Evaluation of Lempel-Ziv Complexity for Brain Region Segmentation Using fMRI Data

Sijie Li[1], Yushan Wei[2] Xintong Shen[3], Yidan Yao[4],

June 2, 2024

[1] 12212859, 12212859@mail.sustech.edu.cn to Sijie Li.
[2] 12212860, 12212860@mail.sustech.edu.cn to Yushan Wei.
[3] 12210754, 12210754@mail.sustech.edu.cn to Xintong Shen.
[4] 12210163, 12210163@mail.sustech.edu.cn to Yidan Yao.

Abstract

Functional Magnetic Resonance Imaging (fMRI) has become a pivotal tool in neuroscience, providing insights into the functional connectivity and activity patterns of the human brain. Traditional approaches to brain region segmentation using fMRI data often rely on features derived from functional connectivity, requiring large datasets and significant computational resources. In contrast, Lempel-Ziv Complexity (LZC) has shown promising bio-interpretability in other applications, but its use in fMRI data analysis remains underexplored. This study aims to fill this gap by employing LZC for brain region segmentation. By utilizing both hierarchical clustering and K-means clustering, we achieve accurate and efficient brain region segmentation with minimal data and computation. Our results demonstrate that LZC-based features significantly outperform traditional time series data in clustering performance, as validated by metrics such as correlation, homogeneity, completeness, silhouette score, and neural network prediction accuracy. This study highlights the potential of LZC as a robust and interpretable feature for brain region segmentation, providing an effective approach that is both computationally efficient and highly accurate.

1 Introduction

The human brain's complexity is reflected in its intricate network of functional connectivity, which can be measured and analyzed through functional Magnetic Resonance Imaging (fMRI). Identifying distinct brain regions based on their activity patterns is crucial for understanding cognitive processes and diagnosing neurological disorders. Traditional approaches often base on the connectivity of brain region evaluating by time series data, which may not fully capture the underlying dynamics patterns. Our motivation is to explore whether analyzing alternative data features – complexity can provide a more accurate atlas. The goal of this study is to figure out the better data features to obtain the atlas.

Numerous studies have explored machine learning techniques for brain region segmentation using fMRI data. Traditional methods such as Support Vector Machines (SVM) [1] and Random Forest (RF) have achieved high accuracy, often exceeding 90% in some cases. However, these methods typically require large, well-labeled datasets and can struggle with the high dimensionality and variability inherent in fMRI data. Deep learning approaches, including Convolutional Neural Networks (CNN), [2, 3, 4]have also shown promise but require significant computational resources and extensive training data.

In this work, we focus on extracting meaningful features from fMRI data using Lempel-Ziv Complexity (LZC) and applying both hierarchical clustering and KMeans clustering to obtain the atlas. Our approach aims to address the limitations of previous methods by leveraging feature extraction to reduce dimensionality and improve clustering performance.

Our contributions are summarized as follows:

- Despite its strong interpretability, LZC has not been extensively utilized in previous studies for fMRI data analysis. We fill the significant gap by employing LZC for clustering fMRI data.
- By leveraging LZC, we demonstrate that accurate brain region segmentation can be achieved using simple computational methods, a small amount of data, and basic clustering techniques, highlighting the efficiency and effectiveness of our approach.
- Comparative analysis with clustering results using time series data shows that our method significantly outperforms traditional approaches, as evaluated by various metrics including correlation, homogeneity, completeness, silhouette score and neural network prediction accuracy.

2 Related work

There are many methods to choose from when processing brain signals, and in existing research, we can choose complexity clustering because it can better capture the complexity, nonlinearity, and dynamic changes of data, and has strong robustness and adaptability.

