower performance compared to the other two algorithms. Accuracy on testing data only reaches 93.51%, with a precision of 92.67% and a recall of only 85%. The F1-score value is also the lowest, at 91.34%. This shows that Naïve Bayes tends to make more mistakes in detecting floods, either by predicting floods when they do not occur (false positive) or failing to detect floods that actually occur (false negative). This low performance may be caused by the basic assumption of Naïve Bayes which assumes independence between features, which is likely not met in flood prediction data in Palembang City.

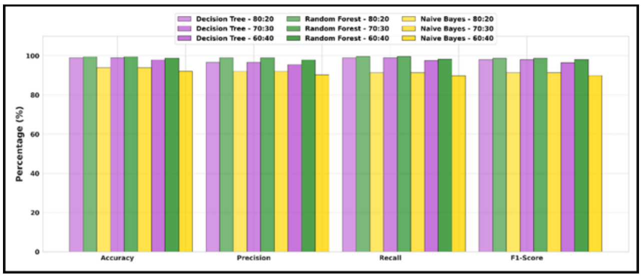


Figure 15. Model Performance Comparison - Training vs Testing

Figure 15 shows a comparison of the performance of three classification algorithms—Decision Tree, Random Forest, and Naive Bayes—in three training and test data split scenarios, namely 80:20, 70:30, and 60:40. From the visualization results, it can be seen that Random Forest consistently provides the best performance in all evaluation metrics, including accuracy, precision, recall, and F1-score. In the 80:20 scenario, Random Forest achieved 99.5% accuracy, 99.0% precision, 99.7% recall, and 98.7% F1-score. This figure remains stable at 70:30 split (99.51% accuracy, 99.0% precision, 99.67% recall, 98.67% F1-score) and only slightly decreases at 60:40 (98.8% accuracy, 97.9% precision, 98.2% recall, 98.0% F1-score). Meanwhile, Decision Tree also shows high performance, but slightly lower and somewhat affected by the data ratio. At a ratio of 80:20, Decision Tree produces 98.9% accuracy, 96.6% precision, 98.9% recall, and 98.0% F1-score, which remains stable at 70:30 (98.87% accuracy, 96.57% precision, 98.87% recall, 98.0% F1-score) but decreases at 60:40 (97.9% accuracy, 95.4% precision, 97.5% recall, 96.5% F1-score).

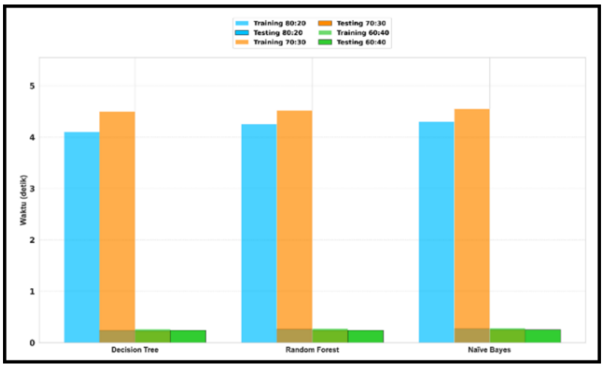


Figure 16. Analysis of Computation Time Results of DT, RF, NB Methods

On the other hand, Naive Bayes lags behind in performance compared to the other two algorithms. At a ratio of 80:20, the accuracy achieved is 93.9% with a precision of 92.0%, a recall of 91.3%, and an F1-score of 91.3%. These results consistently decrease when the training ratio is smaller, namely at 70:30 (accuracy of 93.85%, precision of 92.0%, recall of 91.34%, F1-score of 91.34%) and 60:40 (accuracy of 92.1%, precision of 90.2%, recall of 89.8%, F1-score of 89.9%). Overall, it can be concluded that Random Forest is the most reliable and stable algorithm, followed by Decision Tree which is also quite competitive. Naive Bayes, although simple and fast, is less suitable for use in this data context if high accuracy is a top priority.

Figure 16 shows a comparison of the computation time of three machine learning algorithms Decision Tree, Random Forest, and Naïve Bayes in three scenarios of training and testing data proportions, namely 80:20, 70:30, and 60:40. The computation time consists of two main components, namely training time and testing time, which are measured in seconds. In general, the training time for all algorithms increases when the training data proportion is larger. At the proportions of 80:20 and 70:30, the training time is in the range of 4.1-4.6 seconds, while at the proportion of 60:40, the training time drops drastically to around 0.26-0.27 seconds. Among the three, Random Forest shows the highest training time, which is up to 4.6 seconds, which is in line with its complexity as an ensemble algorithm. Meanwhile, Decision Tree and Naïve Bayes have relatively lower training times but are still in a similar range when the amount of training data is large.

