Journal Scientific of Mandalika (jsm) e-ISSN: 2745-5955, p-ISSN: 2809-0543, Vol. 6, No.8, 2025
website: http://ojs.cahayamandalika.com/index.php/jomla
Accredited Sinta 5 based on SK. No. 177/E/KPT/2024

Waveform Analysis of Broadband Seismic Station Using Machine Learning Python Based on Morlet Wavelet

Simon Simarmata
Universitas Pamulang, Indonesia
*Corresponding Author e-mail: dosen02300@unpam.ac.id

Abstract: *Wavelet signal processing is broadly used for analysis of real time seismic signal. The numerous wavelet filters are developed by spectral synthesis using machine learning python to realize the signal characteristics. Our paper aims to solve and evaluating the frequencies-energy characteristic of earthquake. The wavelet method by Continuous Wavelet Transform (CWT) is able to clearly and simultaneously of amplitudes and frequency-energy from component between the seismogram which seismic sensor broadband recorded in the January 16, 2017 in Medan, North Sumatra. Finally, from machine learning python with morlet wavelet allows good time resolution for high frequencies, and good frequency resolution for low frequencies.*
Key Words:*earthquake clustering analysis, developing tourism, clustering algorithms—K-means, DBSCAN*

Introduction

Tourism at Lake Toba in North Sumatra, Indonesia, has become a primary focus for regional economic development in recent years. Known as the largest volcanic lake in the world, Lake Toba not only offers stunning natural beauty but is also rich in cultural and historical significance. However, despite its substantial potential as a tourist destination, efforts to increase visitor numbers and attract interest from both local and international communities remain significant challenges.The increase in tourist visits to Lake Toba has become a central concern for the government and tourism managers. In recent years, efforts to promote Lake Toba as a premier tourist destination have seen significant enhancement. However, the public's response and sentiment toward these developments vary, encompassing various social, economic, cultural, and environmental aspects. One important aspect of evaluating the impact of tourism is understanding how the community responds to the promotion of Lake Toba through social media. Social media has become a primary platform for sharing information, opinions, and experiences related to tourism. Sentiment analysis on social media can provide valuable insights into how promotional campaigns are understood and accepted by the public, as well as how this influences tourists' interest and decisions to visit Lake Toba. In this context, it is essential to comprehend how public sentiment on social media affects the enthusiasm and number of visitors to Lake Toba. This evaluation should not only encompass positive responses to tourism promotion but also address concerns regarding environmental preservation, sustainable tourism, and the economic benefits for local communities. Such understanding is necessary to develop more effective strategies for promoting Lake Toba as a leading and sustainable tourist destination with high global competitiveness.

Considering these various aspects, an in-depth analysis of public sentiment regarding tourism at Lake Toba on social media will provide a strong foundation for developing tourism policies that are more inclusive, sustainable, and oriented toward the interests of both local and global communities. Wavelet transform is a method used to analyze signals of a certain time period and frequency. A very common model used in analyzing seismic signals is the Morlet wavelet with a very sharp ability to capture changes in frequency and time in seismic signals
Morlet wavelet in combining sinusoidal waves after being paired with a Gaussian window so that it produces better signal analysis and has variations and combinations of fast or slow variations
It can be concluded

The results of the discussion on how the Morlet wavelet transform can contribute to providing benefits in identifying and analyzing hidden frequencies in broadband seismic data, which basically cannot be understood by traditional spectrum analysis such as the Fourier transform

Points that are highly prioritized in the discussion

1. Application of wavelet transformation with Molet wavelet
2. Machine learning for seismic analysis
3. Broadband Seismic Data Processing
4. Advantages and challenges if using the python application in analyzing
5. Validation and the importance of model evaluation
6. Simple applications and limitations and suggestions for further research

Background reasons and purposes for undertaking the project

One of the most notable geological features in the world is the Toba Caldera, which is situated in North Sumatra, Indonesia. The caldera, which was created by a volcanic explosion around 74,000 years ago, has a rich geological past in addition to possible volcanic hazards. Thanks to significant scientific study and technical advancements, our understanding of seismic and volcanic dynamics has expanded significantly in recent decades. The importance of protecting geopark areas like Kaldera Toba is becoming more and more important, along with the need to protect natural areas while ensuring visitor safety. One of the most important lessons learned from this experience is to recognize and understand the role that abiotic activities play in affecting the cleanliness and health of the affected area. With the development of data analytics and artificial intelligence (AI), new avenues for addressing these issues are opening up. Deeper understanding of the seismic patterns and hazards in the area can be obtained by applying AI to earthquake clustering analysis. We can predict possible hazards, find patterns in earthquake activity, and create more efficient mitigation plans by using this data-driven method.

