IMPERIAL

IMPERIAL COLLEGE LONDON

SCHOOL OF PUBLIC HEALTH

Predicting dementia using machine learning approach on toxic aggregate images

*CID:*01911899
*Name:*Yuting Gu
*Word Count:*7829
*Supervisor:*Dr. Yu Ye

Submitted in partial fulfillment of the requirements for the MSc degree in Health Data Analytics and Machine Learning of Imperial College London

25th August 2024

Predicting dementia using machine learning approach on toxic aggregate images

Abstract

Dementia is a rapidly growing public health concern characterised by protein aggregates in the brain. Its morphological difference is suggested to be the potential parameter to diagnose its various types. As there is currently no clinically applicable diagnostic tool that uses aggregate morphology to classify dementia, this project seeks to establish a machine learning pipeline using single aggregate images for prediction of four types of dementia (Parkinson's disease (PD), dementia with Lewy bodies (DLB), frontotemporal dementia (FTD), and Alzheimer's disease (AD)). It also aims to differentiate the morphological features of aggregate between diseases. In this study, the pipeline incorporated two feature extraction methods, Convolutional Autoencoder (CAE) and computational shape measurements via the IMEA package. The extracted features were reduced using UMAP and subsequently clustered via K-Means. Significant clusters specific to different dementia types were identified through Dirichlet regression and characterised using shape measurements, revealing disease-specific morphological differences. Among the four classifiers employed, random forest demonstrates the best balance between performance and reliability, by achieving a noteworthy weighted average AUC on both pipelines (0.64 for CAE and 0.68 for IMEA). To conclude, this project is novel in the its development of the pioneering ene-to-end machine learning pipeline. The results successfully demonstrated the feasibility of the proposed pipeline and provided an important foundational framework for developing a scalable, high-throughput diagnostic tool using aggregate morphology. Moreover, the identified clusters validate the morphological distinctions of protein aggregates specific to different dementia types, warranting future investigations using aggregate morphology in early diagnosis and targeted interventions.

1. Introduction

1.1 Dementia and Public Health

Dementia is an umbrella term for symptoms affecting memory and cognition, often due to ageing-related conditions. Currently, the global population is ageing at an unprecedented rate. It was predicted that the number of people aged 60 years and above is expected to reach over 2 billion by 2050 [1]. Every year, nearly 10 million new cases of dementia are diagnosed globally [2]. Around 5-8% of people aged 65 and older have dementia and nearly 50% of those aged 85 and over [3]. As of now, more than 55 million people live with dementia worldwide. By 2030, this number is projected to reach 78 million and is estimated to be doubled by 2050 [4]. Normally, people living with dementia experience symptoms that affect memory, language, behaviours and problem-solving [5]. These include a progressive deterioration in cognitive abilities and a decrease in mobility which affects daily functioning, independence and quality of life. Therefore, it poses an emergent challenge to public health due to its increasing prevalence and the substantial care requirements it entails, especially in several sectors like healthcare, social services, and economics [6].

1.2 Diagnosis and Biomarker of Dementia

Currently, diagnosing neurodegenerative conditions and dementia involves a comprehensive assessment combining medical history, physical exams, neurological evaluations, cognitive testing, and various laboratory and imaging tests. Brain imaging techniques such as magnetic resonance imaging (MRI), computed tomography (CT) scan and positron emission tomography (PET) scan provide valuable insights into the structure and function of the brain, helping to identify characteristic changes associated with different types of dementia [7]. This allows the potential possibility of early detection of dementia, enabling timely diagnosis and intervention. However, brain imaging is difficult in a high-throughput manner and is limited to providing a definitive diagnosis for dementia including Alzheimer's disease (AD), Parkinson’s disease (PD), dementia with Lewy bodies (DLB), and frontotemporal dementia (FTD).

Recent advances combining biomarker data with clinical assessments and neuroimaging results provide a more comprehensive and accurate diagnosis [8-10]. The progressive accumulation of misfolded protein aggregates in the brain is a hallmark biomarker in neurodegenerative diseases. The individual misfolded proteins form aggregated structures called oligomers, then further aggregate and assemble into long and insoluble fibrils [11] (Figure 1). Toxic misfolded protein aggregates lead to disruption of cellular functions, such as impaired protein degradation, mitochondrial dysfunction, and disruption of intracellular trafficking, and can ultimately cause cell death [12, 13].

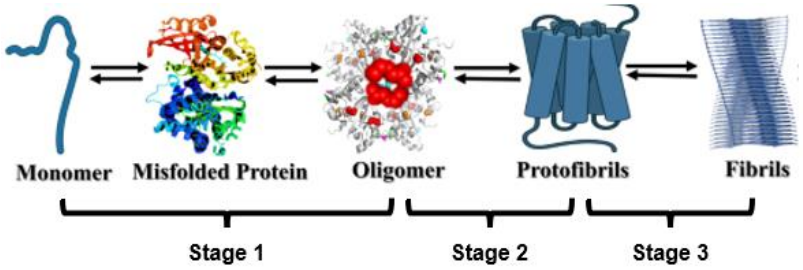


Figure 1: Schematic representation of the protein aggregation pathway, adapted from [11]. In the first stage, a normal functional protein (a single properly folded polypeptide chain) might be misfolded under certain conditions, then these misfolded proteins can cluster into small complexes known as oligomers which are more toxic than monomers. After stage one, oligomers can further aggregate to form an intermediate form, i.e. protofibrils, and eventually assemble into long and insoluble fibrils.

Different types of dementia are associated with overlapping protein aggregates, which makes it challenging to distinguish them, especially in the early stage. Specifically, AD and FTD are both characterised by the deposition of tau protein aggregates, while PD and DLB involve the aggregation of α-synuclein protein [14]. Additionally, DLB presents amyloid-β plaques in the brain, similar to those found in AD. However, recent findings suggest there are specific morphological variations of the same proteins distinct to each disease [13, 15-17]. This new perspective, considering such high-order structural information of aggregates, offers possibilities for differentiating between diseases with common proteins aggregating.

1.3 Distinct Aggregate Morphologies for Each Disease

The aggregates imaging technique is important to capture these morphological differences. Previous work from the Ye lab [18] quantified aggregate sizes ranging from 20 nm to 1 μm using single-molecule localisation microscopy (SMLM), a technique capable of resolving structural features down to ~25 nm (an example image is shown in Figure 2A). They recorded the toxicity responses of cells after incubation with aggregates isolated from donor brain tissues and aggregate concentrations were normalised to account for aggregate sizes. Their findings revealed distinct toxicity profiles of aggregates between PD and DLB (as shown in Figure 2B). Such high resolution, which is much smaller than the diffraction limit of a typical optical microscope, along with the observed disease-specific trends, provides opportunities to capture detailed morphological features of the aggregates and potentially distinguish disease-specific characteristics.

1.4 Deep Learning in Medical Images Analysis

To leverage aggregate images for distinguishing dementia types, extracting meaningful features is an irreplaceable computational process [19, 20]. Compared with traditional shape measurements, the non-linearity introduced by activation functions in deep learning allows models to learn more complex patterns and representations from data [21, 22].

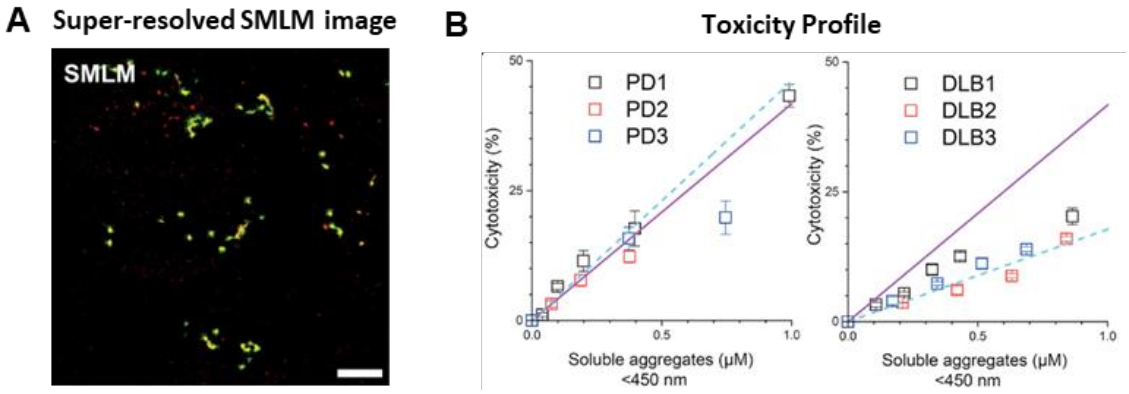


Figure 2: (A) An example of a super-resolved SMLM image of aggregates. (B) Plots of the fitted relationship between cytotoxicity values and aggregate concentration, colour-coded by patients, illustrate distinct toxicity profiles of aggregates between PD and DLB patients. Both figures are adapted from Morten et al. [18].

The autoencoder (AE) is a type of unsupervised deep-learning model that can extract features from images, by reconstructing the input images as similarly as possible [23]. AE contains two main components, encoder and decoder. The encoder is the part of the network that compresses inputs into lower-dimensional representations known as embedded features. Then these embedded features are used by the decoder to produce reconstructed outputs. During the training process, AE learnt to prioritise which aspects of inputs are more useful in reconstruction, and this bottleneck structure enables AE the ability to serve as a feature extractor.

The convolutional autoencoder (CAE) is a special form of AE that integrates the architecture of convolutional neural networks (CNN). CNN is a deep learning model that performs particularly superior in computer vision tasks [24, 25]. The integration of CNN in medical image analysis has demonstrated remarkable performance in various applications such as classification, segmentation, detection, etc [26-30]. The adaptation of convolutional layers from CNN is the key improvement of CAE compared to AE [31]. Unlike the fully connected layers used in AE, the mechanism of the shared kernel in convolutional layers leads to a focus on the local regions of the input image, so better preserving of spatial relationships between pixels. Moreover, this weight sharing also reduces the number of parameters, making CAE more efficient and less prone to overfitting. Thereby, CAE can generate better reconstruction and produce improved image representation than AE, which supports CAE for feature extraction of medical images in many recent works. For example, a CAE-based model, HistoCAE, was proposed to segment tumour regions in histopathological images of the liver [32]. Another CAE-based framework extracts latent representations from brain MRI for predicting Alzheimer's disease [33]. These applications prove the usage of deep learning in medical image analysis and also promise the possibility of using our aggregate images for dementia classification.

1.5 Hypothesis and Aims

Given that the clinical dementia diagnosis is mainly based on observed symptoms and a more accurate pathological diagnosis may only occur post-mortem, there is a real need for methods to aid more accurate and

timely diagnosis. As computational methods offer powerful techniques to characterise dementia biomarkers, therefore, in this project, I will assess whether deep learning analyses are able to classify dementia types, by defining morphological differences of aggregates from brain samples.

Specifically, the hypothesis is that each type of dementia has a subset of morphological disease-specific aggregates. Thus, I will achieve the following aims:

1. Develop a prediction pipeline for patient-level dementia classification based on aggregate populations from patient samples.
2. Use machine learning/deep learning approaches to analyse aggregate images to understand the differences in aggregate morphology between different diseases.

Similar analytical methods in the future will be used to characterise cerebrospinal fluid and plasma samples from living people to help diagnose dementia in a timely manner, which can lead to potential therapeutics and prevention.

2. Methods

2.1 Data Source and Image Dataset Acquisition

In this project, I analysed a dataset containing over a million images of single aggregates. These aggregates were previously isolated from post-mortem human brain tissue by the Ye Lab, sourced from Parkison's UK and Multiple Sclerosis Society Tissue Banks at Imperial College London and the Queen Square Brain Bank for Neurological Disorders at the University of London. Donors in this dataset were classified by their dementia diagnosis, including PD, DLB, FTD and AD, as well as control donors. Each class was matched by gender and age.

Aggregate images were collected using the single-molecule localisation microscope (SMLM). SMLM is a super-resolution technique that involves immobilising samples and stochastically switching fluorescent dyes to separate out single-molecule emissions temporally [34]. Figure 3 provides an overview of the three stages involved in obtaining single aggregate images from lab samples.

Before imaging, aggregates were extracted from 0.05 g of brain samples that were soaked in artificial cerebrospinal fluid (aCSF) and isolated via ultracentrifugation. The supernatants containing soluble aggregates were stored at -80 °C after 72 hours of dialysis in aCSF.

