Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 2, April 2026, pp. 1521~1534
ISSN: 2302-9285, DOI: 10.11591/eei.v15i2.10374

Flood mapping using Res-Q and machine learning on imbalanced data

Siti Yuliyanti[1], Vega Purwayoga[1,4], Andi Nur Rachman[2], Zakwan Gusnadi[3,4]
[1]Department of Informatics, Faculty of Engineering, Silliwangi University, Tasikmalaya, Indonesia
[2]Department of Information System, Faculty of Engineering, Siliwangi University, Tasikmalaya, Indonesia
[3]Department of Civil Engineering, Faculty of Engineering, Siliwangi University, Tasikmalaya, Indonesia
[4]Spatial Intelligence for Climate and Disaster Resilience, Research Group, Siliwangi University, Tasikmalaya, Indonesia

Article Info

Article history:

Received Mar 27, 2025
Revised Feb 23, 2026
Accepted Mar 5, 2026

Keywords:

Decision tree
Flood mapping
Imbalanced data
Machine learning
Reverse sort filter skyline

ABSTRACT

Flood disaster mapping requires accurate methods to support early warning and mitigation planning. To address common issues such as imbalanced data distribution and limited attribute handling, this study proposes an improved approach. The methodology includes: i) modification of the spatial sort filter skyline method with reverse normalization based on attribute preferences, applied when an attribute has minimal preference to ensure balanced consideration during skyline filtering; ii) data labeling and balancing, where initial flood potential labeling is generated using Res-Q, followed by K-Means clustering to group data into four classes (low, moderate, high, and very high) and SMOTE to further balance the dataset with 558 data points per class; iii) model evaluation using the C5.0 algorithm under three schemes, showing high and consistent accuracy with 89.24% on imbalanced data (Schema 2) and 93.3 % on balanced data (Schema 3), while Schema 1 shows overfitting due to extreme imbalance; and iv) the main contribution, integrating reverse normalization with skyline filtering combined with clustering and resampling, enhancing both accuracy and robustness in identifying flood-prone areas. This structured approach highlights methodological improvements, reliable results, and practical contributions for effective flood disaster management.

Corresponding Author:

Vega Purwayoga
Department of Informatics, Faculty of Engineering, Silliwangi University
Tasikmalaya, Indonesia
Email: vega.purwayoga@unsil.ac.id

1. INTRODUCTION

Indonesia is one of the countries highly vulnerable to natural disasters. The types of disasters that frequently occur include earthquakes, floods, landslides, tsunamis, and volcanic eruptions [1]. The frequent occurrence of disasters presents a challenge that needs to be addressed through various approaches to reduce their impacts. The impacts of natural disasters include infrastructure damage, disease outbreaks, as well as social and economic disruptions [2]. Among these disasters, flooding is particularly devastating, underscoring the need for effective disaster preparedness and management. Floods not only cause physical damage but also lead to economic, psychological, and health-related consequences [3]. As demonstrated in a previous study [4], disaster mitigation can be addressed through the formulation of disaster management policies. These policies include mapping disaster-prone areas, predicting disasters, and distributing disaster relief logistics [5]. Based on several prior studies, disaster management can be effectively implemented through a cross-regional collaboration framework [6]. This framework involves mapping potential disasters as a preventive measure, executing response strategies during disasters, and managing post-disaster recovery efforts [6]-[8].

The disaster mapping process can be conducted using a recommendation algorithm, namely the skyline query [9]. The skyline query is capable of identifying optimal locations based on specific location characteristics [10]-[12]. Its ability to map potential disaster-prone areas makes it a valuable tool for disaster management. However, three significant challenges associated with the skyline query algorithm include the flexibility of user queries, data processing speed, and the validity of the recommendation results [13]-[16] study conducted in [12] revealed a weakness in the validity of the recommendation results due to a sorting process that does not consider the concept of maximum and minimum preferences [14]. The studies in [12], [14] implemented the skyline filter sort (SFS) algorithm, where relevant objects are selected based on entropy calculations. However, a major limitation of previous studies is their failure to account for minimum or maximum preference values. The SFS algorithm used in prior research only considers the values within an attribute. If an attribute has a large value, the entropy value increases. However, in the skyline query framework, a minimum preference exists, meaning that a smaller value is sometimes the preferred choice.

A prediction model can enhance the flood mapping process and be utilized as an early warning system. Several algorithms can be used for prediction, including decision tree models. Tanyu *et al*. [17] applied machine learning algorithms to predict landslide disasters. The decision tree models used in the study included C4.5, C5.0, and random forest. The findings in [17] indicated that C5.0 outperformed both random forest and C4.5 when applied to imbalanced data.

Based on the limitations identified in previous studies, this study modifies the SFS algorithm to incorporate both minimum and maximum preferences, influencing the entropy value to generate objects that best match the flood potential criteria. If a criterion has a minimum preference, reverse normalization is applied. Consequently, the modified algorithm in this study is named reverse spatial sort filter skyline (Res-Q). Additionally, this study applies the C5.0 machine learning algorithm to compare the performance of balanced, unbalanced, and highly unbalanced data. The results of object searches using skyline queries often yield highly imbalanced data; therefore, this study also aims to reduce the degree of data imbalance. Res-Q generates recommendations for areas most likely to be affected by flooding. As with the general skyline concept, the skyline query identifies only the best objects without considering other classes or the balance of the data. However, because classification requires balanced data, the data is grouped based on the priority value derived from the entropy results obtained using the Res-Q model.

The process of grouping data to assist in labeling flood-prone areas is performed using an unsupervised machine learning approach, specifically the K-Means algorithm. K-Means is capable of labeling flood potential by determining the number of flood potential classes [18]. The entropy value of each area's object is measured for proximity using the Euclidean distance formula. While K-Means can classify areas based on flood potential, it cannot regulate object density within each group. As a result, data imbalance may still occur. To address this issue, this study also applies the synthetic minority over-sampling technique (SMOTE), which balances class distributions without losing data [19]. SMOTE effectively generates realistic and diverse synthetic data. The novelty of this study can be summarized as follows:

- Modification of the SFS algorithm into Res-Q by introducing reverse normalization, which enables attributes with different preference directions to be fairly considered in flood-prone area labeling.
- A hybrid balancing strategy that integrates Res-Q labeling with K-Means clustering and SMOTE resampling to effectively address severe class imbalance in flood datasets.
- Systematic evaluation with the C5.0 decision tree algorithm under three data schemes (Res-Q, K-Means, and SMOTE), providing evidence of consistent and robust performance across both imbalanced and balanced conditions.

2. METHOD

This research consists of several stages, as illustrated in Figure 1. These stages include data acquisition, data pre-processing, labeling of flood-prone areas using the spatial skyline query, data balancing, flood potential mapping, application of the C5.0 algorithm, and model evaluation. The study focuses on areas in Central Java Province. According to data from the National Disaster Management Agency (BNPB), 94% of the regions in Central Java Province have a high level of flood vulnerability.

2.1. Data acquisition

This stage involves collecting spatial and non-spatial data required for flood analysis. The data are obtained from various official sources to support accurate and comprehensive assessment.

- The acquired data includes administrative boundaries in shapefile format, digital elevation model (DEM) data in raster format, climatological data, and land cover data. The administrative boundary data from the Geospatial Information Agency (BIG) includes county-level administrative boundaries.

