# Random forest algoritm to predict landslide based rainfall parameters

Awang Hendrianto Pratomo[a,1*], Wilis Kaswidjanti[a,2], Eko Teguh Paripurno[a,3], Johan Danu Peasetyo[a,4], Octavina Yenni Siregar[a,5]

[a]Jurusan Informatika Fakultas Teknik Industri UPN "Veteran" Yogyakarta
[1]awang@upnyk.ac.id *; [2] wilis.kas@gmail.com, [3]vynspermadi@upnyk.ac.id, [4]sylvert@upnyk.ac.id, [5]paripurno@upnyk.ac.id, [6]johandanu@upnyk.ac.id, [7]octavina.yenni@gmail.com
* Corresponding Author

Received 8 September 2021; accepted 21 December 2021; published 21 January 2022

**ABSTRACT**

Landslides are one of the most destructive natural disasters because they can cause drastic changes in environmental morphology and damage to natural and artificial structures on earth. In the Special Region of Yogyakarta Province (DIY), landslides have generated in large economic losses and casualties in the 2015-2019 period. By looking at this, efforts are needed to minimize the impact by detecting early signs and possibility of landslides. Random Forest (RF) is a machine learning algorithm that can be used to predict landslides. In this study, the RF algorithm is applied to two models, which are a model with default parameters and a model with tuning parameters. Disaster event data, consisting of 3,848 data records and 21 attributes, are processed and analyzed to determine the correlation between landslide conditioning factors and disaster events. Processing and analyzing the data generates a landslide event dataset with 826 data records and 8 attributes consisting of 206 data records of landslide class and 620 data records of non-landslide class. Prediction parameters used are daily rainfall (CHH), three-days cumulative rainfall (CHK_3H), one-month cumulative rainfall (CHK_1B), soil type, slope, area elevation, and land use. The results of the performance test show that the RF algorithm can be applied to predict landslides triggered by rainfall in the Special Region of Yogyakarta Province. Of the two RF models applied, the model with tuning parameters obtained the best performance results with 87.65% accuracy, 89.66% precision, and 60.47% recall in system testing using test data. The model also produces the highest accuracy and precision with values of 84.62% and 45.95% respectively in system testing using validation data. Meanwhile, the highest recall value of 46.34% is obtained by the model with default parameters.

**KEYWORDS**
Machne Learning
prediction
landslide
random forest



## 1. Introduction

Landslides are described as forms and processes resulting from the displacement of hillsideforming materials downwards and outwards such as soil, debris, or rocks driven by gravitational forces, which are also often assisted by water [1]. [2] mentioned that in most cases of landslides, rainfall is the main triggering factor for landslides because it can encourage an increase in pore water pressure in the soil. High-intensity rainfall in short durations can trigger shallow landslides, while most others are caused by long-lasting rainfall. [3] found that 2- and 3-day (medium-term) rainfall had the strongest influence on large landslides, while 30-day (long-term) rainfall had some additional influence on the Loznica, Western Serbia case study.

The impact of landslide disasters not only results in huge economic losses and casualties, but can also result in drastic changes in environmental morphology and damage to the natural and artificial structures of the earth [4]. In the period 2015-2019, there have been 145 cases of landslide disasters in the Yogyakarta Special Region (DIY) Province. Of all the incidents, 55 people died and were missing, 53 people were injured, and more than 300,000 people were affected by the disaster and displaced. In addition, 517 houses and 51 public facilities were damaged by landslides [5]. Due to landslide disasters that often occur in the DIY Province area, efforts are needed to minimize the impact by early detecting signs and possibilities of landslides.

A large number of studies have tried to analyze and understand the triggers of landslides by establishing an empirical relationship between rainfall and the characteristics of landslides [6], [7], [8], [9]. In addition, several other studies also focus on the observation and analysis of spatial patterns of landslides with various methods and techniques for modeling landslide vulnerability such as *fuzzy* mathematics [10] and *Analytical Hierarchy Process* (AHP) [11], [12]. *Fuzzy* mathematics applied to produce landslide vulnerability maps can obtain accuracy results of up to 93.75%. However, this method is not conducive to the establishment of a rapid landslide hazard assessment model [13] as the membership level is difficult when facing multiple *fuzzy* sets. The application of the AHP method to predict the level of landslide vulnerability has an accuracy value of 75.8% to 79.2%. This method has an easier formation process than other methods, but it has a strong subjectivity and can be easily influenced by human factors. This is because the index weight must be calculated by manual intervention before the AHP method can be used [14].

These methods have achieved good results in certain fields of study, but there are still some aspects that need to be improved. To overcome this, it can be done by applying *a machine learning* approach. This approach paves the way for detecting new patterns, especially in large, incomplete, or multi-parameter data sets such as information about landslide events. *Support Vector Machine* (SVM), *Naïve Bayes* (NB), dan *Decision Tree* (DT) [3], [15] are some *of the machine learning* algorithms that have been applied in this field. These algorithms have managed to significantly improve computation and can better solve non-linear problems, but still show some drawbacks. For example, SVM is such a complex mathematical function that it is quite difficult for human users to understand [14] and difficult to use on a large scale. The need for attribute independence became the main drawback of the NB algorithm. Whereas, DT algorithms are prone to data *noise* and that some output attributes are not allowed [15]. *Random Forest* (RF) is a *machine learning* algorithm consisting of a structured collection of tree classifiers in which each tree relies on random vector values that are sampled independently and distributed identically and provide unit sounds for the most popular classes on *inputs* [16]. This algorithm has several advantages, such as: (1) being able to avoid *overfitting*, (2) having low biases and variants, (3) the correlation of each tree is low because forest diversity is built using a number of variables/factors, (4) strong error estimates using *Out-Of-Bag* (OOB) data, and (5) higher prediction performance [17]. The RF algorithm has been widely applied in various fields, such as research conducted [18] where RF can produce the best prediction accuracy of up to 88.19% against player statistics data *Player Unknown Battleground*. In addition, on the study [13], Rf produced an OOB generalization error of less than 0.08 when its sample data count was greater than 50 and obtained an area value below the ROC kuva of 0.093 to analyze the vulnerability of landslides triggered by earthquakes in the fault zone in the Longmen Mountains, Southwestern China. Meanwhile, in the research conducted [14], RF showed the average training and testing error rates from the five-fold cross-validation results were 8.04% and 8.76% respectively with sample sizes of 5000 and 10000 classification trees in flood hazard assessment.

