Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 3, **Number** 2, November 2023
https://doi.org/10.47709/brilliance.v3i2.3501

Advanced Seismic Data Analysis: Comparative study of Machine Learning and Deep Learning for Data Prediction and Understanding

Gregorius Airlangga[1*]
[1]Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia
[1]gregorius.airlangga@atmajaya.ac.id

open access

***Corresponding Author**

Article History:
Submitted: 04-12-2023
Accepted: 24-12-2023
Published: 28-12-2023

Keywords:
Seismic Data Analysis, Machine Learning, Deep Learning, Autoencoder, Clustering

Brilliance: Research of Artificial Intelligence is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

ABSTRACT

This study delves into the application of machine learning (ML) and deep learning (DL) techniques for the analysis of seismic data, aiming to identify and categorize patterns and anomalies within seismic events. Using a robust dataset, we applied three distinct clustering approaches: K-Means, DBSCAN, and an Autoencoder-based method, each offering unique perspectives on the data. K-Means clustering provided a fundamental partitioning of the data into five predefined clusters, facilitating the identification of broad seismic patterns. DBSCAN, a density-based clustering algorithm, offered insights into the spatial distribution and density of seismic events, adeptly pinpointing anomalies and outliers that signify unusual seismic activity. The Autoencoder, leveraging deep learning, excelled in capturing complex and non-linear relationships within the data, revealing subtle patterns not immediately apparent through traditional methods. The effectiveness of these clustering techniques was quantitatively evaluated using the Silhouette Score and the Davies-Bouldin Score, alongside visual assessments through PCA and t-SNE for dimensionality reduction. The results indicated that while K-Means provided clear partitioning, DBSCAN excelled in outlier detection, and the Autoencoder offered a balanced approach with its nuanced analysis capabilities. Our comprehensive analysis underscores the significance of employing a multi-methodological approach in seismic data analysis, as each method contributes uniquely to the understanding of seismic events. The insights gained from this study are valuable for enhancing predictive models and improving disaster risk management strategies in seismology. Future research directions include the integration of additional seismic features, validation against larger datasets, and the development of hybrid models to further refine the predictive accuracy of seismic event analysis.

INTRODUCTION

The study of seismic activities, traditionally a domain of geology and physical sciences, has undergone a significant transformation with the advent of computational methodologies. This transformation represents a pivotal shift in seismic research, moving from basic statistical models to advanced machine learning (ML) and deep learning (DL) techniques (AlAli & Anifowose, 2022; Dramsch, 2020; Mousavi & Beroza, 2022). In the early stages of computational seismic analysis, efforts were primarily centered around statistical methods for pattern recognition in seismic data (Gerstenberger et al., 2020). These initial approaches provided fundamental insights but often struggled to effectively manage and interpret complex and voluminous datasets (Jiao & Alavi, 2020; Mousavi & Beroza, 2022). With the emergence of ML algorithms such as K-Means and DBSCAN, a new era in seismic analysis was ushered in. These algorithms enabled a more refined categorization and understanding of seismic zones, marking a stark contrast to earlier methodologies (Aranda, Sele, Etchanchu, Guyt, & Vaara, 2021). Alongside these developments, deep learning techniques, particularly neural networks such as Autoencoders, began to gain prominence (Zhao, Han, Ouyang, & Burke, 2023). Their ability to detect subtle and complex patterns in seismic datasets, a task that posed significant challenges to traditional methods, showcased the potential of DL in revolutionizing seismic data analysis (Ros, Riad, & Guillaume, 2023). The urgency and significance of this research cannot be overstated. The increasing global incidence of seismic events, coupled with their potential for catastrophic consequences, underscores the need for advanced research in this area (Xu & Haq, 2022). The primary objective of this research is to develop methodologies that can accurately analyze and predict seismic events, thereby playing a crucial role in enhancing preparedness and mitigating risks associated with these natural phenomena (Asming et al., 2022). This research is particularly significant as it bridges the gap between traditional seismic analysis methods and modern computational techniques. Such a bridge is essential for improving the accuracy of earthquake predictions and for advancing disaster management strategies, ultimately contributing to the safety and well-being of populations in seismically active regions (Johnson, Ben-Zion, Meng, & Vernon, 2020).

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 3, **Number** 2, November 2023
https://doi.org/10.47709/brilliance.v3i2.3501

The current state of the art in seismic data analysis involves a combination of traditional geophysical methods and advanced computational techniques. While traditional methods have laid a strong foundation for understanding seismic phenomena, the introduction of ML and DL has propelled the field into a new era of possibilities. Techniques such as K-Means clustering and DBSCAN have opened new avenues for understanding the clustering of seismic events (Ghasemi & Stephens, 2022). Deep learning techniques, particularly Autoencoders, have shown great promise in identifying complex seismic patterns that were previously undetectable with conventional methods. Despite significant advancements in the field, there remains a notable gap in the comparative study of diverse ML and DL techniques in seismic research (Soh & Demiris, 2015). Many existing studies have focused on the application of individual methods without fully exploring the synergistic potential of combining various computational algorithms (X. Liu et al., 2021). This research aims to address this gap by providing an integrated, multifaceted approach that leverages the strengths of various computational techniques in unison.

