ISSN:2302-8149
e-ISSN:2540-9719

Spatial Variability Analysis of Earthquake Magnitude and Depth in Indonesia using Random Forest based on Historical Data

[1]Holmes Heriyanto Silalahi*, [2]Anief Fauzan Rozi
[1,2]Information Systems, Faculty of Information Technology, Universitas Mercu Buana Yogyakarta, Indonesia 55283
*e-mail: holmesheriyantosilalahi@gmail.com

(***received***: 3 April 2026, ***revised***: 24 June 2026, ***accepted***: 28 June 2026)

Abstract

Indonesia lies at the convergence of three major tectonic plates, resulting in high seismic activity and complex spatial earthquake patterns. This study analyzes the spatial variability of earthquake magnitude and depth in Indonesia using a Random Forest Regressor based on historical earthquake data from the United States Geological Survey (USGS). A total of 24,981 earthquake records (1976–2026) were collected, and 21,886 valid observations were obtained after data cleaning. To reduce data heterogeneity caused by different magnitude measurement methods, the dataset was grouped into six magnitude types (Mb, Ms, Mw, Mwb, Mwc, and Mww), and separate Random Forest models were developed for each group using longitude and latitude as input features. The novelty of this study lies in the magnitude-type-specific modeling approach, which allows spatial patterns to be analyzed more consistently across different earthquake measurement scales. The models were evaluated using 10-fold cross-validation and standard regression metrics. Results show that spatial coordinates have limited ability to explain earthquake magnitude, as indicated by low R^2 values across all models. In contrast, depth prediction shows better performance, with the best result achieved by Mw (R^2 = 0.6801), indicating stronger spatial structure in earthquake depth distribution. Overall, the findings demonstrate that geographic location is more informative for modeling earthquake depth than magnitude. The proposed approach effectively captures spatial depth variability and provides a more structured analysis of earthquake characteristics across Indonesia.

Keywords: earthquake depth, earthquake magnitude, multi-output regression, random forest, spatial analysis, usgs.

1 Introduction

Indonesia is geographically located at the convergence of three major tectonic plates, resulting in a high level of seismic activity [1]. Historical earthquake records indicate significant variations in both magnitude and depth across different regions of the country [2]. Information regarding earthquake magnitude and depth plays a crucial role in seismic hazard analysis and risk assessment. With the advancement of technology and the increasing availability of open-access seismic data from sources such as the United States Geological Survey (USGS), data-driven approaches have become increasingly feasible for analyzing earthquake patterns [3]. Among these approaches, machine learning methods, particularly the Random Forest algorithm have demonstrated strong capability in modeling complex and non-linear relationships among variables [4].

In earthquake datasets, magnitude is commonly measured using different magnitude types, including body-wave magnitude (Mb), surface-wave magnitude (Ms), moment magnitude (Mw), body-wave moment magnitude (Mwb), centroid moment magnitude (Mwc), and W-phase moment magnitude (Mww). These magnitude types are derived using different measurement techniques and seismic wave characteristics. For example, Mb is calculated from body waves, Ms is based on surface waves, while moment magnitude-based measurements (Mw, Mwb, Mwc, and Mww) are derived from seismic moment estimations using different computational approaches and waveform observations. As a result, earthquake events represented by different magnitude types may exhibit distinct characteristics and spatial distributions. Ignoring these differences may introduce heterogeneity into the dataset and potentially reduce the effectiveness of data analysis. Therefore, understanding the spatial characteristics associated with different magnitude types is an important aspect of earthquake studies.

ISSN:2302-8149
e-ISSN:2540-9719

This study investigates the spatial variability of earthquake magnitude and depth in Indonesia using historical earthquake data obtained from USGS. The analysis focuses on the relationship between geographic location, represented by longitude and latitude coordinates, and earthquake characteristics, represented by magnitude and depth. To accommodate the differences among magnitude types, the dataset is grouped according to six dominant magnitude classifications, namely Mb, Ms, Mw, Mwb, Mwc, and Mww. Separate Random Forest models are then developed for each magnitude type to analyze their respective spatial patterns. This approach is intended to reduce data heterogeneity and provide a more detailed understanding of how geographic location is associated with variations in earthquake magnitude and depth across Indonesia. It should be noted that the present study does not consider temporal, geological, tectonic, or other environmental factors that may influence seismic activity. Consequently, the findings should be interpreted within the scope of a spatial analysis based solely on historical earthquake coordinates.

Overall, this study contributes to the application of machine learning techniques in geospatial and seismic data analysis. By examining spatial variability across different magnitude types, the research provides insights into the distribution of earthquake magnitude and depth throughout Indonesia. The findings are expected to enhance the understanding of regional seismic characteristics and provide a foundation for future studies involving more comprehensive geospatial and seismological parameters.

2 Literature Review

Numerous studies have explored the application of data-driven and machine learning approaches for earthquake analysis in Indonesia. Various types of data have been utilized, including large-scale historical datasets, spatial attributes such as latitude, longitude, and depth, as well as time-series seismic records [5-7]. In addition, several studies have employed datasets obtained from global and open-access sources such as the USGS, which provide standardized and comprehensive earthquake records suitable for geospatial and seismic analysis [8].

A wide range of methods has been implemented to investigate earthquake-related phenomena. Machine learning algorithms such as Random Forest, Support Vector Regression, XGBoost, LightGBM, Multi-Layer Perceptron, and Decision Tree have been widely used and compared in terms of analytical and predictive performance [6, 9-11]. In addition, clustering techniques such as DBSCAN and K-Means have been applied to identify seismic zones and earthquake-prone areas, while time-series models such as ARIMA have been utilized to analyze seismic trends and temporal patterns [5, 7, 12-13]. Some studies have also proposed hybrid approaches that combine multiple algorithms, such as Random Forest and Multi-Layer Perceptron, to improve model performance and capture more complex earthquake patterns [14]. The objectives and outputs of these studies vary considerably, including earthquake magnitude estimation, frequency forecasting, seismic risk classification, and spatial mapping of earthquake-prone regions. Overall, the findings indicate that machine learning models, particularly Random Forest and its variants, are effective in modeling complex relationships within earthquake datasets, as demonstrated through evaluation metrics such as MAE, RMSE, R^2, and F1-score [5, 9, 15]. Furthermore, several studies highlight the potential of machine learning techniques to support disaster mitigation efforts, seismic hazard assessment, and the identification of earthquake-prone regions [7-8].

Despite these advancements, previous studies generally focus on predictive performance, classification tasks, clustering methods, or temporal earthquake analysis. Comparatively fewer studies investigate the spatial variability of earthquake characteristics based primarily on geographic coordinates. In addition, earthquake datasets often contain multiple magnitude types, which may exhibit different spatial distributions and seismic characteristics. However, these magnitude types are frequently analyzed as a single homogeneous dataset, potentially obscuring important spatial patterns. Consequently, studies that specifically examine spatial variability across different magnitude types remain limited, particularly in the context of Indonesian earthquake data.

