JOURNAL OF INFORMATION TECHNOLOGY AND ITS UTILIZATION, VOLUME 9, ISSUE 1, MAY 2026
EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.6.2.5259. SUBMITTED: JANUARY 04, 2026; REVISED: APRIL 2, 2026; ACCEPTED: APRIL 28, 2026

Earthquake Classification using Descriptive Statistics, Pearson Correlation, and Ensemble Machine Learning Model

Nunu Ariatmi[1], Alwin Sande[2], Komang Nopa Sudarma[3], Rismayani[4*]

[1,2,3,4]Department of Information Systems, Universitas Dipa Makassar, Indonesia
rismayani@undipa.ac.id
*Corresponding author

***Abstract*--This study explores earthquake classification in Indonesia, a region known for its high seismic activity due to the convergence of the Indo-Australian, Eurasian, and Pacific tectonic plates. The analysis uses earthquake data from the Meteorology, Climatology, and Geophysics Agency and covers earthquakes for the year 2024, combining descriptive statistics, Pearson correlation, and ensemble machine learning approaches. Descriptive statistics are applied to observe patterns, such as frequency, magnitude, depth, and spatial distribution of earthquakes, while Pearson correlation is used to examine relationships between variables. In addition, three ensemble models, Random Forest, XGBoost, and LightGBM, are implemented for earthquake classification. The results show that Random Forest and LightGBM achieved a perfect accuracy of 100%, whereas XGBoost reached 99.4% accuracy but showed slightly lower precision in identifying hazardous events. The correlation analysis indicates that magnitude is the most influential variable in determining hazard levels, with a correlation coefficient of r = 0.3605. In contrast, temporal variables such as hour and month contribute little. These findings highlight the importance of magnitude in earthquake classification and demonstrate the potential of combining statistical analysis.**

***Keywords*: Earthquake Classification; Machine Learning; Person Correlation; Statistics.**

I. INTRODUCTION

Indonesia is one of the countries with the highest level of seismic activity in the world due to its location at the confluence of three major tectonic plates: Indo-Australian, Eurasian, and Pacific [1]. Indonesia is a country most at risk of being affected by earthquakes. The intensity of earthquakes throughout 2022 in Indonesia was recorded at 10,792. A total of 22 earthquake events were destructive earthquakes [2]. Although the government has developed early warning systems, such as InaTEWS, their effectiveness is still questionable. Some earthquakes have triggered sirens in the absence of a tsunami, or vice versa, tsunamis have occurred without proper warning [3].

Previous research has focused on correlations between seismic parameters to accelerate magnitude estimation. Among those used are the dominant period of the P wave and the rupture duration, which have a strong correlation with the moment magnitude. A local study in North Sumatera produced a linear regression equation with an R^2 of 71.96% [4] and the national study yielded a correlation of 76.2% [5]. However, most of these studies only rely on conventional statistical methods, such as linear regression, and have not integrated more complex models to capture nonlinear relationships.

In the context of technology, the use of machine learning, especially ensemble methods such as Random Forest, XGBoost, and LightGBM, has shown superior predictive performance. One study compared five models for earthquake magnitude classification and found that LightGBM had the best accuracy, followed by Random Forest and XGBoost [6]. In addition, the XGBoost model was successfully used for tsunami potential classification based on earthquake data with an accuracy of up to 85%, especially after data balancing using the SMOTE method [2].

On the other hand, text-based approaches, such as natural language processing (NLP), are also being used, for example, to analyze the distribution of earthquake locations from Twitter data, which can complement official seismic data and support real-time visualization of earthquake distribution [7].

However, most studies have not thoroughly combined descriptive statistics, Pearson correlation, and ensemble machine learning methods in one predictive framework. Therefore, this study aimed to develop an earthquake magnitude classification model in Indonesia by

JOURNAL OF INFORMATION TECHNOLOGY AND ITS UTILIZATION, VOLUME 9, ISSUE 1, MAY 2026
EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.6.2.5259. SUBMITTED: JANUARY 04, 2026; REVISED: APRIL 2, 2026; ACCEPTED: APRIL 28, 2026

utilizing earthquake depth and geolocation data, using a combination of descriptive statistics, Pearson correlation, and random forest, XGBoost, and LightGBM models.

This research is expected to provide benefits in the form of improving the accuracy of early warning systems, supporting decision-making in disaster mitigation, and providing scientific contributions through the integration of statistical and machine learning methods that are adaptive to the characteristics of earthquakes in Indonesia.

II. Method

The research methodology is designed to develop a predictive model for classifying earthquake events in Indonesia using a binary classification approach based on seismic data obtained from the Meteorology, Climatology, and Geophysics Agency and covers earthquakes. The model aims to distinguish between earthquakes and non-earthquakes by using magnitude along with spatial and temporal characteristics. To ensure a systematic and reliable process, the study is conducted through several interconnected stages, including data collection, preprocessing, model development, validation, and result interpretation.

