BAREKENG: Journal of Mathematics and Its Applications
June 2026 Volume 20 Issue 2 Page 1019–1034
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss2pp1019-1034

OPTIMIZING LANDSLIDE SUSCEPTIBILITY MAPPING IN CENTRAL SULAWESI WITH RECURSIVE FEATURE ELIMINATION AND RANDOM FOREST ALGORITHM

Indra Rivaldi Siregar [1*], **Anik Djuraidah** [2], **Agus Mohamad Soleh** [3]

[1,2,3]*School of Data Science, Mathematics, and Informatics, IPB University*
Jln. Meranti, Bogor, 16680, Indonesia

*Corresponding author's e-mail: *siregarindra@apps.ipb.ac.id*

Article Info

Article History:

Received: 28th April 2025
Revised: 10th June 2025
Accepted: 7th October 2025
Available online: 26th January 2026

Keywords:

Central Sulawesi;
Landslide;
Mitigation;
Random forest;
Recursive feature elimination.

ABSTRACT

Landslides are among the most destructive natural hazards, causing severe casualties, economic losses, and environmental degradation. Central Sulawesi, characterized by active tectonics such as the Palu-Koro fault, is highly susceptible to landslides, as tragically demonstrated in 2018. Therefore, developing accurate landslide susceptibility maps is essential to support comprehensive landslide mitigation efforts in this region. While machine learning, particularly Random Forest (RF), has proven highly effective for landslide modeling, previous studies around Palu have often overlooked model simplification through feature selection and hyperparameter optimization. This study proposes an integrated approach combining RF with Recursive Feature Elimination (RFE) to reduce model complexity and enhance predictive accuracy. This research utilizes 498 landslide events with fifteen conditions, including topography, environment, geology, and anthropogenic influences. The RFE-RF model achieves superior classification performance, with accuracy, balanced accuracy, and F1-scores exceeding 0.81, outperforming the RF without RFE and Logistic Regression baselines. These findings underscore the urgent need to integrate feature selection methods such as RFE into landslide modeling frameworks to improve predictive accuracy. High accuracy enables government authorities and stakeholders to develop more targeted and effective mitigation priorities. Spatial analysis indicates that Donggala, Palu, and Sigi are the most critical areas requiring prioritized mitigation, with over 9% of their territories classified as highly susceptible. Feature importance analysis reveals that elevation, slope, and land cover are the most influential factors. This study suggests that mitigation efforts should focus on the hills and mountainous areas on both sides of the Palu Valley, with recommended strategies emphasizing land cover management practices, such as reforestation, to enhance slope stability and reduce landslide risk.

How to cite this article:

I. R. Siregar, A. Djuraidah and A. M. Soleh., "OPTIMIZING LANDSLIDE SUSCEPTIBILITY MAPPING IN CENTRAL SULAWESI WITH RECURSIVE FEATURE ELIMINATION AND RANDOM FOREST ALGORITHM", *BAREKENG: J. Math. & App.*, vol. 20, no. 2, pp. 1019-1034, Jun, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · **Open Access**

1. INTRODUCTION

Landslides are among the most destructive geological hazards, resulting from the downslope movement of soil, rock, and other earth materials driven by gravitational forces. This disaster not only results in loss of life but also substantial economic losses and widespread social and environmental impacts [1], [2]. One of the most critical mitigation efforts to reduce the risks posed by landslides is the development of landslide susceptibility mapping, which involves delineating areas based on their likelihood of experiencing landslide events in the future.

Alongside advancements in technology and geospatial science, landslide susceptibility mapping methods have evolved from conventional opinion-based and statistical approaches toward the broader application of machine learning techniques, which offer more adaptive and accurate predictive capabilities [3]. In a comprehensive review, [3] summarized numerous studies that applied machine learning and deep learning algorithms for landslide susceptibility mapping, concluding that tree-based ensemble algorithms, particularly Random Forest (RF), consistently delivered superior classification performance, offering high accuracy and robustness against non-linear data characteristics [4]. The effectiveness of RF for landslide susceptibility mapping has also been demonstrated by [5], who reported that the combination of RF and cross-validation techniques outperformed 40 other models in classifying landslide and non-landslide areas in Kendari City. This further solidifies RF's position as one of the most reliable methods for landslide classification tasks.

On the other hand, Central Sulawesi is one of Indonesia's regions with a high susceptibility to landslides. The presence of the Palu-Koro active fault, with a slip rate ranging from 20–40 mm per year based on GPS observations up to 2016, is a key contributing factor to landslide hazards in the region [6], [7]. In 2018, the region experienced one of its most devastating landslide incidents, which caused significant damage in Palu City, Donggala Regency, and the surrounding areas.

In Central Sulawesi, various landslide susceptibility mapping efforts have been undertaken as part of disaster mitigation strategies, including a study by [8] that applied the weighted overlay technique. While the method is frequently employed and relatively simple to implement, its weighting process is inherently subjective, as it relies heavily on expert judgment or literature references, which can affect the objectivity and accuracy of the final mapping results [9]. In a more recent study, [10] utilized the RF algorithm for landslide susceptibility mapping near Palu City; however, the study did not incorporate hyperparameter optimization nor implement feature selection methods to simplify the model structure.

In response to these gaps, this study develops a landslide susceptibility model for Palu and its surrounding areas using Random Forest (RF) with hyperparameter tuning combined with the Recursive Feature Elimination (RFE) technique. RFE is a feature selection method that iteratively discards less significant variables based on their importance values derived from the RF algorithm [11]. This is particularly important because the absence of hyperparameter tuning can result in suboptimal model calibration, whereas omitting feature selection may retain redundant or irrelevant variables, leading to unnecessarily complex models and potentially diminished classification accuracy [12]. To date, the integration of RF with RFE for landslide susceptibility mapping in Palu and its surrounding areas has not been explored. This methodology is expected to enhance predictive accuracy and reduce model complexity, thereby providing more reliable and actionable information to support targeted landslide mitigation strategies in the region.

2. RESEARCH METHODS

2.1 Random Forest (RF)

Random Forest (RF) is an ensemble method that combines multiple decision trees to improve predictive accuracy and stability. RF implements the principle of bootstrap aggregating (bagging) as shown in Fig. 1. Each tree is constructed from a randomly selected subset of data through a bootstrap mechanism (sampling with replacement), and the predictions of all trees are then aggregated to produce the final decision [4]. This technique reduces the risk of overfitting and yields a more stable model compared to using a single decision tree.

RF is an extension of the bagging method, with the primary difference being the predictor selection process during tree construction. In pure bagging, each tree considers all available predictors. In contrast, RF

randomly selects a subset of predictors for consideration at each split. This approach aims to prevent the trees from becoming too similar, as can occur in standard bagging, thereby reducing inter-tree correlation and improving model generalization. In classification tasks, the final prediction is determined by majority voting across all decision trees.

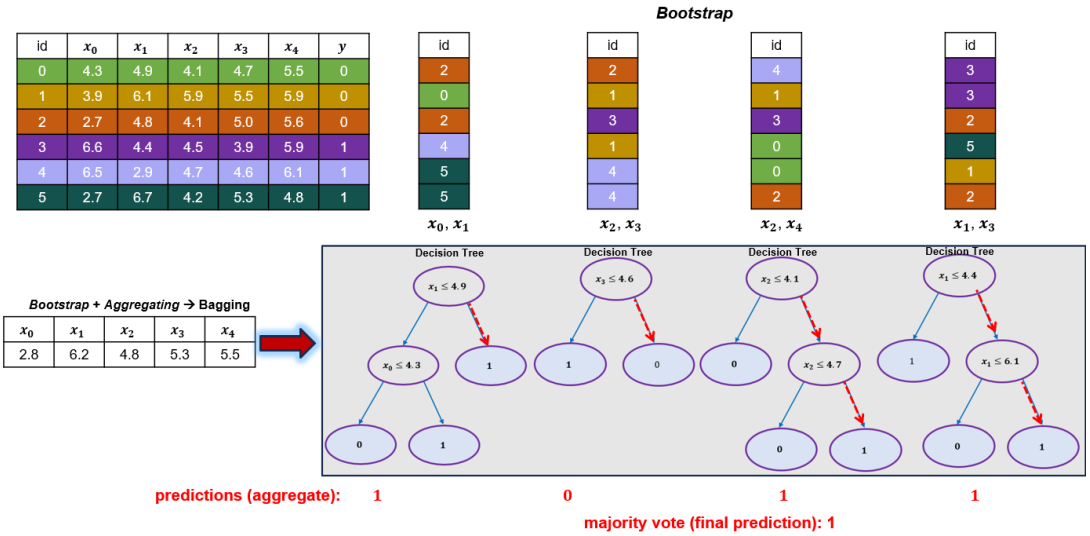


Figure 1. Illustration of Random Forest Algorithm

The construction of each decision tree in RF depends on an optimal splitting criterion, typically based on the Gini index, entropy, or other measures. The Gini index for a dataset D with k classes is calculated as defined in Eq. (1):

$$Gini\ (D) = 1 - \sum_{i=1}^{k} p_i^2, \tag{1}$$

where p_i denotes the probability of class i, a Gini value of 0.5 indicates a perfectly balanced class distribution, while zero reflects complete purity within a node.

