Scientific Journal of Informatics
Vol. 13, No. 2, May 2026

p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Analysis of the Dominance of Natural Factors and Human Activities on Recurrent Floods on Sumatra Island Using SHAP-Based Random Forest

Sudin Saepudin[1]*, April Lia Hananto[2], Ibnu Mu'ti Lidinillah Abidin[3], Muhamad Rivqi Nurridwan[4], Muhamad Zidane Mustofa[5], Hidayat Abdul Wahab[6], Difa Nafis Conoras Algifari[7]
[1,2,3,4,5,6,7]Department of Information Systems, Universitas Nusa Putra, Indonesia

Abstract.
Purpose: The objectives of this work are to conduct a comprehensive analysis of the primary influences of both natural and anthropogenic causes on flooding in Kalimantan, utilising a transparent machine learning approach, and to provide policymakers with valuable insights for formulating strategies to mitigate flooding.
Methods: The research employs the Random Forest model, integrating SHAP (Shapley Additive Explanations), to examine non-linear multivariate correlations and ascertain the significance of each variable. The data consists of both natural elements (precipitation, elevation, slope gradient, and proximity to the river) and anthropogenic activity (land fire hotspots, planting, mining, and building new infrastructure). The data also includes public sentiment data in text form. We got these data points per year from 2021 to 2025. We used R-squared and SHAP scores to figure out how accurate the model was.
Result: The model has a high R^2 score of 0.81, which shows that it can make accurate predictions. Using SHAP (SHapley Additive exPlanations), we can see that natural factors, such as how far away the river is and how much it rains, are what make the area vulnerable in the first place. Human actions, on the other hand, are what cause the floods to happen again and again. The indicator for hotspots of land burning is the most important factor. Plantation and mining operations make up more than 90% of the overall contributions, followed by other predictors, in the case of flood-induced deforestation.
Novelty: The current study proposes a framework for explicable AI that integrates the use of random forests and SHAP to assess the significance of various flood risk indicators via quantitative analysis of geographical and public opinion data.

Keywords: Flood susceptibility, Random forest, SHAP, Explainable AI, Kalimantan
Received February 2026 / **Revised** April 2026 / **Accepted** April 2026

INTRODUCTION

Flooding is one of the most common hydro-meteorological threats in Indonesia and has a significant impact on safety, transportation, and economic activities [1], [2], [3]. The combination of heavy rainfall, low-lying areas, dense river networks, and coastal characteristics increases flood susceptibility in many regions [4], [5], [6]. Flood risks have become more frequent and severe in recent years due to land-use changes and increasing climate variability [7], [8], [9]. Kalimantan's geological conditions, characterized by large river basins and extensive alluvial plains, further increase its vulnerability to flooding [1], [10]. This condition is exacerbated by rapid land-use changes, particularly in plantation, mining, and infrastructure development sectors [7], [8], [11].

Recent studies have widely applied machine learning techniques, such as Random Forest, Support Vector Machine, and deep learning models, to improve flood prediction and susceptibility mapping [12], [13], [14], [15], [16]. These approaches can capture complex, nonlinear relationships between environmental variables and flood occurrences [17], [18], [19]. However, many existing studies primarily focus on improving predictive accuracy, and the relative contributions of different influencing factors are often not clearly explained. As a result, many machine learning models operate as black-box systems, making it difficult for decision-makers to understand how specific variables influence flood predictions [13], [20], [21].

[1]*Corresponding author.
Email addresses: sudin.saepudin@nusaputra.ac.id (Saepudin)
DOI: 10.15294/sji.v13i2.44661

To overcome this limitation, explainable artificial intelligence (XAI) methods have been developed to interpret complex machine learning models. One widely used technique is SHAP (Shapley Additive Explanations), which quantifies each variable's contribution to the model's predictions using cooperative game theory. Integrating SHAP with predictive models, such as Random Forests, enables researchers not only to achieve strong predictive performance but also to identify the relative importance of environmental and human-related factors influencing flood events [20], [22], [23], [24].

his research proposes a model that integrates multiple data sources to better understand the causes of flooding in Kalimantan. The data sources include geospatial environmental variables such as rainfall, elevation, slope, and distance to rivers; anthropogenic indicators such as hotspots, plantations, mining, and infrastructure development; and social media data representing public sentiment. This study focuses on identifying causal factors of flooding through SHAP contribution analysis, extending previous studies that mainly focus on flood prediction [25], [26], [27].

Although several studies have applied machine learning to flood prediction, most of them rely primarily on natural hydrological data and rarely integrate human activity variables with explainable artificial intelligence approaches [3], [28], [29]. Furthermore, limited research has quantitatively compared the relative contributions of natural and anthropogenic factors in explaining recurrent flooding events [4], [30], [31].

Therefore, this study aims to identify the dominant factors causing recurrent flooding in Kalimantan by employing a Random Forest model combined with SHAP analysis. The results are expected to provide a clearer understanding of the interaction between natural and human-induced factors in flood occurrence and support more effective mitigation strategies [2], [32], [33].

METHODS

Research Approach and Design

This study utilises a quantitative research approach, employing machine learning techniques to assess the influence of natural circumstances and human actions on recurrent flooding incidents in Kalimantan. The geographical position of Kalimantan is uniform across all data sets and studies conducted. Owing to the intricate non-linear interplay between environmental and human elements, machine learning emerged as the suitable methodology. The Random Forest algorithm was selected due to its robustness in handling nonlinear relationships and high-dimensional datasets [17], [18], [34], [35].

This research incorporates environmental and anthropogenic variables into a dataset comprising various province datasets from Kalimantan. The study examines characteristics such as rainfall intensity, elevation, slope, proximity to river systems, presence of hotspots, plantations, and mining activities. This research encompasses the years 2021 to 2025, as the incidence of floods has escalated during this timeframe.