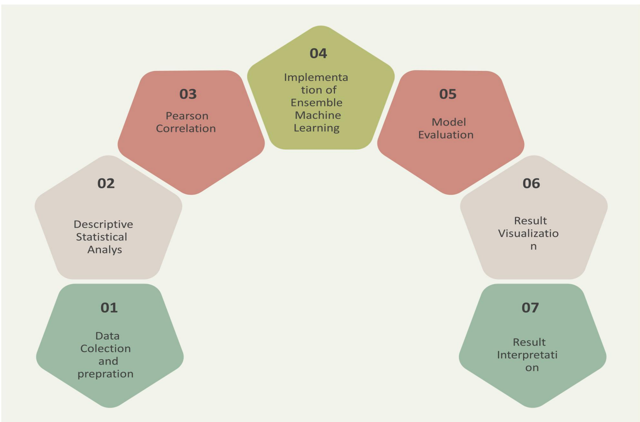

Fig. 1. Research methodology stage

This research aims to predict earthquakes in Indonesia using a combination of statistical and machine learning approaches. There are five main stages carried out: descriptive statistical analysis, Pearson correlation analysis, and implementation of three ensemble learning models (Random Forest, XGBoost, and LightGBM).

A. Data Collection and Preprocessing

The dataset used in this study was obtained from the Meteorology, Climatology, and Geophysics Agency and covers earthquake events recorded throughout 2024. It includes detailed information for each event, such as the date and time of occurrence (GMT), geographic coordinates (latitude and longitude), focal depth, and magnitude.

The first step in this research is to explore the data using descriptive statistical methods. This stage aims to understand the characteristics of earthquake data based on variables such as magnitude, depth, time, and earthquake location. The statistical results provide an initial overview, such as the mean, minimum, maximum, and standard deviation of each variable. This approach is in line with research using descriptive statistics as an initial step in modeling earthquake frequency in the Sumatera subduction zone [8].

Before modeling, several data preprocessing steps were obtained. First, the dataset was filtered by date to retain only events within the specified period, ensuring that the model focused on a consistent temporal context. Next, incomplete records or entries containing missing (null) values were removed to maintain the dataset's reliability. The date and time columns were standardized to facilitate the derivation of new temporal attributes, specifically month and hour, which were considered potentially relevant in identifying earthquake patterns.

Next, numerical features were normalized using Z-score standardization based on the following Equality (1):

$$Z = \frac{X - \mu}{\sigma} \quad (1)$$

where X is the original value, μ is the average, and σ is the standard deviation of the feature.

Finally, the dataset was labeled for classification as follows:

$$\text{Label} = 1 \text{ if Magnitude} \geq 5.0,$$
$$0 \text{ if Magnitude} < 5.0$$

This threshold follows commonly accepted standards for earthquake classification.

B. Descriptive Statistical Analysis

Descriptive statistical analysis was performed to summarize the key numerical features in the dataset, including depth, magnitude, latitude,

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.6.2.5259. SUBMITTED: JANUARY 04, 2026; REVISED: APRIL 2, 2026; ACCEPTED: APRIL 28, 2026

longitude, month, and hour. That step provides an overview of the distribution, tendency of the data center, and potential extreme values that can affect the performance of the classification model.

Some statistical measures calculated include average (mean) used to determine the middle value of each feature equality (2):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} xi \quad (2)$$

where $\bar{x}$ is the mean value of the feature, xi represents the i observation, and n is the total number of observations.

The results of this analysis were calculated automatically using the built-in describe () function in the Pandas library and visualized tabularly to display the minimum, maximum, quartile, mean, and standard deviation of each numerical feature. In addition, this analysis helps in assessing the normality of the data distribution, identifying features that have high or low variation, and detecting initial statistical anomalies before further transformation. This step also underlies the decision to apply Z-score normalization so that the model can process data on a uniform scale and avoid the dominance of one feature over another.

C. Pearson correlation analysis

After performing the descriptive statistical analysis, the next step was to assess the linear relationship between numerical features, specifically to determine which features were most correlated with the hazard labels. For this purpose, the Pearson correlation method, which is commonly used in analyzing relationships between quantitative variables, was used.

The results of this analysis are used as the basis for feature selection for machine learning modeling. Pearson correlation was also used in the study to select relevant features to improve the performance of earthquake classification models using random forest [9].

The Pearson correlation coefficient was calculated to evaluate the linear relationship between features and the target label. The correlation coefficient between variables X and Y is given equality (3):

$$r_{XY} = \frac{\sum_{i-1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (3)$$

where r_{XY} is Pearson correlation coefficient between variable X and Y, xi, yi is the i value of variable X and Y, $\bar{x}, \bar{y}$ is average of the variables Y and n is the amount of data.

This analysis is performed to measure the strength and direction of the relationship between features, identify the most relevant features for classification, and avoid features that are redundant or uncorrelated with the target. Correlation calculations are performed using the Corr() function on pandas data frames, and the results are visualized in the form of heatmaps using the Seaborn library. The calculation results show that the magnitude feature has the highest correlation to the hazard label, with r equal to 0.3605, followed by the depth and latitude features, which have relatively low correlation values. Time features such as month and hour show weak or even statistically insignificant relationships. This correlation visualization is used as the basis for feature selection at the modeling stage, so that only relevant and nonoverlapping features are included in the classification algorithm.

D. Dataset Preparation for Modeling

The random forest algorithm is used to build a binary classification model between dangerous and nondangerous earthquakes. Random Forest is a decision tree-based ensemble learning method that works by voting from many decision trees. It has the advantage of being able to handle unbalanced data and prevent overfitting [10].

After preprocessing and correlation analysis, the next step is to prepare the dataset for use in training the classification model. This stage includes feature selection, training and test data division, and feature scale normalization.

1. Feature Selection

   Based on the results of the Pearson correlation analysis and consideration of the earthquake domain, the six most relevant numerical features for modeling were selected as cross (°), adventure (°), depth (KM), magnitude (M), month, and hour. These features reflect the spatial (location), seismic (magnitude and depth), and temporal (month and hour) characteristics of the earthquake event.
2. Dataset Split

   The dataset is divided into two subsets:
   - Training: 70% of data.
   - Testing: 20% of data.

