BAREKENG: Journal of Mathematics and Its Applications
March 2026 Volume 20 Issue 1 Page 0017–0030
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss1pp0017-0030

COMPARISON OF CLUSTERING EARTHQUAKE PRONE AREA IN SUMATRA ISLAND USING K-MEANS AND SELF-ORGANIZING MAPS

Faradilla Ardiyani [1], Winita Sulandari [2*], Yuliana Susanti [3]

[1,2,3]*Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret*
Jln. Ir. Sutami 36A, Surakarta, 57126, Indonesia

*Corresponding author's e-mail: * **winita@mipa.uns.ac.id***

Article Info

Article History:

Received: 1st January 2025
Revised: 30th April 2025
Accepted: 18th June 2025
Available online: 24th November 2025

Keywords:

Algorithm of K-Means;
Algorithm of SOM;
Clustering;
Earthquake.

ABSTRACT

An earthquake is a sudden vibration on the earth's surface caused by the shifting of tectonic plates. One region in Indonesia that is particularly prone to earthquakes is Sumatra Island, due to its geographical location at the convergence of two tectonic plates, namely the Indo-Australian plate, which is actively subducting beneath the Eurasian plate. While earthquakes cannot be prevented or avoided, effective disaster mitigation strategies can help minimize the impact. The purpose of this research is to classify earthquake-prone areas on Sumatra Island based on depth and magnitude, allowing for further analysis to determine the characteristics of the clustering results. The study employs two clustering methods to analyze earthquake data from 1973 to 2024: the K-means and Self-Organizing Maps (SOM) algorithm. K-means algorithm is preferred for its simplicity and efficiency in handling large datasets, and suitability for numerical earthquake data analysis. Conversely, the SOM algorithm offers more stable clustering results and preserves the topological structure of the data, making it advantageous for exploring spatial patterns. The research findings indicate that the K-means algorithm provides better grouping, achieving a Silhouette Coefficient of 0.53, compared to 0.47 for the SOM algorithm. The K-means clustering resulted in two clusters: Cluster 1 contains 1,213 members and is characterized by shallow depths (3.9 km-41 km) and larger magnitudes (5 m_b- 8.92 m_b), indicating a higher risk level. In contrast, Cluster 2 includes 412 members and represents areas with greater depths (40.8 km-70 km) and smaller magnitudes (5 m_b- 6.85 m_b), corresponding to a lower risk level. This research aims to support the government in its earthquake disaster mitigation efforts, especially on Sumatra Island.

How to cite this article:

F. Ardiyani, W. Sulandari and Y. Susanti., "COMPARISON OF CLUSTERING EARTHQUAKE PRONE AREA IN SUMATRA ISLAND USING K-MEANS AND SELF-ORGANIZING MAPS," *BAREKENG: J. Math. & App.*, vol. 20, iss. 1, pp. 0017-0030, Mar, 2026.

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

Research Article · Open Access

1. INTRODUCTION

An earthquake is a vibration originating from within the earth that propagates to the surface, caused by sudden and violent shifts in the earth's crust. It can result from various geological activities, including tectonic movements, volcanic eruptions, meteorite impacts, underwater landslides, or underground explosions [1]. Earthquakes are among the most frequent natural disaster in Indonesia [2]. It is primarily due to the country's location at the convergence of three major tectonic plates and nine smaller plates, forming a highly complex and active tectonic region. One of the most earthquake-prone areas in Indonesia is the island of Sumatra. This vulnerability is attributed to the island's geographical conditions, which include active faults, volcanic paths, and subduction zones [3]. According to the Ministry of Energy and Mineral Resources, 6 of the 25 earthquake-prone areas in Indonesia are located in Sumatra. These areas include Aceh, Jambi, Bengkulu, Lampung, West Sumatra, and North Sumatra. Data from the Indonesian Meteorology, Climatology, and Geophysics Agency (BMKG) indicates that between 2009 and 2018, Sumatra experienced 5,937 earthquakes, with 36% of these classified as large-magnitude events.

Earthquakes are natural disasters that cannot be prevented or avoided. Therefore, disaster mitigation is essential to minimize the impact. According to Article 1, Section 6 of Government Regulation No. 21 of 2008 regarding the Implementation of Disaster Management, disaster mitigation refers to a series of efforts aimed at reducing disaster risk through physical development, raising awareness, and enhancing the capacity to confront disaster threats. One form of mitigation involves identifying and classifying earthquake-prone areas using earthquake-related variables. This grouping can be carried out using clustering techniques [4].

Clustering is the process of organizing data into several groups or clusters. In this process, data within each group has a high level of similarity, while data between different groups shows minimal similarity [5]. There are several grouping methods including partition methods and model-based methods. Partition-based methods, divide the data into *K* parts, where each part represents a group [6]. A popular and frequently used partition algorithm is K-means. On the other hand, model-based methods include neural network approaches, one of the most prominent being Self-Organizing Maps (SOM).

According to [7], the K-means algorithm is effective for processing large datasets, making it suitable for application to earthquake data, which often consists of a substantial volume of records. K-means is also widely recognized for the simplicity of its algorithm. However, this algorithm has shortcomings in determining centroid or cluster center, as the initial selection is made randomly. This randomness can lead to varying clustering results depending on the initial starting points [8]. In contrast, the SOM algorithm provides a more stable approach to clustering, as it tends to generate fewer branches where the group center value remains consistent across each neuron, and cluster assignment is determined by the minimum distance between the input data and the neurons [9]. Nevertheless, the SOM algorithm typically requires longer training time and involves more complex computations [10].

Both K-means and SOM algorithms have been successfully applied in previous studies related to earthquake data clustering. For instance, research by [11], [12], and [13] applied the K-means algorithm to group earthquake-prone areas in Bengkulu, Java Island, and across Indonesia, respectively. Additionally, studies utilizing the SOM algorithm include those by [14], who used earthquake data from across Indonesia, and by [15], who analyzed earthquake data along the South Coast of Java and Lampung. However, these studies mainly focused on earthquake magnitudes without a detailed consideration of the earthquake source depth, such as whether the events were shallow, medium, or deep.