Earthquakes, as one of the most devastating natural phenomena, pose significant threats to human life, infrastructure, and economies. The region of North Sumatra as shown in fig. 1, situated along the Pacific Ring of Fire, is particularly prone to frequent and severe seismic activities due to its tectonic setting[1]. Understanding the patterns and behaviors of earthquakes in this region is critical for disaster preparedness, risk mitigation, and the formulation of effective response strategies. Clustering algorithms, which group similar data points based on specified characteristics, have proven to be powerful tools in seismology for analyzing seismic events. These techniques help identify patterns, trends, and anomalies within earthquake data, enabling researchers to classify events based on parameters such as location, magnitude, and depth. By doing so, clustering algorithms can reveal hidden structures within the data that might be indicative of underlying geological processes or potential future seismic hazards.

Figure 1. Distribution of Earthquake in northern Sumatra 2019-2022

In addition to enhancing the geopark's sustainability and safety, this strategy intends to provide visitors with more precise and current information on seismic and geological activity. Toba Caldera has the potential to become an exemplar of intelligent and sustainable geopark management through the integration of AI technology, seismic data, and regional development plans. The purpose of this study is to investigate how the creation of the Toba Caldera Geopark may be aided by the use of AI techniques for earthquake clustering. It is envisaged that by applying machine learning techniques and geospatial data analysis, more effective management plans may be developed to safeguard and make the most use of this priceless geological heritage.

This study focuses on the application of three prominent clustering algorithms—K-means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and Fuzzy C-Means—to earthquake data from North Sumatra. Each of these algorithms offers unique advantages: K-means is renowned for its simplicity and efficiency in partitioning data into clusters; DBSCAN excels in identifying clusters of varying shapes and densities while effectively handling noise; and Fuzzy C-Means allows for overlapping clusters, providing a more nuanced understanding of data points that may belong to multiple clusters.

Hence, the objective of this study were:

- to compare the effectiveness of these fundamentally different clustering algorithms in classifying seismic events in North Sumatra
- By evaluating K-means, DBSCAN, and Fuzzy C-Means, this study aims to determine which method is best suited for clustering earthquakes in this region. The analysis will reveal how much the results differ across the algorithms or if they yield similar outcomes.
- This insight could enhance earthquake prediction models and improve disaster readiness in regions susceptible to earthquakes.

Method

Related Research

Numerous studies have demonstrated the application and efficacy of various clustering techniques in analyzing seismic data. Partition-based clustering methods such as K-means and Fuzzy C-Means have been used to classify seismic events based on location, magnitude, and depth, finding K-means to be particularly effective [2]. The Spatiotemporal Extended Fuzzy C-Means (SEFCM) algorithm has been applied to earthquake data from Southern Italy, effectively detecting and predicting seismic hotspots compared to ST-DBSCAN [3]. In Indonesia, K-Affinity Propagation and K-means clustering have been employed for classifying earthquake data, validated with several clustering indices [4]. A two-stage clustering model using K-means and a variable DBSCAN algorithm was proposed to analyze seismic activities in the Himalaya and Sumatra–Andaman regions, demonstrating the method's capability in declustering earthquake catalogs [5]. Additionally, a spatial cluster analysis of land seismicity in Northern Sumatra using the K-Medoids algorithm analyzed seismic data from January 2019 to 2023, identifying optimal cluster configurations that emphasized distinctions in depth and geographical location, significantly contributing to seismic hazard assessment in North Sumatra [6]. Similarly, the K-Medoids clustering method was used to analyze seismic activity in West Java, identifying significant seismic clusters with implications for earthquake monitoring and hazard assessment in the region [7]. By building on these foundational studies, the present research aims to provide a comprehensive comparison of K-means, DBSCAN, and Fuzzy C-Means in the context of

earthquake clustering in North Sumatra, thereby contributing to the enhanced understanding and prediction of seismic activities in the region.

