# Classification challenge on Alzheimer’s Disease using MRIs and Gene Expression data

May 3, 2023

# The problem

- AD affects about 55M people in the world*

  → need for early diagnosis

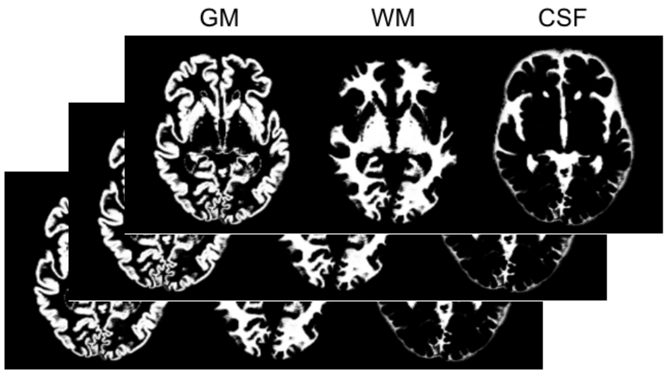
- AD (macro-)stages
  - CTL (Controls): no deficit
  - MCI (Mild Cognitive Impairment): few deficits
  - AD (Alzheimer's Disease): dementia

*[WHO2021]

# “Hints” from different types of data

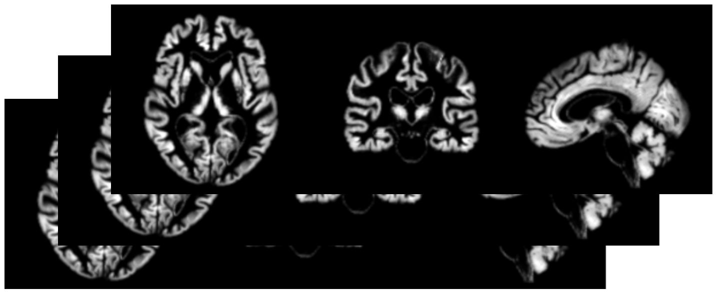
- Demographic (age, gender, instruction, …)
- Clinical evaluation (cognitive tests)
- CSF (Cerebrospinal fluid)
- Medical imaging (MRIs, PETs, DTIs, …)
- Transcriptomics (gene expression, …)

# Features from MRIs



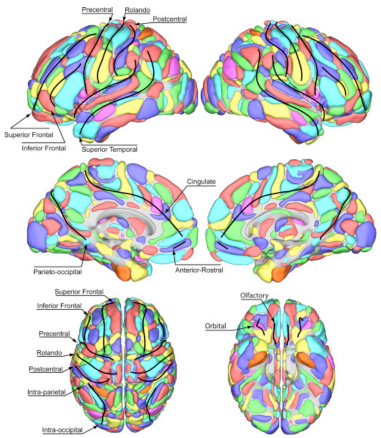

Tissue segmentation, bias correction and spatial normalization

Inter-subject registration (group template)



All images mapped to a common space (MNI), providing a voxel-wise correspondence across subjects

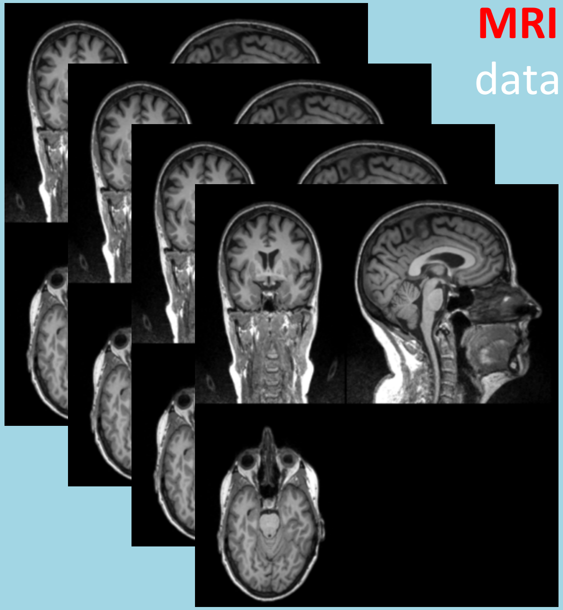
Average grey matter densities obtained from each anatomical region defined in an atlas




