GEOMATICS AND ENVIRONMENTAL ENGINEERING • Volume 17 • Number 6 • 2023

https://doi.org/10.7494/geom.2023.17.6.129

Rido Dwi Ismanto[1], Hana Listi Fitriana[2], Johannes Manalu[3],
Alvian Aji Purboyo[4], Indah Prasasti[5]

Development of Flood-Hazard-Mapping Model Using Random Forest and Frequency Ratio in Sumedang Regency, West Java, Indonesia

Abstract: Flooding, often triggered by heavy rainfall, is a common natural disaster in Indonesia, and is the third most common type of disaster in Sumedang Regency. Hence, flood-susceptibility mapping is essential for flood management. The primary challenge in this lies in the complex, non-linear relationships between indices and risk levels. To address this, the application of random forest (RF) and frequency ratio (FR) methods has been explored. Ten flood-conditioning factors were determined from the references: the distance from a river, elevation, geology, geomorphology, lithology, land use/land cover, rainfall, slope, soil type, and topographic wetness index (TWI). The 35 flood locations from the flood-inventory map were selected, and the remaining 18 flood locations were used for justifying the outcomes. The flooded areas from the RF model were 28.39%; the rest (71.61%) were non-flooded areas. Also, the flooded areas from the FR method were 8.02%, and the non-flooded areas were 91.98%. The AUC for both methods was a similar value – 83.0%. This result is quite accurate and can be used by policymakers to prevent and manage future flooding in the Sumedang area. These results can also be used as materials for updating existing flood-susceptibility maps.

Keywords: flood-susceptibility assessment, random forest, frequency ratio, Sumedang, remote sensing

Received: 27 February 2023; accepted: 31 August 2023

[1] National Research and Innovation Agency (BRIN), Research Center for Computing, Cibinong, Indonesia, email: rido001@brin.go.id (corresponding author), https://orcid.org/0000-0002-2903-6731
[2] National Research and Innovation Agency (BRIN), Research Center for Remote Sensing, Pekayon, Indonesia, email: hana003@brin.go.id, https://orcid.org/0000-0001-6752-8414
[3] National Research and Innovation Agency (BRIN), Research Center for Remote Sensing, Pekayon, Indonesia, email: rjoh001@brin.go.id, https://orcid.org/0009-0000-1761-3725
[4] Indonesia University of Education, Department of Geographic Information Science, Bandung, Indonesia, email: alvianaji17@upi.edu, https://orcid.org/0009-0009-3870-2327
[5] National Research and Innovation Agency (BRIN), Research Center for Remote Sensing, Pekayon, Indonesia, email: inda003@brin.go.id, https://orcid.org/0000-0001-5332-6061

1. Introduction

Among the different sorts of natural hazards, a flood is one of the most damaging types of disasters; it causes significant damage [1–3] and is regarded as a serious natural danger. The extent of its harm is unquantifiable [4], it is the most prevalent natural calamity in the world according to most experts [5], and is a common natural calamity that kills people and destroys property around the world [6]. Between 1996 and 2015, the United Nations Office for Disaster Risk Reduction (UNISDR) reported 150,061 flood-related deaths worldwide; this accounted for 11.1% of all disaster deaths [7]. Floods are caused by severe rainfall that overflows into rivers and flood plains, briefly flooding the surrounding areas [8]. Therefore, a flood is linked to other disasters that might spread and cause a disastrous chain reaction [9].

In Indonesia, floods are one of the most common disasters – especially when the rainfall intensity is significant. They produce flooding in various regions – particularly in metropolitan areas with poor drainage. According to the National Agency for Disaster Countermeasure (BNPB) database (https://dibi.bnpb.go.id/), flooding has dominated all existing disasters by 37% over the last 20 years, resulting in the deaths of 2,184 people, the disappearance of 39,030 people, and the destruction of 267,198 properties. According to the same data, there were 598 flood events in Indonesia out of a total of 2401 disasters in 2022 (24.9%) (Fig. 1). Based on this, it can be seen that floods have caused a lot of losses and damage in terms of the environment, social issues, and even the economic stability of communities.

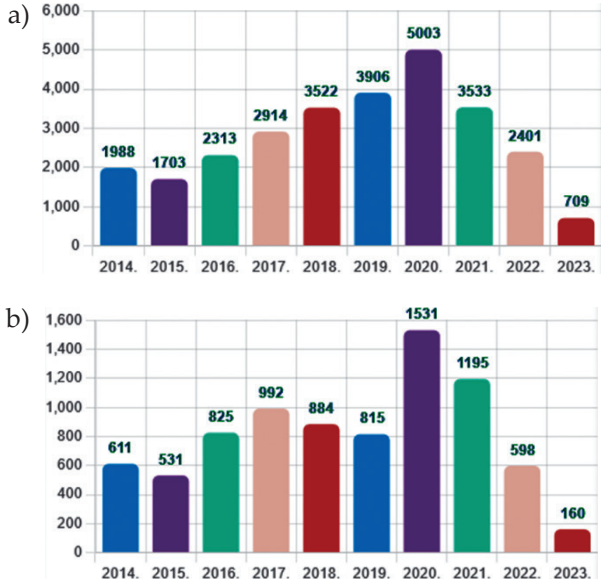

Fig. 1. Number of disasters (a) and flood events (b) in Indonesia from January 2014 to July 2023
Source: BNPB database

In addition, floods are the third-most-common type of disaster in Sumedang Regency (after landslides and forest/land fires). During the period of 2000–2021, there were 103 flood events in the Sumedang district. The data on the flood events for the period of 2000–2011 was obtained from the BNPB database, while for the 2012–2021 period, it was obtained from the West Java provincial government database (https://opendata.jabarprov.go.id/id/dataset/jumlah-kejadian-bencana-banjir-berdasarkan-kabupatenkota-di-jawa-barat). In this regency, the most floods occurred in 2010 (11 events), 2016 (13 events), 2019 (16 events), 2020 (21 events), and 2021 (14 events). Based on the report from the Health Crisis Center, Indonesian Ministry of Health (https://pusatkrisis.kemkes.go.id/), as many as five villages in the Jatinangor sub-district were inundated with floods as high as 70–250 cm on January 6, 2010. This flood occurred due to high rainfall intensity. Floods also occurred in the village of Sidamulya in the Sumedang Utara sub-district; these occurred as a result of the construction of the Cisumdawu toll road. This flood affected seven houses, and one part of the road collapsed.

Prediction, preparation, prevention, and damage assessment are the four steps to flood management [4]. Flood-susceptibility mapping is widely recognized as a necessary step in preventing and managing future flooding [2], as it can cover all four steps. However, flood susceptibility cannot be modeled by the basic and non-linear hydrological approaches because of the complexities of catchments [8, 10]. Therefore, traditional flood-modeling methods were unreliable regarding accurate predictions [11–13]; we need to use the most up-to-date geographic information system (GIS) techniques to process and analyze complicated planning strategies, decision-making, and integrated management [7]. In contrast to traditional flood-control theories, modern flood-risk management constantly attempts to use limited resources (social, environmental, and financial) [9].

According to [7], there are three types of flood-susceptibility assessments. There are hydrological models, statistical and data-driven approaches, and non-linear machine learning algorithms (such as support vector machine [SVM], decision trees [DT], and artificial neural network [ANN]). The main difficulty that is associated with the flood-susceptibility-assessment process is the multi-variable and non-linear relationship between the indices and the risk levels [5]. Traditional hydrological methods are not robust nor automated, as their model design, construction, and parameterization are time-consuming [7]. SVMs are complex mathematical functions that are difficult for humans to understand [5]. DT necessitates extensive pre-treatment and easily falls into local optimization [5, 14]. The ANN method shows over-learning and slow convergence speed problems [5, 15]. In addition, ANN cannot estimate the contribution of each variable to the model [5].

On the other hand, random forest (RF) has been widely used for flood-susceptibility research. This method has been used to map the risk of widespread flooding in Calcasieu Parish, Louisiana, U.S. [16], to develop a spatial prediction of flood susceptibility in the Seoul, South Korea, metropolitan area [17], for a flood-susceptibility analysis

in Gresik Regency, Indonesia [18], for flood mapping and identifying the essential conditioning factors at Fredericton, New Brunswick, Canada [19], for flood-hazard mapping at the Galikesh River basin in northern Iran [20], and for flood-hazard and flood-insurance claims in southeast Texas [21]. In addition, the RF technique has significant advantages when compared to the other more commonly used multivariate regression or classification approaches [22]. It is capable of accounting for the interactions and non-linearity between variables. Second, it allows for the mixed use of categorical and numerical variables without reverting to indicator (or dummy) variables. Third, it does not require assumptions on the distribution of explanatory variables. Moreover, RF can be used to rank the importance of variables using the mean decrease impurity (MDI) measure of significance [23].

The other methods that were considered for use in this study were statistical techniques, as numerous techniques of this type have been employed in earlier studies. Frequency ratio (FR) was used by [24] to map the flood vulnerability in the Kulik River basin in the Indo-Bangladesh Barind region. To map the flood susceptibility of the Kopai River basin in eastern India, [25] used FR, Shannon's entropy, and weight of evidence (WoE) techniques. FR, entropy index, and WoE were used by [26] to analyze the flood risk of the Raiganj subdivision in eastern India. [27] assessed the flood susceptibility of the Patna district in Central Bihar, India, using FR and Shannon's entropy. WoE and Shannon's entropy were also used by [28] to map the flood susceptibility of eastern India. According to these, the FR method was chosen as the comparison for the RF machine learning method. This technique is one of the most widely used bivariate statistical techniques that are applied to various natural hazard studies [8]. This model has the advantages of being easily implemented and producing completely understandable results [8]. In addition, the models produce good flood-risk maps, and the analysis process is simple to grasp [4]. Furthermore, some research has indicated that bivariate statistical models can occasionally be more accurate than machine learning models [29]. Therefore, the aims of this study are (1) to successfully analyze the flood-susceptibility distribution of the study area using RF and FR and (2) to compare the RF machine learning technique to the bivariate statistical FR method in flood-susceptibility assessment. The results can help planners for the Sumedang regional government in preparing a flood-management plan in an effort to manage regional flood risk. These findings can also be used to update the flood-susceptibility maps that are already in existence.