Based on the background and previous research, this study aims to compare the performance of the K-means and SOM algorithms in clustering earthquake data from Sumatra Island, with a focus on key variables such as earthquake depth and magnitude. This research incorporates updated data up to 2024, emphasizing shallow earthquakes (≤70 km) due to their proximity to the earth's surface, which result in stronger ground shaking and more severe damage. Additionally, only earthquakes with magnitudes equal to or greater than 5.0 m_b are included, as these are more likely to have significant impacts. The objective is to provide a comprehensive comparison of the two clustering methods and determine which algorithm performs better using the Silhouette Coefficient (SC) as a performance metric. Furthermore, it produces detailed visual maps of the clustering results to address the lack of visual representations of earthquake-prone areas on Sumatra Island, offering valuable insights to inform disaster mitigation efforts and future urban planning in high-risk regions.

2. RESEARCH METHODS

2.1 Dataset

The dataset used in the study is secondary data obtained from the United States Geological Survey (USGS) website (https://earthquake.usgs.gov/earthquakes/search/). It consists of 1,625 earthquake occurrence records on Sumatra Island from January 1973 to December 2024, comprising three variables: depth, magnitude, and geographic location. This study focuses on shallow earthquakes (depth ≤ 70 km), as they occur closer to the earth's surface and are therefore more likely to produce intense ground shaking and significant damage. Additionally, only earthquake with a magnitude of $M \geq 5.0$ m_b are included, since events of this scale are considered to have substantial impacts. The sample of dataset is shown in Table 1.

Table 1. Sample of Earthquake Data on Sumatra Island

Location	Depth (km)	Magnitude (m_b)
79 km WNW of Bengkulu, Indonesia	59.84	5.00
95 km SSW of Sibolga, Indonesia	64.77	5.10
253 km S of Sinabang, Indonesia	8.00	5.86
103 km SSW of Pagar Alam, Indonesia	53.39	5.00
82 km WNW of Meulaboh, Indonesia	61.14	5.20

***Data source**: https://earthquake.usgs.gov/earthquakes/search/*

2.2 Theoretical Review

2.2.1 Elbow Method

The elbow method is a method used to produce information in determining the best number of clusters by looking at the percentage of the comparison between the number of clusters that will form an elbow at a point [16]. If the value of the first cluster and the value of the second cluster create a noticeable angle on the graph, or if there is a significant reduction in value, this indicates that the chosen number of clusters is appropriate. The comparison is obtained by calculating the Sum of Square Error (SSE) for each cluster value, and it decreases as the number of group K increases. The elbow rule can give a suggestion for selecting the value of K based on the analysis of SSE image [17]. The SSE for each K value is calculated using Eq. (1).

$$SSE = \sum_{k=1}^{K} \sum_{x_i \in S_k} |x_i - c_k|^2 , \tag{1}$$

with

K : number of clusters
x_i : i-th data
c_k : k-th centroid (center of cluster)
S_k : k-th cluster
$|x_i - c_k|$: distance between members of cluster with centroid

2.2.2 K-Means Algorithm

K-means is one of the most widely used clustering techniques due to its simplicity of implementation and its effectiveness [18]. In the K-means algorithm, each cluster is represented by a centroid (a group center), and each data point is assigned to the cluster with the nearest centroid. The steps of the K-means algorithm are as follows [6]:

1. Determine K as the number of clusters to be formed.
2. Select the initial centroids for the K clusters.
3. Calculate the distance from each data point to each centroid using a similarity measure; the most common measure used is Euclidean distance which is calculated using Eq. (2).

$$d = \sqrt{\sum_{i=1}^{n} (x_i - c_i)^2} , \tag{2}$$

with

d : Euclidean distance
x_i : i-th data
c_i : i-th centroid
n : the number of objects that are members of the group.

4. Assign each data point to the nearest centroid. Each object is declared as a group member by measuring the distance of its proximity to the center point of the group.
5. Recalculate the centroid of each cluster based on the new group members with Eq. (3).

$$c_i = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{3}$$

with
c_i : i-th centroid
x_i : i-th data
n : the number of objects that are members of the group.

6. Repeat steps 3-5 until no data points change clusters.

2.2.3 Self-Organizing Maps (SOM) Algorithm

The SOM algorithm is a type of unsupervised learning commonly used as a clustering tool. This technique is trained using an unsupervised method and is capable of organizing different kinds of inputs by the data samples into groups of the cluster with several characteristics [19]. SOM requires the initialization of several parameters in its early stages, including the learning rate, neighborhood function, and map size.

Learning rate is a multiplier that influences the adjustment of connection weights, typically ranging from 0 to 1. A higher learning rate results in faster adaptation of the weights, meaning the input vector has a greater influence on the weight changes. Over time, the learning rate gradually decreases with each iteration. As it approaches zero, weight updates become smaller, enabling better mapping of input vectors [20].

Neighborhood function determines the degree of weight adjustment for neurons surrounding the winning neuron (the neuron closest to the input vector). It proportionally affects how neighboring neurons update their weights in relation to the winner. The most commonly used function is the Gaussian Neighborhood [21], which applies a decreasing influence based on the distance from the winning neuron. Gaussian functions are considered more reliable because different initializations often converge to the same map structure.

In clustering applications, the map size represents the number of output nodes, which can correspond to the number of resulting clusters. The map size is influenced by the number of input samples and the number of variables in the data. If the map is too small, the data may not be distributed accurately or evenly across the nodes. Therefore, the optimal map size is typically determined through trial and error until the best result is achieved [22].

The stages of the SOM algorithm for grouping are as follows [23]:

1. Initialize the following parameters: map size, weight w_{ij}, learning rate (α), neighborhood radius (σ), and maximum iterations (T).
2. Perform steps 3-9 if the stop condition has not been met.
3. Perform steps 4-6 for each input vector x_i ($i = 1, \dots, n$).
4. Calculate the Euclidean distance between the weight w_{ij} and the input vector x_i for each neuron j ($j = 1, \dots, m$) using Eq. (4).

$$\boldsymbol{D(j) = \sqrt{\sum_{i=1}^{n} \left(w_{ij} - x_i\right)^2}}, \tag{4}$$

with
$\boldsymbol{w_{ij}}$: weight that connects the input vector $\boldsymbol{x_i}$ to unit $\boldsymbol{y_j}$
$\boldsymbol{x_i}$: input vector.

5. Determine the Best Matching Unit (BMU) or winning neuron by selecting the vector that has the smallest Euclidean distance.
6. Update the weight w_{ij} value of the BMU and each vector neighboring the BMU using (5).

$$\boldsymbol{w_{ij}(t+1) = w_{ij}(t) + \alpha(t) h^{c}_{i,j}(t)\left[x_i - w_{ij}(t)\right]} \quad (5)$$

with
$\boldsymbol{w_{ij}(t+1)}$: new weight $\boldsymbol{w_{ij}}$
$\boldsymbol{w_{ij}(t)}$: initial weight $\boldsymbol{w_{ij}}$
$\boldsymbol{\alpha(t)}$: learning rate
$\boldsymbol{h^{c}_{i,j}(t)}$: neighborhood function
$\boldsymbol{t}$: iteration.

