Distance-Based Analysis of Functional Connectivity Data for ADHD Classification

Introduction

This project uses functional connectivity data from an ADHD dataset to study how different distance measures describe similarity between participants and how useful these representations are for classification. Each participant is described by a very large number of connectivity features, which makes direct analysis difficult. To address this, this project uses a distance-based approach. Instead of working directly with the original features, we first compute distance matrices that describe how similar each pair of participants is.

Using these distance matrices, we then examine whether the data show meaningful structure and whether this structure can help predict ADHD status. For each distance measure, we follow the same analysis pipeline. This includes visualizing participant structure using multidimensional scaling (MDS) and evaluating classification performance using Logistic Regression and Random Forest models. By comparing results across different distances under the same pipeline, this project aims to understand how distance choice affects both data representation and model performance.

The Data Set

The dataset contains two main components: a functional connectivity feature table and a label table. Each row in the connectivity table represents one participant, and each column represents a brain connectivity feature. The label table contains the participant ID and the ADHD outcome label. These two tables were merged using participant IDs to create the full dataset.

To avoid bias caused by class imbalance, we created a balanced subset by randomly selecting the same number of ADHD and non-ADHD participants. The final dataset used for modeling includes 600 participants. Each participant is represented by approximately 19,900 connectivity features. Because the number of features is much larger than the number of participants, the choice of distance measure and dimensionality reduction method is especially important.

Before modeling, we computed distance matrices for each distance definition and visualized them using heatmaps. These heatmaps help check whether the distances are reasonable and whether the data show any clear structure before applying MDS or classification models.

Results

For each distance measure, we followed the same analysis steps. First, we visualized the distance matrix using a heatmap (Figure 1). Second, we used the distance matrix as input to MDS and created two-dimensional scatter plots to visualize participant structure (Figure 2). Third, we varied the MDS dimensionality from 2 to 10 and used GridSearchCV with five-fold cross-validation to evaluate Logistic Regression and Random Forest models.

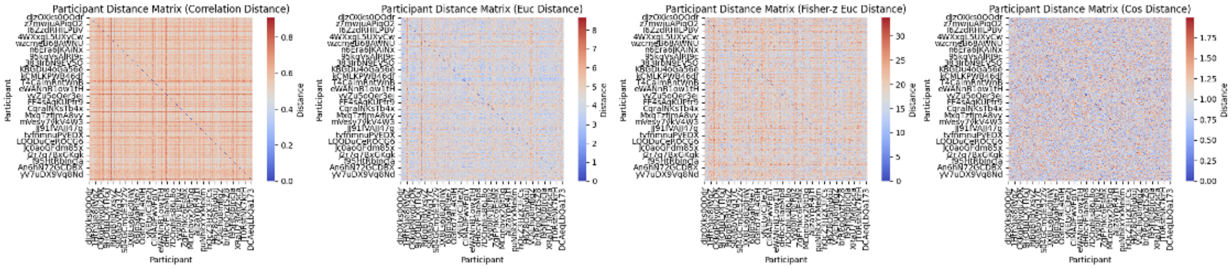

Figure 1. Distance Matrix heatmap. From left to right with the sequence of Correlation Distance, Euclidean Distance, Euclidean Distance with fisher-Z, and Cosine Distance.

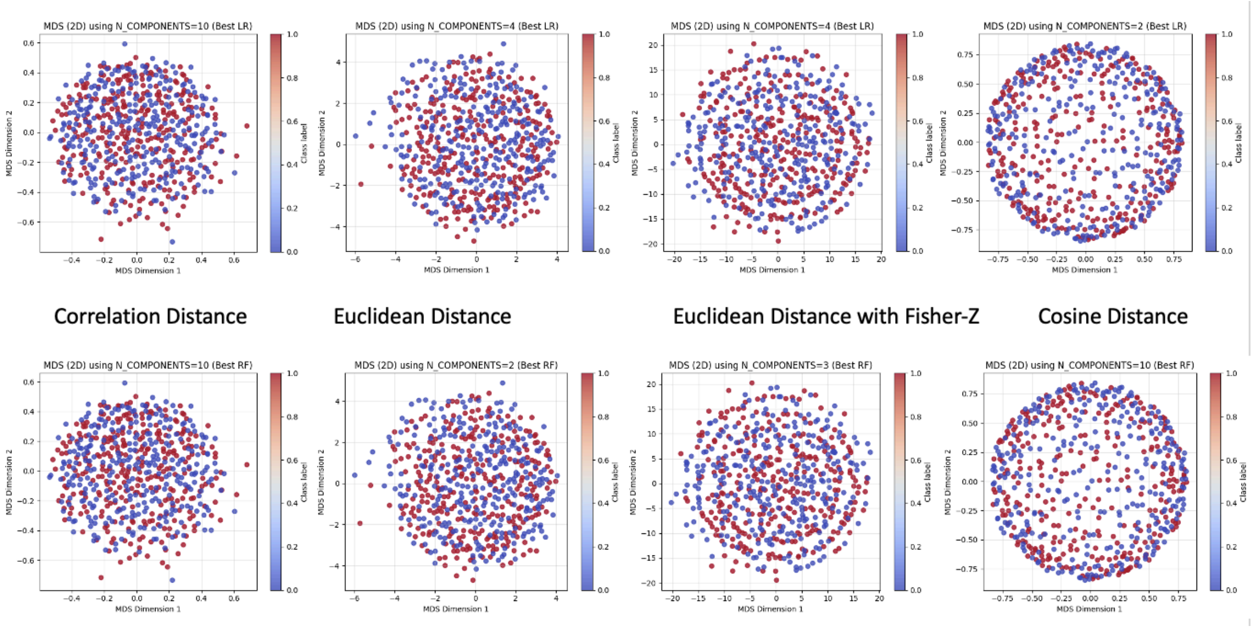

Figure 2. MDS first two dimension visualization. From left to right with the sequence of Correlation Distance, Euclidean Distance, Euclidean Distance with fisher-Z, and Cosine Distance.