2. Study Area

The study area that was chosen as the case study was Sumedang Regency in West Java Province, Indonesia. According to Central Bureau Statistics (BPS), Sumedang had 26 sub-districts and 270 villages in 2021 with a total population of 1,152,507 people and a population density of 739 people/km^2. Sumedang is a hilly and mountainous region with elevations that range from 18 to 1996 meters above

sea level. Mountainous terrain dominates much of the Sumedang region, with some flat areas in the north. Based on these details, Sumedang was selected as the initial study location for the RF and FR algorithm-based flood-susceptibility assessment in Indonesia; there is a chance that we can continue this study in other areas as well. The study area can be seen in Figure 2.

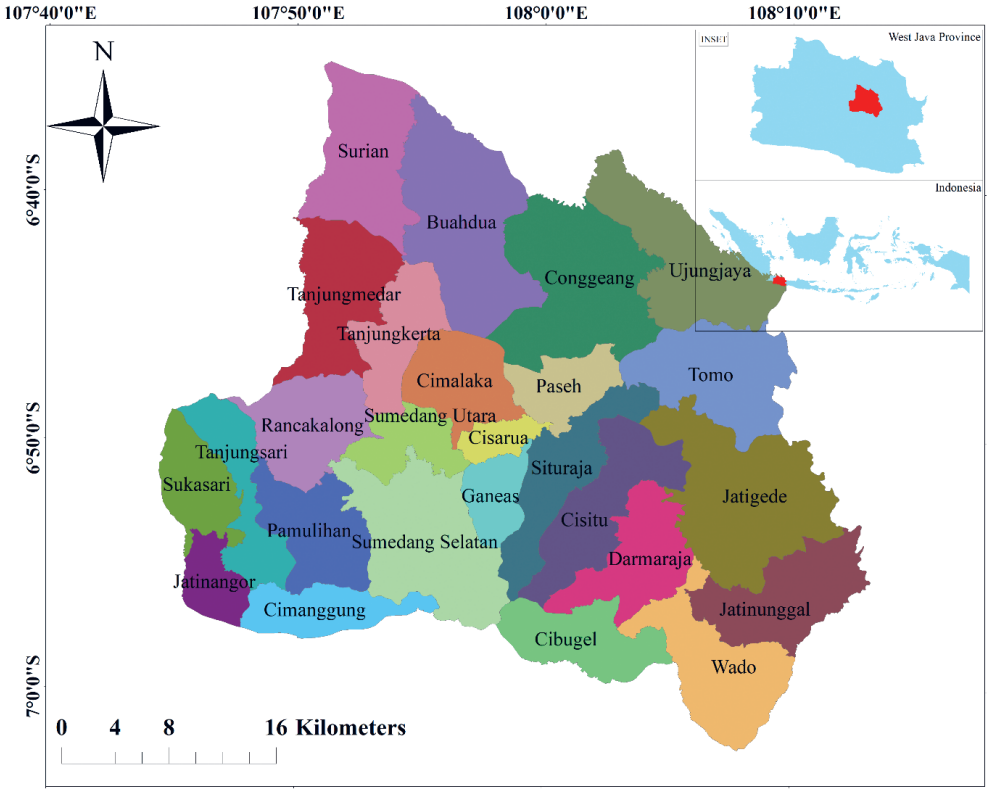

Fig. 2. Twenty-six sub-districts of Sumedang Regency (study site)

3. Materials and Methods

3.1. Flood-Inventory Map

The flood-inventory database is a key component of flood-vulnerability mapping [13]. The suggested methodological initial step is to make a flood-inventory map by collecting the data on previous flood events [7]. Due to changes in land use, historical flood data may provide limited spatial and temporal precision. Consequently, the results may be more ambiguous [7, 30]. A flood-inventory map was created by using historical data sets for specific events in Sumedang Regency. In this study,

a flood-inventory map was created using a collection of flood-event data archives from the BNPB database. The used data was the date of each incident and the coordinates of the floods in Sumedang Regency from 2000 to 2021. Based on the historical data, there are records of 53 coordinates of flood events; 35 coordinates of these events were used to develop the model, and 18 coordinates were used to validate the model. Then, the 53 non-flood coordinates were also meticulously created. Determinations of the non-flood points were made randomly outside the existing flood coordinates.

The accuracy of the data (training sample data) had an important impact on the creation of the flood models [19]; therefore, choosing the right training dataset was critical for ensuring the overall quality and effectiveness of the model [5]. In this study, an equal number of flooded and non-flooded points were generated in the RF classification method in order to avoid the issue of class imbalance [19]. The training and validation data from the flood inventory is shown in Figure 3. RF used both classes (flood and non-flood) to model the susceptibility, but FR merely needed the flood class to build the model.

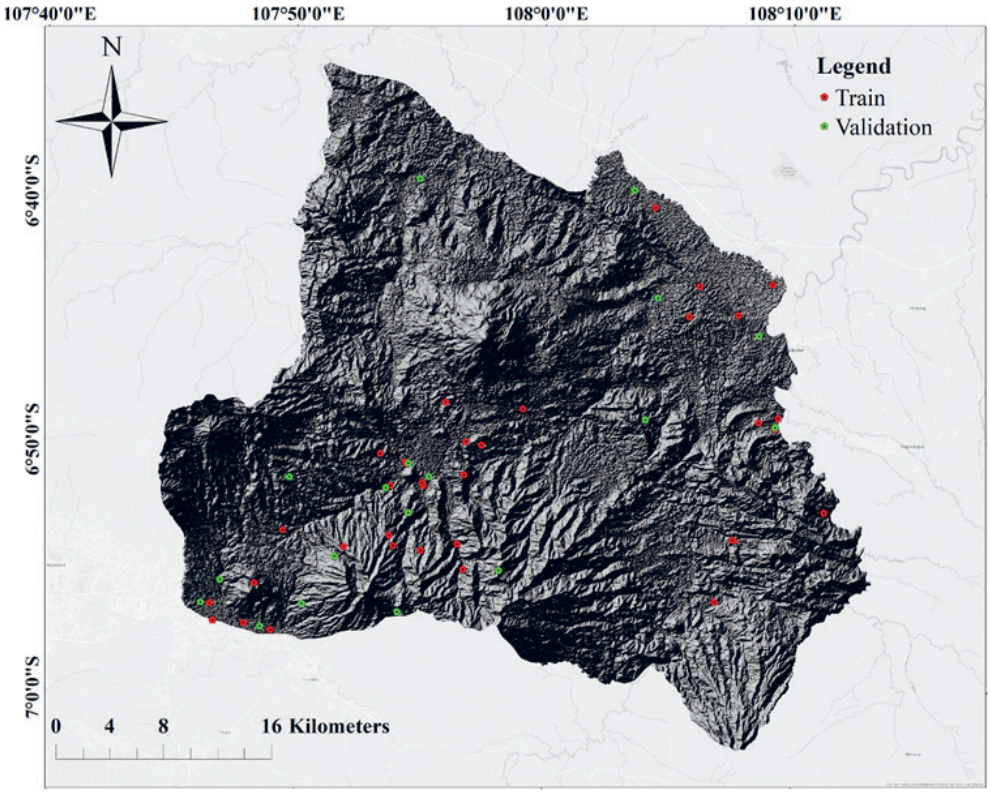

Fig. 3. Coordinates of flood events used as training (35 points) and validation (18 points) mapped on National DEM (DEMNAS) for Sumedang region, West Java