This study addresses this gap by examining the spatial variability of earthquake magnitude and depth across Indonesia using historical USGS data and the Random Forest algorithm. The analysis is conducted using longitude and latitude as the primary spatial variables, while earthquake magnitude and depth are treated as the target variables. Furthermore, the dataset is separated according to six dominant magnitude types, allowing each group to be analyzed independently. This approach reduces

Sistemasi: Jurnal Sistem Informasi
Volume 15, Nomor 7, 2026: 2512-2524
ISSN:2302-8149
e-ISSN:2540-9719

data heterogeneity and enables a more comprehensive understanding of how earthquake magnitude and depth vary spatially across Indonesia. The findings are expected to contribute to geospatial earthquake analysis and provide insights into regional seismic characteristics based on different magnitude measurement methodologies.

3 Research Method

This study adopts a quantitative approach by applying machine learning techniques to analyze the spatial variability of earthquake magnitude and depth in Indonesia using historical earthquake data. The research workflow consists of several sequential stages, beginning with data collection and preprocessing, followed by dataset grouping based on magnitude type, model development using the Random Forest algorithm, model evaluation through statistical metrics and visual analysis, and case study analysis using selected sample locations. The overall research framework is illustrated in Fig 1.

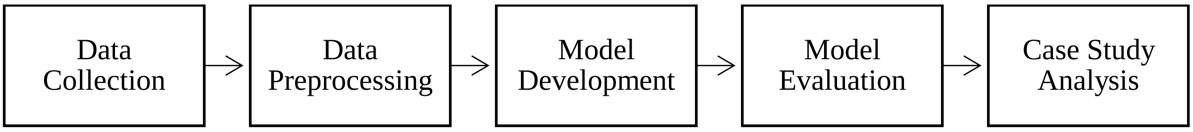

Fig 1 Research method framework

3.1. Data Collection

The dataset used in this study was obtained from the United States Geological Survey (USGS) in the form of open-access earthquake records. The dataset includes key attributes such as geographical coordinates (longitude and latitude), earthquake magnitude, depth, and magnitude type. To ensure relevance, the data were filtered to include only earthquake events occurring within the geographical boundaries of Indonesia. The study focuses on six dominant magnitude types available in the dataset, namely Mb, Ms, Mw, Mwb, Mwc, and Mww. These magnitude types were selected because they represent the majority of earthquake records and provide sufficient observations for model development and analysis.

3.2. Data Preprocessing

Data preprocessing was conducted to improve data quality and ensure the reliability of the analysis. This stage involved data cleaning, removal of duplicate records, handling of missing values, and consistency checking. Records containing incomplete information in the selected attributes were excluded from further analysis. After the cleaning process, the dataset was grouped according to magnitude type. This grouping was performed to reduce data heterogeneity caused by differences in magnitude measurement methodologies and to enable a more detailed analysis of spatial variability within each magnitude type. Subsequently, each magnitude-type dataset was divided into training and testing subsets using an 80:20 ratio. The training data were used to construct the Random Forest models, while the testing data were used to evaluate model performance on unseen observations.

3.3. Model Development

The analytical model was developed using the Random Forest Regressor algorithm implemented in Python through the scikit-learn library. As shown in Fig. 2, Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions to improve accuracy and reduce overfitting [16-17]. As shown in equation (1), the prediction of the Random Forest model can be expressed as [18]:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(x) \tag{1}$$

where, $\hat{y}$ is the predicted value, N is the number of decision trees, $T_i\,(x)$ represents the prediction of the i^{th} decision tree for input x.

ISSN:2302-8149
e-ISSN:2540-9719

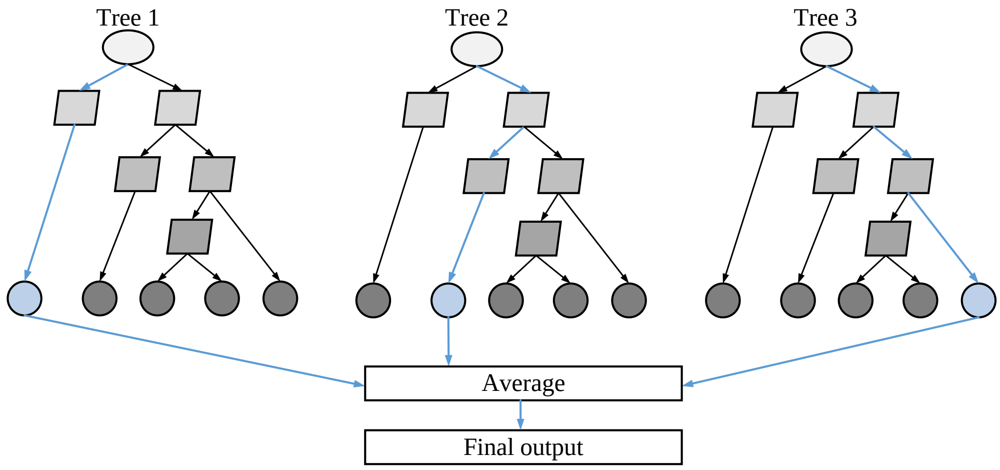

Fig 2 Scheme of random forest method [19]

In this study, longitude and latitude were used as the input variables, while earthquake magnitude and depth were used as the target variables. Rather than constructing a single model for the entire dataset, six independent Random Forest models were developed corresponding to the six magnitude types. This approach allows the models to capture spatial patterns specific to each magnitude type and reduces the influence of heterogeneity resulting from different magnitude measurement methods. The developed models learn the relationships between geographical coordinates and earthquake characteristics based on historical earthquake records across Indonesia.

3.4. Model Evaluation

The performance of the models was evaluated using standard regression metrics, namely Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). MAE measures the average absolute error between actual and predicted values, RMSE provides greater penalties for larger prediction errors, and R^2 indicates the proportion of variance explained by the model. Together, these metrics provide a comprehensive assessment of model accuracy and reliability. As shown in equation (2) to (4), the MAE, RMSE, and R^2 are defined as [20]:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

$$\text{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{4}$$

where, y_i is the actual value, $\hat{y}_i$ is the predicted value, $\bar{y}_i$ is the mean of actual values, n is the number of observations.

To assess model robustness and generalization capability, a 10-fold cross-validation procedure was applied during model training. In this approach, the dataset was divided into ten subsets, where nine subsets were used for training and one subset was used for validation. The process was repeated ten times so that each subset served as the validation set once. The mean and standard deviation of the cross-validation R^2 scores were calculated to evaluate the consistency and stability of the Random Forest models across different magnitude types. In addition to numerical evaluation metrics, Actual versus Predicted plots were generated for both earthquake magnitude and depth across all magnitude-type-specific models. These plots compare the observed values with the corresponding values estimated by the Random Forest models. Data points located near the diagonal reference line indicate strong agreement between actual and predicted values, whereas larger deviations suggest prediction errors or unexplained variability. The visual analysis provided by these plots complements MAE, RMSE, and R^2 by enabling a more intuitive assessment of model behavior, prediction consistency, and potential outliers within each magnitude-type dataset.