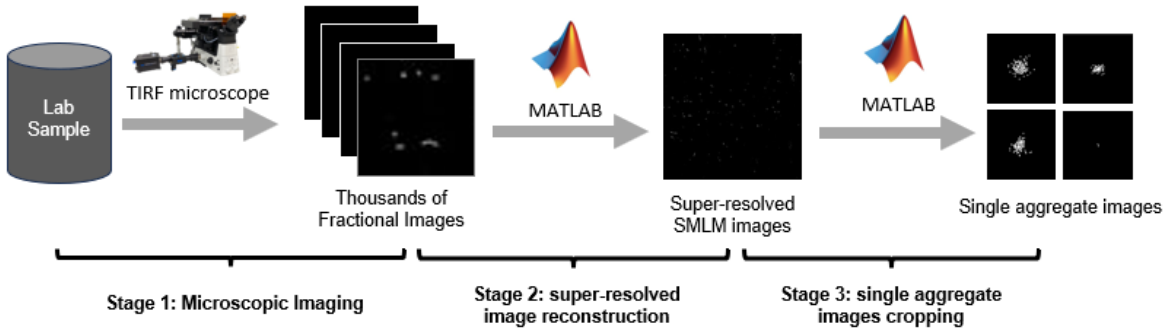


Figure 3: Flowchart illustrating the acquisition of single aggregate images from lab samples. Stage 1: A prepared lab sample was recorded by a TIRF microscope to collect thousands of fractional images per field of view. Stage 2: A set of fractional images was processed by MATLAB to reconstruct a super-revolved SMLM image. Stage 3: High-resolution images of single aggregates were cropped from the reconstructed image.

In the first stage, movies of immobilised aggregates labelled with a fluorescent dye were recorded, by an inverted total internal reflection fluorescence (TIRF) microscope. These movies captured the fluorescent bursts from single molecules, and from each fluorescent burst the position of each molecule is calculated with nanometre precision [18]. Movies were typically recorded for 2000 frames with each pixel in size of 73 nm, and ten fields of view, i.e. movies, were collected per person.

In the second stage, each movie with 2000 frames was passed to a MATLAB script to reconstruct one super-revolved SMLM image using the positions of all individual fluorophores. The resulting image decreased the

pixel size to 12 nm, increasing the resolution > 5 times compared to the diffraction-limited image. Each super-resolved image contains up to one thousand single aggregates. A more detailed description of the imaging technique can be found in Morten et al [18].

In the third stage, images of single aggregates were identified from the super-revolved SMLM images. A rolling-ball filter and a bandpass filter were applied to each reconstructed image to remove background noise. Each aggregate was identified using a bounding box defined by the 'regionprops' function in MATLAB. All images are in black and white only (i.e. binary pixel values), where white pixels represent the shape of each aggregate (as shown in the images on the right-hand side in Figure 3). These raw single aggregate images were further pre-processed before being fed into the computational method.

2.2 Image Dataset Preprocessing

Previous research showed smaller aggregates were more toxic than larger stable aggregates since they could cause additional aggregation and spreading by inducing more oxidative stress and activating pro-inflammatory cell signalling [35]. Therefore, this project focused on raw single aggregate images with sizes smaller than approximately 1.5 μm (128 pixels) in both height and width. A small proportion of images larger than this threshold are excluded directly. Moreover, to avoid including noise, the images of aggregate with sizes less than and equal to 4 pixels are also excluded.

The deep learning approach requires input images to be of the same size. To avoid distorting any information, instead of resizing, all raw single aggregate images were padded with a black border to 128x128 pixels. All processed single aggregate images were saved as .png files and are ready for use.

2.3 Dataset Split

The dataset was split at the patient-level for training and testing. To ensure each disease is represented consistently in the training and testing set, 80% of donors of each disease were randomly picked as the training set, and the remaining 20% were kept as the testing set. Furthermore, 1/4 of the training set is randomly split to be a validation set for model selection. Theoretically, the dataset can be split differently by selecting distinct donors as training sets, to evaluate the stability of methods. In this project, I experimented with the method pipeline in one specific split of the dataset as a proof-of-principle.

2.4 Method Pipeline

The overview of the analysis of single aggregates is shown in Figure 4. The pipeline is split into two stages, the feature extraction stage and the prediction stage. The first is to characterise each individual aggregate, and the second is to cluster aggregate structures and define distinct aggregate morphologies. The distribution of

aggregate morphologies within individual donors will be used in the proposed pipelines to eventually predict diagnosis at the patient-level.

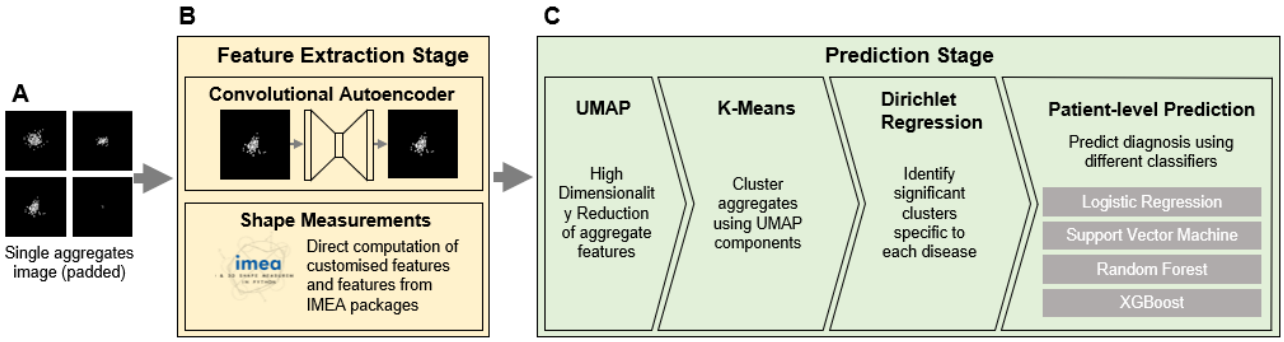


Figure 4: Overview of method pipeline. (A) SMLM images of individual aggregates padded to 128x128 pixels. (B) These images were analysed using deep learning approaches (CAE) or measured using metrics described in the IMEA package. (C) The features measured from the embedded space of the CAE or the IMEA package were used to define clusters of aggregates with similar morphologies A Dirichlet regression method identified which clusters were populated significantly differently in each disease compared to age-matched control samples. Finally, these results were used to predict if test donors were diagnosed with AD, PD, DLB or FTD.

The first stage is the feature extraction stage. Feature extraction in image analysis is a critical process that can capture the key representation of an image, i.e. features, to enhance various downstream computer vision tasks, such as classification, object detection and image segmentation. Image features often can be either manually extracted by some quantitative measurements or automatically extracted by deep learning approaches. In this stage, I used two different approaches to extract features, convolutional autoencoder (CAE) and direct computation of shape measurements.

The second stage is the prediction stage employs a series of machine learning and statistical methods to condense and assemble image features in an interpretable and intelligent way for patient-level disease prediction. I also compared the results from the CAE and the shape measurement feature extraction methods to gain insights into the differences in aggregate morphology between different diseases. In the following subsections, each method utilised is introduced and described in detail.

2.4.1 Feature Extraction Stage

a. Convolutional Autoencoder (CAE)

As mentioned in the introduction, the same protein aggregates are shared among different types of dementia. Annotating each individual aggregate image with the corresponding diagnosis of that patient may lead to inaccurate labels. This issue could impair the performance of supervised methods in feature extraction. Therefore, I choose to use a CAE as the feature extractor since its unsupervised nature has the advantage of learning image features without incorrect labels. The ability of CAE to distil complex medical images into

meaningful features without the need for labelled data provides a powerful tool for analysing and understanding patterns that may be indicative of specific diseases.

The CAE implemented in this project has architecture shown in Figure 5. The input images are the pre-processed (padded) single aggregate images with size of 128x128x1. They are black-and-white images so each image has only one channel with binary pixel values. Both encoder and decoder respectively have three convolutional layers with kernel sizes of 3x3, utilising strides of 2 and padding of 1. In the encoder, the number of filters is increased along layers (8 filters for Conv1, 16 filters for Conv2, 32 filters for Conv3) to reduce the spatial dimensions and capture more complex features. The output from the final convolutional layers is flattened into a 1D vector which is then passed through two fully connected layers to produce *h*-dimensional embedded features. Each convolutional layer is followed by batch normalisation to reduce the risk of overfitting and LeakyReLU activation function. In the decoder, the embedded features are reshaped and passed through a series of three deconvolutional layers, ultimately reconstructing the original image dimensions.

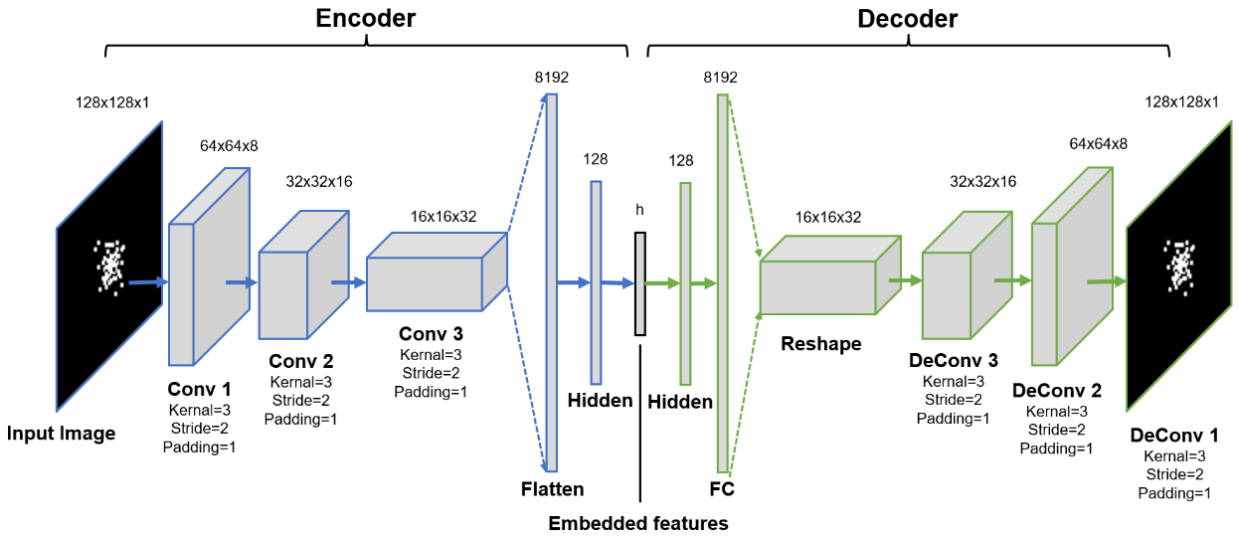


Figure 5: Model Architecture Overview: the designed CAE consists of encoder and decoder modules, and h is the compressed low dimensional embedded features of the input image. The encoder has three convolutional layers and two fully connected layers. The decoder has a mirrored deconvolution structure.

The CAE was trained over a dynamic number of epochs with a batch size of 128, using the AdamW optimiser with a learning rate of 0.001 and a Mean Squared Error (MSE) loss function. AdamW is a variant of the widely used adaptive moment estimation (Adam) optimiser. It is a stochastic gradient descent method in optimisation using first-order and second-order moments to adaptively adjust the learning rate for each parameter, where weight decay is directly applied on parameter update [36]. For efficient convergence of optimal solution and generalisation, a Reduce on Plateau learning rate scheduler is used alongside the optimiser. This scheduler would decrease the learning rate by a factor of 0.5 if the validation loss stopped decreasing for 5 epochs. Moreover, to prevent overfitting (i.e. the issue that a model performs well on the training set but is not generalised well on the validation set), a regularisation technique, early stopping, was used, with the maximal

number of epochs set to 500. Early stopping condition would not be checked for the first 5 epochs to allow some initial training, and the CAE would stop training if validation loss stopped decreasing after 10 consecutive epochs.

To choose the optimal dimensionality of embedded features, CAEs were trained with six different dimensionalities, i.e. $h \in [8, 16, 32, 64, 128, 256, 512]$. The validation MSE was used as the evaluation metric, where the lowest value represented the best performance. Subsequently, the encoder of CAE with the best dimensionality was used to extract condensed image features for the entire dataset.

b. Feature Extraction Stage – Shape Measurements

While the CAE is proven to excel at capturing complex and abstract patterns in data [37, 38], these patterns are often difficult to interpret and may overlook critical domain-specific information. In comparison, shape measurements, such as area, perimeter, and eccentricity, are more easily understood by humans. Shape measurements also have strength in efficiency and robustness, since the calculation is often computationally less intensive compared to training a CAE. Thus, in this project, shape measurement features were also obtained to compare with CAE-embedded features regarding prediction performance. Moreover, it was also used to assist in interpreting the clustering results from the CAE.

I collected 58 different shape measurements, including 5 self-defined metrics and 53 remaining from an open source Python package 'IMEA' [39]. The first 5 commonly used shape measurements, including area, solidity, eccentricity, number of branches and skeleton size, were calculated by self-defined functions based on the Scikit-image library. The definitions of 5 metrics are shown in Table 1, where both the number of branches and skeleton size are defined based on the skeletonisation of an aggregate image. Skeletonisation is a technique to preserve essential morphology while thinning the shape. The skeleton of an aggregate is obtained by interactively removing pixels from its shape's boundary until it is reduced to the minimal form [40].