The study pioneers an innovative comparative study of various ML and DL techniques, offering a novel and comprehensive approach in the realm of seismic data analysis. By providing an in-depth comparative analysis of different algorithms, this research adds significant insights into their effectiveness and suitability for various aspects of seismic analysis. Furthermore, it advances methodologies in seismic risk assessment by incorporating cutting-edge computational techniques, thus enhancing the precision and reliability of seismic event predictions. The detailed structure of this journal article includes an expanded introduction that sets the broader context and highlights the importance of the research within the larger field of seismic studies. The methodology section offers an in-depth explanation of the data preprocessing steps, the implementation of each ML and DL algorithm, and a thorough rationale for the selection of each technique. The results and comprehensive discussion section present an extensive analysis of the findings, including intricate visual representations and a detailed discussion of the implications of these results in the broader context of seismic pattern recognition and anomaly detection.

A robust evaluation and metrics section provides a thorough evaluation using various metrics such as the Silhouette (Shutaywi & Kachouie, 2021) and Davies-Bouldin (Posamentier, Paumard, & Lang, 2022) scores, offering a quantitative and objective assessment of the methodologies employed. The in-depth comparative analysis contrasts the integrated computational approach with traditional seismic analysis methods, highlighting the improvements and advancements offered by the former. The implications and practical applications section explores the broader implications of the findings, discussing potential real-world applications in seismic risk assessment and disaster management strategies. This section showcases the practical utility of the research, emphasizing its relevance to real-world scenarios. The article concludes with a comprehensive summary of the research, its key contributions, and potential avenues for future research in the field. An extensive bibliography is included, providing extensive citations of sources and literature integral to the research. Appendices provide additional supporting data, charts, and technical details, catering to readers seeking an in-depth analytical understanding of the methodologies and findings.

LITERATURE REVIEW

The study of seismic activities has undergone a remarkable evolution, transitioning from traditional geophysical methods to advanced computational techniques. This literature survey traces this transformation, highlighting key research contributions and setting the stage for our comparative study in seismic data analysis. Initially, seismic analysis was predominantly rooted in geology and physical sciences, as delineated by (Johnson et al., 2020). These traditional methods focused on the physical models of the Earth's crust and involved manual interpretation of seismic data, primarily through seismic wave analysis and fault line studies. While these methods were instrumental in laying the groundwork for understanding seismic activities, their qualitative nature often resulted in a lack of precision and difficulty in handling large datasets. This limitation underscored the need for more advanced, computational approaches in seismic analysis.

The introduction of statistical methods in seismic data analysis marked a significant shift from purely geological methods. (Ji, Wen, Ren, & Dhakal, 2020) explored the use of statistical models, focusing on pattern recognition in seismic data to identify common features in seismic events. These models provided more quantitative insights but were often limited in managing the complexity and volume of seismic datasets. The comparative of statistical models was a positive step; however, it was still inadequate for the vast and intricate nature of seismic data, highlighting the potential for more sophisticated computational techniques. The early application of machine learning algorithms like K-Means in seismic analysis, as discussed by (W. Liu & Hu, n.d.), represented a significant advancement. Their study demonstrated how K-Means clustering could categorize seismic events, enhancing the understanding of seismic zones and allowing for a more systematic analysis compared to traditional techniques. However, this study also indicated the necessity for more complex algorithms capable of identifying subtler patterns in the data, beyond the capabilities of K-Means.

Further advancement came with the introduction of density-based clustering algorithms, such as DBSCAN, in seismic analysis. (Ma & Mei, 2021) demonstrated how DBSCAN could effectively identify clusters of seismic events and outliers in spatial data. This method offered a more nuanced approach than K-Means, as it did not require predefined cluster numbers and was particularly adept in handling noisy datasets. However, the performance of

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 3, **Number** 2, November 2023
https://doi.org/10.47709/brilliance.v3i2.3501

DBSCAN in varying seismic contexts was not fully explored, indicating an area ripe for further research. The application of deep learning, specifically Autoencoders, for seismic data analysis marked a new frontier. (Mousavi & Beroza, 2023) were pioneers in demonstrating how Autoencoders could detect complex patterns in seismic datasets, a task challenging for traditional and simpler ML methods. This study showcased the potential of deep learning in seismic analysis, particularly in anomaly detection, but it also highlighted the need for comparative study of these techniques with other ML methods for a more comprehensive analysis.

A notable comparative study by (Noureldin, Ali, Sim, & Kim, 2022; Sun, Burton, & Huang, 2021) assessed various machine learning algorithms in seismic data analysis. This research compared different ML techniques, providing insights into the strengths and limitations of each method. While offering a comparative view, the study pointed towards the necessity of an comparative study that leverages the strengths of multiple ML and DL techniques for a holistic understanding of seismic data. Our research builds upon these foundational studies, aiming to synthesize diverse approaches into a cohesive framework. We compare K-Means and DBSCAN for advanced pattern recognition, leveraging their strengths in different seismic contexts. Our use of Autoencoders goes beyond anomaly detection, aiming to uncover deeper insights into the characteristics of seismic events. By combining these methods with PCA and t-SNE for data visualization, we provide a more intuitive understanding of complex seismic patterns. Our contribution lies in this comparative study of computational techniques, addressing the gaps identified in previous studies. We offer a comprehensive approach that not only enhances the accuracy of seismic event predictions but also provides a more nuanced understanding of seismic activities. This research advances the field of seismic data analysis by effectively combining machine learning and deep learning techniques, setting a new benchmark for future studies in this domain.

METHOD

In this research, we employed a comprehensive methodology that compares various machine learning (ML) and deep learning (DL) techniques to analyze seismic data. Our approach is structured to maximize the insights gained from earthquake datasets, leveraging the strengths of different computational methods to enhance pattern recognition, anomaly detection, and data visualization.

