university of groningen / faculty of science and engineering / biomedical engineering

Cluster Analysis of FDG-PET Imaging of a Dementia Cohort

Monideepa Deepak Gupta
S 3795950

rijksuniversiteit groningen

Department of Nuclear Medicine and Molecular Imaging

Period: 03/02/2020 - 27/08/2020

Master Project

Supervisor:
Dr. Antoon T. M. Willemsen, Medical Physicist
Debora E. Peretti, PhD Canditate
Department of Nuclear Medicine and Molecular Imaging

Mentor:
Dr. Marcel Greuter, Head Medical Physics

umcg

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners. I understand that my thesis may be made electronically available to the public.

Abstract

Dementia refers to a clinical syndrome characterized by a progressive cognitive decline that interferes with the ability to function independently and its subtypes are classified according to the cause of dementia. Alzheimer's disease is the most common subtype of dementia. It is a neurodegenerative disease causing dementia, which comprises about 60% to 80% of cases. The sensitivity and specificity of the clinical diagnosis of these conditions suggest a substantial amount of misdiagnosis. The objective of this study was to perform a quantitative analysis of FDG-PET images, a reliable biomarker showing synaptic dysfunction and neurodegeneration, from patients experiencing dementia. This study will form a basis to explore the potential of eventually developing a classification model. For this, two clustering analysis, HCA and K-means were investigated, first, on the data matrix of Healthy controls and Alzheimer's disease and later mild cognitive impairment, an objective cognitive impairment condition with the preserved function, subject type was also included. Principal component analysis, a feature extraction unsupervised machine learning algorithm, was performed on image data to transform the high dimensional image to low dimensional principal component space, to be then used for clustering. K-means clustering resulted in a good separation between Healthy controls and Alzheimer's disease. From the results, it can be inferred that quantitative analysis of functional images from dementia cohort holds potential to be utilized in the development of a classification model.

Acknowledgement

I would like to thank my project supervisors Dr. Antoon Willemsen and Debora Peretti for their thorough guidance throughout the project. All the weekly meetings, regular feedback and timely directions made this process a holistic learning experience and thus, enjoyable. I would like to thank you especially, for the last few weeks which were exceptionally difficult. You have motivated me and helped me throughout in completing my thesis successfully.

Dedication

This thesis is dedicated to my family, friends, teachers and my dearest Bhakti.

Contents

List of Figures

List of Tables

1 Introduction

Dementia is an umbrella term referring to a clinical syndrome characterized by a progressive cognitive decline that interferes with the ability to function independently (Sheehan, 2012). Alzheimer's disease (AD) affects more than 30 million people around the world (Alzheimer's.Assoc., 2018) and is the most common neurodegenerative disease responsible for dementia, comprising 60% to 80% of cases (Duong et al., 2017). The sensitivity of current clinical diagnostic criteria for AD ranged from 71% to 87% and specificity from 44% to 71%, suggesting substantial rates of AD misdiagnosis among patients with cognitive impairment (Phung et al., 2009). Patients with mild deficits who do not meet the criteria for dementia are considered to have mild cognitive impairment (MCI), an objective cognitive impairment with preserved function (Duong et al., 2017). Annually, it is estimated that 10–15% of patients diagnosed with MCI progress to AD dementia (Farias et al., 2009). With the advent of new treatment options for challenging brain diseases, the need for accurate early detection and differential diagnosis is becoming increasingly evident. However, the early stages of many neurodegenerative disorders may be essentially asymptomatic or clinically non-specific because of the shared involvement of common final pathways (Spetsieris et al., 2009). Various classification approaches have been developed but visual comparisons or quantitative differentiation using conventional statistical methods such as ANOVA and linear discriminant analysis are often inconclusive because of several factors and probable loss of information with univariate approaches (Spetsieris et al., 2009). 18F-fluorodeoxyglucose positron emission tomography (FDG-PET) is considered a useful tool in the evaluation of patients with neurodegenerative disorders (Matias-Guiu et al., 2017). FDG-PET shows synaptic dysfunction and neurodegeneration and, hence, is a reliable biomarker, since it depicts specific brain regions impaired in each patient (Matias-Guiu et al., 2018). Thus, FDG-PET images of subjects were used to perform a novel approach of classification for this study.

One possible approach for classification is the unsupervised method of Clustering analysis (Hennig et al., 2016), which is a set of data exploratory technique that groups the data into clusters, where a cluster refers to a collection of data points aggregated together because of certain characteristics. These characteristics can be distances, similarities or differences within the attributes of data. Clustering is one of the most commonly used unsupervised machine learning algorithms for processing data. Cluster analysis of a multivariate dataset aims to partition a large data set into meaningful subgroups of subjects, such that each data point is assigned to a cluster where there are high intra-cluster (within cluster) similarity and low inter-cluster (between cluster) similarity. Several types of clustering techniques are available for use, such as hierarchical clustering, soft/hard partitional clustering, density-based clustering, model-based clustering and grid-based clustering (Xu and Tian, 2015). As every technique utilizes a different and specific optimization method, two techniques that preferred the data most, were used for clustering. Namely, Agglomerative hierarchical clustering (HCA), and K-means clustering techniques, both unsupervised machine learning algorithms are applied for cluster analysis.

The first clustering method to be used for this study was a hierarchical method, that takes into account the linkage between data points (Jain, 2010). Hierarchical clustering algorithms divide or merge a particular dataset into a sequence of nested divisions. The hierarchy of these nested partitions can be of two types: agglomerative (i.e. bottom-up) or divisive (i.e. top-down). In the agglomerative method, clustering begins by considering every data point as a cluster on its own, so each cluster is a singleton. It then progressively merges two data points that are closest to each other based on the distances from the distance matrix, then the distances are recalculated between the new and old clusters and again the closest clusters are merged. This is repeated until all clusters are merged into one single cluster including all points. This procedure of the hierarchical clustering involves the construction of a hierarchy of treelike structure known

as a dendrogram, a reference is shown in Figure 1 (Jafarzadegan et al., 2019). One advantage of HCA is that the number of clusters is not specified in advance and a dendrogram may aid with determining the optimal number of clusters by visual analysis. However, the ward method for merging the clusters in HCA may also be used. Ward criterion minimizes the total within-cluster variance and finds the pair of clusters that leads to a minimum increase in total within-cluster variance after merging.

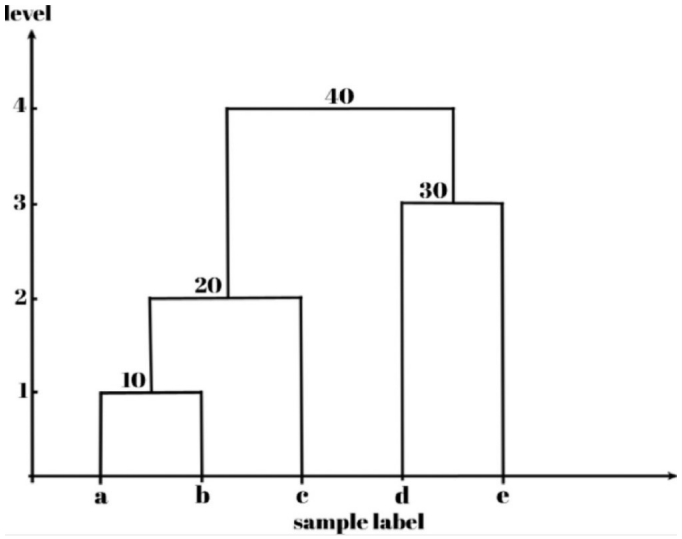

Figure 1: Dendrogram (Jafarzadegan et al., 2019).

Another clustering method used in this study is K-means clustering, it is one of the simplest and most used clustering algorithms, aiming to cluster similar data points together (Xu and Tian, 2015). K-means is a hard partitioning clustering method as it segregates the data in such a way that each data point belongs to only one cluster. K-means starts with the first group of randomly selected centroids, which are used as the beginning points for every cluster, and then assigns every data point to the cluster with the closest centroid. The distance of centroid to data point is calculated using the specified method for calculating distance. With the addition of data points, the centroid of every cluster is updated. It can be calculated as the median or mean of the cluster. Now, the distance of every data point to the updated centroids are calculated again, this is done for every cluster and data points are reassigned to the closest centroid cluster. Iterative calculations are performed to optimize the positions of the centroids, which means that there are no longer data points which switch from cluster to cluster, thus fixing the centroid position. There are different options to measure the distance, such as the Euclidean distance and the city block or Manhattan distance. In Euclidean distance, each centroid is the mean of the points in that cluster and for city-block, each centroid is the median of the points in that cluster. For evaluating the cluster consistency the silhouette evaluation criterion was utilized. Each cluster is represented by a so-called silhouette coefficient, which is based on the comparison of its tightness and separation and shows which objects lie well within their cluster, and which ones are merely somewhere in between clusters (Rousseeuw, 1987).

FDG-PET images used as input contain a high number of voxels, where each voxel holds some functional information. Every voxel is a variable, and with this vast number of variables, it is difficult to study the relationship between them. By reducing the dimension of the variable space, that is by systematically narrowing the number of variables, there are fewer relationships between variables to consider. Broadly there are two ways to reduce dimensionality; feature extraction and feature elimination. For feature elimination, the variables which are thought to best predict the result are kept and rest are dropped and so, the contribution of unused variables is eliminated. Whereas, for feature extraction new independent variables are created, where each

new independent variable is a combination of each of old independent variables. This way the contribution from all the variables is sustained. Principal Component Analysis (PCA) is a statistical multivariate analysis tool for dimensionality reduction and data visualisation, which is used for feature extraction for this study. It utilizes the innate multivariate information associated with neuronal connectivity to assess the spatial covariance structure of the data and attribute relevant portions of the total variance to statistically independent (orthogonal) metabolic patterns (networks) (Rencher, 1995; Petersson et al., 1999). PCA takes the multivariate data matrix as input, uses an orthogonal transformation to produce a set of linearly independent output called principal components (PC's) or Eigenvectors. This transformation projects the high-dimensional data into a low-dimensional space composed of PCs. PCA defines a new orthogonal coordinate system that best describes the intrinsic variability of the data, where few PC's retain most of the variability of the data. These high variation PC's and their corresponding eigenvalues may be further used for cluster analysis.

The goal of this study was to perform clustering analyses on a cohort of dementia patient data to eventually develop a robust classification model that may identify and predict different dementia conditions. An automated model for quantitative analysis of imaging data is expected to be a useful tool in aiding a clinician in the diagnostic process.

2 Materials and Methods

2.1 Patient information

A cohort of sixty-six subjects was selected from a larger ongoing study at the memory clinic of the University Medical Centre Groningen (UMCG), Groningen, The Netherlands. The study was conducted in agreement with the Declaration of Helsinki and subsequent revisions. Patients with an MMSE score higher than 18 were considered mentally competent to give informed consent. This cohort of subjects had a minimum MMSE score of 22, therefore all subjects were considered mentally competent to give informed consent, which was approved by the Medical Ethical Committee of the UMCG (2014/320).