The metric of complexity proposed by Lempel and Ziv complexity(LZC) to evaluate the randomness of finite sequences has been extensively used to solve information theoretic problems and applications such as coding, data compression, and generation of test signals. This complexity measure is related to the number of distinct substrings (i.e., patterns) and the rate of their occurrence along a given sequence. In recent years, LZC has been applied extensively in biomedical signal analysis as a metric to estimate the complexity of discrete-time physiologic signals. For instance, LZC has been used for emotion recognition based on EEG signals [5, 6],the etiology of Alzheimer's disease and mild cognitive impairment [7], recognition of EEG signal features in epilepsy patients [8, 9],analysis of the characteristics of motor imagery EEG signals [10] etc. They extracted complexity features through LZC, and obtained available data.

For the methods of brain region segmentation, existing literature has adopted different algorithms. An article adopts recent brain region segmentation methods based on connection patterns and uses edge weighted spectral clustering algorithm to segment the BA46 region. This article collected 20 subjects and used data from three modalities: dMRI, T1, and fMRI and they aim to divide the right BA46 area based on connection mode. When the

number of clusters is determined to be 4, the accuracy reaches 0.6909. [11] Another study used supervised clustering algorithms to divide brain regions based on function, using fMRI data from 20 subjects with an accuracy of up to 0.8000, but it rigidly required prior knowledge. [12] A survey used data from 154 Parkinson's disease patients and 109 healthy control group fMRI data to classify cerebellar regions based on functional usage spectrum clustering. When the segmentation number was 130, the average accuracy was the highest, at 0.7265. [13]However, the data needs to undergo DPABI 6.0 preprocessing to achieve spatial alignment and improve the signal-to-noise ratio of the data.

Although LZC has extensive applications in many fields, there is currently no research on its use for brain region segmentation. We use LZC for brain region segmentation here, without the need for any prior knowledge or other software to preprocess the data. Just with a little computation, small amount of single modality data, and simple clustering algorithms, the clustered result from LZC data reaches accuracy of 0.6900 and very stable results.

3 Method

We start with raw fMRI data from 50 Chinese and 50 foreign participants, each having data for two resting states and seven task states. The preprocessing involves extracting LZC features from the fMRI time series of the 50 Chinese participants' REST1 state data. To account for the temporal dynamics, we also apply sliding window techniques with various window lengths and steps, ensuring the representation of dynamic brain activity.

Next, we normalize the data to mitigate individual differences, followed by clustering using two methods: hierarchical clustering with the Ward method and KMeans clustering, both set to identify seven clusters corresponding to distinct brain regions. Finally, we evaluate the clustering results using multiple metrics, including correlation, homogeneity, completeness, and prediction accuracy with neural networks, as well as the silhouette score to evaluate cluster compactness and separation.

The framework of our work is illustrated in Figure 1.