Consider a node in a dataset containing 10 instances, where each instance represents an observation characterized by one or more predictor variables. In this example, based on a certain predictor, 6 instances belong to class A, and 4 instances belong to class B. The probability p_i of an instance belonging to each class is calculated as the proportion of instances of that class in the node, following Eq. (2):

$$p_A = \frac{6}{10} = 0.6, p_B = \frac{4}{10} = 0.4. \tag{2}$$

The Gini index for this node is then:

$$Gini\ (D) = 1 - (0.6^2 + 0.4^2) = 0.48. \tag{3}$$

Based on Eq. (3), a value of 0.48 indicates that the node is relatively impure, containing a mixture of instances from both classes. A value of 0 corresponds to a perfectly pure node, while higher values reflect greater heterogeneity. During tree construction, the algorithm selects splits that minimize the Gini index, producing nodes that are more homogeneous and improving model classification accuracy.

2.2 Recursive Feature Elimination (RFE)

Recursive Feature Elimination (RFE) is a feature selection technique designed to identify the most relevant predictors for distinguishing between target classes [11]. It iteratively removes the least important feature based on importance scores from a RF classifier, producing a reduced feature set while maintaining overall predictive accuracy. In each iteration, a new RF model is trained, and cross-validation is used to evaluate its accuracy. The feature with the lowest importance is removed, and the process continues recursively until the desired number of features is reached. The cross-validated accuracy at each iteration provides a quantitative measure to identify the optimal subset of features.

Illustrative example: consider a dataset with m predictors $[x_1, x_2, \ldots, x_m]$. RFE is configured to remove one predictor at each iteration, starting from all m predictors and continuing until only a single predictor remains.

1. Iteration 1: All m predictors are used to train the Random Forest model. Feature importance scores $\{I_1, I_2, \ldots, I_m\}$ are computed, and k-fold cross-validation is performed to calculate the average accuracy A_1. The predictor with the lowest importance, say x_l, is removed.
2. Iteration 2: The model is retrained with the remaining $m - 1$ predictors. Importance scores and cross-validated accuracy A_2 are calculated. The least important feature is removed.
3. Iteration j: Only one feature remains, and the corresponding cross-validated accuracy A_j is calculated.

This iterative process produces a sequence of accuracies $[A_1, A_2, \ldots, A_j]$ corresponding to decreasing feature subsets. By visualizing this sequence, the optimal subset of features can be identified as the smallest set that achieves the highest or near-highest cross-validated accuracy, balancing predictive performance with model simplicity.

2.3 Data and Pre-Processing

The study area encompasses four regencies in Central Sulawesi, covering a total of 1,491.68 km² (Fig. 2). Landslide inventory data, comprising 498 events induced by the 7.5 Mw earthquake in September 2018, were obtained from the USGS database [13]. Non-landslide points are generated randomly following the approach of [14], with a minimum spatial separation of 30 m to prevent points from falling within the same raster cell, which is consistent with the resolution of the conditioning factor data. This random sampling strategy is computationally efficient, reduces selection bias, and ensures that non-landslide points uniformly cover the full range of conditioning factors across the study area.

Unlike approaches that restrict non-landslide points to areas of low slope, this study intentionally allows them to occur even in steep terrain. This reflects the fact that steep slopes do not always fail; slope instability is governed by a combination of conditioning factors rather than slope alone. Including non-landslide points in high-slope areas enables the model to better discriminate which factor combinations truly lead to landslides, rather than overestimating slope as a single dominant trigger. We utilize fifteen conditioning factors classified into four categories: topography, environment, anthropology, and geology (Table 1).

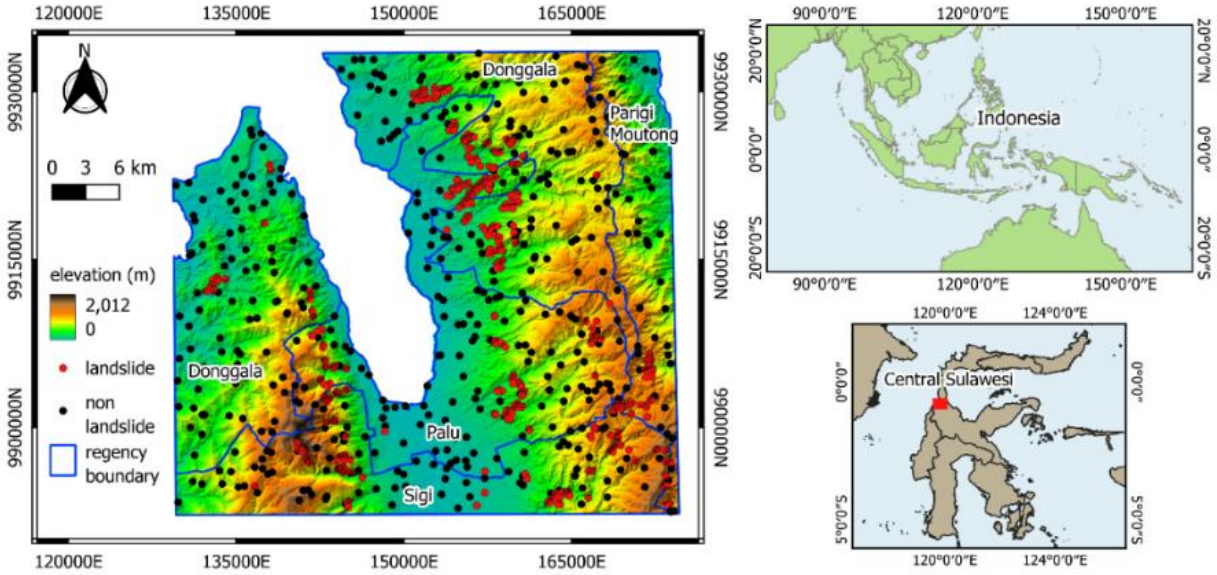


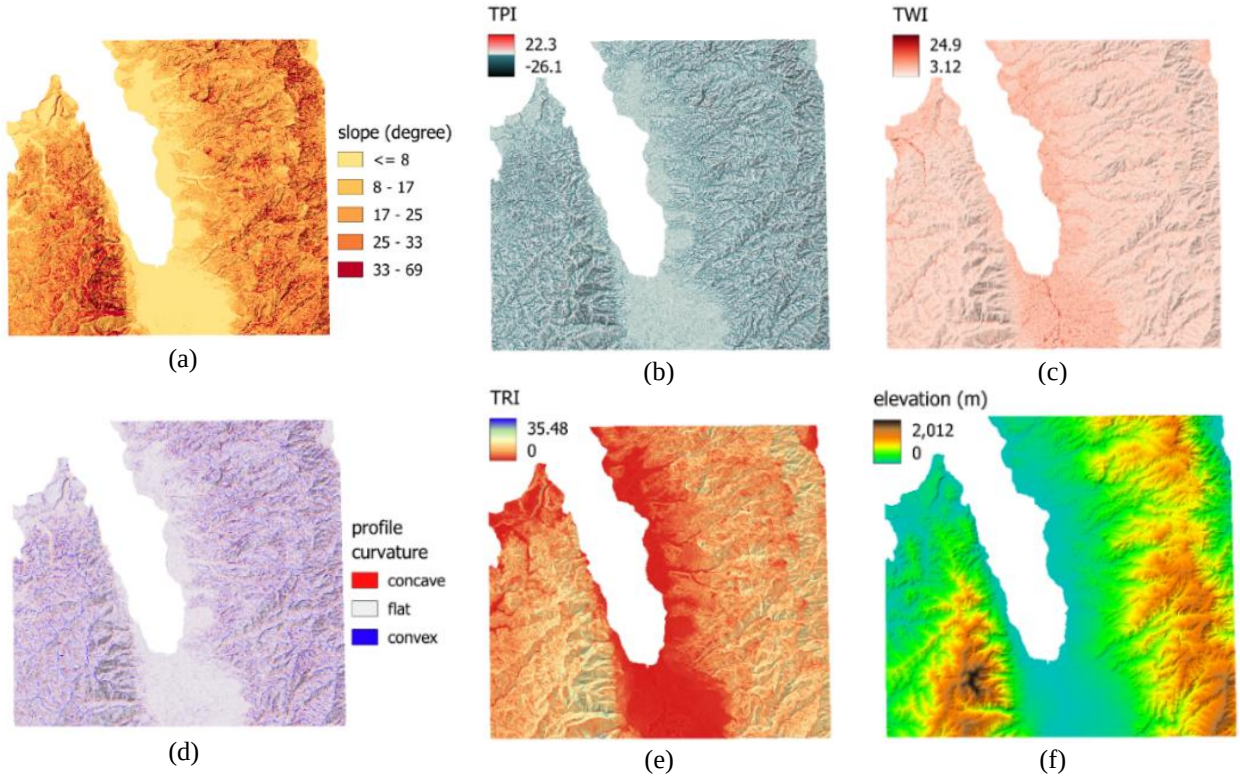
Figure 2. Study Area in Central Sulawesi

