Application of Multimodal Machine Learning methods for studying the heterogeneity of brain ageing using neuroimaging and genetic data

Γιώργος Αϊδίνης

Καθηγήτρια: Κωνσταντίνα Νικήτα

Εργαστήριο Βιοϊατρικών Προσομοιώσεων και Απεικονιστικής Τεχνολογίας (BIOSIM)

29/06/2022

The goal of this thesis

Apply ML and DL techniques in order to classify imaging and genetic data into:

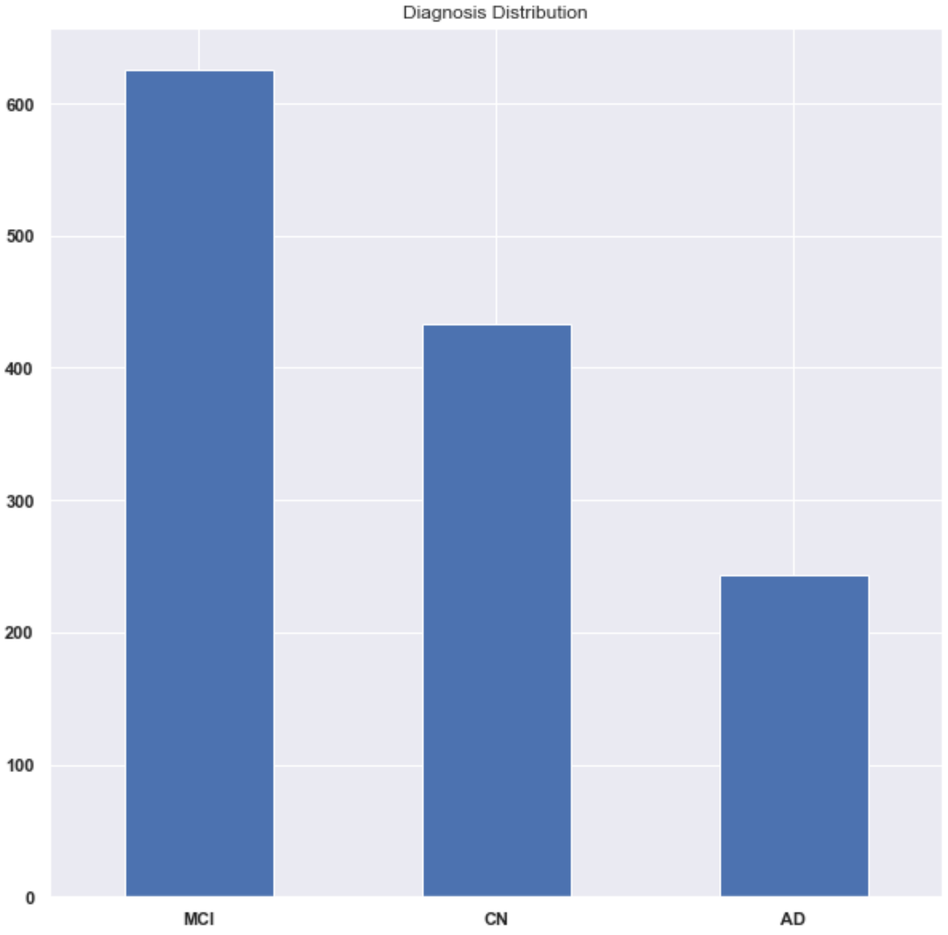
- Cognitive Normal (**CN**)
- Mild Cognitive Impairment (**MCI**)
- Alzheimer's Disease (**AD**)