Data and Model

The earthquake data for Northern Sumatra, recorded between January 2019 and December 2022, was provided by the Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG). During this period, a total of 1069 seismic events were documented. The depths of these earthquakes ranged from the shallowest at 1 kilometer to the deepest at 209 kilometers. The recorded magnitudes varied, with the highest being 5.4 and the lowest at 0.9. The distribution of earthquake data in North Sumatra is visualized through histograms (fig. 2), providing insights into the spread of various features: longitude, latitude, magnitude, and depth. The longitude distribution shows that most earthquake occurrences are concentrated between 97.5 and 99.0 degrees, with a slight skewness indicating specific longitudinal regions where seismic events are frequent. The latitude distribution reveals that earthquakes predominantly occur between 1.5 and 2.5 degrees, with multiple peaks suggesting distinct regions of seismic activity within this latitude range.

The magnitude distribution follows a near-normal curve, with most earthquakes having magnitudes between 2.5 and 3.5. This indicates that moderate earthquakes are more common in this region, while both very weak and very strong earthquakes are less frequent. The depth distribution is highly skewed towards shallow depths (0-50 km), indicating that most earthquakes occur at these depths. There is a sharp decline in the frequency of deeper earthquakes, with very few events occurring at depths greater than 100 km.

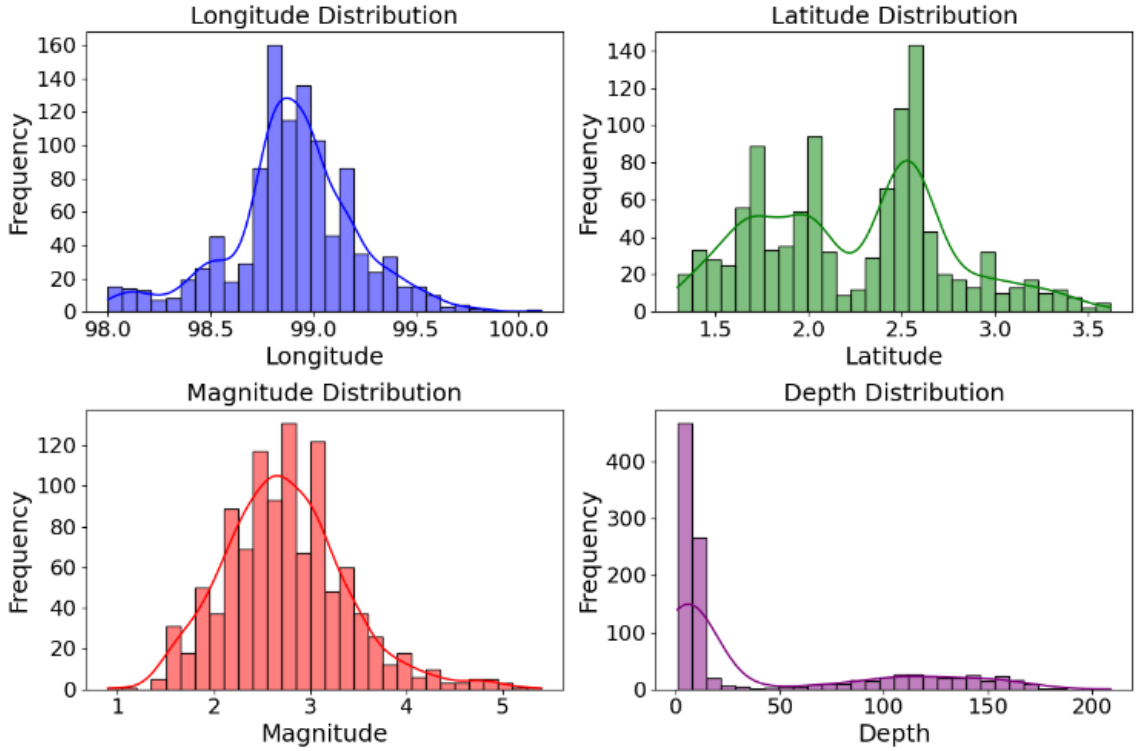

Figure. 2. Earthquake distributions plot

K-means clustering is a widely used partition-based algorithm that aims to divide a dataset into K clusters, where each data point belongs to the cluster with the nearest mean. The algorithm iterates between two steps: assigning data points to the nearest cluster mean and

updating the cluster means based on these assignments. To determine the optimal number of clusters (*K*), the elbow method is used, examining the within-cluster sum of squares for *K* values ranging from 2 to 10. Additionally, a grid search is performed for different random states to enhance the robustness of the clustering solution. Recent studies have enhanced K-means in various ways. One approach proposes an unsupervised learning schema for K-means that automatically determines the optimal number of clusters, addressing a key limitation of traditional K-means [8]. Another improvement involves developing a new similarity calculation method based on weighted Euclidean distance, enhancing both efficiency and correctness [9]. Additionally, a robust deep K-means model has been proposed that leverages deep learning to improve clustering performance by learning hierarchical representations, allowing for more effective clustering of complex datasets [10].