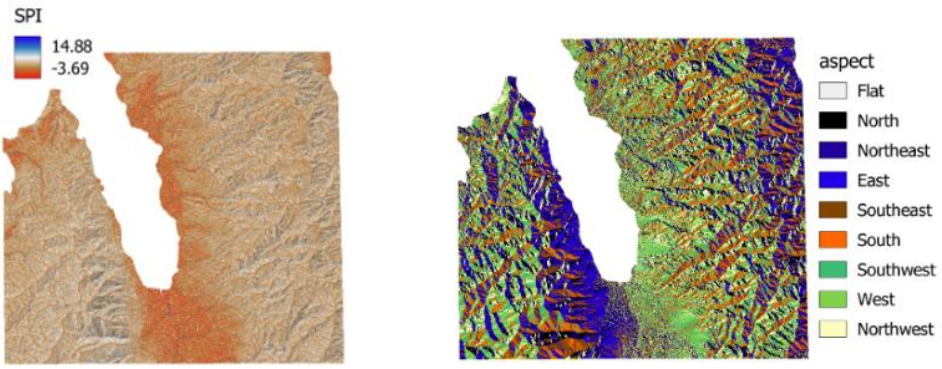
Table 1. Landslide Conditioning Factors

Category	Data Source	Data Type	Variable (Factor)	Scale/Resolution
Topography	Digital Elevation Model (SRTM)	Raster	Elevation Slope Profile Curvature Aspect TWI TPI SPI TRI	±30m
Environment	LANDSAT 8 OLI	Raster	NDVI	±30m
	OpenStreetMap	Vector	Distance to river River density	±30m -
Geology	ESDM Republic of Indonesia KLHK Republic of Indonesia	Vector	Formation Distance to fault	1:250,000 ±30m
Anthropology	KLHK Republic of Indonesia OpenStreetMap	Vector	Land cover Road density	±30m -

Topographic factors (Fig. 3), including slope, topographic position index (TPI), topographic wetness index (TWI), profile curvature, terrain ruggedness index (TRI), elevation, stream power index (SPI), and aspect, are derived from SRTM data (±30 m) using GIS-based processing [15]. Slope is calculated using a numerical differentiation technique that estimates elevation changes by considering the neighboring cells within the DEM grid. The hilly and mountainous regions on the western and eastern sides of the Palu Valley exhibit extreme slopes, with values ranging from 33 to 69 degrees. These areas also show positive values of the TPI, indicating that they are located at higher elevations relative to their surroundings. In contrast, the central part of the study area—namely, the Palu Valley, which includes Palu and Sigi—generally has slopes of less than 8 degrees, along with lower elevation and TPI values compared to adjacent areas.

The Palu Valley also predominantly shows low values of the TRI, which reflects smooth terrain. This contrasts sharply with the surrounding hilly and mountainous areas, which exhibit higher TRI values, indicating rougher topography [5]. Additionally, TWI, which represents the potential for water accumulation [15], tends to be higher in the Palu Valley, suggesting greater water retention capacity.





(g) (h)

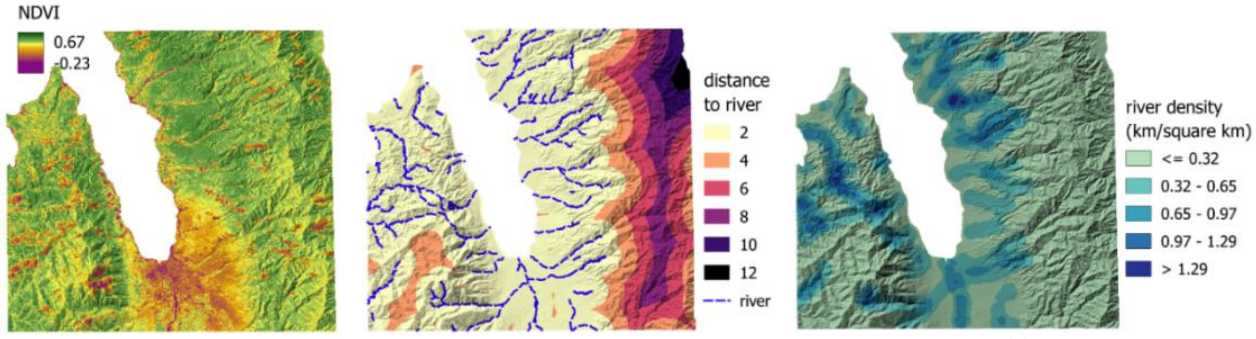
Figure 3. Topographic Factors
(a) Slope, (b) TPI, (c) TWI, (d) Profile Curvature, (e) TRI, (f) Elevation, (g) SPI, (h) Aspect

Profile curvature indicates whether a slope accelerates or decelerates the flow of water and materials such as soil and sediment. Positive values represent convex surfaces, negative values indicate concave surfaces, and values close to zero suggest relatively flat terrain [16]. The Palu Valley mostly exhibits flat curvature, while the adjacent hills and mountains are dominated by convex and concave forms.

SPI is a composite topographic attribute used to measure the erosive power of surface water flow [17]. It assumes that water discharge is proportional to the specific catchment area. Higher SPI values—found in the western and eastern hills of the Palu Valley—suggest stronger erosive forces that may threaten soil stability. Finally, the aspect, which describes the compass direction of the steepest slope (i.e., the direction of maximum elevation change) [15], shows that the study area is predominantly characterized by slopes facing southwest and west.

Environmental factors (Fig. 4) consist of the normalized difference vegetation index (NDVI) derived from LANDSAT 8 OLI imagery [18], distance to the river, and river density. NDVI is a remote sensing index employed to monitor variations in land cover by quantifying vegetation greenness on the Earth's surface. In general, the more positive the NDVI value, the healthier and denser the vegetation. Within the study area, mountainous and hilly regions tend to exhibit relatively high and positive NDVI values.

Water flow also plays a crucial role in landslide classification [19]. Therefore, this study incorporates two potential hydrological variables: distance to the river and river density. Distance to the river is categorized into six buffer zones at 2 km intervals, while river density is calculated as the total length of rivers divided by the pixel area. River systems in this region originate from the mountainous areas in the west and east, and flow toward the central part of the study area.