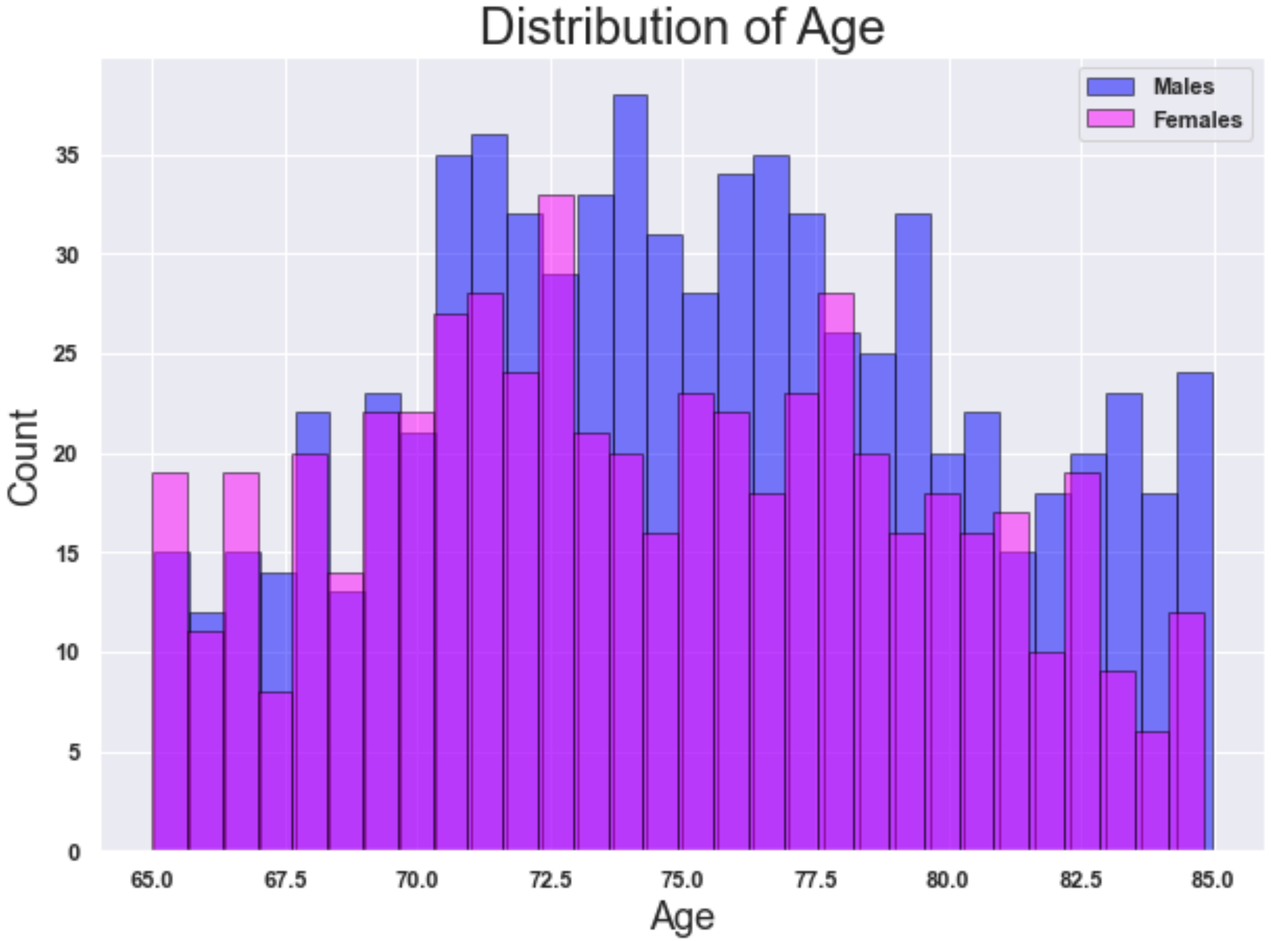
Steps:

1. **Study** the data
2. Perform **preprocessing steps**, train and apply models (**DCCA, MCA, OPNMF, FAMD**)
3. Perform **Classification** using various algorithms (**SVM**, **Ensembles**)
4. Draw **conclusions**, compare results