DATA COLLECTION AND PREPROCESSING

The initial phase of our research methodology involved a meticulous process of data collection and preprocessing, which are critical steps in ensuring the accuracy and reliability of our analysis. The dataset at the heart of our research encompassed an extensive range of seismic activities. It was carefully curated to include key parameters that are vital for a comprehensive analysis of seismic events. These parameters included the latitude and longitude, which provide spatial coordinates of the seismic events, as well as the depth and magnitude, which offer insight into the severity and potential impact of these events. The collection of this data involved aggregating information from various reliable seismic activity databases. Each source was vetted for accuracy and consistency, ensuring that the data we compiled was of the highest quality. This meticulous approach to data collection is crucial in the field of seismic analysis, as the precision of the input data directly affects the reliability of the study's conclusions.

Once collected, the data underwent a thorough preprocessing stage, which is essential in transforming raw data into a format suitable for analysis. One of the primary steps in this stage was the conversion of the date and time of each seismic event into a unified datetime format. This step is crucial for several reasons such as temporal consistency, ease of analysis, and data integrity. Firstly, seismic data often comes from multiple sources, each potentially using different formats for recording date and time. Unifying these into a single datetime format ensures consistency across the dataset, which is essential for accurate temporal analysis. After that, a unified datetime format simplifies subsequent data processing and analysis. It allows for more straightforward temporal comparisons and aggregations, which are often necessary in seismic data analysis. Furthermore, converting to a single datetime format helps maintain data integrity, especially when dealing with large datasets that span long periods. It ensures that all temporal data adhere to the same standards, reducing the likelihood of errors in analysis.

Following the datetime conversion, the original date and time columns were dropped from the dataset. This decision was made to streamline the dataset and focus the analysis on the most pertinent features. Reducing the dimensionality of the data in this manner helps to mitigate potential issues of overfitting and computational complexity in later stages of analysis. The preprocessing stage sets the foundation for the subsequent application of machine learning and deep learning techniques. By ensuring that the data is accurate, consistent, and suitably formatted, we lay the groundwork for a robust and reliable analysis of seismic activities.

FEATURE STANDARDIZATION

In the realm of machine learning, especially when dealing with datasets comprising various features, standardization is a crucial preprocessing step. The seismic dataset used in our research presented features (latitude, longitude, depth, and magnitude) with varying scales and units of measurement. Standardization plays a pivotal role in harmonizing these diverse scales, ensuring that each feature contributes equally to the analysis and model training. The

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 3, **Number** 2, November 2023
https://doi.org/10.47709/brilliance.v3i2.3501

process of standardization involves scaling the features of the dataset so that they have a mean of zero and a standard deviation of one. This normalization of data is essential for multiple reasons such as equal importance, improved algorithm performance, numerical stability, and faster convergence.

Firstly, equal importance, if the experiment is conducted without standardization, features with larger magnitudes or wider ranges could dominate the model training process, leading to biased results. Standardization ensures that each feature contributes equally, regardless of their original scale. Many machine learning algorithms, particularly those involving distance calculations (like K-Means clustering) or gradient descent optimization (common in deep learning), perform better when the input data is standardized. Secondly, algorithms can become numerically unstable when dealing with features of vastly differing magnitudes. Standardization helps mitigate this issue, leading to more reliable and stable algorithm performance. In addition, in optimization algorithms, standardized features often lead to faster convergence towards the minimum, as it ensures a more balanced and uniform scale for all features. In our methodology, we employed the StandardScaler from the sklearn library to perform this task. The StandardScaler works by calculating the mean and standard deviation for each feature in the dataset and then transforming the data according to the equation (1).

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma} \tag{1}$$

In this equation, X represents the original feature values, μ is the mean of the feature, and σ is the standard deviation. The scaled feature is represented as X_{scaled}, thus it has a mean of zero and a standard deviation of one. This transformation was applied to each feature in our dataset. Applying the StandardScaler effectively transformed our dataset into a format more suitable for the application of various machine learning algorithms. By ensuring that all features were on a comparable scale, we mitigated the risk of bias towards certain features and improved the overall effectiveness and reliability of our subsequent analyses.

K-MEANS CLUSTERING

K-Means clustering is a widely used algorithm in machine learning for partitioning a dataset into distinct groups, and we utilize it to discern inherent groupings within our seismic data. The fundamental principle of K-Means is to categorize the data into a specified number of clusters, with each cluster representing a grouping of data points that are like each other. In our study, we chose to divide the seismic data into five distinct clusters. The process of K-Means clustering involves several key steps:

1. Initialization: The algorithm starts by initializing 'k' centroids, which are the central points of the clusters. In our case, 'k' was set to five. These centroids can be initialized randomly or based on specific criteria.
2. Assignment: Each data point in the dataset is assigned to the nearest centroid, based on a distance metric, typically Euclidean distance. This step categorizes each seismic event into one of the five clusters.
3. Centroid Update: After all data points have been assigned to clusters, the centroids are recalculated as the mean of all points in the cluster. This step adjusts the position of the centroid to better represent the members of its cluster.
4. Iteration: Steps 2 and 3 are repeated iteratively until the centroids stabilize and no further changes occur in the membership of the clusters.

The objective of K-Means is to minimize the variance within each cluster, which in essence, is minimizing the sum of squared distances between each data point and its corresponding centroid. The objective function, or cost function, for K-Means is given in the equation (2).