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based clustering algorithm capable of identifying clusters of arbitrary shape and handling noise. It defines clusters as areas of high density separated by areas of low density. DBSCAN identifies core points, which have a sufficient number of neighboring points within a specified radius (ϵ), and expands clusters from these core points. The algorithm is governed by two parameters: ϵ and the minimum number of points required to form a dense region (MinPts). Key advancements in DBSCAN include the introduction of KNN-BLOCK DBSCAN, an approximate algorithm that uses k-nearest neighbors to enhance clustering speed for large datasets [11]. Additionally, fuzzy extensions of DBSCAN have been developed to better identify clusters with variable density distributions and overlapping borders, increasing flexibility and accuracy [12]. To determine the initial value of ϵ, K-nearest neighbors (KNN) with *K* ranging from 4 to 21 are used to compute the distances to the nearest neighbors, which helps in identifying a suitable radius for clustering. The optimal values for ϵ and min_samples are found through grid search. A range of ϵ values is tested to determine the best distance threshold that defines neighborhoods, while min_samples is varied to find the number of points required to form dense regions. Evaluation is done by ensuring valid clustering results (more than one cluster and not all points in one cluster) and optimizing the silhouette score.

Fuzzy C-Means (FCM) is a clustering algorithm that allows each data point to belong to multiple clusters with varying degrees of membership. This algorithm minimizes an objective function that represents the weighted distance between data points and cluster centers. Recent advancements include revisions to handle unequal cluster sizes, noise, and outliers, improving robustness and accuracy [13]. Enhancements to FCM have been made by optimizing the initialization of cluster centers and the merging process, with applications in mental health intelligent evaluation systems [14]. The optimal number of clusters (*c*) and the fuzziness parameter (*m*) are determined through grid search. Various values for *c* are tested to find the best number of clusters, while different *m* values are evaluated to control the degree of membership in clusters. The best parameters are those that provide the highest silhouette score, indicating the most effective clustering.

In this HPC-integrated clustering approach, we aim to optimize the execution of K-means, DBSCAN, and Fuzzy C-Means algorithms by running them concurrently across a high-performance computing (HPC) environment. The process begins with the initialization of the HPC environment, where necessary modules and libraries are loaded, and parallel or distributed computation frameworks like Dask or MPI are set up. This environment allows for the efficient distribution of tasks across multiple cores, nodes, or GPUs, ensuring that computational resources are fully utilized from the outset.

The core of the process involves parallel grid searches for each clustering algorithm. These searches are launched simultaneously, with each algorithm's parameter combinations distributed across the available HPC resources. For K-means, DBSCAN, and Fuzzy C-Means, the algorithm-specific processes loop through different parameter combinations, fitting models, calculating Silhouette scores, and returning the results. This parallel execution significantly reduces the time required to identify the best parameters for each algorithm, as the tasks are handled concurrently rather than sequentially. Once the grid searches are complete, the results are aggregated, and the best-performing models are identified.

Following the parallel execution, the best models from each clustering algorithm are applied to the dataset to generate clusters. A post-processing step ensures that noise points are filtered out (specifically for DBSCAN), and cluster labels are adjusted as needed. The final results, including clustered data and any necessary visualizations, are then outputted. This approach not only accelerates the clustering process but also leverages the full power of HPC to handle large datasets and complex computations, making it ideal for tasks like earthquake clustering analysis in regions like North Sumatra.

The K-Medoids algorithm is an extension of K-means, and falls within the domain of partitional clustering. Its main objective is to minimise the distances between the data points within a cluster and a particular data point, which is referred to as the medoid. The medoid, in this context, represents the central and most representative unit within the cluster. In particular, K-Medoids is robust to noise and outliers in the data set. This makes it a reliable clustering method. Unlike K-Means, which uses cluster centroids, K-Medoids uses actual data points as cluster representatives, making it versatile for various data grouping tasks [3][4].