- An essential dataset is the DEM, which consists of raster or grid data. Each grid is acquired to cover the area defined by the administrative boundary data. DEM data is used to create maps that classify elevation and land slope in a region [20]. Several DEM datasets are needed for each county to fully cover the area and classify its elevation and slope [21].
- In addition to DEM data, climatological data is crucial, particularly about flood disasters [22]. Climatological data includes rainfall, temperature, humidity, wind speed, wind direction, and more. According to research [23], [24], rainfall is the most significant factor influencing flood events—the higher the rainfall volume over a specific period, the greater the potential for flooding in that area. Climatological data, especially rainfall data, was acquired from the meteorology, climatology, and geophysics agency (BMKG) and collected from several weather stations in Central Java and the Special Region of Yogyakarta.
- In addition to elevation, relief, and climatological factors, land cover is another crucial factor contributing to flooding, as highlighted by research [25]. Changes in land use leads to increased flood vulnerability. According to research [26], land cover is classified into five categories: class 1, forests; class 2, plantations; class 3, bushes, shrubs, and reeds; class 4, agriculture, and settlements; and class 5, rice fields, ponds, and open land.

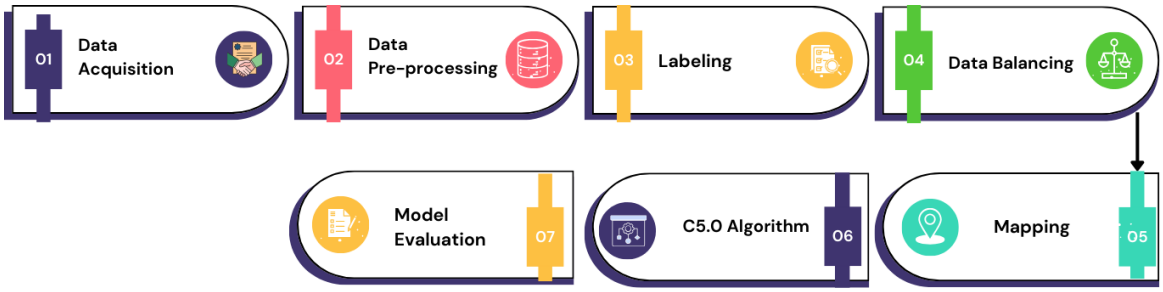


Figure 1. Research stages

2.2. Data pre-processing

The acquired data are processed to ensure compatibility for integration and further analysis. This stage includes several preprocessing steps to prepare the data for subsequent analysis stages.

- The administrative data acquired from BIG was filtered to include only regions within Central Java Province. The selected regions have the highest flood risk in Central Java. In addition to selecting areas with the highest flood potential, neighboring regions were also included for comparative analysis. In line with cross-regional collaboration, neighboring areas with spatial proximity can support each other, particularly in disaster management efforts [6].

The determination of neighboring areas was based on proximity search functions and distance measurement functions. Distance measurement was performed using the Haversine Distance formula, as presented in (1). The Haversine Distance method offers higher accuracy than other distance measurement methods [27].

$$d = \sqrt{2r.\arcsin\left(sin^2\left(\frac{\varphi_2-\varphi_1}{2}\right) + \cos(\varphi_1).\cos(\varphi_2).sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)\right)} \quad (1)$$

Where φ: earth's latitude coordinates (latitude)
$\boldsymbol{\lambda}$: earth's longitude coordinates (longitude)
r: earth's radius length (*radius*)

- The acquired DEM data was merged into a single raster, which was then clipped according to the administrative boundaries of the flood potential study area. The clipped DEM data underwent a classification process for elevation and land slope.
- Mapping the rainfall values in flood potential areas is conducted using interpolation techniques. The interpolation method employed is ordinary co-kriging (OCK). Based on previous studies, OCK demonstrates good accuracy in rainfall interpolation. Rainfall is categorized into several classes: low (<250 mm), slightly low (250-300 mm), slightly high (301-350 mm), and high (>350 mm) [28].
- A selection process is conducted in the land cover data to identify land cover areas that correspond to the selected study area.

2.3. Labeling using spatial skyline query

Annisa and Khairina [10] applied SFS to search nearby public facilities, including food courts, supermarkets, health clinics, and places of worship. SFS was also utilized in the study for the distribution of personal protective equipment (PPE) [29]. SFS represents an improvement over the block nested loops (BNL) algorithm [29]. In BNL, the first object is stored as a skyline object; if another object dominates the stored object in memory, it is eliminated and replaced by the dominating object. In contrast to BNL, SFS first performs sorting before dominance testing. This sorting aims to identify the first skyline object that dominates all other objects. A comparison of several skyline query algorithms is presented in Table 1. From the table, it can be seen that the performance of SFS is influenced by the number of skyline objects obtained, which is highly dependent on the number of dimensions in the data.

Table 1. Comparison of skyline research

Ref.	Skyline model	Research findings
[30]	BNL and SFS	Number of comparisons for BNL (n=1000) ⇒ 499,500 comparisons Number of comparisons for SFS (n=1000), if the skyline size is 1% ⇒ 1,000 comparisons Number of comparisons for SFS (n=1000), if the skyline size is 10% ⇒ 10,000 comparisons Number of comparisons for SFS (n=1000), if the skyline size is 50% ⇒ 500,000 comparisons
[29]	SFS	In cases where the number of data points (n) < 1000, the sorting results have not been tested against the dominance criteria.
[31]	BNL, SFS, and Pro-SKY	SFS experiences a decline in computational performance when 3–4 additional dimensions are added.

The sorting process is conducted through normalization and entropy value calculation. Each attribute, such as rainfall, elevation, and others, is normalized using min-max normalization. The normalized results for each attribute are then inputted into the entropy formula to calculate entropy values; the higher the entropy value, the more likely the data will dominate other data. Normalization of values for each attribute is conducted using the min-max normalization method into the range [0,1] in (2, where represents the normalization result, denotes the object, represents all objects attached to, signifies the smallest value, and denotes the highest value for each attribute of the object. The entropy calculation is presented in (3).

$$f = \frac{D_{[a_i]} - min_{(a)}}{max_{(a)} - min_{(a)}} \quad (2)$$

$$E(D) = \sum_{i=1}^{d} \ln\left(D_{[a_i]} + 1\right) \quad (3)$$

The skyline query is implemented using two schemes: the default SFS model [29] and Res-Q. The implementation of these two skyline models aims to compare the objects recommended by SFS and Res-Q. The Res-Q algorithm, which is a modification of the default SFS, is presented in Algorithm 1.

Algorithm 1. Res-Q

```
Input: Dataset D
Output: The Set of skyline points of D
 1:  D ← if ("D[ai]= minimum preference") then
 2:         1 – normalize D[ai]
 3:     else
 4:        normalize D[ai]
 5:     end if
 6:  E(D) ← calculate entropy of D[ai] using (3)
 7:  D ← sort dataset by descending D[ai]
 8:  S ← data with the highest entropy D
 9:  From 1 to D
10:     if ("D is not dominated") then
11:        write (S, D)
12:     else
13:        remove (S, D)
14:     end if
15:  end
```

Res-Q verifies whether the preference used is minimum or maximum. If the minimum preference is used, the normalization process is reversed. The reverse normalization is obtained by subtracting the normalization result of D[ai] from 1, as shown in lines 1–2.