The IMEA package provided the other 53 features which can be split into four categories [41]. The first category, macro descriptors, covers broad geometric features like perimeter, area, various diameter calculations, and bounding dimensions, which provide a large-scale overview of the shape's structure. The second category, meso descriptors, focuses on features such as erosions, and capturing intermediate details. The third is micro descriptors which examine finer details, including the roughness of particle contours and specific diameter measurements like Feret, Martin, and Nassenschein diameters. Lastly, the Statistical lengths section provides a rich set of measurements that capture the distribution and variation in lengths within the shape. This includes maximum, minimum, median, mean, mode, and standard deviation (std) values for various chord lengths, such as Feret diameters, Martin diameters, Nassenschein diameters, Max chords, and All chords. The full list of all shape measurement features can be found in the appendix.

Table 1: Definitions of 5 self-defined shape measurement metrics

Feature Name	Definition
Area	the total number of white pixels of an aggregate
Solidity	the ratio of the area to the area of a convex hall (i.e. the smallest polygon that aggregates region) of an aggregate, representing the density of this shape
Eccentricity	the ratio of the distance between the foci of the best-fit ellipse to its major axis length, measuring how much the shape deviates from being a perfect circle [42]
Number of branches	the number of pixels in the skeletonised aggregate that are surrounded by three or four other pixels
Skeleton size	the number of pixels of the skeleton of an aggregate

2.4.2 Prediction Stage

After the feature extraction stage, the image features are at the aggregate-level and each donor has a different number of aggregates, so to make predictions at the patient-level, it is necessary to summarise aggregate information for each donor in a more sophisticated way other than simply summing or averaging. However, since the image features obtained (either CAE-embedded features or shape measurements) are high-dimensional, before any further data manipulation, a dimensionality reduction technique, Uniform Manifold Approximation and Projection (UMAP), was applied on image features to enable the feasibility of clustering tools. K-Means clustering was employed across 20 UMAP dimensions to group aggregates with similar morphologies, and a percentage of aggregates in each cluster was reported with respect to the total number of aggregates from each donor.

Before predicting the dementia type of each donor, I used a Dirichlet regression analysis to identify clusters that are significantly differently populated within each type of dementia compared to clusters in age-matched control donors. Only these significant clusters would be used to predict the dementia type of each donor. Below are the details for the UMAP, K-means clustering, and prediction methods used in this study.

a. UMAP for Dimensionality Reduction

UMAP is a dimensionality reduction algorithm aiming to capture the underlying manifold based on the concept that assumes data lies on a lower-dimensional manifold within the high-dimensional space [43]. Compared with Principle Component Analysis (PCA), UMAP can capture non-linear relationships in the data; and it also outperforms other non-linear dimensionality reduction techniques like t-distributed Stochastic Neighbour Embedding (t-SNE) by efficiency in large datasets.

The UMAP model was implemented using the 'umap' package in Python. It was configured to reduce high-dimensional data to 20 components, with 15 neighbours considered for each point and the minimum distance between points set to 0.1. Euclidean distance metric was used to measure similarity and the model was initialised with a random starting point, seeded with a chosen random state to ensure reproducibility.

b. K-Means for Clustering of Aggregates

K-Means is a popular clustering algorithm that partitions the data into K distinct clusters by minimising the variance within each cluster, used here to group similar aggregates together [44]. The algorithm involves four key steps: initialisation, assignment, update, and iteration. In the initialisation step, K centroids were randomly selected to represent the initial positions of the clusters. Then each aggregate was assigned to clusters whose centroid was nearest in Euclidean distance. After the assignment of all aggregates, the centroids were recalculated by taking the mean of all aggregates in each cluster. The second and third steps were repeated until the centroids no longer changed significantly, and then the algorithm terminated.

During this process, the chosen number of clusters K can be tuned to reach better performance of clustering, so the most common evaluation metric, average silhouette score, was used to determine the best K. The definition of silhouette score $S(i)$ for each data point i is shown in equation (1), where $a(i)$ is the average distance between i and all other points in the same cluster, and $b(i)$ is the average distance between i and all points in the nearest cluster [45]. The average silhouette score across all data points of the dataset will give an overall measurement of the performance of a clusterer.

$$S(i) = \frac{b(i) - a(i))}{max(a(i), b(i))} \tag{1}$$

Since the complexity of computing the silhouette score for the entire dataset is $O(n^2)$, which is very computationally expensive for a large dataset, subsampling was applied for efficiency. Silhouette score was calculated for various K ranging from 10 to 150, and repeated for six different splits of the dataset to ensure stable results. The optimal solution of K with the highest silhouette score was used for final clustering. Because the large-scale dataset is too computationally intensive for K-Means, in practice, the 'MiniBatchKMeans' method was used instead of the traditional 'KMeans' method [46]. This 'MiniBatchKMeans' fitted the clusterer iteratively with one batch in size of 200 data points.

After obtaining the cluster membership of each aggregate using the best K, for each donor, the cluster percentages were calculated. The equation (2) shows the definition of cluster percentage for each donor, where C_i is the number of aggregates affiliated to cluster i, and K is the total number of clusters. So the cluster percentage P_i is the proportion of how many aggregates of one cluster out of the total number of aggregates of each donor.

$$\text{Cluster Percentage: } P_i = \frac{C_i}{\sum_{j=1}^{K} C_j} \tag{2}$$

c. Dirichlet Regression for Significant Clusters Selection

Dirichlet regression is a statistical tool used to model compositional data where all variables are proportions summing to 1 [47]. Dirichlet regression assumes the data following a Dirichlet distribution which is a high-dimensional generalisation of beta distribution. It takes each donor's cluster percentages as the high-dimensional outcome and the diagnosis as the dependent variable (one-hot encoded). The p-value reported between each type of dementia and the percentage P_i of cluster i can indicate whether there is a significant association between this cluster i and this disease. This technique is therefore particularly suitable to model the cluster percentages, which can help in determining which clusters are statistically important and relevant to the outcome of interest, allowing for targeted analysis and interpretation.

d. Patient-level Prediction

For the exploratory purpose of classifiers, four different classic machine learning algorithms were implemented, including Multinomial Logistic Regression (MLR), Support Vector Machine (SVM), Random Forest (RF) and Extreme Gradient Boosting (XGBoost) tree. The exposure variables are the cluster percentages of significant clusters identified by Dirichlet regression, and the outcome variable is the diagnosis of dementia types.

Due to the categorical nature of the outcome (types of dementia), MLR is the simplest approach to handle multiple classes. Moreover, the SVM is another powerful classification method that seeks to find the optimal hyperplane by separating the different classes in the data. There were also two tree-based methods implemented. RF is an ensemble learning method that constructs multiple decision trees and combines their predictions to improve accuracy and prevent overfitting [48]. XGBoost is an advanced gradient-boosting algorithm that builds models sequentially, optimising them iteratively to enhance prediction performance [49].

All models were configured by default, except using polynomial kernel for SVM and 'entropy' criterion for random forest. As the number of patients is different for each disease, each training data were assigned a different sample weight to handle the imbalanced dataset. For model performance evaluation, accuracy and weighted average Area Under the Curve (AUC) were used to measure overall model performance. Accuracy is the percentage of donors whose diagnoses were correctly predicted. AUC is the area under the Receiver-operating characteristic (ROC) curve which plots the true-positive rates against the false-positive rates. Since this is a multiclassification problem with an imbalanced dataset (i.e. different number of patients for each disease), the weighted average AUC was calculated using the following equation (3), where AUC_i is the one-vs-rest AUC for class i and Wi is the proportion class i in the test set. In addition, precision, recall, f1 score were also reported to measure model performance for each type of dementia.

$$\text{Weighted Average AUC} = \sum_{i=1}^{5} AUC_i \times W_i \tag{3}$$

3. Results

3.1 Dataset Overview

The dataset provided by the Ye lab included images of aggregates from patients diagnosed with four different types of dementia, PD, DLB, FTD, AD, and healthy donors as a control group. Overall, 41 donors' data were collected from their post-mortem brain tissues. More specifically, there are 10 healthy controls, 10 PD patients, 5 DLB patients, 8 FTD patients and 8 AD patients. I pre-processed the wide field SMLM images to crop areas containing individual aggregates using a MATLAB script (written by the Ye lab) [18]. After pre-processing, single aggregate images were padded so that all images were the same size of 128x128 pixels (examples shown in Figure 6).

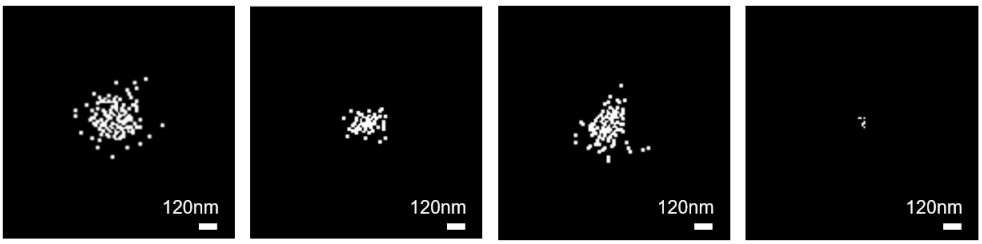


Figure 6: Examples of pre-processed single aggregation images. Aggregates were imaged using SMLM, producing binary images of pixels recording the presence of an amyloidogenic-fluorescent dye. These Images of single aggregates were cropped from the reconstructed super-resolved images.

After the exclusion criterion (I excluded images larger than 128x128, and images with the number of white pixels smaller than 4), there were a total of 1,328,812 images of single aggregates: 125,205 images from control, 147,638 images from PD, 201,400 images for DLB, 645,898 images for FTD and 208,671 images for AD.

The donors selected for training, validation and testing sets can be found in A.Table 1(A/B/C) in Appendix. These tables also show the number of aggregates identified for each donor and demographic characteristics, gender and age.

3.2 Feature Extraction Stage

3.2.1 CAE Training and the Embedded Features

Aiming to extract feature images using a deep learning approach, I used CAE to extract image features in an unsupervised manner. I trained six CAE models with increasing dimensions of the embedded space (embedded feature dimension = 4, 8, 16, 32, 64, 128, 256, 512), and used the mean squared error (MSE) from each model to choose which model to use for further investigation. Figure 7A displays the plot of MSE against different

dimensionality of the embedded features of CAE. The CAE with the lowest MSE was the model with embedded features in the dimension of 256. Each model was trained over a different number of epochs due to the early stopping technique to avoid overfitting. The 256 model was terminated during the training process at 127 epochs, after no improvement of validation loss over 10 epochs. The model's learning performance is shown in Figures 7B and 7C which demonstrate the training and validation loss (MSE) curve. The consistent decrease in training loss indicated that the model was effectively learning from the training data, while the continuous downward trend in validation loss suggested that the model generalised well on unseen data, without signs of overfitting.

To further visualise how well the CAE can summarise image information in embedded features, I checked the output single aggregate images from the decoder part of the CAE model and compared them with their corresponding input images used by the encoder. The model demonstrates a strong ability to generate accurate output images, particularly excelling at capturing the details of smaller aggregates, as illustrated in Figures 7D and 7E.