 ISSN:2302-8149
e-ISSN:2540-9719

3.5. Case Study Analysis

To demonstrate the practical application of the developed models, a case study analysis was conducted using five sample locations representing different geographical regions in Indonesia. The longitude and latitude coordinates of these locations were used as input variables for each of the six Random Forest models corresponding to the magnitude types. For each sample location, the models generated estimates of earthquake magnitude and depth based on the spatial patterns learned from historical earthquake records. The resulting outputs were then compared and analyzed to identify potential spatial differences among regions and magnitude types. This analysis was not intended to forecast future earthquake events but rather to illustrate how the developed models capture spatial variability in earthquake characteristics across Indonesia. The case study provides additional interpretation of the model outputs and supports the examination of regional seismic characteristics based on geographical coordinates and magnitude-type-specific patterns.

4 Results and Discussion

4.1. Data Collection

A total of 24,981 earthquake records were collected from the USGS within the Indonesian region, covering a 50-year observation period from 1976 to 2026. Only earthquake events with magnitudes greater than or equal to 4.8 were included in the analysis to ensure data consistency and focus on events with greater seismic significance. The collected dataset contains information on earthquake location, magnitude, depth, and magnitude type, which were subsequently used for model development and spatial variability analysis. Fig 3 presents a sample of the historical earthquake records obtained from the USGS database. After the data collection and filtering processes, six magnitude types were identified as the dominant categories within the dataset, namely Mb, Ms, Mw, Mwb, Mwc, and Mww.

time	latitude	longitude	depth	magnitude	magType	nst	gap	dmin	rms	net	id	updated	place	type	horizontalError	depthError	magError	magNst	status	locationSource	magSource
2026-03-16T23:06:42.276Z	-7.0385	129.6327	157.734	4.8	mb	81	45	2.514	0.8	us	us6000sgqm	2026-03-17T00:10:30.040Z	Kepulauan Babar, Indonesia	earthquake	11.91	5.863	0.065	72	reviewed	us	us
2026-03-16T11:53:25.233Z	-2.2256	138.9878	10	4.9	mb	50	69	1.738	0.57	us	us6000sgku	2026-03-16T12:25:22.040Z	187 km WNW of Abepura, Indonesia	earthquake	8.34	1.855	0.07	65	reviewed	us	us
2026-03-16T11:41:58.609Z	-9.2403	118.1226	79.634	4.9	mb	47	73	0.534	0.84	us	us6000sgkr	2026-03-16T12:41:25.875Z	86 km SSW of Dompu, Indonesia	earthquake	7.39	5.602	0.092	37	reviewed	us	us
2026-03-11T02:12:13.489Z	-0.5352	121.8871	10	5	mb	47	68	4.934	1.02	us	us7000s3jg	2026-03-11T02:30:38.040Z	110 km WNW of Luwuk, Indonesia	earthquake	9.97	1.874	0.082	48	reviewed	us	us
2026-03-10T21:45:36.137Z	-9.4883	120.5923	49.917	5	mb	60	59	2.857	0.86	us	us7000s3i3	2026-03-11T02:42:20.040Z	40 km ENE of Waingapu, Indonesia	earthquake	11.13	5.228	0.071	64	reviewed	us	us
2026-03-06T20:51:28.711Z	-9.2498	120.1385	44.17	4.9	mb	39	61	2.37	0.76	us	us7000s2n1	2026-03-07T18:27:48.040Z	47 km NNW of Waingapu, Indonesia	earthquake	5.81	8.604	0.088	40	reviewed	us	us
2026-03-04T18:08:05.896Z	-3.1822	130.9385	10	4.9	mb	50	47	1.334	0.64	us	us7000s215	2026-03-04T20:11:10.040Z	153 km W of Fakfak, Indonesia	earthquake	6.24	1.79	0.082	47	reviewed	us	us
2026-03-04T06:44:25.033Z	-3.2467	130.9963	10	4.9	mb	53	56	1.292	1.07	us	us7000s1xq	2026-03-04T07:22:30.040Z	149 km WSW of Fakfak, Indonesia	earthquake	6.79	1.881	0.08	49	reviewed	us	us
2026-03-02T17:40:14.559Z	-7.541	127.6711	139.681	5	mww	64	60	3.732	0.78	us	us7000s1gx	2026-03-14T19:01:50.040Z	131 km NE of Lospalos, Timor Leste	earthquake	8.65	6.852	0.083	14	reviewed	us	us
2026-03-01T18:33:23.740Z	2.7568	128.2361	143.254	4.8	mb	79	106	5.036	0.94	us	us7000s19v	2026-03-14T17:08:12.040Z	116 km NNE of Tobelo, Indonesia	earthquake	10.24	7.239	0.055	101	reviewed	us	us
2026-02-25T17:49:30.311Z	-0.392	98.9797	61.488	4.9	mb	48	160	2.194	0.87	us	us7000s09j	2026-03-10T21:32:44.040Z	129 km W of Pariaman, Indonesia	earthquake	6.54	7.77	0.087	41	reviewed	us	us
2026-02-19T07:02:16.000Z	-5.6588	133.7698	10	4.9	mb	53	75	3.115	0.37	us	us6000sa54	2026-03-14T05:49:37.040Z	112 km E of Tual, Indonesia	earthquake	7.1	1.876	0.082	47	reviewed	us	us
2026-02-17T18:00:52.699Z	-8.2804	108.1062	85.027	4.9	mb	83	95	0.93	0.59	us	us6000s9u1	2026-03-04T15:59:04.327Z	100 km S of Kawalu, Indonesia	earthquake	6.81	5.771	0.07	65	reviewed	us	us
2026-02-16T23:47:21.153Z	-5.6101	133.7994	10	5	mb	65	65	3.088	0.44	us	us6000s9mt	2026-03-03T02:37:39.040Z	116 km E of Tual, Indonesia	earthquake	9.59	1.799	0.07	66	reviewed	us	us
2026-02-16T11:42:15.432Z	-8.5584	118.2753	134.386	4.8	mb	53	68	0.561	0.65	us	us6000s9hg	2026-03-10T14:52:35.040Z	20 km W of Dompu, Indonesia	earthquake	6.52	5.95	0.083	45	reviewed	us	us
2026-02-16T06:00:41.678Z	-0.4045	99.3142	80.621	5	mb	99	60	2.429	0.59	us	us6000s9gm	2026-03-12T12:22:42.040Z	92 km WNW of Pariaman, Indonesia	earthquake	7.73	6.495	0.059	92	reviewed	us	us
2026-02-14T14:21:51.948Z	-2.9274	128.0985	35	4.9	mb	41	77	2.399	0.57	us	us6000s97v	2026-03-09T13:41:26.040Z	85 km N of Ambon, Indonesia	earthquake	6.99	1.818	0.069	65	reviewed	us	us
2026-02-14T12:23:07.187Z	-3.0904	139.8249	10	4.9	mb	37	58	1.047	0.83	us	us6000s97d	2026-03-01T07:04:09.040Z	105 km WSW of Abepura, Indonesia	earthquake	8.23	1.874	0.069	65	reviewed	us	us
2026-02-14T03:09:46.682Z	-2.959	127.8342	35	4.8	mb	40	102	2.584	1.28	us	us6000s95i	2026-02-28T23:37:25.040Z	90 km NNW of Ambon, Indonesia	earthquake	11.06	1.924	0.095	36	reviewed	us	us
2026-02-14T01:02:18.346Z	4.3005	126.6484	66.71	4.8	mb	72	92	2.949	1.04	us	us6000s94b	2026-02-28T23:11:32.040Z	179 km SE of Sarangani, Philippines	earthquake	9.89	7.163	0.062	82	reviewed	us	us
2026-02-13T17:39:08.174Z	4.2371	128.2355	8.65	4.9	mb	111	74	3.861	0.92	us	us6000s91e	2026-03-11T17:05:44.040Z	278 km N of Tobelo, Indonesia	earthquake	9.2	4.157	0.052	115	reviewed	us	us

