Andrew Vattuone
Highest NMI Score: 0.3409
Leaderboard Ranking: 2nd
May 29, 2025

Program 3 Report

In this project, we explored various clustering techniques to group high-dimensional time-series samples based on underlying structure. Using a dataset of 500 samples for each of 23 subjects performing 5 different activities (for a total of 11,500 samples), we needed to cluster each point into 115 different clusters labeled 1-115, using any form of preprocessing, postprocessing, and clustering techniques we chose. Throughout the project, I changed the various techniques I used for preprocessing and clustering numerous times before eventually arriving at my final iteration of a traditional KMeans clustering algorithm using features extracted from autocoder embeddings, MiniRocket, and equal width partitioning with mean and standard deviation, all reduced using t-SNE. This combination of features led to the best outcome with an NMI score of 0.3409 , although it took numerous iterations to achieve this.

To evaluate the quality of my clustering results, I used several internal and external metrics, including the Silhouette Score, Calinski-Harabasz Index, Davies-Bouldin Index, and Normalized Mutual Information (NMI). The Silhouette Score measures how similar each point is to its own cluster compared to other clusters, with higher values indicating better-defined clusters. The Calinski-Harabasz Index evaluates the ratio of between-cluster dispersion to within-cluster dispersion, where higher scores indicate better separation. The Davies-Bouldin Index assesses average similarity between each cluster and its most similar one, with lower values being better. However, these often highly varied, and they didn't perfectly correspond with higher NMI values. I found that plotting histograms of the number of points assigned to each cluster tended to be better (see Figure 6 for an example).

For my first submission, the only feature extraction I chose to do was to standardize the features, since it ensures that each feature contributes equally to distance or variance-based computations, and standardizing the samples is incorrect. I just used K-Means for this as a benchmark to see what I got for my first submission. I got an NMI of only 0.1657 for this. Later on, I decided to add in the Equal Width Partitioning with Mean and Standard Deviation from Lab 7, which increased my score to around 0.2169 and, in general, was slightly better. I noticed that decreasing my bucket size to 2 worked better than when I had a bucket size of 5, as my NMI score was slightly higher for a few submissions with this. Later on, I decided to add in unsupervised nonlinear dense embeddings using a shallow autoencoder trained with MSE reconstruction loss, as this helped to denoise inputs and capture nonlinear structure that enhances separation and compactness of clusters, in addition to performing some dimensionality reduction. I combined this with the Equal Width Partitioning features, and I got some good results. Finally, I decided to use MiniRocket, a time series feature extraction method designed to efficiently capture temporal patterns using convolutional kernels. By combining MiniRocket features with the Equal Width Partitioning and autoencoder embeddings, I was able to create a rich and diverse feature set that captured both global statistical trends and local temporal structure. This combination, when followed by dimensionality reduction techniques like

UMAP or t-SNE and clustering with KMeans, produced my highest scores—reaching an NMI of up to 0.3409 (Figure 4). This demonstrated that carefully combining different types of features and leveraging nonlinear structure led to significant improvements over my initial baseline.

While initially I used no dimensionality reduction, I knew I wanted to test out some methods since I knew it would very likely improve my clustering quality and align results with true labels. I experimented with several dimensionality reduction techniques, including PCA, UMAP, and t-SNE. Among these, PCA applied to the raw features (Equal Partition) significantly boosted internal clustering metrics, achieving the highest silhouette score (0.5108) and Calinski-Harabasz index (97920.95) while maintaining a low Davies-Bouldin score (0.5104). However, when evaluating against ground truth using Normalized Mutual Information (NMI), UMAP and t-SNE applied to a combined feature set (MiniRocket, Equal Partition, and embeddings) outperformed PCA-based methods, with t-SNE achieving the highest NMI of 0.3409. This trade-off revealed that UMAP and t-SNE are better suited for preserving class structure, while PCA better preserves internal compactness. As a result, I decided to use t-SNE since it gave me the best NMI, as even though it's primarily used for visualization, it happened to work quite well in this case.