Subjects were first diagnosed by consensus of a multidisciplinary team based on clinical assessment following the guidelines of the National Institute on Aging Alzheimer's Association criteria (NIA-AA) (McKhann et al., 2011) for the AD patients, and on the Petersen criteria (Petersen et al., 2001) for the MCI patients. Healthy subjects presented no cognitive complaints and a mini-mental state exam score (MMSE) higher than 28. All subjects underwent standard dementia screening. Multimodal neuroimaging was also performed, including PIB and FDG PET scans, as well as T1-3D magnetic resonance imaging (MRI). After this, clinical diagnoses were reconsidered under the National Institute on Aging and the Alzheimer's Association Research Framework (Knopman et al., 2018). Subjects were then reclassified as AD, MCI, or healthy controls (HC) based on the PET images. A summary of the demographic characteristics is shown in Table 1.

Table 1: Demographic characteristics of patients

Diagnosis	**HC**	**AD**	**MCI**
Number of Subjects	18	18	24
Age (Years)	68 ± 5	66 ± 8	65 ± 7
MMSE Score	30 ± 1	24 ± 4	27 ± 2

2.2 Image Acquisition

All subjects underwent a static FDG-PET examination. Scans were performed with either a Siemens Biograph 40mCT or 64mCT scanner (Siemens Medical Solution, USA). Since both systems were of the same vendor and the same generation, the acquisition and reconstruction protocols were harmonized, and the calibration of the systems was equally done, there were no differences between data acquired by the scanners. Patients were in standard resting conditions with eyes closed during the scans. The radiotracer was synthesized at the radiopharmacy facility at the Nuclear Medicine and Molecular Imaging department at the UMCG, according to Good Manufacturing Practice, and it was administered via venous cannula.

Static FDG-PET scans were acquired 30 min after injection (203 8) and lasted for 20 min. All subjects were fasted for at least 6 h before injection, and glucose levels in plasma were measured before the scan, and the PET scan was only performed if glucose levels were lower than 7 mmol/l (Boellaard et al., 2010). All PET images were reconstructed from list-mode data using 3D OSEM (3 iterations and 24 subsets), point spread function correction, and time-of-flight. The resulting images had a matrix of 400 400 111, with isotropic 2-mm voxels, and smoothed 2-mm Gaussian filter at full width and half maximum (FWHM).

2.3 Pre-processing

Image registration was performed using PMOD software package (version 3.8; PMOD Technologies LLC). The T1 3D MRI was normalized to the Montreal Neurologic Institute (MNI) space using probability tissue maps, and the Hammers atlas was used to define the grey matter of the cerebellum volume of interest (VOI). FDG-PET images were aligned to the MRI of each subject respectively. Then, PET images were smoothed using a Gaussian filter of 6-mm at full width and half maximum, and all voxels out of the brain were removed from the image. Standardized uptake value ratios (SUVR) images were generated by normalizing the uptake of each voxel to the mean uptake of the cerebellum VOI. All PET images were transformed into atlas space for further analyses.

2.4 Data Analysis

Data analysis was performed using MATLAB (version R2020a). First, the individual patient images in the standard MNI space were filtered using a binary mask (Figure 2), defining the volume of interest of the original image, to eliminate any voxels outside the brain region. These masked images were then read in a data matrix, one patient at a time sequentially, so that every row of the data matrix represented a subject and every column, a voxel from the subject's brain, thus making dimension of data matrix as number of Subject X number of Voxel. The entire brain volume was transformed into a row vector for every subject, before combining all subject data in one matrix. Afterwards, the data matrix was normalized using standard Z-score per subject (i.e. row) and per voxel (i.e. column).

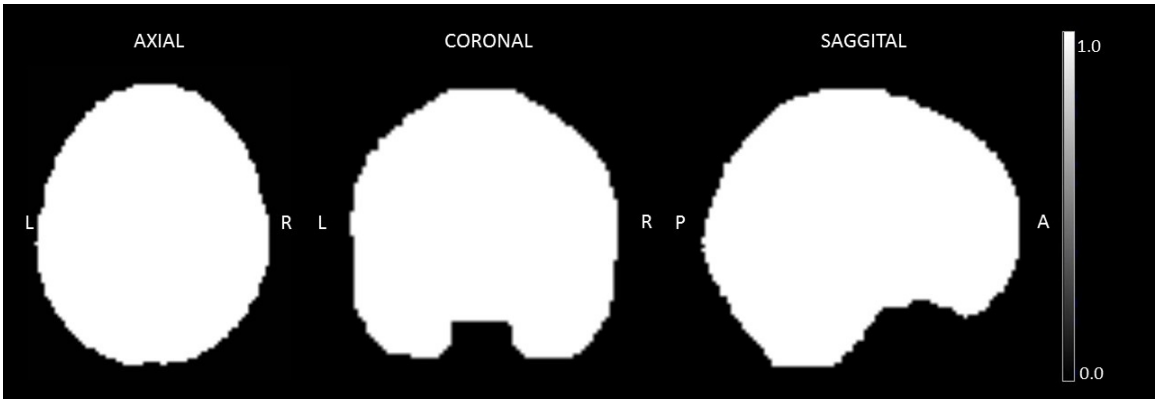

Figure 2: Axial, coronal and Sagittal view of the binary mask used for filtering voxels.

By performing principal component analysis on the normalized data matrix, subject data were translated from the high dimensional voxel space to a lower-dimensional, but high variance, principal component space. For every principal component (PC) (i.e. eigenvector), there was an associated score (i.e. eigenvalue). Scores were generated for every subject per PC's by a dot product between the PC and the subject's image. These scores were further used for the analysis of the images with unsupervised clustering techniques of HCA and K-means. Silhouette evaluation criterion was used to determine the optimal number of PC-score to be used. Silhouette coefficient for two to eight number of clusters was generated when the input for evaluation was only first two PC-score. Later, this was done again but now for three PC-score, and every time the silhouette co-efficient for two clusters was noted. This resulted in a plot of the number of PC-score vs silhouette coefficient, but for two clusters. From this plot, the highest value of silhouette co-efficient was chosen for the optimal value of PC-score.

Lastly, Clustering analysis was performed on optimal PC-score. For HCA clustering, the ward method for linkage was utilized to perform clustering on optimal PC-score and generate a dendrogram. On the X-axis of dendrogram every data point is replaced with its final clinical

diagnosis for a clearer insight into merging and Y-axis, as per usual, is the distance. To utilize the cluster information from dendrogram a labelling criterion was drawn, where based on the majority of the subject type present in a particular cluster, that cluster was labelled. Labels were HC-Cluster, AD-Cluster, MCI-Cluster or non-determined if no type was dominant. Clusters from the dendrogram were plotted in principle component space, for a different visualization. For K-means clustering, two methods of calculating distances were investigated: Euclidean and City-block distance. The number of iterations was kept as 200 and the entire clustering, using new initial cluster centroid positions, was repeated 5 times. Using the resulting indices, clusters were plotted in principle component space for K-means. As, with K-means results, it was difficult to label the clusters, inference and cluster position from HCA clustering result were formed as the basis for cluster labelling.

Initially, the data matrix with eighteen HC and twenty-four AD patients were subjected to clustering. Silhouette evaluation criteria for determining optimal PC-score was used for two clusters. Later, twenty-four MCI patients were added to the previous data matrix of HC and AD patients and the entire process of generating principal components to be used for clustering was performed again, but this time the silhouette evaluation criteria for three clusters was generated for determining optimal PC-score. For data matrix with HC, AD and MCI subject data put together, alongside K-means clustering result, expected K-means clusters were plotted for comparative analysis. This was done only for K-means, clusters were generated using the final clinical diagnosis and city-block distance, Euclidean distance was not used as it showed fluctuating results while investigating on HC and AD data matrix. Data points going against the general trend of its type (i.e. placed in a different cluster rather than with the data points of the same type) were marked as misclassified data points. This was determined by comparing the resulting cluster indices with the final clinical diagnosis that was available.

3 Results

3.1 For data matrix with HC and AD subject type

Initially, clustering analysis was performed on the data matrix containing eighteen HC and twenty-four AD patients. Principle component analysis of this data matrix resulted in 41 PC's. Figure 3 shows the amount of variance of individual PC's with respect to the total variance for all 41 PC'S.

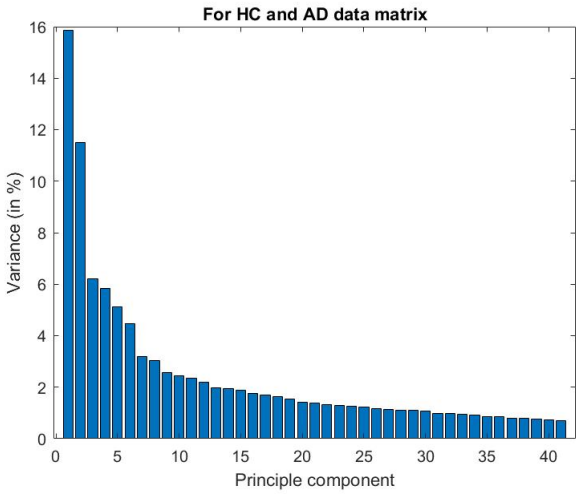

Figure 3: Scree plot for all 41 PC's obtained form the data matric of HC and AD subject type.