The problems outlined above underlie the application of the RF algorithm as a solution offered in this study. The data used is in the form of secondary data which is then processed and categorized into prediction parameters, such as daily rainfall, three days cumulative rainfall, one month cumulative rainfall, soil type, slope slope, area height, and land use. The processing and categorization of the data is carried out because the linkage of data to each other is not yet relevant, so it is necessary to make a

correlation between conditioning factors to disaster events that are in accordance with the needs of the system. The results of the process are then used in research methods to predict landslide events. The RF algorithm in this study was applied into two forms of models, namely models with *default* parameters and models with *tuning* parameters. Furthermore, the results of the performance appraisal of the two models are compared to obtain the most optimal model. This study aims to analyze data and implement rf algorithms in modeling predictions of landslide events. The RF algorithm is expected to produce predictions to assist the government in overcoming and determining better steps in handling landslide disasters.

## 2. Method

The research methodology applied in this study is a quantitative research method using *prototyping* as a system development method. The stages of the methodology carried out are seen in Fig 1.

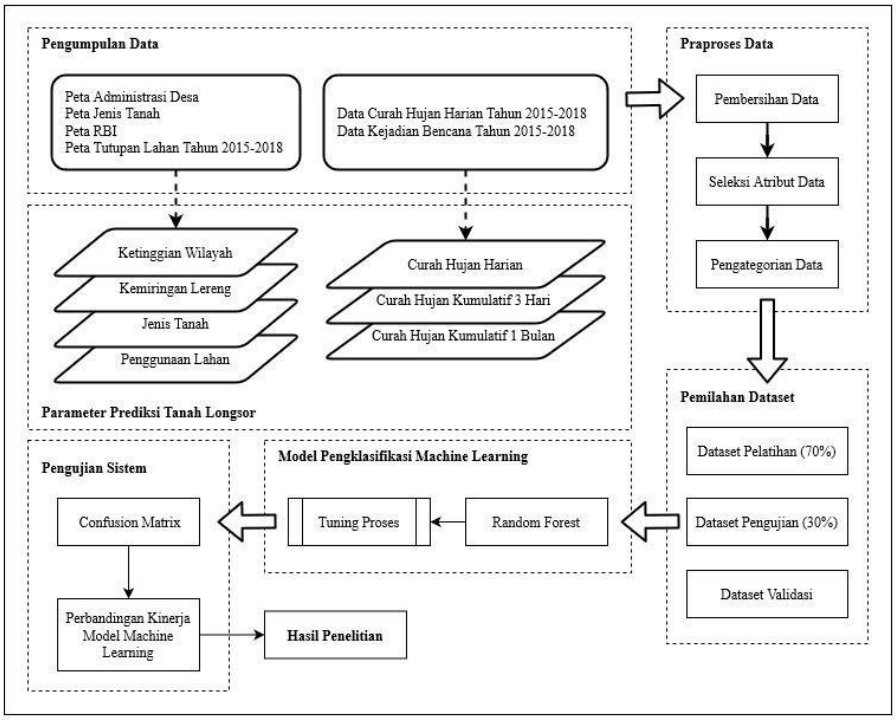


**Fig 1.** Research Methodology

### 2.1. Data Collection

This study used secondary data, where village administration maps and soil type maps of the research area were collected from the National Disaster Management Agency (BNPB) DIY. The diy province map of indonesia's earth (RBI) is provided by the Geospatial Information Agency (BIG) and can be accessed at the https://tanahair.indonesia.go.id/portal-web. From the map, a contour map is obtained that will be used to extract landslide conditioning factors, such as the height of the area and the slope of the slope. The 2015-2016 DIY Provincial land cover map was obtained from the Ministry of Environment and Forestry (KLHK). The map data collected was in the form of a .shp *file* with a scale of 1:250,000 which was then spatially analyzed using ArcGIS 10.8. Spatial analysis was carried out by overlaying each landslide conditioning factor map with the village administration map, resulting in a map with new information attributes as shown in Fig 2. Each map attribute is then *exported* into an *excel file* to preprocess the data. Daily rainfall data for 2015-2018 in river areas in 29 rainfall stations located in three districts of DIY Province, namely Gunungkidul, Bantul, and Sleman were collected from the Public Works, Housing and Mineral Resources Energy Office (Dinas PUP-ESDM) In addition, disaster event data in Gunungkidul Regency and Sleman Regency, each of which data was collected from the Regional Disaster Management Agency (BPBD) of Sleman Regency and Gunungkidul Regency was used as sample data in this study. The data consists of various detailed information such as victim data, time of

occurrence, type of disaster, location of the incident, and impact of the disaster with a total of 3848 records (rows) data.

## 2.2. Data Prepocessing

In order to get a *dataset* that matches the *machine learning* model on all the data that has been collected, manually rapprocessing the data using *the Microsoft Excel* application goes through several stages as follows.

a. Data cleansing

The data cleansing process is carried out on disaster event data to eliminate *missing values* and duplicate data based on the location of the event.