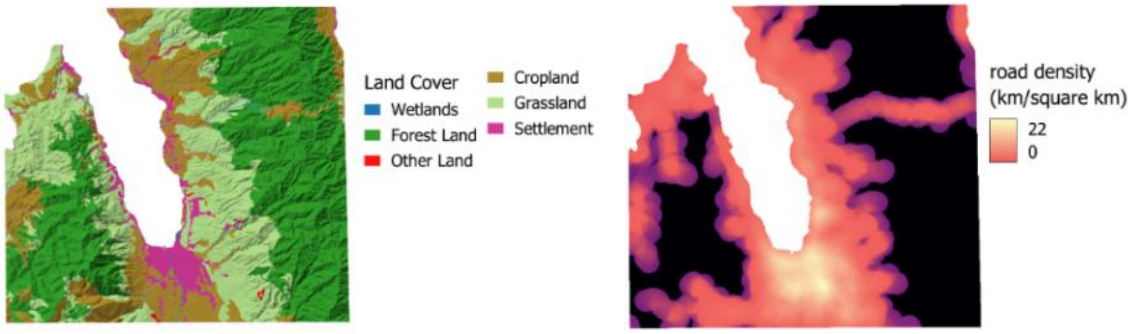


(a) (b) (c)

Figure 4. Environmental Factors
(a) NDVI, (b) Distance to River, (c) River Density

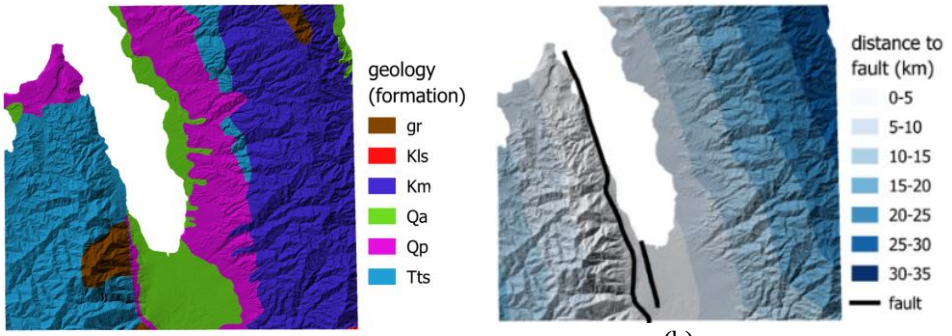
The anthropological factor is represented by two variables: land cover and road density (Fig. 5). Land cover is reclassified into six categories based on the guidelines provided by the intergovernmental panel on climate change (IPCC) [20]. This reclassification aims to reduce the number of unique land cover values and to eliminate categories that are not represented by either landslide or non-landslide samples. The central part of the study area, particularly the city of Palu, is predominantly characterized by settlements, indicating a high level of anthropogenic activity. In contrast, the western and eastern margins of the region are primarily

covered by forest land and grassland. Consistent with the presence of settlements, the lowland areas of Palu and Sigi also exhibit relatively high road density compared to the surrounding regions.



(a) (b)

Figure 5. Anthropological Factors
(a) Land Cover, (b) Road Density



(a) (b)

Figure 6. Geological factors
(a) Formation, (b) Distance to Fault

Geological factors (Fig. 6) encompass six formation classes and distance to fault lines. Previous studies [21], [22] have also incorporated geological factors into landslide and non-landslide classification models, highlighting their relevance in capturing terrain stability and susceptibility. Among these factors, geological formations—representing the classification and distribution of rock types—are particularly influential, as they reflect the underlying lithological diversity across the study area. Different rock types exhibit varying levels of resistance to weathering, erosion, and slope failure, thereby exerting differential impacts on landslide potential.

In this study, geological information is further enriched by the inclusion of fault line data, which serves as a proxy for seismic activity—a well-known landslide-triggering mechanism. To systematically quantify proximity to potential seismic sources, the distance to the nearest fault is categorized into seven buffer zones at 5 km intervals. This classification facilitates a more refined spatial analysis of the influence of tectonic factors on landslide occurrence.

Data from various sources are reprojected to a specific coordinate system, namely UTM Zone 51S (Fig. 7). Subsequently, vector-format data are converted into raster format. All rasterized datasets are then stacked to ensure spatial alignment and consistency for further analysis. The data are randomly divided (70% train, 30% test), and this process is repeated 5 times to ensure more comprehensive results. In each iteration of the data split, both models undergo hyperparameter tuning using k-fold cross-validation [23] on the training data before evaluation on the test data. We compare the performance of the RF model without RFE and the RF model with RFE (RFE-RF) using accuracy, balanced accuracy, and F1-score metrics on the test data. As a baseline, we also implement a Logistic Regression (LR) [21] model for comparison in this study.

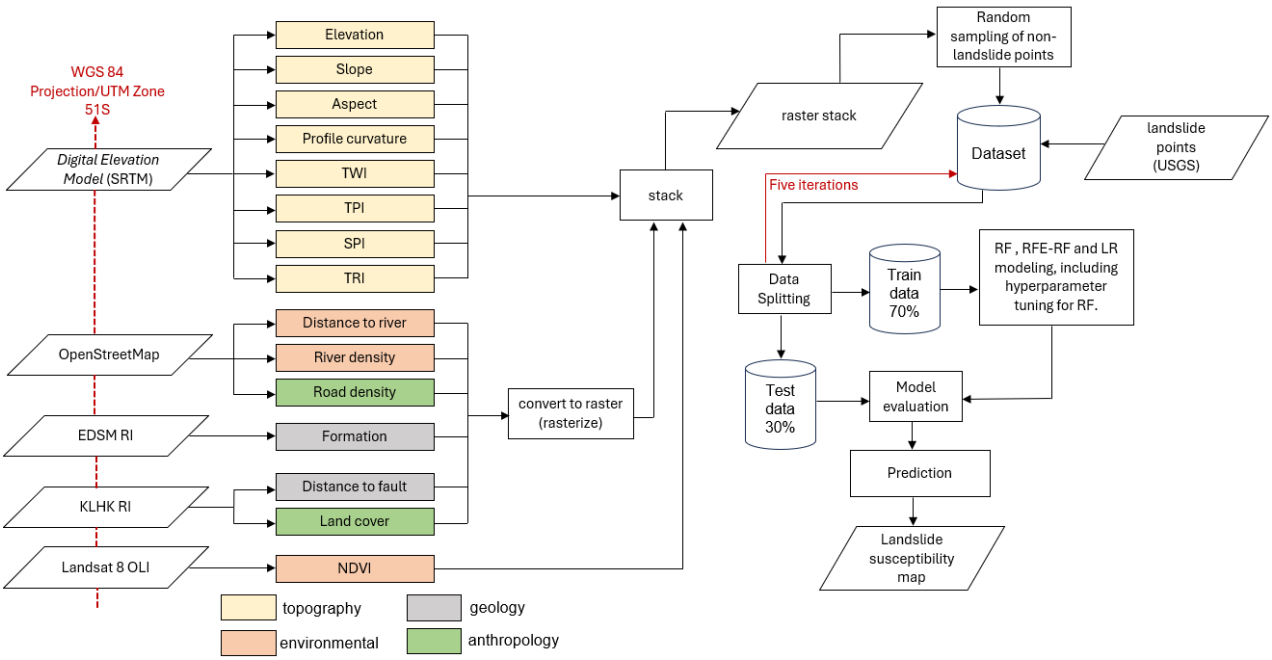


Figure 7. Workflow of Data Preprocessing and Modeling

Table 2 summarizes the hyperparameters explored for tuning the Random Forest (RF) model. Three key hyperparameters are considered: the *mtry*, the *trees*, and the *min_n*. The *mtry*, representing the number of predictor variables randomly selected at each split, is tested from 1 to 4 to control the diversity among individual trees. The *trees*, defining the total number of trees in the forest, are varied from 5 to 500 to balance predictive stability and computational efficiency.

Table 2. Hyperparameters of the RF Model

Hyperparameter
mtry: [1, 2, 3, 4]
trees: [5, 10, 15, 20, 25, …, 500]
min node size: [2, 3, 5, 7, 10, 12, 15, 17, 20]

Finally, the *min_n*, which specifies the minimum number of samples required in a node to attempt a split, is explored from 2 to 20 to regulate tree growth and prevent overfitting. A grid search combined with k-fold cross-validation is used to identify the optimal combination of these hyperparameters.

3. RESULTS AND DISCUSSION

3.1 Exploratory Data Analysis (EDA)

In the initial stage, we perform an exploratory analysis of all available landslide conditioning factors. For continuous features, Pearson correlation analysis (Fig. 8) reveals no highly correlated pairs ($r > 0.8$); however, several feature pairs exhibit moderate correlations ($r > 0.5$), such as road density–elevation, TWI–slope, and TWI–TPI, which may indicate potential redundancy. Although previous study [24] suggests that the RF algorithm can handle multicollinearity, simplifying the model by removing redundant features remains a prudent strategy to enhance model interpretability and computational efficiency.

Subsequently, a visual analysis using boxplots (Fig. 9) provides initial insights into the discriminative ability of each continuous feature in separating landslide and non-landslide classes. Features such as elevation, slope, TWI, and river density exhibit distinct differences in median values across classes, indicating a strong potential to improve classification performance. In contrast, features such as SPI and TPI show relatively similar distributions across classes, suggesting limited discriminative capability.