Fig 3 Sample of historical data by USGS

Table 1 summarizes the distribution of magnitude types across different magnitude ranges. Overall, Mb constitutes the largest proportion of the dataset (71.38%), followed by Mwc (13.46%) and Mww (7.14%). Meanwhile, Mw, Mwb, and Ms represent smaller portions of the total observations. This distribution indicates that body-wave magnitude measurements dominate historical earthquake records in Indonesia, particularly for lower-magnitude events. A more detailed examination reveals a notable shift in magnitude-type composition as earthquake magnitude increases. For earthquakes within the 4.8–5.0 range, Mb accounts for approximately 88.68% of observations, indicating that smaller earthquake events are predominantly recorded using body-wave magnitude measurements. However, the proportion of Mb decreases substantially as magnitude increases, reaching only 3.33% for earthquakes greater than magnitude 6.0. In contrast, the proportions of Mw, Mwc, and Mww increase progressively with magnitude. For earthquakes exceeding magnitude 6.0, Mww (26.67%), Mw (25.07%), and Mwc (24.40%) become the dominant magnitude types. This trend reflects the common seismological practice of utilizing moment-based magnitude measurements for larger earthquakes because they provide more reliable representations of seismic energy release than traditional body-wave magnitude scales.

 ISSN:2302-8149
e-ISSN:2540-9719

Table 1 Distribution of magnitude data

Magnitude Range	Percentage (%)					
	Mb	Ms	Mw	Mwb	Mwc	Mww
All	71.38	0.97	5.03	2.01	13.46	7.14
4.8 to ≤ 5.0	88.68	0.36	0.3	0.03	7.21	3.43
5.0 to ≤ 5.2	66.39	0.86	4.52	0.66	19.4	8.17
5.2 to ≤ 5.5	41.64	1.51	14.94	4.91	25.62	11.37
5.5 to ≤ 6.0	19.78	3.44	20.16	11.26	25.46	19.89
> 6.0	3.33	5.47	25.07	15.07	24.4	26.67

Table 2 presents the distribution of magnitude types across different earthquake depth intervals. Similar to the magnitude distribution, Mb dominates shallow and intermediate-depth earthquakes, accounting for approximately 68–82% of observations between 0 and 300 km depth. The dominance of Mb gradually increases with depth up to approximately 250 km, indicating that body-wave magnitude measurements are widely used for earthquakes occurring within the subduction zones surrounding Indonesia. Another notable observation is the absence of Mb records for earthquakes deeper than 300 km, where the dataset is dominated by Mw, Mwc, and Mww measurements. This pattern suggests that deep-focus earthquakes are more frequently characterized using moment-based magnitude scales, which are generally more suitable for describing large and complex seismic sources. The increasing proportion of Mw-related magnitude types at greater depths further highlights the heterogeneity of earthquake measurements within the dataset.

Table 2 Distribution of depth data

Depth (km)	Percentage (%)					
	Mb	Ms	Mw	Mwb	Mwc	Mww
0 to ≤ 50	68.53	1.54	4.35	2.24	15.58	7.70
50 to ≤ 100	74.72	0.30	6.27	1.74	11.38	5.58
100 to ≤ 150	74.18	0.00	5.69	1.60	10.19	8.34
150 to ≤ 200	80.27	0.00	5.01	1.49	7.99	5.25
200 to ≤ 250	81.59	0.00	4.66	1.40	9.32	3.03
250 to ≤ 300	75.90	0.00	4.22	1.20	10.84	7.83
> 300	0.00	0.00	7.72	1.98	9.60	8.04

These findings demonstrate that the composition of magnitude types varies considerably across both magnitude and depth categories. Such variability indicates that different magnitude scales are associated with distinct earthquake characteristics and observation conditions. Consequently, treating all magnitude types as a single homogeneous dataset may obscure important spatial patterns. This observation supports the methodological decision of this study to develop separate Random Forest models for each magnitude type, thereby enabling a more detailed analysis of spatial variability in earthquake magnitude and depth across Indonesia.

4.2. Data Preprocessing

The preprocessing stage was performed to improve data quality and ensure the reliability of subsequent analyses. Several procedures were applied, including duplicate removal, magnitude-type filtering, missing-value inspection, and dataset partitioning. The initial dataset consisted of 24,981 earthquake records collected from the USGS database. During the cleaning process, 3,086 duplicate records were identified and removed, reducing the dataset to 21,895 observations. Subsequently, records containing magnitude types outside the scope of this study were excluded, resulting in a final dataset of 21,886 earthquake events. A summary of the preprocessing results is presented in Table 3. A missing-value inspection was then conducted on the selected attributes, namely longitude, latitude, magnitude type, magnitude, and depth. The results showed that no missing values were present in any of the selected variables. Consequently, no additional imputation procedures were required, and all remaining records were retained for further analysis.