Neighborhood function (Gaussian Neighborhood) is formulated as Eq. (6).

$$\boldsymbol{h^{c}_{i,j}(t) = exp\left(-\frac{{d^{c}_{i,j}}^{2}}{2\sigma^{2}(t)}\right)}, \quad (6)$$

with $\boldsymbol{d^{c}_{i,j}}$ is the Euclidean distance between BMU and neighboring neurons and $\boldsymbol{\sigma(t)}$ is the width or radius of the neighborhood.

7. Update the learning rate so that it decreases as the training progresses, using Eq. (7).

$$\boldsymbol{\alpha(t) = \alpha_0\left(1 - \frac{t}{T}\right)}, \quad (7)$$

with
$\boldsymbol{\alpha_0}$: initial learning rate
$\boldsymbol{t}$: iteration
$\boldsymbol{T}$: maximum iteration.

8. Update the neighborhood radius so that the distance to neighboring vectors decreases Eq. (8).

$$\boldsymbol{\sigma(t) = \sigma_0\, exp\left(-\frac{t}{T}\right)}, \quad (8)$$

with
$\boldsymbol{\sigma_0}$: initial neighborhood radius
$\boldsymbol{t}$: iteration
$\boldsymbol{T}$: maximum iteration.

9. Test stopping conditions. The stop condition is carried out when the number of iterations is maximum.
10. Perform grouping by calculating the Euclidean distance between the final weights and the input vectors, then identify the smallest distance to determine the final group.

2.2.4 Silhouette Coefficient (SC)

The SC is a method for validating the results of clustering analysis through internal validation, introduced by Rousseeuw in 1986. The SC provides a value that measures how well an object is positioned within its group and the degree of separation between different groups. In other words, SC evaluates whether an object properly belongs to its assigned group. Eq. (9) is used for determining SC [24]:

$$SC = \frac{1}{n}\sum_{i=1}^{n}\frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}, \quad (9)$$

with
$a(i)$: the average distance of a specific object (Mi) to all other objects in the same cluster
$b(i)$: the average distance of the object (Mi) to all objects in each cluster (excluding the cluster containing Mi)
n : amount of data.

2.3 Research Steps

In this research, Jupyter Notebook software is used with the Python programming language to help with the analysis. The following steps were taken throughout the study.

1. Collected earthquake data on Sumatra Island from the USGS website.

2. Conducted descriptive analysis of the data.
3. Performed data preprocessing, including checking missing values and normalizing data.
4. Implemented clustering using the K-means algorithm.
5. Implemented clustering using the SOM algorithm.
6. Compared the clustering results of K-means and SOM using the SC.
7. Identified the characteristics of each cluster.
8. Mapped the best-performing cluster.

3. RESULTS AND DISCUSSION

3.1 Descriptive Statistics

The dataset utilized in this study was sourced from the USGS website and comprises 1,625 earthquake events that occurred on Sumatra Island from 1973 to 2024. It includes key variables such as depth and magnitude. The descriptive statistics for this dataset are presented in Table 2.

Table 2. Descriptive Statistics of Earthquake Data on Sumatra Island

	Depth (km)	Magnitude (m_b)
Mean	34.35	5.34
Std. Deviation	14.19	0.41
Minimum	3.90	5.00
Median	33.00	5.20
Maximum	70.00	8.92

Table 2 shows that the average depth of earthquakes that occurred on Sumatra Island is 34.35 km, with an average magnitude of 5.34 mb. The depth ranges from 3.9 km to 70 km, while the magnitude ranges from 5 mb to 8.92 mb. These figures indicate that the variables have significantly different maximum and minimum values. Therefore, data normalization is necessary to ensure that all variables operate on the same scale, preventing any single variable from disproportionately influencing the clustering results.

The normalization method used in this study is Min-Max Scaling. This technique transforms features to a scale of [0, 1], ensuring that each variable contributes equally to the clustering process and preventing any single variable from dominating due to its original scale [25]. This normalization is essential for producing meaningful and unbiased clustering results. Table 3 presents the descriptive statistics of the dataset after normalization using the Min-Max Scaling method.

Table 3. Descriptive Statistics After Min-Max Scaling Normalization

	Depth (km)	Magnitude (m_b)
Mean	0.46	0.09
Std. Deviation	0.21	0.1
Minimum	0	0
Median	0.44	0.05
Maximum	1	1

Based on Table 3, the minimum value is 0 for both variables, while the maximum value is 1. The average value for the depth variable is 0.46, and for the magnitude variable is 0.09. These results confirm that the data has been standardized to the same scale, ensuring that no variable dominates the clustering process. Therefore, the dataset is now ready for the next stage, which is clustering analysis using the K-means and SOM algorithms.

3.2 K-means Clustering

Grouping using the K-means algorithm requires determining the number of clusters (K) in advance. The value of K significantly influences the clustering results, so selecting the most appropriate number of clusters is essential. There are several methods to determine the optimal K, and this study, the Elbow method is used. The elbow method involves calculating the SSE for different values of K, and then analyzing how the SSE changes as K increases. Table 4 presents the SSE values for K = 1 to K = 9.

Table 4. The Sum of Square Error Value

K	SSE	Difference
1	92.70	-
2	**43.06**	**49.64**
3	27.85	15.21
4	19.60	8.25
5	15.35	4.25
6	13.03	2.32
7	10.88	2.15
8	9.10	1.78
9	8.10	1.00

The method identifies the point where the SSE experiences the most significant drop. Based on Table 4, the most substantial decrease in SSE occurs between 0 and 49.64, where the SSE drops by 43.06. This indicates that $K = 2$ is the optimal number of clusters. A plot of the SSE values against the number of clusters (K) is shown in Fig. 1 below. It is obtained that the optimal number of clusters is identified at $K = 2$, where the curve shows a noticeable bend or "elbow", indicating a significant reduction in SSE to 43.06.