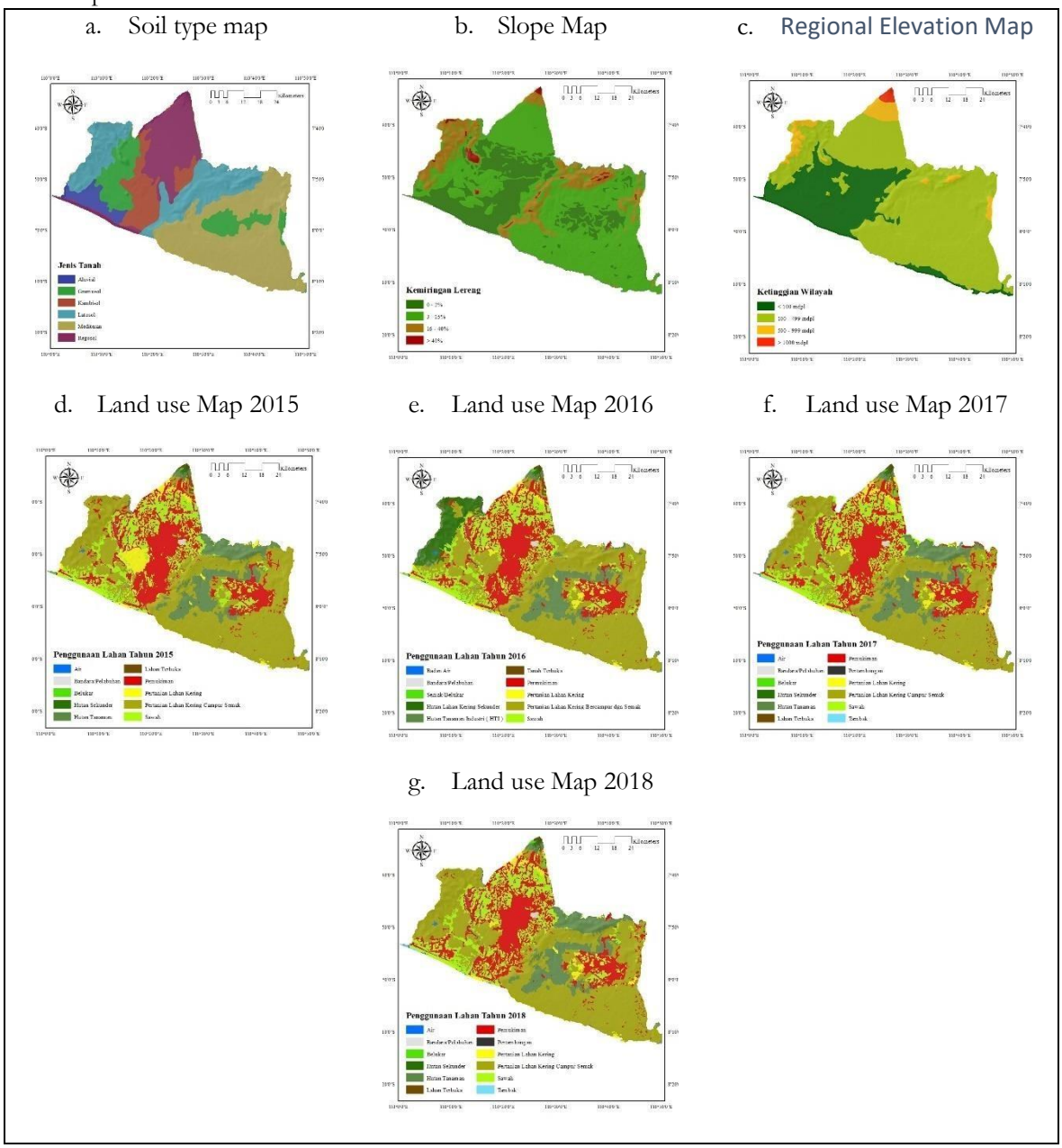


**Fig 2.** Spatial Analysis Results Of Landslide Conditioning Factor Map

At this stage, the event data is correlated with rainfall data to extract landslide conditioning factors, such as daily rainfall, three days cumulative rainfall, and one month cumulative rainfall (30 days). The value of these factors is obtained by calculating the daily and cumulative values of rainfall on the date of the disaster at the rainfall station closest to the disaster site. The determination of rainfall stations is

done manually by looking at the proximity of the location of the station area to the location of the disaster on *Google Maps* based on the travel time. In addition, disaster event data is also correlated with the results of spatial analysis of maps that have been obtained previously based on the location of the incident and produce landslide datasets with new attributes, such as soil type, slope slope, area height, and land use. Next, attribute selection is carried out.

**Table 1.** List of Attributes Used in Research

| No | Attribute Name | Information |
|---|---|---|
| 1 | CHH | Daily rainfall; precipitation value on the date of the disaster |
| 2 | CHK_3H | Three-day cumulative rainfall; cumulative rainfall value on the date of the disaster occurred up to two days earlier |
| 3 | CHK_1B | One month's cumulative rainfall; cumulative rainfall value within 30 days to the date of disaster |
| 4 | Soil Type | Types of soil at the site of the occurrence of the disaster |
| 5 | Slope | The degree of slope of the slope at the site of the disaster |
| 6 | Height of the Territory | The level of height of the territory at the site of the disaster |
| 7 | Land Use | Land use around the disaster site |
| 8 | Types of Disasters | Types of catastrophic events that occur |

