IMPERIAL

Predicting dementia types using machine learning approach on aggregate images

Yuting Gu

MSc HDAML Research Project

Supervisor: Dr. Yu Ye (Ye Lab)

19th September 2024

Protein Aggregates Associated with Different Types of Dementia

Dementia

Alzheimer's Disease

Dementia with Lewy bodies

Frontotemporal Dementia

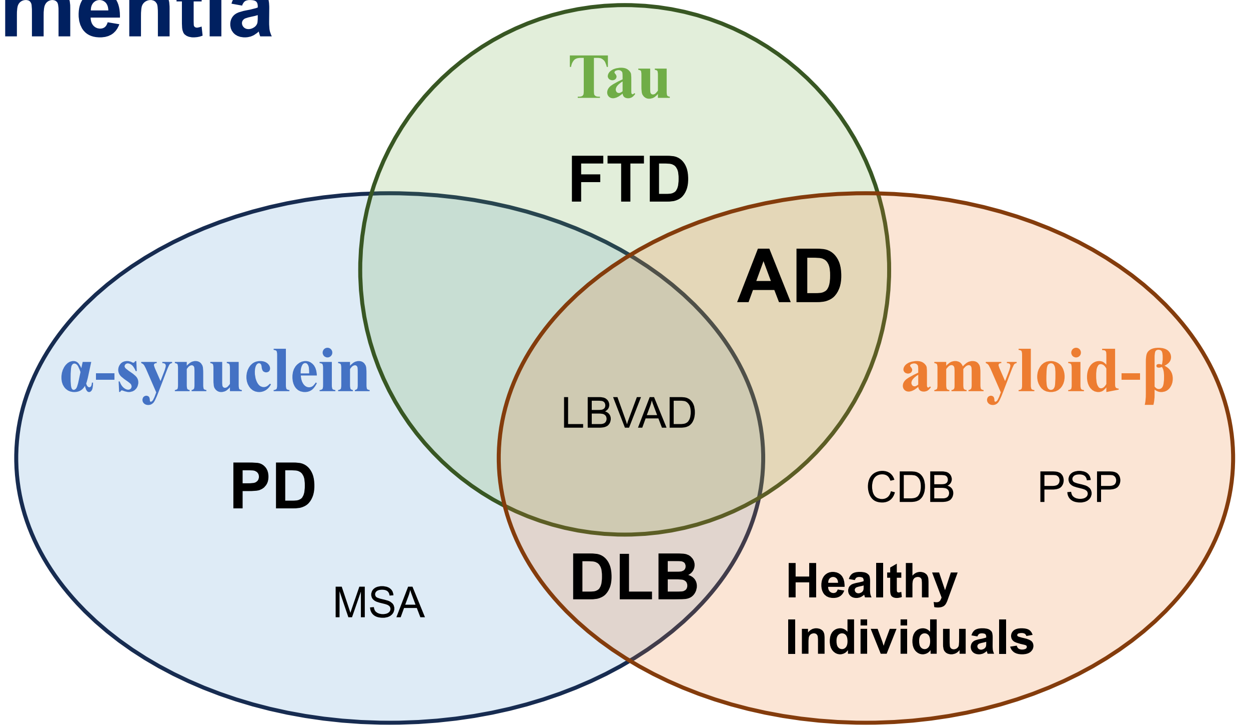
Parkinson's Disease

hallmark

Progressive accumulation

Misfolded protein aggregates

Misfolded Protein



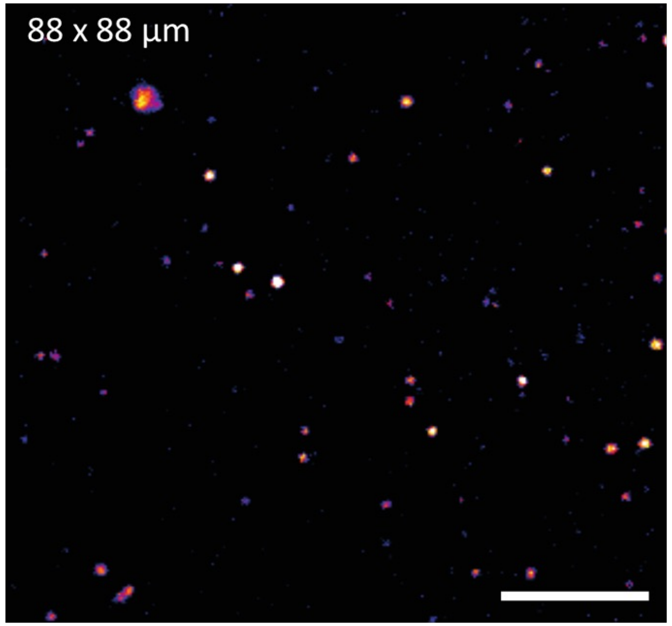
Challenge

Hard to distinguish different disease associated with same protein aggregates