Multiple clustering algorithms were tested to identify the best fit for my transformed feature space, including KMeans, Agglomerative Clustering, and Spectral Clustering. The performance of Agglomerative Clustering was comparatively poor across all metrics, with the lowest silhouette score (0.1735) and NMI (0.2169). Spectral Clustering, while showing promising results when paired with PCA (particularly on Equal Partition features), struggled when applied directly to UMAP or embeddings, showing highly variable performance and instability (e.g., a DB index of 5.54 on embeddings PCA). In contrast, KMeans consistently performed well, particularly when applied to the UMAP-transformed combined feature set, where it achieved a strong balance of internal metrics (silhouette: 0.3887, DBI: 0.7572) and the second-highest NMI (0.3291). These results confirmed that KMeans, when used with an effective feature reduction pipeline, is the most reliable clustering method for this task.

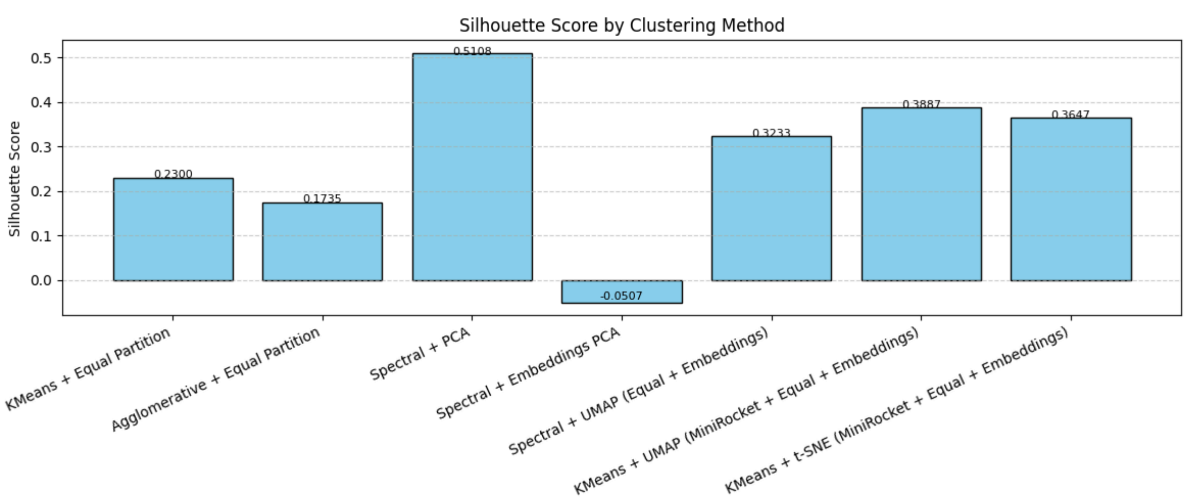

Figure 1: Silhouette Score by Clustering Method