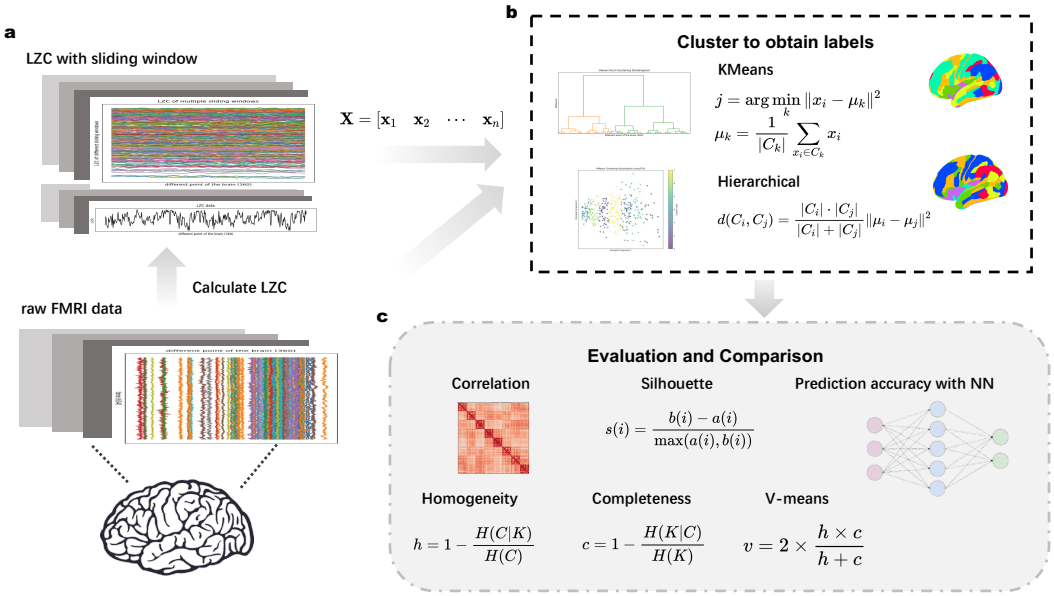


Figure 1: Framework

4 Data and Experiments

4.1 Dataset

In this study, we ran experiments on two fMRI datasets, including Human Connectome Project (HCP) [14] and Chinese HCP (CHCP) [15]. We randomly selected 50 Chinese subjects from CHCP and 50 foreign subjects from HCP. Each subject has two sessions of resting-state fMRI (i.e., rest1 and rest3), and seven sessions of task fMRI (i.e., emotion, gambling, language, motor, n-back, relation, and social tasks). The fMRI data was preprocessed by the HCP pipeline, and we extracted the averaged time courses in 360 brain regions using MMP atlas.

4.2 Experiments

4.2.1 Data Preprocessing

The preprocessing includes both feature extraction and data generalization.

Series Averaging To account for individual differences in brain activity, we average the fMRI time series data across the 50 Chinese participants. Specifically, we take the REST1 state data, resulting in a dataset of shape 364 (time points) x 360 (brain regions).

LZC Feature Extraction We compute the LZC for the entire time series, resulting in a dataset of shape 50 (participants) x 360 (brain regions). Additionally, we average these features across participants to obtain a 1x360 dataset.

Sliding Window LZC To capture dynamic changes in brain activity, we also apply sliding window techniques with various window lengths. Specifically, window length of 200, 300, 400, 500, 600 is covered, and all of the moving step is 1. These sliding window operations produce multiple datasets, each capturing different aspects of the temporal dynamics of the brain regions and thus enables further exploration.

4.2.2 Clustering

We applied two clustering methods to classify brain regions based on the extracted features: KMeans Clustering and Hierarchical Clustering. These methods are widely used, and the reasons for selecting them in this project is their capability to achieve the specified cluster number K=7, matching the 7 cognitive functional categories in our project and its simplicity but effectiveness highlighting the power the LZC features.

KMeans Clustering KMeans clustering is a widely used algorithm for partitioning large datasets into K distinct clusters by minimizing the within-cluster sum of squares (WCSS). The algorithm works as follows: it starts by randomly selecting K initial centroids. Each data point is then assigned to the nearest centroid, forming K clusters. The centroids are recalculated as the mean of all points in a cluster. These steps are repeated until the centroids stabilize. The objective function for KMeans is:

$$\arg\min_{S} \sum_{i=1}^{K} \sum_{x \in S_i} \|x - \mu_i\|^2$$

where S represents the set of clusters, and μ_i is the centroid of cluster S_i. This method is effective for organizing data into predefined clusters, such as the seven cognitive functional categories in the dataset.

Agglomerative Hierarchical Clustering Agglomerative hierarchical clustering is a bottom-up method for clustering data. It starts with each data point as an individual cluster and iteratively merges the closest clusters until all points form a single cluster. Initially, each point is its own cluster. The algorithm calculates the distance between all pairs of clusters, often using single linkage (minimum distance):

$$d(A, B) = \min_{a \in A, b \in B} \|a - b\|$$

In each step, the two closest clusters are merged, and distances are recalculated. This process continues until only one cluster remains, forming a dendrogram that illustrates the hierarchical structure of the data. This method is useful for understanding relationships between brain regions and their cognitive functions, complementing KMeans clustering results. The distance measurement we use within the agglomerative hierarchical clustering process is the Ward method:

$$d\left(C_i, C_j\right) = \frac{|C_i| \cdot |C_j|}{|C_i| + |C_j|} \left\|\mu_i - \mu_j\right\|^2$$

It specifically accounts for the number of samples within each cluster to prevent clusters with very few samples. This consideration ensures the robustness of the clustering results by avoiding over-reliance on clusters with limited samples. Hence, employing the Ward method for distance calculation enhances the quality and reliability of clustering outcomes.