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.6.2.5259. SUBMITTED: JANUARY 04, 2026; REVISED: APRIL 2, 2026; ACCEPTED: APRIL 28, 2026

- Validation: 10% data.

This division was stratified by label so that the proportion of earthquakes and nonearthquakes remained equal in both subsets.

$$Training\ Data\ Size = 0{,}7 \times N,$$
$$Testing\ Data\ Size = 0{,}2\ N, Validation\ Data\ Size = 0{,}1 \times N$$

where N is the total number of samples in the dataset. The technical implementation utilizes the train_test_split function from the sklearn.model_selection module:

$$X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, stratify = y)$$

3. Feature Normalization (Z-Score Standardization)

Before the model is trained, the numerical features are first normalized to be on a uniform scale. It is important to avoid the dominance of features with large absolute values on the model learning process. Normalization is performed using the Z-score technique, using equality (4):

$$Z = X - \mu\sigma Z = \frac{X-\mu}{\sigma} Z = \sigma X - \mu \qquad (4)$$

with:
X = original value of the feature
μ = average feature value
σ = standard deviation of features

This normalization is performed using Standard Scaler from the sklearn preprocessing library (import train test split). This step ensures that all features have a mean of zero and a standard deviation of one, so that the model is not biased toward any particular scale.

E. Implementation of Ensemble Learning Models

At this stage, the processed data are used to build a binary classification model using an ensemble learning approach. The three algorithms applied are Random Forest, XGBoost, and LightGBM, which were chosen for their proven ability to handle classification problems with complex and imbalanced data.

Each model was trained using normalized training data, with default parameters as the baseline for initial evaluation.

Random Forest is a bagging algorithm that forms a set of decision trees and combines the results through majority voting. This model is effective in reducing overfitting and has robustness to outliers.

$$from\ sklearn.ensemble\ import\ Random\ ForestClassifier$$
$$model\_rf = RandomForestClassifier()$$
$$model\_rf.fit(X\_train\_scaled, y\_train$$

XGBoost (Extreme Gradient Boosting) was used due to its ability to speed up training and improve accuracy through boosting. The model was trained using cleaned and normalized earthquake data, with 80% data split for training and 20% for testing.

Parameters such as learning rate, max depth, and estimators were adjusted using grid search. The research successfully showed that XGBoost could produce magnitude classifications with a very small RMSE (0.01) and R^2 of 0.98 on test data [11].

LightGBM (Light Gradient Boosting Machine) is a histogram-based boosting algorithm designed for high efficiency. LightGBM uses a leaf-wise growth strategy, which allows for flexible tree depth and more accurate classification results. The LightGBM algorithm was used as a comparison with XGBoost and Random Forest. LightGBM has high efficiency in training and classification on large datasets. LightGBM was implemented for earthquake frequency analysis using data from the Sumatera subduction zone and showed strong classification performance through the comparison of four data sharing scenarios 80:20 to 95:5 as well as MAAPE and RMSE evaluation [12]. LightGBM is optimized for speed; it uses histogram-based learning and leaf-wise tree growth for greater efficiency. LightGBM characteristics tree growth based on maximum gain information at each leaf. Faster than XGBoost in the case of large and high-dimensional data. All models were initially trained using default parameters as a performance baseline.

F. Model Evaluation and Validation

After the model training process, the next step is to perform evaluation and validation to measure the performance of each ensemble algorithm in classifying earthquakes. The evaluation was conducted using test data that were not used

JOURNAL OF INFORMATION TECHNOLOGY AND ITS UTILIZATION, VOLUME 9, ISSUE 1, MAY 2026
EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.6.2.5259. SUBMITTED: JANUARY 04, 2026; REVISED: APRIL 2, 2026; ACCEPTED: APRIL 28, 2026

during training to provide an objective picture of the generalization ability of the model.

The performance evaluation is based on the classification results represented in the confusion matrix. The following are the classification matrix Equality (5)–(8) used:

a. Accuracy
Shows the proportion of correct classifications over all data:
$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+F} \quad (5)$$

b. Precision
Indicates how much of the earthquake class classification is actually correct:
$$\text{Precision} = \frac{TP}{TP+} \quad (6)$$

c. Recall
Measures the ability of the model to detect all earthquake events:
$$\text{Recall} = \frac{TP}{TP+FN} \quad (7)$$

d. F1-Score
The harmonic means of precision and recall:
$$F1 = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (8)$$

e. Mean Squared Error (MSE)
Measure of numerical classification error equality (9):
$$MSE = 1n \sum i = 1n\left(yi - y^i\right) \quad (9)$$

f. K-fold and Temporal Validation

Evaluation was conducted on the test set, which was not used during model training, to ensure objective validation. To address potential bias from a single train, test split, this study employs a more robust validation strategy using k-fold cross-validation (k=5). The dataset is partitioned into five subsets, with four folds used for training and one for testing in each iteration. Final performance is reported as the average across all folds.

Additionally, due to the temporal aspect of earthquake data, a temporal validation scheme is employed. The dataset is organized chronologically, with earlier data used for training and later data for testing. This method ensures that the model is tested in a realistic forecasting context and minimizes the chance of information leakage from future data. Unlike conventional evaluations that use a single data partition, the model performance in this study was assessed using holdout validation and k-fold cross-validation. The use of cross-validation helps ensure that reported performance is independent of a specific data partition and provides more stable estimates of model generalization.