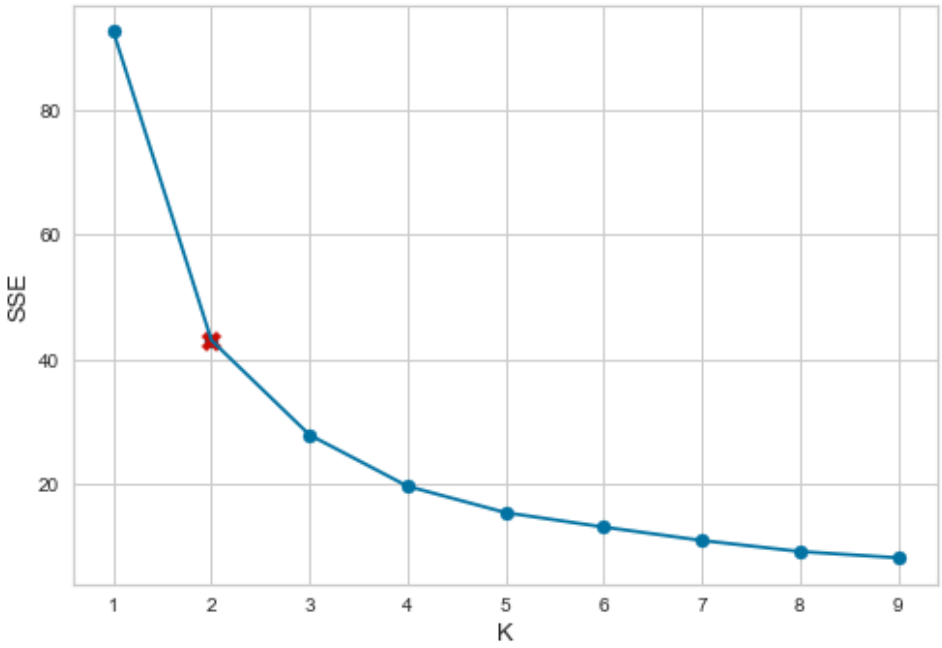


Figure 1. Determining the Number of *K* using the Elbow Method

The visualization results of the K-means algorithm clustering using Python are shown in Fig. 2.

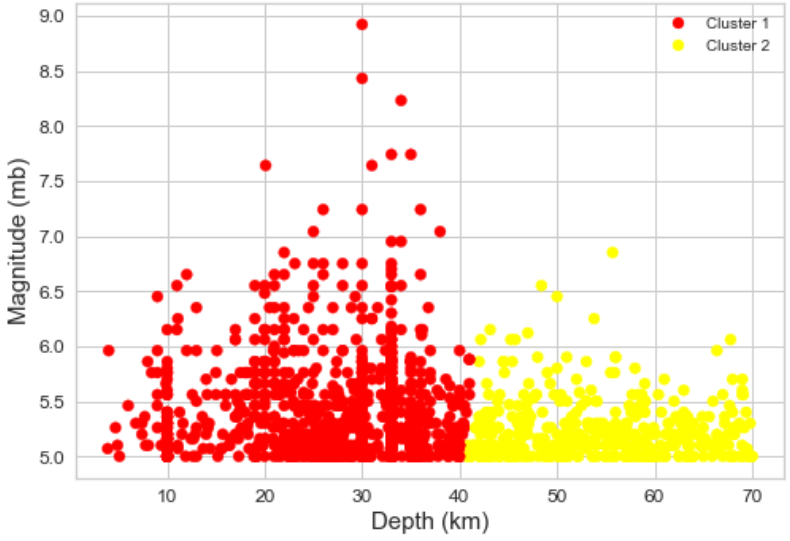


Figure 2. Results of clustering earthquake data on Sumatra Island using the K-means algorithm

The centroid of the results of clustering earthquake data on the island of Sumatra using the K-means algorithm after being returned to the initial data (denormalization) is shown in Table 5 below.

Table 5. Center of Cluster K-means Algorithm

Cluster	Depth (km)	Magnitude (m_b)	Number of Members
1	39.44	5.37	1,213
2	45.12	5.36	412

Table 5 shows that the center of Cluster 1 has a smaller depth value (39.44 km) compared to Cluster 2 (45.12 km), and slightly higher magnitude (5.37 m_b) than Cluster 2 (5.36 m_b). Cluster 1 also contains more members, with a total of 1,213 objects, while Cluster 2 consists of 412 objects.

3.3 SOM Clustering

Clustering using the SOM algorithm requires the initialization of several parameters, including map size, learning rate, neighborhood radius, maximum iterations, and initial weights. In this research, the Gaussian Neighborhood function is utilized. SOM clustering is implemented using the MiniSom library in Python. The parameter selection for modeling follows the default settings provided by MiniSom, such as a neighborhood radius of 1.0 and a learning rate of 0.5. In addition, a trial-and-error approach is applied to identify the combination of parameters that produces the best modeling based on the SC value.

Table 6. Comparison of Self-Organizing Maps Algorithm Parameters

Map Size	Learning Rate	Neighborhood Radius	SC
1×2	0.3	1.0	0.36
	0.5		0.36
	0.7		0.36
1×3	0.3	1.0	0.46
	0.5		0.45
	0.7		0.32
1×4	0.3	1.0	0.46
	0.5		**0.47**
	0.7		0.46
2×2	0.3	1.0	0.46
	0.5		0.46
	0.7		0.46

Based on the trials conducted in Table 6, the parameter initialization with the highest SC value is 0.47. The one-dimensional (1D) SOM can outperform the two-dimensional (2D) SOM, as demonstrated in the studies by Ramos et al. and Ullah et al. [26], [27]. This is because the 1D SOM adapts more easily to the distribution of the underlying dataset than to the 2D SOM. Furthermore, to determine the number of iterations based on cost-time considerations, more iterations require more time for computation. In this study, the maximum number of iterations was set to 500 times. The summary of the initial parameter settings is shown in Table 7.

Table 7. Initialization of Self-Organizing Maps Algorithm Parameters

Parameters	Value
Learning rate	0.5
Neighborhood Function	Gaussian
Neighborhood Radius	1.0
Map Size	1x4
Number of Clusters	4
Maximum Iteration	500

The SOM algorithm clustering using Python's visualization results produced 4 clusters, as shown in Fig. 3.