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \tag{2}$$

In this equation, J represents the cost function, $x_i^{(j)}$ denotes the data points in cluster j, and c_j is the centroid of cluster j. The goal is to minimize J, thereby ensuring that each cluster is as compact as possible. By employing K-Means clustering on our seismic data, we were able to identify patterns and correlations that may not have been immediately apparent. This algorithm was particularly effective in revealing the spatial distribution of seismic events, grouping them into clusters based on their similarities in features such as location, depth, and magnitude. The resulting clusters provided valuable insights into the characteristics of seismic activities in different geographical regions, aiding in the understanding of seismic patterns and trends.

DBSCAN FOR ANOMALY DETECTION

In our seismic data analysis, we incorporated DBSCAN (Density-Based Spatial Clustering of Applications with Noise), a robust and versatile clustering algorithm, to complement the insights gained from K-Means clustering. DBSCAN stands out in the realm of clustering algorithms for its unique approach to defining clusters, based not on predetermined cluster numbers but on the density of data points. This characteristic of DBSCAN makes it especially adept at identifying outliers or anomalies, which are crucial in seismic analysis for flagging unusual or significant

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 3, **Number** 2, November 2023
https://doi.org/10.47709/brilliance.v3i2.3501

seismic events. DBSCAN classifies data points into clusters based on their density, fundamentally differing from algorithms like K-Means, which require the number of clusters to be specified a priori. The core concepts of DBSCAN revolve around two parameters such as Epsilon (ε): This parameter defines the radius around each data point to search for neighboring points and Minimum Points (MinPts): This parameter specifies the minimum number of points required to form a dense region, which is considered a cluster. Furthermore, there are several steps that have to be conducted in order to implement this algorithm such as:

1. Core Points Identification: Initially, the algorithm identifies 'core points' as those that have at least MinPts within their ε-neighborhood. These core points are considered the heart of the clusters.
2. Border Points Allocation: Points that are within the ε-neighborhood of a core point but do not have enough neighbors to be core points themselves are classified as 'border points'.
3. Noise Identification: Points that are neither core nor border points are labeled as 'noise' or outliers. These are points that do not belong to any cluster due to their low density.
4. Cluster Formation: The algorithm forms clusters by connecting core points that are within each other's ε-neighborhood, along with their associated border points.

This methodology enables DBSCAN to detect areas of high density, which are classified as clusters, and points that fall outside these dense areas, classified as anomalies or noise. The ability to identify such outliers is particularly valuable in seismic data analysis, as it allows for the detection of seismic events that deviate significantly from typical patterns. The mathematical representation of the DBSCAN criteria for core points can be expressed as presented in the equation (3).

$$\text{If } |N_\epsilon(x_i)| \geq \text{MinPts, then } x_i \text{ is a core point} \tag{3}$$

In this equation, $N_\epsilon(x_i)$ denotes the number of points within an ε-radius of point x_i, and MinPts is the minimum number of points specified to define a core point. The symbol $|\cdot|$ represents the cardinality, or the count, of points within the ε-neighborhood. Applying DBSCAN to seismic data allowed us to explore the dataset beyond conventional clustering. By identifying outliers and anomalies, we could flag seismic events that might indicate unique or significant geological phenomena. This capability is crucial for a comprehensive seismic analysis, as it provides insights into potentially high-impact seismic events that might be overlooked by other clustering methods. DBSCAN's flexibility in not requiring a predefined number of clusters also made it particularly suitable for our dataset, where the number of distinct seismic groupings was not known in advance. This approach provided a more natural and data-driven method of clustering, leading to potentially more meaningful and insightful results in the context of seismic activity analysis.

AUTOENCODER FOR DEEP LEARNING ANALYSIS

In our seismic data analysis, we incorporated an Autoencoder, a specialized type of neural network, to perform a deeper, more nuanced analysis. Autoencoders are particularly effective in unsupervised learning scenarios, where the goal is to learn a representation (encoding) for a set of data, typically for dimensionality reduction or anomaly detection. In our case, the Autoencoder was instrumental in detecting anomalies within the seismic data. The architecture of the Autoencoder used in our study consisted of several layers:

1. Input Layer: The first layer of the Autoencoder, where the seismic data is input into the network. The size of this layer corresponds to the number of features in our dataset.
2. Hidden Layers: We employed two hidden layers, one for encoding and one for decoding. The encoding layer compresses the input data into a lower-dimensional representation, capturing the most salient features of the data. The decoding layer then reconstructs the data back to its original dimensionality.
3. Output Layer: The final layer, where the reconstructed data is output. The size of this layer matches the size of the input layer, allowing for a comparison between the original and reconstructed data.

The training process of the Autoencoder involves adjusting its weights to minimize the difference between the input data and the reconstructed output. This process, known as reconstruction, is crucial for anomaly detection. The principle behind this is that the Autoencoder learns to efficiently represent typical data patterns during training. When presented with anomalous data, the network struggles to reconstruct it accurately, resulting in a higher reconstruction error. The mean squared error (MSE) is used to quantify this reconstruction error and is computed as presented in the equation (4).

$$\text{MSE} = \frac{1}{n}\sum\left(X - \hat{X}\right)^2 \tag{4}$$

In this equation, X represents the original input data, $\hat{X}$ denotes the reconstructed data output by the Autoencoder, and n is the number of data points. The MSE essentially measures the average squared difference between the original and reconstructed data. Seismic events that yield a high MSE are flagged as potential anomalies, indicating deviations from typical seismic patterns learned by the Autoencoder. The Autoencoder's ability to detect anomalies in

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 3, **Number** 2, November 2023
https://doi.org/10.47709/brilliance.v3i2.3501

seismic data is a significant asset in our analysis. It allowed us to identify seismic events that differed markedly from the norm, which could be indicative of unusual or significant seismic activity. This aspect of deep learning analysis is particularly valuable, as it provides an automated, data-driven approach to pinpointing anomalies that might require further investigation.