Source of DEMNAS: https://tanahair.indonesia.go.id/demnas/#/demnas

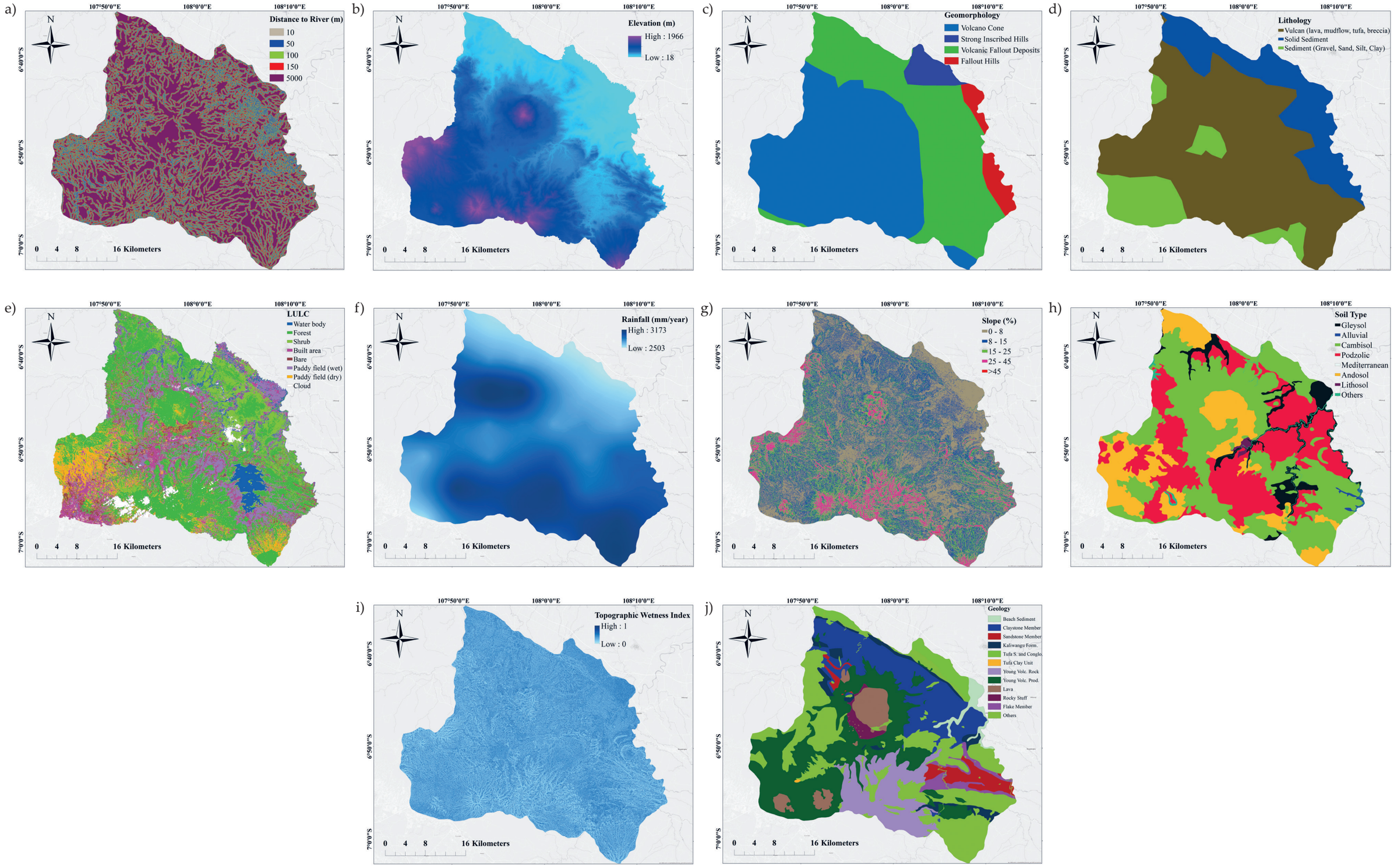

Fig. 4. Flood-conditioning factors: a) distance from rivers [m]; b) elevation [m]; c) geomorphology; d) lithology; e) land use/land cover; f) rainfall [mm/year]; g) slope [%]; h) soil type; i) TWI; j) geology

3.2. Flood-Conditioning Factors

Understanding and determining flood-conditioning factors were critical for this study when creating the flood model [6]. Although some factors were important when determining one flood-susceptibility area, they may be insignificant in another [6, 31]; therefore, the selection of conditioning factors depended on the study area and its characteristics [6, 19]. The conditioning factors of this study were the distance from a river, elevation, geology, geomorphology, lithology, land use/land cover, rainfall, slope, soil type, and TWI (Fig. 4 on the interleaf).

The distance from the rivers of the study area significantly impacts the speed and extent of floods [2, 32]. TWI measures topographic influences on hydrologic processes and is closely linked with groundwater depth and soil moisture [21, 33]. The TWI values are in the form of a 0–1 range (the values are normalized); a value of 0 indicates that the land does not accumulate water, while a value of 1 indicates that the land accumulates water. Land use/land cover directly or indirectly influence infiltration, evapotranspiration, and surface-runoff generation [13]. Surface runoff and water-flow intensity are controlled by the slope, which causes soil erosion and vertical percolation [29, 34]. Lithology is a key flooding conditioning parameter because it directly impacts land permeability and surface runoff [13, 35]. Moreover, the importance of soil data in predicting excess precipitation and infiltration is particularly important [7]. At the same time, geomorphology has the potential to provide deterministic methods for detecting flood risk or hazard [36]. Geology parameters influence flood susceptibility because of their sensitivity to lithological units [37]. Furthermore, elevation is the most critical component for flood-susceptibility mapping according to the sensitivity analysis [6]. Last, rainfall has a direct relationship with river discharge according to the literature, and a substantial amount of rain in a short period of time can cause flash floods in semi-arid locations [29].

Land use/land cover were derived from Landsat-8 image data using supervised classification with the support vector machine (SVM) algorithm at an 85% accuracy level. Data from the National DEM (DEMNAS, https://tanahair.indonesia.go.id/demnas/#/demnas) was used to make the slope (processing uses slope features) and TWI (uses formula calculations based on references). TWI uses DEMNAS image data to know the trend-mapping method of water accumulation on topographical control [38]. DEMNAS was built from several data sources, including IFSAR data (5-metre resolution), TERRASAR-X (5-metre resampling resolution from 5–10 m original resolution) and ALOS PALSAR (11.25-metre resolution) by adding the mass point data that is used in making topographical maps. The spatial resolution of DEMNAS was 0.27-arcseconds when using the EGM2008 vertical datum. This DEMNAS data was used as a reference by the authors because it was issued by the national authority. The DEMNAS data had a smaller root mean square error (2.79 m) as compared to the DTM data (3.24 m) and DSM (3.71 m), with bias errors of −0.13, −0.63, and 2.21 m for the DEMNAS, DTM, and DSM data, respectively.

Furthermore, rainfall was obtained by downloading the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS) satellite monthly average rainfall raster file for 30 years (1990–2020). The CHIRPS data is a re-analysis of rainfall data with a spatial resolution of 0.05° × 0.05° that has been available since 1981 (and introduced by [39]). The data will be processed by converting all of the bulk data from originally a monthly average to an annual one and interpolating using the inverse distance weight (IDW) method. All of the data sources are listed in Table 1. All of the data was then resampled to a spatial resolution of 30 m.

Table 1. Flood-conditioning factor (data sources)

No.	Factor	Format	Resolution [m]	Source
1	Geomorphology	polygon	–	Indonesian Ministry of Energy and Mineral Resources
2	Geology	polygon	–	Indonesian Ministry of Energy and Mineral Resources
3	Lithology	polygon	–	Indonesian Ministry of Energy and Mineral Resources
4	Soil type	polygon	–	Indonesian Ministry of Agriculture
5	Elevation	raster	8	DEMNAS
6	Slope	raster	8	DEMNAS
7	Rainfall	raster	5,500	CHIRPS
8	TWI	raster	8	DEMNAS
9	Land use/land cover	raster	30	Landsat-8 image
10	Distance from river	line vector	–	DEMNAS

3.3. Random Forest

Random forest (RF) classification is a machine learning algorithm for non-parametric multivariate classification that was first developed by Leo Breiman in 2001 [22]. RF uses a non-parametric estimating framework in which an input random vector $\vec{X} \in \mathcal{X} \subset \mathbb{R}^p$ is observed [23]. The purpose is to estimate function $m(\vec{x}) = \mathbb{E}[Y \mid \vec{X} = \vec{x}]$ in order to anticipate random response $Y = \{0, 1\}$. With this goal in mind, we assume we are given a training sample $\mathcal{D}_n = \left((\vec{X}_1, Y_1), \ldots, (\vec{X}_n, Y_n)\right)$ of independent random variables distributed as independent prototype pair $(\vec{X}, Y)$. The purpose is to create an estimate $m_n : \mathcal{X} \to \mathbb{R}$ of function m using data set.

RF is a predictor that is composed of M randomized regression trees [23]. The predicted value at query point $\vec{x}$ is represented as $m_n(\vec{x}; \Theta_j, \mathcal{D}_n)$ for the j-th tree in the family, where $\Theta_1, \ldots, \Theta_M$ are independent random variables.

If a leaf represents region *A*, the randomized tree classifier takes the following form:

$$m_n(\vec{x};\Theta_j,\mathcal{D}_n)=\begin{cases}1, \text{ if } \sum\limits_{i:\vec{x}_i\in D_n^*(\Theta_j)} 1_{\vec{x}_i\in A,Y_i=1} > 1_{\vec{x}_i\in A,Y_i=0},\ \vec{x}\in A\\ 0, \text{ otherwise}\end{cases} \tag{1}$$

where $\mathcal{D}_n^*(\Theta_j)$ denotes the data points that are chosen during the resampling process. The RF classifier is a result of a majority vote among classification trees; that is:

$$m_{M,n}(\vec{x};\Theta_1,\ldots,\Theta_M,\mathcal{D}_n)=\begin{cases}1, \text{ if } \frac{1}{M}\sum\limits_{j=1}^{M} m_n(\vec{x};\Theta_j,\mathcal{D}_n) > \frac{1}{2}\\ 0, \text{ otherwise}\end{cases} \tag{2}$$

Consider a single tree that has not been subsampled. Let $N_n(A)$ be the total number of data points that fall into *A*. A cut in *A* is pair (*j*, *z*), where *j* is some value (dimension) from {1, ... , *p*}, and the position of the cut along the *j*-th coordinate within the limits of *A*. Let C_A denote the collection of all possible cuts in *A*. Let $p_{0,n}(A)$ (resp., $p_{1,n}(A)$) be the empirical probability given a data point in cell *A* that has label 0 (reps., label 1). By noticing that $\overline{Y}_A = p_{1,n}(A) = 1 - p_{0,n}(A)$, the classification and regression trees (CART) split the criterion reads for any $(j,z)\in C_A$:

$$L_{class,n}(j,z)=p_{0,n}(A)p_{1,n}(A)-\frac{N_n(A_L)}{N_n(A)}\cdot p_{0,n}(A_L)p_{1,n}(A_L)-\frac{N_n(A_R)}{N_n(A)}\cdot p_{0,n}(A_R)p_{1,n}(A_R) \tag{3}$$

This criterion is based on the so-called *Gini impurity measure* [23]. For each cell *A*, the best cut (j_n^*, z_n^*) is selected by maximizing $L_{class,n}(j, z)$ over a subset $M_{try}\subset\{1,\ldots,p\}$ and C_A; that is:

$$(j_n^*,z_n^*)\in \arg\max_{\substack{j\in M_{try}\\(j,z)\in C_A}} L_{class,n}(j,z).$$

RF can be utilized to determine the significance of parameters in regression or classification problems using the mean decrease impurity (MDI) measure of significance [23]. MDI is calculated by averaging the total decrease in the node impurity that is caused by variable splitting across all trees. Set $\vec{X}=(X^{(1)},\ldots,X^{(p)})$, for a forest that is formed by the joining of *M* tresses, the MDI of the variable $X^{(j)}$ is defined by the following:

$$\widehat{MDI}(X^{(j)})=\frac{1}{M}\sum_{l=1}^{M}\sum_{\substack{t\in T_l\\ j_{n,t}^*=j}} p_{n,t}L_{class,n}(j_{n,t}^*,z_{n,t}^*) \tag{4}$$

where $p_{n,t}$ is the fraction of observations that fall within node *t*, $\{T_l\}_{1\le l\le M}$ is the collection of trees in the forest, and $(j_{n,t}^*, z_{n,t}^*)$ is the split that maximizes $L_{class,n}(j, z)$ in node *t*.