AICHA Atlas

# Combined Data

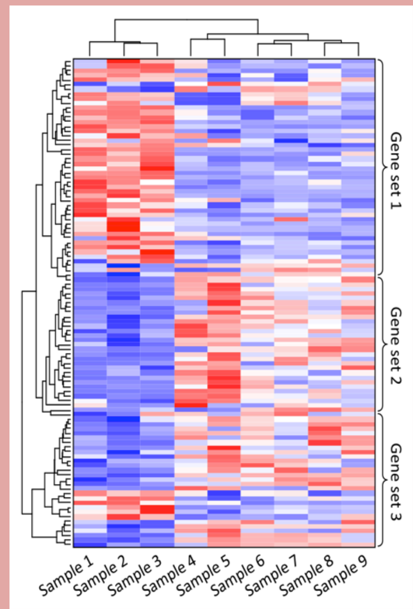
**MRI** data



| Subject_id | Background | G_Frontal_Sup-1-L | G_Frontal_Sup-1-R | G_Frontal_Sup-2-L | G_Frontal_Sup-2-R | G_Frontal_Sup-3-L |
|---|---|---|---|---|---|---|
| 002_S_0413 | 0.021984403 | 0.470479202 | 0.365683574 | 0.42932241 | 0.401926473 | 0.497643487 |
| 002_S_0685 | 0.020607992 | 0.47969718 | 0.218352588 | 0.380672979 | 0.362764797 | 0.370832544 |
| 002_S_1261 | 0.021845324 | 0.456640816 | 0.348445856 | 0.398290689 | 0.383856007 | 0.550132618 |
| 002_S_1280 | 0.017489938 | 0.348388934 | 0.2053103 | 0.300627906 | 0.291697004 | 0.473903188 |
| 002_S_4213 | 0.021635527 | 0.431245375 | 0.295218306 | 0.410807503 | 0.371577452 | 0.433755406 |
| 002_S_4225 | 0.028633343 | 0.5985744 | 0.637216532 | 0.51010088 | 0.549204757 | 0.640226531 |
| 002_S_4270 | 0.020788799 | 0.443955183 | 0.327883199 | 0.36906867 | 0.319275917 | 0.390957816 |
| 003_S_0907 | 0.022751048 | 0.427923298 | 0.370895854 | 0.37581052 | 0.345848585 | 0.393171344 |

+

**GE** data




| Subject_id | ACKR2 | ACP5 | ACRBP | ADAMTS15 | ADAP2 | ADD2 | ADIPOR1 | AHSP | AIMP2 | ALAS2 | ALDH5A1 | ALYREF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 002_S_0413 | 3.9135 | 3.979 | 8.761 | 3.254 | 5.7315 | 3.2205 | 10.5 | 7.634 | 7.4315 | 11.407 | 4.963 | 5.428 |
| 002_S_0685 | 2.9675 | 3.924 | 7.974 | 3.3715 | 5.543 | 2.867 | 9.7755 | 6.121 | 7.108 | 9.178 | 4.391 | 5.954 |
| 002_S_1261 | 3.777 | 3.865 | 8.89 | 3.202 | 5.2415 | 2.8285 | 10.6265 | 6.684 | 7.24 | 10.077 | 4.969 | 5.364 |
| 002_S_1280 | 3.372 | 4.429 | 8.382 | 3.3975 | 4.869 | 3.0875 | 9.4835 | 6.65 | 7.145 | 10.681 | 4.981 | 5.491 |
| 002_S_4213 | 3.4565 | 3.539 | 9.186 | 3.3835 | 5.0205 | 3.067 | 10.944 | 7.162 | 7.452 | 11.02 | 4.8525 | 5.363 |
| 002_S_4225 | 3.679 | 3.827 | 7.676 | 3.052 | 5.243 | 3.13 | 10.6015 | 7.741 | 7.32 | 10.658 | 5.505 | 5.921 |
| 002_S_4270 | 3.8005 | 3.421 | 7.89 | 3.1215 | 5.207 | 2.9305 | 9.973 | 6.475 | 7.4695 | 10.14 | 5.2035 | 5.113 |
| 003_S_0907 | 3.598 | 3.935 | 7.956 | 3.4655 | 6.419 | 3.172 | 11.0555 | 8.431 | 7.1595 | 11.95 | 4.9155 | 5.892 |

# You will receive

- 3 training datasets for 3 different binary classification problems. These are to be used for training 3 separate classifiers, experimenting different classifiers, feature selection methods, etc. and validating them through cross-validation
- 3 test datasets (one for each binary problem) to test the classifier and provide the obtained predictions

# Training data

**1) ADCTLtrain.csv**: training data for binary classification problem to discriminate AD vs. CN patients

a) First column: ID of the patient
b) Columns from 2 to 430: MRI+GE features
c) Last column: Label (patient classification: ‘AD’ or ‘CTL’)

Overall 164 patients: 81 AD and 83CTL

**2) ADMCItrain.csv**: training data for binary classification problem to discriminate AD vs. MCI patients

a) First column: ID of the patient
b) Columns from 2 to 64: MRI+GE features
c) Last column: Label (patient classification: ‘AD’ or ‘MCI’)

Overall 172 patients: 82 AD and 90 MCI

**3) MCICTLtrain.csv**: training data for binary classification problem to discriminate MCI vs. CTL patients

a) First column: ID of the patient
b) Columns from 2 to 594: MRI+GE features
c) Last column: Label (patient classification: ‘MCI’ or ‘CTL’)