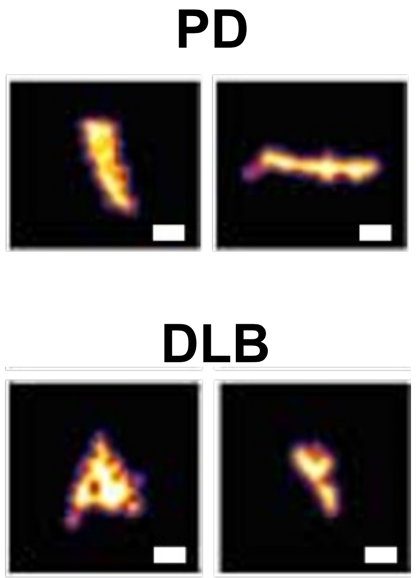
IMPERIAL

Morphological Difference of Aggregates

Super-resolved SMLM images



down to ~25 nm
Beyond the diffraction limit of
a typical optical microscope



Previous work from the lab

Super-resolved imaging technique

Some shape more prevalent to one disease than another

Use images capturing aggregate morphology to classify different types of dementia

Morten, M.J. et al. (2022) 'Quantitative super-resolution imaging of pathological aggregates reveals distinct toxicity profiles in different synucleinopathies', Proceedings of the National Academy of Sciences, 119(41), p. e2205591119.