4.2.3 Evaluation and Comparison

Qualitative analysis Initially, to evaluate the effectiveness of LZC data in this specific task, qualitative analysis is carried out. KMeans result is visualized by Principle Component Analysis (PCA) methods and LZC correlation matrices of different brain regions under different state are plotted.

Homogeneity Homogeneity measures whether each cluster contains only members of a single class. It quantifies the purity of the clustering result by calculating the uncertainty of class labels given the cluster assignment. The formula is:

$$\text{Homogeneity} = 1 - \frac{H(C|K)}{H(C)}$$

where $H(C|K)$ is the conditional entropy of the class distribution given the cluster assignments, and $H(C)$ is the entropy of the class distribution.

Completeness Completeness measures whether all members of a given class are assigned to the same cluster. It assesses the extent to which samples of a class are assigned to a single cluster by calculating the uncertainty of the cluster assignments given the class labels. The formula is:

$$\text{Completeness} = 1 - \frac{H(K|C)}{H(K)}$$

V-Measure V-Measure is the harmonic mean of homogeneity and completeness, providing a balanced evaluation of clustering performance. It ensures that both the purity and the completeness of the clustering are considered. The formula is:

$$\text{V-Measure} = 2 \times \frac{\text{Homogeneity} \times \text{Completeness}}{\text{Homogeneity} + \text{Completeness}}$$

This metric provides a comprehensive measure to evaluate the clustering algorithm, balancing between homogeneity and completeness.

Silhouette score Silhouette score is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette score ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

a(i) is the average distance from i to all other points in the same cluster. b(i) is the smallest average distance from i to points in a different cluster, minimized over clusters.

Prediction accuracy To assess the stability and reliability of our clustering results, we attempted to predict whether a segment of brain fMRI data belonged to Chinese or foreign individuals based on our clustering outcomes using neural networks. To achieve this, we constructed neural networks and partitioned 50 fMRI datasets from Chinese individuals and 50 from foreign individuals into training and testing sets. Utilizing the mapping obtained from clustering, we divide the fMRI data into seven brain regions and assessed complexity, then trained and tested the neural networks accordingly. For each clustering method, we developed a separate neural network, and higher accuracy in the testing results of the neural network indicated better clustering performance.

The neural network contains 6 layers. One convolutional layer with a one-dimensional convolution operation. This layer has 64 filters, a kernel size of 3, and uses the ReLU activation function. One max-pooling layer used to reduce the dimensionality of the output from the convolutional layer and extract important features. One flatten layer flattens the output from the convolutional layer into a one-dimensional array. One fully connected layer with 128 neurons and ReLU activation function. One dropout layer with a dropout rate of 0.5. And one Dense layer with 2 neurons and softmax activation function