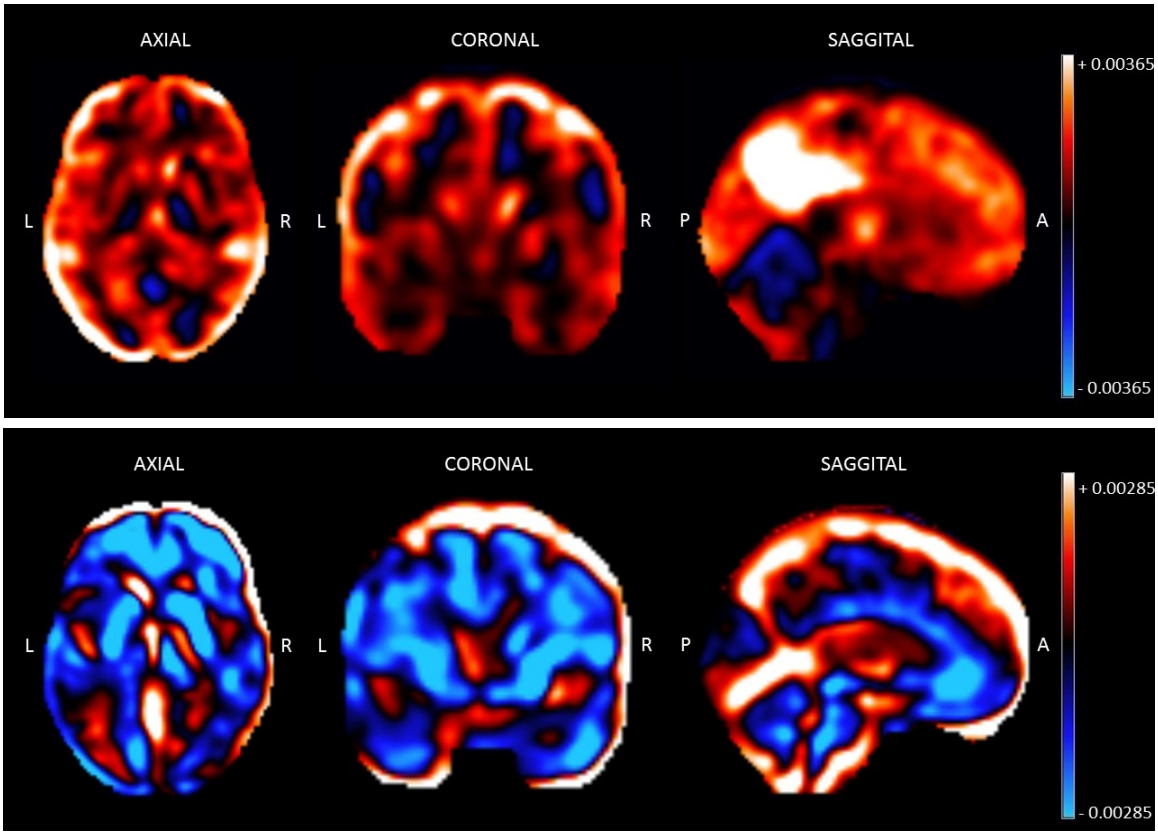

Figure 4: (a): shows highest variance, that, is PC-1 (eigenvector) for a set of HC subjects and AD patients in all three planes. (b): is PC-2 from the same data matrix.

Two highest variance PC's generated from HC and AD data matrix are shown in Figure 4. The range of the voxel values was made symmetrical setting the background of the image to zero. The blue region highlights the high variance brain region with negative values and red region highlights the high variance brain region but with positive values. Figure 4a is the highest variance PC, comprising approximately 16% of the total variance, whereas Figure 4b is the second-highest variance PC, comprising approximately 12% of the total variance.

Silhouette evaluation criterion for linkage method (HCA) concluded two highest variance PC-score, as the optimal number of PC-score to be used for generating two clusters. Figure 5 displays the silhouette coefficient vs the number of PC's.

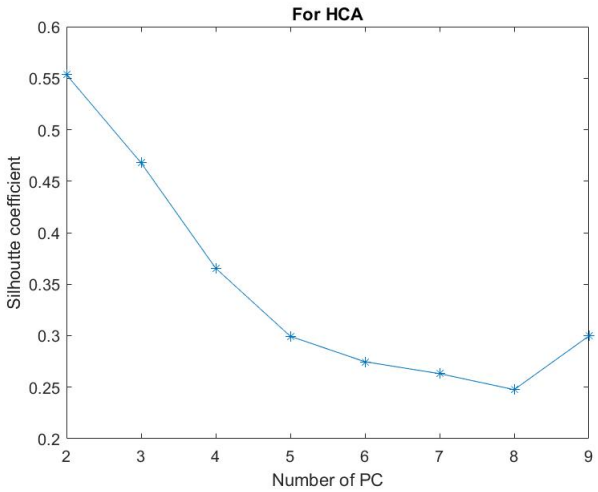

Figure 5: Silhouette coefficient for two cluster (y-axis) for matrix of different PC-score, ranging from 2 to 9 high variance PC-score (x-axis).

The dendrogram generated for ward linkage method using two PC-score is shown in Figure 6a. On X-axis, final clinical patient diagnosis is displayed, where 1 is for HC and 2 is for AD. Dendrogram clearly shows the presence of two clusters, where the red cluster can be labelled as HC-Cluster and blue as AD-Cluster, based on labelling criteria. Figure 6b. presents the translation of dendrogram as clusters in principle component space, for a different visual display. This cluster display also highlights the misclassified data points, marked as a black cross.

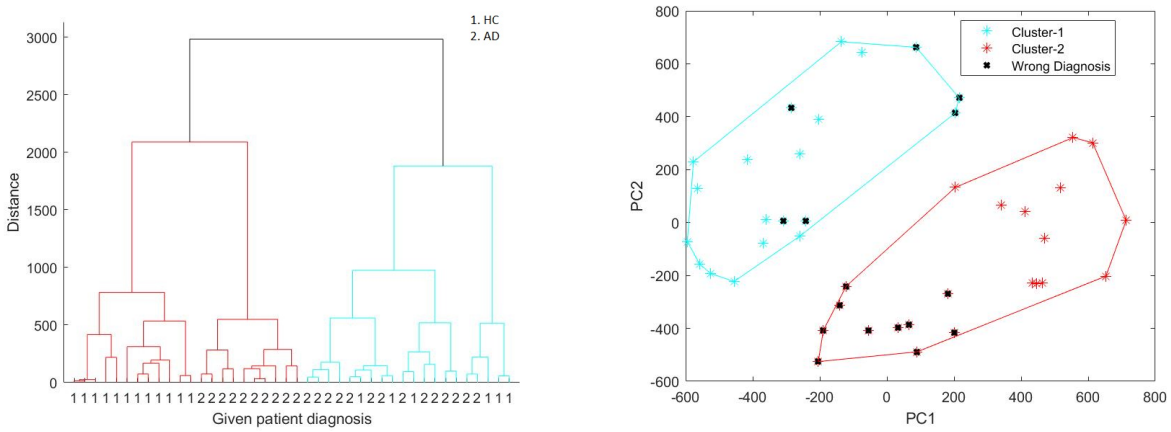

Figure 6: (a): Dendrogram generated using ward linkage method for HC and AD patient group. On the X-axis is the final clinical diagnosis and Y-axis is the distance between clusters. (b): cluster visualized in principle component space. Misclassified data points are highlighted.

HCA used two PC-score comprising 28% of the total variance and based on misclassification criterion resulted in 16 misclassifications, that is, 38% of data points were wrongly placed in the clusters. A Confusion matrix for HCA is shown in Table 2. Using HCA, there was 33% of misclassification and corresponding 66% of correct classification of HC subject type within the training data set. Whereas for AD patient type, the misclassification was 42% and the corresponding correct classification was 58% within the training data set.

Table 2: Confusion matrix for HCA clustering results for data matrix containing HC and AD patient type

For HCA technique		**True diagnosis**	
		HC	**AD**
Predicted diagnosis	**AD**	6	14
	HC	12	10
Number of subjects		18	24

Silhouette evaluation criterion for K-means clustering concluded that the two highest variance PC-score were optimal for classifying data into two clusters, for both Euclidean and city-block distance. Figure 7a and Figure 7b displays the silhouette coefficient vs the number of PC's for Euclidean distance and city-block (Manhattan) distance, respectively.

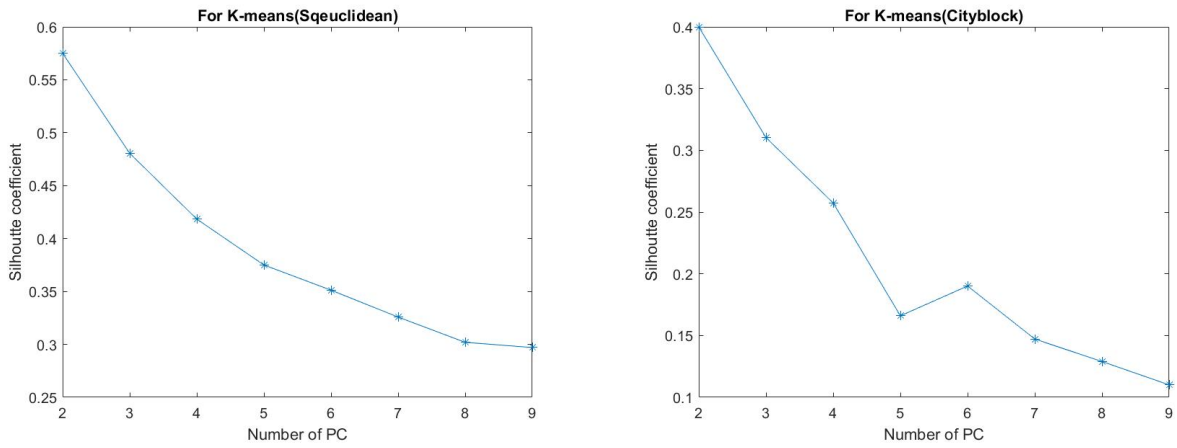

Figure 7: Silhouette coefficient for two clusters(y-axis) for increasing set of PC's (x-axis). (a): for the Euclidean distance and (b): for the city-block distance.

K-means clustering using two PC-score resulted in the clusters shown in Figure 8 for Euclidean method for generating distance matrix. Using knowledge of labelled cluster from HCA clustering result in Figure 6, the red cluster was labelled as AD-Cluster and blue as HC-Cluster. Multiple clustering results for the same input were observed with different misclassified data points ranging from 5 to 9 misclassification.

K-means clustering using two PC-score resulted in the clusters shown in Figure 9 for City-block method for generating distance matrix. Using knowledge of labelled cluster from HCA clustering result in Figure 6, the red cluster was labelled as AD-Cluster and blue as HC-Cluster.