2.4. Data balancing

The SFS and Res-Q models generate recommendations for areas most likely to be affected by flooding. Following the skyline concept, the skyline identifies the best objects without considering other classes or the balance of the data. Since the classification process must account for data balance [19], the data is grouped based on the priority values derived from the entropy results obtained through the Res-Q model.

The K-Means algorithm is used in the grouping process to assist in labeling flood potential. K-Means assigns labels by categorizing data into flood potential classes such as Low, Moderate, High, and Very High. The value of K in the K-Means algorithm is 4, which is determined based on the flood potential classes presented in the study [32]. However, a limitation of K-Means lies in data distribution, as it cannot ensure a balanced number of members in each class, potentially leading to unbalanced data. This issue is addressed using the synthetic minority over-sampling technique (SMOTE), which adds data to the minority class to improve balance. The parameters were set as k_neighbors 5, sampling_strategy="auto", and random_state=42.

2.5. Mapping potential flood areas

One of the benefits of disaster maps, particularly flood disaster maps, is their ability to identify disaster vulnerability levels in a given area. Regions with high disaster vulnerability can prioritize infrastructure reinforcement, preparedness measures, and disaster response efforts [32].

2.6. Classification using decision tree C5.0

Before classification, the data is divided into two parts: training data and testing data. In this study, data partitioning is performed using K-fold cross-validation, with K set to 10. The choice of K = 10 was made because this value yields more stable performance compared to other K values [29]. The data partitioning scheme is presented in Table 2.

Table 2. Data partition

Fold on test data	Fold on train data
1	2,3,4,5,6,7,8,9,10
..	…
10	1,2,3,4,5,6,7,8,9

The decision tree model has a basic tree-like structure consisting of rules for making decisions [33]. The C5.0 algorithm was developed as an improvement over previous decision tree models. In building a classification model, the C5.0 algorithm requires several key variables, such as Gain and Entropy. The formulas for calculating Gain and Entropy values can be found in (4) to (6):

$$Entropy(D) = -\sum_i^m p_i log_2 p_i \quad (4)$$

$$Entropy_A(D) = -\sum_i^m \frac{|D_j|}{|D|} \times Entropy(D_j) \quad (5)$$

$$Gain(A) = Entropy(D) - Entropy_A(D) \quad (6)$$

2.7. Model evaluation

Model evaluation is performed using a confusion matrix. The confusion matrix consists of key factors used to measure model performance, including true positive (TP), true negative (TN), false positive (FP), and false negative (FN), as explained in (7) [34]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

Spatial validation was performed by overlaying the predicted flood-prone areas with the actual observed flood events using GIS analysis. This step allowed a spatial interpretation of the model's accuracy, showing where the model correctly or incorrectly predicted flood occurrences.

3. RESULTS AND DISCUSSION

3.1. Data acquisition

The results of data acquisition describe the characteristics, sources, and coverage of the datasets used in this study. These datasets provide the necessary spatial and temporal information to support flood potential analysis and ensure the reliability of the subsequent processing stages.

- The study area used in this research is Central Java Province. The available administrative boundary data comprises all regions in Indonesia, necessitating the selection of areas located in Central Java Province.
- The DEM data consists of 77 grids. The entire grid is merged to perform the clipping process with the selected study area. The DEM data consists of 77 grids. The entire grid is merged to perform the clipping process with the selected study area. The DEM data, with 8 m spatial resolution, act as remote sensing inputs providing detailed topographic information essential for flood potential mapping.
- Rainfall data were obtained from 9 weather stations, with six stations located in Central Java Province and three in the Special Region of Yogyakarta (DIY). Data were collected throughout the rainy season, from October 2023 to March 2024, and accumulated monthly to capture temporal variations in rainfall intensity. These weather stations serve as a sensor network, providing time-sensitive hydrological input that reflects changes in flood potential over the rainy season.
- Land cover data were obtained from the Ministry of Forestry and Environment. According to research [27], land cover is classified into five categories: class 1, forests; class 2, plantations; class 3, bushes, shrubs, and reeds; class 4, agriculture and settlements; and class 5, rice fields, ponds, and open land. The land cover maps act as remote sensing inputs supporting flood potential analysis.

The combination of sensor-based rainfall measurements and remote sensing-derived DEM and land cover data ensures sufficient spatial and temporal detail. This integrated dataset can be utilized in embedded or IoT-based platforms for real-time urban flood monitoring and decision support.

3.2. Data pre-processing

The results of data preprocessing present the transformed datasets after applying various preparation techniques. These processes ensure that the data are properly structured and ready for subsequent spatial analysis.

- The proximity of each region was measured using the Haversine Distance. Neighboring regions are considered potential contributors to flood management, as closer areas are more likely to assist. Based on BNPB data, Demak Regency has the largest flood-affected area. This study includes Demak Regency, its neighboring regions, and other nearby areas, identified using the spatial join (sjoin) function in the Geopandas library [35]. Distance measurements to determine regions adjacent to Demak Regency were carried out using the Haversine Distance formula, assisted by the math library [36]. Only regions within an average distance of <= the proximity of any area to Demak Regency were considered.
- The DEM data, consisting of several rasters, were merged and clipped to fit the study area using the clip function in ArcGIS. After clipping, a classification process was conducted using the spatial analyst tools and the reclassify function. The results of the preprocessing are presented in Figure 2 (in Appendix). The slope classes are divided into five categories, ranging from flat to very steep. The flat category dominates the relief classification results. According to research, flat relief is the most flood-prone category [29]. The classification results of elevation and slope can be seen in Figures 2(a) and (b).
- The rainfall data from several weather stations is distributed across various regions using interpolation techniques. The interpolation results are categorized into several classes: low, slightly low, slightly high, and high [29]. In terms of rainfall, the dominant class is high, leading to the conclusion that most of the selected study area experiences high levels of rainfall. The rainfall interpolation results are presented in Figure 2(c).
- The classification of land use and land cover is divided into five classes, as presented in Figure 2(d).

3.3. Labeling using spatial skyline query

Table 3 presents the first skyline object generated by the SFS and Res-Q models. The first object identified by SFS is the area with ID 885, whereas Res-Q designates the area with ID 667 as the skyline object. The skyline object refers to the area with the highest potential impact in the event of a flood disaster.

Table 3. Skyline object search results

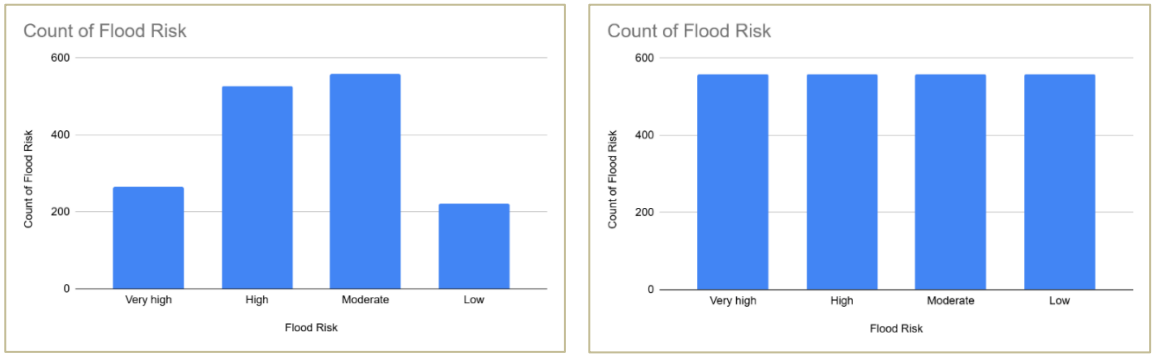
Default SFS	Score
First object	855
Region name	Magelang
Object characteristics (region)	Rainfall: high (4), slope: very steep (5), land use/land cover: rice fields, ponds, open land (5), elevation: high (4), area size: large (3)
Res-Q	Score
First object	667
Region name	Batang
Object characteristics (region)	Rainfall: high (4), slope: flat (1), land use/land cover: rice fields, ponds, open land (5), elevation: very low (1), area size: large (3)

Based on the flooding criteria, the Res-Q model produces the best recommendation. It selects object 667 due to the area's high rainfall, flat slope, high-risk land cover, very low elevation, and large area size. Since Res-Q provides an appropriate recommendation, its results are used in the next stage.