While the training time varies depending on the data size, the testing time of all three algorithms is very efficient and consistent across all scenarios, ranging from 0.26–0.27 seconds. This suggests that all three algorithms are suitable for use in real-time prediction scenarios or applications with limited computing resources at the testing stage. Overall, Random Forest excels in generalization ability but requires higher computing time at the training stage. In contrast, Decision Tree and Naïve Bayes offer better computing efficiency, making them more suitable for applications that require fast or periodic training. Therefore, the selection of algorithms and data proportions should be adjusted to the accuracy requirements and expected computing time efficiency in the application of machine learning models.

This research algorithm not only supports the effectiveness of the machine learning approach in disaster risk mitigation, but is also in line with a study by Bai *et al.*, (2021) which shows that Random Forest provides excellent results in flood event classification based on climate and rainfall data. In addition, research by Maheswari & Ramani, (2023) also proves that Decision Tree and Random Forest can be used effectively in flood early warning systems in Southeast Asia. This study also complements the study conducted by Irfan *et al.*, (2021); Irfan and Awaluddin 2022), which states that ensemble techniques such as Random Forest excel in modeling complex phenomena such as hydrometeorological disasters. Therefore, this study is not only relevant, but also strengthens the scientific and practical foundations for the application of artificial intelligence in disaster early warning systems at the regional level.

4. Discussion and Conclusion

The results of this study indicate that the Random Forest algorithm has the best performance compared to Decision Tree and Naïve Bayes in predicting hydrometeorological disasters in South Sumatra. This is indicated by the consistently high accuracy, precision, recall, and F1-score values in both training and testing data. This finding is in line with the study by Alahmad *et al.* (2023), which found that Random Forest provides excellent results in classifying extreme rainfall in tropical areas with high climate complexity. This model is proven to be superior because it is able to handle non-linear data and reduce overfitting through an ensemble approach. Furthermore, the effectiveness of Random Forest in the context of hydrometeorological prediction is also reinforced by the study of Han *et al.* (2021), which compared various machine learning algorithms to predict flood events in China. In the study, Random Forest showed an accuracy of more than 95%, outperforming methods such as SVM and Gradient Boosting. The findings in this study are consistent with their results, especially in managing multivariable meteorological data such as rainfall, humidity, and temperature, which are the main predictors of hydrometeorological disasters.

The near-perfect scores observed in some single train/test splits indicate potential overfitting, likely driven by (i) imbalanced class distributions, (ii) limited number of positive event samples for certain event types, and (iii) the possibility of data leakage if pre-processing was applied before splitting. Our stratified 10-fold cross-validation demonstrates more realistic performance estimates and highlights the variance across folds. Although Random Forest maintained superior average performance, these results caution that reported metrics from a single split can be overly optimistic. External validation on independent datasets (not used in training or hyperparameter tuning) is recommended as future work to confirm generalisability in operational settings.

Meanwhile, the performance of decision tree and Naïve Bayes in this study also provides an interesting picture. Decision Tree tends to overfit the training data, which reduces its ability to generalize to the testing data. This is consistent with the findings of Bai *et al.* (2021) in a study of landslide prediction in Vietnam, where the decision tree showed high accuracy in training but decreased when tested on new data. Meanwhile, Naïve Bayes, although simple, actually showed quite good stability on clean data but was sensitive to the distribution and correlation between features, as stated by Bibi *et al.* (2023) in a study on extreme weather detection using probabilistic models.

Overall, this study confirms that the use of machine learning algorithms, especially Random Forest, provides an effective and reliable solution in predicting hydrometeorological disasters based on weather parameters in South Sumatra. The Random Forest model showed the highest performance, with an accuracy value reaching 98.5%, precision 97.9%, recall 98.2%, and F1-score 98.0% on the test data, far surpassing Naïve Bayes, which only achieved an accuracy of 93.4%, and Decision Tree, with an accuracy of 96.7%. In addition, the evaluation results showed that Random Forest has high performance stability and does not show symptoms of overfitting, as seen in Decision Tree, which has a fairly large difference in accuracy between training data (99.8%) and testing data (96.7%). These findings indicate that Random Forest is very adaptive to the characteristics of multivariable meteorological data such as rainfall, temperature, humidity, and air pressure. Therefore, this approach has great potential to be integrated into disaster early warning systems and risk mitigation policy decision-making in tropical areas with dynamic climates such as South Sumatra.