3.4. Frequency Ratio

A frequency ratio (FR) is a bivariate statistical analysis method that uses a spatial distribution-dependent (probability-dependent) factor (flood location) as well as flood-triggering and causal factors [29]. The bivariate probability of each independent flood-triggering factor was determined by its relationship with flood occurrence [29]. The higher the bivariate probability (greater than 1), the stronger the correlation between the flood incidence and the flood-triggering factors; the lower the probability (less than 1), the weaker the correlation [29].

Let L and F stand for floods and a specific flood-related factor, respectively. The frequency ratio for the i-th type of factor $F(F_i)$ can be stated as follows [13]:

$$FR_i = \frac{PL_i}{PF_i} = \frac{\dfrac{area\ of\ floods\ in\ F_i\ region}{area\ of\ F_i\ region}}{\dfrac{area\ of\ floods\ in\ study\ region}{area\ of\ study\ region}} = \frac{p(L \mid F_i)}{p(L)} \tag{5}$$

Because "the probability of floods in the study region" $p(L)$ is predefined based on the flood and factor data, frequency ratio FR_i is entirely determined by "the probability of floods in the F_i area" $p(L \mid F_i)$, which is actually "the conditional probability of L given F_i" [37]. A larger conditional probability $p(L \mid F_i)$ means that the occurrence probability of floods is larger in the i-th type or the i-th class of factor $F(F_i)$.

Consider an arbitrary flood-related factor $F^{(j)}$ ($j = 1, 2, 3, \ldots, m$); the frequency ratios for the different classes, $FR_i^{(j)}$ ($i = 1, 2, 3, \ldots, n$; $j = 1, 2, 3, \ldots, m$), can be computed by using Equation (5). The frequency ratio of this factor at this location $FR^{(j)}$ will be $FR_i^{(j)}$ if the class $F^{(j)}$ of at this location is $F_i^{(j)}$. At this location, the flood-susceptibility index (FSI) will then be the sum of the frequency ratios of various flood-related factors [37]:

$$FSI = \sum_{j=1}^{m} FR^{(j)} \tag{6}$$

3.5. Flood-Susceptibility Assessment based on RF

A flowchart of flood-susceptibility assessment based on RF was created (shown in Figure 5). The work starts with collecting the flood-inventory data and the factors that caused the flood events. The 53 flood coordinates were found based on the historical data. For the RF training sample, 53 non-flood coordinates were also generated and made randomly outside the existing flood coordinates. The precision of the data that was used to generate the flood model has a significant impact on its accuracy [19]. Selecting an appropriate training data set is critical for ensuring the overall quality of the model and efficacy [5]. The BNPB database was used to acquire several sample points. An equal number of flooded and non-flooded points were generated to avoid the issue of class imbalance [19].

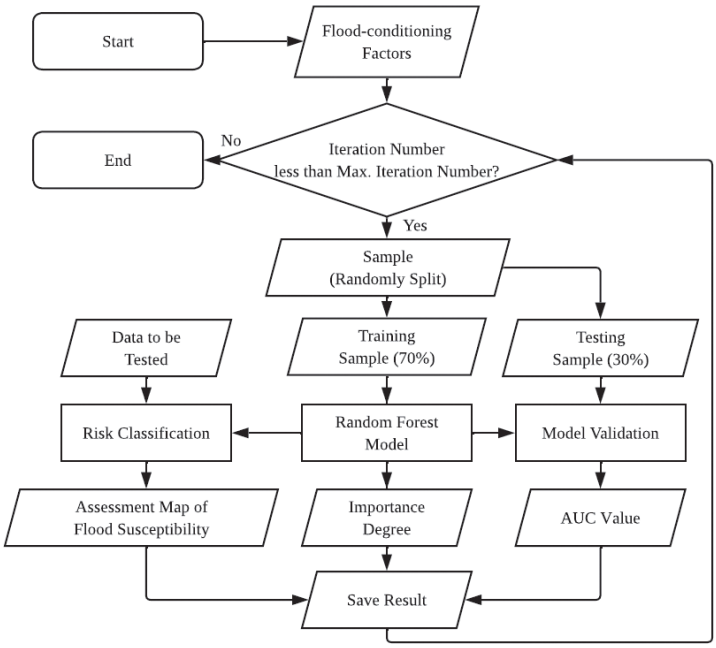

Fig. 5. Flowchart of flood-susceptibility assessment using random forest

From the flood and non-flood coordinates, the training sample was generated by taking all of the digital numbers on the flood-conditioning factor. The data set was randomly split into training and testing data – both of which were utilized to develop the RF model and test its accuracy [5]. The training sample was randomly separated as training data and test data at a ratio of 70:30. Using the training data and the RF algorithm, the RF model was earned. From the model, the importance degree was generated. The assessment map of the flood susceptibility was created with all of the digital numbers of the flood-conditioning factor and the RF model. Next, the validation was conducted using the test data and flood susceptibility.

The RF method was implemented using the open-source *scikit-learn* python package (https://scikit-learn.org). The model hyper-parameters of the RF algorithm (which regulate the structure of the forest and the level of the randomization) needed to be defined before it could be used [40]. Therefore, the two different scenarios were conducted based on this. The first scenario was to vary the number of trees from 100, 200, through 1000, with all of the other parameters set by default. The second scenario was the ten-times cross-validation; this was achieved by varying the training and test data by randomly separating each running. These two scenarios evaluated the performance of the RF model. The area under the receiving operating characteristic curve (AUC) was used to evaluate the performance of the model. According to [26], the AUC value was classified into five classes: 0.9–1.0 (excellent); 0.8–0.9 (very good); 0.7–0.8 (good); 0.6–0.7 (average); and 0.5–0.6 (poor).

3.6. Flood-Susceptibility Assessment based on FR

To compare the performance of the RF method, the bivariate FR method is explained in Figure 6. The used data (training and test data) was the same data that was used in the RF method. In the FR method, the used data was only the food data without the non-food data. Since the used data was absolutely the same as the data that was used in RF, the performance can be compared. Perceiving the feature-ÐÔportance value, the normalized prediction rate was used. Normalization helped ÙÌmove the efect of variation in the scale of the data set; i.e., a data set with large values can be easily compared to a data set with smaller values. Besides, AUC was also derived from the FR of the food susceptibility using the test data.

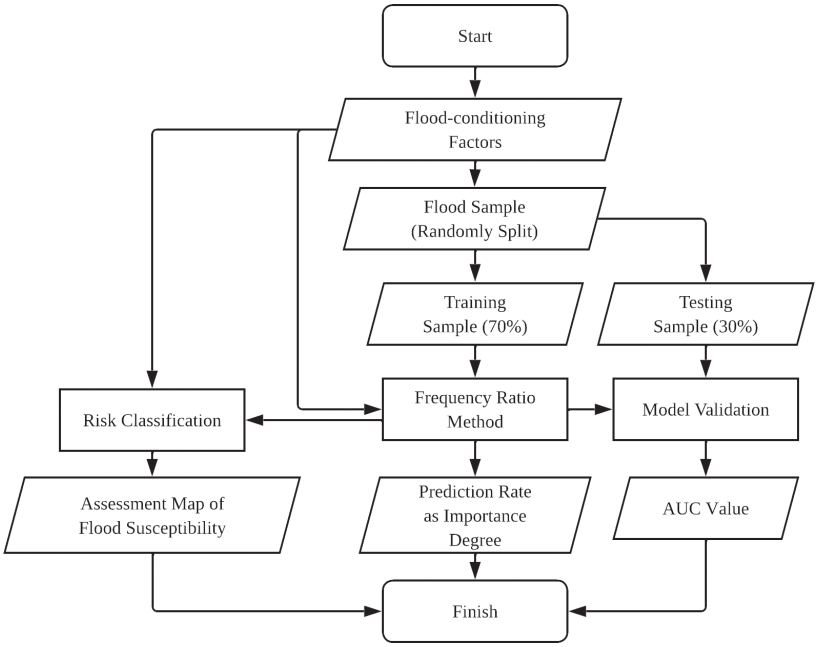

Fig. 6. Flowchart of food-susceptibility assessment using frequency ratio

4. Results

4.1. Random Forest

Two scenarios were carried out to see the RF performance. The AUC values that were based on the results by varying the number of trees are shown in Figure 7. The AUC values that were obtained from this scenario were divided into two values: 82.0, and 83.0%. The results of several numbers of these trees were very close, and no signifcant diference could be seen. The AUC value of 83.0% was achieved

by the following numbers of trees: 200, 300, 400, 500, 700, and 900. The number of trees of 200 was chosen (for the ten-times cross-validation scenario) because it was the smallest number of trees but had the highest AUC value.

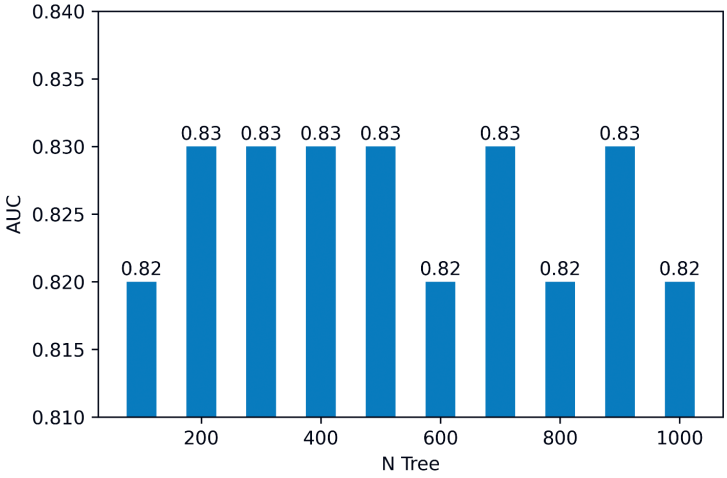

Fig. 7. AUC values based on N-tree sensitivity

The following scenario was ten-times cross-validation with the number of trees set to 200 while the other hyper-parameters were held constant. From this process, the performance of the model could be observed. The receiver operating characteristic (ROC) curves and AUC for each run are shown in Figure 8.