DIMENSIONALITY REDUCTION AND VISUALIZATION

In our approach to analyzing the high-dimensional seismic data, dimensionality reduction played a crucial role, particularly for visualization purposes. We employed two prominent techniques: Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), each serving a distinct purpose in our analysis. PCA is a statistical technique used to simplify the complexity in high-dimensional data by transforming it into a lower-dimensional space. This transformation is achieved while retaining as much of the variability in the dataset as possible. There are two kinds of PCA steps, firstly is PCA Process, this works by identifying the directions, called principal components, along which the variation in the data is maximal. In mathematical terms, PCA seeks to find a set of orthogonal vectors that define a subspace where the data variance is maximized. Secondly, PCA Implementation, in our study, PCA was used to reduce the dimensions of the seismic data, transforming the original features into a new set of linearly uncorrelated components (principal components). This reduced dataset retains most of the original data's variability, making it easier to visualize and interpret. The PCA transformation can be mathematically represented in the equation (5).

$$X_{\text{PCA}} = XW \tag{5}$$

Here, X is the original data matrix, W is the matrix of principal components, and X_{PCA} is the transformed data in the PCA subspace. The columns of W are the eigenvectors of the covariance matrix of X, ordered by their corresponding eigenvalues in descending order.

Conversely, t-SNE is a non-linear technique used for dimensionality reduction and is particularly well-suited for the visualization of high-dimensional datasets. It works by converting similarities between data points into joint probabilities and then minimizing the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. Unlike PCA, t-SNE does not rely on linear projections. Instead, it maintains the local structure of the data, making it effective in unfolding complex manifolds and revealing hidden structures within the data. Furthermore, in our seismic data analysis, t-SNE was employed to provide a more nuanced, non-linear dimensionality reduction. It was particularly useful in visualizing clusters and anomalies, as it preserves the local relationships between data points. The mathematical representation of the t-SNE process involves two key equations. The similarity of datapoint x_j to datapoint x_i, in the original high-dimensional space is represented as a conditional probability. We defined the method in equation (6) and (7).

$$p_{(j|i)} = \frac{\exp\left(-\left|\left|x_i - x_j\right|\right|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\left|\left|x_i - x_k\right|\right|^2 / 2\sigma_i^2\right)} \tag{6}$$

$$q_{ij} = \frac{\left(1 + \left|\left|y_i - y_j\right|\right|^2\right)^{-1}}{\sum_{k \neq l}\left(1 + \left|\left|y_k - y_l\right|\right|^2\right)^{-1}} \tag{7}$$

Both PCA and t-SNE played pivotal roles in visualizing our seismic data. PCA provided a broad overview of the data structure, highlighting the directions of maximal variance. In contrast, t-SNE allowed us to delve deeper into the local structures of the data, revealing clusters and anomalies that were not immediately apparent. These visualization techniques were instrumental in offering intuitive insights into the spatial and temporal patterns of the seismic events, thus significantly enhancing our understanding of the data's underlying structure.

EVALUATION METRICS

To assess the effectiveness of our clustering algorithms, we employed two key metrics: the Silhouette Score and the Davies-Bouldin Score. The Silhouette Score measures the similarity of an object to its own cluster compared to other clusters, providing an indication of the appropriateness of the cluster assignments. The Davies-Bouldin Score evaluates the average similarity between each cluster and the cluster most like it, with lower scores indicating better clustering. We describe the equation in (8)-(10).

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{8}$$

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 3, **Number** 2, November 2023
https://doi.org/10.47709/brilliance.v3i2.3501

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} R\,(i,j) \tag{9}$$

$$R(i,j) = \frac{s_i + s_j}{d_{ij}} \tag{10}$$

RESULT

In this section, the researcher will explain the results of the research obtained.