3.4. Data balancing

The Res-Q model produces only two classes of flood potential: high-risk areas and low-risk areas. There are 18 high-risk polygons or regions, while 1574 areas are classified as low-risk. Due to the significant imbalance in class proportions, the Res-Q results are considered highly imbalanced data. This extreme imbalance indicates a potential risk of overfitting, as models trained on such data may overly favor the majority class while underrepresenting the minority class.

One approach to addressing highly imbalanced data is to increase the number of flood potential classes by applying the clustering concept. The clustering process is conducted using the K-Means algorithm, with the number of clusters (K) set to four. The value of K corresponds to the flood potential classification, as explained in the study [32]. The determination of the number of clusters was also validated using the Elbow method, which indicated that the optimal number of clusters is four for grouping flood potential data in this study. The clustering process is based on the entropy value of each area or polygon. The visualization of the class distribution generated by clustering and SMOTE is presented in Figure 3. Figure 3(a) and (b) illustrate the results of K-Means clustering and SMOTE, respectively.



(a) (b)

Figure 3. Distribution of flood risk classes; (a) K-Means and (b) SMOTE

Figure 3 shows that K-Means can mitigate data imbalance, but the results are still not optimal. The low class consists of 202 data points, moderate 558, high 528, and very high 266. Since the K-Means results are not optimal, this study employs SMOTE to address the issue of unbalanced data. The data distribution generated by SMOTE results in 558 instances per class. SMOTE operates by augmenting the minority class with synthetic data [19]. The findings of study [37] confirm that SMOTE is not only used to balance the distribution of flood potential data, but also to maintain the validity of the augmented data. SMOTE was applied using its default configuration, with *k_neighbors*=5 and *sampling_strategy*="auto", as provided by the *imbalanced-learn* library.

3.5. Mapping potential flood areas

The mapping of potential flood areas is carried out by aggregating flood potential classes across all regions. If a region is predominantly composed of areas classified as having a very high flood potential, it is categorized as being at very high flood risk.

Based on Figure 4, the number of areas with very high flood potential is one, high is seven, medium is twelve, and low is one. That indicates that most areas have a moderate flood risk. The flood potential map clearly shows areas at different levels of flood risk. This visualization can support smart city planning, helping authorities identify high-risk zones and make timely decisions for flood management.

3.6. Classification using decision tree C5.0

The C5.0 algorithm is applied to three schemes. Schema 1 is performed on highly imbalanced data generated from Res-Q. Schema 2 is applied to imbalanced data produced by the K-Means algorithm, while Schema 3 is performed on balanced data generated using the SMOTE technique. Details of each schema are presented in Table 4. Each schema is divided into ten folds, which illustrate the involvement of attributes in the decision tree and the rules generated. Table 4 shows that all attributes are included in the decision tree.

Figure 4. Mapping of potential flood areas

Table 4. Classification results

Schema	Fold on test data	Fold on test data	Attribute usage	Number of rules
Schema 1	1	2,3,4,5,6,7,8,9,10	Rainfall, slope, elevation, land use and land cover, area size	16
	..	…		…
	10	1,2,3,4,5,6,7,8,9	Rainfall, slope, elevation, land use and land cover, area size	16
Schema 2	1	2,3,4,5,6,7,8,9,10	Rainfall, slope, elevation, land use and land cover, area size	514
	…	…		…
	10	1,2,3,4,5,6,7,8,9	Rainfall, slope, elevation, land use and land cover, area size	532
Schema 3	1	2,3,4,5,6,7,8,9,10	Rainfall, slope, elevation, land use and land cover, area size	499
	…	…	…	…
	10	1,2,3,4,5,6,7,8,9	Rainfall, slope, elevation, land use and land cover, area size	541

The average number of rules generated in Schema 1 is 16, in Schema 2 is 532, and in Schema 3 is 537. The number of classes plays a crucial role in determining the number of rules based on the three schemas applied.

- If area size=small, land cover=plantation, elevation=very low, rainfall=very low then flood risk=low
- If area size=small, land cover=forest, elevation=very low, rainfall=slightly low, slope=slope then flood risk=moderate
- If area size=small, land cover=rice fields, ponds, open land, elevation=very low, rainfall=slightly low, slope=slope then flood risk=high
- If area size=small, land cover=rice fields, ponds, open land, elevation=very low, rainfall=high, slope=flat then flood risk=very high.

Based on Table 3, Schema 1 produces fewer rules compared to Schema 2 and Schema 3. This condition may be attributed to the data imbalance in Schema 1, which results in a simpler decision tree but tends to be biased toward the majority class [38].

3.7. Model evaluation

Figure 5 presents the test results of the three C5.0 algorithm schemes. The knowledge derived from C5.0 can enhance disaster preparedness in areas identified as having flood potential. Figure 5(a) shows the confusion matrix for Schema 1, Figure 5(b) presents the confusion matrix for Schema 2, while Figure 5(c) presents the confusion matrix for Schema 3.

As shown in Figure 5, the best-performing fold in Schema 2 is Fold 2, with an accuracy of 89.24%, while in Schema 3, it is Fold 2, with an accuracy of 93.30%. Figure 5(a) shows that all data were classified correctly, with no misclassifications. This happened because the Skyline results already clearly separate the High and Low classes. However, the data are highly imbalanced, with the Low class being much larger than the High class. Therefore, the high accuracy is mainly due to the dominance of the Low class rather than the

model's balanced performance. In Schema 1, the data used to build the C5.0 model was derived from the Res-Q recommendations. These recommendations produced only two classes, namely high and low, which increases the risk of overfitting in the C5.0 model. Moreover, the data distribution generated by Res-Q is highly imbalanced, with 18 instances in the high class and 1556 instances in the low class. Based on the confusion matrices for Fold 2 in Schema 2 and Schema 3, the model demonstrates good and consistent classification performance. Most samples are correctly classified, with only a few misclassifications, primarily between the high–moderate and low–moderate classes.

The evaluation of the model's ability to distinguish between positive and negative classes was conducted using the receiver operating characteristic (ROC) curve, as shown in Figure 6. Figures 6(a)-(c) correspond to Scheme 1, Scheme 2, and Scheme 3, respectively. In Scheme 1, an AUC value of 0 indicates that the model was unable to distinguish between classes effectively, which was caused by a highly imbalanced data distribution. However, in Scheme 2 and Scheme 3, the ROC curves show a significant improvement, with AUC values above 0.96 across all classes. This indicates that the model has been able to classify the data much more accurately. Furthermore, the results in Scheme 3 appear better than those in Scheme 2, suggesting an improvement in model performance after the data balancing process was applied.