In k-fold cross-validation, the dataset $\mathcal{D}$ is partitioned into k equally sized subsets (folds):
$$\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \ldots, \mathcal{D}_k\} \quad (10)$$

For each iteration $i = 1,2, \ldots, k$ one-fold $\mathcal{D}_i$ is used as the validation set, whereas the remaining $k-1$ folds are used for training:
$$\mathcal{D}_{train}^{(i)} = \frac{\mathcal{D}}{\mathcal{D}_i} \quad (11)$$
$$\mathcal{D}_{test}^{(i)} = \mathcal{D}_i \quad (12)$$

The model is trained and evaluated across all k iterations, producing performance scores $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \ldots, \mathcal{M}_k$. The overall performance is then calculated as the average of all folds:
$$\mathcal{M}_{avg} = \frac{1}{k} \sum_{i-1}^{k} \mathcal{M}_i \quad (13)$$

Given that earthquake data is inherently time-dependent, a temporal validation strategy is applied to preserve chronological order and simulate real-world prediction scenarios. The dataset used is a series arranged chronologically:
$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

where x_i represents the feature vector, y_i the corresponding label, and t_i the timestamp of the $i-th$ observation. The dataset is then split into training, testing, and validation sets based on a temporal cutoff index mmm, such that
$$\mathcal{D}_{train} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$$
$$\mathcal{D}_{test} = \{(x_{m+1}, y_{m+1}), \ldots, (x_n, y_n)\}$$

where:
$$m = [\alpha \times n], 0 < \alpha < 1$$

and α denotes the proportion of data used for training ($\alpha = 0.8$).

This temporal formulation ensures that the evaluation framework reflects real-world application conditions, where future earthquake events must be predicted using only past observations.

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.6.2.5259. SUBMITTED: JANUARY 04, 2026; REVISED: APRIL 2, 2026; ACCEPTED: APRIL 28, 2026

G. Visualization and Interpretation of Results

Visualization of evaluation results is an important step in analyzing the performance of the applied classification model. The purpose of this visualization is to make it easier to understand the performance of each model, describe the accuracy and precision of the model in classifying earthquakes and non-earthquakes, and explain the influence of each feature on the classification process.

1. Model Performance Comparison

   Evaluation metrics such as accuracy, precision, recall, and F1-score are compared in the form of bar graphs for the three models: Random Forest, XGBoost, and LightGBM. The visualization results show that Random Forest and LightGBM achieved a perfect F1 score of 1.00 for both hazardous and non-hazardous classes. XGBoost has a slightly lower performance in the dangerous class, especially on the precision metric, although recall remains high. This comparison reinforces that Random Forest and LightGBM are more stable and accurate in handling unbalanced datasets like this case (earthquakes are only about 2.3% of the total data).
2. Comparison of Mean Squared Error (MSE)

   Visualization of Mean Squared Error (MSE) values shows that Random Forest and LightGBM recorded very low MSE, whereas XGBoost had a slightly higher MSE value, indicating a certain amount of class classification error. The low MSE values indicate that the classifications are very close to the actual labels, which is important in the context of risk classification.
3. Correlation between Features

   Prior to model training, a Pearson correlation analysis between features was conducted. These results were visualized in the form of a heatmap, which shows the magnitude feature has the strongest correlation to the hazard label r = 0.3605. Other features, such as depth and latitude, have low correlations, negative correlations, and close to zero indicate that these features have less direct influence on the classification label. This finding provides a strong foundation for the feature selection process for machine learning models.
4. Feature Importance Analysis

   Visualization of feature importance generated from Random Forest and LightGBM models shows that magnitude is the dominant feature in determining classification, followed by depth, latitude, and hour of occurrence. The moon feature has a relatively small contribution. This visualization confirms the importance of seismic factors (magnitude and depth) in distinguishing between hazardous and non-hazardous earthquakes, which is also consistent with previous correlation results and literature.

III. Result and Discussion

A. Model Performance Comparison

This study applied three ensemble learning algorithms, namely, Random Forest, XGBoost, and LightGBM, to earthquake classification events into two categories: class 1 earthquake and class 0 non- earthquake. The dataset consisted of 1570 earthquake records, with a highly imbalanced class distribution: 1534 records, 97.7%, were labeled as non- earthquake and only 36 records, 2.3%, as earthquake. The performance of the three models is summarized in Table I.

TABLE I
Confusion Matrix for Classification Model Evaluation

Model	**Class**	**Precision**	**Recall**	**F1-Score**	**Support**
Random Forest	0	1.00	1.00	1.00	307
	1	1.00	1.00	1.00	7
Accuracy				**1.00**	314
XGBoost	0	1.00	0.99	1.00	307
	1	0.78	1.00	0.88	7
Accuracy				**0.994**	314
LightGBM	0	1.00	1.00	1.00	307
	1	1.00	1.00	1.00	7
Accuracy				**1.00**	314

The results are presented comprehensively to assess the relative merits of each model in dealing with complex earthquake data characteristics. The selection of these models is based on their ability to manage data with non-linear, unbalanced, and multidimensional properties.