Research Rationale

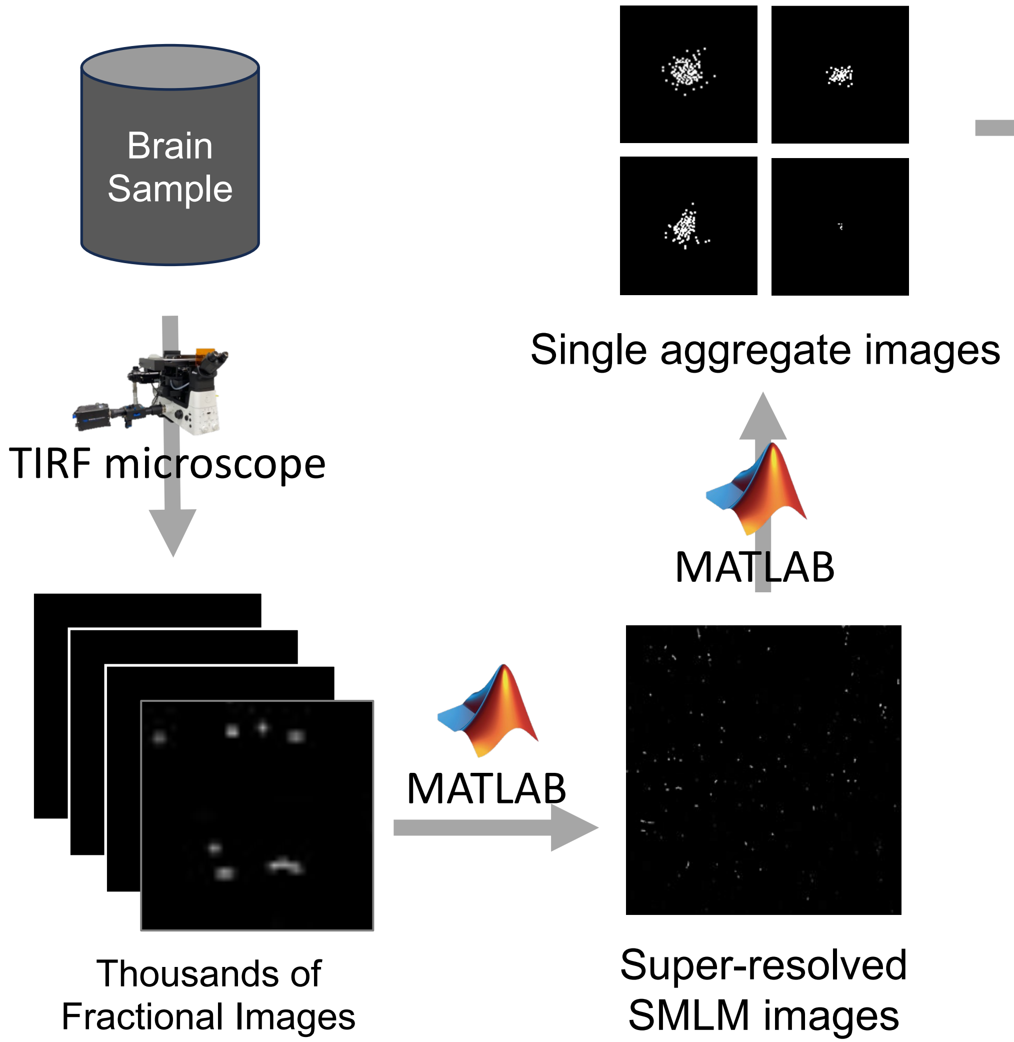
Hypothesis

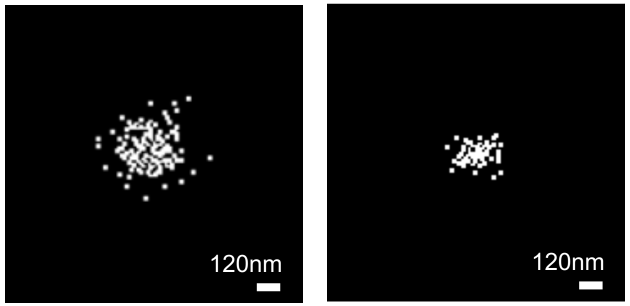
Each type of dementia has a subset of **morphological disease-specific** aggregates

Aims

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

Data Acquisition and Preprocessing





Single aggregate images (preprocessed)

- Exclude larger than 128 pixels
- Exclude aggregate with sizes <= 4 pixels
- Padded all to 128 x 128

- Randomly assign subset of patients to Training/Validation/Testing Set

Train Set

C1, C2, C4, C6, C7, C8
PD1, PD2, PD4, PD5, PD9, PD10
DLB1, DLB4, DLB5, FTD2, FTD3, FTD5
FTD6, FTD7, AD4, AD5, AD6, AD7
AD8