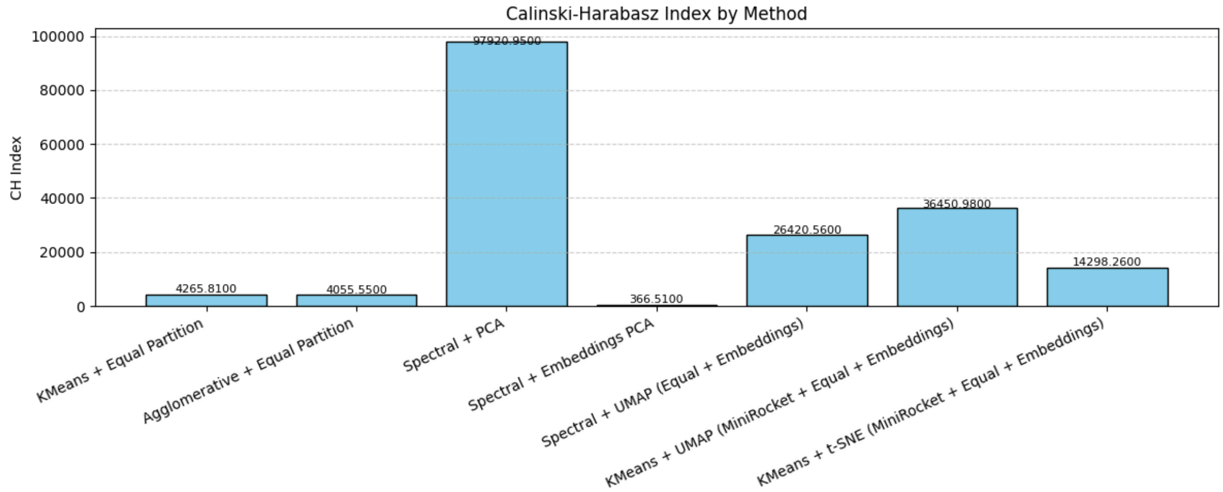

Figure 2: Calinski-Harabasz Index by Clustering Method

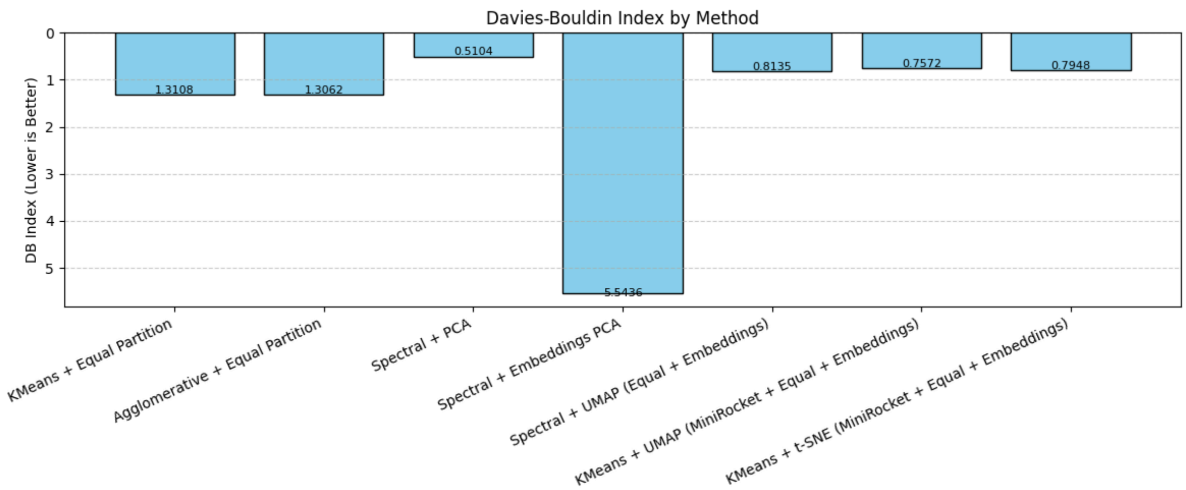

Figure 3: Davies-Douldin Index by Clustering Method

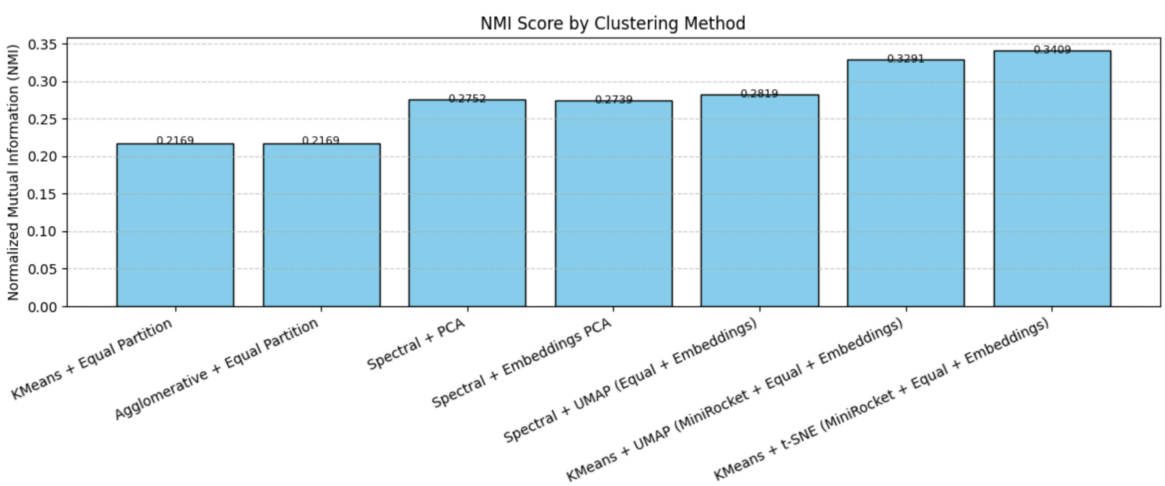

Figure 4: NMI Score by Clustering Method

Based on the internal evaluation metrics, I ultimately chose t-SNE as my dimensionality reduction technique because it produced strong overall results and aligned well with the structure of the underlying labels. Using t-SNE followed by KMeans clustering gave me a Silhouette Score of 0.3647 (Figure 1), a Calinski-Harabasz Index of 14,298.26 (Figure 2), and a Davies-Bouldin Index of 0.7948 (Figure 3). While these internal scores were slightly lower than some UMAP-based runs, this combination yielded my highest Normalized Mutual Information (NMI) score of 0.3409, indicating the best alignment with ground truth. This confirmed that t-SNE was particularly effective at preserving class structure in the embedded space, making it the best choice for my final submission.