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.6.2.5259. SUBMITTED: JANUARY 04, 2026; REVISED: APRIL 2, 2026; ACCEPTED: APRIL 28, 2026

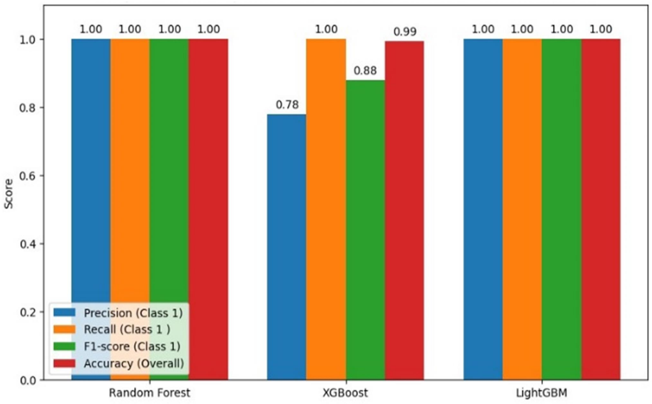

Fig. 2. Comparison of the Classification Matrix

B. Random Forest

The Random Forest model produced a perfect accuracy of 100%, with precision, recall, and F1-score values of 1.00 each for both classes (hazardous and non-hazardous). This shows that the model successfully predicted all earthquake events correctly, including all hazardous events, without any misclassification.

This result is in line with the findings of [13], who reported a low RMSE value of 0.31758 when using Random Forest for earthquake magnitude classification. Moreover, Random Forest achieved high classification accuracy, ranging from 83.77% to 89.66% in earthquake classification in Indonesia, demonstrating its effectiveness in handling complex seismic data patterns [10].

The study also confirmed that Random Forest is a method that consistently provides high accuracy in disaster classification [14]. This finding is consistent with the study which states that ensemble-based statistical models, such as Random Forest, are very effective for earthquake nowcasting applications in areas with high geological diversity, such as Sulawesi [15]. They emphasized that the complex statistical characteristics of earthquakes in the region are better handled by ensemble methods than traditional methods.

C. XGBoost

XGBoost displays a total accuracy of 99.36%, but the precision for the dangerous class is only 0.78, whereas the recall remains perfect at 1.00. This indicates some false positive classifications, i.e. events classified as dangerous when they were not. The F1 score also dropped to 0.88. This shows that although all earthquake events were successfully detected with recall = 1.00, the model also generated some false positives. This result is in line with the findings of [16], who observed that XGBoost provides near-optimal predictive performance MAE = 0.4932; RMSE = 0.6471, but tends to be more sensitive to class imbalance in the dataset.

Research also observed similar challenges, showing that XGBoost can perform well in earthquake classification, but its performance is more dependent on class distribution [17]. This result highlights the weakness of XGBoost in handling class imbalance, especially when the minority data is much less than the majority. XGBoost is known for its ability to adjust the weights of each feature and build the model incrementally, but without proper parameter adjustments, the model can exhibit classification imbalance.

Models such as XGBoost can be very accurate, but require supporting techniques such as balancing and optimal parameter tuning to achieve maximum performance, especially in the classification of geophysical disasters, such as tsunamis and earthquakes [18]. In real disaster scenarios, model performance must be prioritized for recall so that no dangerous events escape detection, even if it has to sacrifice a little precision.

D. LightGBM

The LightGBM model also produced perfect scores for precision, recall, F1-score, and accuracy 100%. This advantage of LightGBM can be attributed to its efficient model in handling large datasets and numerical features, such as earthquake magnitude and depth. LightGBM uses a histogram-based decision tree approach, which speeds up the training process and makes it more efficient than XGBoost on a large scale. LightGBM has the advantage of exploring complex relationships between geospatial and geophysical features, thus providing very precise classification results in the context of disaster classification. With the right parameter settings and selection of relevant features, LightGBM is an ideal candidate for use in natural disaster early detection systems [18].

E. Interpretation of the Results

To determine the influence of each feature on

JOURNAL OF INFORMATION TECHNOLOGY AND ITS UTILIZATION, VOLUME 9, ISSUE 1, MAY 2026
EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.6.2.5259. SUBMITTED: JANUARY 04, 2026; REVISED: APRIL 2, 2026; ACCEPTED: APRIL 28, 2026

the classification results, a Pearson correlation analysis between variables in the dataset was conducted. The results were visualized in the form of a heatmap. The correlation coefficient values between numerical features are presented in Table II.

TABLE II
Pearson Correlation between Numerical Features

	Latitude (°)	**Longitude (°)**	**Depth (Km)**	**Magnitude (M)**	**Month**	**Day**	**Hour**	**Label**
Latitude (°)	1.000.000	0.053245	-0.026014	0.141116	-0.0043	-0.0195	0.0389	0.0454
Longitude (°)	0.053245	1.000.000	0.100978	0.163540	0.0364	0.0393	0.0176	0.0013
Depth (Km)	0.026014	0.100978	1.000.000	0.183253	0.0427	0.0161	0.0128	0.0144
Magnitude (M)	0.141116	0.163540	0.183253	1.000.000	0.0769	0.0706	0.0302	0.3605
Month	-0.0043	0.0364	0.0427	0.0769	10.000	-0.2075	0.0622	0.0102
Day	-0.0195	-0.0393	-0.0161	-0.0706	-0.2075	10.000	0.0037	-0.0443
Hour	0.0389	-0.0176	-0.0128	-0.0302	0.0622	0.0037	10.000	-0.0489
Label	0.0454	-0.0013	0.0144	**0.3605**	0.0102	-0.0443	-0.0489	10.000