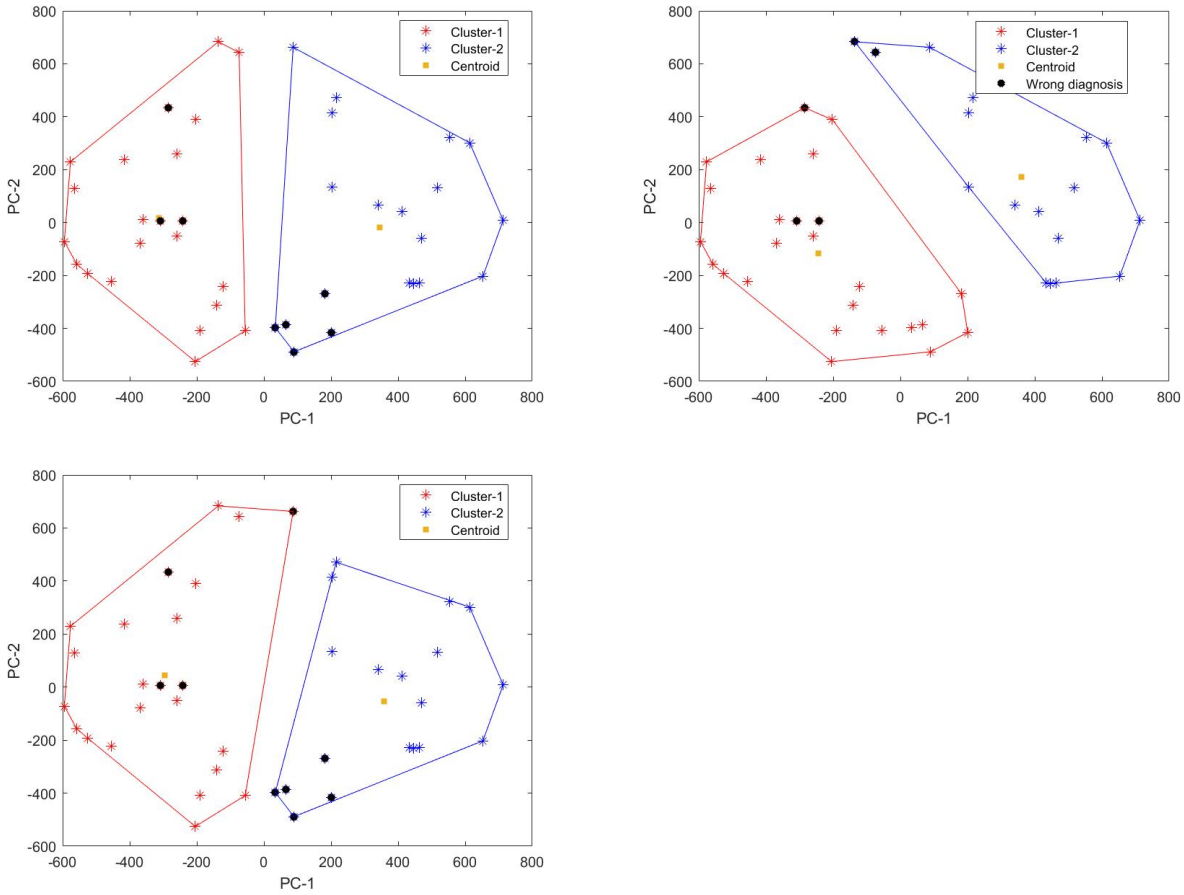

Figure 8: K-means clustering in principal component space using Euclidean distance. Clustering utilized two high variance PC-score. Misclassified data points are highlighted.

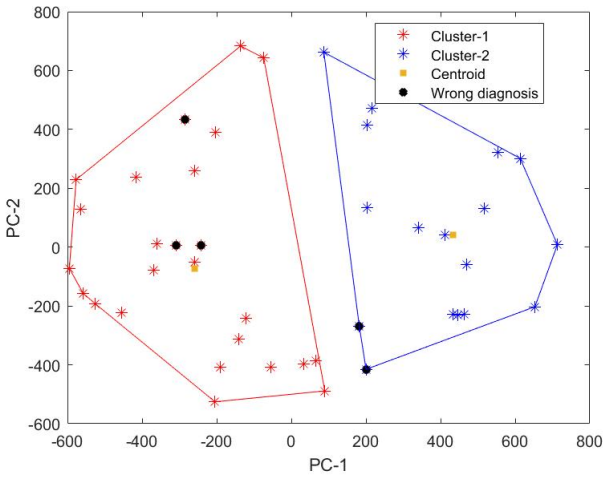

Figure 9: K-means clustering in principal component space using City-block distance. Clustering utilized two high variance PC-score. Misclassified data points are highlighted.

Both the distance approach used two PC-score for clustering, comprising 28% of the total variance. Multiple clustering results were seen for Euclidean distance, ranging from 5 to 9 misclassifications, whereas for city-block distance results were constant with 5 misclassifications. A Confusion matrix for K-means using city-block distance is shown in Table 3. Based

on misclassification criterion, there was 5 misclassification for city-block distance, that is, approximately 12% of data points were wrongly placed in the clusters. There was 16% of misclassification and corresponding 83% of correct classification of HC subject type within the training data set. Whereas for AD patient type, the misclassification was approximately 8% and the corresponding correct classification was 92% within the training data set.

Table 3: Confusion matrix for K-means clustering results for data matrix containing HC and AD patient type

For K-means technique (City-block distance)		**True diagnosis**	
		HC	**AD**
Predicted diagnosis	**AD**	3	22
	HC	15	2
Number of subjects		18	24

3.2 For data matrix with HC, AD and MCI subject type

Clustering analysis was performed on a data matrix containing eighteen HC subjects, twenty-four AD patients and twenty-four MCI patients, a total of sixty-six subjects. Principle component analysis on this normalized data matrix resulted in 65 PC's. Figure 10 shows the amount of variance of individual PC's with respect to the total variance for all 65 PC's.

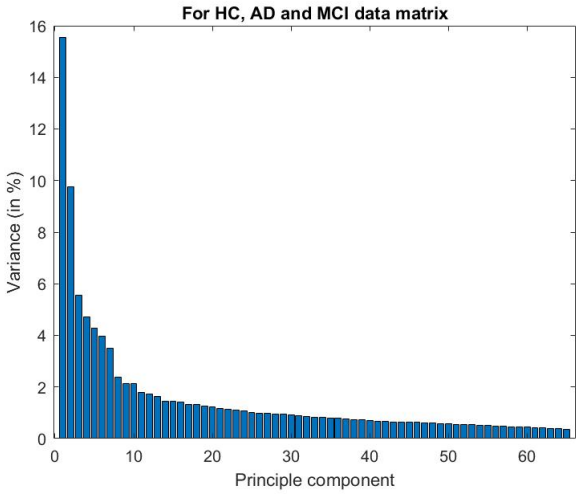

Figure 10: Scree plot for all 65 PC's obtained form the data matric of HC, AD and MCI subject type.

Two highest variance PC's generated from HC, AD and MCI data matrix are shown in Figure 11. The range of the voxel values was made symmetrical setting the background of the image to zero. The blue region highlights the high variance brain region with negative values and red region highlights the high variance brain region but with positive values. Figure 11a is the highest variance PC, comprising approximately 16% of the total variance, whereas Figure 11b is the second high variance PC, comprising approximately 10% of the total variance.

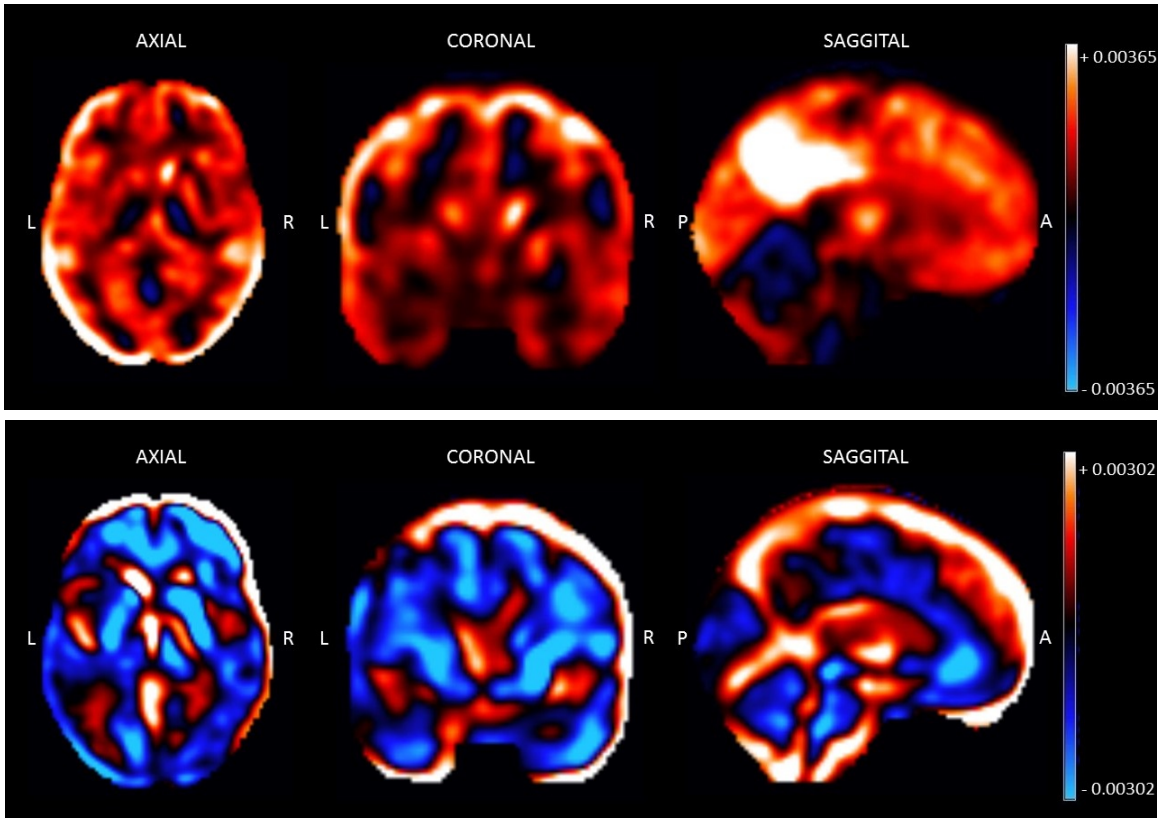

Figure 11: (a): shows highest variance, that, is PC-1 (eigenvector) for a set of HC subjects and AD, MCI patients in all three planes. (b): is PC-2 from the same data matrix.

Silhouette evaluation criterion for linkage method (HCA) concluded two PC-score, as optimal number of PC-score to be used for generating clusters. Figure 12 displays the silhouette coefficient vs the number of PC's.

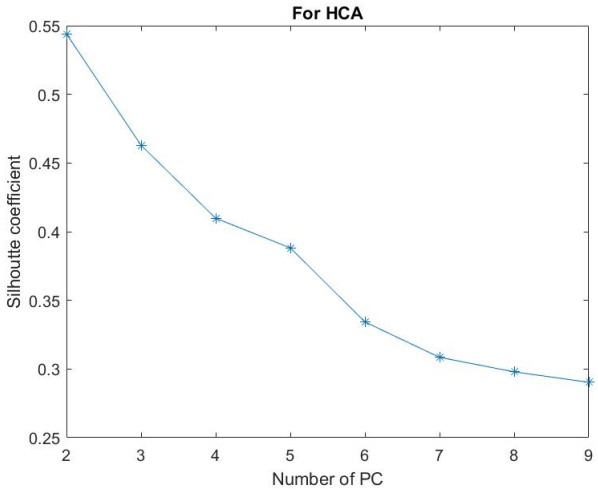

Figure 12: Silhouette coefficient for three cluster (y-axis) for matrix of different PC-score, ranging from 2 to 9 high variance PC-score (x-axis).