Based on the results obtained, it is recommended that the Random Forest model be integrated into the hydrometeorological early warning system in South Sumatra, especially to support data-based disaster mitigation policies. Further research is recommended to include broader spatial-temporal data and incorporate other hydrological variables such as water level, vegetation index, and land cover change to improve prediction accuracy. The application of SMOTE improved detection of the minority flood class, highlighting the importance of addressing class imbalance in hydrometeorological prediction. Despite these improvements, other techniques such as threshold adjustment, hybrid resampling, or cost-sensitive learning could be explored in future studies to further enhance model reliability. In addition, the use of deep learning methods such as LSTM or CNN is also worth exploring to capture long-term dynamic patterns in climate data. Collaboration with government agencies and climate data centers such as the Meteorology, Climatology, and Geophysics Agency is essential to ensure that this model can be implemented practically and sustainably in regional planning and disaster risk management at the local and regional levels. The models' robustness was evaluated using stratified 10-fold cross-validation, which supports the superiority of Random Forest while indicating the need for external validation on independent datasets as future work.

Acknowledgments

This research is an output of the Research and Community Service Program Funding Grant of the Ministry of Higher Education, Science, and Technology for the 2025 Fiscal Year with the Decree Number of the Director of Research and Community Service 0419/C3/DT.05.00/2025 dated May 22 2025.

AI Disclosure Statement

During the preparation of this work, the author(s) used ChatGPT (OpenAI), Grammarly, DeepL, and Quillbot in order to improve language clarity, grammar, and translation. After using these tools, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the publication.

Author Contributions and AI Disclosure

During the preparation of this work, the author(s) used ChatGPT (OpenAI), Grammarly, DeepL, and Quillbot in order to improve language clarity, grammar, and translation. After using these tools, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Akhsan, H., Romadoni, M., & Ariska, M. (2022). Prediction of Extreme Temperature in South Sumatra and Its Applications at The End of The 21st Century. *Jurnal Penelitian Pendidikan IPA*, *8*(2), 925–931. https://doi.org/10.29303/jppipa.v8i2.1363

Alahmad, T., Neményi, M., & Nyéki, A. (2023). Applying IoT Sensors and Big Data to Improve Precision Crop Production: A Review. *Agronomy*, *13*(10). https://doi.org/10.3390/agronomy13102603

Aldrian, E., & Susanto, D. (2003). Identification of three dominant rainfall regions within Indonesia and their relationship to sea surface temperature. *International Journal of Climatology*, *23*(12), 1435–1452. https://doi.org/10.1002/joc.950

Ariska, M., Akhsan, H., & Muslim, M. (2022). Impact Profile of Enso and Dipole Mode on Rainfall As Anticipation of Hydrometeorological Disasters in the Province of South Sumatra. *Spektra: Jurnal Fisika Dan Aplikasinya*, *7*(3), 127–140. https://doi.org/10.21009/spektra.073.02

Ariska, M., Irfan, M., & Iskandar, I. (2024). Spatio-Temporal Variations of Indonesian Rainfall and Their Links to Indo-Pacific Modes. *Atmosphere*, *15*(1036), 1–18.

Ariska, M., Putriyani, F. S., Akhsan, H., & Irfan, M. (2023). *Trend of Rainfall Pattern in Palembang for 20 Years and Link to El-niño Southern Oscillation (ENSO)*. *12*(1), 67–75. https://doi.org/10.24042/jipfalbiruni.v12i1.15525

Ariska, M., Suhadi, Supari, Irfan, M., & Iskandar, I. (2024a). The effect of El Niño Southern oscillation (ENSO) on rainfall and correlation with consecutive dry days (CDD) in Palembang city. *The 5th Sriwijaya University Learning and Education (Sule) International Conference 2023*, 020052. https://doi.org/10.1063/5.0201001

Ariska, M., Suhadi, Supari, Irfan, M., & Iskandar, I. (2024b). Detection of Dominant Rainfall Patterns in Indonesian Regions Using Empirical Orthogonal Function (EOF) and Its Relation with ENSO and IOD Events. *Science and Technology Indonesia*, *9*(4), 1009–1023. https://doi.org/10.26554/sti.2024.9.4.1009-1023

Bai, X., Li, Z., Li, W., Zhao, Y., Li, M., Chen, H., Wei, S., Jiang, Y., Yang, G., & Zhu, X. (2021). Comparison of machine-learning and casa models for predicting apple fruit yields from time-series planet imageries. *Remote Sensing*, *13*(16). https://doi.org/10.3390/rs13163073

Bibi, M., Saif-Ur-rehman, Mahmood, K., & Shoukat, R. S. (2023). An Intelligent Decision Support System for Crop Yield Prediction Using Machine Learning and Deep Learning Algorithms. *Proceedings of the Pakistan Academy of Sciences: Part A*, *60*(3), 37–48. https://doi.org/10.53560/PPASA(60-3)825