**Table 2.** Parameter Categories and Frequency Ratio Analysis of Landslide Event Datasets

| No | Attribute Name | Classification | Category | *Di* | *% Di* | *Ai* | *% Ai* | *FR* |
|---|---|---|---|---|---|---|---|---|
| 1 | CHH | < 5 mm | Very light | 238 | 28.81 | 55 | 26.70 | 0.93 |
| | | 5 – 20 mm | Light | 224 | 27.12 | 50 | 24.27 | 0.90 |
| | | 21 – 50 mm | Keep | 212 | 25.67 | 57 | 27.67 | 1.08 |
| | | 51 – 100 mm | Dense | 120 | 14.53 | 33 | 16.02 | 1.10 |
| | | > 100 mm | Very bushy | 32 | 3.87 | 11 | 5.34 | 1.38 |
| 2 | CHK_3H | < 50 mm | Low | 451 | 54.60 | 83 | 40.29 | 0.74 |
| | | 50 – 99 mm | A bit low | 234 | 28.33 | 79 | 38.35 | 1.35 |
| | | 100 – 199 mm | Keep | 121 | 14.65 | 39 | 18.93 | 1.29 |
| | | 200 – 300 mm | A bit high | 15 | 1.82 | 2 | 0.97 | 0.53 |
| | | > 300 mm | Tall | 5 | 0.61 | 3 | 1.46 | 2.41 |
| 3 | CHK_1B | < 100 mm | Low | 57 | 6.90 | 7 | 3.40 | 0.49 |
| | | 100 – 300 mm | Intermediate | 277 | 33.54 | 73 | 35.44 | 1.06 |
| | | 301 – 500 mm | Tall | 366 | 44.31 | 91 | 44.17 | 1.00 |
| | | > 500 mm | Very high | 126 | 15.25 | 35 | 16.99 | 1.11 |
| 4 | Soil Type | Aluvial | Insensitive | 0 | 0.00 | 0 | 0.00 | 0.00 |
| | | Latosol | Somewhat sensitive | 144 | 17.43 | 103 | 50.00 | 2.87 |
| | | Mediteran | Lack of sensitivity | 97 | 11.74 | 32 | 15.53 | 1.32 |
| | | Grumosol, Kambisol | Sensitive | 184 | 22.28 | 19 | 9.22 | 0.41 |
| | | Litosol | Very sensitive | 401 | 48.55 | 52 | 25.24 | 0.52 |
| 5 | Slope | 0 – 2 % | Ramps | 161 | 19.49 | 14 | 6.80 | 0.35 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 3 – 15 % | A bit steep | 521 | 63.08 | 85 | 41.26 | 0.65 |
| | | 16 – 40 % | Steep | 134 | 16.22 | 102 | 49.51 | 3.05 |
| | | > 40 % | Very steep | 10 | 1.21 | 5 | 2.43 | 2.00 |
| 6 | Regional Heights | < 100 mdpl | Low | 37 | 4.48 | 2 | 0.97 | 0.22 |
| | | 100 – 499 mdpl | Intermediate | 739 | 89.47 | 191 | 92.72 | 1.04 |
| | | 500 – 999 mdpl | Tall | 47 | 5.69 | 12 | 5.83 | 1.02 |
| | | > 1000 mdpl | Very high | 3 | 0.36 | 1 | 0.49 | 1.34 |
| 7 | Land Use | Dense forests / vegetation and water bodies | Low | 87 | 10.53 | 58 | 28.16 | 2.67 |
| | | Mixed garden / shrub | A bit medium | 0 | 0.00 | 0 | 0.00 | 0.00 |
| | | Irrigated plantations and rice fields | Keep | 303 | 36.68 | 81 | 39.32 | 1.07 |
| | | Industrial estates and settlements/townships | A bit high | 434 | 52.54 | 67 | 32.52 | 0.62 |
| | | Vacant lots | Tall | 2 | 0.24 | 0 | 0.00 | 0.00 |

On a landslide *dataset* where unnecessary attributes on the event data are removed. The attributes to be removed are those that are not relevant to the research needs. The list of attributes to be used in this study can be seen in Table 1.Data categorization. Data categorization is a stage where the values in each attribute are categorized into multiple classes based on the parameter categories found in Table 2. In the disaster type attribute, the data is categorized into two classes, namely the 'Yes' class for landslide disaster type data and the 'No' class for disaster type data other than landslides. This attribute will then be used as a response variable in this study. The data preprocessing stage that has been carried out, produces a dataset of landslide events with 826 records (rows) and 8 attributes (columns). Furthermore, the *dataset* is divided into 70% of the training data of 540 data used to train the *machine learning* classifier model, while 30% of that data and validation data of 286 data are used to test the model.

## 2.3. Machine Learning

### 2.3.1 Random Forest

*Random forest* (RF) is a statistical model equipped with *ensemble* learning for classification and prediction of phenomena [19]. The RF algorithm is based on the concept of *bagging* and uses *the decision tree* as a basic classifier. This method consists of a collection of treestructured classifiers in which each tree relies on random vector values that are sampled independently and distributed identically and provides unit sounds for the most popular classes on inputs [16]. In *dataset* D with *variable p*, rf operating principle is summarized in the following stages and illustrated in Fig 3.

1. (*Bootstrap stage*) Sample data $D_1$, $D_2$, ..., $D_k$ is created by randomly retrieving from *the D dataset* by retrieval (*replacement*), where generally 2/3 of the data is selected.
2. (*Random feature selection* stage) Using samples *bootstrap*, the tree is built until it reaches its maximum size (without pruning). At each node, the selection of sorters is carried out by randomly selecting the variable *m*, where $m < p$, then the best sorter is selected based on *m* variables.
3. Steps 1 and 2 are repeated *k* times, so that a forest is formed consisting of *k* decision trees.
4. Each decision tree votes the class unit most often appears as a result of predictions. The concept of the election is called *majority vote*.