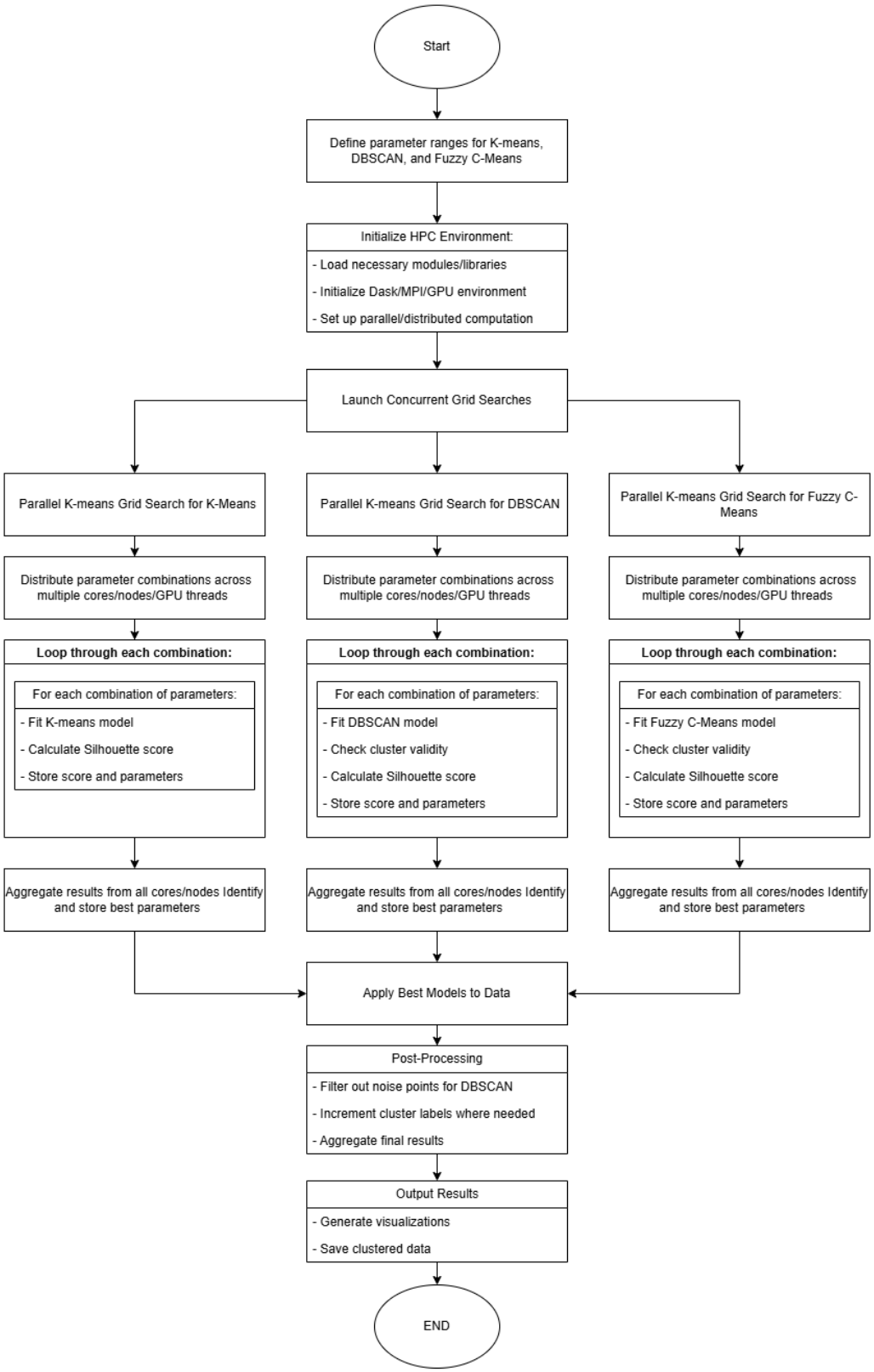

Figure. 3. Earthquake distributions plot

We assess clustering models using three pivotal metrics: the Silhouette Score, Calinski-Harabasz Index (CH), and Davies-Bouldin Index (DB). The Silhouette Score gauges the similarity of each data point to its own cluster versus others, with higher scores indicating clearer clusters. Its utility in determining optimal cluster numbers has been validated, and its applicability has been extended to complex cluster shapes[15], [16]. The Calinski-Harabasz Index measures clustering effectiveness by comparing between-cluster to within-cluster dispersion. It's computed as,