60%

Validation Set

C3, C5
PD3, PD8
DLB2, FTD4
AD3

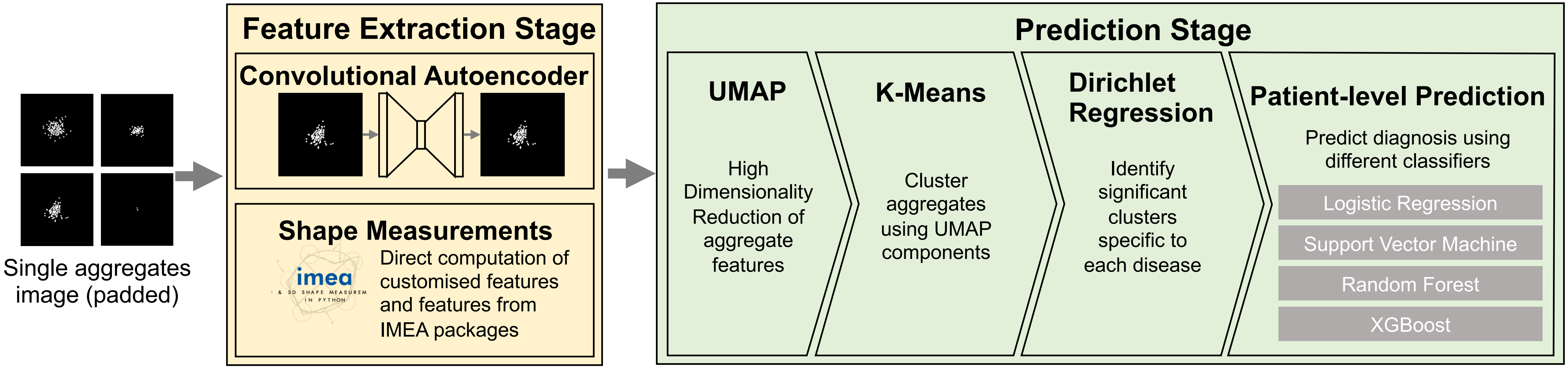
20%

Test Set

C9, C10
PD6, PD7
DLB3, FTD1
FTD8, AD1
AD2

20%

Overview of Method Pipeline

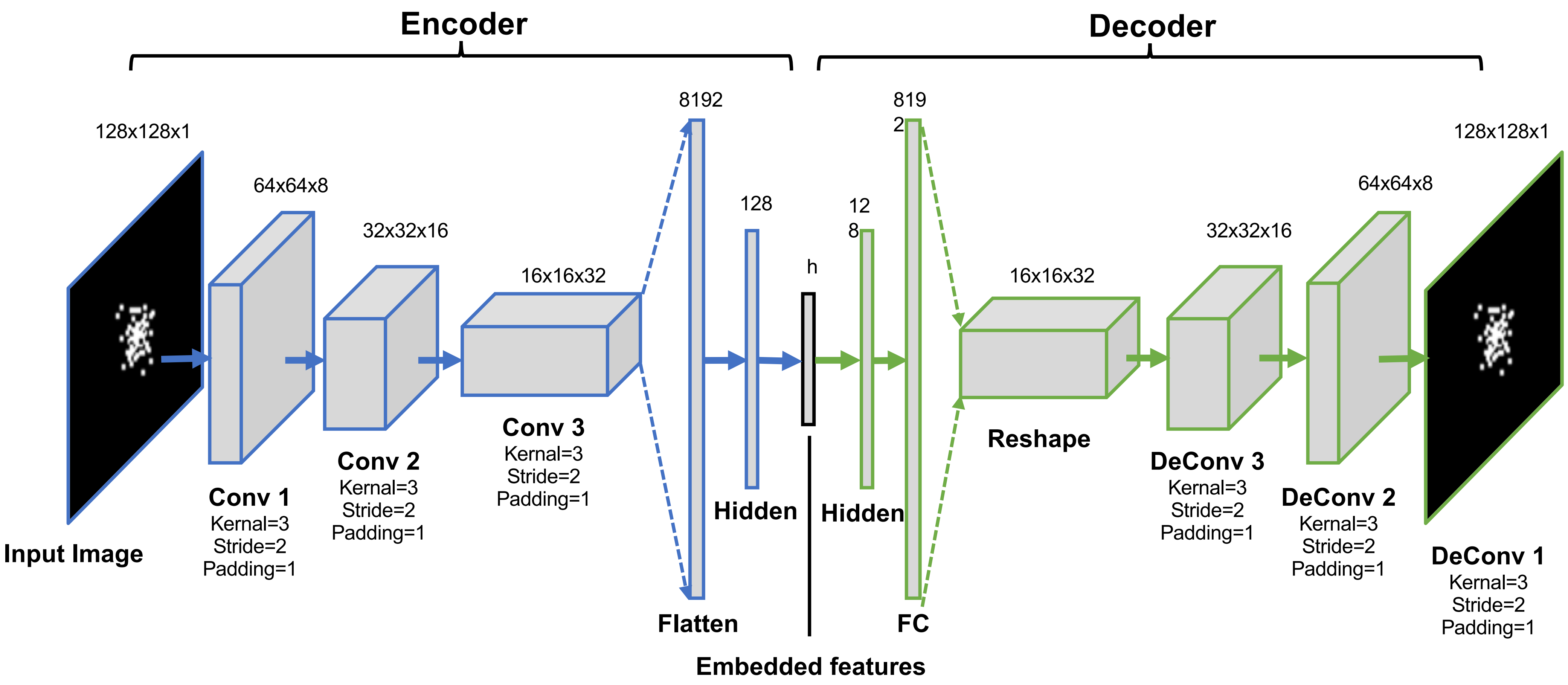


Compare two pipelines using different feature extraction approaches

Characterise aggregates specific to distinct disease

Evaluate classifiers

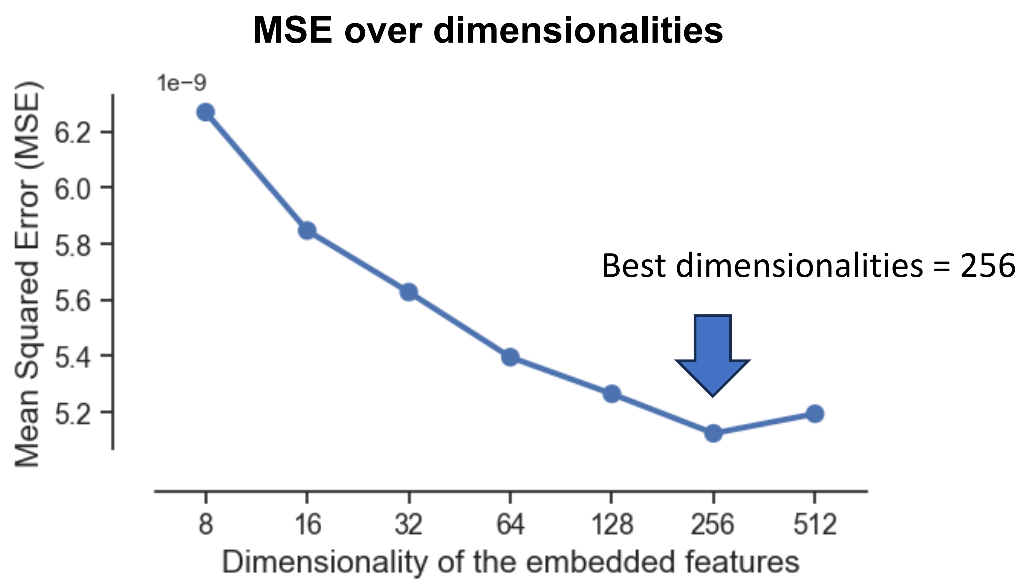
Convolutional Autoencoder(CAE)



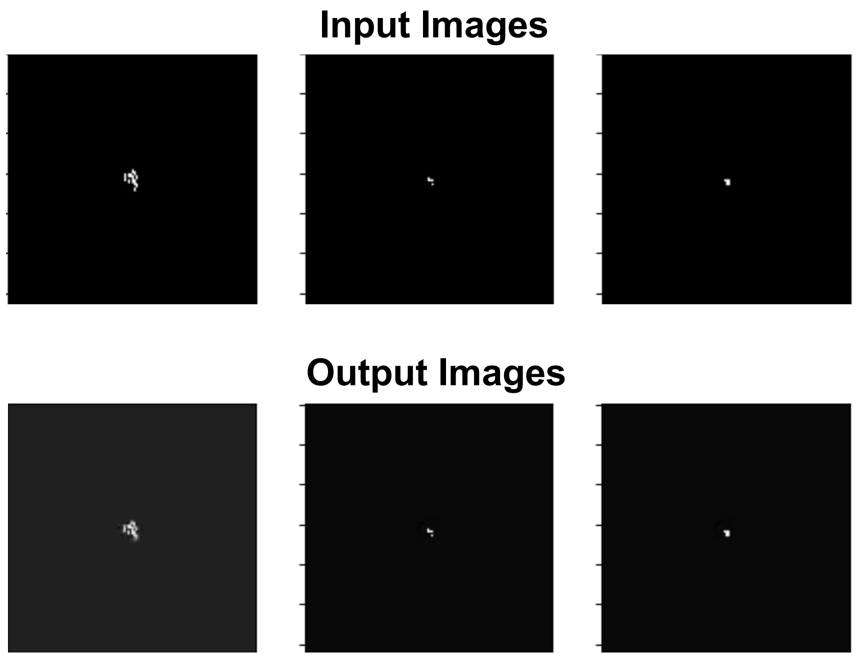
Convolutional Autoencoder(CAE)

IMPERIAL

Use Mean Squared Error (MSE) to select best dimensionality of reduction



Example of CAE performance at Best dim=256



IMPERIAL

Shape Measurements

56 shape measurement features collected for each image

5 commonly used shape measurement features

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
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

51 computed using IMEA package

macro descriptors
geometric features

perimeter, area, diameter and etc.

meso descriptors
intermediate details

like erosions

micro descriptors
finer details

roughness of particle contours, specific diameter measurements like Feret and etc.