The dendrogram generated for ward linkage method using two PC-score, the optimal number of PC-score to be used for two clusters, is shown in Figure 13a. On X-axis is the final clinical patient diagnosis where 1 is for HC, 2 is for AD and 3 for MCI. The dendrogram is presented in a way, to show the formation of three clusters, as three is the desired cluster number. Red,

as well as a blue cluster, can be labelled as AD-Cluster and green as HC-Cluster based on labelling criteria. Figure 13b. presents the translation of the dendrogram as clusters in principle component space.

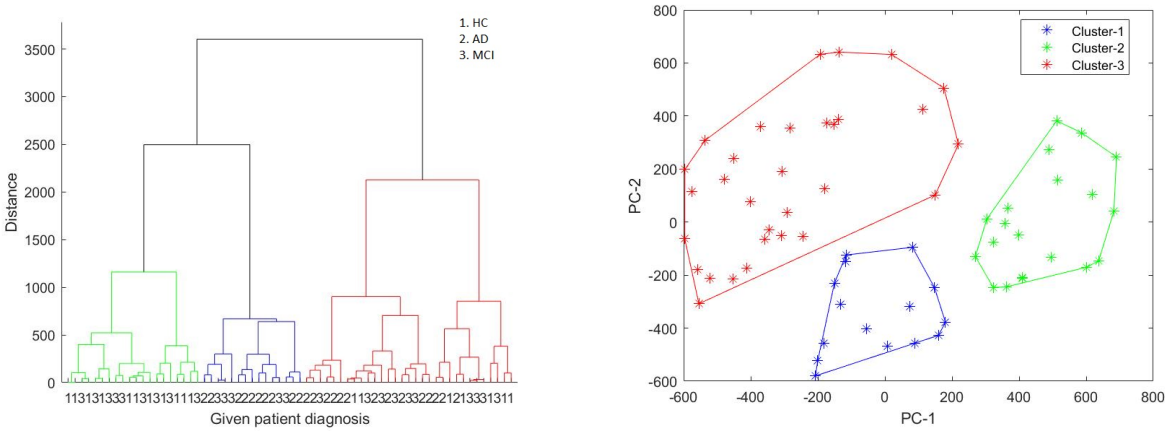

Figure 13: (a): Dendrogram generated using ward linkage method for HC, AD and MCI patient group. On the X-axis is final diagnosis and Y-axis is the distance between clusters. (b): cluster visualized in principle component space.

A Confusion matrix for HCA is shown in Table 4. HCA used 26% variance and resulted in 31 misclassifications based on misclassification criterion. Considering both AD-Cluster, red and blue cluster, there was a 100% correct classification for AD patients, combined. Within the red AD-Cluster, there was an overall 55% misclassification, within which 23% were HC and 32% were MCI misclassified to be AD. Within the blue AD-Cluster, there was an overall 33% misclassification, all of which were MCI misclassified to be AD. For HC-Cluster, green cluster, the correct classification was 55% and 45% of misclassifications were MCI as HC within the training data set. Within the HC-Cluster, no AD was misclassified as HC. Overall, MCI was completely misclassified.

Table 4: Confusion matrix for HCA clustering results for data matrix containing HC, AD and MCI patient type

For HCA technique		**True diagnosis**			
		HC	**AD**	**MCI**	**Number of subjects per cluster**
Predicted diagnosis	**AD (Red cluster)**	7 (23%)	14 (45%)	10 (32%)	31
	HC (Green cluster)	11 (55%)	0	9 (45%)	20
	AD (Blue cluster)	0	10 (66%)	5 (33%)	15
Number of subjects		18	24	24	66

Since K-means using Euclidean did not display the consistent result, it was excluded and only city-block was investigated for K-means for data matrix containing HC, AD and MCI patient type. Silhouette evaluation co-efficient was generated for K-means using city-block

which concluded that two high variance PC-score are optimal for three clusters. Figure 14 displays the silhouette coefficient vs the number of PC's for city-block (Manhattan) distance.

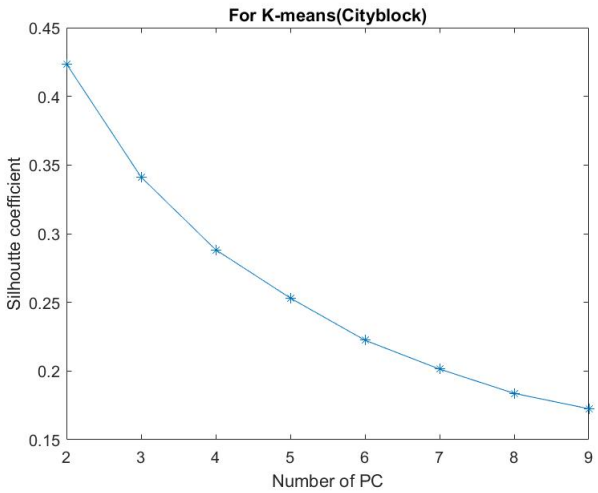

Figure 14: Silhouette coefficient using the city-block distance for three clusters (y-axis) for increasing set of PC's (x-axis).

K-means clustering using two PC-score resulted in the clusters shown in Figure 15a. for city-block method. Expected clusters was also plotted in Figure 15b alongside the final clustering result for comparative analysis of the outcome.

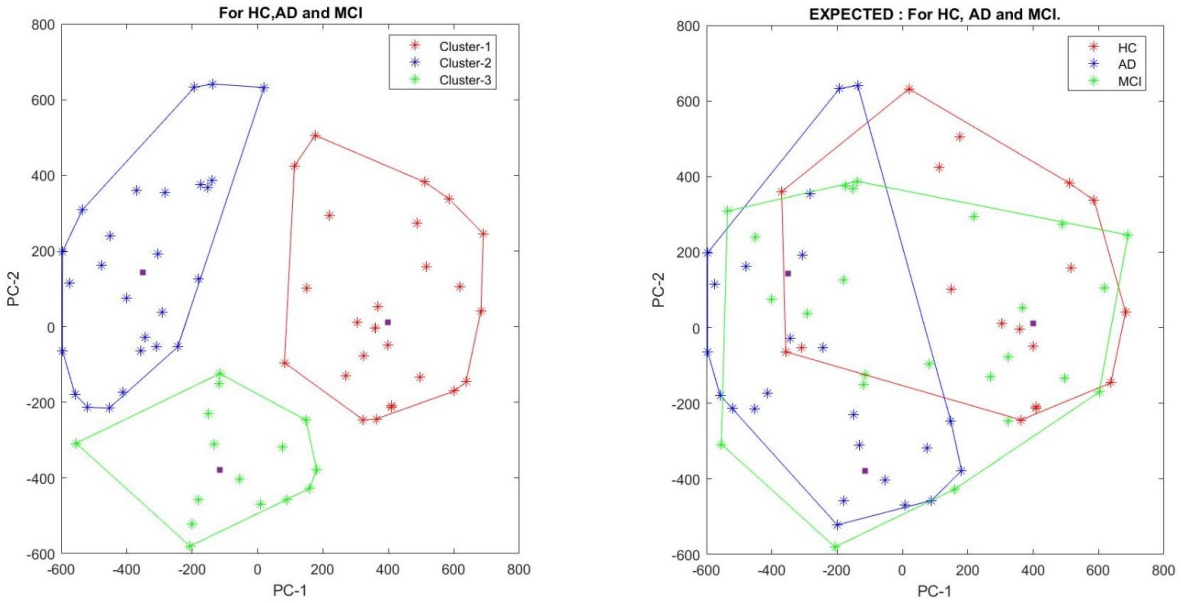

Figure 15: Cluster outcome using the K-means technique with city-block distance. (a): is the obtained clustering result using K -means and (b): is the expected clustering outcome for K-means.

A Confusion matrix for K-means using city-block is shown in Table 5. K-means used 26% variance and resulted in 28 misclassifications based on misclassification criterion. Considering both AD-Cluster, blue and green cluster, there was 100% correct classification for AD patients, combined. Within the blue AD-Cluster, there was an overall 46% misclassification, within

which 15% were HC and 31% were MCI misclassified to be AD. Within the green AD-Cluster, there was an overall 33% misclassification, all of which were MCI misclassified to be AD. For HC-Cluster, red cluster, the correct classification was 56% and 44% of misclassifications were MCI as HC within the training data set. Within the HC-Cluster, no AD was misclassified as HC. Overall, MCI was completely misclassified.

Table 5: Confusion matrix for K-means clustering results for data matrix containing HC, AD and MCI patient type

For K-means technique (City-block distance)		**True diagnosis**			
		HC	**AD**	**MCI**	**Number of subjects per cluster**
Predicted diagnosis	**AD (Blue cluster)**	4 (15%)	14 (54%)	8 (31%)	26
	HC (Red cluster)	14 (56%)	0	11 (44%)	25
	AD (Green cluster)	0	10 (66%)	5 (33%)	15
Number of subjects		18	24	24	66

4 Discussion

In this study, data-driven clustering analysis was performed on patient data with neurodegenerative disorders classified as Alzheimer's disease and Mild cognitive impairment. The objective was to explore within the cluster similarity and between the cluster differences and the extent of it, eventually to develop a classification model. The principal component analysis was performed on the data to reduce the dimension and bring it to an operative dimension. Later, two methods of clustering were investigated, HCA and K-means for cluster analysis.

Clustering analysis performed on a matrix containing HC subject and AD patient data resulted in different clustering results for both techniques (Figure 6 and Figure 9). The dendrogram of HCA technique reflected the presence of two broad clusters and, based on labelling criterion, HC-cluster and AD-cluster were labelled in principle component space. For K-means clustering technique, Euclidian and city-block distance were investigated, which resulted in different clustering outcomes. From K-means clustering result, it was difficult to label which cluster was HC-Cluster or AD-cluster. Thus, using the information of labelled HC and AD cluster from HCA clustering result and also the corresponding cluster positions in principle component space of these labelled clusters (Figure 6) clusters were labelled for K-means technique in Figure 8 and Figure 9. For City-block based clustering, the centroid is a median, while for Euclidean distance its centroid is the mean of the data points within the cluster. Therefore, the possibility of the centroid to change with a slight change in the cluster points happens more frequently for Euclidean distance-based clustering, which resulted in different clusters and also in different misclassifications every time, seen in Figure 8. Thus, the city-block distance was preferred as a parameter when performing k-means over Euclidian distance. HCA was performed using ward as a method for minimum variance linkage, where default distance is Euclidean distance. In HCA, data points simply merge to form a broader cluster. Here, merging is progressive and the minimum distance between cluster is the criteria for merging. Whereas, for K-means the centroid controls the cluster constitution, which is determined by all the data points with minimum distance. Thus, the clustering using K-means is a reflection of individual data points rather than small clusters as is in the case of HCA. Hence, clustering using the K-means method was better than HCA. K-means clustering with city-block resulted in the best clustering for HC and AD data matrix, considering diagnosis as a reference.