In the RF method, the formation of a single classification tree structure can be organized by specifying several modeling parameters. *Tuning* (tuning) parameters is carried out in order to obtain parameter optimization (*hyperparameters*) which aims to improve predictability or simplify the training process on the model. According to [20], Rf modeling parameters that are commonly performed *tuning* are as follows.

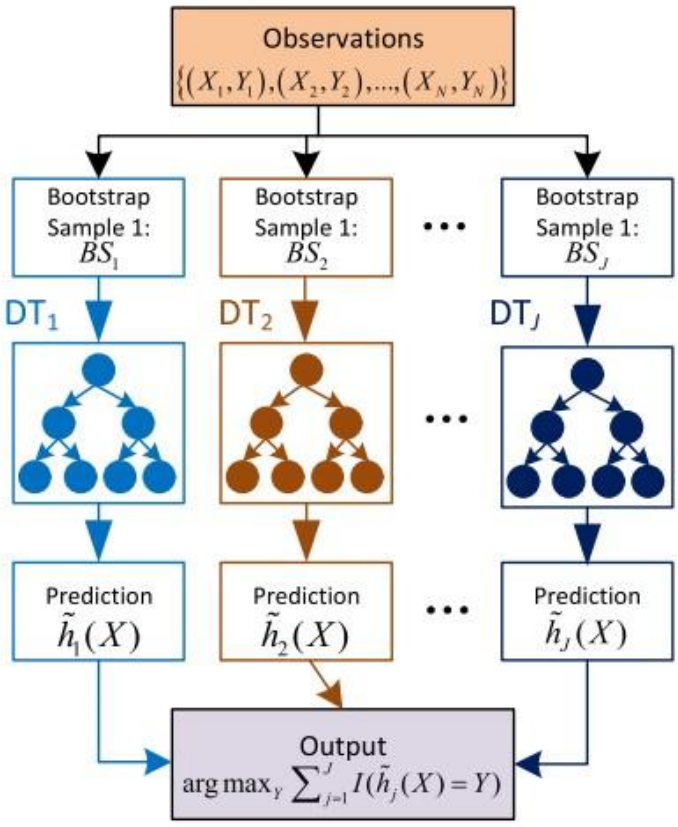


**Fig 3.** Stages of Random Forest

a. The number of random variables used as separators (*mtry*)
The standard (*default*) size *of the mtry* parameter in the RF classification algorithm is where p is the total of the variables with the smallest node value being $1\sqrt{p}$ [21]. The size of this parameter greatly affects the correlation and strength of each tree. If the size is small, the tree will have a lower correlation that has the potential to be greater for *variance* reduction. Whereas as the size increases, the effect of reducing *the variance* of randomization decreases. And if the size is equal to p, then RF is reduced to *bagging*.

b. The multiplicity of single classification trees built (*ntree*)
The sheer number of single classification trees built can be of as much value as possible because the RF classifier is computationally efficient and does not *overfit* [22], [23] recommend experimenting with multiple tree count values and stopping until *the error* values stabilize. Some studies use different *ntree* values , such as 250 [21], 500 [22], [2], [18] which used the number of trees ranging from 10 to 80 trees and obtained results in which 70 trees showed the highest accuracy values.

c. Tree size (*nsplit*, *maxnodes*)
The size of the tree can be controlled in two ways, namely explicitly by limiting the number of separations (*maxnodes*) and adaptively by telling the algorithm when the growth of the tree should stop splitting the leaf *nodes* from the tree (*nsplit*) [20].

### 2.3.2 Tuning Parameters

*Process tuning* on rf models is performed tuning parameters based on the results of the search for the best parameters. The search process uses the *Grid Search* method with 5-fold *Cross Validation*. The best parameter search is performed on *n_estimator* parameters with a range of values ranging from 10 trees to 1000 trees with multiples of 10, *max_features* ranging from 1 to 7, and *max_leaf_nodes* ranging from 2 to 50.

### 2.4 System Testing

System testing with classification method is generally carried out using *the confusion matrix* method [24]. *Confusion matrix* is used to find out how good or how much accuracy results from a classification model that has been created to predict or classify classes from *testing* data [18]. Based on the value of *the confusion matrix*, accuracy, precision, and recall values can be obtained, each of which is calculated using equations 1, 2, and 3.

$$Akurasi = \frac{Tp + Tn}{Tp + Tn + Fp + Fn} \times 100\% \quad (1)$$

$$Presisi = \frac{Tp}{Fp + Tp} \times 100\% \quad (2)$$

$$Recall = \frac{Tp}{Tp + Fn} \times 100\% \quad (3)$$

Where:
TP (*True Positive*): The number of positive observations classified as positive
FP (*False Positive*): The number of positive observations that are incorrectly classified as negative TN (*True Negative*): The number of negative observations classified as negative
FN (*False Negative*): The number of false negative observations classified as positive.

In addition to assessing performance, in system testing, an error value calculation is also carried out using *the Mean Absolute Percent Error* (MAPE) with the following formula.

$$M\ A\ P\ E = \frac{1}{n}\sum_{t-1}^{n}\frac{|A_t - F_t|}{A_t} \quad (4)$$

where$n$ is the magnitude of the test data, $A_t$ is the actual value, and is the predicted value$F_t$. *There* are two datasets used in the test, namely 30% of the dataset test set obtained from the 20152017 data and the validation set obtained from the 2018 data. Both *datasets* are then fed into the system and predicted using an RF model with *default* parameters and an RF model with *tuning* parameters. The *default* parameters of the RF algorithm consist of the number of trees (*n_estimator*) built as many as 100 trees, the number of random variables (*max_features*) is set '*auto*', that is, and the size of the tree ($\sqrt{n\_features}$ *max_leaf_nodes*) is '*none*' which means the number of leaf *nodes* is unlimited. Testing *of datasets* in the system is carried out on laptops with the following specifications.

