**Final Report**

**ECE-5332 Data Science Project**

By:

**Ahmed Bayoumi**

**Amjad Sultan Awan**

**1. Introduction**

Alzheimer's disease (AD) is a neurological disorder that affects millions of people globally, causing a progressive decline in cognitive function. It is a significant public health concern that demands effective treatment and management. Early detection of AD is critical to provide appropriate interventions, which can help to slow down the progression of the disease. The identification of individuals with cognitive impairment is challenging, and thus, the use of normal control groups is necessary to understand the differences in cognitive performance between healthy and affected individuals. In this report, we will explore the importance of early identification of individuals with AD, and how normal control groups can be used to gain insights into cognitive decline associated with the disease.

Functional magnetic resonance imaging (fMRI) is a powerful tool for studying brain activity and has been used extensively in research on Alzheimer's disease. In this project, we aim to classify fMRI data into two classes: Alzheimer's disease (AD) and Normal control (NC) group. Our dataset consists of fMRI data from 10 subjects for each class.

To achieve this goal, we will employ signal processing and analysis techniques. Specifically, we will preprocess the fMRI data to clean and prepare it for analysis, select features that are relevant for classification, and use machine learning algorithms to classify the data into the two classes. We believe that this approach will provide valuable insights into the differences in brain activity between Alzheimer's disease and Normal control group subjects and help to identify potential biomarkers for early diagnosis and treatment of AD.

**2. Data Preprocessing**

To prepare the fMRI data for analysis, we employed several preprocessing steps using the Statistical Parametric Mapping (SPM) software (version 12). First, we applied motion correction using the realignment algorithm in SPM to correct for any movement-related artifacts. We then performed slice timing correction using the SPM Slice Timing tool to adjust for differences in acquisition time between slices. Figure 1 shows the batch editor window of SPM 12 for slice timing correction.

Next, we applied a high-pass filter with a cutoff of 128 seconds (about 2 minutes) to remove low-frequency noise and drift from the data. We also applied z-scoring to normalize the data within each subject. This technique is useful for removing systematic variation in the data and ensuring that each subject's data is on the same scale.

After that, we performed smoothing on the data using an 8 mm (about 0.31 in) full width at half maximum (FWHM) Gaussian kernel. This step was done to reduce the effect of individual differences in gyral and sulcal patterns.

Graphical user interface, application

Description automatically generated

Figure 1: SPM Batch Editor Interface

To be able to compare the same regions of the brain across different subjects, the data is normalized to a standard space. We used the standard MNI (Montreal Neurological Institute) space normalization method

in SPM.

We then extracted the time series for each voxel within the brain using a mask created from the Automated Anatomical Labeling (AAL) atlas. This allowed us to work with specific regions of interest in the brain. Particularly, hippocampus and entorhinal cortex for our work with Alzheimer's disease classification.

During preprocessing, we encountered some challenges related to motion artifacts and data quality. To address these issues, we visually inspected the data for any obvious artifacts and excluded subjects with poor data quality. There were 2 severe cases, one in each category which were discarded. One such bad instance is shown in figure 2, fourth one from the right in the top row.

A picture containing text

Description automatically generated Figure 2 : Reg check for motion artifacts and data quality

Overall, these preprocessing steps helped to ensure that the fMRI data was of high quality and free from noise or other interference. By normalizing the data and extracting relevant features, we were able to prepare the data for classification analysis and identify potential biomarkers for early diagnosis and treatment of AD.

**3. Feature Selection**

After preprocessing the fMRI data, the next step was to select a subset of relevant features to use in our classification analysis. Feature selection is an essential step in machine learning and data analysis, as it helps to reduce the dimensionality of the data and identify the most important variables for classification.

We flattened the brain volumes to vectors and got a 2D matrix of flattened brain volumes vs time. We then used principal component analysis (PCA) to reduce the dimensionality of the feature matrix. This technique allowed us to identify the most relevant features for classification while reducing the dimensionality of the data. We plotted the scree plot to see how many components were necessary and their relative importance, and it was a bit surprising that the first principal component accounted for more than 99% of the total variance as shown in figure 3. However, while using PCA for some subjects we were encountering numerical instabilities and Matlab was warning that the matrix is singular or badly scaled. Thus, PCA couldn’t be computed for about one third of the total subjects and Matlab returned NaNs as output.

Graphical user interface, text, application, email, Teams

Description automatically generatedFigure 3 : Scree plot

After dropping these subjects and the ones for which we had registration problems we were left with 6 subjects for AD and 4 subjects for NC.

To visually analyze the effectiveness of our selected features we did a UMAP plot to see if subjects from two different groups would form different clusters and we did find these very clearly separated in a few instances and not so clearly segregated in some others. Figure 4 shows a UMAP plot in which the 2 groups are very clearly segregated.

![Diagram

Description automatically generated with medium confidence]()Figure 4 : UMAP plot of AD vs NC using the first principal component

**4. Classification Algorithms**

After preprocessing the fMRI data and selecting relevant features, the next step was to train and evaluate some classification algorithm to predict the class labels of the subjects in our dataset. Support vector machines (SVM) is a popular classification algorithm that works by finding the hyperplane that separates the classes in the feature space with the largest margin. The SVM algorithm can handle non-linear classification problems by mapping the input features to a higher-dimensional space using a kernel function. In our study, we used the fitcsvm function in Matlab to train the SVM model with a radial basis function (RBF) kernel.

One advantage of SVM is its ability to handle high-dimensional feature spaces. However, SVM can be sensitive to the choice of kernel function and the regularization parameter.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedFigure 5 : Classes and Support vectors scatter plot for a few dimensions’ pairs

As can be seen from figure 5, the support vectors in some of the dimensions were able to segregate the class while in some others the classes were inseparable.

**5. Results**

To evaluate the performance of selected classification algorithm, common metrics include the evaluation of accuracy, precision, recall, and F1-score. To evaluate the tradeoff between true positive rate and false positive rate, the receiver operating characteristic (ROC) curves are generated, and the area under the curve (AUC) is calculated. But in order to compute these evaluation matrices we would need more data. The data of only ten instances per class was already very limited to begin with and then after preprocessing the data about a third of the samples were discarded for bad data quality. So a sense of performance evaluation is achieved by the plots of support vectors in figure 5. Where we can notice that in some dimensions class separation is achieved but in others it is not the case. The dimensions are to many to be dealt with manually (about half a million), but some further method can be incorporated in subsequent study that takes out the most relevant dimensions and discards the others for better results.

**6. Conclusion**

The present study aimed to classify Alzheimer's disease (AD and Normal control group using functional Magnetic Resonance Imaging (fMRI) data. The study employed various techniques from signal processing, data science, and machine learning to preprocess, select features, and classify the data. The preprocessing of fMRI data is crucial in reducing noise and improving signal quality. In this study, SPM was used to preprocess the fMRI data. This included motion correction, spatial normalization, and smoothing. The choice of SPM was based on its popularity and ease of use in the neuroimaging community. However, other preprocessing tools such as FSL could also be used with comparable results.

The study's limitations should also be noted, including the small sample size, the use of a single imaging modality, and the limited number of feature selection methods and classification algorithms tested. Further studies with larger sample sizes and multimodal imaging data are needed to confirm the robustness of the findings and evaluate the generalizability of the classification models developed in this study. Additionally, future studies could explore the use of deep learning techniques for fMRI data classification, which could potentially lead to more accurate and robust classification models.