The clustering analysis performed on data matrix containing HC, AD and MCI patient type resulted in different results for both techniques (Figure 13 and Figure 15). For HC, AD and MCI data matrix, both the clustering technique classified all AD patients as a single cluster but failed in generating a separate cluster for MCI patient type. For HCA technique, 38% of MCI were classified as HC and 62% as AD, whereas for the K-means method, 46% of MCI were classified as HC and 54% as AD. This is in agreement with the literature of MCI dementia type (Duong et al., 2017) since it is considered an intermediate condition between HC and AD. Some interesting observation can be drawn from the expected cluster (Figure 15b), all MCI data points are spread throughout HC-Cluster and AD-cluster. Some MCI data points are present very deep in the HC-Cluster and remaining in AD-cluster right in the vicinity of other HC or AD data points in the Principle component space. Thus, the overlapping of MCI over both HC-Cluster and AD-cluster is not easily separated into a distinct cluster.

For further analysis, it will be interesting to look into different PC's for a more detailed insight on high variance brain regions and their association with the dementia type. Furthermore, dendrogram for both the data matrix (Figure 6a and Figure 13a) reflected on the presence of sub-clusters within the broad clusters. This might especially be interesting to investigate using data matrix of HC, AD and MCI, with MCI being an intermediate, progressive condition, as the expected cluster in Figure 15a shows the wide stretch of PC-score for this

condition. Another interesting result was the presence of high variance regions but with positive as well as negative values. It will be compelling to understand the relevance of these, and how the presence of these high variance positive and negative values is associated with the neurodegenerative condition and the brain region it impacts or its clinical symptoms. For example, in Figure 4a, one region of high variance with positive value is the precuneus and one with a negative value is the cerebellum. Another line of investigation would be to validate the clustering results using a new set of data matrix or by bootstrapping method for validation, for example. The longitudinal information of the misclassified patient, especially the HC's might also be an interesting line of investigation. Soft partitioning method like Fuzzy C-means can also be explored, particularly, for HC, AD, MCI data matrix, as there is a very strong overlapping of MCI over HC-Cluster and AD-Cluster. Fuzzy C-means delivers a membership function which indicates a data point belonging to multiple clusters.

The original dataset also contained frontotemporal dementia and Lewy body dementia patient data. However, since there were not may patients, they were not included in the analysis. But this can also make into an interesting line of investigation, with enough number of patient data for all other neurodegenerative condition.

5 Conclusions

The objective of this study was to perform clustering analysis on a dementia cohort to validate the potential of developing an automated tool for quantitative analysis of FDG-PET images. This tool might assist at the clinical level, eventually improving the diagnosis, by aiding in minimizing the late diagnosis and misdiagnosis. Different methods were used for clustering, which resulted in different results. Overall the results with K-means displayed a good separation between Healthy controls and Alzheimer's disease. Moreover, the results for Mild cognitive impairment classification resonated with its literature and showed potential for further investigation. In general, it can be concluded that by quantitatively analyzing FDG-PET images of dementia cohort, there is a possibility of the development of an automated classification algorithm.

A Ethics

The study was conducted in agreement with the Declaration of Helsinki and subsequent revisions, at the memory clinic of the University Medical Centre Groningen (UMCG), Groningen, The Netherlands. Subjects were considered competent to give informed consent based on their mini mental state exam score. The present study will set the track for further investigation in quantitative analysis of functional images of dementia cohort. The Long term benefit would be an automated clinical tool for diagnosis and prognosis of neurodegenerative conditions. This will help in improving sensitivity and specificity of diagnosis at clinical level, thus will contribute in healthy ageing.

B MATLAB Code: Read and transform data

```
clc;
clear all;
% reading the mask image
mask_V = spm_vol('C:\Users\jmdg9\OneDrive\Documents\spm\images\mask\Mask.n
img_mask = spm_read_vols(mask_V);
% storing filename as string
str1 = 'C:\Users\jmdg9\OneDrive\Documents\spm\images\S_';
str2 = '_FDG_ATLAS_SUVR_smoothed.nii';
patient_number = 1;
%file number given as input to i array

%HC,AD,MCI(+,-)
for i=[1 2 3 4 8 15 17 23 24 25 26 28 29 31 32 35 36 37 38 39 40 41 43 44 4

%HC,AD,MCI(+,-) without three HC
%for i=[1 2 3 4 8 15 17 23 24 25 26 28 29 31 32 35 36 37 38 39 40 41 43 44

%HC and MCI
%for i = [8 17 23 24 25 35 38 40 41 43 44 45 46 49 51 52 53 66 67 68 78 79

%for AD and MCI
%for i = [1 2 3 4 8 15 17 23 24 25 26 29 32 36 37 39 41 61 63 67 68 72 75 7

%AD,HC
%for i=[1 2 3 4 15 26 29 32 35 36 37 38 39 40 43 44 45 46 49 51 52 53 61 63

%AD,HC without three hc patients
%for i=[1 2 3 4 15 26 29 32 35 36 37 38 39 40 43 44 45 46 49 51 52 53 61 63

%AD,HC without three hc and two AD wrong patients
%for i=[1 2 3 4 15 26 35 36 37 38 39 40 43 44 45 46 49 51 52 53 61 63 66 72

%for all types
%for i = [1 2 3 4 8 11 15 17 21 23 24 25 26 27 28 29 31 32 34 35 36 37 38 3

%for all types WITHOUT 3 HC
%for i = [1 2 3 4 8 11 15 17 21 23 24 25 26 27 28 29 31 32 34 35 36 37 38 3

%for MCI+ AND MCI-
%for i =[8 17 23 24 25 28 31 41 54 60 64 67 68 71 79 80 83 86 97 104 108 11

%for HC and FTD
%for i = [11 21 27 35 38 40 43 44 45 46 49 51 52 53 66 78 81 85 89 91 95 10

%for DLB and FTD
%for i  = [11 21 27 34 58 76 89 94 95 96 99]

%for DLB and FTD and  all HC
%for i  = [11 21 27 34 35 38 40 43 44 45 46 49 51 52 53 58 66 76 78 81 85 8
```

```
50
51 %for DLB and FTD and HC
52 %for i  = [11 21 27 34 35 38 40 43 44 58 76 89 94 95 96 99]
53
54 %for DLB and FTD and AD
55 %for i  = [1 2 3 4 11 15 21 27 34 58 76 89 94 95 96 99]
56
57 %for DLB and FTD AD and HC
58 %for i  = [1 2 3 4 11 15 21 27 34 35 38 40 43 44 58 76 89 94 95 96]
59
60 %for FTD AD and HC
61 %for i  = [1 2 3 4 11 15 21 27 35 38 40 43 44 89 95]
62
63 %for DLB and FTD AD MCI and HC
64 %for i  = [1 2 3 4 8 11 15 17 21 23 24 25 27 34 35 38 40 43 44 58 76 89 94
65
66 %for DLB and FTD AD MCI(+,-) and HC
67 %for i  = [1 2 3 4 8 11 15 17 21 23 24 25 27 28 31 34 35 38 40 43 44 54 58
68
69 %for DLB and FTD AD MCI HC  all
70 %for i  = [1 2 3 4 11 15 21 26 27 29 32 34 35 36 37 38 39 40 43 44 45 46 49
71
72
73
74
75 %Balanced HC,AD,MCI
76 %for i=[1 2 3 4 8 15 17 23 24 25 26 29 32 35 36 37 38 39 40 41 43 44 45 46
77 %Balanced  HC,MCI
78 %for i = [8 17 23 24 25 35 38 40 41 43 44 45 46 49 51 52 53 66 67 68 78 79
79 %Balanced  AD,MCI
80 %for i = [1 2 3 4 8 15 17 23 24 25 26 29 32 36 37 39 41 61 63 67 68 72 79 8
81 %Balanced  AD,HC
82 %for i=[1 2 3 4  15 26 29 32 35 36 37 38 39 40 43 44 45 46 49 51 52 53 61 6
83
84 %INDIVIDUAL GROPUS
85 %HC
86 %for i =[35 38 40 43 44 45 46 49 51 52 53 66 78 81 85 91 106 110]
87
88 %MCI
89 %for i =[8 17 23 24 25 41 67 68 79 80 83 108 111 117]
90
91       % 04d since image has zero in its name
92       imag_path = {str1 num2str(i,'%04d') str2};
93       % so that file naem can be passed as sting
94       imag_path = cell2mat(imag_path);
95       %spm functions to read voxels
96       %reading header first
97       V = spm_vol(imag_path);
98       img = spm_read_vols(V);
99       %img = img(:);
100      %Filtering with the given mask; element wise multiplication
```

```
101         %binary mask
102         filter_img = img.*img_mask;
103         %making the filtered image a vector
104         all_filter_img(:,patient_number) = filter_img(:);
105         patient_number = patient_number + 1;
106 end
107 %saving the image data as .mat for further use
108 save('all_filter_img.mat','all_filter_img');
109
110 mean_1 = mean(all_filter_img(:,1));
111 mean_2 = mean(all_filter_img(:,2));
```