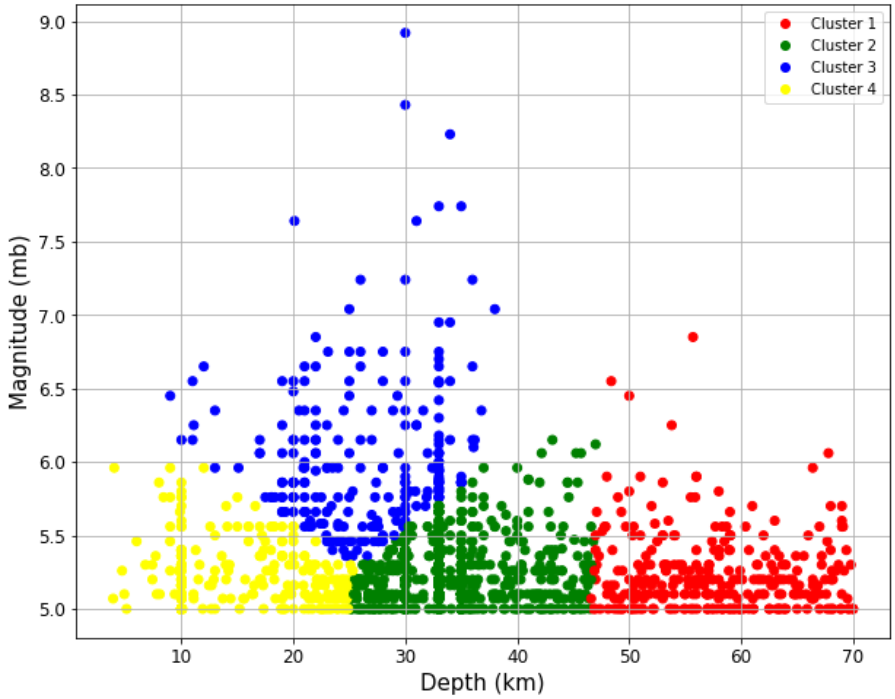
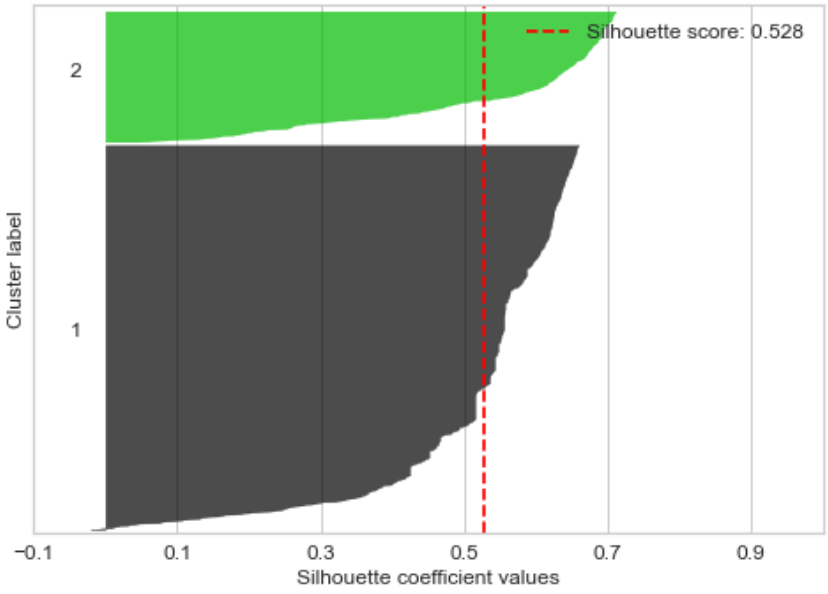


Figure 3. Results of Clustering Earthquake Data on Sumatra Island using the Self-Organizing Maps Algorithm

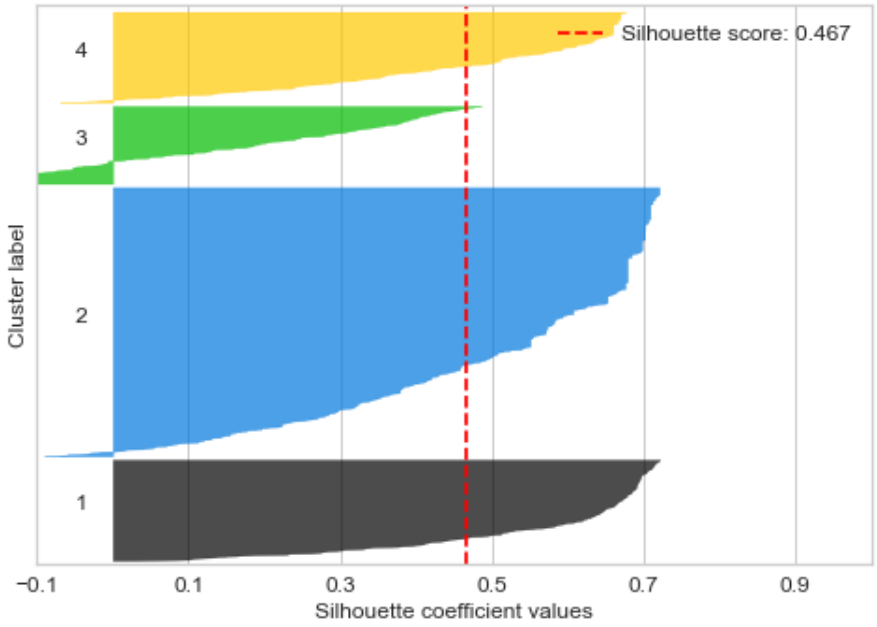
From the results, we can see the number of members in each cluster. Cluster 2 has the most members which is 811 objects, followed by Cluster 1 with 305 members, Cluster 4 with 274 members, and Cluster 3 with the fewest at 235 members.

3.4 Comparison of Clustering K-means and SOM Algorithm

The clustering results from the K-means and SOM algorithms are evaluated using the SC method. A higher SC value indicates better-defined and more cohesive clusters. Fig. 4 presents the python output displaying the SC values for the clustering results obtained from both the K-means and SOM algorithms.



(a)



(b)

Figure 4. Silhouette Coefficient Plot of (a) K-Means Algorithm and (b) Self-Organizing Maps Algorithm

A summary of the comparison of the SC values for both algorithms is presented in Table 8 below.

Table 8. Comparison of Silhouette Coefficient of (a) K-means Algoritm and (b) Self-Organizing Map Algorithm

Algorithm	Number of Clusters	Silhouette Coefficient
K-means	2	0.53
SOM	4	0.47

The clustering results in Table 8 indicate that an SC for the K-means algorithm is 0.53, which is higher than the value obtained for the SOM algorithm, which is 0.47. Based on these values, it can be concluded that the clustering results produced by the K-means algorithm are more optimal. The SC of 0.53 suggests that the clusters formed are fairly strong and well-separated (reasonable clustering). Therefore, the K-means algorithm was selected for the final clustering of earthquake-prone areas on Sumatra Island.