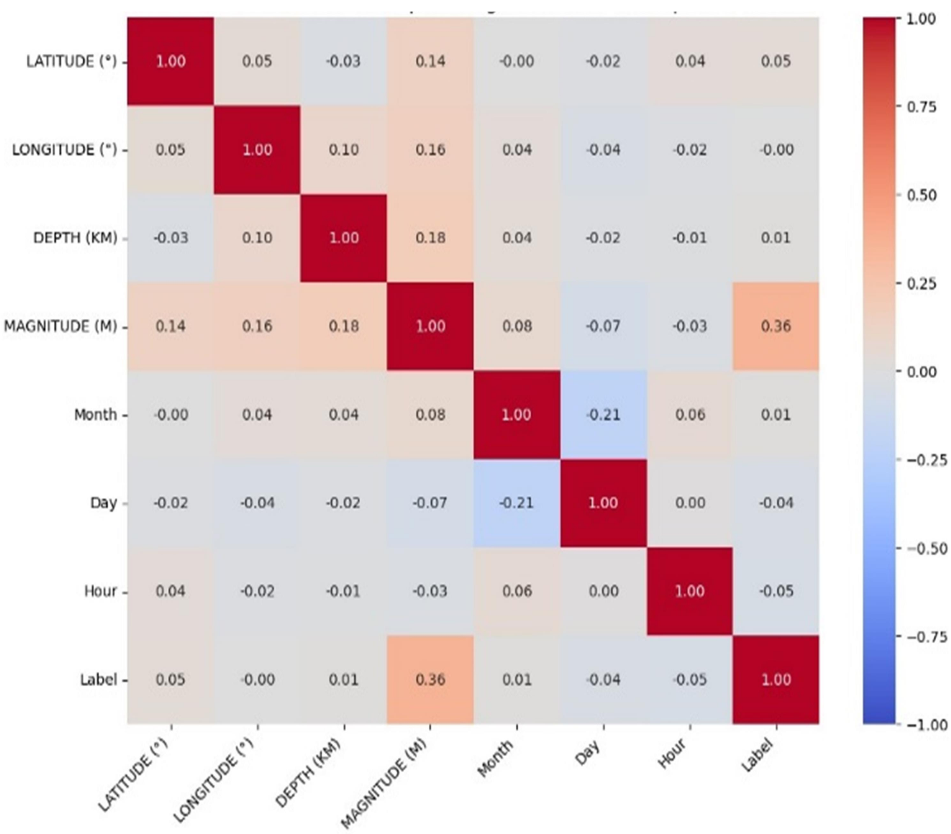

Fig. 3. Heatmap of Pearson correlation between features and earthquakes

Table II shows that:

- Magnitude has the strongest correlation to the hazardous earthquake classification label, with a correlation value of r = 0.36. This suggests that the higher the magnitude, the more likely the earthquake is to be categorized as an earthquake.
- Depth has a very low correlation r = 0.014, whereas latitude also shows a small contribution r = 0.045.
- Temporal features such as hour, day, and month have correlations close to zero, suggesting that the timing of the event on a calendar scale is not a dominant factor in distinguishing earthquake hazard levels.

This finding is reinforced by the research of who used singular spectrum analysis (SSA) to model earthquake frequency in the Sumatera subduction zone [19]. They found that magnitude is the main factor affecting the dynamics of

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.6.2.5259. SUBMITTED: JANUARY 04, 2026; REVISED: APRIL 2, 2026; ACCEPTED: APRIL 28, 2026

earthquake occurrence, and temporal factors do not show significant patterns in predicting frequency or hazard levels. In this context, the validation of correlations between features reinforces the importance of relevant feature selection in building accurate and efficient classification models.

The correlation of seismicity levels in the Mentawai Islands showed that the frequency of large earthquakes is more influenced by magnitude and geographical location than by the time factor [20]. They used the Guttenberg-Richter method to show that the potential seismic hazard can be estimated from the seismicity index and the "b" value of the magnitude distribution parameter. The results provide additional evidence that the magnitude feature is the most significant determinant in earthquake hazard modeling.

In addition to the visualization of feature correlations, additional evaluations were conducted on model performance using accuracy and mean squared error (MSE) metrics. The results are displayed in Fig. 4, which shows that the Random Forest and LightGBM models achieved perfect performance with 100% accuracy and MSE = 0.0000, whereas the XGBoost model had a slightly lower accuracy of 99.36%, with an MSE of 0.0064.

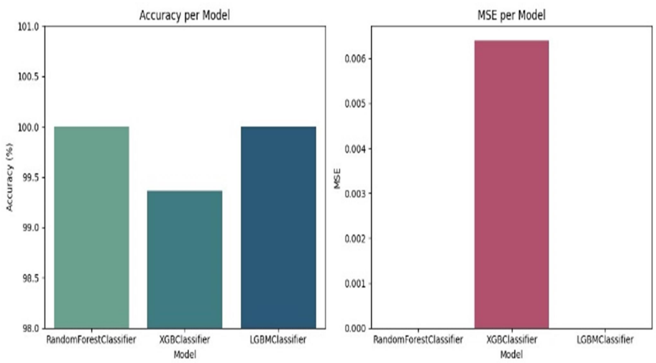

Fig. 4. Comparison of the Performance Model in Earthquake Hazard Status Classification.

The performance shows that while XGBoost remains highly accurate, it tends to produce slight misclassifications, especially on imbalanced data, as also indicated by the lower precision values in the previous evaluation. The average combined accuracy of the three models reached 99.79%, and the average MSE was 0.0021, indicating high stability and effectiveness in predicting hazardous earthquakes.

Fig. 4 shows a comparison of the Mean Squared Error (MSE) values of the three ensemble models. Both Random Forest and LightGBM recorded very low MSE values, indicating that the classifications from both models are very close to the actual class labels. In contrast, the XGBoost model shows a higher MSE value, indicating a larger amount of classification error.