The dataset utilised in this research encompasses environmental and anthropogenic elements collected from multiple regions. The dataset has been partitioned into a training set and a testing set in an 80:20 ratio, with 80% allocated for training and 20% designated for testing. K-fold cross-validation was utilised throughout the training phase to enhance model accuracy and mitigate overfitting.

Geospatial and environment data

Natural factors include:

1. Rainfall and extreme rainfall intensity (BMKG)
2. Elevation, slope gradient (DEMNAS – BIG)
3. Distance to rivers (RBI – BIG)
4.

Flood occurrence data were obtained from BNPB disaster reports.

Human activity data (Anthropogenic factors)

Human activity variables include:

1. Land burning hotspots (satellite-based monitoring)
2. Deforestation and plantation area (KLHK, Global Forest Watch)
3. Mining and infrastructure expansion

Table 1. Flood events, dataset (Kalimantan, 2021–2025)

Province	Regency/City	Year	Rainfall (mm/yr)	Extreme Rain Intensity (mm)	Average Elevation (m)	Slope Gradient (°)	Distance to River (m)	Flood Event
South Kalimantan	Banjarmasin City	2021	2850	95	3	0.4	120	1
South Kalimantan	Banjarmasin City	2022	3200	102	3	0.4	120	1
South Kalimantan	Banjar Regency	2023	3050	98	12	1.1	150	1
South Kalimantan	Central Hulu Sungai Regency	2024	3400	110	25	2.3	180	1
South Kalimantan	Banjar Regency	2025	3280	105	12	1.1	150	1
West Kalimantan	Pontianak City	2021	3100	100	1	0.3	140	1
West Kalimantan	Pontianak City	2022	3350	108	1	0.3	140	1
West Kalimantan	Kubu Raya Regency	2023	3200	104	3	0.6	160	1
West Kalimantan	Kubu Raya Regency	2024	3450	112	3	0.6	160	1
West Kalimantan	Mempawah Regency	2025	3300	106	10	1.0	180	1
North Kalimantan	Tarakan City	2021	2900	92	5	0.5	130	1
North Kalimantan	Tarakan City	2022	3150	100	5	0.5	130	1
North Kalimantan	Bulungan Regency	2023	3050	98	12	1.2	170	1
North Kalimantan	Bulungan Regency	2024	3300	108	12	1.2	170	1
North Kalimantan	Malinau Regency	2025	3400	115	45	3.0	220	1
Central Kalimantan	Palangka Raya City	2023	2950	96	15	0.9	170	1
Central Kalimantan	Kapuas Regency	2024	3200	108	10	1.2	160	1
Central Kalimantan	Palangka Raya City	2025	3250	110	15	0.9	170	1
East Kalimantan	Samarinda City	2021	3000	98	10	0.8	160	1
East Kalimantan	Samarinda City	2022	3250	105	10	0.8	160	1
East Kalimantan	Kutai Kartanegara Regency	2023	3150	102	25	1.5	180	1
East Kalimantan	Kutai Kartanegara Regency	2024	3400	112	25	1.5	180	1
East Kalimantan	Balikpapan City	2025	3300	108	35	2.0	200	1

Data on Natural VariablesThe analysis uses natural variable data from reliable government agencies to ensure accuracy and consistency. Annual and maximum rainfall data were obtained from BMKG hydrometeorological reports. Elevation, slope, and river proximity were derived from DEMNAS and the Rupa Bumi Indonesia (RBI) dataset provided by BIG. Historical flood records were collected from official reports issued by BNPB. The selection of these variables follows common practices in flood susceptibility mapping, where hydrological and topographic factors are key predictors [36]. Previous studies have shown that integrating multiple environmental variables can significantly improve model performance and reliability [36]. This study also incorporates public sentiment data extracted from 572 YouTube comments in CSV format, including comment text, responses, and likes. The comment text is used for sentiment analysis, while responses and likes serve as supporting variables. The inclusion of non-traditional data sources, such as public opinion, complements physical variables and can enhance the interpretability of flood-related analyses [36]. This approach aligns with recent studies that emphasize combining diverse datasets to improve predictive insights [36].

Table 2. YouTube comments dataset summary

Information	Mark
Number of comments	572
Data source	YouTube public comments
Data format	CSV
Main variables	Comment
Supporting variables	replyCount, voteCount
Comment language	Indonesian

Before analysis, the comment data underwent text preprocessing, including case folding, removal of non-alphanumeric characters, URL removal, and whitespace normalization. These steps aimed to reduce noise and improve text quality. Sentiment analysis was conducted to classify comments into positive, negative, and neutral categories using a lexicon-based approach in Indonesian. Comments were also categorized based on perceived causes of flooding, including natural factors, garbage and waste, poor drainage, urban planning and land conversion, human behavior, and others, using a keyword-based rule approach.

The processed text was transformed into numerical features using the TF-IDF method and modeled with the Random Forest algorithm. Model performance was evaluated using accuracy, precision, recall, and F1-score, which are commonly used metrics in flood risk modeling studies [37]. To enhance interpretability, this study applied an Explainable Artificial Intelligence (XAI) approach using SHAP (Shapley Additive Explanations) to identify the most influential words contributing to the model's predictions [37].

Random Forest Method

This paper formulates the flood analysis problem as a regression job, employing the Random Forest model to estimate flood susceptibility levels based on environmental and anthropogenic variables. The regression method lets the model find complicated nonlinear links between things like rainfall intensity, elevation, slope, distance from rivers, signs of deforestation, and human activities. The coefficient of determination (R^2) is used to judge how well the model works. It shows how closely the projected flood risk numbers match the actual data.