1. *Processor Intel Core* i3
2. RAM 4 GB
3. *Hard disk* 500 GB

## 3. Results and Discusion

The results of data analysis and the implementation of the RF algorithm to predict landslide events are described based on the research methodology

### 3.1 Data Analysis

In this study, the data used was disaster event data consisting of 3,848 data *records* and 21 attributes. In this data, there are 951 disaster event data in Gunungkidul Regency and 2,897 disaster event data in Sleman Regency. All data that has been collected is then preprocessed data which goes through several stages, including data cleaning, data attribute selection, and categorization. After the preprocess, a landslide event dataset was obtained with 826 data *records* and 8 attributes consisting of 206 landslide class data records and 620 non-landslide class data records. To find out the relationship of landslide

prediction parameters to the occurrence of landslide disasters, the *Frequency Ratio* (FR) method obtained with the following formula is used.

$$FR = \frac{Di/Ai}{\sum_{t-1}^{N} Di / \sum_{t-1}^{N} Ai} \quad (5)$$

where is the amount of avalanche data on a class in a certain parameter and is the sum of the data of a class in a particular parameter. Each class in each of the landslide prediction parameters has an influence on landslide susceptibility. The FR value$DiAi$ of > 1.00 indicates that the indication of the probability of a landslide is high, while if the FR value < 1.00 then the probability of a landslide occurs is low. The results of the FR analysis for each class of landslide prediction parameters are presented in Table 2. Based on the table, the FR value in the daily rainfall (CHH) parameter is getting bigger as the CHH class increases. This indicates that the higher the intensity of daily rainfall, the higher the probability of landslides. In the three-day cumulative rainfall parameter (CHK_3H), the probability of landslides is higher when the intensity is in the classes of 50 – 99 mm, 100 – 199 mm, and > 300 mm with FR values of 1.35, 1.29, and 2.41, respectively. Meanwhile, in the cumulative rainfall parameters of one month (CHK_1B), landslide events are likely to occur in classes with an intensity of > 100 mm. In this case, it can be seen that these three parameters can be predicted parameters for landslide disaster events due to the influence of rainfall which can encourage an increase in pore water pressure in the soil so that it can trigger soil movement.

In the soil type parameters, the Latosol class is the class that most influences the occurrence of landslides with an FR value of 2.87 and followed by the Mediteran class with an FR value of 1.32. In the study area, 51.94% of landslide disasters occurred on slope parameters of class 16 – 40% (FR = 3.05) and > 40% (FR = 2.00). Meanwhile, 98.04% of landslide disasters occur on the height parameters of the region class 100 – 499 masl (FR = 1.04), 500 – 999 masl (FR = 1.02), and > 1000 masl (FR = 1.34). This shows that there is an indication of a high probability of landslides in areas with a slope level of > 16% and an altitude of > 100 meters above sea level. In land use parameters, the highest probability of landslides occurring in forests / dense vegetation and water bodies with an FR value of 2.67 which is followed by plantation areas and irrigated rice fields with an FR value of 1.07.

### 3.2. Turning Results

The best parameter search results with the *Grid Search* method are displayed into the graph in Fig 4. Based on the graph, it can be seen that the model performed well and stably for each parameter value *max_features* at the time of training, but the model was *overfitting* when *the max_features* value < 2 and *max_features* > 2 at the time of testing. Therefore, it can be concluded that the value of the best *max_features* parameter is 2. In *the max_leaf_nodes* parameters, the model's performance increases for each value during training, but when testing the value *of the max_leaf_nodes* the model performance graph increases to a value of 12 and begins to overfitting afterwards. Thus, it can be concluded that the best value for *max_leaf_nodes* parameter is 12.

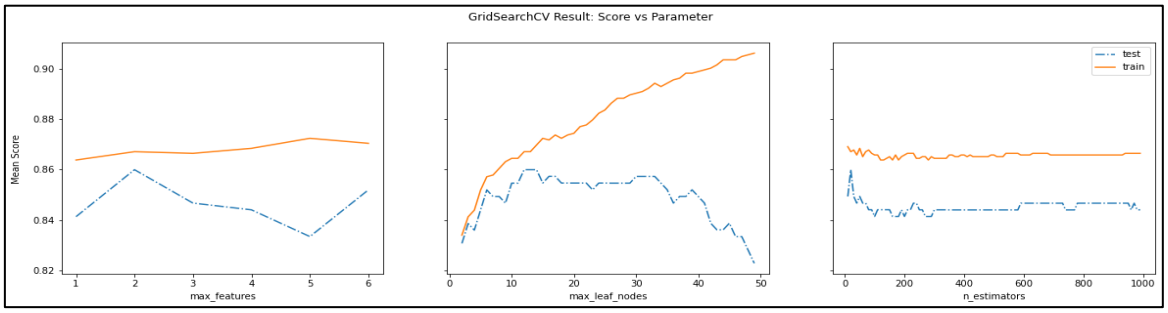


**Fig 4.** Best Parameter Search Results with Grid Search Method

Meanwhile, the best value for *the n_estimator* parameter is 20. This is because the model in the *n_estimator* parameter begins to overfitting when the value is > 20 and does not experience a significant increase in performance.

### 3.3 Machine Learning

Performance appraisal using the confusion matrix method using test data and validation data from RF models with *default* parameters (Fig 5) and RF models with *tuning* parameters (Fig 6) are repeated 100 times. The results of the calculation of the average performance value of each *machine learning* model are compared and presented in Table 3. Based on the results of the comparison of system performance tests using validation data, it can be seen that rf models with *tuning* parameters are superior to overall performance measures compared to RF models with *default* parameters. This model has the shortest runtime, which is 0.01 seconds with *an error* value of 0.15. In addition, the model also obtained the highest performance assessment on accuracy with a value of 84.62% and precision with a value of 45.95%. Nonetheless, the highest recall value of 46.34% was obtained by the *Random Forest* model with *default* parameters.