5 Results

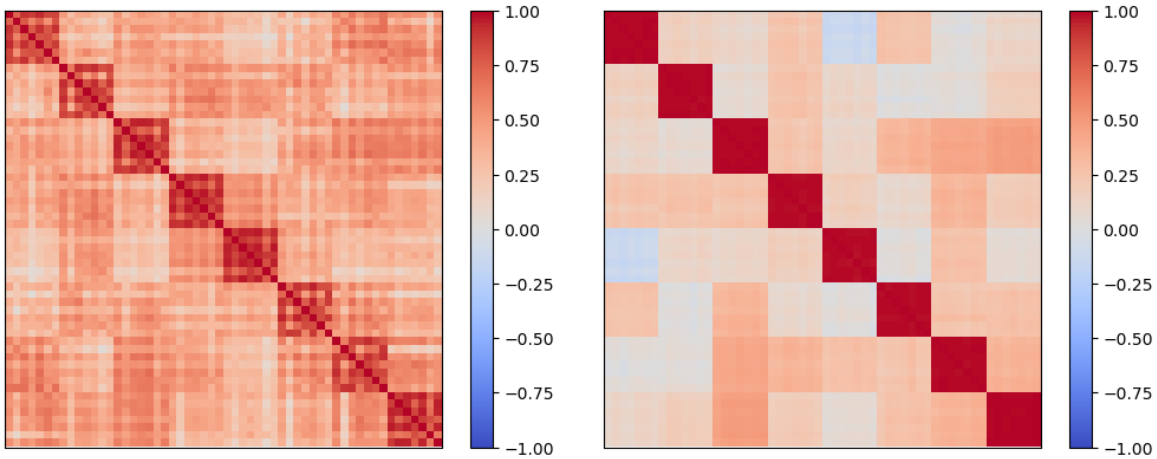
5.1 Qualitative analysis

To initially validate the effectiveness of the LZC feature for clustering, we selected a small batch of data and visualized the clustering results. We employed KMeans clustering and visualized the results by reducing the dimensionality of the feature vectors using Principal Component Analysis (PCA). The clustering outcome, depicted in Figure 2, illustrates the distinct separation under the principal component achieved by the LZC features.

Figure 2: Visualization of KMeans result

Additionally, we assessed the independence of different brain regions under various task

states by plotting correlation matrices. These matrices compare the labels obtained through LZC-based clustering with randomly generated labels. Each value in the correlation matrix represents the relationship between specific brain regions under different task states (a total of eight task states). Lighter colors indicate higher independence. The comparison, presented in Figure 3, demonstrates the efficacy of LZC features in distinguishing brain regions, as evidenced by the presence of more lighter blocks in the central sub-blocks, indicating higher independence of different regions within the same state.



(a) LZC (b) Random labels

Figure 3: Correlation matrix

Also, from the correlation matrix (Figure 3a), lighter sub-blocks are clearly observed, indicating significant differences in brain complexity under different states.

Having validated the efficacy of the LZC feature, we then further explored its advantages by comparing the clustering results with those obtained from the original time series data and several methods are used to quantify the evaluation.

5.2 Homogeneity and completeness

The following tables present the homogeneity, completeness, and V-measure of different brain region divisions obtained by varying the seed of K-means clustering and by comparing clustering methods (hierarchical clustering vs. K-means clustering).

	Time series	LZC_0	LZC_200	LZC_300	LZC_400	LZC_500	LZC_600
Homogeneity	0.8624	1	0.888	0.8978	0.8617	0.8081	0.8485
Completeness	0.8619	1	0.8923	0.9100	0.8661	0.7927	0.8437
V-Measure	0.8622	1	0.8905	0.9039	0.8639	0.8003	0.8461

Table 1: Results of varying K-means clustering seed

Explanation of the table headers: Time series refers to clustering based on the original time series data averaged over 50 participants. LZC indicates the results obtained from

data with Lempel-Ziv Complexity (LZC) features extracted, with the number following LZC representing the window length used for feature extraction. This explanation applies to all the tables presented.

Table 1 demonstrates that data with LZC features generally yield better clustering results, as evidenced by higher homogeneity, completeness, and V-measure scores across different clustering seeds. This suggests that LZC-based features provide more consistent and robust clusters compared to raw time series data.