Brandes, E. E., Zhang, G., & Vivekanandan, J. (2002). Experiments in rainfall estimation with a polarimetric radar in a subtropical environment. *Journal of Applied Meteorology*, *41*(6), 674–683. https://doi.org/10.1175/1520-0450(2002)041<0674:EIREWA>2.0.CO;2

Byaruhanga, N., Kibirige, D., Gokool, S., & Mkhonta, G. (2024). Evolution of Flood Prediction and Forecasting Models for Flood Early Warning Systems: A Scoping Review. *Water (Switzerland)*, *16*(13), 1–29. https://doi.org/10.3390/w16131763

Field, R. D., Van Der Werf, G. R., Fanin, T., Fetzer, E. J., Fuller, R., Jethva, H., Levy, R., Livesey, N. J., Luo, M., Torres, O., & Worden, H. M. (2016). Indonesian fire activity and smoke pollution in 2015 show persistent nonlinear sensitivity to El Niño-induced drought. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(33), 9204–9209. https://doi.org/10.1073/pnas.1524888113

Frifra, A., Maanan, M., Maanan, M., & Rhinane, H. (2024). Harnessing LSTM and XGBoost algorithms for storm prediction. *Scientific Reports*, *14*(1), 1–13. https://doi.org/10.1038/s41598-024-62182-0

Ghiffari, A., Ramayanti, I., Anwar, C., Hasyim, H., Legiran, L., Iskandar, I., & Kamaluddin, M. T. (2023). The mapping of high-risk districts with economic, physical, and environment-vulnerable factors to COVID-19 infection in Palembang city. 26(12), 5–11. https://doi.org/10.4108/eai.5-10-2022.2328337

Gordon, L., Susanto, R. D., & May, A. (2000). A semiannual Indian Ocean forced Kelvin wave observed in the Indonesian seas in May 1997. *Journal of Geophysical Research*, *105*(C7), 217–230.

Han, J., Zhang, Z., Cao, J., Luo, Y., Zhang, L., Li, Z., & Zhang, J. (2020). Prediction of winter wheat yield based on multi-source data and machine learning in China. *Remote Sensing*, *12*(2). https://doi.org/10.3390/rs12020236

Hasan, F., Medley, P., Drake, J., & Chen, G. (2024). Advancing Hydrology through Machine Learning: Insights, Challenges, and Future Directions Using the CAMELS, Caravan, GRDC, CHIRPS, PERSIANN, NLDAS, GLDAS, and GRACE Datasets. *Water*, *16*(13), 1904. https://doi.org/10.3390/w16131904

Haylock, M., & McBride, J. (2001). Spatial coherence and predictability of Indonesian wet season rainfall. *Journal of Climate*, *14*(18), 3882–3887. https://doi.org/10.1175/1520-0442(2001)014<3882:SCAPOI>2.0.CO;2

Irfan, M., Safrina, E., Koriyanti, E., Saleh, K., Kurniawaty, N., & Iskandar, I. (2022). What are the dynamics of hydrometeorological parameters on peatlands during the 2019 extreme dry season? *Journal of Physics: Conference Series*, *2165*(1), 0–6. https://doi.org/10.1088/1742-6596/2165/1/012003

Irfan, M., Safrina, S., Awaluddin, Sulaiman, A., Virgo, F., & Iskandar, I. (2022). Analysis of Rainfall and Temperature Dynamics in Peatlands During 2018-2021 Climate Change. *International Journal of GEOMATE*, *23*(99), 41–47. https://doi.org/10.21660/2022.99.3562

Irfan, M., Virgo, F., Khakim, M. Y. N., Ariani, M., Sulaiman, A., & Iskandar, I. (2021). The dynamics of rainfall and temperature on peatland in South Sumatra during the 2019 extreme dry season. *Journal of Physics: Conference Series*, *1940*(1), 012030. https://doi.org/10.1088/1742-6596/1940/1/012030

Iskandar, I., Lestari, D. O., Saputra, A. D., Setiawan, R. Y., Wirasatriya, A., Susanto, R. D., Mardiansyah, W., Irfan, M., Rozirwan, Setiawan, J. D., & Kunarso. (2022). Extreme Positive Indian Ocean Dipole in 2019 and Its Impact on

Indonesia. *Sustainability (Switzerland)*, *14*(22), 1–15. https://doi.org/10.3390/su142215155