statistical lengths
distribution and variation in lengths

various chord lengths

Prediction Stage: UMAP and K-Means

256 CAE image features

56 IMEA image features

UMAP

20 CAE-UMAP features

20 IMEA-UMAP features

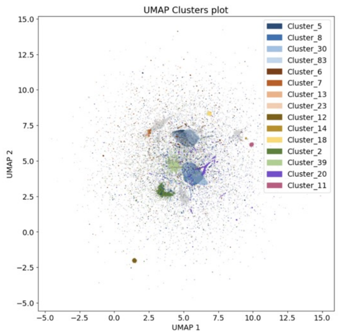
K-Means

Cluster Percentage

$$P_i = \frac{C_i}{\sum_{j=1}^{K} C_j}$$

C_i is the number of aggregates affiliated to cluster i, and K is the total number of clusters

Reduce dimensionalities to 20 features to enable the feasibility of clustering tools



Clustering aggregate morphologies using K-Means

Explore best choice of K

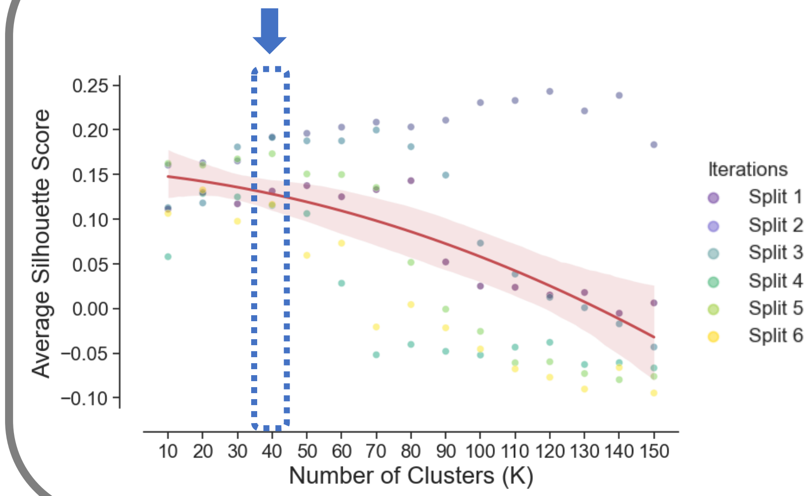
Repeating the clustering for different **6 splits** and K ranging from **10 to 150**.

Evaluate clustering performance by **average silhouette score.** Higher score is better.

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

Solve best K of a fitted quadratic line

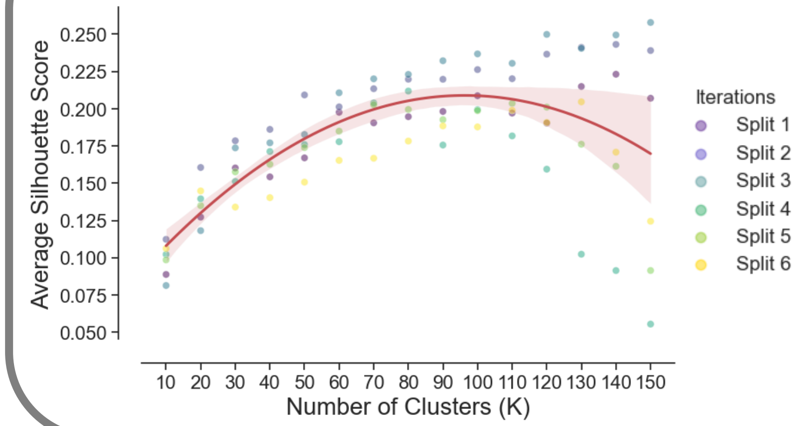
CAE-pipeline



No feasible solution of the fitted quadratic line

Best K = 40

IMEA-pipeline



Fitted quadratic line

$$-1.4\times10^{-5}x^2 + 0.0026x + 0.083 = 0$$

Best K = 96

IMPERIAL

Identifying significant clusters using Dirichlet regression

- Different grouped clusters present very **distinct characteristics** by their polygon shape

- Both feature extraction approaches (CAE and IMEA) 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.**