We first analyzed **correlation distance**. The heatmap for this distance is shown in Figure 1a. Correlation distance focuses on similarity in connectivity patterns rather than absolute values. The heatmap shows some local structure, but there is no clear separation between ADHD and non-ADHD participants. Using this distance, we performed MDS and visualized the two-dimensional embeddings in Figure 2. Each point represents one participant, and colors indicate ADHD status. The two groups overlap strongly, although some local clustering can be seen. We then evaluated classification performance across MDS dimensions. The best Logistic Regression accuracy was 0.54 with 10 components, and the best Random Forest accuracy was 0.565 with 10 components.

Next, we analyzed **Euclidean distance computed on the original features**. The heatmap for this distance is shown in Figure 1b. Compared to correlation distance, this heatmap shows different contrast and scale, since Euclidean distance is sensitive to feature magnitude. The two-dimensional MDS plots are shown in Figure 2. The participant distribution differs from the correlation-based embedding, but clear class separation is still not observed. In the classification results, Logistic Regression achieved a best accuracy of 0.538 with 4 components, while Random Forest achieved a best accuracy of 0.552 with 2 components.

We then analyzed **Euclidean distance with Fisher z transformation**. In this case, the connectivity values were first clipped to avoid extreme values and then transformed using Fisher z before computing Euclidean distance. The heatmap for this distance is shown in Figure 1c and appears more stable than the raw Euclidean distance. The two-dimensional MDS plots are shown in Figure 2. This distance produced the best overall performance. Logistic Regression reached a best accuracy of 0.543 with 4 components, and Random Forest reached a best accuracy of 0.558 with 3 components.

Finally, we analyzed **cosine distance**. The heatmap for cosine distance is shown in Figure 1d. Its overall structure is similar to correlation distance, since both focus on pattern similarity. The MDS visualizations are shown in Figure 2. According to the notebook output, the best Logistic Regression accuracy was 0.532 with 2 components, and the best Random Forest accuracy was 0.540 with 10 components. These values match the correlation distance results and are reported directly from the notebook output without further assumptions.

In addition to these distances, we also computed a **cosine-based distance** and saved it as a separate file. However, this distance produced numerical instability due to invalid log values. As a result, it was not included in the full MDS and classification pipeline. The main analyses therefore focus on the four stable distance measures described above.

We can conclude that distance-based representations can capture some meaningful structure in ADHD functional connectivity data, when using random forest. However, the predictive signal is weak. Model choice has a small effect, while distance choice has a clearer impact on performance. These methods are most useful for exploratory analysis rather than strong prediction. Under logistic regression, original data without distance matrix has higher accuracy.

Distance Measure	Model	Best accuracy	Best MDS component	Best parameters
Correlation	Logistic Regression	0.54	10	C:0.01
	Random Forest	0.565	10	Max_depth:10, N_estimates: 300
Euclidean	Logistic Regression	0.538	4	C:0.01
	Random Forest	0.552	2	Max_depth:None, N_estimates: 500
Euclidean with fisher-z	Logistic Regression	0.543	4	C:0.01
	Random Forest	0.558	3	Max_depth:None, N_estimates: 100
Cosine	Logistic Regression	0.532	2	C:1
	Random Forest	0.540	10	Max_depth:1, N_estimates: 200
Original dataset without distance matrix	Logistic Regression	0.555	/	/
	Random Forest	0.523	/	/

Discussion

From these analyses, we learned how distance matrices can be used to summarize very high-dimensional data in a simple and useful way. Instead of working with thousands of connectivity features directly, a distance matrix describes how similar or different each pair of participants is. We also learned that different distance measures capture different aspects of the data, such as overall pattern similarity or feature magnitude, and that these choices affect both visualization and model performance. Using MDS helped me see the structure of the data in low dimensions, even when the classes were not clearly separated.

We also learned how GridSearchCV and cross-validation help evaluate models in a fair and systematic way. By testing different hyperparameters and embedding dimensions, we could see which model settings worked best and how stable the results were across folds. A simple way to understand the model performance is that participants with more similar connectivity patterns are slightly more likely to share the same ADHD label, but there is still a lot of overlap between groups. This suggests that functional connectivity alone provides a weak signal, but it can still be informative when combined with other features. In a lab or company setting, this kind of analysis can guide decisions about which distance measures and representations are more stable and worth using in later, more complex models.

Additionally, this project shows that distance-based representations can capture some structure in high-dimensional functional connectivity data, but the signal related to ADHD status is weak. Different distance measures emphasize different aspects of the data and lead to different MDS embeddings and model performance. Correlation and cosine distance focus on pattern similarity, while Euclidean distance is more sensitive to magnitude. Applying Fisher z transformation improves the stability and performance of Euclidean distance.

Although classification accuracy is slightly above chance, none of the models show strong separation between ADHD and non-ADHD participants. This suggests that functional connectivity alone may not be sufficient for reliable prediction. However, distance matrices and MDS provide useful tools for exploring data structure and comparing representations. Future work could include additional evaluation metrics and combine connectivity data with other types of features to improve performance.