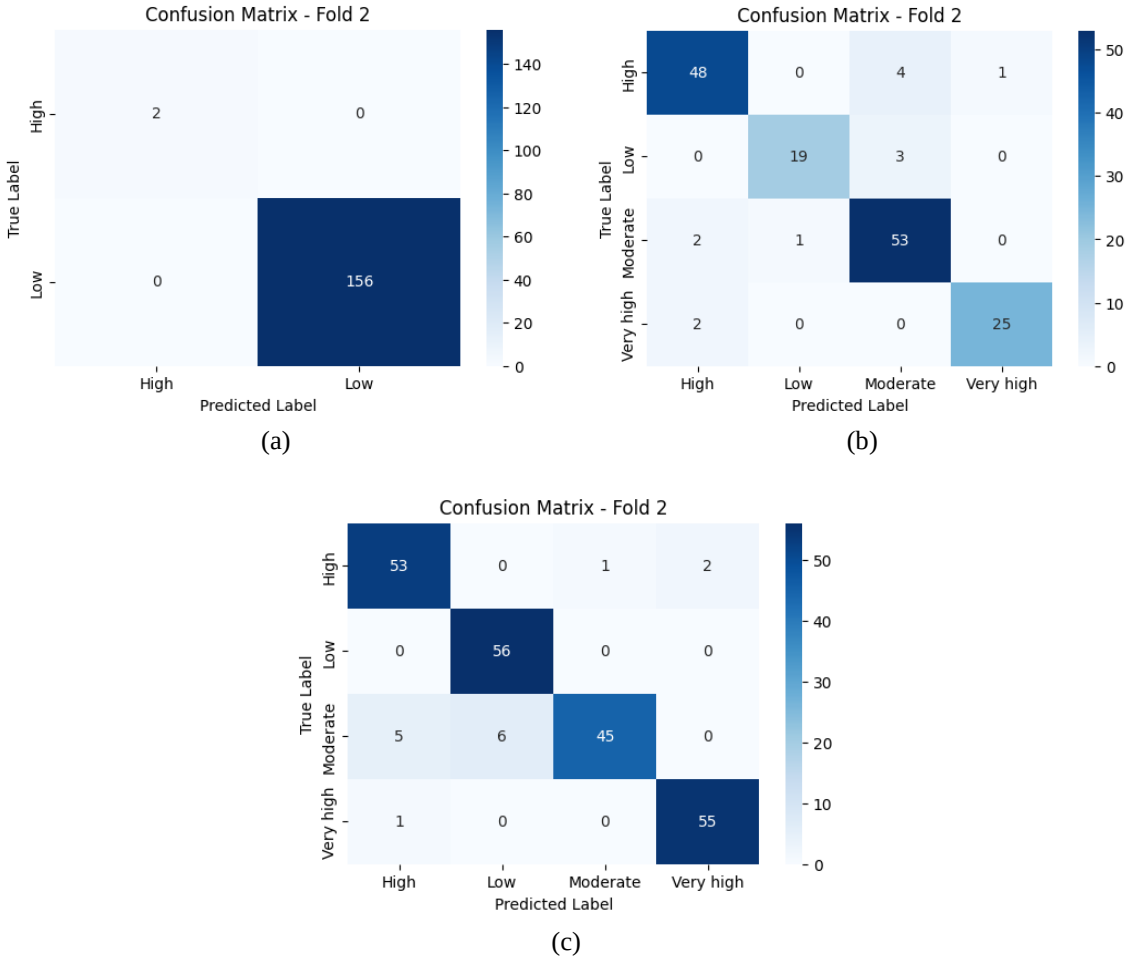


Figure 5. Model evaluation; (a) Schema 1, (b) Schema 2, and (c) Schema 3

Based on the Wilcoxon signed-rank test between Schema 2 and Schema 3, the obtained p-value=0.002. Since the p-value <0.05, there is a statistically significant difference between the two schemas. The average score of Schema 3 is higher than that of Schema 2, indicating that Schema 3 performs significantly better than Schema 2. The strong performance of Schema 2 and Schema 3 indicates that the trained C5.0 model can be applied for real-time flood monitoring, where it can process rainfall data continuously acquired from sensor stations to provide updated flood risk predictions. When integrated with cloud platforms and mobile alert systems, the model can help authorities quickly identify flood-prone areas and support smart city flood management.

In Figure 7, the flood-prone areas predicted by the C5.0 model for the high flood class are presented. The predicted flood areas were overlaid with the administrative map of Semarang city and verified against the official flood map owned by the city government. The verification results show that most of the predictions correspond well with the flood risk map provided by the Semarang city government.