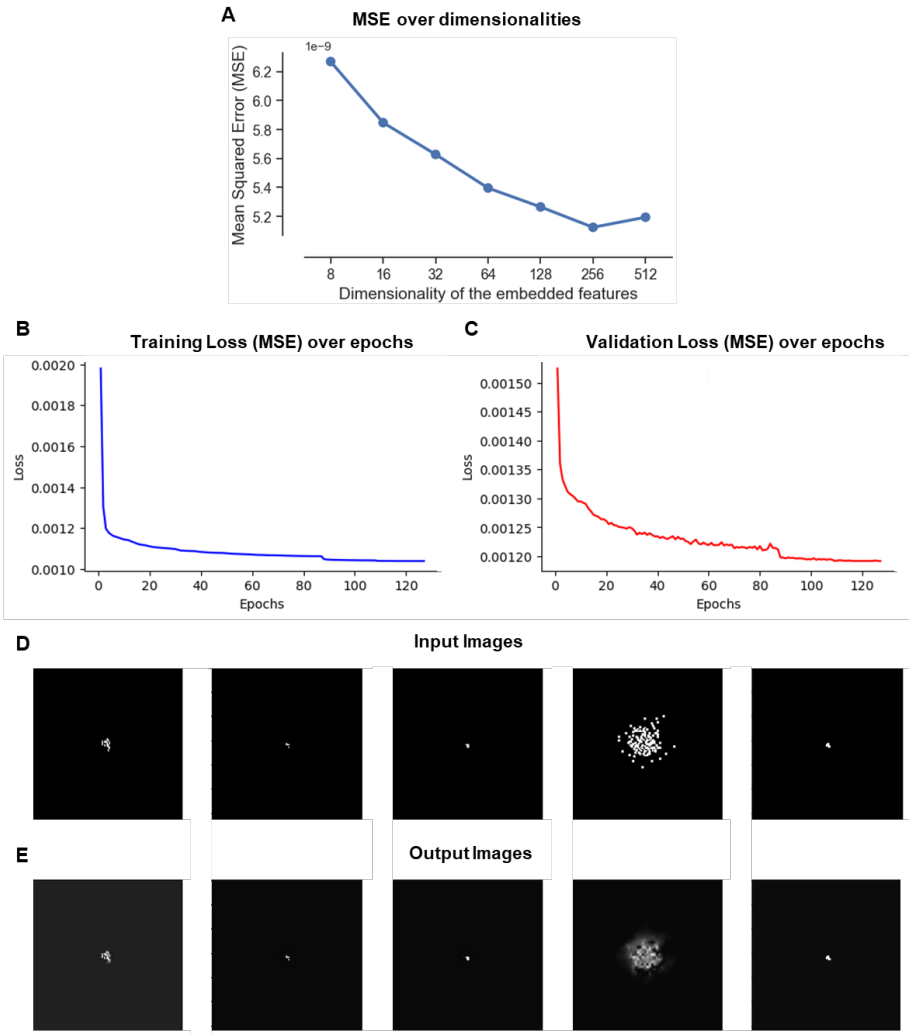


Figure 7: The plot of (A) MSE against different dimensionality of the embedded features of CAE, (B) CAE training loss curve and (C) validation loss curve. Examples of (D) input single aggregate images and (E) output images of CAE.

3.2.2 Shape Measurement Features

Extracting aggregate image features using deep learning approaches is useful as these methods comprehensively analyse images by capturing more variance between images missed by analogue methods. However, deep learning techniques are also famously hard to interpret and have been previously called "black-box" methods. Thus, I sought to compute shape measurements of each single aggregate image to generate more interpretable and robust results. The 'IMEA' analysis used describes 53 shape measurement features and I also measured 5 additional features used previously by the Ye Lab to characterise aggregate morphology. Figure 8 shows the distribution of a few selected shape features for different types of dementia as an example. Among all features obtained from IMEA packages, there are two features having missing values, so they were excluded for the following analysis (i.e. resulting in 56 features left; the excluded features are noted in A.Table 2 shown in the Appendix). They are 'fractal dimension boxcounting method' and 'allchords min'. For convenience, I will use IMEA to refer to these shape features in the following paragraph.

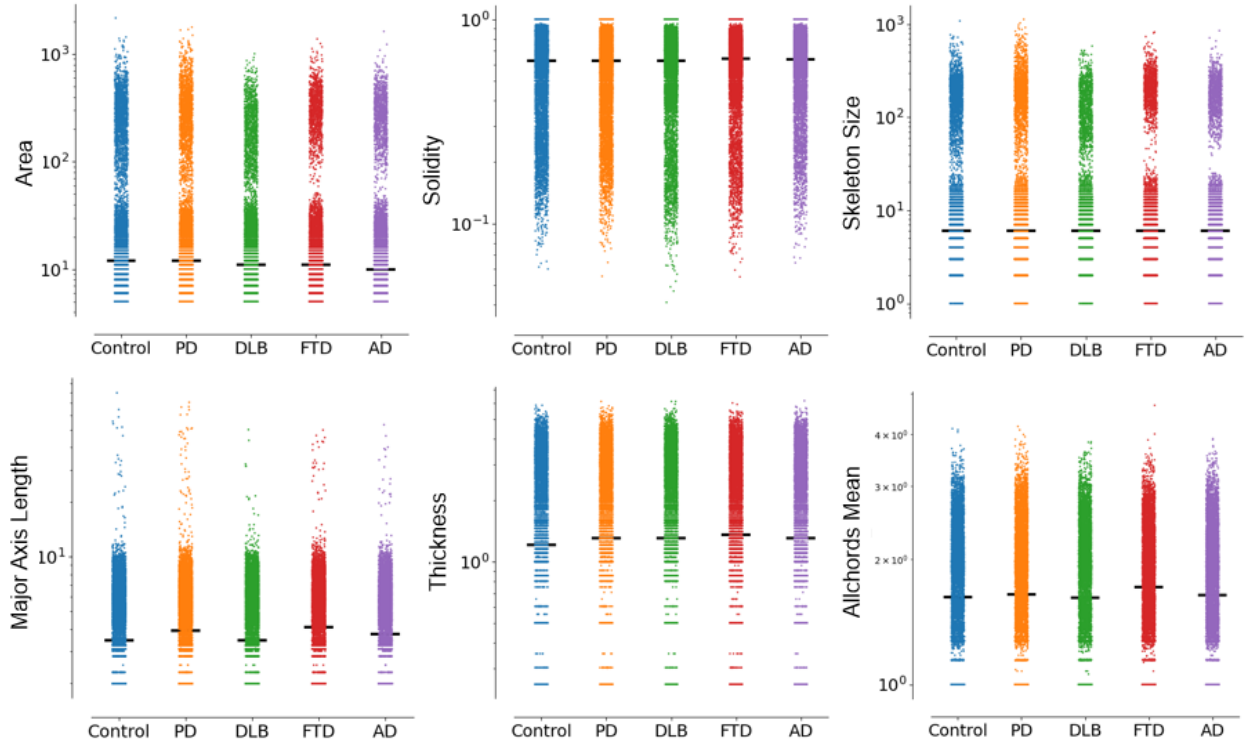


Figure 8: Plots to show the distribution of a few selected shape features for each type of dementia (shaded by different colours). The median value of each shape feature is marked by a black bar for each disease.

3.3 Prediction Stage

3.3.1 Dimensionality reduction using UMAP

After extracting features using both approaches, the dimensionalities of both types of features are too computationally expensive for the downstream task, K-Mean clustering (256 dimensional CAE features and 56 dimensional for IMEA features). Thus, UMAP was used to further condense the feature information to only

20 dimensions, making it feasible to cluster aggregates using K-Means. Figure 9 shows a visualisation of the first and second UMAP components computed from CAE-embedded features and IMEA features.

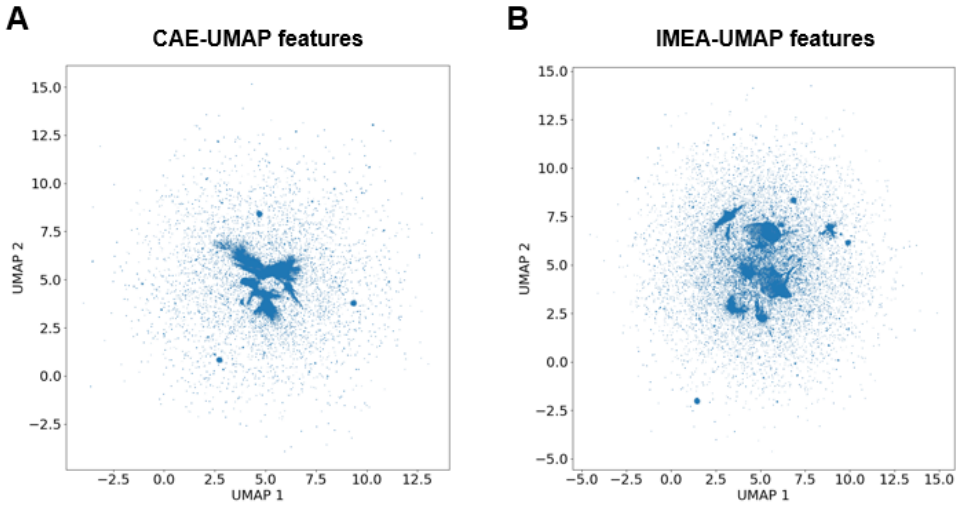


Figure 9: Visualisation of the first (UMAP1) and second (UMAP2) UMAP components for different types of features. (A) UMAP visualisation of CAE-embedded features. (B) UMAP visualisation of IMEA features.

3.3.2. Clustering aggregate morphologies using K-Means

Aggregates of similar morphologies were identified using K-Means clustering. The optimal choice of the number of clusters, K, were determined using the average silhouette score to evaluate overall model performance in separating aggregates. K-Means models were applied to six different splits of the dataset (leave-one-out split), and then models were trained for different K ranging from 10 to 150 with a stepsize of 10. The UMAP condensed CAE and IMEA features were used to train these clusterers. Figures 10A and 10C show the average silhouette score variation for the CAE pipeline and IMEA pipeline respectively, with error bars computed for each K. Then a quadratic line was fitted for each pipeline's results to solve the optimal solution of K.

For the IMEA pipeline, the fitted quadratic function is $-1.4\times10^{-5}x^2 + 0.0026x + 0.083 = 0$ where K=96.3 is the optimal solution (shown in Figure 10D). However, for the CAE pipeline, the fitted quadratic function is $-5.7\times10^{-6}x^2 + -0.00037x + 0.15 = 0$ whose optimal solution K= -31.9 is out of the feasible range(shown in Figure 10B), so the optimal choice of K was determined based on Figure 10A. Therefore, K=40 and K=96 were used to train K-Mean models for the CAE pipeline and IMEA pipeline respectively. The K-Mean models were trained on the training set and then used to predict cluster membership for the whole dataset (both training and testing set).

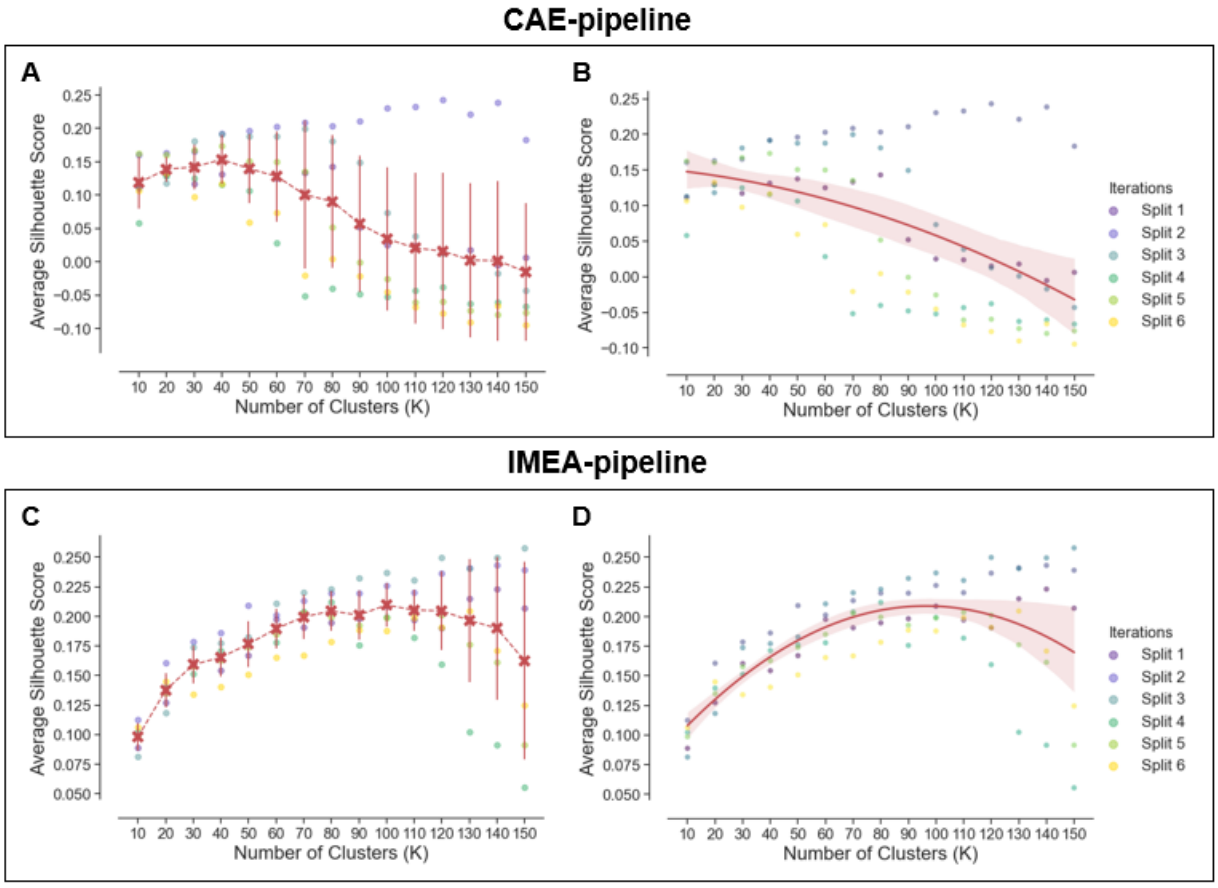


Figure 10: Evaluation of clustering performance using silhouette scores across different numbers of clusters (K) and feature types, CAE-embedded features (Panels A and B) and Shape Measurements (Panels C and D). Panels A and C show the average silhouette scores with error bars across six data splits. Panels B and D present the fitted quadratic lines with confidence intervals for the silhouette scores across the same range of clusters.