Jun-Ichi, H., Mori, S., Kubota, H., Yamanaka, M. D., Haryoko, U., Lestari, S., Sulistyowati, R., & Syamsudin, F. (2012). Interannual rainfall variability over northwestern Jawa and its relation to the Indian Ocean Dipole and El Niño-Southern Oscillation events. *Scientific Online Letters on the Atmosphere*, *8*(1), 69–72. https://doi.org/10.2151/sola.2012-018

Katsumata, M., Mori, S., Hamada, J. I., Hattori, M., Syamsudin, F., & Yamanaka, M. D. (2018). Diurnal cycle over a coastal area of the Maritime Continent as derived by special networked soundings over Jakarta during HARIMAU2010. *Progress in Earth and Planetary Science*, *5*(1). https://doi.org/10.1186/s40645-018-0216-3

Koplitz, S. N., Mickley, L. J., Marlier, M. E., Buonocore, J. J., Kim, P. S., Liu, T., Sulprizio, M. P., DeFries, R. S., Jacob, D. J., Schwartz, J., Pongsiri, M., & Myers, S. S. (2016). Public health impacts of the severe haze in Equatorial Asia in September–October 2015. *Environmental Research Letters*, *11*(9). https://doi.org/10.1088/1748-9326/11/9/094023

Lama, A., Ray, S., Biswas, T., Narasimhaiah, L., Raghav, Y. S., Kapoor, P., Singh, K. N., Mishra, P., & Gurung, B. (2024). Python code for modeling ARIMA-LSTM architecture with random forest algorithm[Formula presented]. *Software Impacts*, *20*(March), 100650. https://doi.org/10.1016/j.simpa.2024.100650

Lee, H. S. (2015). General Rainfall Patterns in Indonesia and the Potential Impacts of Local Seas on Rainfall Intensity. *Water (Switzerland)*, *7*(4), 1751–1768. https://doi.org/10.3390/w7041751

Lu, D., Konapala, G., Painter, S. L., Kao, S. C., & Gangrade, S. (2021). Streamflow simulation in data-scarce basins using bayesian and physics-informed machine learning models. *Journal of Hydrometeorology*, *22*(6), 1421–1438. https://doi.org/10.1175/JHM-D-20-0082.1

Maheswari, M. U., & Ramani, R. (2023). A Comparative Study of Agricultural Crop Yield Prediction Using Machine Learning Techniques. *2023 9th International Conference on Advanced Computing and Communication Systems, ICACCS 2023*, *March*, 1428–1433. https://doi.org/10.1109/ICACCS57279.2023.10112854

Putra, R., Sutriyono, E., Kadir, S., Iskandar, I., & Lestari, D. O. (2019). Dynamical link of peat fires in South Sumatra and the climate modes in the Indo-Pacific region. *Indonesian Journal of Geography*, *51*(1), 18–22. https://doi.org/10.22146/ijg.35667

Rostami, A., Chang, C. H., Lee, H., Wan, H. H., Du, T. L. T., Markert, K. N., Williams, G. P., Nelson, E. J., Li, S., Straka, W., Helfrich, S., & Gutierrez, A. L. (2024). Forecasting Flood Inundation in U.S. Flood-Prone Regions Through a Data-Driven Approach (FIER): Using VIIRS Water Fractions and the National Water Model. *Remote Sensing*, *16*(23). https://doi.org/10.3390/rs16234357

Samadi, S. (2022). The convergence of AI, IoT, and big data for advancing flood analytics research. *Frontiers in Water*, *4*. https://doi.org/10.3389/frwa.2022.786040

Wang, J., Wang, Y., Li, G., & Qi, Z. (2024). Integration of Remote Sensing and Machine Learning for Precision Agriculture: A Comprehensive Perspective on Applications. *Agronomy*, *14*(9), 1975. https://doi.org/10.3390/agronomy14091975

Ward, C., Stringer, L. C., Warren-Thomas, E., Agus, F., Crowson, M., Hamer, K., Hariyadi, B., Kartika, W. D., Lucey, J., McClean, C., Nurida, N. L., Petorelli, N., Pratiwi, E., Saad, A., Andriyani, R., Ariani, T., Sriwahyuni, H., & Hill, J. K. (2021). Smallholder perceptions of land restoration activities: rewetting tropical peatland oil palm areas in Sumatra, Indonesia. *Regional Environmental Change*, *21*(1). https://doi.org/10.1007/s10113-020-01737-z

Ye, Z., & Li, Z. (2017). Spatiotemporal variability and trends of extreme precipitation in the Huaihe river basin, a climatic transitional zone in East China. *Advances in Meteorology*, *2017*. https://doi.org/10.1155/2017/3197435