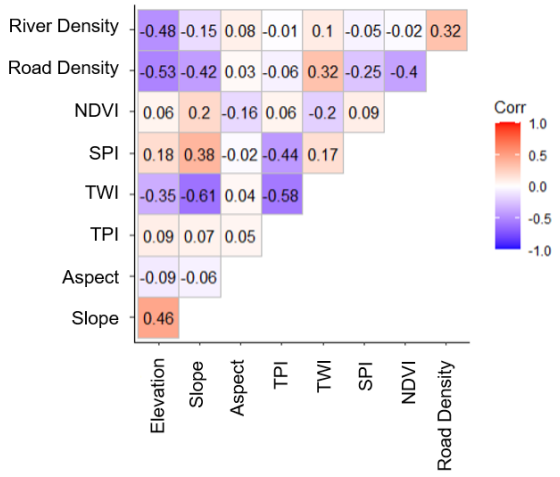


Figure 8. Pearson Correlation Heatmap of Continuous Features

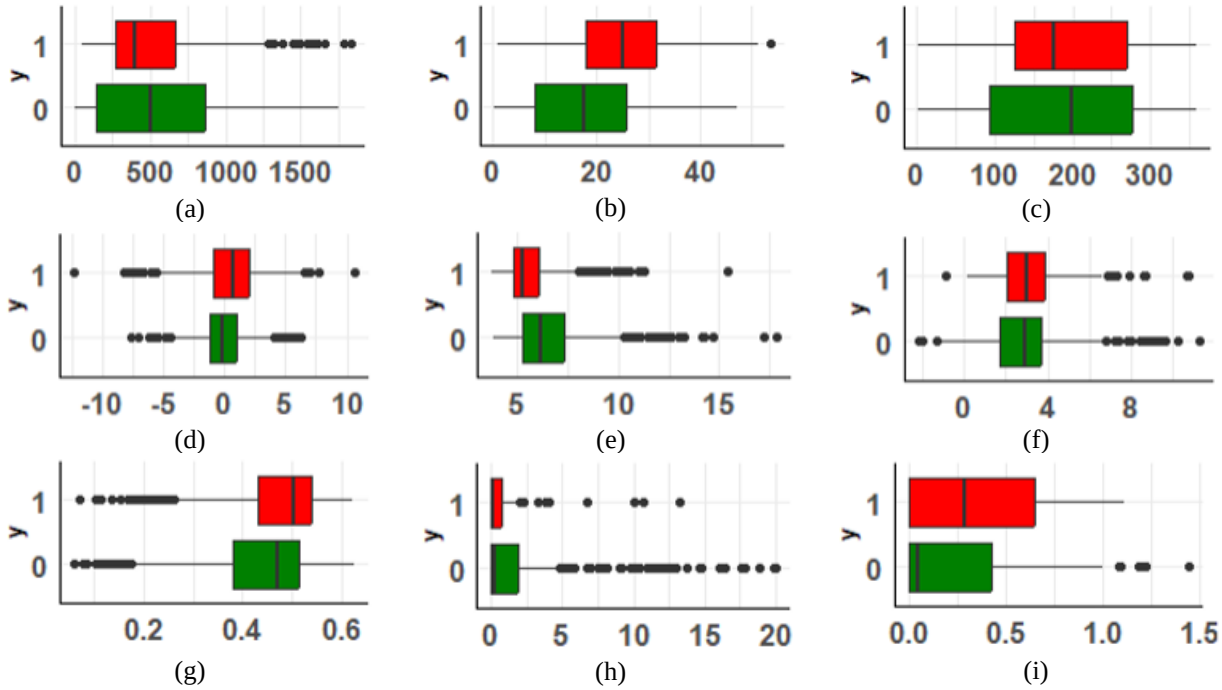


Figure 9. Boxplot of Continuous Features Categorized by Landslide (y=1) and Non-Landslide (y=0) Classes (a) Elevation, (b) Slope, (c) Aspect, (d) TPI, (e) TWI, (f) SPI, (g) NDVI, (h) Road Density, (i) River Density

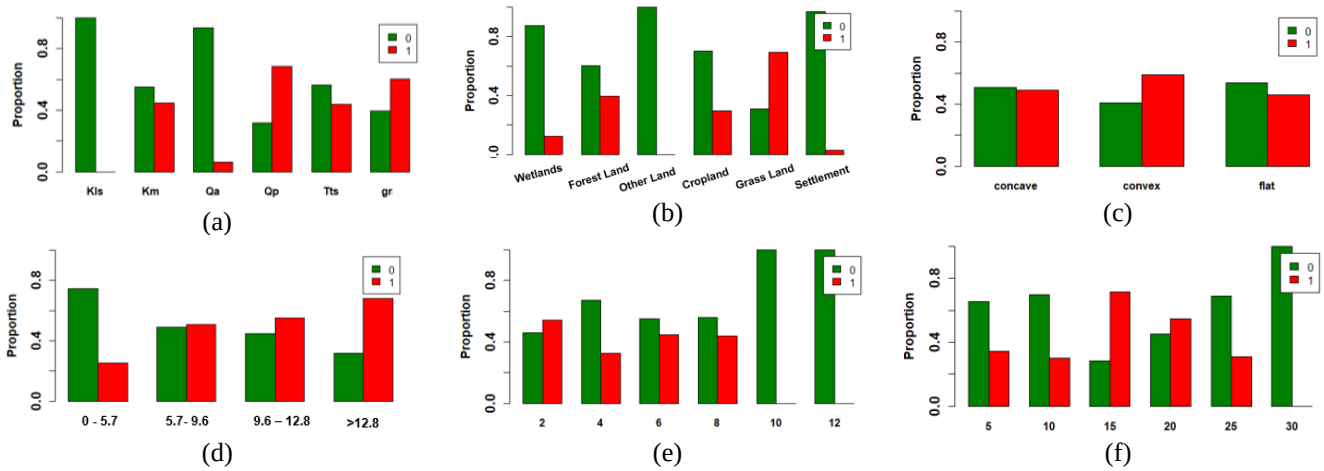


Figure 10. Barplot of Categorical Features Categorized by Landslide (y=1) and Non-Landslide (y=0) Classes (a) Formation, (b) Land Cover, (c) Profile Curvature, (d) TRI, (e) Distance to River, (f) Distance to Fault

For categorical features, the barplot in Fig. 10 reveals that landslide occurrences in the study area are more frequently associated with Qp and GR geological formations, grassland land cover, and a TRI value exceeding 9.6. Conversely, profile curvature appears to have limited influence on classification, as there is no substantial difference in the proportion of landslide and non-landslide cases across its categories (concave, convex, and flat).

3.2 Modelling and Landslide Susceptibility Map

We incorporate a feature selection process using RFE into the RF model applied to the training data, with the evaluation using 5-fold cross-validation to identify the optimal combination of features from all available features. The results of the RFE-RF are presented in Fig. 11, which shows that the highest classification performance—indicated by the red point representing the maximum accuracy—is achieved using nine features: elevation, slope, distance to fault, road density, land cover, TWI, formation, TRI, and river density. These nine features represent the most significant factors for modeling landslide susceptibility for this study.

The RFE-RF model is subsequently compared with the RF model without RFE and the LR model on the test data, as presented in Table 3. The results show that the RFE-RF model achieves the best performance, with all metric values consistently exceeding 0.81 and outperforming both the RF and LR models on the test data. These findings indicate that the application of RFE contributes to selecting more relevant features for distinguishing between landslide and non-landslide classes. As highlighted by [25], RFE can assist in eliminating redundant features or those that are highly correlated. In this study, for instance, the correlation between TPI and TWI is approximately -0.58 (Fig. 8), and RFE advises the exclusion of the TPI. In the boxplot visualization shown in Fig. 9, TWI values exhibit a sharper separation univariately between landslide and non-landslide classes compared to TPI. Thus, in this context, RFE effectively discards less relevant features and reduces potential redundancy among variables.