CAE-pipeline | **IMEA-pipeline**

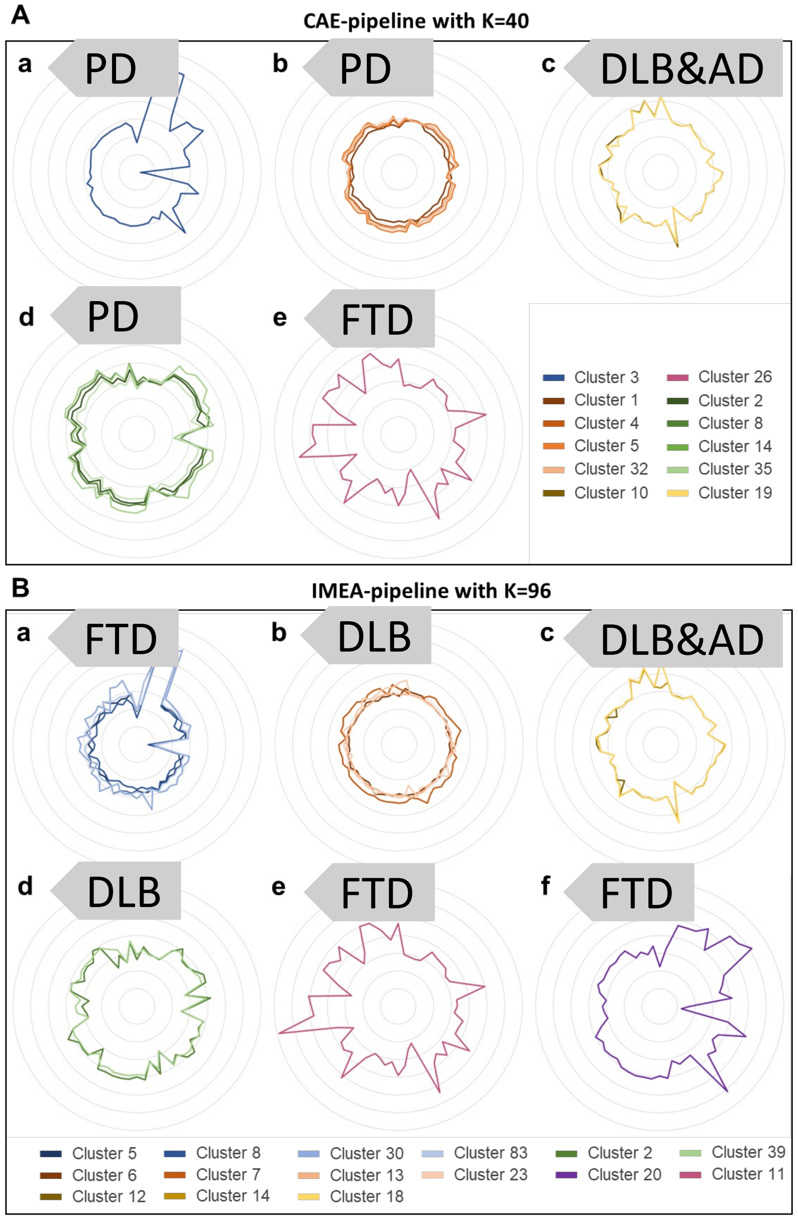
Cluster Percentage

Dirichlet regression

12 significant clusters were selected | **15 significant clusters were selected**

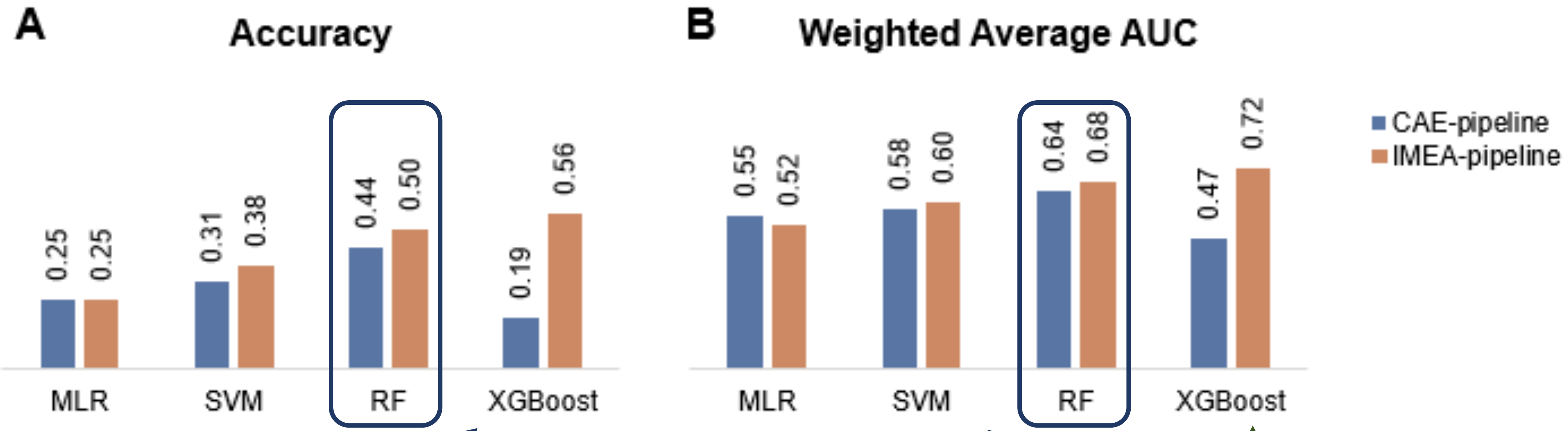
use the 53 IMEA features to assist in the interpretation of the characteristics of clusters