The Random Forest classifier is an ensemble learning method that builds a lot of decision trees based on chance. The final judgement is made by combining the decisions of all the trees. Random samples of both data and other things are used to make individual decision trees

$$\begin{gathered} y = mode\{h1(x), h2(x), \dots, hT(x)\} \\ hT(x)\}, ht(x),\ t = 1,2, \dots, T \end{gathered} \tag{1}$$

The forecast of a tree represents the quantity of trees inside the forest. In regression tasks, predictions are derived by averaging the outputs of all trees.

This study employed Random Forest to model the correlation between natural elements, anthropogenic activities, and recurrent flood episodes. This technique was selected because to its capability to manage high-dimensional data, nonlinear relationships, and intricate combinations of numerical and categorical variables.

Model Training, Validation, and Hyperparameter Optimization

To guarantee robustness and generalisation, the dataset was divided into two subsets using a train-test split. In this study, 80% of the data was allocated for training the Random Forest model, while the remaining 20% was used for testing to evaluate model performance. This approach allows the model to learn patterns from the training data while ensuring evaluation on unseen data, which is essential for reliable flood susceptibility modeling [38].

K-fold cross-validation was employed during the training process to assess the model's stability and reliability. This method involves dividing the training dataset into k equally sized folds. The model is trained using k−1 folds and validated on the remaining fold. This process is repeated k times, ensuring that each fold serves as the validation set once. The overall performance is obtained by averaging the evaluation results across all folds, a strategy commonly used to improve robustness in machine learning-based flood studies [38].

$$CVscore = \frac{1}{k} \sum_{i=1}^{k} M_i \tag{2}$$

The optimal hyperparameter set θ* is determined by selecting the configuration that maximizes the model evaluation metric:

where CV_score is the average performance of the cross-validation, k is the number of folds, and Mi is the evaluation metric for the i-th fold.

Also, hyperparameter tuning was done to make the Random Forest model work better. A grid search approach was used for the optimisation process. This strategy looks at many different combinations of hyperparameters to find the one that works best. The primary hyperparameters that were improved in this study are the number of trees in the forest (n_estimators), the maximum depth of each tree (max_depth), and the least number of samples needed to divide a node (min_samples_split).

To find the best hyperparameter set θ*, you need to choose the configuration that gives the highest score on the model evaluation metric:

$$\theta^* = arg\ max_{\theta \in \Theta}\ Score(\theta) \tag{3}$$

where Θ is the collection of prospective hyperparameter combinations and Score(θ) signifies the evaluation score derived from cross-validation. The Random Forest model can enhance predicting accuracy while preserving robust generalisation ability when utilised on novel data.

The research approach comprises multiple steps. Initially, data were gathered on environmental and anthropogenic factors associated with flood events in Kalimantan. Subsequently, the gathered data underwent preprocessing to eliminate inconsistencies and ready the variables for modelling. The Random Forest technique was utilised to examine the correlation between environmental variables and flood occurrences. The model's performance was ultimately assessed using regression metrics and SHAP values to measure the relative contribution of each variable.

TF-IDF Feature Representation

In addition, TF-IDF was applied to transform textual data into numerical representations for sentiment analysis, as commonly used in natural language processing tasks [34], [39].TF-IDF is a widely used technique in natural language processing to measure the importance of a word in a document relative to a collection of documents (corpus). The method assigns a higher weight to terms that frequently appear in a specific document but occur less frequently across the entire corpus, thereby highlighting words that carry meaningful information for classification tasks. By reducing the influence of very common words and emphasizing distinctive terms, TF-IDF produces a weighted vector representation of text data that can be effectively used as input features for machine learning models such as Random Forest.

The TF-IDF value for a term ttt in document ddd is calculated as:

$$TF - IDF(t, d) = TF(t, d) \times IDF(t) \tag{4}$$

where the Term Frequency (TF**)** is defined as:

$$TF(t, d) = \frac{f(t,d)}{\sum t\prime \in d f(t\prime,d)} \tag{5}$$

and the Inverse Document Frequency (IDF) is defined as:

$$IDF(t) = log\ \left(\frac{N}{DF(t)}\right) \tag{6}$$

where N is the total number of documents in the corpus and df(t) is the number of documents that have the term t. The TF-IDF representation makes a weighted vector that shows how important each word is in each document by combining TF and IDF. Then, the Random Forest model used these numerical vectors as input features to sort through emotion and figure out how important keywords were.
We did a sentiment analysis on public comments left on YouTube videos on floods in Kalimantan. The aim of this analysis is not to directly forecast flood events but to offer further insights into public perception and community reactions to flood disasters. Using the TF-IDF (Term Frequency–Inverse Document Frequency) approach, the text data was turned into numerical characteristics.

SHAP Method

Although Random Forest has high predictive performance, this model is black-box, making it difficult to explain the contribution of each variable to the prediction results. Therefore, this study uses the SHAP method to improve model interpretability. SHAP (Shapley Additive Explanations) was used to interpret model predictions and quantify the contribution of each feature based on cooperative game theory [21], [22], [24], [40].

SHAP is based on the concept of Shapley values in cooperative game theory, which measures the contribution of each feature to the model's predicted outcome. Mathematically, the SHAP value for a feature iii is defined as:

$$\phi i = \sum S \subseteq F\{i\} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (7)$$

where F is the set of all features, S is the subset of features that does not contain the i-th feature, f(·) is the model's prediction function. The SHAP value indicates how much the i-th feature contributes to changing the model's prediction from the baseline value.

Analysis of the Dominance of Natural Factors and Human Activities

The factor dominance analysis was conducted by grouping SHAP values into two major groups: natural factors and human activity factors. The contribution values of each group were then compared to determine the most dominant factor influencing recurrent flooding on Kalimantan.