ISSN:2302-8149
e-ISSN:2540-9719

Table 3 Summary of data preprocessing

Processing Stage	Number of Records
Initial dataset	24,981
Duplicate records removed	3,086
Remaining after duplicate removal	21,895
Records removed after magnitude-type filtering	9
Final dataset	21,886

After data cleaning, the distribution of earthquake records across the six magnitude types is presented in Table 4. Mb represents the largest category with 15,304 records, accounting for approximately 69.93% of the final dataset. Mwc and Mww constitute the second and third largest groups with 3,003 and 1,613 records, respectively. In contrast, Ms and Mwb contain relatively fewer observations, which may influence model stability and predictive performance in subsequent analyses. Following the grouping process, each magnitude-type dataset was divided into training and testing subsets using an 80:20 ratio. This partitioning strategy ensures that sufficient data are available for model learning while maintaining an independent testing set for performance evaluation. The resulting training and testing datasets for each magnitude type are summarized in Table 5.

Table 4 Distribution of cleaned dataset by magnitude type

Distribution	Magnitude Type						Total
	Mb	Ms	Mw	Mwb	Mwc	Mww	
Number of Records	15,304	230	1,236	500	3,003	1,613	21,886
Percentage (%)	69.93	1,05	5,65	2,28	13,72	7,37	100,00

Table 5 Training and testing data distribution

Distribution	Magnitude Type					
	Mb	Ms	Mw	Mwb	Mwc	Mww
Total Data	15,304	230	1,236	500	3,003	1,613
Training Data	12,243	184	988	400	2,402	1,290
Testing Data	3,061	46	248	100	601	323

The preprocessing results indicate that the final dataset is dominated by Mb observations, whereas Ms, Mw, and Mwb contain substantially fewer samples. This imbalance reflects the natural distribution of earthquake magnitude measurements in historical Indonesian earthquake records. Therefore, separating the dataset by magnitude type provides a more consistent basis for analyzing spatial variability and developing magnitude-type-specific Random Forest models.

4.3. Model Development

The Random Forest Regressor algorithm was implemented using the scikit-learn library in Python. Each model was trained separately using longitude and latitude as predictor variables, while earthquake magnitude and depth were simultaneously used as target variables in a multi-output regression framework. In this study, each model was configured with a total of 500 decision trees to ensure stable ensemble learning and reduce variance in predictions. The maximum depth of each tree was limited to 20 to prevent overfitting while still allowing the model to capture nonlinear spatial patterns. In addition, a node must contain at least 5 samples before it can be split further, while each terminal leaf node contains a minimum of 2 samples, improving generalization performance. The models were trained using random_state = 42 to ensure reproducibility of results, and n_jobs = -1 was set to utilize all available CPU cores for parallel computation, thereby improving computational efficiency.

During training, the Random Forest algorithm constructs multiple decision trees using bootstrap samples of the training data. Each tree is trained on a randomly selected subset of the data and learns the relationship between geographic coordinates and earthquake characteristics independently. The final prediction is obtained by averaging the outputs of all trees for regression tasks, as defined in Equation (1). Through this ensemble learning approach, the model is able to capture complex

ISSN:2302-8149
e-ISSN:2540-9719

nonlinear relationships between spatial variables and earthquake magnitude and depth while reducing the risk of overfitting. The model development process resulted in six trained Random Forest models, each consisting of 500 decision trees with controlled tree depth and splitting constraints, which were subsequently used for performance evaluation and case study analysis.

4.4. Model Evaluation

The performance of the developed Random Forest models was evaluated using 10-fold cross-validation and several regression metrics, including MAE, RMSE, and R^2. In addition, Actual versus Predicted plots were analyzed to visually assess the agreement between observed and estimated values.

4.4.1 Cross-Validation Results

To assess model robustness and generalization capability, 10-fold cross-validation was applied to each magnitude-type dataset during model training. The resulting mean and standard deviation of the R^2 scores are presented in Table 6. The cross-validation results indicate that most magnitude-type-specific models achieved positive mean R^2 values, suggesting that geographic coordinates contain useful spatial information related to earthquake characteristics. Among all models, Mw produced the highest mean R^2 value (0.3562), followed by Mb (0.3399) and Mwc (0.3029). These results indicate that the spatial patterns associated with these magnitude types are relatively more consistent and can be learned effectively by the Random Forest algorithm.

Table 6 Cross-Validation (CV) results of random forest models

Results	Magnitude Type					
	Mb	**Ms**	**Mw**	**Mwb**	**Mwc**	**Mww**
CV Mean R^2	0.3399	-0.1369	0.3562	0.1886	0.3029	0.2903
CV Std R^2	0.0257	0.1179	0.0705	0.1017	0.0512	0.0593

The Ms model produced a negative mean R^2 value (-0.1369), which may be attributed to the limited number of available observations. With only 230 records, the model has fewer examples from which to learn stable spatial relationships, resulting in weaker generalization performance during validation. Nevertheless, the relatively low standard deviation values observed for most magnitude types indicate that model performance remains reasonably consistent across different folds.

4.4.2 Prediction Performance

The predictive performance of the developed models was evaluated on the testing dataset for both earthquake magnitude and depth as summarized in Table 7. For earthquake magnitude estimation, all models achieved relatively low MAE values ranging from 0.1583 to 0.5865 magnitude units. The Mb model produced the lowest prediction error with an MAE of 0.1583 and an RMSE of 0.2112, indicating that the predicted magnitudes generally differ from observed values by around 0.2 magnitude units on average. The Ms model exhibited the largest prediction errors, which is consistent with its limited dataset size. Furthermore, the corresponding R^2 values are generally close to zero or negative across all magnitude types. The negative R^2 values obtained in some magnitude models indicate that the model performs worse than a baseline mean predictor. This suggests that spatial coordinates alone are insufficient to explain the variability of earthquake magnitude.

Table 7 Prediction performance of random forest models

Parameters	Performance	Magnitude Type					
		Mb	**Ms**	**Mw**	**Mwb**	**Mwc**	**Mww**
Magnitude	MAE	0.1583	0.5865	0.3620	0.3566	0.3115	0.3888
	RMSE	0.2112	0.7620	0.4694	0.4526	0.4335	0.5070
	R^2	-0.0139	-0.2333	-0.0353	-0.0067	-0.0036	0.0233
Depth	MAE	29.3787	6.4624	36.5776	32.8599	24.2389	32.6275
	RMSE	58.5599	8.9373	74.0104	70.5768	50.9291	77.7868
	R^2	0.6685	-0.0410	0.6801	0.3565	0.5691	0.4637

ISSN:2302-8149
e-ISSN:2540-9719

For earthquake depth estimation, the models generally achieved stronger predictive performance than for magnitude estimation. The highest depth prediction performance was obtained by the Mw model with an R^2 value of 0.6801, closely followed by Mb (0.6685) and Mwc (0.5691). These results indicate that spatial coordinates contain more distinguishable information related to earthquake depth than to earthquake magnitude. The superior performance observed for depth prediction may be associated with the tectonic structure of Indonesia, where subduction zones and seismic belts generate spatially identifiable depth patterns. In contrast, earthquake magnitude is influenced by a wider range of geological processes that may not be fully represented by geographic coordinates alone.