3.3.3. Identifying significant clusters using Dirichlet regression

The distributions of aggregates in each cluster for individual donors were passed to Dirichlet regression to identify the significant clusters for each disease. Figures 11A and 11B show the significant clusters selected for the CAE pipeline and IMEA pipeline respectively. The significant clusters denote those that have statistically different distributions (p-values<0.05) of aggregates in each cluster for that disease compared to age-matched controls.

In total, there are 12 significant clusters from the CAE pipeline and 15 significant clusters for the IMEA pipeline across all diseases. The total aggregate counts in each cluster for each disease and the average cluster percentage of each disease were measured and are shown in Figure 11. The percentage differences of each cluster from donors diagnosed with dementia compared to age-matched controls are colour-coded to denote if that type of dementia typically has a higher (blue) or lower (red) proportion of aggregates in that cluster compared to controls. The CAE pipeline was able to identify significant clusters for PD, DLB and AD, while the IMEA pipeline identified clusters significant for DLB, FTD and AD. I also computed the total aggregate counts for each cluster to identify if the significant clusters were rare clusters. The results show that all

significant clusters are rather larger clusters: for the CAE pipeline, all significant clusters are within the top 20; and for the IMEA pipeline all significant clusters are within the top 40.

Since this project aims to understand the difference in morphology of aggregates that are specific to different diseases, I also used the 53 IMEA features to assist in the interpretation of the characteristics of clusters. The mean value of each IMEA feature for all significant clusters was computed and visualised in radar plots to provide an easier interpretation (shown in Figure 12). By comparing the shape of each cluster, some significant clusters were grouped together based on the observed similarities.

There are five grouped clusters (a-e) shown in Figure 12A which were identified for clusters obtained from the CAE pipeline, and there are six grouped clusters (a-f) shown in Figure 12B identified from clusters obtained from the IMEA pipeline. Clusters belonging to the same grouped cluster were set in the same series of colours for better recognition. It can be observed that different grouped clusters present very distinct characteristics by their polygon shape. Moreover, both feature extraction approaches result in almost the same grouped clusters, except the IMEA pipeline separated one more grouped cluster (f) than the CAE pipeline. Some grouped clusters were consistently identified and selected as significant clusters for specific dementia types. For instance, Cluster (e) was related to FTD, while Cluster (c) was linked to both DLB and AD. Nevertheless, some clusters displayed variable specificity. Cluster (a) was associated with either PD or FTD, while Clusters (b) and (d) were relevant to either PD or DLB.

A

CAE-pipeline with K=40

	Aggregate Counts					Cluster Percentage					Percentage Difference				Total Aggregate Counts
	CTRL	PD	DLB	FTD	AD	CTRL	PD	DLB	FTD	AD	PD	DLB	FTD	AD	
Cluster 3	16,394	19,618	16,937	66,396	15,056	15.64%	16.22%	9.79%	11.36%	9.66%	0.59%				166,582
Cluster 1	4,885	5,283	8,440	31,191	8,514	4.66%	4.37%	4.88%	5.34%	5.46%	-0.29%			0.81%	66,408
Cluster 4	5,663	6,710	8,890	37,623	9,206	5.40%	5.55%	5.14%	6.44%	5.91%	0.15%				78,018
Cluster 5	11,886	13,931	21,047	65,510	18,538	11.34%	11.52%	12.17%	11.21%	11.90%	0.18%				152,932
Cluster 32	5,470	6,110	8,822	36,549	9,315	5.22%	5.05%	5.10%	6.25%	5.98%	-0.16%				75,606
Cluster 10	1,860	1,387	3,596	11,425	3,628	1.77%	1.15%	2.08%	1.95%	2.33%		0.30%		0.55%	23,880
Cluster 19	1,973	1,447	3,770	11,896	3,748	1.88%	1.20%	2.18%	2.04%	2.41%	-0.69%	0.30%		0.52%	24,748
Cluster 2	8,446	9,502	14,540	49,450	12,957	8.06%	7.86%	8.41%	8.46%	8.32%	-0.20%				110,523
Cluster 8	13,705	16,457	24,036	71,175	19,937	13.07%	13.61%	13.90%	12.18%	12.80%	0.54%				171,558
Cluster 14	15,090	18,241	28,575	81,372	22,383	14.39%	15.09%	16.52%	13.92%	14.37%	0.69%				195,344
Cluster 35	10,951	13,818	18,773	69,037	16,969	10.45%	11.43%	10.85%	11.81%	10.89%	0.98%				150,064
Cluster 26	1,749	1,342	3,517	11,037	3,439	1.67%	1.11%	2.03%	1.89%	2.21%		0.37%			22,827

B

IMEA-pipeline with K=96

	Aggregate Counts					Cluster Percentage					Percentage Difference				Total Aggregate Counts
	CTRL	PD	DLB	FTD	AD	CTRL	PD	DLB	FTD	AD	PD	DLB	FTD	AD	
Cluster 5	7,895	8,252	6,076	30,472	10,072	7.36%	7.94%	5.24%	6.28%	5.83%			-1.08%		79,106
Cluster 8	7,064	7,727	4,936	27,137	8,507	6.59%	7.43%	4.25%	5.59%	4.92%			-1.00%		80,074
Cluster 30	5,225	5,725	3,825	18,450	6,327	4.87%	5.51%	3.30%	3.80%	3.66%			-1.07%		43,964
Cluster 83	5,976	6,234	4,692	22,087	7,673	5.57%	6.00%	4.04%	4.55%	4.44%			-1.02%		46,937
Cluster 6	3,152	2,901	4,073	15,221	5,683	2.94%	2.79%	3.51%	3.14%	3.29%		0.57%			36,262
Cluster 7	1,638	1,603	1,909	7,013	2,602	1.53%	1.54%	1.64%	1.45%	1.50%		0.12%			22,545
Cluster 13	1,669	1,554	2,229	8,096	3,113	1.56%	1.49%	1.92%	1.67%	1.80%		0.36%			35,051
Cluster 23	256	211	403	1,866	625	0.24%	0.20%	0.35%	0.38%	0.36%		0.11%			4,000
Cluster 12	1,851	1,247	3,535	9,484	3,887	1.73%	1.20%	3.05%	1.95%	2.25%		1.32%		0.52%	20,574
Cluster 14	333	237	571	1,688	658	0.31%	0.23%	0.49%	0.35%	0.38%		0.18%			4,093
Cluster 18	1,790	1,254	3,435	9,155	3,827	1.67%	1.21%	2.96%	1.89%	2.21%		1.29%		0.54%	43,580
Cluster 2	8,352	7,919	9,538	45,281	15,393	7.79%	7.62%	8.22%	9.33%	8.90%		0.43%			105,496
Cluster 39	6,954	6,615	7,742	34,379	12,049	6.49%	6.36%	6.67%	7.09%	6.97%		0.18%			98,205
Cluster 20	5,310	5,453	5,086	25,665	8,157	4.95%	5.25%	4.38%	5.29%	4.72%			0.34%		54,690
Cluster 11	1,694	1,210	3,406	9,019	3,647	1.58%	1.16%	2.93%	1.86%	2.11%		1.35%		0.53%	23,531

Figure 11: Table of cluster aggregate counts, cluster percentage and cluster percentage difference across different disease classes for clusters obtained from (A) CAE pipeline and (B) IMEA pipeline respectively.

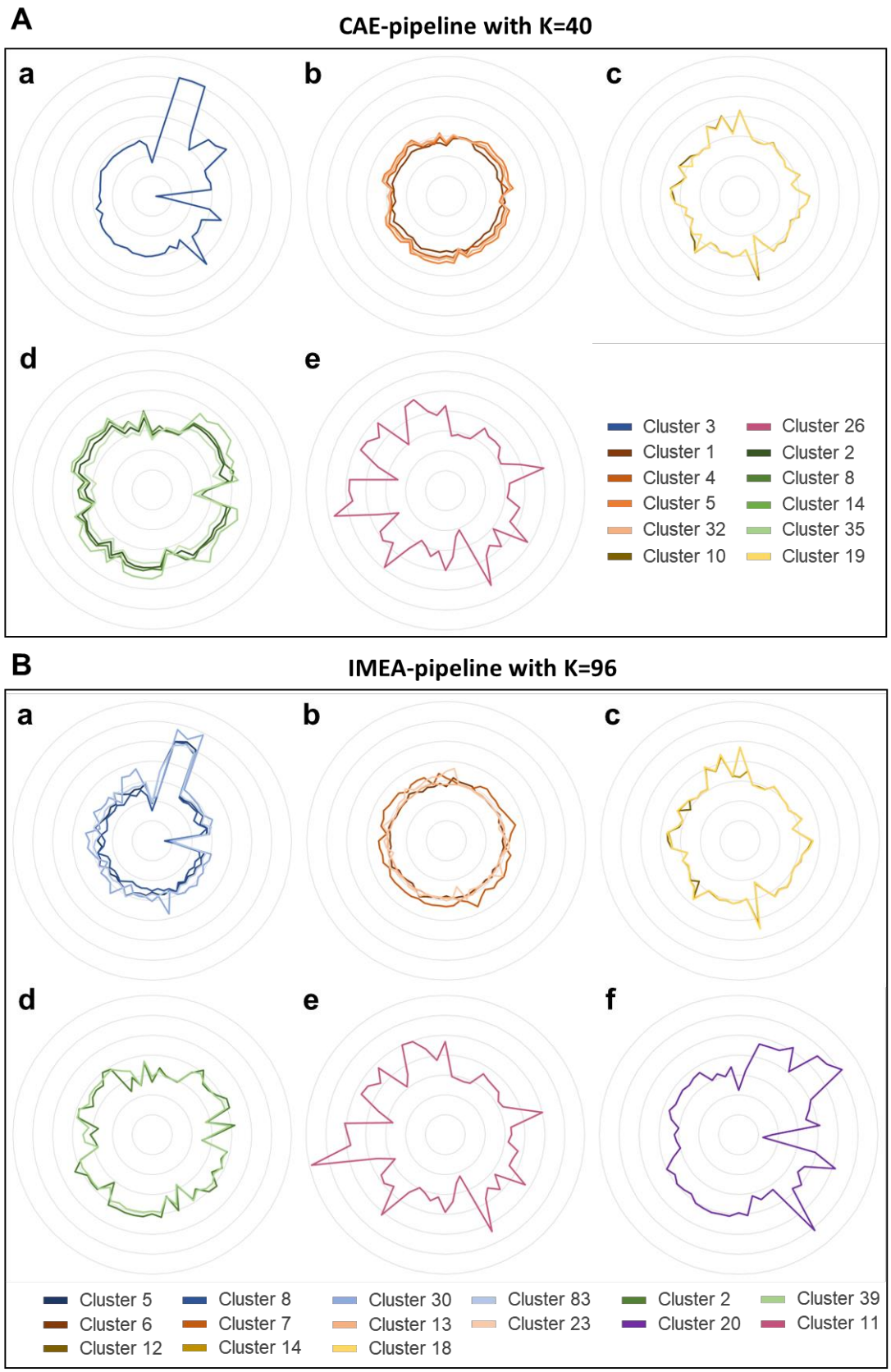


Figure 1: Radar plots of cluster characteristics for clusters obtain from (A) CAE and (B) IMEA pipeline. The axis labels of IMEA feature are hidden for better visualisation, and the detailed values and corresponding IMEA labels can be found in the heatmap in Appendix A.Figure 1 and A.Figure 2.

3.3.4 Prediction-level Prediction

After selecting significant clusters, the cluster percentage of these significant clusters from donors in the training set was used to train four different classifiers: Multinomial logistic regression (MLR), Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) tree. Then different evaluation metrics were computed on the testing set. Figure 13 shows the overall accuracy and weighted average AUC measured for the evaluation of these models' overall performance. Since there are five classes (i.e. four types of dementia and healthy control), 0.25 of accuracy and 0.5 of AUC are the reference performance, which means random guessing.

When comparing the models based on accuracy (shown in Figure 13A), for the IMEA-pipeline, XGBoost emerges as the top performer, achieving the highest accuracy of 0.56, followed by RF which shows competitive accuracy of 0.50. For CAE-pipeline, RF becomes the best model with an accuracy of 0.44, while XGBoost presents the worst performance across all four models where the accuracy of 0.19 is even lower than random guessing. SVM and MLR demonstrate poorer accuracy across both pipelines, with MLR consistently performing only at 0.25 for both pipelines.