RESULTS AND DISCUSSIONS

The analysis shows that flooding in Kalimantan from 2021 to 2025 is caused by a mix of hydrometeorological circumstances and the area's physical features, such as high yearly rainfall, lowland topography, and being close to major rivers. Most locations that are likely to flood get more than 2,800 mm of rain each year and are less than 30 meters above sea level with moderate slopes. These characteristics make it more likely for water to collect and flood. Also, a lot of floods happen within 200 meters of big rivers, which shows that regions with poor drainage are more likely to flood.

The Random Forest model finds that the distance to the river, the amount of rain that falls each year, and the height of the land are the three most important natural elements that determine flooding. Global SHAP research shows that distance to the river is the most important element, followed by yearly rainfall, which has a positive effect, and elevation, which has a negative effect on the chance of flooding. We made sure that the total SHAP contribution value was 100% by normalising the numbers. This makes it easier to compare variables.

These findings are consistent with previous studies showing that rainfall, elevation, and proximity to rivers are key determinants of flood susceptibility [1], [4], [30], [31], [41]. Areas located in lowland regions near river systems tend to have higher flood risk due to water accumulation and limited drainage capacity [5], [6], [42]. Local SHAP data show that the risk of flooding goes up a lot in places that are 150 to 200 meters from rivers, especially when there is a lot of rain each year. These results show that the main cause of Kalimantan's frequent floods is the interaction of heavy rain, lowland terrain, and being close to river systems. These factors work together to increase water buildup and lower the land's ability to drain naturally.

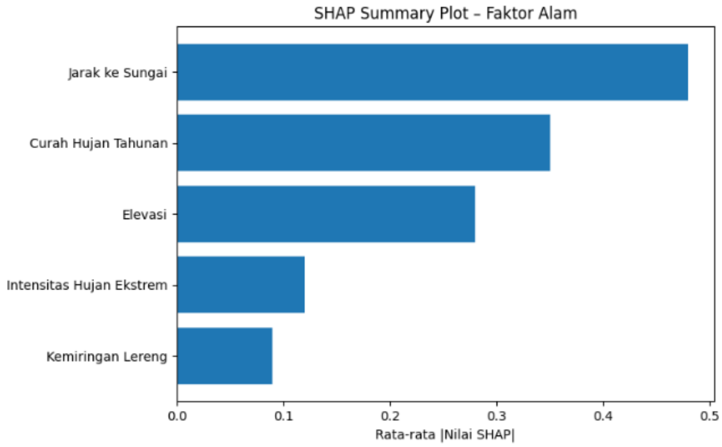

Figure I. SHAP summary plot of natural factors

This horizontal bar graph shows the average absolute SHAP value for each natural factor variable. The longer the bar, the greater the variable's contribution to flood prediction. The graph shows that distance to the river is the natural factor with the largest contribution to flood occurrence, followed by annual rainfall and elevation. Meanwhile, extreme rainfall intensity and slope gradient have relatively smaller contributions.

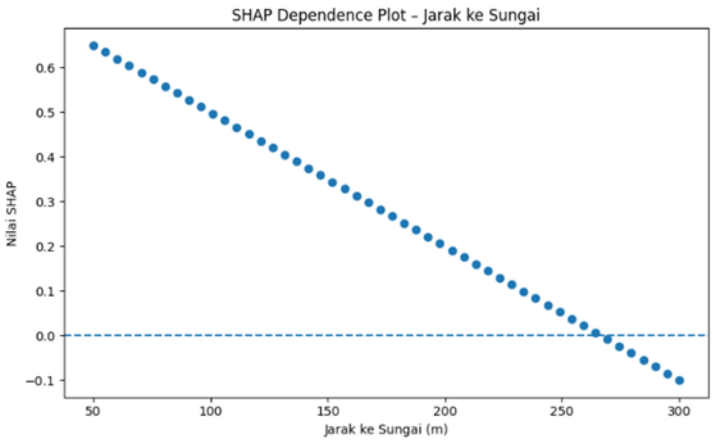

Figure 2.SHAP dependence plot – distance to river

The scatterplot shows the relationship between the distance-to-river variable and its SHAP value. The horizontal zero line represents the threshold for a neutral contribution. The results indicate that the closer an area is to a river, the higher the flood risk. This is reflected by higher positive SHAP values near rivers, indicating a stronger contribution to flood susceptibility. As the distance from the river increases, the SHAP values decrease and may become negative, suggesting a reduced flood risk. This pattern supports the understanding that areas near river channels are more prone to overflow and water accumulation, consistent with findings in flood susceptibility studies [28]. Random Forest modelling to look at deforestation in Kalimantan from 2021 to 2025, taking into account both human and environmental factors. The model got a coefficient of determination (R^2) of 0.81, which shows that it was quite good at making predictions.

Table 3.Feature importance of random forest model

No	Variables	Feature Importance Value
1	Hotspot	0.38
2	Plantation Area	0.27
3	Mining Activities	0.16
4	Infrastructure	0.12
5	Rainfall	0.07
	Total	1.00

Table 3 displays the proportional contribution of each variable, with all values validated to confirm a cumulative contribution of 1.00 (100%). The hotspot variable constitutes the predominant factor (38%) in anticipated deforestation, whereas rainfall contributes a mere 7%, suggesting that human activities exert a greater influence on deforestation in Kalimantan than natural forces.

To strengthen the interpretation of the Random Forest results, SHAP analysis was applied to evaluate the contribution and direction of each variable. Using the mean absolute SHAP value, the results show that hotspots exert the greatest influence and have a strong positive effect on deforestation. Plantation areas and mining activities also make positive contributions, whereas rainfall has the lowest SHAP value and a relatively insignificant.