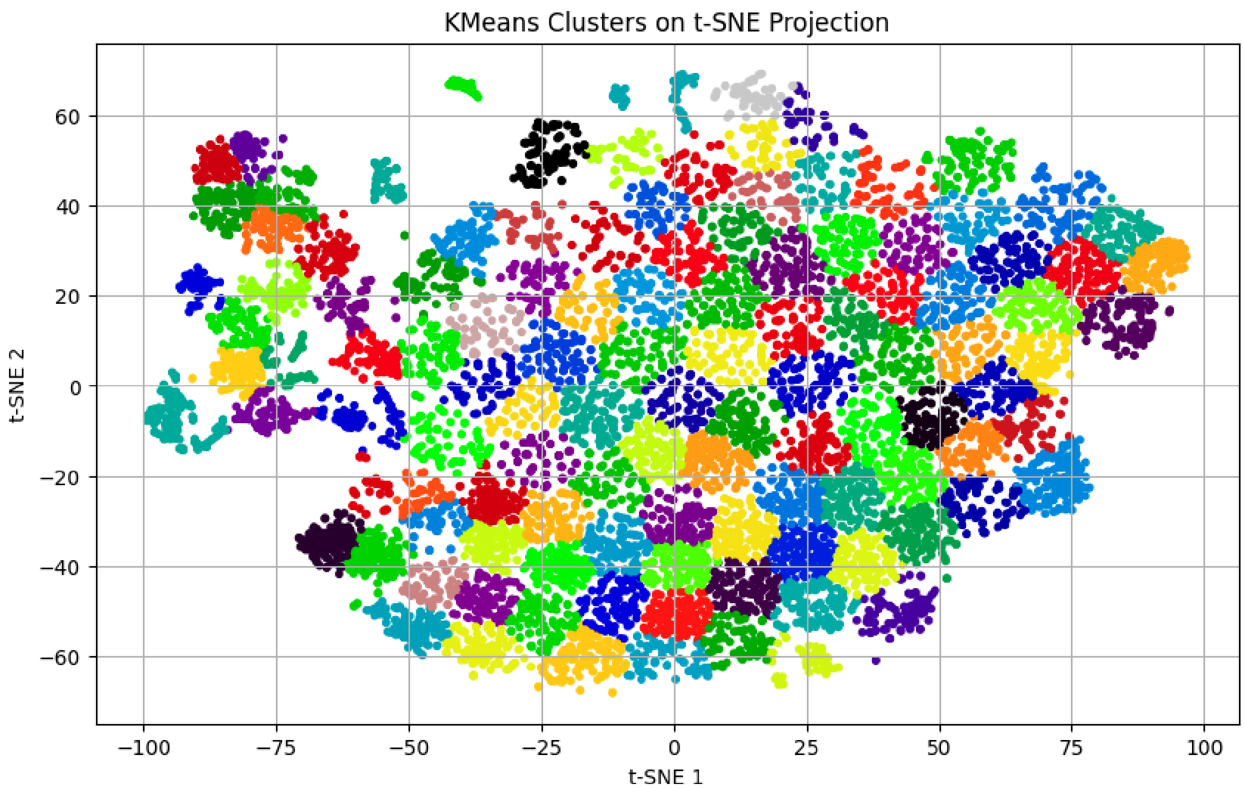

Figure 5: KMeans Clusters on t-SNE Projection