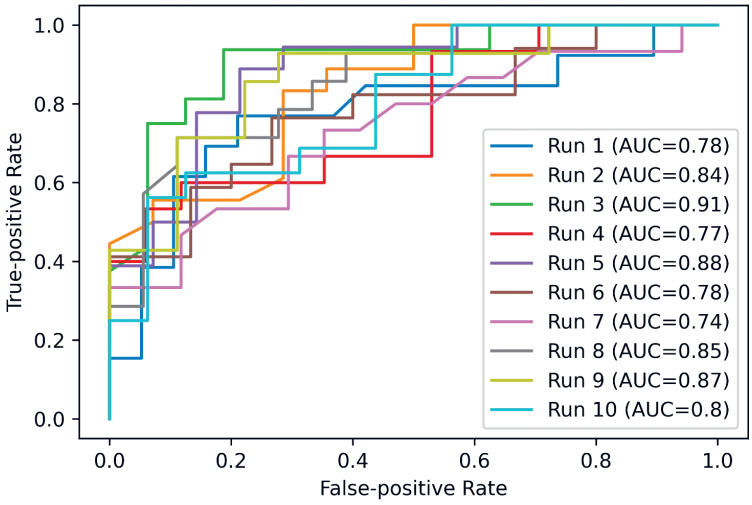

Fig. 8. ROC curves and AUC values based on ten-times cross-validation

The AUC values varied quite a bit on each run; these were different from the previous scenario. The smallest AUC was 74.0% and occurred on the seventh run, while the highest AUC was 94.0% and was on the third run. The average AUC value of all of these cross-processes was 83.0%.

In addition to the AUC value, feature-importance values were also derived from the RF model. The importance value of this feature determined which feature of the existing data was most helpful in the RF model. Table 2 shows the average feature-importance values from ten cross-validation runs. The most important parameter was geology (with an average value of 19.42%), followed by land use/land cover and soil type (with average appearance importance values of 15.84 and 10.28%, respectively); the lowest significant values were geomorphology, TWI, and lithology (with averages of 4.62, 7.24, and 7.78%, respectively).

Table 2. Random forest importance values

Feature	Importance value [%]	Rank
Geology	19.42	1
Land use/land cover	15.84	2
Soil type	10.28	3
Rainfall	9.61	4
Elevation	8.53	5
Distance from river	8.47	6
Slope	8.21	7
Lithology	7.78	8
TWI	7.24	9
Geomorphology	4.62	10

After applying the RF algorithm to various combinations of the training and validation data, each implementation produced a probability map with values that ranged from 0 to 1. Each value of the pixel showed the likelihood of that pixel becoming flooded or not. The flood-probability map is shown in Figure 9. Using Jenks natural breaks categorization approach [19, 41], the flood-susceptibility map was derived from the probability map (shown in Figure 10) with five susceptibility classes; namely, very low, low, moderate, high, and very high.

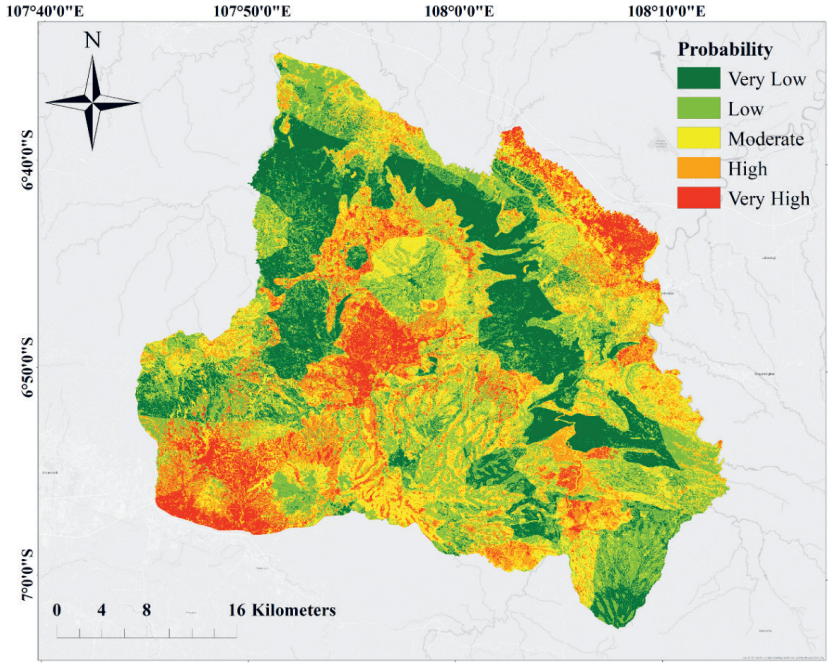

Fig. 9. Random forest method probability

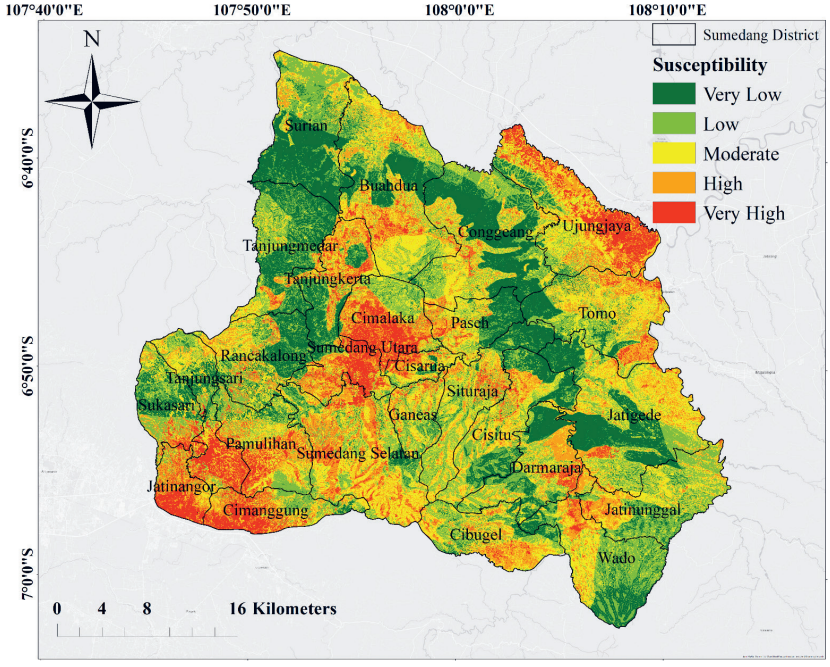

Fig. 10. Random forest method susceptibility

4.2. Frequency Ratio

Bivariate statistical RF was used to compare the performance of the RF method. The separated training and test data that was used in this model was the same as that which was used in second run from the cross-validation scenario in the RF model. The data that was used in the FR model was only the flooded inventory data.

Table 3. Frequency ratio summary

Feature	Number of pixels – each class	Total pixels – area	Number of flood pixel	Frequency ratio (FR)
Distance from river [m]				
0–10	1,162,164	22,886,661	4	2.25
10–50	4,488,188		7	1.02
50–100	4,866,632		10	1.34
100–150	3,753,107		6	1.05
>150	8,616,570		8	0.61
Elevation				
Lowland	779,252	22,886,670	3	2.52
Lowland hills	2,227,794		5	1.47
Low hills	2,574,683		0	0.00
Hills	7,197,657		12	1.09
High hills	10,107,284		15	0.97
Geomorphology				
Volcano cone	12,056,378	22,887,214	19	1.03
Strong inscribed hills	807,846		1	0.81
Volcanic fallout deposits	9,077,103		13	0.94
Fault hills	945,887		2	1.38
Lithology				
Lava, volcanic mudflow, tuffa, brecci	15,025,648	22,887,214	16	0.70
solid sediment	4,704,503		9	1.25
Semi-solid sediment (gravel, sand, silt, clay)	3,157,063		10	2.07

Table 3. cont.

Land use/land cover				
Waterbody	1,109,199	22,612,717	3	1.75
Forest	8,207,183		7	0.55
Shrubs	4,750,051		6	0.82
Building area	2,191,050		11	3.24
Open-space area	1,278,003		3	1.52
Wet agriculture	2,842,272		5	1.14
Dry agriculture	1,687,997		0	0.00
Cloud	546,962		0	0.00
Rainfall [mm/year]				
2503–2668	2,911,214	22,886,661	8	1.80
2668–2776	3,704,360		1	0.18
2776–2873	5,420,916		13	1.57
2873–2973	5,217,407		6	0.75
2973–3173	5,632,764		7	0.81
Slope [%]				
0–8	7,986,698	22,887,110	21	1.72
8–15	7,214,156		5	0.45
15–25	5,239,389		8	1.00
25–45	2,428,127		1	0.27
>40	18,740		0	0.00
Soil type				
Gleysol	1,391,467	22,887,214	3	1.41
Alluvial	65,439		0	0.00
Cambisol	9,742,493		18	1.21
Podzolic	7,140,751		10	0.92
Mediterranean	6,184		0	0.00
Andosol	4,229,586		4	0.62
Lithosol	80,565		0	0.00
Others	230,729		0	0.00

Table 3. cont.