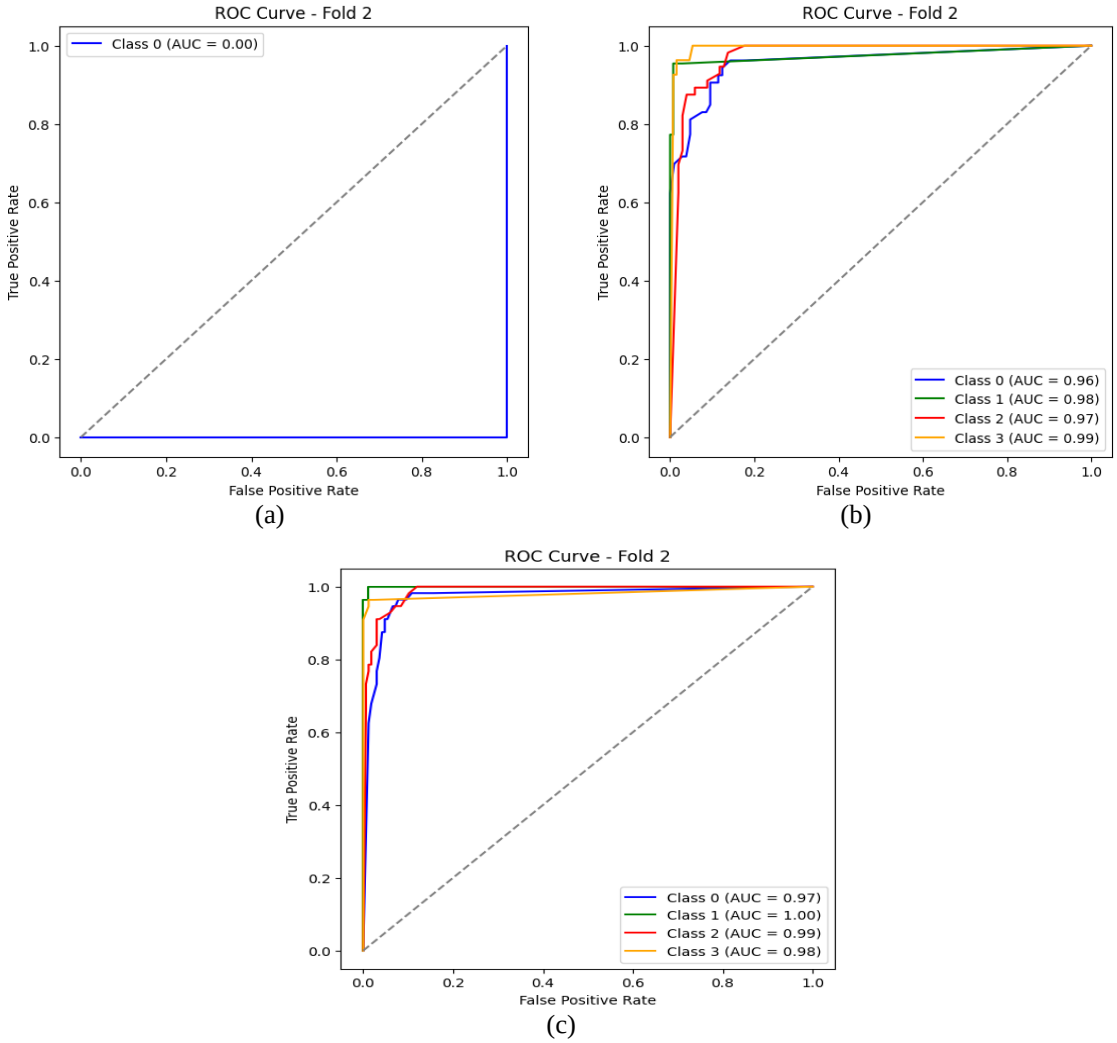


Figure 6. ROC curve; (a) Schema 1, (b) Schema 2, and (c) Schema 3

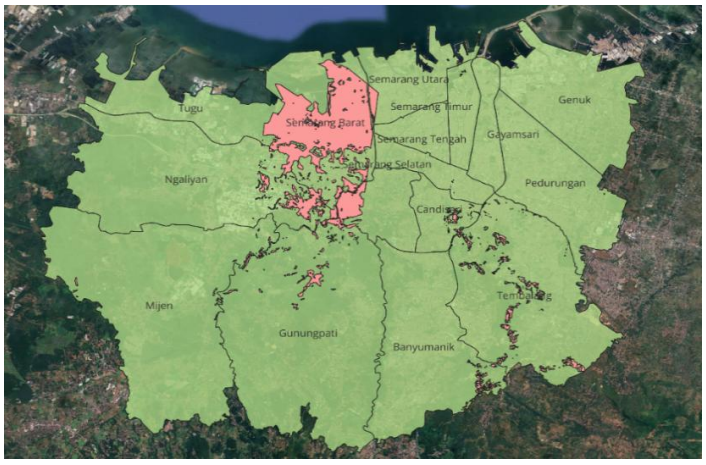


Figure 7. Flood zone prediction map

4. CONCLUSION

This study successfully developed a new model that addresses the shortcomings of the default SFS in mapping potential flood areas. The new model, named Res-Q, demonstrates higher accuracy in recommending potential flood areas compared to the default SFS model. Additionally, this study effectively compares several classification schemes using the C5.0 algorithm on different data distributions, including balanced data generated from SMOTE, imbalanced data from K-Means, and highly unbalanced data from Res-Q. The results show that the C5.0 algorithm performs well on both imbalanced data (Schema 2) and balanced data (Schema 3). However, when applied to the highly imbalanced dataset in Schema 1, the algorithm had difficulty building a stable model, and the large gap between classes led to signs of overfitting. The performance difference between Schema 2 and Schema 3 was further examined using the Wilcoxon signed-rank test, which produced a p-value of 0.002. Since this value is below the 0.05 significance level, it indicates a statistically significant difference in performance between the two schemas. This study is expected to contribute to improving disaster preparedness, particularly for flood events. Flood potential maps can help determine how flood relief should be distributed, so assistance can be sent from safer areas to locations that are most vulnerable. Future studies should include both historical and real-time climate data to improve monitoring of changing weather conditions. In addition, the proposed system could be integrated into smart city infrastructure, enabling local governments to combine flood monitoring with broader urban management and IoT systems.

FUNDING INFORMATION

We extend our gratitude to the Directorate of Research, Technology, and Community Service (DRTPM) of the Ministry of Education, Culture, Research, and Technology for funding this research in the 2024 fiscal year under Grant Contract No. 088/E5/PG.02.00.PL/2024. We would also like to thank the Institute for Research and Community Services at Siliwangi University for supporting this research.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Siti Yuliyanti	✓	✓	✓	✓	✓				✓	✓			✓	✓
Vega Purwayoga	✓	✓		✓		✓		✓	✓	✓	✓	✓		
Andi Nur Rachman	✓					✓	✓			✓	✓		✓	✓
Zakwan Gusnadi	✓			✓	✓	✓	✓		✓				✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data derived from this study that support its findings may be requested from the corresponding author.

REFERENCES

[1] F. Faradiba, St. F. Azzahra, T. Guswantoro, L. Zet, and N. G. Manullang, "Assessing Natural Disaster Vulnerability in Indonesia Using a Weighted Index Method," *Nature Environment and Pollution Technology*, vol. 24, no. 1, p. D1683, Mar. 2025, doi: 10.46488/NEPT.2025.v24i01.D1683.

[2] M. Fuady, R. Munadi, and M. A. K. Fuady, "Disaster mitigation in Indonesia: between plans and reality," in *The 10th Annual International Conference on Science and Engineering (10th AIC 2020)*, vol. 1087, no. 1, p. 012011, Feb. 2021, doi: 10.1088/1757-899X/1087/1/012011.