$$CH = \frac{SSB}{SSW} x \frac{N-k}{k-1} \quad (1)$$

where is the SSB between-cluster dispersion, SSW is the within-cluster dispersion, N is the total number of data points, and k is the number of clusters. Combine this index with the Silhouette Score results in more robust cluster evaluation practical scenarios [17], [18].
The Davies-Bouldin Index evaluates cluster similarity, aiming for lower values indicating better clustering quality. It's expressed as

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{j \neq i}\left(\frac{S_i+S_j}{d_{ij}}\right) \quad (2)$$

where S_i and S_j represent the average distances of points within clusters i and j, respectively, and d_{ij} is the distance between the centroids of clusters i and j. This index is used to optimize clustering quality by minimizing the score. It has been applied to various scenarios, including optimizing school facility clustering and developing fuzzy variants for overlapping clusters [19], [20]. These metrics collectively provide a robust framework for evaluating clustering accuracy and reliability. By leveraging these sophisticated measures, we gain deeper insights into seismic activity patterns in North Sumatra, ensuring our analysis is rigorous and impactful.

Innovation and practical application

This study investigates the efficacy of three clustering algorithms—K-means, DBSCAN, and Fuzzy C-Means—in classifying seismic events in North Sumatra. Earthquake data spanning from January 2019 to December 2022, provided by the Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG), was analyzed to evaluate how well each algorithm identifies and organizes clusters based on earthquake characteristics such as longitude, latitude, magnitude, and depth. K-means excelled in producing well-separated, spherical clusters; DBSCAN effectively detected clusters of varying densities and identified noise points; Fuzzy C-Means offered insights into overlapping clusters and gradual transitions. Despite some differences in the clustering results, all methods provided similar outcomes, with DBSCAN uniquely highlighting noise points. These insights contribute to a better understanding of seismic activity in North Sumatra and could improve earthquake prediction models and disaster preparedness strategies.

Conclusion

This study shows that the Morlet wavelet transform and machine learning using Python will provide a short and effective way to analyze broadband seismic signals. The combination

used can help researchers improve their understanding of seismic signal behavior, noise filtering and how to make more accurate predictions about seismic events.

V. Proposed development (time line)

1. Year One: Data collection, literature study, research proposal, Prototype for cluster of earthquakes. (Spatial Cluster Analysis of Land Seismicity in Geopark Kaldera Toba using The K-Medoids Algorithm)
2. Year two: Completing Automation of Earthquake Clustering, Publication, literature study. (Clustering Analysis of earthquake based on K-means, DBSCAN, and Fuzzy C-Means in Geopark Kaldera Toba)

Reference:

[1] E. Darnila, K. Tarigan, F. Grantianus Nafiri Larosa, and M. Sinambela, "Cluster Analysis And Seismicity Partioning For Northern Sumatera Using Machine Learning ApproacH 1*," *J Theor Appl Inf Technol*, vol. 31, no. 2, 2021, [Online]. Available: www.jatit.org

[2] R. Alom, A. Mazumdar, R. Kr Prasad, G. Basumatary, and B. Baruah, "Analysis of Seismic Data Using Partition-Based Clustering Techniques," in *2022 IEEE Global Conference on Computing, Power and Communication Technologies (GlobConPT)*, 2022, pp. 1–6. doi: 10.1109/GlobConPT57482.2022.9938362.

[3] F. Di Martino, W. Pedrycz, and S. Sessa, "Spatiotemporal extended fuzzy C-means clustering algorithm for hotspots detection and prediction," *Fuzzy Sets Syst*, vol. 340, pp. 109–126, Jun. 2018, doi: 10.1016/j.fss.2017.11.011.

[4] M. Muhajir and N. N. Sari, "K-Affinity Propagation (K-AP) and K-Means Clustering for Classification of Earthquakes in Indonesia," in *2018 International Symposium on Advanced Intelligent Informatics (SAIN)*, IEEE, Aug. 2018, pp. 6–10. doi: 10.1109/SAIN.2018.8673344.