Feature	Number of pixels – each class	Total pixels – area	Number of flood pixel	Frequency ratio (FR)
TWI				
1.14–4.59	4,669,778	22,601,689	6	0.83
4.59–5.74	9,056,324		13	0.93
5.74–7.16	6,132,616		12	1.26
7.16–9.32	2,057,843		3	0.94
9.32–18.38	685,128		1	0.94
Geology				
Coastal sediment	409,057	22,887,214	3	4.80
Claystone	3,744,312		1	0.17
Sandstone	851,190		1	0.77
Kaliwangu formation	754,499		2	1.73
Tufa sandstone and conglomerate	729,714		2	1.79
Tufaan clay unit	8,741		1	74.81
Young volcanic rock	2,530,552		3	0.78
Young volcanic product	6,006,891		18	1.96
Lava	1,078,535		1	0.61
Rocky tuff	275,507		1	2.37
Flake	533,315		2	2.45
Others	5,964,901		0	0.00

Using the digital numbers from all of the conditioning factors that were extracted in the training data coordinates as well as Equation (5), the FR model was created; the summary is presented in Table 3. From this model, the flood-probability map can be derived using Equation (6); the result is shown in Figure 11. Using Jenks natural breaks categorization approach [19, 42], the flood-susceptibility map was derived from the probability map (shown in Figure 12) with the same five susceptibility classes as were in the RF model.

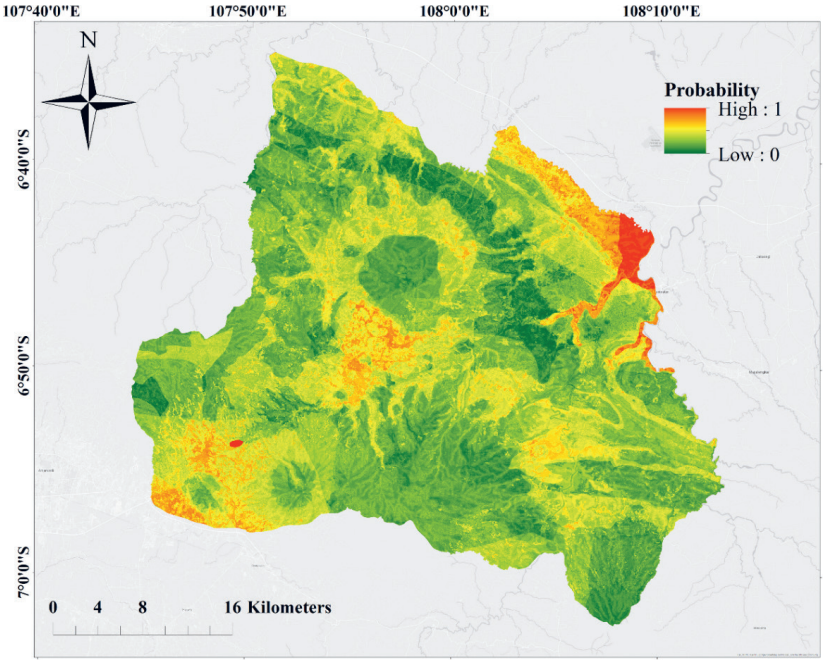

Fig. 11. Frequency ratio probability

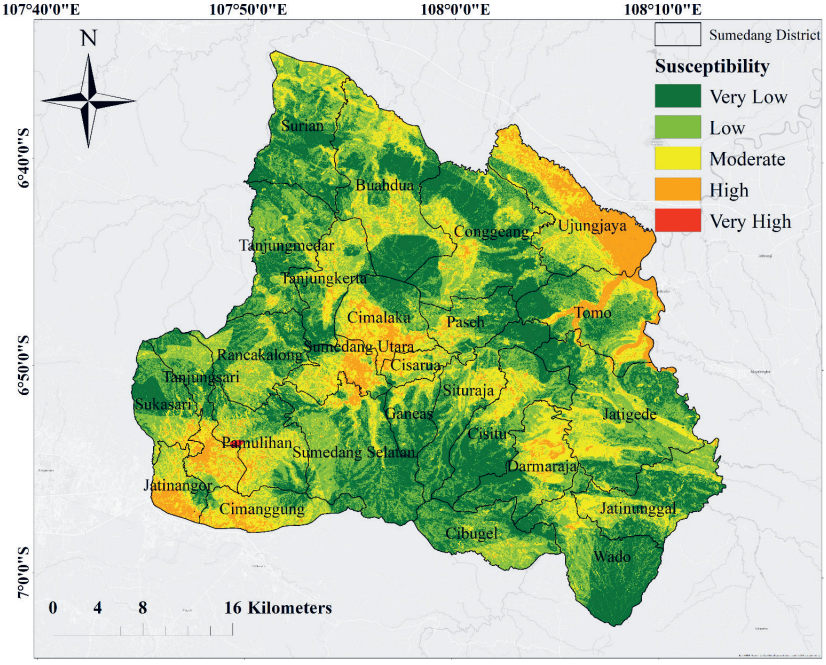

Fig. 12. Frequency ratio susceptibility

The validation data (Fig. 3) was used to measure the performance of the model. The ROC values are shown in Figure 13 (with an AUC value of 83.0%). In addition, the feature-importance values were also derived from the FR model using the normalized prediction rates. Table 4 shows the feature-importance values from the FR model. The most important parameter was the geology (just the same as in the RF model), with a feature-importance value of 19.11%, followed by lithology and slope (with values of 12.15 and 11.78%, respectively). The lowest in importance were the TWI, geomorphology, and slope type, with importance values of 6.07, 7.83, and 8.00%, respectively. Like in the RF model, geomorphology was always one of the lowest factors.

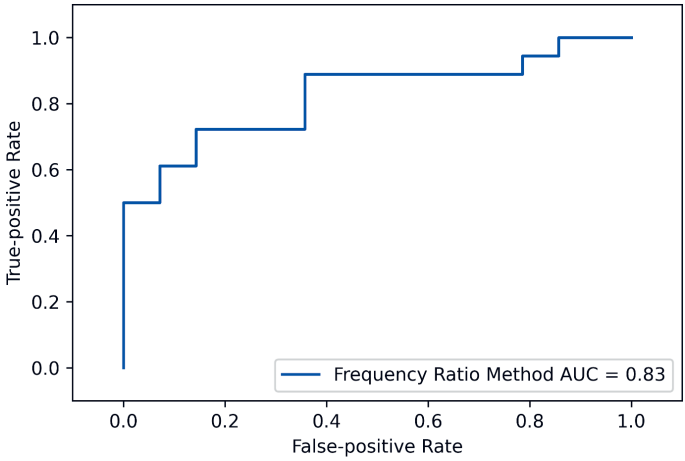

Fig. 13. ROC of frequency ratio result

Table 4. Frequency ratio importance values

Feature	Importance value [%]	Rank
Geology	19.11	1
Lithology	12.15	2
Slope	11.78	3
Elevation	9.81	4
Land use/land cover	8.48	5
Distance from river	8.46	6
Rainfall	8.29	7
Soil type	8.00	8
Geomorphology	7.83	9
TWI	6.07	10

5. Discussions

5.1. Random Forest

In this study, the flooded areas (susceptible zones) were defined as the high- and very-high-susceptibility classes. After the total area of each class was calculated, the regions with the highest flood-susceptibility zones were the Ujungjaya (4423.14 ha), Sumedang Selatan (4072.05 ha), and Cimalaka (2775.6 ha) sub-districts (shown in Figure 10). It can be seen in Figure 10 that these regions have a lot of red and orange pixels. This makes sense, as these regions have a lot of flood-event history (as is shown in Figure 3). Likewise, the Tanjungmedar, Surian, and Ganeas sub-districts were the regions with the lowest flood-susceptibility zones, as they have almost no flood-event histories (as can be seen in Figure 10). Overall, the map (Fig. 10) contains five probability map classes: very low (21.12%), low (25.15%), moderate (25.34%), high 18.44%, and very high (9.95%). Therefore, the flooded susceptibility zone from the RF model was 28.39%, and the remaining 71.61% consisted of non-flooded zones. Very high flood-susceptibility zones are located in the Ujungjaya, Cimalaka, Cimanggung, and Sumedang Utara sub-districts, while high flood-susceptibility zones are in the Sumedang Selatan, Buahdua, and Ujungjaya sub-districts.

In addition, the highest flood-susceptibility-zone regions based on the percentage (the ratio of the susceptible area to the sub-district area) were Jatinangor (81.05%), Sumedang Utara (67.61%), and Cimanggung (58.15%). It can be seen in Figure 10 that each of these areas is dominated by red and orange pixels; despite the fact that these regions are small, they have a lot of flood-event history (Fig. 3). Meanwhile, the regions with the lowest percentages were the Surian, Tanjungmedar, and Conggeang sub-districts.

Based on total area and percentage, it can be generally concluded that Ujungjaya and Cimalaka are the most-susceptible zones in Sumedang, while Surian and Tanjungmedar are the least-susceptible zones. This is acceptable according to the fact that the average AUC for RF was 83.0% (acc. to [26], this value can be considered to be very good), with the highest AUC being 94.0%. This value is said to be relatively good when compared to the works of [21] (with an AUC value of 89.5%) and [18] (with an AUC value of 97.0%); it is better than the work of [17] (with an AUC value of 79.2%).

5.2. Frequency Ratio

As shown in Table 3, the highest FR value was reached by the tuff clay unit class, followed by the coastal sediment class for the geology factor. This was because this region was small but had a flood-event history. Besides, the lowland and lowland hills classes from the elevation factor had rather high FR values. It makes sense that the low areas have higher susceptibility than the higher areas. The same is true for slopes of less than 8%; this region had an FR value that was greater than 1, while the other areas (with slopes that were greater than 8%) had the same or less than 1 (showing

that these areas are not susceptible to floods). This was also acceptable, since flat areas can hold more water than sloping areas. Moreover, the building area class on the LULC factor looked to have a higher FR value that the others; this was because the building areas had a small catchment area. Furthermore, the highest FR for the rainfall factor was generated by the smallest rainfall class. Despite this fact, this was in line with the fact that the rainfall factor was less important than the others (Table 4).