Using the K-means algorithm, two clusters were formed: Cluster 1 with 1,213 data points and Cluster 2 with 412 data points. Table 9 provides descriptive statistics for both clusters, focusing on the variables of earthquake depth and magnitude.

Table 9. Descriptive Statistics of K-means Algorithm Clustering Results

Variable	Category	Cluster 1	Cluster 2
Depth (km)	Minimum	3.9	40.8
	Mean	27.64	54.11
	Maximum	41.0	70
Magnitude (m_b)	Minimum	5.0	5.0
	Mean	5.37	5.24
	Maximum	8.92	6.85

The analysis reveals a notable difference between the two clusters in terms of earthquake depth, while the magnitude differences are less pronounced. Cluster 1 is characterized by shallower earthquake depths and slightly higher magnitudes compared to Cluster 2. Specifically, the average magnitude in Cluster 1 is 27.64 km, whereas Cluster 2 averages 54.11 km. The average magnitude in Cluster 1 is 5.37 m_b, slightly higher than the 5.24 m_b in Cluster 2. These results suggest that Cluster 1 is more vulnerable to seismic activity and poses a greater risk than Cluster 2. The spatial distribution of the clusters identified using the K-means algorithm is shown in Fig. 5.

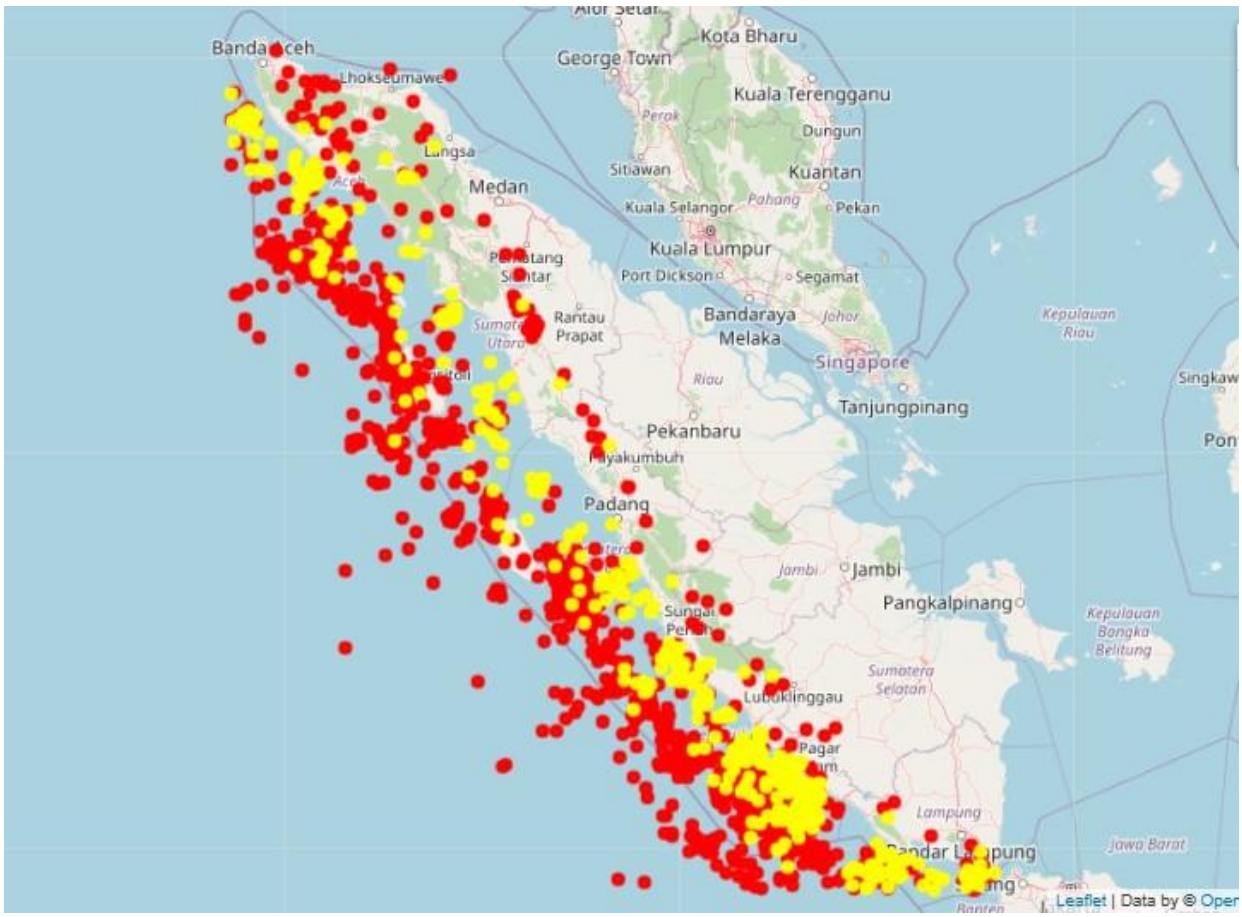


Figure 5. Map of Earthquake Clusters on Sumatra Island using the K-Means Algorithm

Fig. 5 illustrates that Cluster 1, represented in red, indicates areas with a high risk of earthquake occurrences. This group is primarily located in the offshore areas of the Indian Ocean, to the west of Sumatra Island, encompassing high-risk regions such as Aceh, North Sumatra, West Sumatra, and Bengkulu. These areas have historically experienced significant seismic events, including the Bengkulu earthquake (magnitude 7.4) on February 14, 2001; the catastrophic Aceh earthquake (magnitude 9.0) on December 26, 2004; the North Sumatra earthquake (magnitude 8.6) on March 28, 2005; the Mentawai Islands earthquake in West Sumatra (magnitude 7.8) on March 2, 2016; and the North Sumatra earthquake (magnitude 6.9) on March 14, 2022 [28]. Cluster 2, represented in yellow, corresponds to areas with a relatively lower seismic risk. These regions are mainly located inland or near the eastern part of Sumatra Island, including provinces such as Lampung, South Sumatra, Bangka Belitung, Jambi, Riau, and the Riau Islands. Although the seismic risk is lower in these regions, several notable earthquakes have still occurred, such as the Lampung earthquake (magnitude 6.6) on February 16, 1994; the Jambi earthquake (magnitude 6.7) on October 7, 1995; and the South Sumatra earthquake (magnitude 5.4) on March 31, 2014 [28].