5 grouped clusters were identified | **6 grouped clusters were identified**



Patient-level Prediction

Tree based method, RF and XGBoost, tended to show superior results



Random forest

Best balance between performance and reliability by achieving best performance in CAE-pipeline and second best for IMEA-pipeline

Performed best in IMEA-pipeline but worst in CAE-pipeline

IMPERIAL

Strengths, Limitations and Future Work

Strengths	• **Identified clusters of disease-specific aggregates**
	  • Observed resemblance in characteristics of these aggregate subsets in both pipelines, indicating stability of this aggregate identification method
	• **Developed end-to-end prediction pipeline**
	  ○ unsupervised feature extraction method (CAE)
	  ○ fine-tune the models
	  ○ practically more efficient
	• **Evaluation of different feature extraction approaches and classifiers**

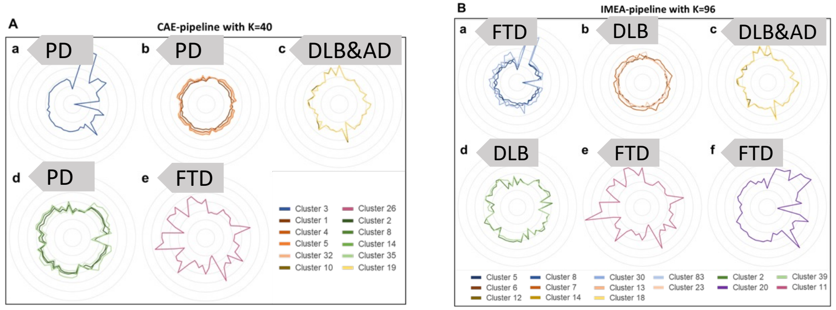
Limitations	• Limited sample size (41 donors) • Binary (black-and-white) images, limit the data's richness. • One split of dataset, need more evaluation of generalisation
Future Work	• Collect more donors' sample • Non-binary images with pixel intensity or coloured images • Optimise models and replicate pipeline

IMPERIAL

Conclusion

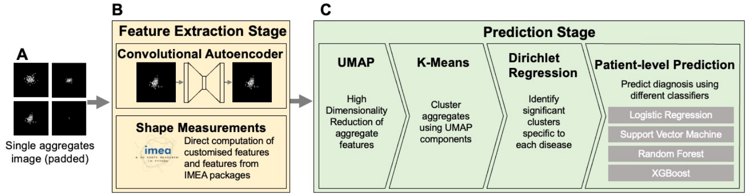
Aim1: Aggregates morphology

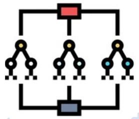
Identified and characterised the morphological differences of disease-specific aggregates



Aim2: Prediction pipeline

Novel development of the comprehensive end-to-end machine learning pipeline for dementia prediction using aggregate images





Random forest

weighted average AUC

- CAE-Pipeline: 0.64
- IMEA-Pipeline: 0.68

Important foundational framework for developing a scalable, high-throughput diagnostic tool using aggregate morphology

IMPERIAL

Thank you for your attention!

Acknowledgement

IMPERIAL

Dr. Yu Ye
Main Supervisor
Principal Investigator
Senior Lecturer in Molecular Neuroscience at Imperial College London

Hailey Gu
HDA alumni
Research Technician in Computational Biology

Dr. Michael Morten
Secondary Supervisor
Research Associate in Biophysics