Similarly to the RF result, the susceptible zone from FR was defined as the high- and very-high-probability map classes. Based on the total area, the region with the highest susceptible zone was Ujungjaya (3693.60 ha), followed by the Tomo (1344.87 ha) and Cimalaka (1077.75 ha) sub-districts (as can be seen in Figure 12). It can also be seen in Figure 12 that these regions are dominated by orange pixels. This makes sense, as these regions have had a lot of historical flood events (according to Figure 3). Since they have had almost no flood history, the Wado, Tanjungmedar, and Cibugel sub-districts were the regions with the lowest numbers of susceptible zones (this can also be seen in Figure 12). Overall, the map (Fig. 12) contains five probability map classes: very low (31.63%), low (36.88%), moderate (23.47%), high (7.98%), and very high (0.04%). A very high susceptibility is associated with Pamulihan, and a high susceptibility is mainly associated with the Ujungjaya, Tomo, and Cimalaka sub-districts. Therefore, the flood-susceptibility zones from the FR model accounted for 8.02% of the zones (located in the Pamulihan, Ujung Jaya, Tomo, and Cimalaka sub-districts), while the remaining 91.98% were non-flooded zones.

In addition, the areas with the highest percentages of susceptible zones were Ujungjaya (43.61%), Jatinangor (35.51%), and Sumedang Utara (25.86%). It can be seen from Figure 12 that each of these areas is dominated by orange pixels; this is because a history of flood events has occurred in these areas even though they are not very large (Fig. 3). Meanwhile, the areas with the lowest flood-area percentages were Surian, Tanjungmedar, and Conggeang.

Based on these two categories (flood area and flood-area percentage), areas such as Ujungjaya, Cimalaka, and North Sumedang can be generally said to be the most-prone areas to flooding in Sumedang (because these areas were included in the top five based on both categories), while Cibugel, Wado, and Tanjungmedar are the safest areas from flooding. These results are acceptable because the AUC value for the FR model is 83%. This value can be said to be very good according to [26], although it is lower than [24] (with an AUC of 90.1%), [25] (with an AUC of 96.5%), and [26] (with an AUC of 91.1%).

5.3. Comparison

Based on RF and FR, the flood-susceptible areas in Sumedang were relatively the same; namely, Ujungjaya and Cimalaka. For those areas that are not prone to flooding, RF and FR concluded that Tanjungmedar was not a flood-susceptible area; however, the results of the susceptible areas based on the two methods showed relatively different results. The flooded susceptibility zones from the RF model

accounted for 28.39%, and the remaining 71.61% were non-flooded zones, whereas the flood-susceptibility zones for FR accounted for 8.02%, and the remaining 91.98% were non-flooded zones. It can be seen that the susceptibility zone was smaller based on FR than the results from RF – especially for very high susceptibility areas. This happened because, in the FR method, the FR value was dominated by one class; namely, tuff clay units (whose value was quite high). Thus, the distribution of the pixel values collected on the left (lower) and caused the Jenks natural breaks to categorize fewer classes for the very high class.

In both models, the results of the importance value stated that geological parameters are essential parameters for flood-risk assessment (Fig. 4j). This is similar to [43, 44], where geology was determined to be one of the significant flood-determining events. If the training data is overlaid with the geology data in Figure 4j, it can be seen that not all classes in geology have flood events. In those classes where there was flooding, the distribution of the flood points was not evenly distributed in each class (Table 3, 'Number of flood pixel' column, 'Geology' row). Thus, the class differences in geology can determine the differences in the levels of flood susceptibility quite well. This is what caused geology to be the most important parameter.

6. Conclusions

The study of flood-hazard risk in the Sumedang area was conducted using the machine learning RF algorithm as compared to the bivariate statistical FR method. This study chose ten flood-conditioning factors: distance from a river, elevation, geology, geomorphology, lithology, land use/land cover, rainfall, slope, soil type, and TWI. There were two scenarios: the first was the number of trees sensitivity, and the second was the training and test data (cross-validation) sensitivity. Based on the first scenario, the number of trees that were employed throughout the cross-validation process was 200, and the other hyper-parameters were fixed to default. By using the *feature-importance* method on *scikit-learn*, the significant parameters were derived. From the important average values, the ranking of the essential parameters was as follows: geology, land use/land cover, elevation, rainfall, soil type, distance from a river, TWI, slope, geomorphology, and lithology (in order from highest to lowest). Therefore, geology was the highest, and lithology was the most subordinate significance to the RF model. Using the FR method, the important parameters ranked from the highest to the lowest were as follows: geology, lithology, slope, elevation, land use/land cover, distance from a river, rainfall, soil type, geomorphology, and TWI. Here, TWI scored the lowest significance, but the highest was the same as with the RF method (geology).

Furthermore, the flood-susceptibility map was created using the model and all of the digital numbers from conditioning factors. Using the Jenks natural break classification method, a flood-susceptibility map was derived from that map, including five classes: very low, low, moderate, high, and extremely high. Flooded areas were

defined as having high and very high probabilities. The flooded areas from the RF model were 28.39%, and the remaining 71.61% were non-flooded areas. The susceptibility zones from RF were mainly from the Ujungjaya, Sumedang Selatan, Cimalaka, Buahdua, and Cimanggung sub-districts. The flooded areas from the FR method were 8.02%, with the most being located in the Ujungjaya, Tomo, Cimalaka, Jatinangor and Sumedang Utara sub-districts; the non-flooded areas accounted for 91.98%. The RF model resulted in more flooded areas than the FR model; this happened because the FR value was dominated by one class in the FR method (namely, tuff clay units), whose value was quite high. Thus, the distribution of the pixel values collected on the left (lower) and caused the Jenks natural breaks to categorize fewer classes for the very high class.

The area under the receiver operating characteristic curve (AUC) was used to assess the model. The average AUC from the ten-times cross-validation RF model was 83.0%, with the best AUC being 94.0% and the worst – 74.0%. At the same time, the AUC average from the FR method was also 83.0% (just the same as the average AUC value from RF). According to [26], both results had the 'very good' title; however, these results can still be developed when compared with the previous study. Overall, the flood-susceptibility examination in the Sumedang area using the RF algorithm compared to the bivariate statistical FR both yielded relatively good results.

Author Contributions

The distribution of authors contributions to this project is as follows:

- Rido Dwi Ismanto: 32%.
- Hana Listi Fitriana: 20%.
- Johannes Manalu: 10%.
- Alvian Aji Purboyo: 10%.
- Indah Prasasti: 28%.

Declaration of Competing Interests

We all certify that none of us have any competing interests with this paper.

Acknowledgments

The authors are thankful to all of the parties who have provided data for no charge and would like to thank the National Research and Innovation Agency (BRIN) for supporting this study. The authors also want to express their appreciation to the editor and the anonymous reviewers for their insightful criticisms and recommendations, which all helped to make the manuscript better.

References

[1] Youssef A.M., Pradhan B., Hassan A.M.: *Flash flood risk estimation along the St. Katherine road, southern Sinai, Egypt using GIS based morphometry and satellite imagery*. Environmental Earth Sciences, vol. 62(3), 2011, pp. 611–623. https://doi.org/10.1007/s12665-010-0551-1.

[2] Tehrany M.S., Pradhan B., Mansor S., Ahmad N.: *Flood susceptibility assessment using GIS-based support vector machine model with different kernel types*. Catena, vol. 125, 2015, pp. 91–101. https://doi.org/10.1016/j.catena.2014.10.017.

[3] Sutradhar S., Mondal P.: *Prioritization of watersheds based on morphometric assessment in relation to flood management: A case study of Ajay river basin, Eastern India*. Watershed Ecology and the Environment, vol. 5, 2023, pp. 1–11. https://doi.org/10.1016/j.wsee.2022.11.011.

[4] Tehrany M.S., Pradhan B., Jebur M.N.: *Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS*. Journal of Hydrology, vol. 504, 2013, pp. 69–79. https://doi.org/10.1016/j.jhydrol.2013.09.034.

[5] Wang Z., Lai C., Chen X., Yang B., Zhao S., Bai X.: *Flood hazard risk assessment model based on random forest*. Journal of Hydrology, vol. 527, 2015, pp. 1130–1141. https://doi.org/10.1016/j.jhydrol.2015.06.008.

[6] Kia M.B., Pirasteh S., Pradhan B., Mahmud A.R., Sulaiman W.N.A., Moradi A.: *An artificial neural network model for flood simulation using GIS: Johor River Basin, Malaysia*. Environmental Earth Sciences, vol. 67(1), 2012, pp. 251–264. https://doi.org/10.1007/s12665-011-1504-z.

[7] Hong H., Panahi M., Shirzadi A., Ma T., Liu J., Zhu A.X., Chen W., Kougias I., Kazakis N.: *Flood susceptibility assessment in Hengfeng area coupling adaptive neuro-fuzzy inference system with genetic algorithm and differential evolution*. Science of the Total Environment, vol. 621, 2018, pp. 1124–1141. https://doi.org/10.1016/j.scitotenv.2017.10.114.

[8] Khosravi K., Nohani E., Maroufinia E., Pourghasemi H.R.: *A GIS-based flood susceptibility assessment and its mapping in Iran: A comparison between frequency ratio and weights-of-evidence bivariate statistical models with multi-criteria decision-making technique*. Natural Hazards, vol. 83(2), 2016, pp. 947–987. https://doi.org/10.1007/s11069-016-2357-2.

[9] Chen J., Li Q., Wang H., Deng M.: *A machine learning ensemble approach based on random forest and radial basis function neural network for risk evaluation of regional flood disaster: A case study of the Yangtze River Delta, China*. International Journal of Environmental Research and Public Health, vol. 17(1), 2020, 49. https://doi.org/10.3390/ijerph17010049.

[10] Sahoo G.B., Schladow S.G., Reuter J.E.: *Forecasting stream water temperature using regression analysis, artificial neural network, and chaotic non-linear dynamic models*. Journal of Hydrology, vol. 378(3–4), 2009, pp. 325–342. https://doi.org/10.1016/j.jhydrol.2009.09.037.

[11] Tehrany M.S., Pradhan B., Jebur M.N.: *Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS*. Journal of Hydrology, vol. 512, 2014, pp. 332–343. https://doi.org/10.1016/j.jhydrol.2014.03.008.