In terms of Weighted Average AUC (shown in Figure 13B), XGBoost again leads with the highest AUC value of 0.72 when using the IMEA-pipeline, indicating that it not only excels in accuracy but also in distinguishing between classes. Random Forest also performs well in terms of AUC, particularly with the IMEA-pipeline (0.68), showing that it is a more robust model across both metrics. SVM and MLR show similar AUC performance, with values hovering around 0.55 to 0.60, indicating moderate performance.

When comparing the two pipelines, the IMEA-pipeline consistently outperforms the CAE-pipeline across the more complex models, i.e. XGBoost and Random Forest, where it shows significant improvements in both accuracy and AUC.

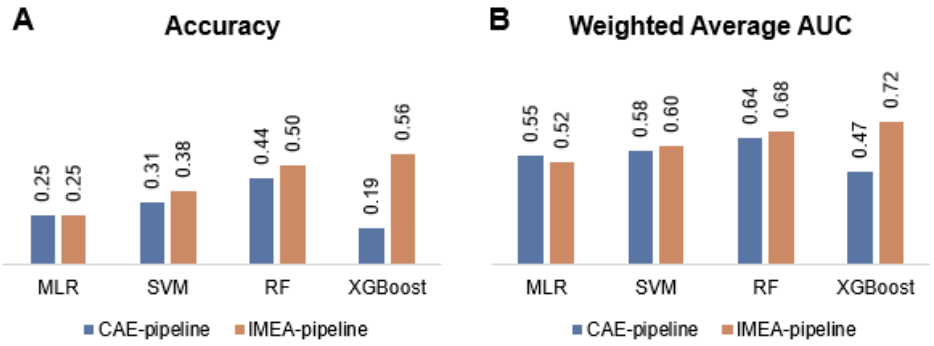


Figure 13: Performance comparison of (A)Accuracy and (B)Weighted AUC for MLR, SVM, RF and XGBoost, using CAE-embedded features and Shape Measurements.

To assess these four models' performance in each class, the precision, recall, f1 and the One-vs-rest Prediction AUC were computed (shown in Table 2). One-vs-rest AUC was also plotted in bar plots shown in Figure 14.

In the CAE pipeline, the overall best model, random forest, presents performance particularly well on PD, DLB and FTD, achieving One-vs-rest AUCs 0.83, 0.71 and 0.63 respectively, but it performs moderately in healthy control and worse than random guessing in AD. On the other hand, in the IMEA pipeline, the overall best mode, XGBoost, reaches 0.88 for the control group, and 0.63 to 0.75 for FTD, AD and PD, indicating strong predictive capability.

However, the precision of 0.00 and One-vs-rest AUC of 0.5 reveal some performance issues that there are some types of dementia never predicted by models in both pipelines. For example, for both pipelines, control, FTD and AD were never predicted by MLR and FTD was never predicted by SVM. Also, DLB was never predicted by random forest and XGBoost in the IMEA pipeline; while FTD was never predicted by XGBoost in the CAE pipeline.

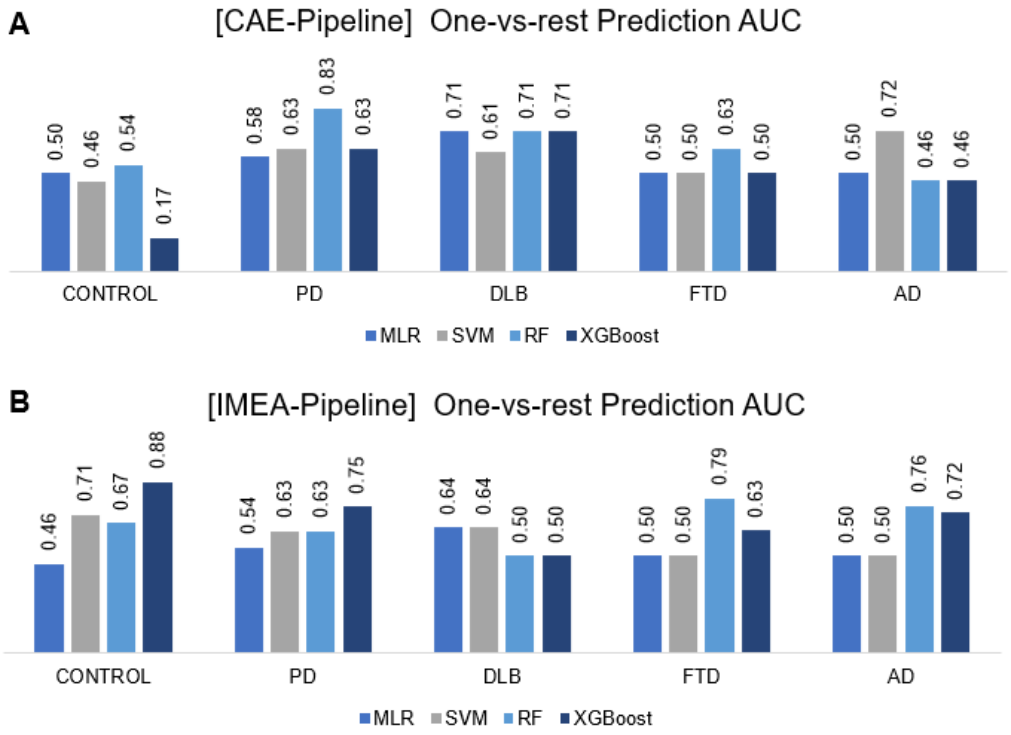


Figure 14: Comparison of One-vs-Rest Prediction AUC across different models for (A)CAE-pipeline and (B)IMEA-pipeline

Table 2: Comparison of classification metrics for CAE-Pipeline and Shape Measurements-Pipeline across different models and *disease classes.*

CAE-Pipeline

Multinomial Logistic Regression

	Precision	Recall	F1	one-vs-rest AUC
CONTROL	0.00	0.00	0.00	0.50
PD	0.33	0.50	0.40	0.58
DLB	0.20	1.00	0.33	0.71
FTD	0.00	0.00	0.00	0.50
AD	0.00	0.00	0.00	0.50

Support Vector Machine

	Precision	Recall	F1	one-vs-rest AUC
CONTROL	0.20	0.25	0.22	0.46
PD	1.00	0.25	0.40	0.63
DLB	0.20	0.50	0.29	0.61
FTD	0.00	0.00	0.00	0.50
AD	0.40	0.67	0.50	0.72

Random Forest

	Precision	Recall	F1	one-vs-rest AUC
CONTROL	0.29	0.50	0.36	0.54
PD	0.75	0.75	0.75	0.83
DLB	0.50	0.50	0.50	0.71
FTD	0.50	0.33	0.40	0.63
AD	0.00	0.00	0.00	0.46

XGBoost

	Precision	Recall	F1	one-vs-rest AUC
CONTROL	0.00	0.00	0.00	0.17
PD	0.40	0.50	0.44	0.63
DLB	0.50	0.50	0.50	0.71
FTD	0.00	0.00	0.00	0.50
AD	0.00	0.00	0.00	0.46

IMEA-Pipeline

Multinomial Logistic Regression

	Precision	Recall	F1	one-vs-rest AUC
CONTROL	0.00	0.00	0.00	0.46
PD	0.27	0.75	0.40	0.54
DLB	0.25	0.50	0.33	0.64
FTD	0.00	0.00	0.00	0.50
AD	0.00	0.00	0.00	0.50

Support Vector Machine

	Precision	Recall	F1	one-vs-rest AUC
CONTROL	0.36	1.00	0.53	0.71
PD	1.00	0.25	0.40	0.63
DLB	0.25	0.50	0.33	0.64
FTD	0.00	0.00	0.00	0.50
AD	0.00	0.00	0.00	0.50

Random Forest

	Precision	Recall	F1	one-vs-rest AUC
CONTROL	0.50	0.50	0.50	0.67
PD	0.40	0.50	0.44	0.63
DLB	0.00	0.00	0.00	0.50
FTD	0.67	0.67	0.67	0.79
AD	0.50	0.67	0.57	0.76

XGBoost

	Precision	Recall	F1	one-vs-rest AUC
CONTROL	1.00	0.75	0.86	0.88
PD	0.50	0.75	0.60	0.75
DLB	0.00	0.00	0.00	0.50
FTD	0.50	0.33	0.40	0.63
AD	0.40	0.67	0.50	0.72

4. Discussion

Dementia is a significant global public health concern due to its rapidly increasing prevalence, especially in ageing populations. It places a heavy burden on not only individual families but the whole society and the healthcare system, emphasising the urgency to develop early diagnostic tools, preventive strategies, and improved care to mitigate its impact worldwide.

By leveraging a combination of machine learning and deep learning approaches, the pipeline outlined in this project represents a critical advancement in not only predicting the diagnosis of dementia but also in differentiating between various types of dementia, which is crucial for tailored interventions and treatments.

4.1 Interpretation of Results

The advantage of a series of different models was considered when I was developing a comprehensive pipeline to diagnose dementia using aggregate images, successfully overcoming potential mislabelling challenges, and optimising the training process for efficiency. To leverage the information from aggregate images, the pipeline built in this project first intended to extract aggregate image features, which would then be put into a prediction stage that includes a series of models designed to identify dementia in patients. However, because some aggregates may co-exist across different diseases and single aggregate images were labelled according to the diagnosis of the patients from whom they were derived, a potential mislabelling issue arose. This could possibly explain the poor performance of the other supervised models I initially attempted to extract image features. The unsupervised nature of the CAE offered a major advantage in overcoming this difficulty by allowing the extraction of image features without the need for labels, while making use of neural networks' capacity to extract intricate information from images. In terms of training efficiency, the original CAE scripts used in previous work on this topic required extensive training times, spanning several days, due to the large scale of the dataloader. These scripts were optimised by integrating both the dataloader and training module into a single script, significantly reducing the training time to just several hours. Moreover, the practical implementation of Mini-batch K-Means and subsampling of silhouette score computation made it feasible to efficiently cluster millions of data points within half an hour, allowing for extensive experimentations to determine the optimal number of clusters (K).

Furthermore, each stage of the models was customised and evaluated using relevant performance metrics to fine-tune the models, resulting in a more accurate and dependable pipeline. In the first feature extraction stage, through experiments with varying dimensionalities of the embedded features in the CAE, the optimal model was identified, effectively extracting image features in 256 dimensions while maintaining high-quality reconstruction in single aggregate images. In the second prediction stage, the dimensionality of the extracted features was reduced using UMAP, which facilitates the clustering method by addressing the computational challenges posed by very high-dimensional data. For K-Means methods designed to define clusters of morphologically distinct aggregates, the optimal number of clusters (K) ranging from 10 to 150 was identified

according to its average silhouette score. Clustering enables the aggregation of single aggregate information at the patient-level by summarising it as cluster percentages, representing the proportion of each cluster among all aggregates. This approach provides a more intelligent way to profile patients based on their aggregate characteristics, thereby facilitating patient-level prediction. The Dirichlet method was employed to select disease-specific clusters of aggregates, successfully identifying clusters that were significant for each disease. Additionally, shape measurement features were utilised to characterise these clusters, revealing inherent differences that are more interpretable. This enabled the grouping of algorithmically derived clusters into more meaningful clusters.

Recognising that the most commonly used MLR typically performs moderately as a baseline classification model, several other classic machine learning approaches, SVM, RF, and XGBoost, were preliminarily used to explore more competitive models and optimise predictive ability. By reviewing the performance of different classification models, tree-based methods, RF and XGBoost, tended to show superior results. XGBoost, which performed best in the IMEA pipeline but worst in the CAE pipeline, suggests that while there may be opportunities to enhance performance, it could also lead to increased instability. In contrast, RF demonstrated more stable overall performance, ranking highest for the CAE pipeline and second best for the IMEA pipeline, emerging as the most reliable model among all those explored. Previous exploration conducted on the same topic in the Ye lab mainly focused on aggregate-level prediction [50, 51], so to make the computational model eventually scalable in the real clinical setting, the dataset was split at the patient-level with some patients holding out as the testing set to mimic true scenarios. Thus, these results successfully demonstrate the technical feasibility of the proposed computational pipeline and its future potential deployment of this pipeline as a dementia diagnostic tool.

Beyond the deep learning approach, I also conducted a comparative study to evaluate two pipelines differentiated in feature extraction methods. When comparing the two pipelines, it was initially hypothesised that the CAE pipeline might outperform the IMEA pipeline because of its ability to extract complex features from images; however, both pipelines demonstrated comparable prediction accuracy and weight average AUC. Moreover, both consistently assigned aggregates into clusters with similar characteristics, which were identified as significant to different types of dementia. This finding strongly supports the hypothesis that specific subsets of aggregates are relevant to respective types of dementia.