This result supports the findings shown in Table III and Fig. 4, where XGBoost has a lower precision value of 0.78 for the hazardous earthquake class. This pattern reported that LightGBM achieved the lowest RMSE value of 0.6284 [16], followed by Random Forest with an RMSE of 0.31758. The XGBoost with an RMSE of 0.6471. This finding confirms that LightGBM and Random Forest are more reliable in producing accurate and stable classifications, especially when used on earthquake datasets with high class imbalance. This shows that tree-based ensemble models such as CatBoost, LightGBM, and XGBoost work optimally if the features used are relevant and the risk of data leakage can be controlled [17].

This result reinforces the findings that emphasized the importance of a balance between accuracy and generalization ability in disaster classification models [18]. In the context of early warning systems, models such as Random Forest and LightGBM are proven to be accurate and stable in the face of unbalanced class distributions, as exemplified in the study for the context of earthquakes in Sulawesi [15].

Therefore, in addition to considering feature correlations in Fig. 3, the accuracy and MSE comparisons in Fig. 4 provide an overall picture of the reliability of each model in real classification practice.

F. Challenges and Limitations

Although the model performs very well, there are some important constraints that can affect the interpretation and generalization of the result. The number of earthquake data in the dataset is very small, only 7 out of 314, creating a significant class imbalance. This increases the risk of classification bias toward the majority class, where the model tends to classify more events as

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.6.2.5259. SUBMITTED: JANUARY 04, 2026; REVISED: APRIL 2, 2026; ACCEPTED: APRIL 28, 2026

non-hazardous. This imbalance can cause evaluation metrics, such as accuracy, to become unrepresentative; thus, it is necessary to apply balancing techniques, such as SMOTE, to overcome this problem, in the context of tsunami classification [18]. Models like XGBoost have a tendency to overfit the training data if not accompanied by proper cross-validation or pruning techniques. This can result in excellent performance on training data but poor performance on new data. In addition, since the dataset only covers a specific region and period, external testing is needed using data from other regions, such as the Sumatera or Sulawesi megathrust zones, that have different geological characteristics. Research shows that earthquake dynamics in Sulawesi, although different from Sumatera, still show consistency in earthquake statistical patterns, so the model needs to be tested across regions to test its predictive stability [15]. Some features available in the have very low relevance to classification. These features can add complexity without significantly contributing to the accuracy of the model. Therefore, it is important to perform statistical feature selection or use feature engineering methods to improve predictive relevance. Feature tuning is also important to avoid high dimensionality that can worsen model performance and trigger the "curse of dimensionality" phenomenon. Local geological conditions influence earthquakes, such as rock types, active fault systems, and the depth of subduction zones. Research discusses that while statistical data, such as magnitude and depth, play an important role, local spatial dynamics also influence the occurrence pattern [19]. The combination of more detailed spatial and temporal data allows the model to learn from the context of the area and not just from the numerical values of the features. Considering these challenges, the development of earthquake classification models should not only focus on basic performance metrics but also on the validation of generalizability across regions, class balancing, and selection of physically and geologically relevant features. A thorough evaluation of this approach would not only strengthen the reliability of the model but also support decision-making in disaster risk mitigation nationwide in Indonesia. Through a comprehensive and sustainable approach, this earthquake classification system is expected to be used as part of an adaptive and accurate natural disaster early warning system.

While the models showed perfect accuracy on the hold-out test set, cross-validation offers a more reliable assessment. The average accuracy across different folds is a bit lower but more consistent, indicating that the initial perfect results might be affected by dataset peculiarities or sampling bias. This underscores the necessity of using robust validation methods in earthquake classification, particularly when working with imbalanced and small datasets.

A limitation of this study is its dependence on a single training-test data split, which can result in overly optimistic performance estimates and restrict the model's ability to generalize. Although we added validation with k-fold cross-validation and temporal splitting, the small dataset size and class imbalance continue to challenge robust generalization. Future research should use larger, multi-year datasets and validate externally across various seismic regions.

IV. CONCLUSION

This study compares the performance of three ensemble machine learning models: Random Forest, XGBoost, and LightGBM in classifying earthquake events in Indonesia into hazardous and non-hazardous categories. The results show that Random Forest and LightGBM achieve perfect accuracy of 100%, whereas XGBoost achieves 99.4% accuracy but is slightly less accurate in identifying hazardous events. Pearson correlation analysis indicates that magnitude is the most influential factor in determining earthquake hazard, with $r = 0.3605$, confirming its central role in the classification process. In contrast, temporal variables such as hour and month show little or no significant contribution in distinguishing hazard levels, suggesting that time-based features are less relevant in this context. Although Random Forest and LightGBM performed admirably, additional validation through cross-validation and temporal testing demonstrated that model performance may fluctuate under varying data distributions. Consequently, the proposed methodology exhibits significant potential; however, further validation

EISSN 2654-802X ; PISSN 2985-4067
DOI: https://doi.org/10.56873/jitu.6.2.5259. SUBMITTED: JANUARY 04, 2026; REVISED: APRIL 2, 2026; ACCEPTED: APRIL 28, 2026

is required across larger and more diverse datasets.

V. ACKNOWLEDGMENT

We would like to thank Universitas Dipa Makassar. We would like to express our gratitude for all the support provided to the authors in completing this research. We also thank those who provided information and data availability.