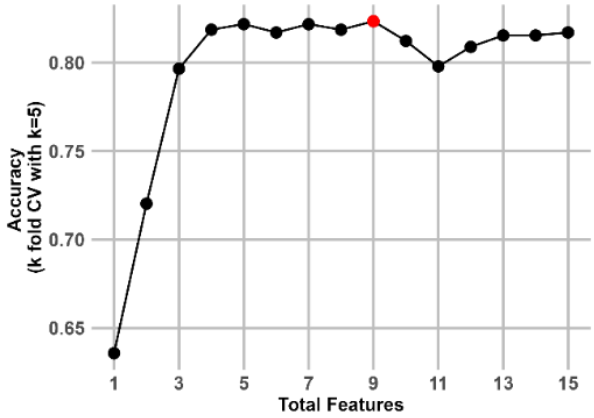


Figure 11. Recursive Feature Elimination on RF Using Train Data

Table 3. Comparison of RFE-RF, RF, and LR on the Test data (Splitting Repeated 5 Times)

Number of Features	Model (after hyperparameter tuning)	Best Hyperparameters	Mean Accuracy ± SD	Mean Balanced Accuracy ± SD	Mean F1-Score ± SD
15	LR	-	0.75 ± 0.029	0.75 ± 0.029	0.73 ± 0.033
15	RF	*mtry*: 3 *trees*: 150 *min node size*: 5	0.80 ± 0.026	0.80 ± 0.026	0.78 ± 0.02
9	RFE-RF	*mtry*: 2 *trees*: 100 *min node size*: 5	0.82 ± 0.01	0.82 ± 0.01	0.81 ± 0.01

We categorize landslide susceptibility into five levels, ranging from very low to very high, based on probability values generated by the RFE-RF model (Fig. 12). The predictions predominantly fall into the very low (427.3 km²) and low (504.8 km²) categories, while areas with high and very high susceptibility only cover 172.5 km² (Table 4). Parigi Moutong and northern Donggala are mainly predicted as very low susceptibility zones, which aligns with the low number of recorded landslide events in these regions. Overall, landslide

points are successfully identified within high and very high susceptibility zones, indicating good model performance.

However, the model exhibits limitations in certain areas, particularly in the eastern part of Sigi, where several recorded landslide locations are still classified as low susceptibility. This indicates that while the RFE-RF model performs well for earthquake-induced (coseismic) landslides, it may not generalize to landslides triggered by other factors, such as rainfall. Recognizing this limitation provides direction for future research, including the incorporation of rainfall and other triggering factors to improve model generalizability across different landslide types.

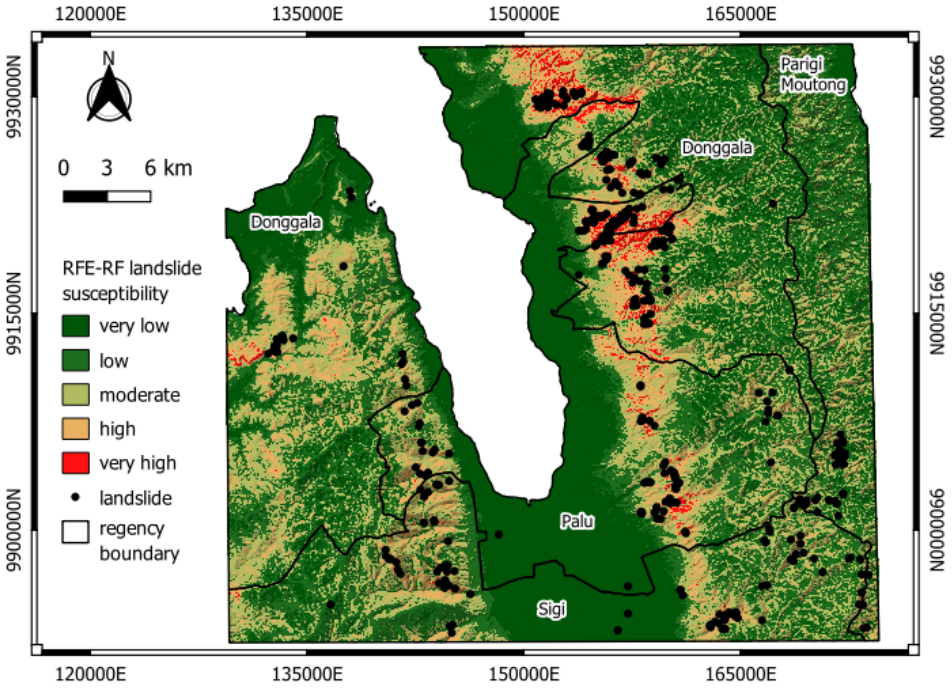


Figure 12. Landslide Susceptibility Map Using RFE-RF Model. Susceptibility Classes are Based on Probability: Very Low (0-0.2), Low (0.2-0.4), Moderate (0.4-0.6), High (0.6-0.8), Very High (0.8-1)

Table 4. Area Size per Category from RFE-RF Prediction

Susceptibility Class	RFE-RF	
	Pixel count	Area (km²)
Very low	449,225	427.3
Low	530,706	504.8
Moderate	406,928	387.1
High	153,723	146.2
Very high	27,597	26.3

Table 5 depicts the proportion of areas with high and very high landslide susceptibility relative to the total area of each regency, as predicted by the RFE-RF model. This information is crucial for establishing priorities in disaster mitigation planning. Donggala, Palu, and Sigi have more than 9% of their territory classified as highly susceptible, indicating a significant elevation in landslide risk. The spatial analysis illustrated in Fig. 12 reveals a concentration of susceptible areas along the administrative boundary between Donggala and Palu. The geomorphological interpretation further indicates that landslides are predominantly distributed along the hilly and mountainous zones flanking the eastern and western margins of the Palu Valley, reflecting structural control on landslide distribution patterns. In contrast, Parigi Moutong exhibits the lowest susceptibility levels among the regencies analyzed.

Table 5. Summary of High and Very High Landslide Susceptibility Areas per Regency from RFE-RF Model

Regency	Total		"High" and "Very High" Landslide Category		Percentage Area (km²) of "High" and "Very High"
	Pixel Count	Area (km²)	Pixel Count	Area (km²)	
Donggala	686,491	653	96,643	91.9	14.1%
Palu	410,545	390.5	51,120	48.6	12.4%
Parigi M.	175,326	166.7	44,64	4.2	2.5%
Sigi	295,817	281.2	29,271	27.8	9.9%

To identify the features that contribute most significantly to model performance, we extract feature importance from the RFE-RF model. This analysis highlights the key features that enhance the model's predictive capability and provides insights into the factors influencing landslide occurrence. Fig. 13 presents the features ordered according to their relative importance. Elevation shows the highest importance, while geological formation exhibits the lowest.

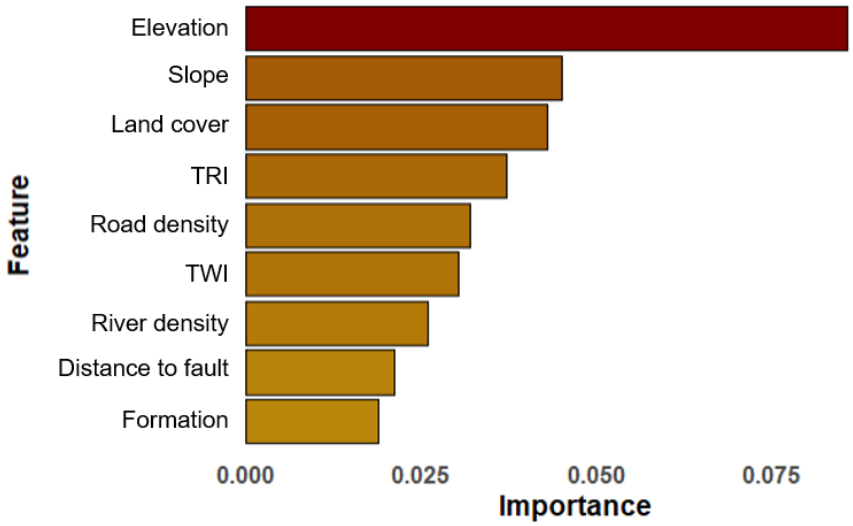


Figure 13. Feature Importance of RFE-RF Model

Fig. 13 shows that the elevation median associated with landslide events is 375 meters, compared to 500 meters for non-landslide events. This observation is particularly noteworthy, considering that previous studies, such as [21], have indicated that landslides are more prevalent at higher elevations relative to their surroundings. Moreover, [21] demonstrated a linear correlation between elevation and landslide frequency, implying that higher elevations are associated with an increased probability of landslides. Nevertheless, susceptibility to landslides is not solely determined by elevation but also by local geological characteristics.