One significance of this project is its novelty in addressing the research gaps in the characterisation of aggregate morphology. Prior studies have concentrated on the molecular architecture of protein aggregates, primarily using cryo-electron microscopy (cryo-EM) to examine their conformation, such as the internal structure of α-synuclein or tau filaments [52, 53]. However, these studies often neglected or used experimental extraction methods which potentially disturbed higher-order structural features (i.e. morphology), such as whether they are fibrous, spherical, sheet-like, or amorphous structures. Furthermore, while fluorescence microscopy and atomic force microscopy (AFM) are useful, they fail to capture the complete complexity of

aggregate morphology, limiting the extent of morphological-contextual insights [54]. This project overcomes these limitations by employing advanced image analysis methods to focus on larger ultra-complex morphological molecular features.

4.2 Limitations and Future Works

Despite the novel pipeline developed in this project, several limitations should be acknowledged, followed by corresponding recommendations for future works.

Though acquiring 41 donors is noteworthy in this domain, machine learning methods often need a larger dataset to achieve acceptable accuracy and avoid overfitting. The limited sample size results in both the validation and testing sets together containing 16 individuals, with the smallest group, DLB, having two patients. Consequently, some of the extreme performance metrics observed may be attributed to these particular circumstances. With more patient or donor samples collected in the future, more robust prediction performance could be generated by replicating this pipeline on a larger dataset. In addition, more comprehensive experiments involving multiple dataset splits should be conducted to replicate this pipeline and assess the robustness and generalisation of the predictive performance.

Another area for improvement lies in the amount of information contained within the aggregate images. Currently, the images are binary (black-and-white), which may limit the data's richness. Creating non-binary images with pixel intensity or coloured images with three channels to convey more information could increase analytical depth and lead to better application of advanced computational models.

The CAE training stage also presents the potential for further optimisation. To efficiently explore the entire pipeline and streamline the training process, data augmentation was not performed due to the already substantial size of the dataset, leading to insufficient analysis of rotational sensitivity. Since images of aggregates were captured at fixed angles by the microscope, it is important to ensure that morphologically identical aggregates, even when rotated or flipped, are recognised as the same. Future work could enhance this aspect by incorporating data augmentation techniques. Furthermore, the imbalance in the dataset, particularly the larger number of images of smaller aggregates, has resulted in the CAE performing better on these small aggregates. There is scope to further optimise the CAE by adjusting the loss function, which could improve reconstruction performance across all aggregate sizes.

In terms of the prediction stage, the current classification models, which rely on conventional machine learning classifiers using significant cluster percentages, could be extended to a multi-modal model. A multi-modal model, which integrates various types of data, such as aggregate composition and patient information (e.g., age, gender, lifestyle), might enhance predictive accuracy. Additionally, alternatives to Dirichlet and K-Means clustering could be explored to optimise the entire pipeline.

Finally, while the project successfully identified certain grouped clusters that are consistently significant for particular types of dementia, other clusters were found to be associated with multiple types. This finding underscores the need for further analysis to deepen the understanding of aggregates that are specific to more than one cluster, potentially revealing novel insights into the pathology of these diseases.

Overall, the insights gained from this project lay a strong foundation for continued advancement in identifying dementia types by aggregate morphologies. By tackling these limitations, subsequent research can expand on present discoveries to improve the model's strength and adaptability. This effort aims to create a diagnostic instrument to aid in the clinical detection of diseases.

4.3 Conclusion

In conclusion, as dementia becomes more prevalent in the population, the demand for a high-throughput, stable, and accurate tool for the early diagnosis of dementia has become increasingly pressing. This project builds a critical hypothesis that there exist morphologically distinct subsets of aggregates specific to different dementia types. Based on the SMLM imaging technique developed in the Ye Lab, single aggregate images were obtained to pave the opportunities to predict different types of dementia based on aggregate morphological differences.

This project successfully employed a series of machine learning and deep learning approaches to establish a comprehensive end-to-end pipeline for dementia prediction. Among all models exploited, random forest with weighted average AUC of 0.64 and 0.68 in two pipelines demonstrates the potential and feasibility of this framework in clinical application. Moreover, subsets of aggregates specific to different diseases were well identified by statistical analyses, and further quantified by shape measurements. The observed resemblance in characteristics of these aggregate subsets in both pipelines indicates the stability of this aggregate identification method. Although the findings are preliminary and the study faces several limitations, the results underscore significant potential for optimising the pipeline. Such optimisations are expected to culminate in a more comprehensive diagnostic tool, with future possible applications aimed at facilitating high-throughput testing of cerebrospinal fluid (CSF) and plasma samples, thereby enhancing early diagnosis of various dementia types in large populations.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Yu Ye, for his invaluable guidance, support, and encouragement throughout this exciting project. His expertise and insight have been instrumental in shaping my research, and I am deeply appreciative of the opportunity to learn under his supervision.

I also wish to extend my sincere thanks to Dr. Michael Morten and Hailey Gu for their unwavering support, detailed feedback, and constant encouragement which have greatly enhanced my work. I am truly grateful for the time and effort you both have invested in my project.

Furthermore, I would like to thank the entire team at the Ye Lab. The lab's rich resources and excellent research environment have been pivotal to my success. The knowledge and skills I have gained during my time here have profoundly contributed to both my personal and professional development.

Thank you all for making this an invaluable chapter in my academic journey and for providing a solid foundation for my future career.

References

[1] Noto S. Perspectives on Aging and Quality of Life. Healthcare. 2023;11(15):2131.

[2] Organization WH. Dementia [https://www.who.int/news-room/fact-sheets/detail/dementia [Accessed:10th August 2024]

[3] Clinic C. Dementia: Symptoms, Types, Causes, Treatment & Risk Factors 2022 [https://my.clevelandclinic.org/health/diseases/9170-dementia [Accessed:10th August 2024]

[4] (ADI) AsDI. ADI - Dementia statistics [https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/ [Accessed:10th August 2024]

[5] Choices N. What is dementia [https://www.nhs.uk/conditions/dementia/about-dementia/what-is-dementia/ [Accessed:10th August 2024]

[6] Olivari BS, French ME, McGuire LC. The public health road map to respond to the growing dementia crisis. Innovation in aging. 2020;4(1):igz043.

[7] Chouliaras L, O'Brien JT. The use of neuroimaging techniques in the early and differential diagnosis of dementia. Mol Psychiatry. 2023;28(10):4084-97.

[8] Bayer AJ. The role of biomarkers and imaging in the clinical diagnosis of dementia. Age Ageing. 2018;47(5):641-3.

[9] Blanco K, Salcidua S, Orellana P, Sauma-Perez T, Leon T, Steinmetz LCL, et al. Systematic review: fluid biomarkers and machine learning methods to improve the diagnosis from mild cognitive impairment to Alzheimer's disease. Alzheimers Res Ther. 2023;15(1):176.

[10] Ahmadzadeh M, Christie GJ, Cosco TD, Arab A, Mansouri M, Wagner KR, et al. Neuroimaging and machine learning for studying the pathways from mild cognitive impairment to alzheimer's disease: a systematic review. BMC Neurol. 2023;23(1):309.

[11] Chopra G, Shabir S, Yousuf S, Kauts S, Bhat SA, Mir AH, et al. Proteinopathies: Deciphering Physiology and Mechanisms to Develop Effective Therapies for Neurodegenerative Diseases. Mol Neurobiol. 2022;59(12):7513-40.

[12] Wen JH, He XH, Feng ZS, Li DY, Tang JX, Liu HF. Cellular Protein Aggregates: Formation, Biological Effects, and Ways of Elimination. Int J Mol Sci. 2023;24(10).

[13] Wu J, Cao C, Loch RA, Tiiman A, Luo J. Single-molecule studies of amyloid proteins: from biophysical properties to diagnostic perspectives. Q Rev Biophys. 2020;53:e12.

[14] Spires-Jones TL, Attems J, Thal DR. Interactions of pathological proteins in neurodegenerative diseases. Acta Neuropathol. 2017;134(2):187-205.

[15] Yang Y, Shi Y, Schweighauser M, Zhang X, Kotecha A, Murzin AG, et al. Cryo-EM structures of α-synuclein filaments from Parkinson's disease and dementia with Lewy bodies. bioRxiv. 2022:2022.07.12.499706.

[16] Yang Y, Shi Y, Schweighauser M, Zhang X, Kotecha A, Murzin AG, et al. Structures of alpha-synuclein filaments from human brains with Lewy pathology. Nature. 2022;610(7933):791-5.

[17] Cendrowska U, Silva PJ, Ait-Bouziad N, Muller M, Guven ZP, Vieweg S, et al. Unraveling the complexity of amyloid polymorphism using gold nanoparticles and cryo-EM. Proc Natl Acad Sci U S A. 2020;117(12):6866-74.

[18] Morten MJ, Sirvio L, Rupawala H, Mee Hayes E, Franco A, Radulescu C, et al. Quantitative super-resolution imaging of pathological aggregates reveals distinct toxicity profiles in different synucleinopathies. Proc Natl Acad Sci U S A. 2022;119(41):e2205591119.

[19] Archana R, Jeevaraj PSE. Deep learning models for digital image processing: a review. Artificial Intelligence Review. 2024;57(1):11.

[20] Fomalont EB, editor Image analysis. Synthesis Imaging in Radio Astronomy II; 1999.

[21] Bishop CM. Chapter 5.1: Feed-forward Network Functions. Pattern recognition and machine learning. 22006. p. 1122-8.

[22] Bishop CM, Bishop H. Chapter 6.2: Multilayer Networks. Deep learning: Foundations and concepts: Springer Nature; 2023.

[23] Goodfellow I, Bengio Y, Courville A. Chapter 14: Autoencoders. Deep Learning: MIT Press; 2016.

[24] Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998;86(11):2278-324.

[25] Li Z, Liu F, Yang W, Peng S, Zhou J. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. IEEE Transactions on Neural Networks and Learning Systems. 2022;33(12):6999-7019.

[26] Debelee TG, Schwenker F, Ibenthal A, Yohannes D. Survey of deep learning in breast cancer image analysis. Evolving Systems. 2020;11(1):143-63.

[27] Suganyadevi S, Seethalakshmi V, Balasamy K. A review on deep learning in medical image analysis. International Journal of Multimedia Information Retrieval. 2022;11(1):19-38.

[28] Li Q, Cai W, Wang X, Zhou Y, Feng DD, Chen M, editors. Medical image classification with convolutional neural network. 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV); 2014 10-12 Dec. 2014.

[29] Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical Image Analysis using Convolutional Neural Networks: A Review. Journal of Medical Systems. 2018;42(11):226.

[30] Ker J, Wang L, Rao J, Lim T. Deep Learning Applications in Medical Image Analysis. IEEE Access. 2018;6:9375-89.

[31] Zhai J, Zhang S, Chen J, He Q, editors. Autoencoder and Its Various Variants. 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC); 2018 7-10 Oct. 2018.

[32] Roy M, Kong J, Kashyap S, Pastore VP, Wang F, Wong KCL, et al. Convolutional autoencoder based model HistoCAE for segmentation of viable tumor regions in liver whole-slide images. Scientific Reports. 2021;11(1):139.

[33] Yagis E, Herrera AGSd, Citi L, editors. Convolutional Autoencoder based Deep Learning Approach for Alzheimer's Disease Diagnosis using Brain MRI. 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS); 2021 7-9 June 2021.

[34] Lelek M, Gyparaki MT, Beliu G, Schueder F, Griffié J, Manley S, et al. Single-molecule localization microscopy. Nature Reviews Methods Primers. 2021;1(1):39.

[35] Sideris DI, Danial JSH, Emin D, Ruggeri FS, Xia Z, Zhang YP, et al. Soluble amyloid beta-containing aggregates are present throughout the brain at early stages of Alzheimer's disease. Brain Commun. 2021;3(3):fcab147.

[36] Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv preprint arXiv:171105101. 2017.

[37] Seyfioğlu MS, Özbayoğlu AM, Gürbüz SZ. Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities. IEEE Transactions on Aerospace and Electronic Systems. 2018;54(4):1709-23.

[38] Li P, Pei Y, Li J. A comprehensive survey on design and application of autoencoder in deep learning. Applied Soft Computing. 2023;138:110176.

[39] Kroell N. imea: A Python package for extracting 2D and 3D shape measurements from images. Journal of Open Source Software. 2021;6(60):3091.

[40] Fisher R, Perkins S, Walker A, Wolfart E. Morphology - Skeletonization/Medial Axis Transform 2003 [https://homepages.inf.ed.ac.uk/rbf/HIPR2/skeleton.htm [Accessed:12th August 2024]

[41] Kroell N. Current available shape measurements 2021 [https://imea.readthedocs.io/en/latest/shape_measurements/ [Accessed:12th August

2024]

[42] Mazet V. Measure [https://vincmazet.github.io/bip/mm/measure.html [Accessed:12th August 2024]

[43] McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426. 2018.

[44] Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms; New Orleans, Louisiana: Society for Industrial and Applied Mathematics; 2007. p. 1027–35.

[45] Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987;20:53-65.

[46] Sculley D. Web-scale k-means clustering. Proceedings of the 19th international conference on World wide web; Raleigh, North Carolina, USA: Association for Computing Machinery; 2010. p. 1177–8.