[5] R. K. Vijay and S. J. Nanda, "A Variable \varepsilon -DBSCAN Algorithm for Declustering Earthquake Catalogs," 2019, pp. 639–651. doi: 10.1007/978-981-13-1592-3_50.

[6] H. H. Arrizal, M. Sinambela, Widodo, S. P. Adi, and H. T. Frianto, "Spatial Cluster Analysis of Land Seismicity in the Northem Part of Sumatra Using the K-Medoids Algorithm," in *2023 International Conference on Information Technology and Computing (ICITCOM)*, 2023, pp. 181–185. doi: 10.1109/ICITCOM60176.2023.10442535.

[7] N. F. Saragih, Y. Y. Pratiwi, I. K. Jaya, I. M. Sarkis, H. H. Arrizal, and M. Sinambela, "Clustering Earthquakes in West Java Using Machine Learning Algorithm," in *2023 International Conference of Computer Science and Information Technology (ICOSNIKOM)*, 2023, pp. 1–4. doi: 10.1109/ICoSNIKOM60230.2023.10364541.

[8] K. P. Sinaga and M.-S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[9] Y. Zhao and X. Zhou, "K-means Clustering Algorithm and Its Improvement Research," *J Phys Conf Ser*, vol. 1873, no. 1, p. 012074, Apr. 2021, doi: 10.1088/1742-6596/1873/1/012074.

[10] S. Huang, Z. Kang, Z. Xu, and Q. Liu, "Robust deep <math altimg="si1.svg"> <mi>k</mi> </math> -means: An effective and simple method for data clustering," *Pattern Recognit*, vol. 117, p. 107996, Sep. 2021, doi: 10.1016/j.patcog.2021.107996.

[11] Y. Chen *et al.*, "KNN-BLOCK DBSCAN: Fast Clustering for Large-Scale Data," *IEEE Trans Syst Man Cybern Syst*, vol. 51, no. 6, pp. 3939–3953, Jun. 2021, doi: 10.1109/TSMC.2019.2956527.

[12] D. Ienco and G. Bordogna, "Fuzzy extensions of the DBScan clustering algorithm," *Soft comput*, vol. 22, no. 5, pp. 1719–1730, Mar. 2018, doi: 10.1007/s00500-016-2435-0.

[13] S. Askari, “Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development,” *Expert Syst Appl*, vol. 165, p. 113856, Mar. 2021, doi: 10.1016/j.eswa.2020.113856.
[14] S. Hu, “Fuzzy C-means Clustering Algorithm and Its Application in Mental Health Intelligent Evaluation System,” in *2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS)*, IEEE, Dec. 2022, pp. 1305–1309. doi: 10.1109/TOCS56154.2022.10016042.
[15] K. R. Shahapure and C. Nicholas, “Cluster Quality Analysis Using Silhouette Score,” in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, Oct. 2020, pp. 747–748. doi: 10.1109/DSAA49011.2020.00096.
[16] D. Cheng, Q. Zhu, J. Huang, Q. Wu, and L. Yang, “A Novel Cluster Validity Index Based on Local Cores,” *IEEE Trans Neural Netw Learn Syst*, vol. 30, no. 4, pp. 985–999, Apr. 2019, doi: 10.1109/TNNLS.2018.2853710.
[17] V. M. Vergara, M. Salman, A. Abrol, F. A. Espinoza, and V. D. Calhoun, “Determining the number of states in dynamic functional connectivity using cluster validity indexes,” *J Neurosci Methods*, vol. 337, p. 108651, May 2020, doi: 10.1016/j.jneumeth.2020.108651.
[18] X. Wang and Y. Xu, “An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index,” *IOP Conf Ser Mater Sci Eng*, vol. 569, no. 5, p. 052024, Jul. 2019, doi: 10.1088/1757-899X/569/5/052024.
[19] Y. Arie Wijaya, D. Achmad Kurniady, E. Setyanto, W. Sanur Tarihoran, D. Rusmana, and R. Rahim, “Davies Bouldin Index Algorithm for Optimizing Clustering Case Studies Mapping School Facilities,” *TEM Journal*, pp. 1099–1103, Aug. 2021, doi: 10.18421/TEM103-13.
[20] A. A. Vergani and E. Binaghi, “A Soft Davies-Bouldin Separation Measure,” in *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, IEEE, Jul. 2018, pp. 1–8. doi: 10.1109/FUZZ-IEEE.2018.8491581.