	Time series	LZC_0	LZC_200	LZC_300	LZC_400	LZC_500	LZC_600
Homogeneity	0.0830	0.8947	0.7775	0.7533	0.7397	0.6687	0.7784
Completeness	0.0701	0.8892	0.7933	0.7794	0.7661	0.6976	0.7966
V-Measure	0.0760	0.8919	0.7853	0.7661	0.7527	0.6828	0.7874

Table 2: Results of varying clustering methods

Table 2 shows that regardless of whether hierarchical clustering or K-means clustering is employed, the clustering results based on LZC features consistently exhibit significantly higher similarity metrics than those based on time series data.

Higher similarity scores indicate that the data used for clustering are more inherently grouped and distinguishable. Therefore, the results suggest that LZC-based data provide a more distinct and separable representation of brain activity compared to time series data, leading to more effective clustering outcomes.

5.3 Silhouette score

We employed the silhouette coefficient to evaluate the differences in clustering performance between the LZC-based method and the time-based method. To enhance the silhouette coefficient of the LZC-based method, we utilized a sliding window approach to capture more dynamic information. The results are presented in the Table 3 below.

	Time series	LZC_0	LZC_200	LZC_300	LZC_400	LZC_500	LZC_600
KMeans	0.5150	0.5753	0.4697	0.4977	0.5116	0.5284	0.5443
Hierarchical	-0.2462	0.5609	0.4406	0.4727	0.4820	0.5184	0.5087

Table 3: Result of Silhouette Score

From the Table, it can be observed that the effectiveness of time-series clustering is significantly inferior to that of LZC clustering. When using the Hierarchical clustering method, negative values even appear, indicating that a substantial number of data points were incorrectly devided.

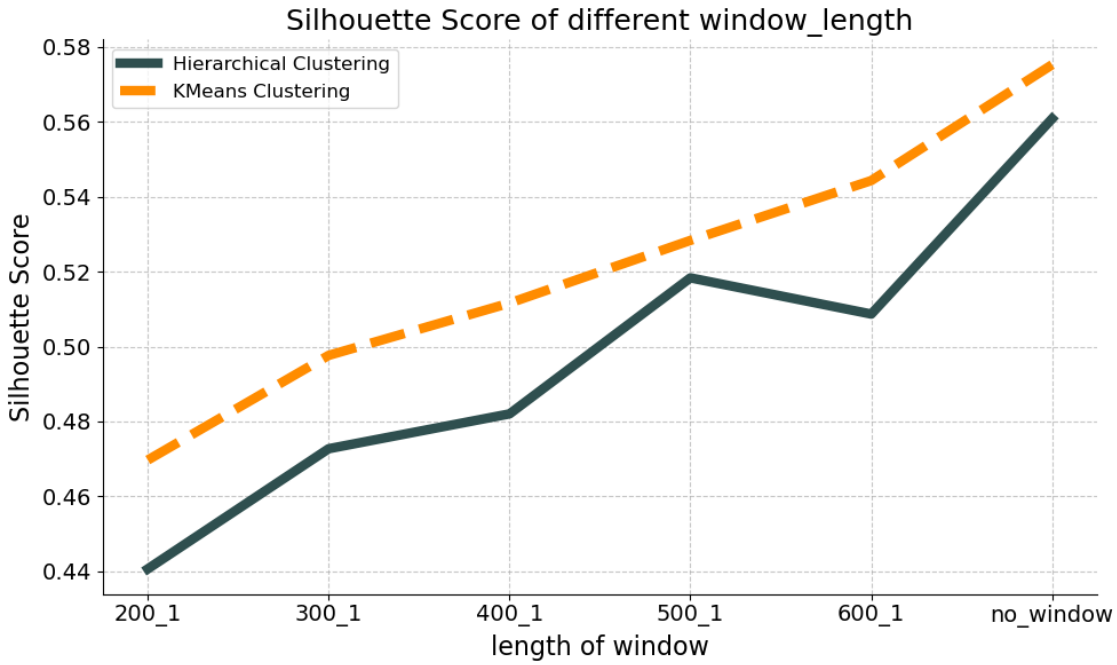


Figure 4: Result of Silhouette Score

From the Figure 4, it can be observed that the silhouette coefficient exhibits an upward trend as the window length increases, which aligns to some extent with our expectations. Interestingly, when the windowing is entirely removed, resulting in an overall LZC measure, the silhouette coefficient reaches its maximum. This might indicate that reducing dynamic information to some extent may enhance the clustering performance.