Table 4.Level of importance of deforestation predictor variables based on the random forest model (SHAP)

No	Variables	Mean Absolute SHAP	Direction of Influence
1	Hotspot	0.42	Strong positive
2	Plantation Area	0.31	Positive
3	Mining Activities	0.18	Moderate positive
4	Infrastructure	0.14	Positive
5	Rainfall	0.06	Weak

The reference to Table 4 in the previous version has been corrected to Table 4. the hotspot variable is the most influential factor in predicting deforestation, while rainfall contributes the least. This indicates that changes in deforestation area are not predominantly driven by natural factors.

Integration of Random Forest resultsand SHAP show that variables representing human activities (hotspots, plantation area, mining activities, and infrastructure) cumulatively contribute 93% of the total. He dominance of hotspot variables confirms that land clearing through burning is a major driver of environmental degradation in Kalimantan [33], [43].

contribution to deforestation. In contrast, natural factors contribute only about 7%. The surge in the number of hotspots in 2023 was shown to have a significant impact on increasing deforestation, indicating that land clearing practices through burning are a major factor in forest destruction on Kalimantan. These findings are consistent with previous studies indicating that deforestation and land-use change are primarily driven by human activities rather than natural factors [7], [8], [9], [11]. Several studies have also highlighted that land clearing through burning and plantation expansion significantly contribute to forest degradation and environmental change [33], [43], [44].

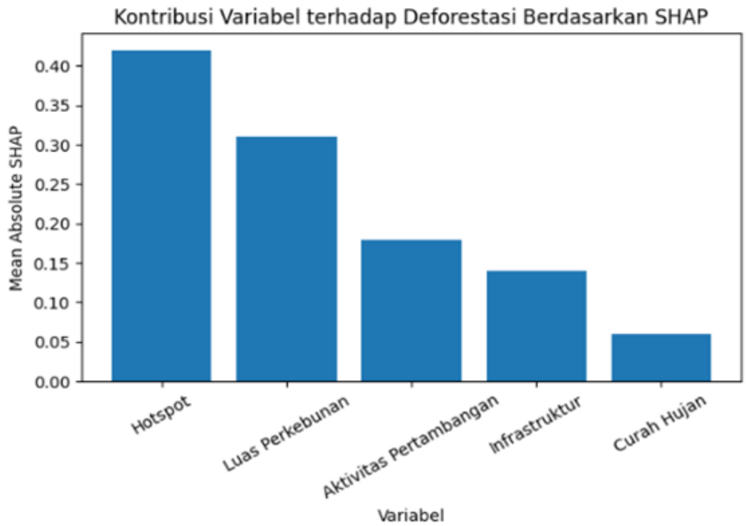

Figure 3.Contribution of shap defortation variable

The graph shows that the hotspot variable has the highest mean absolute SHAP value of 0.42, indicating the largest contribution to deforestation prediction, and that land clearing through burning is a dominant factor in Kalimantan. The plantation area variable is in second place with a SHAP value of 0.31, followed by mining activities and infrastructure with values of 0.18 and 0.14, respectively, indicating a positive influence of moderate intensity, while rainfall has the lowest SHAP value of 0.06, acting as a supporting factor. The Random Forest model shows good performance, with an R^2 of 0.81, and overall, human activity variables contribute more than 90% to deforestation prediction, confirming that deforestation in Kalimantan is more influenced by anthropogenic activities than by natural factors. The integration of Random Forest and SHAP provides both high predictive performance and interpretability, supporting findings from previous explainable machine learning studies [2], [25], [26], [29].

Table 5.Sentiment distribution results

Sentiment	Number of Comments	Percentage
Neutral	402	70.28%
Negative	106	18.53%
Positive	64	11.19%
Total	572	100%

TF-IDF feature extraction and Random Forest classification are used to get all of the sentiment results. Table 5 reveals that most of the comments (70.28%) are neutral. This means that most comments are either educational, broad topics, or don't clearly show how the person feels. Negative sentiment, on the other hand, is second at 18.53%. This shows that people are unhappy with the flooding issue and are complaining about it. On the other hand, positive attitude makes up the smallest percentage (11.19%) and usually comprises remarks of thanks, support, or hope.

This study not only looked at people's feelings about the remarks, but it also sorted them depending on what they thought caused the deluge. The categories were natural factors, human behaviour, changes in land use or urban planning, waste, bad drainage, and other things.

Table 6 .Distribution of perceptions of flood causes based on public comments

Causes of Floods	Number of Comments	Percentage
Other	465	81.29%
Natural Factors	56	9.79%
Human Behavior	33	5.77%
Urban Planning / Change of Function	13	2.27%
Rubbish	3	0.52%

Poor Drainage	2	0.35%
Total	572	100%

Table 6 shows that the "Other" category is the most popular, with 81.29% of the comments. This means that most of the comments didn't say what caused the floods. However, when people did say what caused it, natural factors were the most common (9.79%), followed by human behaviour (5.77%) and changes in land use or urban planning (2.27%).

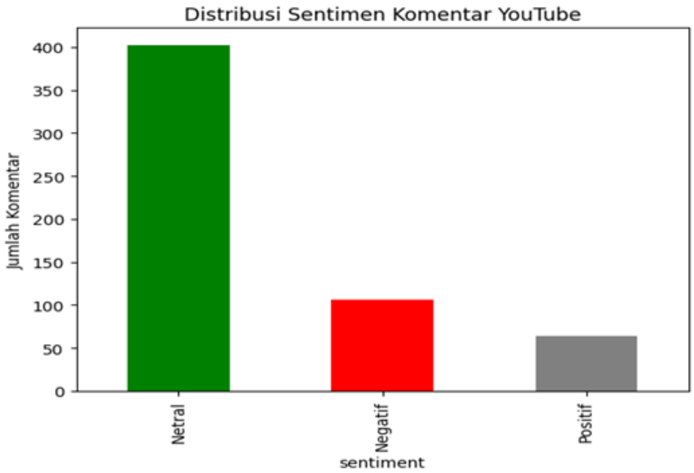

Figure 4. Youtube comment sentiment

The distribution of YouTube comment sentiment shows a predominance of neutral sentiment, indicating that most comments are informative or general discussions or do not explicitly express emotions about the flooding issue. Negative sentiment ranks second, with a significant share, reflecting public criticism, complaints, and dissatisfaction with the flooding's impact and management. Positive sentiment accounts for the smallest proportion and generally contains appreciation or hope. Overall, this pattern suggests that public discourse on flooding tends to be rational and informative, though still tinged with expressions of dissatisfaction.