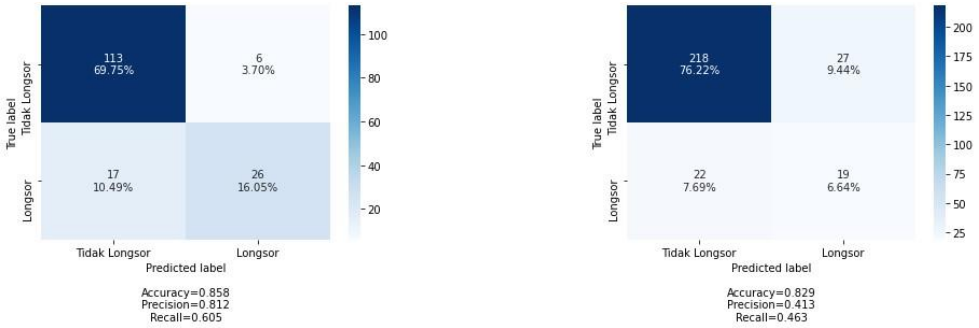


**Fig 5.** Random Forest Model Test Results with Default Parameters

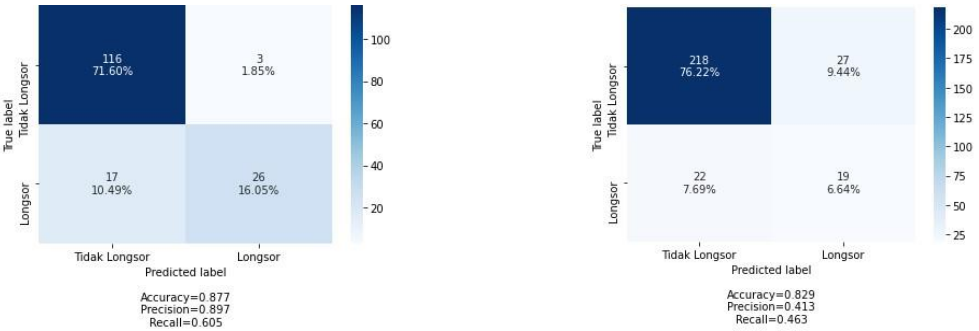


**Fig 6.** Random Forest Model Test Results with Tuning Parameters

### 4. Conclusion

Based on the results and discussions that have been described in the previous chapter, the following conclusions can be drawn. The RF algorithm can be applied to predict the possibility of landslides triggered by rainfall in the Special Region of Yogyakarta Province. The prediction parameters used include daily rainfall (CHH), three-day cumulative rainfall (CHK_3H), one-month cumulative rainfall (CHK_1B), soil type, slope slope, area height, and land use. The test results showed that the RF algorithm on models with *tuning* parameters performed best than other models in predicting landslide events triggered by precipitation. This is evidenced by the acquisition of RF model results with *tuning* parameters that have the highest values in accuracy and precision with values of 87.65% and 89.66% respectively in the test data and 84.62% and 45.95% in the validation data with a processing time of 0.01 seconds and *an error* value between 0.12 – 0.15. The system that has been built in this study still has some problems that have not been able to be resolved and do not rule out the possibility of further

development. One of them is the determination of the nearest rainfall station to the location of the landslide disaster which is still done manually. To overcome this, an algorithm can be applied that can solve the problem automatically. In addition, the data used is unbalanced data and there has been no process related to the data. Therefore, algorithms capable of handling unbalanced data can be applied to the system to support the improvement of the performance of the RF algorithm in further development. To make it easier for users to obtain information on predicting landslide events, this research can also be developed by applying a geographic information system.

## Declarations

**Author contribution.** All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.
**Funding statement.** None of the authors have received any funding or grants from any institution or funding body for the research.
**Conflict of interest.** The authors declare no conflict of interest.
**Additional information.** No additional information is available for this paper.

## References

[1] H. Wang, L. Zhang, K. Yin, H. Luo, and J. Li, "Landslide identification using machine learning," *Geosci. Front.*, vol. 12, no. 1, pp. 351–364, Jan. 2021, doi: 10.1016/J.GSF.2020.02.012.

[2] J. Dou *et al.*, "Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan," *Sci. Total Environ.*, vol. 662, pp. 332–346, Apr. 2019, doi: 10.1016/J.SCITOTENV.2019.01.221.

[3] M. Marjanović, M. Krautblatter, B. Abolmasov, U. Đurić, C. Sandić, and V. Nikolić, "The rainfall-induced landsliding in Western Serbia: A temporal prediction approach using Decision Tree technique," *Eng. Geol.*, vol. 232, pp. 147–159, Jan. 2018, doi: 10.1016/J.ENGGEO.2017.11.021.

[4] T. Kavzoglu, E. K. Sahin, and I. Colkesen, "Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression," *Landslides*, vol. 11, no. 3, pp. 425–439, Feb. 2014, doi: 10.1007/S10346-013-0391-7/METRICS.

[5] "Data Informasi Bencana Indonesia (DIBI)." https://dibi.bnpb.go.id/ (accessed Dec. 28, 2022).

[6] C. W. Chen, T. Oguchi, Y. S. Hayakawa, H. Saito, and H. Chen, "Relationship between landslide size and rainfall conditions in Taiwan," *Landslides*, vol. 14, no. 3, pp. 1235–1240, Jun. 2017, doi: 10.1007/S10346-016-0790-7/METRICS.