4.4.3 Actual versus Predicted Analysis

To complement the quantitative evaluation metrics, Actual versus Predicted plots were generated for both earthquake magnitude and depth across all six magnitude types. Fig. 4 and Fig. 5 illustrate the relationship between observed values and the corresponding predictions produced by the Random Forest models. The red dashed diagonal line represents the ideal condition where predicted values exactly match the actual observations. For earthquake depth, the plots generally demonstrate a clear positive relationship between actual and predicted values. The Mb, Mw, Mwc, and Mww models exhibit data distributions that follow the diagonal reference line reasonably well, indicating that the models successfully captured meaningful spatial patterns associated with earthquake depth. This observation is consistent with the relatively high R^2 values obtained for depth prediction, particularly for Mw (0.6801), Mb (0.6685), and Mwc (0.5691). The Mwb depth model also demonstrates a positive relationship between actual and predicted values, although with a larger dispersion around the reference line. This behavior is reflected in its lower R^2 value (0.3565) compared with the other depth models. Meanwhile, the Ms depth model shows a scattered distribution with no strong alignment along the diagonal line, indicating limited predictive capability. The relatively poor performance of the Ms model is likely influenced by the small dataset size, which contains only 230 observations after preprocessing.

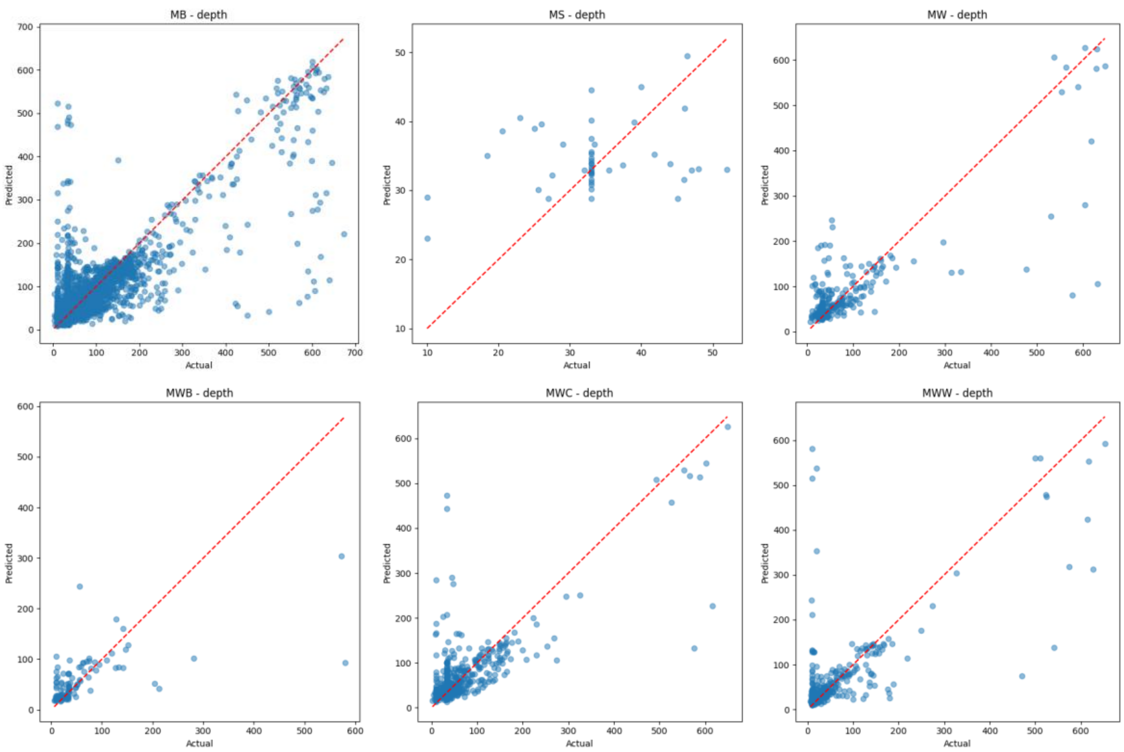

Fig. 4 Actual vs predicted plots for earthquake depth across all magnitude types

ISSN:2302-8149
e-ISSN:2540-9719

In contrast, the magnitude plots reveal a noticeably different pattern. For all magnitude types, the predicted values are concentrated within a relatively narrow range, producing horizontal bands rather than a strong diagonal distribution. For example, the Mb model predicts most magnitude values between approximately 4.8 and 5.2 despite actual magnitudes extending beyond 6.0. Similar patterns are observed for Mw, Mwb, Mwc, and Mww, where the predicted magnitudes tend to cluster around the dominant magnitude classes within each dataset. This behavior explains why the magnitude models achieve relatively low MAE values while simultaneously producing R^2 values close to zero. The models are able to estimate magnitudes near the central tendency of each magnitude-type dataset, thereby reducing average prediction errors. However, they are less effective at capturing the full variability of magnitude values, particularly for larger-magnitude earthquakes. Consequently, the predictions exhibit limited dispersion and tend to underestimate extreme observations.

Overall, the Actual versus Predicted analysis supports the numerical evaluation results and demonstrates that the magnitude-type-specific Random Forest models are capable of identifying meaningful spatial patterns, particularly for earthquake depth. These findings further indicate that spatial location plays a more significant role in explaining depth variability than magnitude variability within the Indonesian earthquake dataset.