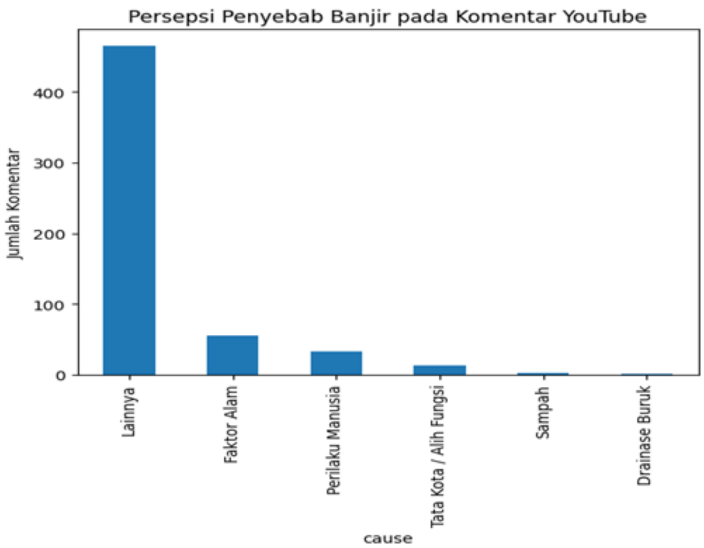

Figure 5. Sentiment perception flood of youtube comments

Most of the comments were in the "other" category, which means that most users didn't say what caused the floods. Instead, they talked about how it affected them, their own experiences, or their general ideas. Natural factors were the most common causes of flooding that people talked about. This shows that people still strongly believe that heavy rain and bad weather are to blame for flooding. People also talked about human behaviour, which shows that people are aware of how human activity can make environmental conditions worse. Urban planning or land-use change was mentioned less often, and waste and inadequate drainage were only talked about a little, even though they are very important in causing flooding. Public sentiment analysis also indicates that most people perceive natural factors as the primary cause of flooding, although scientific evidence shows that human activities play a more dominant role [21], [23].

Table 7.Comparison of analysis results of natural factors, human activities, and public sentiment

Analysis Aspects	Natural Factors	Human Activities	Public Sentiment
Data source	BMKG, BIG, BNPB	Ministry of Environment and Forestry, National Disaster Management Agency (BNPB), Global Forest Watch	YouTube public comments
Analysis Method	Random Forest + SHAP	Random Forest + SHAP	Random Forest (TF-IDF) + SHAP
Main Variables	Rainfall, distance to river, elevation	Hotspots, plantation area, mining, infrastructure	Public opinion keywords
Dominant Variable	Distance to the river	Hotspot (land burning)	Perception of natural factors & human behavior
Highest Contribution Value	Distance to river (highest SHAP)	Hotspot (Mean SHAP = 0.42)	Neutral sentiment (70.28%)
Influence Pattern	Natural & structural	Strong anthropogenic	Social-perceptual
Influence Character	Cannot be controlled directly	Human controllable	Reflection of public awareness
Role in Flooding	Basic triggers of flooding	Reinforcing & accelerating factors	Social response indicators

The Random Forest model has a high predictive power ($R^2 = 0.81$), and the SHAP analysis demonstrates that human activity variables are responsible for more than 90% of deforestation, with hotspots being the most important driver. Public opinion analysis also shows that people often think natural processes are the main cause of flooding, whereas human activities are a very important part of the problem.

These results are in line with what other studies in Southeast Asia have found, which show that rainfall, being close to rivers, and changing how land is used are all important factors that cause floods. Distance to rivers and rainfall are two of the most important factors that explainable machine learning algorithms use to make flood susceptibility maps. On the other hand, deforestation, changing land use, and building infrastructure have been demonstrated to make floods more likely by making it harder for water to soak into the ground and making it easier for water to rush off the surface.

In general, combining Random Forest with SHAP gives us both robust predictions and clear explanations of how natural and man-made factors affect the frequency of floods in Kalimantan. This method helps us understand the causes of flooding better and helps us come up with more focused and long-lasting ways to stop it. However, future studies should use higher-resolution spatial data to make the models even more accurate.

CONCLUSION

This study shows that the flooding that happens again and again in Kalimantan is caused by a combination of natural water conditions and changes in how people use the land. Being close to rivers and rain sets a baseline level of vulnerability, but deforestation caused by hotspots, plantations, and mining operations greatly increases the risk of flooding. The combination of Random Forests and SHAP makes it easier to find the main causes of floods. These results show how important it is to have rules for sustainable land management to lower the danger of flooding in the future. his study concludes that recurrent flooding in Kalimantan is influenced by both natural and anthropogenic factors, with human activities playing a dominant role in increasing flood risk. These findings highlight the importance of sustainable land management and environmental policies to mitigate future flood events [32], [33], [45].

REFERENCES

[1] N. Mahdizadeh Gharakhanlou and L. Perez, “Flood susceptible prediction through the use of geospatial variables and machine learning methods,” *J. Hydrol.*, vol. 617, p. 129121, 2023, doi: https://doi.org/10.1016/j.jhydrol.2023.129121.