[3] F. I. W. Rohmat, A. J. Löhr, F. Pratama, N. S. Burnama, and A. A. Kuntoro, "Quantifying time-dependent flood resilience index in a densely populated urban environment in Manado, Indonesia," *International Journal of Disaster Risk Reduction*, vol. 116, p. 105112, Jan. 2025, doi: 10.1016/j.ijdrr.2024.105112.
[4] D. Ayuningtyas, S. Windiarti, M. S. Hadi, U. U. Fasrini, and S. Barinda, "Disaster Preparedness and Mitigation in Indonesia: A Narrative Review," *Iranian Journal of Public Health*, Jul. 2021, doi: 10.18502/ijph.v50i8.6799.
[5] D. I. Putra and M. Matsuyuki, "Disaster Management Following Decentralization in Indonesia: Regulation, Institutional Establishment, Planning, and Budgeting," *Journal of Disaster Research*, vol. 14, no. 1, pp. 173–187, Feb. 2019, doi: 10.20965/jdr.2019.p0173.
[6] J. Liu, Y. Guo, S. An, and C. Lian, "A Study on the Mechanism and Strategy of Cross-Regional Emergency Cooperation for Natural Disasters in China—Based on the Perspective of Evolutionary Game Theory," *International Journal of Environmental Research and Public Health*, vol. 18, no. 21, p. 11624, Nov. 2021, doi: 10.3390/ijerph182111624.
[7] J. Li, J. Wang, H. Lee, and X. Zhao, "Cross-regional collaborative governance in the process of pollution industry transfer: The case of enclave parks in China," *Journal of Environmental Management*, vol. 330, p. 117113, Mar. 2023, doi: 10.1016/j.jenvman.2022.117113.
[8] Y. Wang, "Multiperiod Optimal Allocation of Emergency Resources in Support of Cross-Regional Disaster Sustainable Rescue," *International Journal of Disaster Risk Science*, vol. 12, no. 3, pp. 394–409, Jun. 2021, doi: 10.1007/s13753-021-00347-5.
[9] R. D. Kulkarni and B. F. Momin, "Skyline computation for frequent queries in update intensive environment," *Journal of King Saud University - Computer and Information Sciences*, vol. 28, no. 4, pp. 447–456, Oct. 2016, doi: 10.1016/j.jksuci.2015.04.003.
[10] A. Annisa and S. Khairina, "Location Selection Based on Surrounding Facilities in Google Maps using Sort Filter Skyline Algorithm," *Khazanah Informatika : Jurnal Ilmu Komputer dan Informatika*, vol. 7, no. 2, pp. 65–72, Jul. 2021, doi: 10.23917/khif.v7i2.12939.
[11] A. Annisa and L. Angraeni, "Location Selection Query in Google Maps using Voronoi-based Spatial Skyline (VS2) Algorithm," *Jurnal Online Informatika*, vol. 6, no. 1, p. 25, Jun. 2021, doi: 10.15575/join.v6i1.667.
[12] N. T. Lapatta, "Ecotourism Recommendations based on Sentiments Using Skyline Query and Apache-Spark," *Journal of Social Science*, vol. 3, no. 3, pp. 534–546, May 2022, doi: 10.46799/jss.v3i3.333.
[13] A. Vlachou, C. Doulkeridis, J. B. Rocha-Junior, and K. Nørvåg, "Decisive skyline queries for truly balancing multiple criteria," *Data & Knowledge Engineering*, vol. 147, p. 102206, Sep. 2023, doi: 10.1016/j.datak.2023.102206.
[14] M. B. Swidan, A. A. Alwan, S. Turaev, and Y. Gulzar, "A Model for Processing Skyline Queries in Crowd-sourced Databases," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 10, no. 2, p. 798, May 2018, doi: 10.11591/ijeecs.v10.i2.pp798-806.
[15] A. Cuzzocrea, P. Karras, and A. Vlachou, "Effective and efficient skyline query processing over attribute-order-preserving-free encrypted data in cloud-enabled databases," *Future Generation Computer Systems*, vol. 126, pp. 237–251, Jan. 2022, doi: 10.1016/j.future.2021.08.008.
[16] D. Yuan, L. Zhang, S. Li, and G. Sun, "Skyline query under multidimensional incomplete data based on classification tree," *Journal of Big Data*, vol. 11, no. 1, p. 72, May 2024, doi: 10.1186/s40537-024-00923-8.
[17] B. F. Tanyu, A. Abbaspour, Y. Alimohammadlou, and G. Tecuci, "Landslide susceptibility analyses using Random Forest, C4.5, and C5.0 with balanced and unbalanced datasets," *CATENA*, vol. 203, p. 105355, Aug. 2021, doi: 10.1016/j.catena.2021.105355.
[18] H. Wang, S. Xu, H. Xu, Z. Wu, T. Wang, and C. Ma, "Rapid prediction of urban flood based on disaster-breeding environment clustering and Bayesian optimized deep learning model in the coastal city," *Sustainable Cities and Society*, vol. 99, p. 104898, Dec. 2023, doi: 10.1016/j.scs.2023.104898.
[19] H. Hairani, T. Widiyaningtyas, and D. D. Prasetya, "Addressing Class Imbalance of Health Data: A Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies," *JOIV : International Journal on Informatics Visualization*, vol. 8, no. 3, p. 1310, Sep. 2024, doi: 10.62527/joiv.8.3.2283.
[20] C. J. Okolie and J. L. Smit, "A systematic review and meta-analysis of Digital elevation model (DEM) fusion: pre-processing, methods and applications," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 188, pp. 1–29, Jun. 2022, doi: 10.1016/j.isprsjprs.2022.03.016.
[21] S. C. Kulkarni and P. P. Rege, "Pixel level fusion techniques for SAR and optical images: A review," *Information Fusion*, vol. 59, pp. 13–29, Jul. 2020, doi: 10.1016/j.inffus.2020.01.003.
[22] C. I. Donatti *et al.*, "Global hotspots of climate-related disasters," *International Journal of Disaster Risk Reduction*, vol. 108, p. 104488, Jun. 2024, doi: 10.1016/j.ijdrr.2024.104488.
[23] K. Breinl, D. Lun, H. Müller-Thomy, and G. Blöschl, "Understanding the relationship between rainfall and flood probabilities through combined intensity-duration-frequency analysis," *Journal of Hydrology*, vol. 602, p. 126759, Nov. 2021, doi: 10.1016/j.jhydrol.2021.126759.
[24] C. Mei *et al.*, "Flood risk related to changing rainfall regimes in arterial traffic systems of the Yangtze River Delta," *Anthropocene*, vol. 35, p. 100306, Sep. 2021, doi: 10.1016/j.ancene.2021.100306.
[25] M. Rahman *et al.*, "Flooding and its relationship with land cover change, population growth, and road density," *Geoscience Frontiers*, vol. 12, no. 6, p. 101224, Nov. 2021, doi: 10.1016/j.gsf.2021.101224.
[26] R. M. F. Hannum, I. P. Santikayasa, and B. D. Dasanto, "Evaluation of Flood Hazard Potency in Jakarta based on Multi-criteria Analysis," *Agromet*, vol. 36, no. 2, pp. 101–111, Dec. 2022, doi: 10.29244/j.agromet.36.2.101-111.
[27] K. Su, R. Yang, Q. Cui, and T. Wang, "The geographic distance of independent directors and stock price crash risk: Evidence from China," *Res. Int. Bus. Finance*, vol. 69, p. 102270, Apr. 2024, doi: 10.1016/j.ribaf.2024.102270.
[28] A. Nurkholis and I. S. Sitanggang, "A spatial analysis of soybean land suitability using spatial decision tree algorithm," in *Sixth International Symposium on LAPAN-IPB Satellite*, Dec. 2019, doi: 10.1117/12.2541555.
[29] V. Purwayoga, S. Yuliyanti, A. Nurkholis, H. Gunawan, S. Sokid, and N. Kartini, "Distribution Model of Personal Protective Equipment (PPE) Using the Spatial Dominance Test and Decision Tree Algorithm," *JOIV : International Journal on Informatics Visualization*, vol. 8, no. 3, p. 1445, Sep. 2024, doi: 10.62527/joiv.8.3.2471.
[30] J. Zhang, W. Wang, X. Jiang, W.-S. Ku, and H. Lu, "An MBR-Oriented Approach for Efficient Skyline Query Processing," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, IEEE, Apr. 2019, pp. 806–817, doi: 10.1109/ICDE.2019.00077.
[31] C. Li *et al.*, "Mining area skyline objects from map-based big data using Apache Spark framework," *Array*, vol. 25, p. 100373, Mar. 2025, doi: 10.1016/j.array.2024.100373.
[32] S. Heo, W. Sohn, S. Park, and D. K. Lee, "Multi-hazard assessment for flood and Landslide risk in Kalimantan and Sumatra: Implications for Nusantara, Indonesia's new capital," *Heliyon*, vol. 10, no. 18, p. e37789, Sep. 2024, doi: 10.1016/j.heliyon.2024.e37789.