[7] C. W. Chen, H. Saito, and T. Oguchi, "Rainfall intensity–duration conditions for mass movements in Taiwan," *Prog. Earth Planet. Sci.*, vol. 2, no. 1, pp. 1–13, Dec. 2015, doi: 10.1186/S40645-015-0049-2/TABLES/2.

[8] G. Satya, A. H. Andriawan, A. Ridho'i, and H. Seputro, "Intensitas Curah Hujan Memicu Tanah Longsor Dangkal di Desa Wonodadi Kulon," *JPM17 J. Pengabdi. Masy.*, vol. 1, no. 01, pp. 65–71, Dec. 2014, doi: 10.30996/JPM17.V1I01.355.

[9] S. A. Rahayu, *Pengaruh Curah Hujan Terhadap Potensi Longsor Di Daerah Aliran Sungai (DAS) Citarum*. LAPAN, 2012.

[10] M. Adi Saputra, M. Lutfi Rayes, I. Nita Jurusan Tanah, F. Pertanian, and U. Brawijaya, "PEMETAAN PREDIKSI SEBARAN KERENTANAN LONGSOR DI KECAMATAN

TAWANGMANGU, KABUPATEN KARANGANYAR MENGGUNAKAN PENDEKATAN FUZZY LOGIC," *J. Tanah dan Sumberd. Lahan*, vol. 6, no. 2, pp. 1353–1359, Jul. 2019, doi: 10.21776/UB.JTSL.2019.006.2.16.

[11] M. Choiroh, "Sistem pendukung keputusan prediksi tingkat kerawanan tanah longsor menggunakan metode Double Exponential Smoothingdan Analytical Hierarchy Process," Jun. 2018.

[12] "(PDF) Assessment of Landslide Susceptibility area in Ponorogo, East Java, Indonesia Using Analytical Hierarchy Process and Natural Breaks Classification." https://www.researchgate.net/publication/318773265_Assessment_of_Landslide_Susceptibility_area_in_Ponorogo_East_Java_Indonesia_Using_Analytical_Hierarchy_Process_and_Natural_Breaks_Classification (accessed Dec. 29, 2022).

[13] H. Li, R. Liu, J. Xie, and Z. Lai, "Random forests methodology to analyze landslide susceptibility: An example in Lushan earthquake," *Int. Conf. Geoinformatics*, vol. 2016-January, Jan. 2016, doi: 10.1109/GEOINFORMATICS.2015.7378570.

[14] Z. Wang, C. Lai, X. Chen, B. Yang, S. Zhao, and X. Bai, "Flood hazard risk assessment model based on random forest," *J. Hydrol.*, vol. 527, pp. 1130–1141, Aug. 2015, doi: 10.1016/J.JHYDROL.2015.06.008.

[15] D. Tien Bui, B. Pradhan, O. Lofman, and I. Revhaug, "Landslide susceptibility assessment in vietnam using support vector machines, decision tree, and nave bayes models," *Math. Probl. Eng.*, vol. 2012, 2012, doi: 10.1155/2012/974638.

[16] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324/METRICS.

[17] M. Amiri, H. R. Pourghasemi, G. A. Ghanbarian, and S. F. Afzali, "Assessment of the importance of gully erosion effective factors using Boruta algorithm and its spatial modeling and mapping using three machine learning algorithms," *Geoderma*, vol. 340, pp. 55–69, Apr. 2019, doi: 10.1016/J.GEODERMA.2018.12.042.

[18] R. A. Haristu, "Penerapan metode Random Forest untuk prediksi win ratio pemain player Unknown Battleground," 2019.

[19] V. K. Pandey, K. K. Sharma, H. R. Pourghasemi, and S. K. Bandooni, "Sedimentological characteristics and application of machine learning techniques for landslide susceptibility modelling along the highway corridor Nahan to Rajgarh (Himachal Pradesh), India," *CATENA*, vol. 182, p. 104150, Nov. 2019, doi: 10.1016/J.CATENA.2019.104150.

[20] B. A. Goldstein, E. C. Polley, and F. B. S. Briggs, "Random Forests for Genetic Association Studies," *Stat. Appl. Genet. Mol. Biol.*, vol. 10, no. 1, p. 32, Jan. 2011, doi: 10.2202/1544-6115.1691.

[21] 14611240 Julia Widiastuti, "KLASIFIKASI PEMBIAYAAN WARUNG MIKRO MENGGUNAKAN METODE RANDOM FOREST DENGAN TEKNIK SAMPLING KELAS IMBALANCED (Studi Kasus: Data Nasabah Pembiayaan Warung Mikro Bank Syariah Mandiri KC Jambi)," May 2018, Accessed: Dec. 29, 2022. [Online]. Available: https://dspace.uii.ac.id/handle/123456789/7690

[22] M. Belgiu and L. Drăgu, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. Photogramm. Remote Sens.*, vol. 114, pp. 24–31, Apr. 2016, doi: 10.1016/J.ISPRSJPRS.2016.01.011.

[23] M. (Maulana) Dhawangkhara and E. (Edwin) Riksakomara, "Prediksi Intensitas Hujan Kota Surabaya dengan Matlab Menggunakan Teknik Random Forest dan CART (Studi Kasus Kota Surabaya)," *J. Tek. ITS*, vol. 6, no. 1, pp. 88–93, 2017, Accessed: Dec. 29, 2022. [Online]. Available: https://www.neliti.com/id/publications/193006/

[24] S. Strecker, A. Kuckertz, and J. M. Pawlowski, "Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab," *ICB Res. Reports*, no. 9, 2014, Accessed: Dec. 29, 2022. [Online]. Available: https://openlibrary.telkomuniversity.ac.id/home/catalog/id/101414/slug/data-mining-mengolah-data-menjadi-informasi-menggunakan-matlab.html