5.4 Prediction with neuron network

We trained and tested four neural networks using a dataset comprising fMRI data from 50 Chinese individuals and 50 foreign individuals. The networks were trained and tested on training and testing sets delineated using KMeans and Hierarchical clustering methods for LZC and temporal clustering, respectively. The accuracy results of predictions on the testing set are presented in the table 4 below.

	Time series	LZC_0
KMeans	0.5722	0.6999
Hierarchical	0.5833	0.6278

Table 4: Prediction accuracy of neuron network

From the results in the table, it can be observed that the prediction accuracy is higher for clustering based on LZC compared to clustering based on time series. Additionally, the accuracy obtained using the KMeans clustering method is higher than that obtained using the Hierarchical clustering method.

6 Conclusion

In this study, we introduced an effective approach for brain region segmentation by leveraging Lempel-Ziv Complexity (LZC) as a feature extraction method on fMRI data. Our methodology addressed the limitations of traditional time series-based clustering by capturing the dynamic nature of brain activity. Through extensive experiments involving hierarchical clustering and K-means clustering, we demonstrated that LZC-based features provide significantly higher accuracy and stability in brain region segmentation.

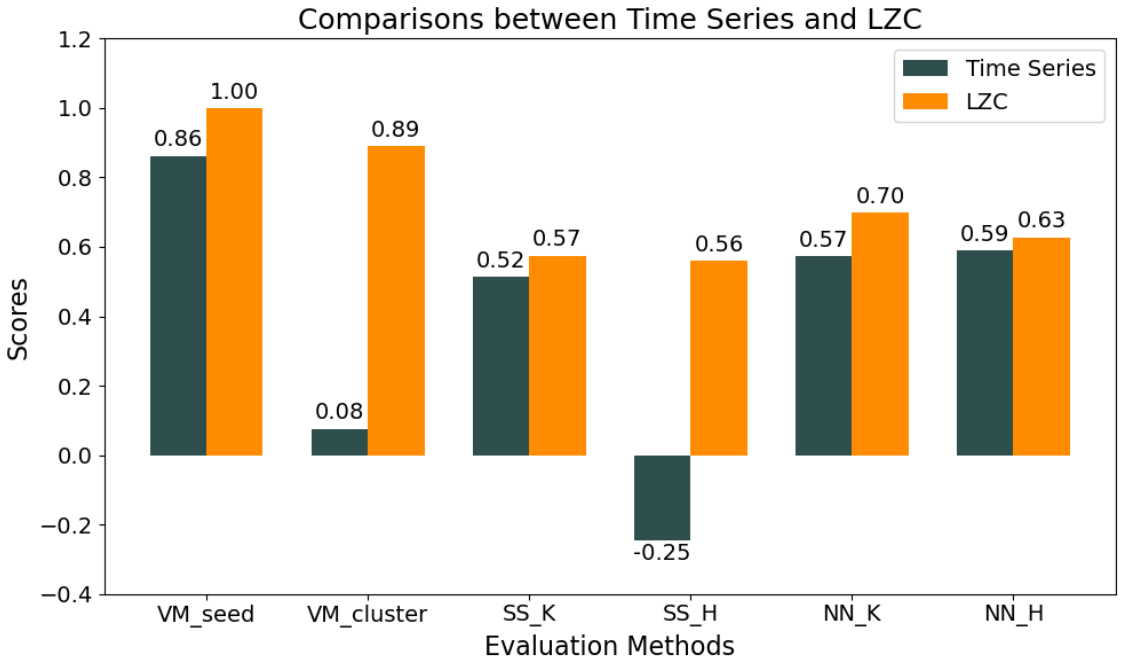


Figure 5: Comparisons between Time Series and LZC

Abbreviation: VM_seed, V-measure score obtained by varying the seed of the K-means clustering algorithm; VM_cluster, V-measure score obtained by comparing hierarchical clustering and K-means clustering methods; SS_K, Silhouette score obtained using the K-means clustering method; SS_H, Silhouette score obtained using the hierarchical clustering method; NN_K, Prediction accuracy of the neural network using clusters obtained from the K-means clustering method; NN_H, Prediction accuracy of the neural network using clusters obtained from the hierarchical clustering method;