Overall 172 patients: 90 MCI and 82 CTL

# Example of training dataset

| ID | Background | Precentral_L | … | ABCA7 | AGTRAP | … | Labels |
|---|---|---|---|---|---|---|---|
| ADCTL001 | ... | ... | | ... | ... | | AD |
| ADCTL002 | ... | ... | | ... | ... | | AD |
| ADCTL003 | ... | ... | | ... | ... | | AD |
| ADCTL004 | ... | ... | | ... | ... | | AD |
| ADCTL005 | ... | ... | | ... | ... | | AD |
| ADCTL006 | ... | ... | | ... | ... | | AD |
| ADCTL007 | ... | ... | | ... | ... | | AD |
| ADCTL008 | ... | ... | | ... | ... | | AD |
| ADCTL009 | ... | ... | | ... | ... | | AD |
| ADCTL010 | ... | ... | | ... | ... | | AD |
| … | … | … | … | … | … | … | … |

# Test data

**1) ADCTLtest.csv**: test data for binary classification problem to discriminate AD vs. CTL patients

a) First column: ID of the patient
b) Columns from 2 to 430: MRI+GE features

Overall 41 patients

**2) ADMCItest.csv**: test data for binary classification problem to discriminate AD vs. MCI patients

a) First column: ID of the patient
b) Columns from 2 to 64: MRI+GE features

Overall 41 patients

**3) MCICTLtest.csv**: test data for binary classification problem to discriminate MCI vs. CTL patients

a) First column: ID of the patient
b) Columns from 2 to 594: MRI+GE features

Overall 43 patients

# Your submission will consist of

For each binary classification problem

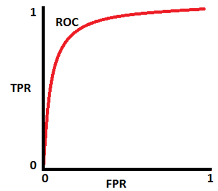
1. Two CSV files whose name are formatted as:
   *StudentRegistrationNumber_FamilyName*_*res.csv
   *StudentRegistrationNumber_FamilyName*_*feat.csv
   where * denotes the classification problem (ADCTL, ADMCI, or MCICTL)
   - 1.a) The first file will contain three columns
     - the IDs of the test observation;
     - the predicted labels;
     - the probabilities of the predicted labels;
   - 1.b) The second file will contain the column index in the training data files (from 2 to end-1) of the selected features. If features are somehow pre-transformed, describe the transformation in the presentation file (see below).
2. A presentation in PDF with up to 6 pages, in which you describe how you obtained the model. It is NOT mandatory to choose the same classification model for the three different binary problems; just choose the one that leads to the most promising results.
3. The R macro (script) named *StudentRegistrationNumber_FamilyName*_solution.R used to obtain the results.

# The winner is…

- For each binary problem, the results will be ranked according to
  - AUC (e.g., auc R function from pROC) and
  - MCC (e.g., mcc R function from mltools)

$$AUC = \int_0^1 Sens(x)\,dx, \quad x = 1 - Spec$$

Uses the ROC curve to exhibit the trade-off between the classifier's TP and FP rates




$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Correlation coefficient between observed and predicted binary classifications

# Other metrics

$$Acc = \frac{TP + TN}{TP + FN + FP + TN}$$

Percentage of correctly classified samples

$$Spec = \frac{TN}{TN + FP}$$

Percentage of negative samples correctly identified

$$Sens = \frac{TP}{TP + FN}$$

Percentage of positive samples correctly classified (*Recall* or *TPR*)

$$Prec = \frac{TP}{TP + FP}$$

Percentage of positive samples correctly classified, considering the set of all the samples classified as positive

$$F_1 = \frac{2 \cdot Prec \cdot Sens}{Prec + Sens}$$

Compromise between sensitivity and precision

$$BA = \frac{Spec + Sens}{2}$$

Mean of Specificity and Sensitivity

# Example results

## A) Performance on the training datasets:

| | Acc | Sens | Spec | Prec | F1 | AUC | MCC | BA |
|---|---|---|---|---|---|---|---|---|
| **AD vs CTL** | 0.902 | 0.926 | 0.880 | 0.882 | 0.904 | 0.969 | 0.806 | 0.903 |
| **AD vs MCI** | 0.977 | 0.976 | 0.978 | 0.976 | 0.976 | 0.993 | 0.953 | 0.977 |
| **MCI vs CTL** | 0.826 | 0.867 | 0.780 | 0.812 | 0.839 | 0.884 | 0.651 | 0.824 |

## B) Performance on the test datasets using the classifier trained on the training data:

| | Acc | Sens | Spec | Prec | F1 | AUC | MCC | BA |
|---|---|---|---|---|---|---|---|---|
| **AD vs CTL** | 0.829 | 0.857 | 0.800 | 0.818 | 0.837 | 0.888 | 0.659 | 0.829 |
| **AD vs MCI** | 0.643 | 0.700 | 0.591 | 0.609 | 0.651 | 0.761 | 0.292 | 0.645 |
| **MCI vs CTL** | 0.628 | 0.545 | 0.714 | 0.667 | 0.600 | 0.801 | 0.263 | 0.630 |