C MATLAB Code: For HC and AD data matrix

```
clc;
clear all;
% reading data from .mat saved file
read_data = matfile('all_filter_img.mat');
%putting the read.mat in to a matrix form
all_filter_img = read_data.all_filter_img;
%saving original data
orig_data = all_filter_img;
%input data as subject X voxel
all_filter_img = all_filter_img.';


%patient diagnosis for HC and AD
pat_diag = [2;2;2;2;2;2;2;2;1;2;2;1;2;1;1;1;1;1;1;1;1;2;2;1;2;2;2;1;1;2;1

%patient diagnosis for HC and AD without wrong HC
%pat_diag = [2;2;2;2;2;2;2;2;1;2;2;1;2;1;1;1;1;1;1;1;1;2;2;1;2;2;2;2;1;2


%NORMALIZE DATA IN THREE DIFFERENT WAY.
%Normalize each subject.
mydata_norm1 = zscore(all_filter_img,[],1);
%Normalize each voxel.
mydata_norm2 = zscore(all_filter_img,[],2);
%Normalize each subject and voxel.
mydata_norm3 = zscore(mydata_norm1,[],2);


%pca on image data.
[coeff,score,latent,tsquared,explained,mu] = pca(mydata_norm3);


% variance percentage using explained
figure
bar(explained(:,1));
xlabel('Principle component');
ylabel('Variance (in %)');
title('For HC and AD data matrix')

% cluster analysis
% score_mat is high variance score matrix
% with first column as patient number
[Row_sc, col_sc] = size(score);
score_mat= [1:Row_sc].' ;
score_mat(:,2:col_sc+1) = score(:,1:col_sc);



%.....%%%HISTOGRAM%%%%%%%%%%
```

```
% figure
% scatterhist(score_mat(:,2),score_mat(:,3),'Group',pat_diag,'kernel','on'
% hold on
% %drawing the ouline around cluster
% %for cluster1
% K1 = convhull(score_mat(pat_diag== 1,2),score_mat(pat_diag== 1,3));
% %for cluster 2
% K2 = convhull(score_mat(pat_diag== 2,2),score_mat(pat_diag== 2,3));
%
% %X,Y Coordinates of hull points for cluster1
% xaxis_clus_1= score_mat(pat_diag== 1,2);
% yaxis_clus_1= score_mat(pat_diag== 1,3);
% plot(xaxis_clus_1(K1),yaxis_clus_1(K1),'b')
% hold on
% %X,Y Coordinates of hull points for cluster2
% xaxis_clus_2= score_mat(pat_diag== 2,2);
% yaxis_clus_2= score_mat(pat_diag== 2,3);
% plot(xaxis_clus_2(K2),yaxis_clus_2(K2),'r')
% hold on
% xlabel('score - 1')
% ylabel('score - 2')
% legend('HC','AD','Location','best')
% title('HC and AD')





%%%%%%%%%%%%%%%%%%.........AGGLOMERATIVE METHOD.......%%%%%%%%%%%%%%%%%
%..................%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%..................

Z =  linkage(score_mat(:,2:3),'ward','euclidean');
figure
[H,T,outperm] = dendrogram(Z,0);
%for plotting patient diagnosis on X axis.
for i = 1:size(pat_diag,1)
    outperm(2,i)= pat_diag(outperm(1,i),1);
end
Clus = cluster(Z,'maxclust',2);
cutoff = median([Z(end,3)]);
figure
h = dendrogram(Z,0,'colorThreshold',cutoff);
xticklabels(outperm(2,:))
xlabel('Given patient diagnosis')
% text(36.8,3080,'On X-axis:','FontSize',10)
% text(37,3000,'1:HC','FontSize',10)
% text(37,2920,'2:AD','FontSize',10)
ylabel('Distance')
%score1 vs Score2
%with cross marked as wrong diagnosis.
%original data
```

```
hold off

figure
plot(score_mat(Clus== 1,2),score_mat(Clus== 1,3),'c*','MarkerSize',7)
hold on
plot(score_mat(Clus== 2,2),score_mat(Clus== 2,3),'r*','MarkerSize',7)
hold on
plot(score_mat((Clus == pat_diag)== 1,2),...
    score_mat((Clus == pat_diag)== 1,3),'kx','MarkerSize',5,'LineWidth',2)
hold on
%drawing the ouline around cluster
%for cluster1
K1 = convhull(score_mat(Clus== 1,2),score_mat(Clus== 1,3));
%for cluster 2
K2 = convhull(score_mat(Clus== 2,2),score_mat(Clus== 2,3));
%X,Y Coordinates of hull points for cluster1
xaxis_clus_1= score_mat(Clus== 1,2);
yaxis_clus_1= score_mat(Clus== 1,3);
plot(xaxis_clus_1(K1),yaxis_clus_1(K1),'c')
hold on
%X,Y Coordinates of hull points for cluster2
xaxis_clus_2= score_mat(Clus== 2,2);
yaxis_clus_2= score_mat(Clus== 2,3);
plot(xaxis_clus_2(K2),yaxis_clus_2(K2),'r')
hold on
% for i = 1:size(score_mat,1)
% str = num2str(i,'%d') ;
% text(score_mat(i,2),score_mat(i,3),str)
% end
xlabel('PC1')
ylabel('PC2')
legend('Cluster-1','Cluster-2','Wrong Diagnosis','Location','best')
hold off





%%%%%%%%%%%%%%%%%%%.........KMEANS METHOD.......%%%%%%%%%%%%%%%%%%%
%...................%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%...................


[idx,C] = kmeans(score_mat(:,2:3),2,'Distance','sqeuclidean'...
            ,'replicates',5,'MaxIter',200);

%score1 vs Score2
%with cross marked as wrong diagnosis.
%original data
figure
plot(score_mat(idx== 1,2),score_mat(idx== 1,3),'r*','MarkerSize',7)
hold on
```

```
plot(score_mat(idx== 2,2),score_mat(idx== 2,3),'b*','MarkerSize',7)
hold on
plot(C(:,1),C(:,2),'s','MarkerSize',3,'LineWidth',2)
hold on
%marking the wrong diagnosed patient
plot(score_mat((idx == pat_diag)== 1,2),...
    score_mat((idx == pat_diag)== 1,3),'kx','MarkerSize',5,'LineWidth',3)
hold on
%drawing the ouline around cluster
%for cluster1
K1 = convhull(score_mat(idx== 1,2),score_mat(idx== 1,3));
%for cluster 2
K2 = convhull(score_mat(idx== 2,2),score_mat(idx== 2,3));
%X,Y Coordinates of hull points for cluster1
xaxis_clus_1= score_mat(idx== 1,2);
yaxis_clus_1= score_mat(idx== 1,3);
plot(xaxis_clus_1(K1),yaxis_clus_1(K1),'r')
hold on
%X,Y Coordinates of hull points for cluster2
xaxis_clus_2= score_mat(idx== 2,2);
yaxis_clus_2= score_mat(idx== 2,3);
plot(xaxis_clus_2(K2),yaxis_clus_2(K2),'b')
hold on
xlabel('PC-1')
ylabel('PC-2')
% title('For HC,AD')
legend('Cluster-1','Cluster-2','Centroid','Location','best')
hold off
% exportgraphics(fig,'k-means_HC_AD.pdf','Resolution',500)


% calculate optimal number of PC to use
for i = 3:10
eva_linkage = evalclusters(score_mat(:,2:i),'linkage','Silhouette','klist'
eval_mat(i-2,1) = i-1;
eval_mat(i-2,2) = eva_linkage.CriterionValues(2);
end
figure
plot(eval_mat(:,1),eval_mat(:,2),'-*')
xlabel('Number of PC')
ylabel('Silhoutte coefficient')
title('For HCA')
hold off

for i = 3:10
eva_linkage = evalclusters(score_mat(:,2:i),'kmeans','Silhouette','klist',
eval_mat(i-2,1) = i-1;
eval_mat(i-2,2) = eva_linkage.CriterionValues(2);
end
figure
plot(eval_mat(:,1),eval_mat(:,2),'-*')
```

```
203 xlabel('Number of PC')
204 ylabel('Silhoutte coefficient')
205 title('For K-means(Cityblock)')
206 hold off
```