[12] Li X., Zhang Q., Shao M., Li Y.: *A comparison of parameter estimation for distributed hydrological modelling using automatic and manual methods.* Advanced Materials Research, vol. 356–360, 2012, pp. 2372–2375. https://doi.org/10.4028/www.scientific.net/AMR.356-360.2372.

[13] Samanta S., Pal D.K., Palsamanta B.: *Flood susceptibility analysis through remote sensing, GIS and frequency ratio model*. Applied Water Science, vol. 8(2), 2018, 66. https://doi.org/10.1007/s13201-018-0710-1.

[14] Liu X., Li X., Liu L., He J., Ai B.: *An innovative method to classify remote-sensing images using ant colony optimization*. IEEE Transactions on Geoscience and Remote Sensing, vol. 46(12), 2008, pp. 4198–4208. https://doi.org/10.1109/TGRS.2008.2001754.

[15] Li X., Yeh A.G.O.: *Neural-network-based cellular automata for simulating multiple land use changes using GIS*. International Journal of Geographical Information Science, vol. 16(4), 2002, pp. 323–343. https://doi.org/10.1080/13658810210137004.

[16] Ganjirad M., Delavar M.R.: *Flood risk mapping using Random Forest and Support Vector Machine*. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. X-4/W1-2022, 2023, pp. 201–208. https://doi.org/10.5194/isprs-annals-X-4-W1-2022-201-2023.

[17] Lee S., Kim J.-C., Jung H.-S., Lee M.J., Lee S.: *Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea*. Geomatics, Natural Hazards and Risk, vol. 8(2), 2017, pp. 1185–1203. https://doi.org/10.1080/19475705.2017.1308971.

[18] Arlisa S.D., Handayani H.H.: *Flood vulnerability analysis using random forest method in Gresik Regency, Indonesia*. IOP Conference Series: Earth and Environmental Science, vol. 1127(1), 2023, 012023. https://doi.org/10.1088/1755-1315/1127/1/012023.

[19] Esfandiari M., Jabari S., McGrath H., Coleman D.: *Flood mapping using random forest and identifying the essential conditioning factors; A case study in Fredericton, New Brunswick, Canada*. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. V-3-2020, 2020, pp. 609–615. https://doi.org/10.5194/isprs-Annals-V-3-2020-609-2020.

[20] Farhadi H., Najafzadeh M.: *Flood risk mapping by remote sensing data and random forest technique*. Water, vol. 13(21), 2021, 3115. https://doi.org/10.3390/w13213115.

[21] Mobley W., Sebastian A., Blessing R., Highfield W.E., Stearns L., Brody S.D.: *Quantification of continuous flood hazard using random forest classification and flood insurance claims at large spatial scales: A pilot study in southeast Texas*. Natural Hazards and Earth System Sciences, vol. 21(2), 2021, pp. 807–822. https://doi.org/10.5194/nhess-21-807-2021.

[22] Catani F., Lagomarsino D., Segoni S., Tofani V.: *Landslide susceptibility estimation by random forests technique: Sensitivity and scaling issues*. Natural

Hazards and Earth System Sciences, vol. 13(11), 2013, pp. 2815–2831. https://doi.org/10.5194/nhess-13-2815-2013.

[23] Biau G., Scornet E.: *A random forest guided tour*. TEST, vol. 25(2), 2016, pp. 197–227. https://doi.org/10.1007/s11749-016-0481-7.

[24] Sarkar D., Mondal P.: *Flood vulnerability mapping using frequency ratio (FR) model: A case study on Kulik river basin, Indo-Bangladesh Barind region*. Applied Water Science, vol. 10(1), 2020, 17. https://doi.org/10.1007/s13201-019-1102-x.

[25] Ghosh R., Sutradhar S., Das N., Mondal P.: *A comparative evaluation of GIS based flood susceptibility models: A case of Kopai River Basin, Eastern India*. Research Square, 2021. https://doi.org/10.21203/rs.3.rs-705204/v1.

[26] Saha S., Sarkar D., Mondal P.: *Efficiency exploration of frequency ratio, entropy and weights of evidence-information value models in flood vulnerabilityassessment: A study of Raiganj Subdivision, Eastern India*. Stochastic Environmental Research and Risk Assessment, vol. 36(6), 2022, pp. 1721–1742. https://doi.org/10.1007/s00477-021-02115-9.

[27] Sarkar D., Saha S., Mondal P.: *GIS-based frequency ratio and Shannon's entropy techniques for flood vulnerability assessment in Patna district, Central Bihar, India*. International Journal of Environmental Science and Technology, vol. 19(9), 2022, pp. 8911–8932. https://doi.org/10.1007/s13762-021-03627-1.

[28] Dutta M., Saha S., Saikh N.I., Sarkar D., Mondal P.: *Application of bivariate approaches for flood susceptibility mapping: A district level study in Eastern India*. HydroResearch, vol. 6, 2023, pp. 108–121. https://doi.org/10.1016/j.hydres.2023.02.004.

[29] Ullah K., Zhang J.: *GIS-based flood hazard mapping using relative frequency ratio method: A case study of Panjkora River Basin, eastern Hindu Kush, Pakistan*. PLoS ONE, vol. 15(3), 2020, e0229153. https://doi.org/10.1371/journal.pone.0229153.

[30] Hagen E., Shroder J.F., Lu X.X., Teufert J.F.: *Reverse engineered flood hazard mapping in Afghanistan: A parsimonious flood map model for developing countries*. Quaternary International, vol. 226(1–2), 2010, pp. 82–91. https://doi.org/10.1016/j.quaint.2009.11.021.

[31] Pradhan B., Lee S., Buchroithner M.F.: *A GIS-based back-propagation neural network model and its cross-application and validation for landslide susceptibility analyses*. Computers, Environment and Urban Systems, vol. 34(3), 2010, pp. 216–235. https://doi.org/10.1016/j.compenvurbsys.2009.12.004.

[32] Glenn E.P., Morino K., Nagler P.L., Murray R.S., Pearlstein S., Hultine K.R.: *Roles of saltcedar (Tamarix spp.) and capillary rise in salinizing a non-flooding terrace on a flow-regulated desert river*. Journal of Arid Environments, vol. 79, 2012, pp. 56–65. https://doi.org/10.1016/j.jaridenv.2011.11.025.

[33] Sørensen R., Zinko U., Seibert J.: *On the calculation of the topographic wetness index: Evaluation of different methods based on field observations*. Hydrology and Earth System Sciences, vol. 10(1), 2006, pp. 101–112. https://doi.org/10.5194/hess-10-101-2006.

[34] Jahangir M.H., Mousavi Reineh S.M., Abolghasemi M.: *Spatial predication of flood zonation mapping in Kan River Basin, Iran, using artificial neural network algorithm*. Weather and Climate Extremes, vol. 25, 2019, 100215. https://doi.org/10.1016/j.wace.2019.100215.

[35] Haghizadeh A., Siahkamari S., Haghiabi A.H., Rahmati O.: *Forecasting flood-prone areas using Shannon's entropy model*. Journal of Earth System Science, vol. 126(3), 2017. https://doi.org/10.1007/s12040-017-0819-x.

[36] Thompson A., Clayton J.: *The role of geomorphology in flood risk assessment*. Proceedings of the Institution of Civil Engineers: Civil Engineering, vol. 150(5), 2002, pp. 25–29. https://doi.org/10.1680/cien.150.5.25.38634.

[37] Gudiyangada Nachappa T., Tavakkoli Piralilou S., Gholamnia K., Ghorbanzadeh O., Rahmati O., Blaschke T.: *Flood susceptibility mapping with machine learning, multi-criteria decision analysis and ensemble using Dempster Shafer Theory*. Journal of Hydrology, vol. 590, 2020, 125275. https://doi.org/10.1016/j.jhydrol.2020.125275.

[38] Pourali S.H., Arrowsmith C., Chrisman N., Matkan A.A., Mitchell D.: *Topography Wetness Index application in flood-risk-based land use planning*. Applied Spatial Analysis and Policy, vol. 9(1), 2016, pp. 39–54. https://doi.org/10.1007/s12061-014-9130-2.

[39] Funk C., Peterson P., Landsfeld M., Pedreros D., Verdin J., Shukla S., Husak G., Rowland J., Harrison L., Hoell A., Michaelsen J.: *The climate hazards infrared precipitation with stations – a new environmental record for monitoring extremes*. Scientific Data, vol. 2(1), 2015, 150066. https://doi.org/10.1038/sdata.2015.66.

[40] Muñoz P., Orellana-Alvear J., Willems P., Célleri R.: *Flash-flood forecasting in an andean mountain catchment-development of a step-wise methodology based on the random forest algorithm*. Water (Switzerland), vol. 10(11), 2018, 1519. https://doi.org/10.3390/w10111519.

[41] Breiman L.: *Random forests*. Machine Learning, vol. 45, 2001, pp. 5–32. https://doi.org/10.1023/A:1010933404324.

[42] North M.A.: *A method for implementing a statistically significant number of data classes in the Jenks algorithm*. [in:] *Sixth International Conference on Fuzzy Systems and Knowledge Discovery: FSKD 2009, Tianjin, China, 14–16 August 2009. Vol. 1*, IEEE, Piscataway 2009, pp. 35–38. https://doi.org/10.1109/FSKD.2009.319.

[43] Regmi A.D., Devkota K.C., Yoshida K., Pradhan B., Pourghasemi H.R., Kumamoto T., Akgun A.: *Application of frequency ratio, statistical index, and weights-of-evidence models and their comparison in landslide susceptibility mapping in Central Nepal Himalaya*. Arabian Journal of Geosciences, vol. 7(2), 2014, pp. 725–742. https://doi.org/10.1007/s12517-012-0807-z.

[44] Bhandari B.P., Dhakal S.: *Lithological control on landslide in the Babai Khola Watershed, Siwaliks Zone of Nepal*. American Journal of Earth Sciences, vol. 5(3), 2018, pp. 54–64.