[33] Q. Su *et al.*, "Landslide Susceptibility Zoning Using C5.0 Decision Tree, Random Forest, Support Vector Machine and Comparison of Their Performance in a Coal Mine Area," *Frontiers in Earth Science*, vol. 9, Dec. 2021, doi: 10.3389/feart.2021.781472.
[34] D. W. Nugraha, A. A. Ilham, A. Achmad, and A. Arief, "Performance Improvement of Deep Convolutional Networks for Aerial Imagery Segmentation of Natural Disaster-Affected Areas," *JOIV : International Journal on Informatics Visualization*, vol. 7, no. 4, p. 2321, Dec. 2023, doi: 10.62527/joiv.7.4.1383.
[35] C. Rojas, R. Linfati, R. F. Scherer, and L. Pradenas, "Using Geopandas for locating virtual stations in a free-floating bike sharing system," *Heliyon*, vol. 9, no. 1, pp. 1-14, Jan. 2023, doi: 10.1016/j.heliyon.2022.e12749.
[36] R. A. Azdy and F. Darnis, "Use of Haversine Formula in Finding Distance Between Temporary Shelter and Waste End Processing Sites," in *3rd Forum in Research, Science, and Technology (FIRST 2019)*, vol. 1500, no. 1, p. 012104, Apr. 2020, doi: 10.1088/1742-6596/1500/1/012104.
[37] Y. Wu, Y. Ding, and J. Feng, "SMOTE-Boost-based sparse Bayesian model for flood prediction," *EURASIP Journal on Wireless Communications and Networking*, vol. 2020, no. 1, p. 78, Dec. 2020, doi: 10.1186/s13638-020-01689-2.
[38] K. M. Sujon *et al.*, "The Effects of Imbalanced Datasets on Machine Learning Algorithms in Predicting Student Performance," *JOIV : International Journal on Informatics Visualization*, vol. 8, no. 3–2, p. 1599, Nov. 2024, doi: 10.62527/joiv.8.3-2.2449.

APPENDIX

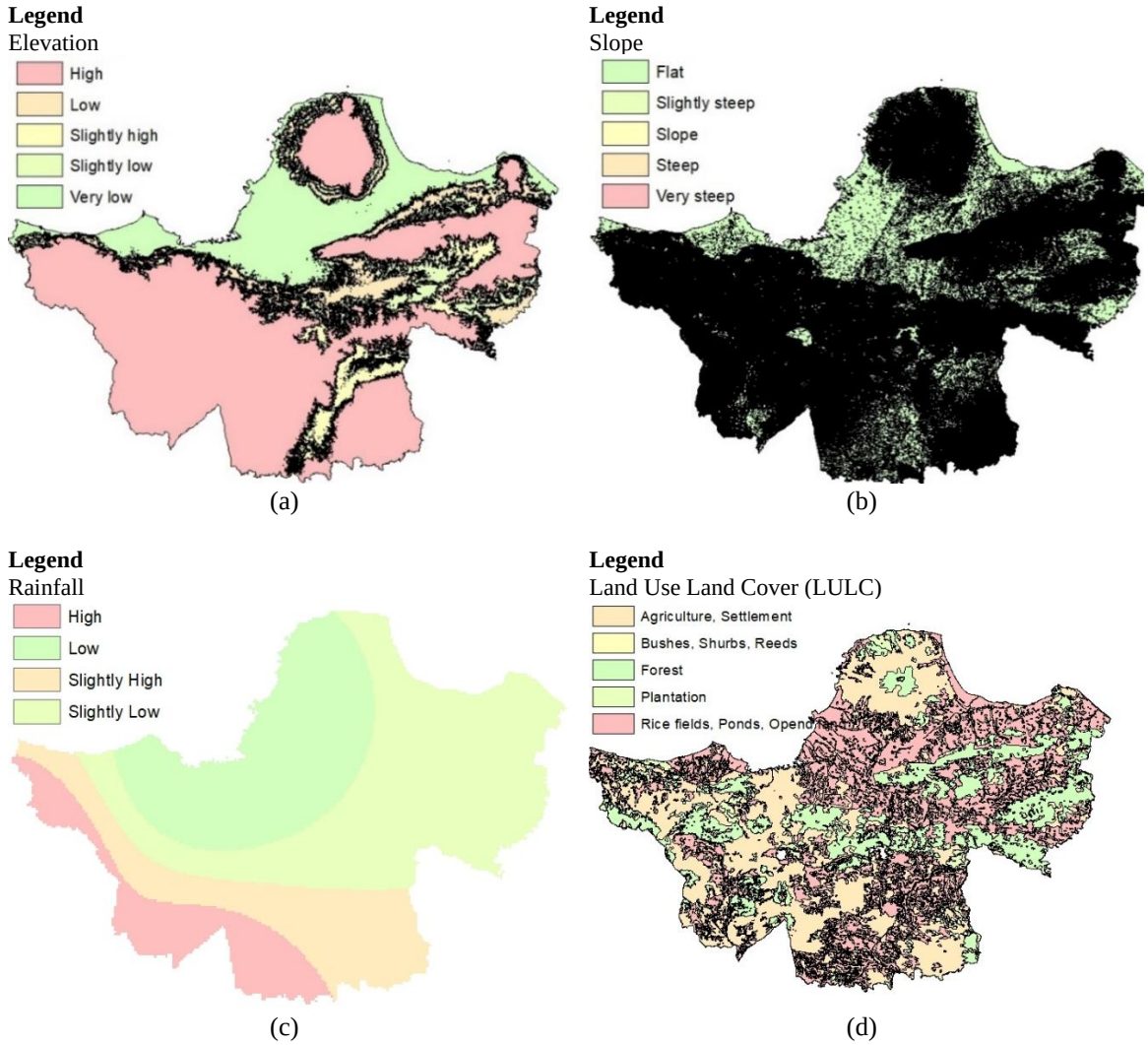


Figure 2. Preprocessing results of factors influencing flood disasters; (a) elevation, (b) slope, (c) rainfall, and (d) LULC

BIOGRAPHIES OF AUTHORS

Siti Yuliyanti earned her bachelor's degree in Informatics Engineering from STMIK Bandung in 2010 and completed her master's degree in Computer Science at IPB University, Bogor, in 2016. She is currently a lecturer in the Department of Informatics at Siliwangi University. Her research interests include multimodal natural language processing, text mining, and data mining. She can be contacted at email: sitiyuliyanti@unsil.ac.id.

Vega Purwayoga was born in March 1995 in Indonesia. He obtained his bachelor's degree in Informatics Engineering from Ahmad Dahlan University, Yogyakarta, in 2017, and completed his master's degree in Computer Science at IPB University, Bogor, in 2019. He is currently a lecturer in the Department of Informatics at Siliwangi University and a member of the Spatial Intelligence for Climate and Disaster Resilience (SPINTER) research group. His research interests include geoinformatics, GeoAI, data mining, text mining, and data visualization. He can be contacted at email: vega.purwayoga@unsil.ac.id.

Andi Nur Rachman graduated with a bachelor's degree in Informatics Engineering from Universitas Siliwangi, Tasikmalaya, in 2009. He pursued a master's degree in Informatics at Institut Teknologi Bandung and completed it in 2015. Currently, he serves as a lecturer in the Information Systems Study Program at Universitas Siliwangi. His research interests focus on information systems. He can be contacted at email: andy.rachman@unsil.ac.id.

Zakwan Gusnadi earned his bachelor's degree in Civil Engineering from Universitas Pendidikan Indonesia, Bandung, in 2018, and subsequently completed his master's degree in Civil Engineering at Universitas Katolik Parahyangan, Bandung, in 2021. He is currently a lecturer in the Civil Engineering Study Program at Universitas Siliwangi and a member of the Spatial Intelligence for Climate and Disaster Resilience (SPINTER) research group. His research interests focus on geotechnics. He can be contacted at email: zakwangusnadi@unsil.ac.id.