D MATLAB Code: For HC, AD and MCI data matrix

```
clc;
clear all;
% reading data from .mat saved file
read_data = matfile('all_filter_img.mat');
%putting the read.mat in to a matrix form
all_filter_img = read_data.all_filter_img;
%saving original data
orig_data = all_filter_img;
%input data as subject X voxel
all_filter_img = all_filter_img.';

%AD is 2, MCI IS 3 and HC is 1
%patient diagnosis for HC,AD and MCI(+,-)
pat_diag = [2;2;2;2;3;2;3;3;3;3;2;3;2;3;2;1;2;2;1;2;1;3;1;1;1;1;1;1;1;3;3

%patient diagnosis for HC,AD and MCI(+,-) without THREE WRONG HC
%pat_diag =  [2;2;2;2;3;2;3;3;3;3;2;3;2;3;2;1;2;2;1;2;1;3;1;1;1;1;1;1;1;3

% FOR MCI + MCI-
%pat_diag = [1;1;1;1;1;2;2;1;2;2;2;1;1;2;1;1;1;2;2;2;1;1;2;1];


%NORMALIZE DATA IN THREE DIFFERENT WAY.
%Normalize each subject.
mydata_norm1 = zscore(all_filter_img,[],1);
%Normalize each voxel.
mydata_norm2 = zscore(all_filter_img,[],2);
%Normalize each subject and voxel.
mydata_norm3 = zscore(mydata_norm1,[],2);

%pca on image data.
[coeff,score,latent,tsquared, explained,mu] = pca(mydata_norm3);

figure
bar(explained(:,1));
xlabel('Principle component');
ylabel('Variance (in %)');
title('For HC, AD and MCI data matrix')

% cluster analysis
% score_mat is high variance score matrix
% with first column as patient number
[Row_sc, col_sc] = size(score);
score_mat= [1:Row_sc].' ;
score_mat(:,2:col_sc+1) = score(:,1:col_sc);



%%%%%%%%%%%%%%%%%%.....AGGLOMERATIVE METHOD.....%%%%%%%%%%%%%%%%%%
```

```
%..................%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%..................

Z =  linkage(score_mat(:,2:3),'ward','euclidean');
figure
[H,T,outperm] = dendrogram(Z,0);
%for plotting patient diagnosis on X axis.
for i = 1:size(pat_diag,1)
    outperm(2,i)= pat_diag(outperm(1,i),1);
end
Clus = cluster(Z,'maxclust',3);
cutoff = median([Z(end-1,3) Z(end-2,3)]);
    %Z(end-3,3)]);
    %Z(end-4,3) Z(end-5,3)]);
figure
h = dendrogram(Z,0,'colorThreshold',cutoff);
xticklabels(outperm(2,:))
xlabel('Given patient diagnosis')
ylabel('Distance')
%score1 vs Score2
%with cross marked as wrong diagnosis.
%original data
figure
plot(score_mat(Clus== 1,2),score_mat(Clus== 1,3),'b*','MarkerSize',7)
hold on
plot(score_mat(Clus== 2,2),score_mat(Clus== 2,3),'g*','MarkerSize',7)
hold on
plot(score_mat(Clus== 3,2),score_mat(Clus== 3,3),'r*','MarkerSize',7)
hold on
%drawing the ouline around cluster
%for cluster1
K1 = convhull(score_mat(Clus== 1,2),score_mat(Clus== 1,3));
%for cluster 2
K2 = convhull(score_mat(Clus== 2,2),score_mat(Clus== 2,3));
%for cluster 3
K3 = convhull(score_mat(Clus== 3,2),score_mat(Clus== 3,3));
%X,Y Coordinates of hull points for cluster1
xaxis_clus_1= score_mat(Clus== 1,2);
yaxis_clus_1= score_mat(Clus== 1,3);
plot(xaxis_clus_1(K1),yaxis_clus_1(K1),'b')
hold on
%X,Y Coordinates of hull points for cluster2
xaxis_clus_2= score_mat(Clus== 2,2);
yaxis_clus_2= score_mat(Clus== 2,3);
plot(xaxis_clus_2(K2),yaxis_clus_2(K2),'g')
hold on
%X,Y Coordinates of hull points for cluster3
xaxis_clus_3= score_mat(Clus== 3,2);
yaxis_clus_3= score_mat(Clus== 3,3);
plot(xaxis_clus_3(K3),yaxis_clus_3(K3),'r')
hold on
xlabel('PC-1')
```

```
ylabel('PC-2')
legend('Cluster-1','Cluster-2','Cluster-3','Location','best')
hold off





%%%%%%%%%%%%%%%%%%.........KMEANS METHOD.......%%%%%%%%%%%%%%%%%%
%..................%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%..................


% [idx,C] = kmeans(score_mat(:,2:3),3,'Distance','cityblock'...
%             ,'replicates',5,'MaxIter',200);
%
% % %score1 vs Score2
% % %with cross marked as wrong diagnosis.
% % %original data
% figure
% subplot(1,2,1)
% plot(score_mat(pat_diag== 1,2),score_mat(pat_diag== 1,3),'r*','MarkerSize
% hold on
% plot(score_mat(pat_diag== 2,2),score_mat(pat_diag== 2,3),'b*','MarkerSize
% hold on
% plot(score_mat(pat_diag== 3,2),score_mat(pat_diag== 3,3),'g*','MarkerSize
% hold on
% plot(C(:,1),C(:,2),'s','MarkerSize',3,'LineWidth',2)
% hold on
% %drawing the ouline around cluster
% %for cluster1
% K1 = convhull(score_mat(pat_diag== 1,2),score_mat(pat_diag== 1,3));
% %for cluster 2
% K2 = convhull(score_mat(pat_diag== 2,2),score_mat(pat_diag== 2,3));
% %for cluster 3
% K3 = convhull(score_mat(pat_diag== 3,2),score_mat(pat_diag== 3,3));
% %X,Y Coordinates of hull points for cluster1
% xaxis_clus_1= score_mat(pat_diag== 1,2);
% yaxis_clus_1= score_mat(pat_diag== 1,3);
% plot(xaxis_clus_1(K1),yaxis_clus_1(K1),'r')
% hold on
% %X,Y Coordinates of hull points for cluster2
% xaxis_clus_2= score_mat(pat_diag== 2,2);
% yaxis_clus_2= score_mat(pat_diag== 2,3);
% plot(xaxis_clus_2(K2),yaxis_clus_2(K2),'b')
% %X,Y Coordinates of hull points for cluster3
% xaxis_clus_3= score_mat(pat_diag== 3,2);
% yaxis_clus_3= score_mat(pat_diag== 3,3);
% plot(xaxis_clus_3(K3),yaxis_clus_3(K3),'g')
% hold on
% xlabel('PC-1')
% ylabel('PC-2')
```

```
% legend('HC','AD','MCI','Location','best')
% title('EXPECTED : For HC, AD and MCI.')
%
% subplot(1,2,2)
% plot(score_mat(idx== 1,2),score_mat(idx== 1,3),'r*','MarkerSize',7)
% hold on
% plot(score_mat(idx== 2,2),score_mat(idx== 2,3),'b*','MarkerSize',7)
% hold on
% plot(score_mat(idx== 3,2),score_mat(idx== 3,3),'g*','MarkerSize',7)
% hold on
% plot(C(:,1),C(:,2),'s','MarkerSize',3,'LineWidth',2)
% hold on
%%drawing the ouline around cluster
%%for cluster1
% K1 = convhull(score_mat(idx== 1,2),score_mat(idx== 1,3));
%%for cluster 2
% K2 = convhull(score_mat(idx== 2,2),score_mat(idx== 2,3));
%%for cluster 3
% K3 = convhull(score_mat(idx== 3,2),score_mat(idx== 3,3));
%%X,Y Coordinates of hull points for cluster1
% xaxis_clus_1= score_mat(idx== 1,2);
% yaxis_clus_1= score_mat(idx== 1,3);
% plot(xaxis_clus_1(K1),yaxis_clus_1(K1),'r')
% hold on
%%X,Y Coordinates of hull points for cluster2
% xaxis_clus_2= score_mat(idx== 2,2);
% yaxis_clus_2= score_mat(idx== 2,3);
% plot(xaxis_clus_2(K2),yaxis_clus_2(K2),'b')
%%X,Y Coordinates of hull points for cluster3
% xaxis_clus_3= score_mat(idx== 3,2);
% yaxis_clus_3= score_mat(idx== 3,3);
% plot(xaxis_clus_3(K3),yaxis_clus_3(K3),'g')
% hold on
% xlabel('PC-1')
% ylabel('PC-2')
% title('For HC,AD and MCI')
% legend('Cluster-1','Cluster-2','Cluster-3','Location','best')
% hold off


% calculate optimal number of PC to use
% for i = 3:10
% eva_linkage = evalclusters(score_mat(:,2:i),'linkage','Silhouette','klist
% eval_mat(i-2,1) = i-1;
% eval_mat(i-2,2) = eva_linkage.CriterionValues(2);
% end
% figure
% plot(eval_mat(:,1),eval_mat(:,2),'-*')
% xlabel('Number of PC')
% ylabel('Silhoutte coefficient')
% title('For HCA')
```

```
% hold off
%
% for i = 3:10
% eva_linkage = evalclusters(score_mat(:,2:i),'kmeans','Silhouette','klist
% eval_mat(i-2,1) = i-1;
% eval_mat(i-2,2) = eva_linkage.CriterionValues(2);
% end
% figure
% plot(eval_mat(:,1),eval_mat(:,2),'-*')
% xlabel('Number of PC')
% ylabel('Silhoutte coefficient')
% title('For K-means(Cityblock)')
% hold off
```

References

Alzheimer's.Assoc. 2018 alzheimer's disease facts and figures. *Alzheimer's Dementia*, 14(3): 367 – 429, 2018. ISSN 1552-5260. doi: https://doi.org/10.1016/j.jalz.2018.02.001. URL http://www.sciencedirect.com/science/article/pii/S1552526018300414.

R. Boellaard, M. J. O'Doherty, W. A. Weber, F. M. Mottaghy, M. N. Lonsdale, S. G. Stroobants, W. J. Oyen, J. Kotzerke, O. S. Hoekstra, J. Pruim, P. K. Marsden, K. Tatsch, C. J. Hoekstra, E. P. Visser, B. Arends, F. J. Verzijlbergen, J. M. Zijlstra, E. F. Comans, A. A. Lammertsma, A. M. Paans, A. T. Willemsen, T. Beyer, A. Bockisch, C. Schaefer-Prokop, D. Delbeke, R. P. Baum, A. Chiti, and B. J. Krause. FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0. *Eur. J. Nucl. Med. Mol. Imaging*, 37(1):181–200, Jan 2010.

S. Duong, T. Patel, and F. Chang. Dementia: What pharmacists need to know. *Can Pharm J (Ott)*, 150(2):118–129, 2017.

S. T. Farias, D. Mungas, B. R. Reed, D. Harvey, and C. DeCarli. Progression of mild cognitive impairment to dementia in clinic- vs community-based cohorts. *Arch. Neurol.*, 66(9):1151–1157, Sep 2009.

C. M. Hennig, M. Marina, F. Murtagh, and R. Rocci. *Handbook of cluster analysis*. CRC Press, 2016.

M. Jafarzadegan, F. Safi-Esfahani, and Z. Beheshti. Combining hierarchical clustering approaches using the pca method. *Expert Systems with Applications*, 137:1 – 10, 2019. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2019.06.064. URL http://www.sciencedirect.com/science/article/pii/S0957417419304737.

A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651 – 666, 2010. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2009.09.011. URL http://www.sciencedirect.com/science/article/pii/S0167865509002323. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR).

D. S. Knopman, S. B. Haeberlein, M. C. Carrillo, J. A. Hendrix, G. Kerchner, R. Margolin, P. Maruff, D. S. Miller, G. Tong, M. B. Tome, et al. The national institute on aging and the alzheimer's association research framework for alzheimer's disease: perspectives from the research roundtable. *Alzheimer's & Dementia*, 14(4):563–575, 2018.

J. Matias-Guiu, M. Cabrera-Martín, J. Matías-Guiu, and J. Carreras. Fdg-pet/ct or mri for the diagnosis of primary progressive aphasia? *American Journal of Neuroradiology*, 2017. ISSN 0195-6108. doi: 10.3174/ajnr.A5255. URL http://www.ajnr.org/content/early/2017/05/25/ajnr.A5255.

J. A. Matias-Guiu, J. D?az-?lvarez, J. L. Ayala, J. L. Risco-Mart?n, T. Moreno-Ramos, V. Pytel, J. Matias-Guiu, J. L. Carreras, and M. N. Cabrera-Mart?n. Clustering Analysis of FDG-PET Imaging in Primary Progressive Aphasia. *Front Aging Neurosci*, 10:230, 2018.

G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, R. C. Mohs, J. C. Morris, M. N. Rossor, P. Scheltens, M. C. Carrillo, B. Thies, S. Weintraub, and C. H. Phelps. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*, 7(3):263–269, May 2011.

R. C. Petersen, R. Doody, A. Kurz, R. C. Mohs, J. C. Morris, P. V. Rabins, K. Ritchie, M. Rossor, L. Thal, and B. Winblad. Current Concepts in Mild Cognitive Impairment. *Archives of Neurology*, 58(12):1985–1992, 12 2001. ISSN 0003-9942. doi: 10.1001/archneur.58.12.1985. URL https://doi.org/10.1001/archneur.58.12.1985.

K. M. Petersson, A. Howseman, S. Zeki, T. E. Nichols, J.-B. Poline, and A. P. Holmes. Statistical limitations in functional neuroimaging ii. signal detection and statistical inference. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 354(1387):1261–1281, 1999. doi: 10.1098/rstb.1999.0478. URL https://royalsocietypublishing.org/doi/abs/10.1098/rstb.1999.0478.

T. K. Phung, B. B. Andersen, L. V. Kessing, P. B. Mortensen, and G. Waldemar. Diagnostic evaluation of dementia in the secondary health care sector. *Dement Geriatr Cogn Disord*, 27 (6):534–542, 2009.

A. C. Rencher. *Methods of multivariate analysis*. John Wiley & Sons, 1995.

P. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):53–65, Nov. 1987. ISSN 0377-0427. doi: 10.1016/0377-0427(87)90125-7. URL https://doi.org/10.1016/0377-0427(87)90125-7.

B. Sheehan. Assessment scales in dementia. *Ther Adv Neurol Disord*, 5(6):349–358, Nov 2012.

P. G. Spetsieris, Y. Ma, V. Dhawan, and D. Eidelberg. Differential diagnosis of parkinsonian syndromes using PCA-based functional imaging features. *Neuroimage*, 45(4):1241–1252, May 2009.

D. Xu and Y. Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, Jun 2015. ISSN 2198-5812. doi: 10.1007/s40745-015-0040-1. URL https://doi.org/10.1007/s40745-015-0040-1.