[47] Maier M. DirichletReg: Dirichlet Regression for Compositional Data in R. [WU Working Paper]. In press 2014.

[48] Jo T. Machine learning foundations. Machine Learning Foundations Springer Nature Switzerland AG https://doi org/101007/978-3-030-65900-4. 2021.

[49] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, California, USA: Association for Computing Machinery; 2016. p. 785–94.

[50] Hequet I. Using Deep Learning to Identify High Specificity Biomarkers in Dementia: Imperial College London; 2022.

[51] Gu J. Deep learning approaches to predict dementia from images of toxic aggregates: Imperial College London; 2023.

[52] Holec SAM, Woerman AL. Evidence of distinct α-synuclein strains underlying disease heterogeneity. Acta Neuropathologica. 2021;142(1):73-86.

[53] Scheres SHW, Zhang W, Falcon B, Goedert M. Cryo-EM structures of tau filaments. Current Opinion in Structural Biology. 2020;64:17-25.

[54] Ruggeri FS, Habchi J, Cerreta A, Dietler G. AFM-Based Single Molecule Techniques: Unraveling the Amyloid Pathogenic Species. Curr Pharm Des. 2016;22(26):3950-70.

Appendix

A.Table 1(A): Dataset detail of the training set

Disease	Patient ID	Split	Number of Aggregates	Sex	Age
Control	C1	train	5,656	F	92
Control	C2	train	3,311	M	82
Control	C4	train	7,987	F	84
Control	C6	train	10,523	M	77
Control	C7	train	1,718	F	92
Control	C8	train	78,022	M	89
PD	PD1	train	8,164	M	76
PD	PD2	train	14,254	F	69
PD	PD4	train	934	M	75
PD	PD5	train	4,771	M	78
PD	PD9	train	56,728	F	83
PD	PD10	train	19,113	F	80
DLB	DLB1	train	3,418	F	80
DLB	DLB4	train	4,721	M	75
DLB	DLB5	train	107,922	M	81
FTD	FTD2	train	47,895	M	48
FTD	FTD3	train	30,568	M	63
FTD	FTD5	train	19,556	F	62
FTD	FTD6	train	40,034	M	66
FTD	FTD7	train	347,181	M	67
AD	AD4	train	12,690	M	92
AD	AD5	train	129,810	M	82
AD	AD6	train	18,547	M	90
AD	AD7	train	8,231	F	87
AD	AD8	train	3,615	F	92

A.Table 1(B): Dataset detail of the validation set

Disease	Patient ID	Split	Number of Aggregates	Sex	Age
Control	C3	validation	6,594	F	91
Control	C5	validation	7,539	F	81
PD	PD3	validation	501	F	85
PD	PD8	validation	6,262	M	73
DLB	DLB2	validation	61,622	M	71
FTD	FTD4	validation	1,761	M	71
AD	AD3	validation	26,620	F	88

A.Table 1(C): Dataset detail of the testing set

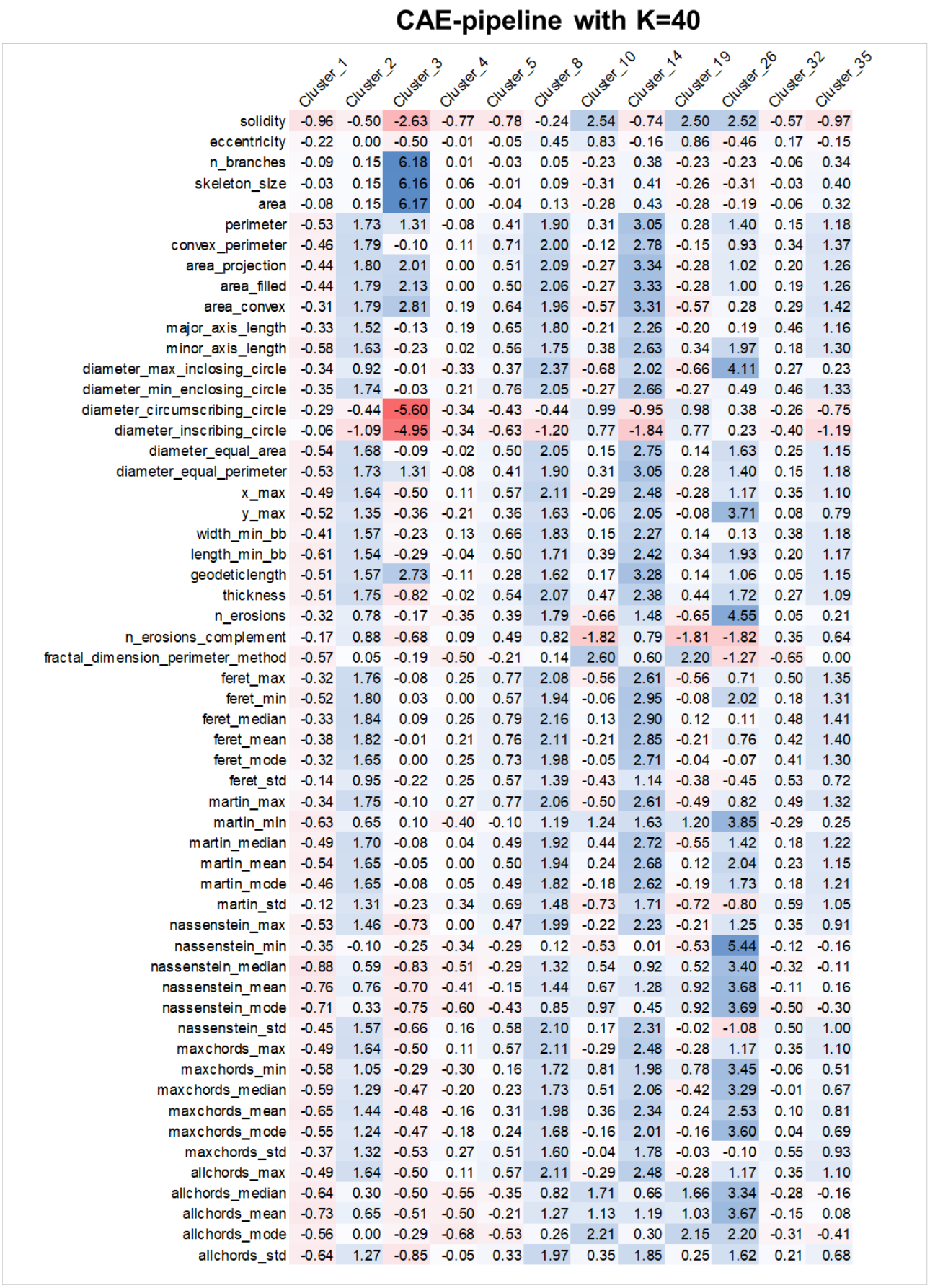
Disease	Patient ID	Split	Number of Aggregates	Sex	Age
Control	C9	test	3,271	F	73
Control	C10	test	584	M	77
PD	PD6	test	36,067	M	84
PD	PD7	test	844	F	83
DLB	DLB3	test	23,717	M	66
FTD	FTD1	test	8,552	F	63
FTD	FTD8	test	150,351	M	70
AD	AD1	test	7,707	M	93
AD	AD2	test	1,451	F	80

A.Table 2: Full list of shape measurements calculated in this project. This list is adapted from [41]. (continued on next page)

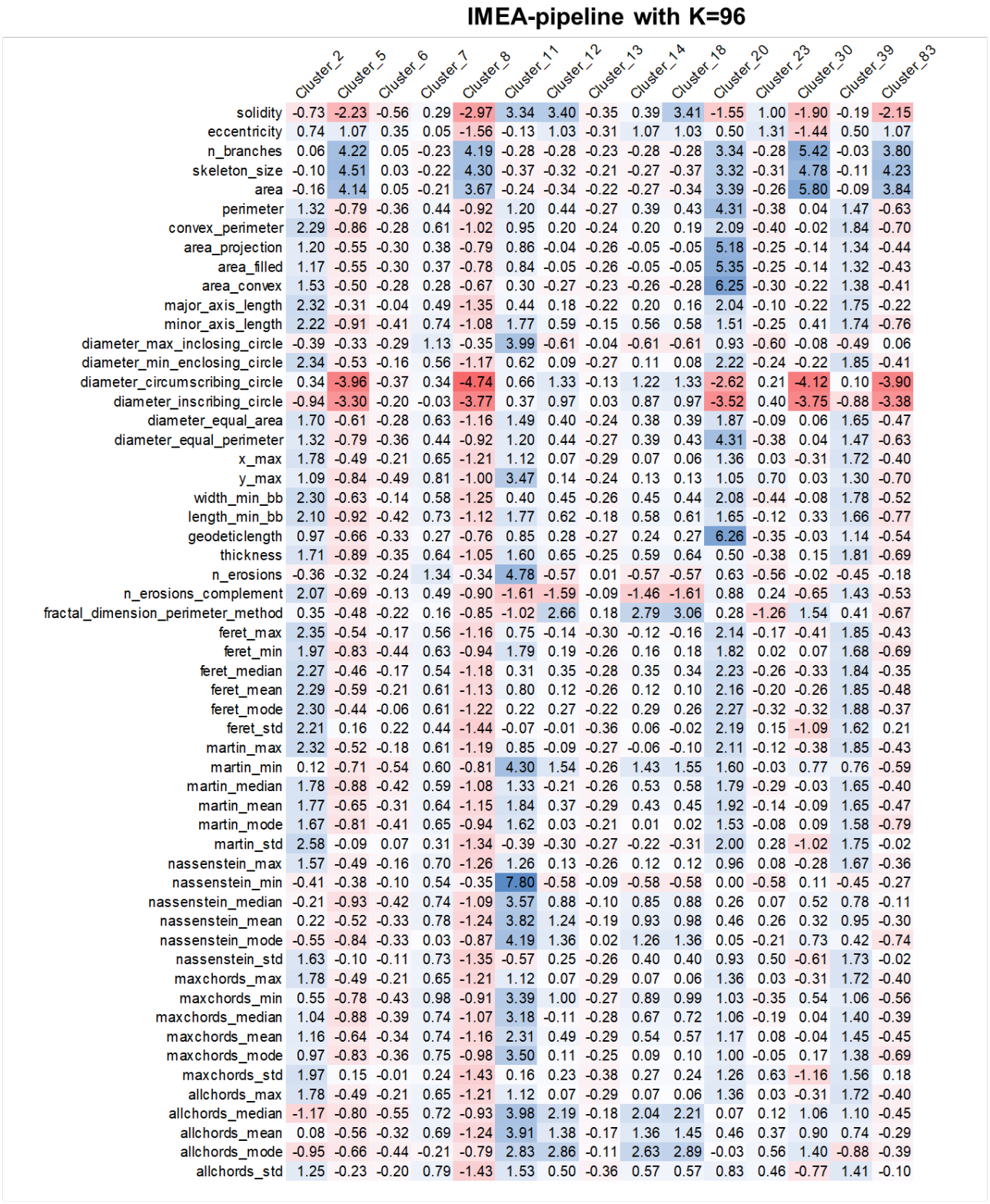
Feature Category		Feature Name
Common shape		Area
		Solidity
		Eccentricity
		Number of branches
		Skeleton size
Macro descriptors	Perimeter	perimeter
		convex perimeter
	Area	area projection
		area filled
		area convex
	Legendre inertia ellipse	major axis length
		minor axis length
	Diameters	diameter max inclosing circle
		diameter min enclosing circle
		diameter circumscribing circle
		diameter inscribing circle
		diameter equal area
		diameter equal perimeter
	Max dimensions	x max
		y max
	Minimal bounding box	width min bb
		length min bb
	Geodetic length and thickness	geodeticlength
		thickness
Meso descriptors	Erosions	n erosions
		n erosions complement
Micro descriptors	Roughness of particle contours	fractal dimension boxcounting method (Excluded)
		fractal dimension perimeter method

A.Table 2: Full list of shape measurements calculated in this project. This list is adapted from [41].

Statistical lengths	Feret diameter	feret max
		feret min
		feret median
		feret mean
		feret mode
		feret std
	Martin diameter	martin max
		martin min
		martin median
		martin mean
		martin mode
		martin std
	Nassenstein diameter	nassenstein max
		nassenstein min
		nassenstein median
		nassenstein mean
		nassenstein mode
		nassenstein std
	Max chords	maxchords max
		maxchords min
		maxchords median
		maxchords mean
		maxchords mode
		maxchords std
	All chords	allchords max
		allchords min (not included)
		allchords median
		allchords mean
		allchords mode
		allchords std



A.Figure 1: Heatmap displaying mean values of shape measurement features for each significant clusters in CAE pipeline.



A.Figure 2: Heatmap displaying mean values of shape measurement features for each significant clusters in IMEA pipeline.