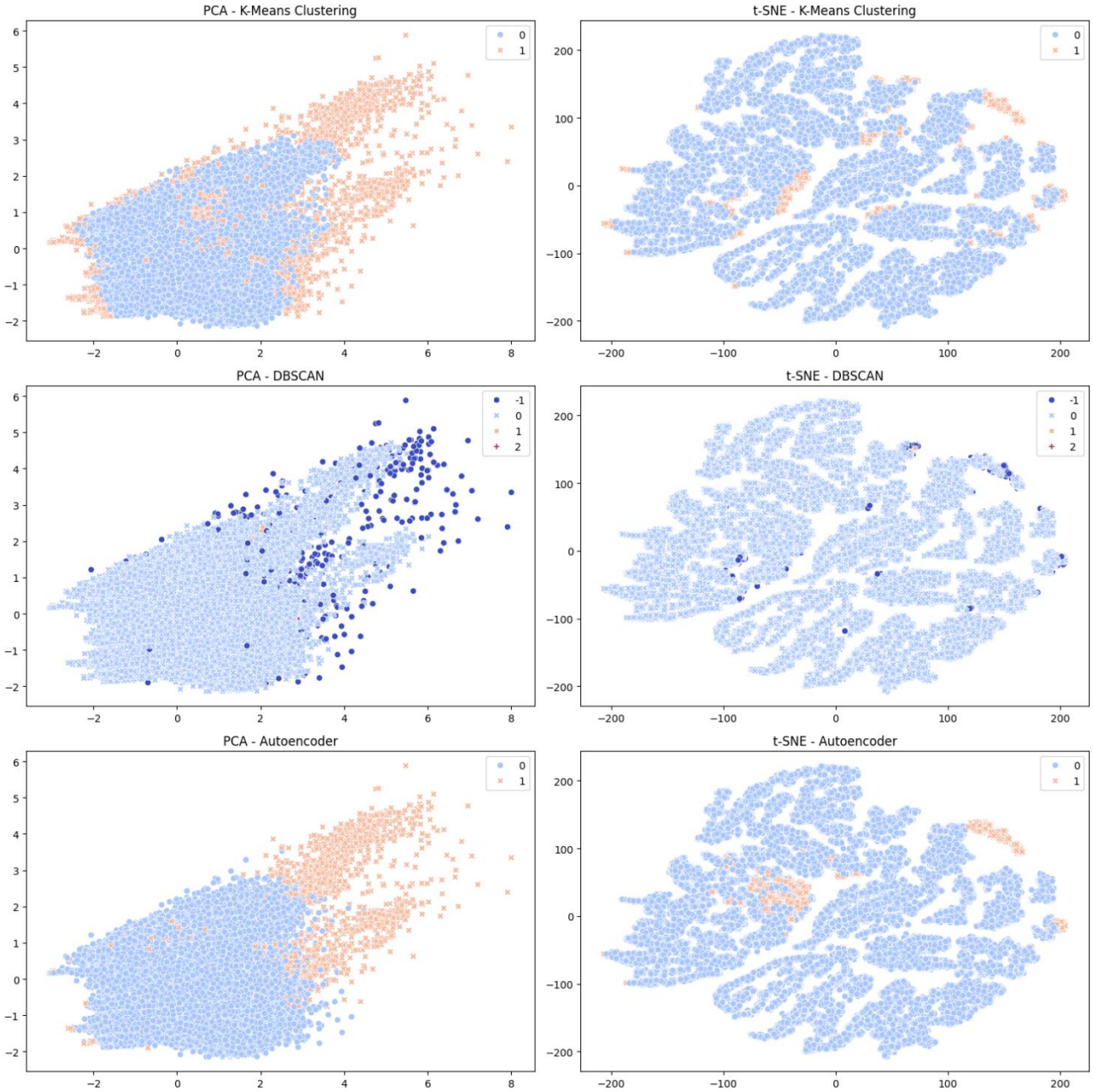


Figure 1. The Anomaly Detection Results

Table 1. The Performance Results of Experiment Method

Methods	Silhouette Score	Davies-Bouldin Score
K-Means	0.38205692782106854	2.2971150438728163
DBSCAN	0.2461437880176781	2.0807131484559562
AutoEncoder	0.31947468300358045	2.364885903813628

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 3, **Number** 2, November 2023
https://doi.org/10.47709/brilliance.v3i2.3501

The provided results as presented in the figure 1 and table 1 showcase the application of three different clustering techniques on a seismic dataset: K-Means, DBSCAN, and an Autoencoder-based approach. Each method was evaluated using two metrics: the Silhouette Score and the Davies-Bouldin Score. These scores, along with the visualizations generated by PCA (Principal Component Analysis) and t-SNE (t-Distributed Stochastic Neighbor Embedding), offer comprehensive insights into the performance and characteristics of each clustering method. The K-Means algorithm, which partitions the data into a predefined number of clusters, achieved a Silhouette Score of approximately 0.382. This score suggests a reasonable structure where data points are, on average, closer to their own cluster center than to centers of other clusters. However, the score is not close to 1, indicating there is room for improvement in cluster definition. The Davies-Bouldin Score for K-Means is approximately 2.297, which is the highest among the three methods tested. Since lower scores are more desirable, indicating better separation between clusters, this score suggests that the clusters formed by K-Means are not as compact and well-separated as we might prefer.

Visualizations via PCA and t-SNE further elucidate the clustering pattern. PCA shows distinct but somewhat overlapping clusters, while t-SNE reveals a more complex structure, suggesting that seismic events are not entirely linearly separable. The orange and blue clusters appear somewhat intermixed in the t-SNE plot, which could explain the moderate Silhouette Score. In addition, DBSCAN, a density-based clustering method, obtained a Silhouette Score of approximately 0.246. This lower score indicates that the clusters defined by DBSCAN are less distinct compared to K-Means. The data points are not as tightly grouped within their own cluster or as well-separated from other clusters. The Davies-Bouldin Score for DBSCAN is approximately 2.081, which, while lower than the score for K-Means, still indicates that the clusters are not as distinct as they could be.

The PCA visualization for DBSCAN shows a dense core cluster with several outliers, which is characteristic of the density-based nature of DBSCAN. On the other hand, the t-SNE visualization reveals that DBSCAN is able to capture a more nuanced structure within the data, with a clear separation of dense clusters and scattered outlier points. The presence of outliers, as indicated by the DBSCAN method, is congruent with its lower Silhouette Score, reflecting the noise and less-defined cluster boundaries. The Autoencoder approach, utilizing a neural network to reconstruct data for anomaly detection, achieved a Silhouette Score of approximately 0.319. This score is higher than that of DBSCAN but lower than K-Means, suggesting that while the Autoencoder clusters are more defined than those from DBSCAN, they are not as cohesive as K-Means clusters. The Davies-Bouldin Score of approximately 2.365 is similar to K-Means, indicating similar challenges in achieving clear separation between clusters.