A study by [26] in the European Alps revealed that shallow landslides can occur at relatively low elevations, particularly in areas covered by grassland. This finding aligns with our observations in Central Sulawesi, where the frequency of landslides is notably high in grassland areas. It can be attributed to the susceptibility of grasslands to soil stability disturbances, such as erosion, especially during periods of high rainfall [26].

The most significant factor after elevation is slope. Landslide occurrences tend to be more frequent in areas with a slope greater than 20°, whereas non-landslide events typically have a slope of less than 20°. Previous studies [27], [28] have revealed that as the slope value increases, the frequency of landslides also rises, with slope playing a crucial role in mass movement driven by gravity [29].

TRI is among the top five factors significantly contributing to landslides in the RFE-RF model. TRI measures the elevation difference between a surface and its neighboring points [5] and reflects the roughness of the terrain. As shown in Fig. 13, the proportion of landslides is higher than that of non-landslides when TRI values exceed 9.6. Associating with the prevalence of rugged topography, it contributes to slope instability and increases the likelihood of landslides. Additionally, high road density plays a key role in triggering landslides due to anthropogenic activities such as vehicles and human presence. As noted by [5], road construction can damage the natural structure of slopes and expose fractures, thereby amplifying the potential for landslides.

High TWI increases soil moisture through groundwater accumulation, contributing to the potential for landslides. With a median TWI of 5.4 at landslide points, this value approaches the value associated with high landslide density [30]. In addition to TWI, another hydrological factor is river density. Fig. 9 illustrates that landslide points tend to have higher river density compared to non-landslide points, which are associated with erosion potential. The dominance of Qp and gr formations also contributes to landslide potential, although to a lesser extent compared to other factors based on RFE-RF modeling.

Based on the analysis, Donggala, Palu, and Sigi are recommended as the primary priorities for landslide mitigation efforts, although another regency in the study area also exhibits significant landslide potential in certain zones. Feature importance analysis reveals that elevation, slope, and land cover are the three main factors influencing landslide susceptibility. Elevation and slope are inherent geomorphological characteristics formed through long-term geological processes, making them difficult to alter significantly for mitigation purposes. Therefore, interventions should focus on land cover management. Strategies such as reforestation, increasing vegetation density, or planting deep-rooted species could enhance slope stability and effectively reduce landslide risk in the future [31].

4. CONCLUSION

Applying the Recursive Feature Elimination technique to the tuned Random Forest model has proven to enhance the performance of landslide and non-landslide classification. The RFE-RF model achieves average accuracy, balanced accuracy, and F1-score exceeding 0.81, outperforming both the RF and Logistic Regression models. The classification results, categorized into five susceptibility classes from very low to very high, generally align with the distribution of landslide locations. In the study area located in Central Sulawesi, the model identifies regions such as Donggala, Palu, and Sigi, where high and very high susceptibility areas exceed 9%, suggesting these regions as priorities for mitigation. Mitigation efforts can concentrate on areas along the hills and mountains on both sides of the Palu Valley. Based on the feature importance analysis, we recommend mitigation strategies such as reforestation and planting deep-rooted vegetation to enhance soil stability. This approach aims to minimize disruptions to slope stability caused by the dominance of grasses, which are highly susceptible to erosion, particularly during periods of heavy rainfall. Future work could build upon the current model by integrating high-resolution rainfall data, which would enable the analysis to account for precipitation patterns that strongly influence landslide occurrence. Incorporating rainfall alongside soil moisture and land cover dynamics may further enhance the accuracy of landslide susceptibility predictions over time.

Author Contributions

Indra Rivaldi Siregar: Conceptualization, Methodology, Data Pre-processing, Data Modelling, Visualization and Interpretation, Writing-Original Draft. Anik Djuraidah: Conceptualization, Interpretation of Spatial Analysis, Reviewing and Editing. Agus Mohamad Soleh: Conceptualization, Interpretation of Machine Learning perspectives, Reviewing and Editing. All authors contributed to manuscript refinement, approved the final version, and agreed to be accountable for all aspects of the work.

Funding Statement

This research was funded by the Indonesia Endowment Fund for Education Agency (LPDP), Ministry of Finance, Republic of Indonesia.

Acknowledgment

We sincerely thank the Indonesia Endowment Fund for Education Agency of the Republic of Indonesia for their financial support of this study. Our gratitude also extends to USGS, ESDM, and KLHK for their significant contributions to the data, which were crucial to the research.

Declarations

The authors declare no conflicts of interest.

Declaration of Generative AI and AI-assisted Technologies

The authors declare that no generative AI or AI-assisted technologies were used in the preparation of this manuscript, including for writing, editing, data analysis, or the creation of tables and figures.

REFERENCES

[1] X. Fan *et al*, "EARTHQUAKE-INDUCED CHAINS OF GEOLOGIC HAZARDS: PATTERNS, MECHANISMS, AND IMPACTS," *Reviews of Geophysics*, vol. 57, no. 2, pp. 421–503, Jun. 2019, doi: https://doi.org/10.1029/2018RG000626.

[2] F. S. Tehrani, M. Calvello, Z. Liu, L. Zhang, and S. Lacasse, "MACHINE LEARNING AND LANDSLIDE STUDIES: RECENT ADVANCES AND APPLICATIONS," *Natural Hazards*, vol. 114, no. 2, pp. 1197–1245, Nov. 2022, doi: https://doi.org/10.1007/s11069-022-05423-7.

[3] A. Merghadi *et al*, "MACHINE LEARNING METHODS FOR LANDSLIDE SUSCEPTIBILITY STUDIES: A COMPARATIVE OVERVIEW OF ALGORITHM PERFORMANCE," *Earth-Science Reviews*, vol. 207, p. 103225, Aug. 2020, doi: https://doi.org/10.1016/j.earscirev.2020.103225.

[4] L. Breiman, "RANDOM FORESTS," *Mach Learn*, vol. 45, pp. 5–32, Oct. 2001, doi: https://doi.org/10.1023/A:1010933404324.

[5] S. Aldiansyah and F. Wardani, "ASSESSMENT OF RESAMPLING METHODS ON PERFORMANCE OF LANDSLIDE SUSCEPTIBILITY PREDICTIONS USING MACHINE LEARNING IN KENDARI CITY, INDONESIA," *Water Practice and Technology*, vol. 19, no. 1, pp. 52–81, Jan. 2024, doi: https://doi.org/10.2166/wpt.2024.002.

[6] [BG] Badan Geologi, *DI BALIK PESONA PALU: BENCANA MELANDA GEOLOGI MENATA*, 1st ed. Bandung: Kementerian Energi dan Sumber Daya Mineral Republik Indonesia, 2018.

[7] A. Sabaruddin., *LAPORAN PROGRESS PENEGASAN ZONA RAWAN BENCANA SESAR PALUKORO PASCA GEMPA PALU 28 SEPTEMBER 2018*. Kementerian Pekerjaan Umum dan Perumahan Rakyat, 2018.

[8] Sunardi, N. Anggraini, S. Alfiandy, and A. F. Ilahi, "IDENTIFIKASI TINGKAT KERAWANAN TANAH LONGSOR DI PROVINSI SULAWESI TENGAH," *Buletin GAW Bariri*, vol. 3, no. 2, pp. 47–57, Dec. 2022, doi: https://doi.org/10.31172/bgb.v3i2.79.

[9] H. Kaur, S. Gupta, S. Parkash, and R. Thapa, "KNOWLEDGE-DRIVEN METHOD: A TOOL FOR LANDSLIDE SUSCEPTIBILITY ZONATION (LSZ)," *Geology, Ecology, and Landscapes*, vol. 7, no. 1, pp. 1–15, Jan. 2023, doi: https://doi.org/10.1080/24749508.2018.1558024.

[10] S. Sukristiyanti *et al*, "MACHINE LEARNING FOR LANDSLIDE SUSCEPTIBILITY MAPPING USING PHYTON IN SIGI BIROMARU AREA (NEAR PALU), CENTRAL SULAWESI, INDONESIA," *IOP Conference Series: Earth and Environmental Science*, vol. 1276, no. 1, p. 012024, Dec. 2023, doi: https://doi.org/10.1088/1755-1315/1276/1/012024.