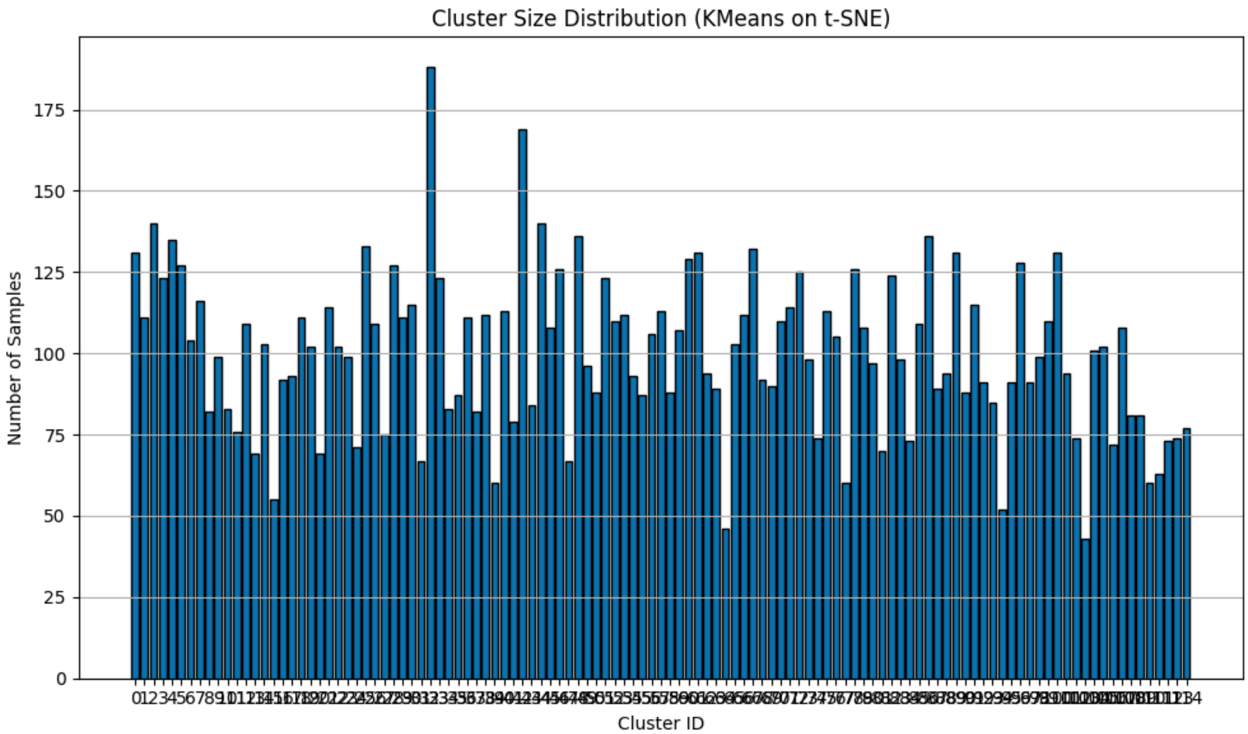

Figure 6: Histogram of Cluster Size Distribution for t-SNE