Shallow earthquakes, defined as seismic events occurring at depths of less than 70 km, tend to generate more intense ground shaking and cause greater damage compared to deeper earthquakes, especially when accompanied by large magnitudes. This heightened impact results from the proximity of the earthquake's focus to the earth's surface, which allows more seismic energy to reach the affected areas. The clustering analysis conducted in this study reveals that 74.6% of earthquakes on Sumatra Island fall into the high-risk category, while the remaining 25.4% are categorized as lower risk. These findings confirm that the majority of earthquakes in the region are shallow in nature, with depths less than 70 km and magnitudes equal to or greater than 5 m_b.

Although earthquakes are inherently unpredictable, this research is expected to support local authorities in enhancing their disaster mitigation strategies. These findings can serve as a scientific basis for strengthening preparedness efforts, such as reinforcing infrastructure resilience, improving early warning systems, and conducting regular public education campaigns. The identification of shallow and high-magnitude earthquake patterns highlights the importance of spatial planning that incorporates seismic risk, particularly in vulnerable communities. Furthermore, integrating this information into regional development

policies can help reduce the impact of future events and support more targeted mitigation and preparedness measures across Sumatra Island.

4. CONCLUSION

Based on the results of the analysis and discussion, it can be concluded that the K-means algorithm resulted in two clusters with an SC of 0.53, while the SOM algorithm produced four clusters with an SC of 0.47. Since the Silhouette Coefficient evaluates how well each data point fits within its assigned cluster, and higher values indicate more coherent and well-separated clusters, the K-means algorithm demonstrated superior clustering performance. Therefore, from the standpoint of both clustering quality and interpretability, the K-means is considered the more suitable method for classifying earthquake-prone areas on Sumatra Island.

Cluster 1, comprising 1,213 earthquake events, is characterized by shallower depths (3.9 km-41 km) and higher magnitudes (5m_b-8.92 m_b), indicating a higher level of seismic risk. Geographically, this cluster is predominantly located offshore in the Indian Ocean to the west of Sumatra Island, covering areas such as Aceh, North Sumatra, West Sumatra, and Bengkulu. Cluster 2, consisting of 412 earthquake events, is marked relatively deeper depths within the shallow earthquake category (40.8 km-70 km) and smaller magnitudes (5 m_b-6.85 m_b), suggesting a lower risk level. This cluster is mainly situated near the mainland of Sumatra Island, including areas such as Lampung, South Sumatra, Bangka Belitung, Jambi, Riau, and the Riau Islands.

Author Contributions

Faradilla Ardiyani: Data Curation, Formal Analysis, Investigation, Validation, Visualization, Writing-Original Draft, Writing-Review and Editing. Winita Sulandari: Conceptualization, Methodology, Supervision, Writing-Review and Editing, Project Administration, Funding Acquisition. Yuliana Susanti: Validation, Visualization. All authors discussed the results and contributed to the final manuscript

Funding Statement

This research is funded by the RKAT of Universitas Sebelas Maret (UNS) for the 2025 Fiscal Year through the Research Scheme PKGR-UNS (Penguatan Kapasitas Grup Riset-UNS) B with Research Assignment Agreement Number: 371/UN27.22/PT.01.03/2025

Acknowledgment

The authors would like to express their sincere gratitude to LPPM UNS for facilitating the publication of this work. The authors also wish to thank the anonymous reviewers for their constructive suggestions and valuable feedback, which have significantly improved the quality of this paper.

Declaration

The authors declare that there is no conflict of interest

REFERENCES

[1] B. A. P. Martadiputra, D. Rachmatin, and A. S. Hidayat, "ANALYSIS OF CHARACTERISTICS OF EARTHQUAKE AREA IN INDONESIA IN 2020 WITH CLUSTER ANALYSIS AS NATURAL DISASTER," *Int. J. Sci. Res.*, vol. 9, no. 11, pp. 1243–1250, 2021, doi: https://10.21275/SR201122121148.

[2] S. Al Faridzi et al., "PENGOLAHAN DATA: PEMAHAMAN GEMPA BUMI DI INDONESIA MELALUI PENDEKATAN DATA MINING," *J. Pengabdi. Kolaborasi dan Inov. IPTEKS*, vol. 2, no. 1, pp. 262–270, 2024, doi: https://doi.org/10.59407/jpki2.v2i1.506.

[3] W. Asnita, D. Sugiyanto, and I. Rusydy, "KAJIAN STATISTIK SEISMISITAS KAWASAN SUMATERA," *J. Nat.*, vol. 16, no. 2, pp. 5–9, 2016, doi: https://doi.org/10.24815/jn.v16i2.4917.

[4] I. H. Rifa, H. Pratiwi, and Respatiwulan, "IMPLEMENTASI ALGORITMA CLARA UNTUK DATA GEMPA BUMI DI INDONESIA," *Semin. Nas. Penelit. Pendidik. Mat. UMT*, pp. 161–166, 2019.