[11] L. Demarchi *et al*, "RECURSIVE FEATURE ELIMINATION AND RANDOM FOREST CLASSIFICATION OF NATURA 2000 GRASSLANDS IN LOWLAND RIVER VALLEYS OF POLAND BASED ON AIRBORNE HYPERSPECTRAL AND LIDAR DATA FUSION," *Remote Sensing*, vol. 12, no. 11, p. 1842, Jun. 2020, doi: https://doi.org/10.3390/rs12111842.

[12] A. R. Barzani, P. Pahlavani, O. Ghorbanzadeh, K. Gholamnia, and P. Ghamisi, "EVALUATING THE IMPACT OF RECURSIVE FEATURE ELIMINATION ON MACHINE LEARNING MODELS FOR PREDICTING FOREST FIRE-PRONE ZONES," *Fire*, vol. 7, no. 12, p. 440, Nov. 2024, doi: https://doi.org/10.3390/fire7120440

[13] B. Zhao, "AN OPEN REPOSITORY OF EARTHQUAKE-TRIGGERED GROUND FAILURE INVENTORIES, U.S. geological survey data release collection."

[14] M. Azarafza, M. Azarafza, H. Akgün, P. M. Atkinson, and R. Derakhshani, "DEEP LEARNING-BASED LANDSLIDE SUSCEPTIBILITY MAPPING," *Scientific Reports*, vol. 11, no. 1, p. 24112, Dec. 2021, doi: https://doi.org/10.1038/s41598-021-03585-1.

[15] N. Saleem, Md. E. Huq, N. Y. D. Twumasi, A. Javed, and A. Sajjad, "PARAMETERS DERIVED FROM AND/OR USED WITH DIGITAL ELEVATION MODELS (DEMS) FOR LANDSLIDE SUSCEPTIBILITY MAPPING AND LANDSLIDE RISK ASSESSMENT: A REVIEW," *ISPRS International Journal of Geo-Information*, vol. 8, no. 12, p. 545, Nov. 2019, doi: https://doi.org/10.3390/ijgi8120545.

[16] S. Lee and J. A. Talib, "PROBABILISTIC LANDSLIDE SUSCEPTIBILITY AND FACTOR EFFECT ANALYSIS," *Environmental Geology*, vol. 47, no. 7, pp. 982–990, May 2005, doi: https://doi.org/10.1007/s00254-005-1228-z.

[17] H. R. Pourghasemi, B. Pradhan, and C. Gokceoglu, "APPLICATION OF FUZZY LOGIC AND ANALYTICAL HIERARCHY PROCESS (AHP) TO LANDSLIDE SUSCEPTIBILITY MAPPING AT HARAZ WATERSHED, IRAN," *Natural Hazards*, vol. 63, no. 2, pp. 965–996, Sep. 2012, doi: https://doi.org/10.1007/s11069-012-0217-2.

[18] K. R. Ahmed and S. Akter, "ANALYSIS OF LANDCOVER CHANGE IN SOUTHWEST BENGAL DELTA DUE TO FLOODS BY NDVI, NDWI AND K-MEANS CLUSTER WITH LANDSAT MULTI-SPECTRAL SURFACE REFLECTANCE SATELLITE DATA," *Remote Sensing Applications: Society and Environment*, vol. 8, pp. 168–181, Nov. 2017, doi: https://doi.org/10.1016/j.rsase.2017.08.010.

[19] H. R. Pourghasemi and O. Rahmati, "PREDICTION OF THE LANDSLIDE SUSCEPTIBILITY: WHICH ALGORITHM, WHICH PRECISION?," *Catena*, vol. 162, pp. 177–192, Mar. 2018, doi: https://doi.org/10.1016/j.catena.2017.11.022.

[20] IPCC, "LAND USE, LAND-USE CHANGE, AND FORESTRY. A special report," 2002.

[21] N. B. Raja, I. Çiçek, N. Türkoğlu, O. Aydin, and A. Kawasaki, "LANDSLIDE SUSCEPTIBILITY MAPPING OF THE SERA RIVER BASIN USING LOGISTIC REGRESSION MODEL," *Natural Hazards*, vol. 85, no. 3, pp. 1323–1346, Feb. 2017, doi: https://doi.org/10.1007/s11069-016-2591-7.

[22] L. Y. Irawan *et al*, "THE USE OF MACHINE LEARNING FOR ACCESSING LANDSLIDE SUSCEPTIBILITY CLASS: STUDY CASE OF PACET SUBDISTRICT, MOJOKERTO REGENCY," in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing Ltd, Nov. 2021. doi: https://doi.org/10.1088/1755-1315/884/1/012006.

[23] S. M. Malakouti, M. B. Menhaj, and A. A. Suratgar, "THE USAGE OF 10-FOLD CROSS-VALIDATION AND GRID SEARCH TO ENHANCE ML METHODS PERFORMANCE IN SOLAR FARM POWER GENERATION

PREDICTION," *Cleaner Engineering and Technology*, vol. 15, p. 100664, Aug. 2023, doi: https://doi.org/10.1016/j.clet.2023.100664.

[24] X. Guo and P. Hao, "USING A RANDOM FOREST MODEL TO PREDICT THE LOCATION OF POTENTIAL DAMAGE ON ASPHALT PAVEMENT," *Applied Sciences*, vol. 11, no. 21, p. 10396, Nov. 2021, doi: https://doi.org/10.3390/app112110396.

[25] L. Demarchi *et al*, "RECURSIVE FEATURE ELIMINATION AND RANDOM FOREST CLASSIFICATION OF NATURA 2000 GRASSLANDS IN LOWLAND RIVER VALLEYS OF POLAND BASED ON AIRBORNE HYPERSPECTRAL AND LIDAR DATA FUSION," *Remote Sensing*, vol. 12, no. 11, p. 1842, Jun. 2020, doi: https://doi.org/10.3390/rs12111842.

[26] C. Geitner *et al*, "SHALLOW EROSION ON GRASSLAND SLOPES IN THE EUROPEAN ALPS – GEOMORPHOLOGICAL CLASSIFICATION, SPATIO-TEMPORAL ANALYSIS, AND UNDERSTANDING SNOW AND VEGETATION IMPACTS," *Geomorphology*, vol. 373, p. 107446, Jan. 2021, doi: https://doi.org/10.1016/j.geomorph.2020.107446.

[27] Asdar *et al*, "ANALYSIS OF THE LANDSLIDES VULNERABILITY LEVEL USING FREQUENCY RATIO METHOD IN TANGKA WATERSHED," *IOP Conference Series: Earth and Environmental Science*, vol. 870, no. 1, p. 012013, Oct. 2021, doi: https://doi.org/10.1088/1755-1315/870/1/012013.

[28] R. Amaliah, A. S. Soma, B. Mappangaja, and F. Mambela, "ANALYSIS OF THE LANDSLIDE SUSCEPTIBILITY MAP USING FREQUENCY RATIO METHOD IN SUB-SUB-WATERSHED MAMASA," *IOP Conference Series: Earth and Environmental Science*, vol. 886, no. 1, p. 012088, Nov. 2021, doi: https://doi.org/10.1088/1755-1315/886/1/012088.

[29] F. E. S. Silalahi, Pamela, Y. Arifianti, and F. Hidayat, "LANDSLIDE SUSCEPTIBILITY ASSESSMENT USING FREQUENCY RATIO MODEL IN BOGOR, WEST JAVA, INDONESIA," *Geoscience Letters*, vol. 6, no. 1, p. 10, Dec. 2019, doi: https://doi.org/10.1186/s40562-019-0140-4.

[30] M. Meinhardt, M. Fink, and H. Tünschel, "LANDSLIDE SUSCEPTIBILITY ANALYSIS IN CENTRAL VIETNAM BASED ON AN INCOMPLETE LANDSLIDE INVENTORY: COMPARISON OF A NEW METHOD TO CALCULATE WEIGHTING FACTORS BY MEANS OF BIVARIATE STATISTICS," *Geomorphology*, vol. 234, pp. 80–97, Apr. 2015, doi: https://doi.org/10.1016/j.geomorph.2014.12.042.

[31] L. Chen, Z. Guo, K. Yin, D. P. Shrestha, and S. Jin, "THE INFLUENCE OF LAND USE AND LAND COVER CHANGE ON LANDSLIDE SUSCEPTIBILITY: A CASE STUDY IN ZHUSHAN TOWN, XUAN'EN COUNTY (HUBEI, CHINA)," *Natural Hazards and Earth System Sciences*, vol. 19, no. 10, pp. 2207–2228, Oct. 2019, doi: https://doi.org/10.5194/nhess-19-2207-2019.