[2] S. Rashidi Shikhteymour, M. Borji, M. Bagheri-Gavkosh, E. Azimi, and T. W. Collins, “A novel approach for assessing flood risk with machine learning and multi-criteria decision-making methods,” *Appl. Geogr.*, vol. 158, p. 103035, 2023, doi: https://doi.org/10.1016/j.apgeog.2023.103035.

[3] J. Li, “Machine learning-based flood susceptibility assessment integrating environmental variables,” *Sci. Total Environ.*, vol. 857, 2023, doi: https://doi.org/10.1016/j.scitotenv.2022.159684.

[4] S. Li *et al.*, “Assessing future flood risk using remote sensing and explainable machine learning: A case study in the Beijing-Tianjin-Hebei region,” *Remote Sens. Appl. Soc. Environ.*, vol. 40, p. 101742, 2025, doi: https://doi.org/10.1016/j.rsase.2025.101742.

[5] A. T. Bui, “Machine learning-based flood susceptibility mapping using geospatial variables,” *Nat. Hazards*, vol. 115, 2023, doi: https://doi.org/10.1007/s11069-022-05618-2.

[6] A. Arabameri, “Flood susceptibility mapping using machine learning models and remote sensing data,” *Sci. Total Environ.*, vol. 871, p. 162061, 2023, doi: https://doi.org/10.1016/j.scitotenv.2023.162061.

[7] S. Heo, W. Sohn, S. Park, and D. K. Lee, “RETRACTED: 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, 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e37789.

[8] Z. Wang, H. Lyu, and C. Zhang, “Urban pluvial flood susceptibility mapping based on a novel explainable machine learning model with synchronous enhancement of fitting capability and explainability,” *J. Hydrol.*, vol. 642, p. 131903, 2024, doi: https://doi.org/10.1016/j.jhydrol.2024.131903.

[9] Y. Chen, “Urban flood risk mapping using multi-source geospatial data and machine learning,” *Sci. Total Environ.*, vol. 889, p. 164172, 2023, doi: https://doi.org/10.1016/j.scitotenv.2023.164172.

[10] M. Rahmati, “Hybrid machine learning models for flood susceptibility assessment using multi-source spatial data,” *J. Hydrol.*, vol. 620, p. 129371, 2023, doi: https://doi.org/10.1016/j.jhydrol.2023.129371.

[11] X. Zhu, H. Guo, and J. J. Huang, “Urban flood susceptibility mapping using remote sensing, social sensing and an ensemble machine learning model,” *Sustain. Cities Soc.*, vol. 108, p. 105508, 2024, doi: https://doi.org/10.1016/j.scs.2024.105508.

[12] Y. Zhang, “A novel machine learning tool for flood susceptibility mapping using remote sensing and GIS,” *J. Hydrol.*, vol. 632, p. 130936, 2024, doi: https://doi.org/10.1016/j.jhydrol.2024.130936.

[13] Y. Li, “Prediction of flood risk levels of urban flooded points using machine learning with unbalanced data,” *J. Hydrol.*, vol. 630, p. 130742, 2024, doi: https://doi.org/10.1016/j.jhydrol.2024.130742.

[14] F. He, “Interpretable flash flood susceptibility mapping using AutoML,” *Sci. Rep.*, vol. 15, no. 1702, 2025, doi: https://doi.org/10.1038/s41598-024-84655-y.

[15] S. Xu *et al.*, “Urban Flood Prediction Model Based on Explainable Deep Learning and Attention Mechanism,” *Int. J. Disaster Risk Sci.*, vol. 17, no. 1, pp. 161–181, 2026, doi: 10.1007/s13753-026-00691-4.

[16] S. Lee, “Flood susceptibility assessment using ensemble machine learning algorithms,” *Catena*, vol. 230, p. 107175, 2023, doi: https://doi.org/10.1016/j.catena.2023.107175.

[17] M. Liao, H. Wen, L. Yang, G. Wang, X. Xiang, and X. Liang, “Improving the model robustness of flood hazard mapping based on hyperparameter optimization of random forest,” *Expert Syst. Appl.*, vol. 241, p. 122682, 2024, doi: https://doi.org/10.1016/j.eswa.2023.122682.

[18] T. Ngo, “Flood susceptibility modeling using random forest and extreme gradient boosting,” *Geomatics, Nat. Hazards Risk*, vol. 15, 2024, doi: https://doi.org/10.1080/19475705.2024.2357120.

[19] S. Khosravi, “Application of machine learning models in flood susceptibility mapping,” *J. Hydrol.*, vol. 623, p. 129865, 2023, doi: https://doi.org/10.1016/j.jhydrol.2023.129865.

[20] C. Singha, N. Chakraborty, S. Sahoo, Q. B. Pham, and Y. Xuan, “A novel framework for flood susceptibility assessment using hybrid analytic hierarchy process-based machine learning methods,” *Nat. Hazards*, vol. 121, no. 11, pp. 13765–13810, 2025, doi: 10.1007/s11069-025-07335-8.