[5] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *INTRODUCTION TO DATA MINING, SECOND EDITION*. New York: Pearson Education, Inc., 2019.
[6] J. Han, J. Pei, and H. Tong, *DATA MINING: CONCEPTS AND TECHNIQUES, FOURTH EDITION*. Elsevier Inc., 2022.
[7] E. Zacchei and R. Brasil, "K-MEANS FOR EARTHQUAKES: DISAGGREGATION ANALYSES OF SMALL EVENTS BY CONSIDERING WAVE COMPONENTS AND SOIL TYPES," *Arab. J. Geosci.*, vol. 17, no. 302, 2024, doi: https://doi.org/10.1007/s12517-024-12113-0.
[8] A. A. Khan, M. S. Bashir, A. Batool, M. S. Raza, and M. A. Bashir, "K-MEANS CENTROIDS INITIALIZATION BASED ON DIFFERENTIATION BETWEEN INSTANCES ATTRIBUTES," *Int. J. Intell. Syst.*, 2024, doi: https://doi.org/10.1155/2024/7086878.
[9] S. Ariani, M. Nusrang, and M. K. Aidid, "APPLICATION OF CLUSTER ANALYSIS OF SELF ORGANIZING MAP (SOM) METHOD IN THE COMMUNITY LITERACY DEVELOPMENT INDEX IN INDONESIA," *J. Appl. Sci. Eng. Technol. Educ.*, vol. 6, no. 1, pp. 56–62, 2024, doi: https://doi.org/10.35877/454RI.asci1571.
[10] T. Kotsiopoulos, P. Sarigiannidis, D. Ioannidis, and D. Tzovaras, "MACHINE LEARNING AND DEEP LEARNING IN SMART MANUFACTURING: THE SMART GRID PARADIGM," *Comput. Sci. Rev.*, vol. 40, no. 100341, 2021, doi: https://doi.org/10.1016/j.cosrev.2020.100341.
[11] P. Novianti, D. Setyorini, and U. Rafflesia, "K-MEANS CLUSTER ANALYSIS IN EARTHQUAKE EPICENTER CLUSTERING," *Int. J. Adv. Intell. Informatics*, vol. 3, no. 2, pp. 81–89, 2017, doi: https://doi.org/10.26555/ijain.v3i2.100
[12] F. Reviantika, C. N. Harahap, and Y. Azhar, "ANALISIS GEMPA BUMI PADA PULAU JAWA MENGGUNAKAN CLUSTERING ALGORITMA K-MEANS," *J. Din. Inform.*, vol. 9, no. 1, pp. 51–60, 2020.
[13] N. Dwitiyanti, S. Ayu Kumala, and S. Dwi Handayani, "PENERAPAN METODE K-MEANS PADA KLASTERISASI WILAYAH RAWAN GEMPA DI INDONESIA IMPLEMENTATION OF K-MEANS METHOD IN CLASSTERIZATION OF EARTHQUAKE PRONE AREAS IN INDONESIA," *Pros. Semin. Nas. UNIMUS*, vol. 6, pp. 1029–1037, 2023.
[14] B. S. Febriani and R. F. Hakim, "ANALISIS CLUSTERING GEMPA BUMI SELAMA SATU BULAN TERAKHIR DENGAN MENGGUNAKAN ALGORITMA SELF-ORGANIZING MAPS (SOMS) KOHONEN," *Pros. Semin. Nas. Mat. dan Pendidik. Mat. UMS*, pp. 715–722, 2015.
[15] T. Ariawan and Supatman, "KLASTERISASI GEMPA BUMI DI PESISIR SELATAN JAWA DAN LAMPUNG MENGGUNAKAN ALGORITMA SELF-ORGANIZING MAPS (SOM) KOHONEN," no. November, pp. 171–177, 2019.
[16] R. Nainggolan, R. Perangin-angin, E. Simarmata, and F. A. Tarigan, "IMPROVED THE PERFORMANCE OF THE K-MEANS CLUSTER USING THE SUM OF SQUARED ERROR (SSE) OPTIMIZED BY USING THE ELBOW METHOD," *J. Phys. Conf. Ser.*, vol. 1361, no. 012015, 2019, doi: https://doi.org/10.1088/1742-6596/1361/1/012015.
[17] H. Zhao, "DESIGN AND IMPLEMENTATION OF AN IMPROVED K-MEANS CLUSTERING ALGORITHM," *Mob. Inf. Syst.*, 2022, doi: https://doi.org/10.1155/2022/6041484.
[18] F. H. Awad, "IMPROVED K-MEANS CLUSTERING ALGORITHM FOR BIG DATA BASED ON DISTRIBUTED SMARTPHONE NEURAL ENGINE PROCESSOR," *Electronics*, vol. 11, no. 883, 2022, doi: https://doi.org/10.3390/electronics11060883.
[19] T. Ahmad and H. Chen, "A REVIEW ON MACHINE LEARNING FORECASTING GROWTH TRENDS AND THEIR REAL-TIME APPLICATIONS IN DIFFERENT ENERGY SYSTEMS," *Sustain. Cities Soc.*, 2019, doi: https://doi.org/10.1016/j.scs.2019.102010.
[20] A. Jamil, A. A. Hameed, and Z. Orman, "A FASTER DYNAMIC CONVERGENCY APPROACH FOR SELF-ORGANIZING MAPS," *Complex Intell. Syst.*, vol. 9, no. 1, pp. 677–696, 2023, doi: https://doi.org/10.1007/s40747-022-00826-2.
[21] S. Licen, A. Astel, and S. Tsakovski, "SELF-ORGANIZING MAP ALGORITHM FOR ASSESSING SPATIAL AND TEMPORAL PATTERNS OF POLLUTANTS IN ENVIRONMENTAL COMPARTMENTS: A REVIEW," *Sci. Total Environ.*, vol. 878, no. March, p. 163084, 2023, doi: https://doi.org/10.1016/j.scitotenv.2023.163084.
[22] M. Scholz, "WETLANDS FOR WATER POLLUTION CONTROL (SECOND EDITION)," in *Wetlands for Water Pollution Control*, Elsevier Inc, 2016, pp. 239–257.doi: https://doi.org/10.1016/B978-0-444-63607-2.00026-5
[23] D. Miljković, "BRIEF REVIEW OF SELF-ORGANIZING MAPS," in *MIPRO*, 2017, pp. 1252–1257, doi: https://doi.org/10.23919/MIPRO.2017.7973581.
[24] M. Shutaywi and N. N. Kachouie, "SILHOUETTE ANALYSIS FOR PERFORMANCE EVALUATION IN MACHINE LEARNING WITH APPLICATIONS TO CLUSTERING," *Entropy*, vol. 23, no. 6, p. 759, 2021, doi: https://doi.org/10.3390/e23060759.
[25] F. Majidi, "A HYBRID SOM AND K-MEANS MODEL FOR TIME SERIES ENERGY CONSUMPTION CLUSTERING," 2023, doi: https://doi.org/10.48550/arXiv.2312.11475.
[26] E. J. Ramos, A. D., Lopez-Rubio, E., and Palomo, "THE ROLE OF THE LATTICE DIMENSIONALITY IN THE SELF-ORGANIZING MAP," *Neural Netw. World*, vol. 1, pp. 57–85, 2018.doi: https://doi.org/10.14311/NNW.2018.28.004
[27] S. W. Ullah, A., Haydarov, K., Ul Haq, I., Muhammad, K., Rho, S., Lee, M., and Baik, "DEEP LEARNING ASSISTED BUILDINGS ENERGY CONSUMPTION PROFILING USING SMART METER DATA," *Sensors*, vol. 20, no. 3, p. 873, 2020.doi: https://doi.org/10.3390/s20030873
[28] Badan Meteorologi Klimatologi dan Geofisika, *KATALOG GEMPABUMI SIGNIFIKAN & MERUSAK TAHUN 1821-2023*. Pusat Gempabumi dan Tsunami, 2024.