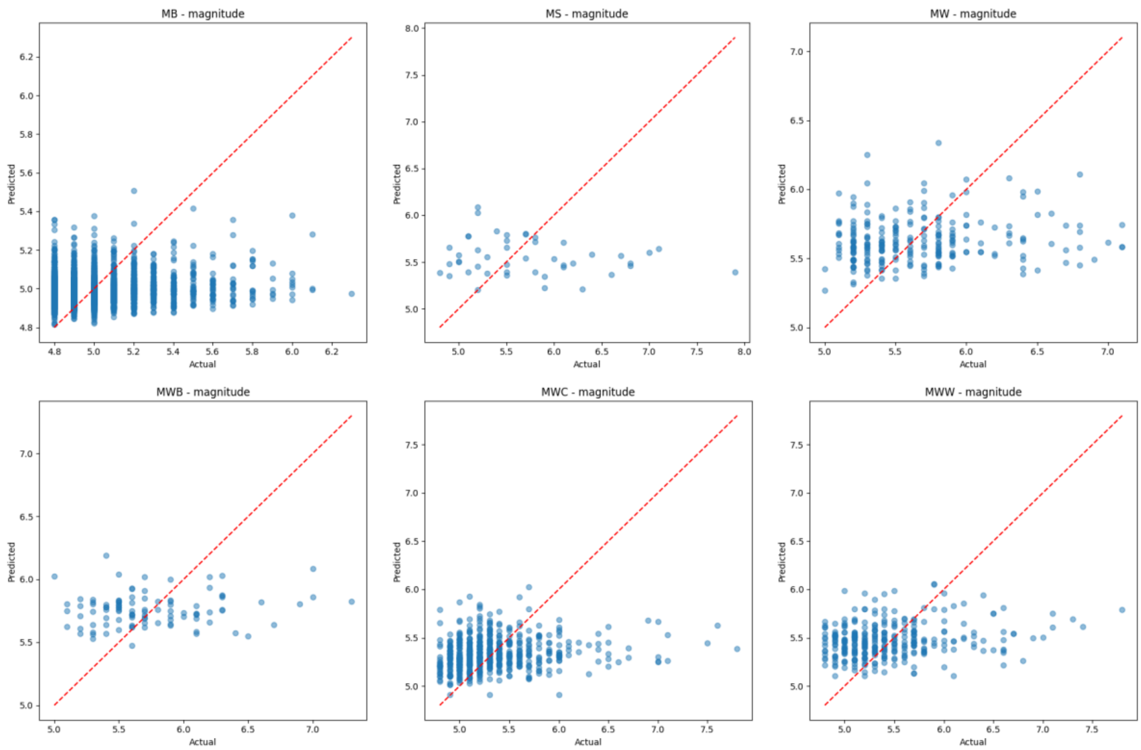

Fig. 5 Actual vs predicted plots for earthquake magnitude across all magnitude types

4.5. Case Study Analysis

The case study analysis was conducted to evaluate the practical implementation of the proposed Random Forest Regressor Multi-Output model using five representative locations in Indonesia. These locations were selected to represent diverse geographical and tectonic conditions, ranging from western to eastern regions of the country, including educational institutions, public facilities, and strategic landmarks. The detailed information of each location, including address and geographic coordinates, is presented in Table 8.

ISSN:2302-8149
e-ISSN:2540-9719

Table 8 Location of case study

No	Address	Longitude	Latitude
1	UMBY Kampus 3, Jl. Ring Road Utara, Ngropoh, Condongcatur, Kec. Depok, Kabupaten Sleman, Daerah Istimewa Yogyakarta 55281	110.4138	-7.7624
2	UMB Jakarta, Jl. Meruya Selatan No.1, RT.4/RW.1, Meruya Sel., Kec. Kembangan, Kota Jakarta Barat, Daerah Khusus Ibukota Jakarta 11650	106.7331	-6.1833
3	Tugu Sawit Mahato KM 24, Mahato, North Tambusai, Rokan Hulu Regency, Riau 28558	100.3366	1.3769
4	Sriti Convention Hall Palu, Jl. Durian No.88, Kamonji, Kec. Palu Bar., Kota Palu, Sulawesi Tengah 94111	119.8496	-0.8946
5	Kampus Uswim Nabire, Kalibobo, Nabire, Nabire Regency, Central Papua 98818	135.4842	-3.3703

The magnitude prediction results shown in Table 9 indicate that the model produces relatively stable outputs across all case studies. The predicted values generally fall within the range of approximately 4.8 to 6.0. Among the different magnitude types, Mwb consistently yields the highest values across all cases, with predictions ranging from 5.7428 to 5.9671. The Mw magnitude, which is widely considered the most representative measure of earthquake size, shows values ranging from 5.3821 to 5.7151 across all cases. The highest Mw prediction is recorded in Case 4, located in Palu, Central Sulawesi, with a value of 5.7151, which aligns with the known high seismic activity in the region due to the presence of active fault systems. In contrast, the lowest Mw value is observed in Case 3. Meanwhile, Mb values tend to be the lowest among all magnitude types, with a relatively narrow range between 4.8735 and 5.0166, indicating consistent lower-bound estimations by the model across all locations.

Table 9 Results of magnitude analysis

Case	Magnitude Type					
	Mb	Ms	Mw	Mwb	Mwc	Mww
1	4.9724	5.4336	5.4746	5.8347	5.3454	5.5000
2	5.0166	5.5389	5.5863	5.7428	5.2854	5.3013
3	4.8735	5.5162	5.3821	5.9671	5.5077	5.4394
4	4.9659	5.4464	5.7151	5.7519	5.3710	5.5094
5	4.9392	5.5680	5.3875	5.8872	5.2496	5.5728

The depth prediction results presented in Table 10 show a significantly higher level of variability compared to magnitude predictions. This variation reflects the complex nature of earthquake depth distribution across different tectonic settings in Indonesia. Cases 1, 2, and 3 generally exhibit deeper earthquake predictions, with Case 3 showing the deepest value of 227.1852 km for the Mb magnitude type. Depth values exceeding 100 km are also observed in several magnitude types within these cases, indicating the presence of deep-focus seismic characteristics influenced by subduction-related historical patterns in the dataset. In contrast, Cases 4 and 5 demonstrate much shallower depth predictions across most magnitude types. Case 4, located in Palu, records a minimum depth of 10.2552 km for Mww, while Case 5 shows depth values predominantly in the range of approximately 21 to 39 km. These shallower depths are typically associated with crustal fault activity, which is more likely to produce strong surface shaking despite moderate magnitude levels. The contrast between shallow and deep predictions across the cases highlights the influence of regional tectonic structures on earthquake depth variability.

Table 10 Results of depth (km) analysis

Case	Magnitude Type					
	Mb	**Ms**	**Mw**	**Mwb**	**Mwc**	**Mww**
1	112.5597	32.0428	134.6983	103.8191	106.7914	83.3781
2	139.2229	33.4541	116.5609	66.9256	120.1035	79.8901
3	227.1852	33.8714	78.8692	46.4490	117.7078	28.8578
4	31.2686	32.5747	38.2977	45.4955	33.1071	10.2552
5	21.2908	31.1619	39.2222	25.6478	25.5720	24.8476

Overall, the results indicate that the model is capable of capturing spatial patterns in earthquake characteristics based on geographic coordinates. The magnitude predictions remain relatively consistent and clustered within a moderate range, while depth predictions exhibit greater variability that reflects differences in tectonic environments across Indonesia. These findings suggest that the model effectively learns the spatial relationships embedded in the historical seismic dataset, making it potentially useful as a supporting tool for preliminary seismic hazard assessment and spatial risk analysis, although it should not be interpreted as a direct prediction of future earthquake events.

5 Conclusion

This study analyzes the spatial variability of earthquake magnitude and depth in Indonesia using a Random Forest Regressor based on 21,886 USGS earthquake records (1976–2026). After data cleaning and grouping into six magnitude types (Mb, Ms, Mw, Mwb, Mwc, and Mww), longitude and latitude were used as input features to predict magnitude and depth. The results show that spatial coordinates have limited ability to explain earthquake magnitude, as indicated by low R^2 values, while they perform better in predicting earthquake depth, especially for Mw which achieved the highest R^2 value (0.6801). Case studies from five locations in Indonesia also show that magnitude values are relatively stable, while depth varies more across regions. Overall, the Random Forest model is more effective for capturing spatial patterns of earthquake depth than magnitude, making it more suitable for spatial seismic analysis rather than direct earthquake prediction. Future work should consider adding tectonic and geological features and testing other machine learning models or deep learning to improve performance.