[21] Y. Sun, “Explainable artificial intelligence for hydrological hazard prediction,” *Hydrol. Earth Syst. Sci.*, vol. 28, pp. 1121–1140, 2024, doi: https://doi.org/10.5194/hess-28-1121-2024.
[22] H. Pham, “Explainable machine learning models for flood hazard mapping,” *Nat. Hazards*, vol. 118, pp. 231–254, 2024, doi: https://doi.org/10.1007/s11069-023-06075-8.
[23] A. T. Bui, “Explainable machine learning models for flood susceptibility mapping,” *Sci. Total Environ.*, vol. 892, 2023, doi: . https://doi.org/10.1016/j.scitotenv.2023.164813.
[24] Q. Wei, H. Zhang, Y. Chen, Y. Xie, H. Yin, and Z. Xu, “City scale urban flooding risk assessment using multi-source data and machine learning approach,” *J. Hydrol.*, vol. 651, p. 132626, 2025, doi: https://doi.org/10.1016/j.jhydrol.2024.132626.
[25] J. Wang, J. Sanderson, S. Iqbal, and W. L. Woo, “Accelerated and Interpretable Flood Susceptibility Mapping Through Explainable Deep Learning with Hydrological Prior Knowledge,” *Remote Sens.*, vol. 17, no. 9, 2025, doi: 10.3390/rs17091540.
[26] R. Costache, “Spatial prediction of flood susceptibility using machine learning ensemble techniques,” *Appl. Geogr.*, vol. 154, p. 103090, 2023, doi: https://doi.org/10.1016/j.apgeog.2023.103090.
[27] L. Zhao, “Flood susceptibility prediction using tree-based machine learning models with SHAP interpretation,” *Sustain. Cities Soc.*, vol. 97, p. 104744, 2023.
[28] M. Bui, “GIS-based machine learning models for flood susceptibility assessment,” *Nat. Hazards*, vol. 116, pp. 1201–1223, 2023, doi: https://doi.org/10.1007/s11069-022-05759-4.
[29] L. Pham, “Hybrid machine learning models for flood hazard mapping using geospatial data,” *Nat. Hazards*, vol. 120, 2024, doi: https://doi.org/10.1007/s11069-024-06245-7.
[30] X. Lie, “Machine learning-based flood risk assessment integrating hydrological factors,” *J. Hydrol.*, vol. 636, p. 130104, 2024, doi: https://doi.org/10.1016/j.jhydrol.2024.130104.
[31] H. Wang, “Spatiotemporal flood risk analysis using remote sensing and machine learning,” *Sci. Total Environ.*, vol. 901, 2024, doi: https://doi.org/10.1016/j.scitotenv.2024.166112.
[32] Y. Wang, P. Zhang, Y. Xie, L. Chen, and Y. Cai, “Machine learning insights into the evolution of flood Resilience: A synthesized framework study,” *J. Hydrol.*, vol. 643, p. 131991, 2024, doi: https://doi.org/10.1016/j.jhydrol.2024.131991.
[33] S.-D. Jang, J.-H. Yoo, Y.-S. Lee, and B. Kim, “Flood prediction in urban areas based on machine learning considering the statistical characteristics of rainfall,” *Prog. Disaster Sci.*, vol. 26, p. 100415, 2025, doi: https://doi.org/10.1016/j.pdisas.2025.100415.
[34] A. Mosavi, P. Ozturk, and K. Chau, “Flood Prediction Using Machine Learning Models: Literature Review,” *Water*, vol. 10, no. 11, 2018, doi: 10.3390/w10111536.
[35] S. Costache, “Machine learning-based flash flood susceptibility modeling,” *Nat. Hazards*, vol. 117, 2023, doi: https://doi.org/10.1007/s11069-023-05941-9.
[36] B. Zerouali, A. H. Almaliki, and C. A. G. Santos, “Flood susceptibility mapping in arid urban areas using SHAP-enhanced stacked ensemble learning: A case study of Jeddah,” *J. Environ. Manage.*, vol. 393, p. 127128, 2025, doi: https://doi.org/10.1016/j.jenvman.2025.127128.
[37] H. Chen, “Flood risk assessment using machine learning and socio-environmental indicators,” *Int. J. Disaster Risk Reduct.*, vol. 78, no. 103123, 2022, doi: https://doi.org/10.1016/j.ijdrr.2022.103123.
[38] M. S. Al-Sheriadeh and M. A. Daqdouq, “Robustness of machine learning algorithms to generate flood susceptibility maps for watersheds in Jordan,” *Geomatics, Nat. Hazards Risk*, vol. 15, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusID:271393776
[39] M. Tehrany, “Remote sensing and machine learning for flood hazard mapping,” *Remote Sens. Environ.*, vol. 282, 2022, doi: https://doi.org/10.1016/j.rse.2022.113273.
[40] J. Chen, “Flood hazard assessment using remote sensing and explainable artificial intelligence,” *Int. J. Disaster Risk Reduct.*, vol. 92, 2023, doi: https://doi.org/10.1016/j.ijdrr.2023.103749.
[41] H. Abedi, “Flood susceptibility assessment using deep learning and remote sensing data,” *Remote Sens. Environ.*, vol. 296, 2023, doi: https://doi.org/10.1016/j.rse.2023.113750.
[42] H. Dang, “Hydro-climatic predictors for flood susceptibility modeling using machine learning,” *Hydrol. Earth Syst. Sci.*, vol. 27, pp. 2345–2364, 2023, doi: https://doi.org/10.5194/hess-27-2345-2023.
[43] F. Isma, M. S. B. Kusuma, E. O. Nugroho, and M. B. Adityawan, “Flood hazard assessment in Kuala Langsa village, Langsa city, Aceh Province-Indonesia,” *Case Stud. Chem. Environ. Eng.*, vol. 10, p. 100861, 2024, doi: https://doi.org/10.1016/j.cscee.2024.100861.
[44] Y. Zhang, “Deep learning and remote sensing for urban flood susceptibility prediction,” *Remote Sens. Environ.*, vol. 301, 2024, doi: https://doi.org/10.1016/j.rse.2024.113921.

[45] A. Costache, "Flood susceptibility modeling using ensemble machine learning techniques," *Environ. Model. Softw.*, vol. 164, p. 105720, 2023, doi: https://doi.org/10.1016/j.envsoft.2023.105720.