In the PCA visualization, the Autoencoder method shows a clear distinction between normal data points and anomalies. The t-SNE visualization further emphasizes the capacity of the Autoencoder to distinguish between typical and atypical seismic patterns, although some overlap is present. Across all methods, the visualizations and scores collectively suggest that while distinct clustering and anomaly patterns exist within the seismic data, there is complexity that challenges the definition of well-separated clusters. This complexity could be due to inherent noise in the data, the presence of sub-clusters, or overlapping distributions of seismic event features. The moderate scores from the K-Means and Autoencoder methods suggest that some seismic events do not conform to clear-cut clusters. This could reflect the reality of seismic data, where events do not always occur in neatly defined groups. The lower scores for DBSCAN emphasize its strength in identifying outliers but also its limitation in defining clusters with high separation. The combination of Silhouette and Davies-Bouldin scores, along with PCA and t-SNE visualizations, provides a multi-faceted view of the clustering methods' performance. K-Means offers a more cohesive clustering structure, while DBSCAN excels in detecting outliers. The Autoencoder finds a middle ground, identifying anomalies while maintaining moderate cluster cohesion. These results underscore the complexities of seismic data and highlight the importance of leveraging multiple methods and metrics to gain the most comprehensive understanding of such data. Each method brings its own strengths and reveals different aspects of the data, emphasizing the multifaceted nature of seismic event analysis.

DISCUSSION

The exploration of seismic data through various clustering techniques has illuminated the multifaceted nature of seismic events and the complexities involved in their analysis. In our study, we employed K-Means clustering, DBSCAN, and an Autoencoder-based approach to partition the data and detect anomalies. Each method was critically evaluated using the Silhouette Score and the Davies-Bouldin Score, alongside visualization techniques like PCA and t-SNE. This section discusses the implications of the findings, the strengths and limitations of the methods, and the insights they provide into seismic patterns. The moderate Silhouette Score achieved by K-Means clustering suggests a fair level of distinction among the clusters. K-Means' objective to minimize within-cluster variance generally results in spherical-shaped clusters, which can be limiting if the natural clusters in the data have non-spherical shapes. The relatively high Davies-Bouldin Score indicates less separation between clusters, a possible indication that the assumption of spherical clusters might not hold for seismic data. The visualizations support this, as the PCA and t-SNE plots show overlapping clusters, suggesting that the underlying distribution of seismic events may not be ideally suited for K-Means without further refinement of the feature space or preprocessing.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 3, **Number** 2, November 2023
https://doi.org/10.47709/brilliance.v3i2.3501

DBSCAN's lower Silhouette Score indicates that the clusters identified are less compact and more interspersed. This outcome could be due to DBSCAN's sensitivity to the density parameter settings, which might require fine-tuning to match the data distribution of seismic events accurately. However, DBSCAN's ability to identify outliers is clearly demonstrated in the visualizations, especially in the t-SNE plots, where isolated points are scattered outside the dense clusters. These outliers could represent significant seismic events that differ from common patterns and may warrant further investigation. The Autoencoder, with its neural network architecture, provided a balanced approach to anomaly detection. While its Silhouette Score was not as high as that of K-Means, the Autoencoder identified a distinct grouping of anomalies. This is particularly visible in the PCA plot, where two clearly differentiated clusters are apparent. The Autoencoder's non-linear nature allows it to capture complex patterns within the data, which linear methods like PCA cannot. However, the Davies-Bouldin Score indicates that the separation between the normal data points and anomalies is not as distinct, suggesting that some seismic events may exhibit characteristics that blur the lines between typical and atypical activity.

The intricate nature of seismic data, influenced by a multitude of factors, results in complex patterns that are challenging to categorize into distinct clusters, and this complexity is compounded by the presence of noise and outliers. Our analysis employed various methods, each illuminating different facets of the data: K-Means provided a clear-cut partitioning of data into groups based on feature similarity; DBSCAN shed light on the density distribution of events, pinpointing potential anomalies; and the Autoencoder's deep learning capabilities facilitated a nuanced understanding of the data, capturing complex and non-linear relationships. The findings highlight the importance of utilizing a combination of clustering techniques and evaluation metrics to gain a comprehensive understanding of seismic data. K-Means is advantageous for initial, broad pattern recognition, DBSCAN excels in a more detailed analysis that focuses on anomaly detection and spatial distribution, and Autoencoders are adept at unraveling the underlying data structure and discerning subtle patterns. Additionally, visualization tools like PCA and t-SNE have proven instrumental in interpreting the high-dimensional seismic data—PCA offers a macroscopic overview of the data's structure, while t-SNE provides an intricate examination of local patterns, enhancing our understanding of the clustering executed by DBSCAN and the Autoencoder.

CONCLUSION

The conclusion of our study reflects upon the application of multiple clustering techniques to seismic data analysis, revealing the nuanced and multifaceted nature of seismic event patterns. Through the implementation of K-Means, DBSCAN, and Autoencoder algorithms, complemented by dimensionality reduction and visualization through PCA and t-SNE, we have gained valuable insights into the clustering and anomaly detection within seismic datasets. K-Means clustering demonstrated its utility in providing an initial structure to the data, segmenting seismic events into broad, discernible groups. However, its performance, as quantified by the Silhouette and Davies-Bouldin scores, suggests that there might be a more complex interplay of features that requires a nuanced approach. DBSCAN's strength in identifying outliers enriched our analysis by highlighting seismic events that deviated from the common patterns, potentially flagging critical incidents that warrant further examination. The Autoencoder, with its deep learning prowess, excelled in capturing the intricate structures within the data, suggesting that there are deep-seated patterns within seismic activities that linear methods may not fully uncover.