VI. REFERENCES

[1] D. Hr, "Korelasi Magnitudo Gempa Bumi Lokal Dengan Periode Dominan Gelombang P Untuk Peringatan Dini Tsunami," vol. 1, no. 1.

[2] I. G. J. Kurniarwan, C. Dewi, and M. A. Rahman, "Penerapan Machine Learning Extreme Gradient Boosting Dalam Klasifikasi Potensi Tsunami Berdasarkan Data Gempa Bumi."

[3] D. I. Purnamasari, "Analisis Korelasi Antara Magnitudo Momen Gempa Bumi Regional Dengan Periode Dominan Gelombang P Di Wilayah Indonesia," vol. 04, 2015.

[4] D. Y. Kurniawati, "Korelasi Periode Dominan (Td) Dengan Magniotudo Momen (Mw) Untuk Gempa Bumi Lokal Di Wilayah Sumatera Utara."

[5] D. I. Purnamasari, "Analisis Korelasi Antara Magnitudo Momen Gempa Bumi Regional Dengan Periode Dominan Gelombang P Di Wilayah Indonesia," vol. 04, 2015.

[6] I. Maulita and A. M. Wahid, "Prediksi Magnitudo Gempa Menggunakan Random Forest, Support Vector Regression, XGBoost, LightGBM, dan Multi-Layer Perceptron Berdasarkan Data Kedalaman dan Geolokasi," *Jurnal Pendidikan dan Teknologi Indonesia*, vol. 4, no. 5, pp. 221–232, Dec. 2024, doi: 10.52436/1.jpti.470.

[7] J. A. Nursiyono and R. K. Gibran, "Natural Language Processing for Unstructured Data: Earthquakes Spatial Analysis in Indonesia Using Platform Social Media Twitter," *IIRoi*, vol. 5, no. 1, Mar. 2023, doi: 10.37058/innovatics.v5i1.6678.

[8] T. Turino, R. E. Saputro, and G. Karyono, "Penerapan Model Ensemble Learning dengan Random Forest dan Multi-Layer Perceptron untuk Prediksi Gempa," *Jurnal Pendidikan dan Teknologi Indonesia*, vol. 5, no. 2, Feb. 2025, doi: 10.52436/1.jpti.667.

[9] H. Tantyoko, D. K. Sari, and A. R. Wijaya, "Prediksi Potensial Gempa Bumi Indonesia Menggunakan Metode Random Forest Dan Feature Selection," *IDEALIS*, vol. 6, no. 2, pp. 83–89, Jul. 2023, doi: 10.36080/idealis.v6i2.3036.

[10] R. M. A. Hutagaol, V. T. Lana, Z. A. Dzunnurain, and R. Kurniawan, "Penerapan Machine Learning dalam Prediksi Klasifikasi Big Data Kedalaman Gempa Bumi di Indonesia Tahun 2015-2024," *SN.SD*, vol. 4, no. 1, pp. 42–51, Sep. 2024, doi: 10.33005/senada.v4i1.156.

[11] A. Wibowo, "Prediksi Kekuatan Gempa Menggunakan Machine Learning Dengan Model Xgboost Sebagai Langkah Strategis Dalam Perencanaan Struktur Bangunan Tahan Gempa Di Indonesia," vol. 6, no. 1, 2022.

[12] V. N. Islamiati, S. Supardiono, Y. H. Perdana, and A. R. Setyahagi, "Analisis Peluruhan Gempa Bumi Susulan Di Ambon Tahun 2019 Dengan Pendekatan Statistik Menggunakan Software Peluruhan V2.0," *IFI*, vol. 9, no. 2, pp. 163–172, Jul. 2020, doi: 10.26740/ifi.v9n2.p163-172.

[13] A. Fauzan and D. Ahmad, "Analisis Hasil Prediksi Magnitudo Gempa Di Wilayah Kota Padang Menggunakan Teknik Random Forest," *SCI TECH ED MATH*, vol. 4, no. 3, pp. 1569–1576, Dec. 2023, doi: 10.46306/lb.v4i3.450.

[14]"Apriani_2021_J._Phys.__Conf._Ser._1951_012057."

[15] S. Pasari, A. V. H. Simanjuntak, Neha, and Y. Sharma, "Nowcasting earthquakes in Sulawesi Island, Indonesia," *Geosci. Lett.*, vol. 8, no. 1, p. 27, Dec. 2021, doi: 10.1186/s40562-021-00197-5.

[16] I. Maulita and A. M. Wahid, "Prediksi Magnitudo Gempa Menggunakan Random Forest, Support Vector Regression, XGBoost, LightGBM, dan Multi-Layer Perceptron Berdasarkan Data Kedalaman dan Geolokasi," *Jurnal Pendidikan dan Teknologi Indonesia*, vol. 4, no. 5, pp. 221–232, Dec. 2024, doi: 10.52436/1.jpti.470.

[17] Y. Zhao and D. Gorse, "Earthquake classification from seismic indicators using tree-based ensemble learning," *Nat Hazards*, vol. 120, no. 3, pp. 2283–2309, Feb. 2024, doi: 10.1007/s11069-023-06221-5.

[18] H. Prasetyo and A. E. Maulana, "Penerapan Algoritma Support Vector Machine untuk Melakukan Prediksi Potensi Tsunami di Indonesia."

[19] S. Yosmar, S. Damayanti, and J. Rizal, "The Application of Singular Spectrum Analysis in Modeling Earthquake Frequency in the Sumatera Subduction Zone," *MMEP*, vol. 12, no. 2, pp. 533–548, Feb. 2025, doi: 10.18280/mmep.120218.