CRediT

Holmes Heriyanto Silalahi: Conceptualization, Methodology, Formal Analysis, Writing – Original Draft & Editing, Funding acquisition. **Anief Fauzan Rozi**: Supervision, Writing – Review.

References

[1] S. J. Hutchings and W. D. Mooney, "*The Seismicity of Indonesia and Tectonic Implications*," *Geochemistry, Geophysics, Geosystems*, Vol. 22, No. 9, Sep. 2021, DOI: 10.1029/2021GC009812.

[2] N. S. Santoso, N. A. Inayah, S. F. Kamilah, and S. N. Aisyah, "*Visualization of the Relationship between Magnitude and Depth of Earthquakes in Southern Indonesia through Curve Interpolation*," *Current STEAM and Education Research*, Vol. 3, No. 3, pp. 183–198, Dec. 2025, DOI: 10.58797/cser.030305.

[3] United States Geological Survey, "*USGS Earthquake Catalog*," Website. Accessed: Mar. 26, 2026. [Online]. Available: https://earthquake.usgs.gov/earthquakes/search/

[4] W. S. Karurung, K. Lee, and W. Lee, "*Assessment of Forest Fire Vulnerability Prediction in Indonesia: Seasonal Variability Analysis using Machine Learning Techniques*," *International Journal of Applied Earth Observation and Geoinformation*, Vol. 138, Apr. 2025, DOI: 10.1016/j.jag.2025.104435.

[5] R. D. Prasetyo, N. A. Maori, and A. K. Zyen, "Pengaruh Optimasi *Hyperparameter Random Forest* terhadap Akurasi Prediksi Magnitudo Gempa Bumi berdasarkan Hasil Klasterisasi

DBSCAN," *Jurnal Teknologi Informasi dan Komunikasi*, Vol. 10, No. 1, pp. 405–415, Jan. 2026, DOI: 10.35870/jtik.v10i1.5555.

[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] F. N. Rifqi, M. Subhan Arobi, and M. K. Nurzaman, "Prediksi Kejadian Gempa Bumi di Indonesia berbasis Pola Runtun Waktu dengan *Autoregressive Intergrated Moving Average*," *Advances in Data Mining and Digital Governance*, Vol. 1, No. 1, pp. 158–168, 2025, [Online]. Available: https://fisip.uinsgd.ac.id/conferences/index.php/ADMDG

[8] H. Utama, A. Masruro, T. Indriyatmoko, and Sudarmanto, "Integrasi Augmentasi Data dan *Machine Learning* dalam Prediksi Magnitudo Gempa Bumi," *BHATARA: Jurnal Multidisiplin*, Vol. 2, No. 3, pp. 97–108, 2025, [Online]. Available: https://earthquake.usgs.gov/earthquakes/feed/v1.0/csv.php

[9] H. Tantyoko, D. Kartika Sari, and A. R. Wijaya, "Prediksi Potensial Gempa Bumi Indonesia menggunakan Metode *Random Forest* dan *Feature Selection*," *Indonesia Journal Information System*, Vol. 6, No. 2, pp. 83–89, Jul. 2023, [Online]. Available: http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/indexHenriTantyoko|http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index|

[10] R. A. Syahputra and M. A. Safitri, "Pemodelan Prediksi Magnitudo Gempa Bumi di Indonesia berdasarkan Data Spasial dan Temporal menggunakan *Algoritma Random Forest* dan *Neural Networks*," Universitas Mercu Buana, Jakarta, 2025. [Online]. Available: https://lib.mercubuana.ac.id

[11] M. Prasetio, H. Sulistiani, O. Y. Inonu, K. Magda, and B. Santosa, "Klasifikasi Tingkat Risiko Gempa di Indonesia menggunakan Pola Spasial dan Temporal berbasis *Decision Tree*," *Bulletin of Computer Science Research*, Vol. 5, No. 5, pp. 1059–1066, Aug. 2025, DOI: 10.47065/bulletincsr.v5i5.624.

[12] Y. S. Yulianti, "Pemetaan dan Klasifikasi Wilayah Rawan Gempa di Indonesia dengan Metode *K-Means* dan *LightGBM*," Universitas Bina Sarana Informatika, 2025.

[13] Y. S. Yulianti, R. A. Purnama, and T. A. A. Sandi, "Pemetaan dan Klasifikasi Wilayah Rawan Gempa di Indonesia dengan Metode *K-Means* dan *LightGBM*," *Jurnal Informatika dan Teknik Elektro Terapan*, Vol. 13, No. 3S1, pp. 1968–1975, Oct. 2025, DOI: 10.23960/jitet.v13i3S1.8152.

[14] 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, pp. 493–504, Feb. 2025, DOI: 10.52436/1.jpti.667.

[15] Risda, R. Irwana, M. Shidiq Permana, I. Rizki Saputra, R. Oktaviana HErtland, and M. Bayu Anggara, "Penerapan *Algoritma Random Forest* untuk Prediksi Jumlah Gempa Bulanan di Indonesia," *Jurnal Ilmiah Binary STMIK Bina Nusantara Jaya Lubuklinggau*, Vol. 7, No. 2, pp. 65–70, Aug. 2025, DOI: 10.52303/jb.v7i2.172.

[16] L. Breiman, "*Random Forests*," *Mach. Learn.*, Vol. 45, pp. 5–32, 2001.

[17] L. Breiman, "*Randomizing Outputs to Increase Prediction Accuracy*," *Mach. Learn.*, Vol. 40, pp. 229–242, 2000.

[18] L. Breiman, "*Some Infinity Theory for Predictor Ensembles*," Jun. 2000.

[19] H. Dabiri, A. Faramarzi, A. Dall'Asta, E. Tondi, and F. Micozzi, "*A Machine Learning-based Analysis for Predicting Fragility Curve Parameters of Buildings*," *Journal of Building Engineering*, Vol. 62, Dec. 2022, DOI: 10.1016/j.jobe.2022.105367.

[20] H. Khoshvaght, R. R. Permala, A. Razmjou, and M. Khiadani, "*A Critical Review on Selecting Performance Evaluation Metrics for Supervised Machine Learning Models in Wastewater Quality Prediction*," *J. Environ. Chem. Eng.*, Vol. 13, No. 6, Dec. 2025, DOI: 10.1016/j.jece.2025.119675.

[21] Suraya and M. Sholeh, "*Designing and Implementing a Database for Thesis Data Management by using the Python Flask Framework*," *International Journal of Engineering, Science and Information Technology*, Vol. 2, No. 1, pp. 9–14, 2022, DOI: 10.52088/ijesty.v1i1.197.