The comparative use of PCA and t-SNE visualizations allowed us to see beyond the limitations of individual clustering techniques, presenting a holistic view of the seismic data structure. These visualization techniques were indispensable in corroborating the clustering analysis, providing a more comprehensive understanding of the seismic events' distribution. Our research underscores the critical role of leveraging a suite of analytical methods to interpret complex datasets effectively. The insights gained through this study have significant implications for the field of seismology, offering a pathway to more accurate predictions and a better understanding of seismic activities. Future research should aim to build upon these findings, exploring the integration of additional features into the analysis and applying these techniques to larger, more diverse seismic datasets to validate and refine the models further. Furthermore, the development of hybrid models that combine the strengths of each clustering technique could prove invaluable in advancing the field.

REFERENCES

AlAli, A., & Anifowose, F. (2022). Seismic velocity modeling in the digital transformation era: a review of the role of machine learning. *Journal of Petroleum Exploration and Production Technology*, 1–14.

Aranda, A. M., Sele, K., Etchanchu, H., Guyt, J. Y., & Vaara, E. (2021). From big data to rich theory: Integrating critical discourse analysis with structural topic modeling. *European Management Review*, *18*(3), 197–214.

Asming, V. E., Asming, S. V, Fedorov, A. V, Yevtyugina, Z. A., Chigerev, Y. N., & Kremenetskaya, E. O. (2022). System for automatic recognition of types of sources of regional seismic events. *Seismic Instruments*, *58*(5), 509–520.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 3, **Number** 2, November 2023
https://doi.org/10.47709/brilliance.v3i2.3501

Dramsch, J. S. (2020). 70 years of machine learning in geoscience in review. *Advances in Geophysics*, *61*, 1–55.

Gerstenberger, M. C., Marzocchi, W., Allen, T., Pagani, M., Adams, J., Danciu, L., … others. (2020). Probabilistic seismic hazard analysis at regional and national scales: State of the art and future challenges. *Reviews of Geophysics*, *58*(2), e2019RG000653.

Ghasemi, A., & Stephens, M. T. (2022). Building clustering for regional seismic response and damage analysis. *Earthquake Spectra*, *38*(4), 2941–2969.

Ji, K., Wen, R., Ren, Y., & Dhakal, Y. P. (2020). Nonlinear seismic site response classification using K-means clustering algorithm: Case study of the September 6, 2018 Mw6. 6 Hokkaido Iburi-Tobu earthquake, Japan. *Soil Dynamics and Earthquake Engineering*, *128*, 105907.

Jiao, P., & Alavi, A. H. (2020). Artificial intelligence in seismology: advent, performance and future trends. *Geoscience Frontiers*, *11*(3), 739–744.

Johnson, C. W., Ben-Zion, Y., Meng, H., & Vernon, F. (2020). Identifying different classes of seismic noise signals using unsupervised learning. *Geophysical Research Letters*, *47*(15), e2020GL088353.

Liu, W., & Hu, W. (n.d.). Oversampling of Tabular Data for Imbalanced Learning Via Denoising Diffusion Probabilistic Models. *Available at SSRN 4673719*.

Liu, X., Li, B., Li, J., Chen, X., Li, Q., & Chen, Y. (2021). Semi-supervised deep autoencoder for seismic facies classification. *Geophysical Prospecting*, *69*(6), 1295–1315.

Ma, Z., & Mei, G. (2021). Deep learning for geological hazards analysis: Data, models, applications, and opportunities. *Earth-Science Reviews*, *223*, 103858.

Mousavi, S. M., & Beroza, G. C. (2022). Deep-learning seismology. *Science*, *377*(6607), eabm4470.

Mousavi, S. M., & Beroza, G. C. (2023). Machine Learning in Earthquake Seismology. *Annual Review of Earth and Planetary Sciences*, *51*, 105–129.

Noureldin, M., Ali, A., Sim, S., & Kim, J. (2022). A machine learning procedure for seismic qualitative assessment and design of structures considering safety and serviceability. *Journal of Building Engineering*, *50*, 104190.

Posamentier, H. W., Paumard, V., & Lang, S. C. (2022). Principles of seismic stratigraphy and seismic geomorphology I: Extracting geologic insights from seismic data. *Earth-Science Reviews*, *228*, 103963.

Ros, F., Riad, R., & Guillaume, S. (2023). PDBI: A partitioning Davies-Bouldin index for clustering evaluation. *Neurocomputing*, *528*, 178–199.

Shutaywi, M., & Kachouie, N. N. (2021). Silhouette analysis for performance evaluation in machine learning with applications to clustering. *Entropy*, *23*(6), 759.

Soh, H., & Demiris, Y. (2015). Spatio-temporal learning with the online finite and infinite echo-state Gaussian processes. *IEEE Transactions on Neural Networks and Learning Systems*, *26*(3), 522–536. https://doi.org/10.1109/TNNLS.2014.2316291

Sun, H., Burton, H. V, & Huang, H. (2021). Machine learning applications for building structural design and performance assessment: State-of-the-art review. *Journal of Building Engineering*, *33*, 101816.

Xu, G., & Haq, B. U. (2022). Seismic facies analysis: Past, present and future. *Earth-Science Reviews*, *224*, 103876.

Zhao, J., Han, X., Ouyang, M., & Burke, A. F. (2023). Specialized deep neural networks for battery health prognostics: Opportunities and challenges. *Journal of Energy Chemistry*.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.