The experimental results Figure 5 showed that clustering outcomes using LZC features consistently outperformed those based on time series data, as validated by various metrics such as correlation, homogeneity, completeness, silhouette score, and neural network prediction accuracy. Specifically, the LZC features enabled us to achieve more distinct and separable representations of brain activity, leading to more effective clustering results. Our approach proved to be computationally efficient, requiring only a small amount of data and simple clustering algorithms to achieve accurate brain region segmentation.

In conclusion, our findings suggest that LZC is a highly effective feature for brain region segmentation, offering a robust and interpretable alternative to traditional time series analysis. Future work could explore the integration of LZC with other advanced machine learning tech-

niques to further enhance brain region segmentation and understanding of brain connectivity.

Author Contribution Sijie Li posted the idea, process data, drew the pipline, and took part in the report and poster writing; Yushan Wei wrote the neural network and evaluate the clustering effect; Xintong Shen carried out the clustering process and the evaluation of the clustering effect; Yidan Yao carried out the clustering process and searched relevant literature and collect relevant research.

References

[1] Lao Huan. Research on alzheimer's disease classification based on brain imaging features, 2022.

[2] Wang Lingdu. Research and development of brain region segmentation method and system based on neural networks, 2022.

[3] Zhu Zhimin. Research on hippocampus segmentation based on deep learning, 2023.

[4] Wang Chenyu. Research on brain tumor and region segmentation based on deep learning, 2023.

[5] Zhang Dong, You Ya Chen Dongwei, and Li Haifang. Analysis of emotional eeg signal features based on adaptive lempel ziv complexity. *Computer Applications and Software*, (9):162–165, 1 2014.

[6] Chen Dongwei and Chen Junjie. Research on lempel ziv complexity of eeg signals in emotional recognition. *Journal of Taiyuan University of Technology*, 45(6):758–763, 1 2014.

[7] Zhu Benju. Eeg study on alzheimer's disease, mild cognitive dysfunction, and normal elderly people based on complexity analysis, 2012.

[8] Mou Xueli. Research on eeg signal characteristics of epilepsy patients based on complexity and power spectrum analysis, 2022.

[9] Xia Deling, Niu Hegong Meng Qingfang, Wei Yingda, and Liu Haihong. Detection method for epileptic eeg signals based on lempel ziv complexity and empirical mode decomposition. *Computational Physics*, (6):709–714, 1 2015.

[10] Luo Zhizeng and Cao Ming. Analysis of motion imagination eeg signal characteristics based on multiscale lempel ziv complexity. *Journal of Sensing Technology*, 24(7):1033–1037, 1 2011.

[11] Lv Xiuxiu. Constructing a brain map of the right ba46 region based on connectivity information, 2015.

[12] Song Dandan. Brain function image segmentation by fusing prior information. *Electronic Technology*, 26(9):4–6,9, 1 2013.

[13] Zhou Weiqi. cerebellar partitioning based on spectral clustering and its application in feature extraction of parkinson's brain networks, 2023.

[14] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

[15] Guoyuan Yang, Jelena Bozek, Stephanie Noble, Meizhen Han, Xinyu Wu, Mufan Xue, Jujiao Kang, Tianye Jia, Jilian Fu, Jianqiao Ge, et al. Global diversity in individualized cortical network topography. *Cerebral cortex*, 33(11):6803–6817, 2023.