Study our Dataset

The ADNI Dataset



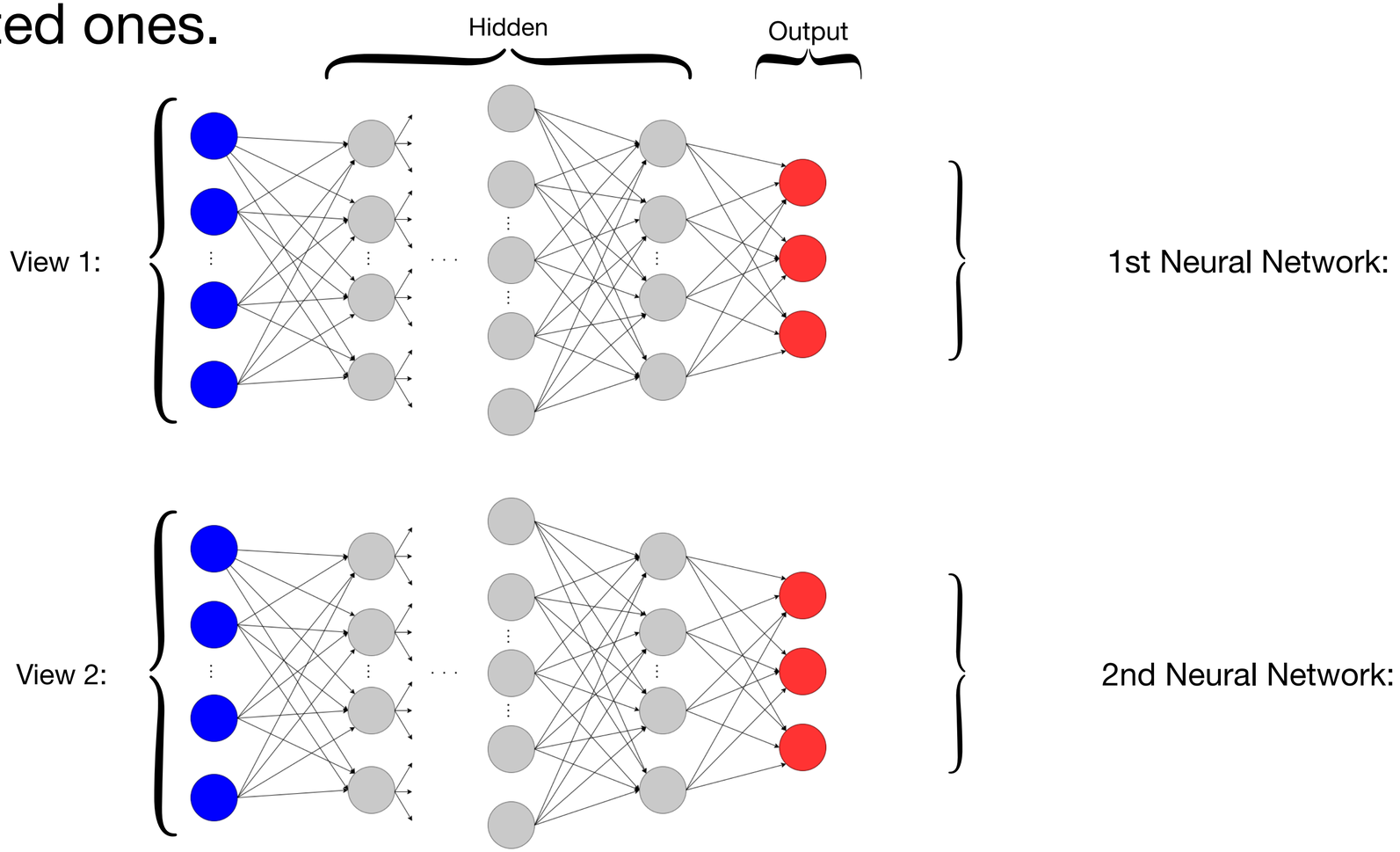


Preprocessing

Deep Canonical Correlation Analysis

What is DCCA (1)

Goal: Transform 2 views of the dataset (e.g. Imaging - Genetic) into linearly correlated ones.



G. Andrew et al., *Deep Canonical Correlation Analysis*, Proceedings of the 30th International Conference on Machine Lear- ning, 2013.

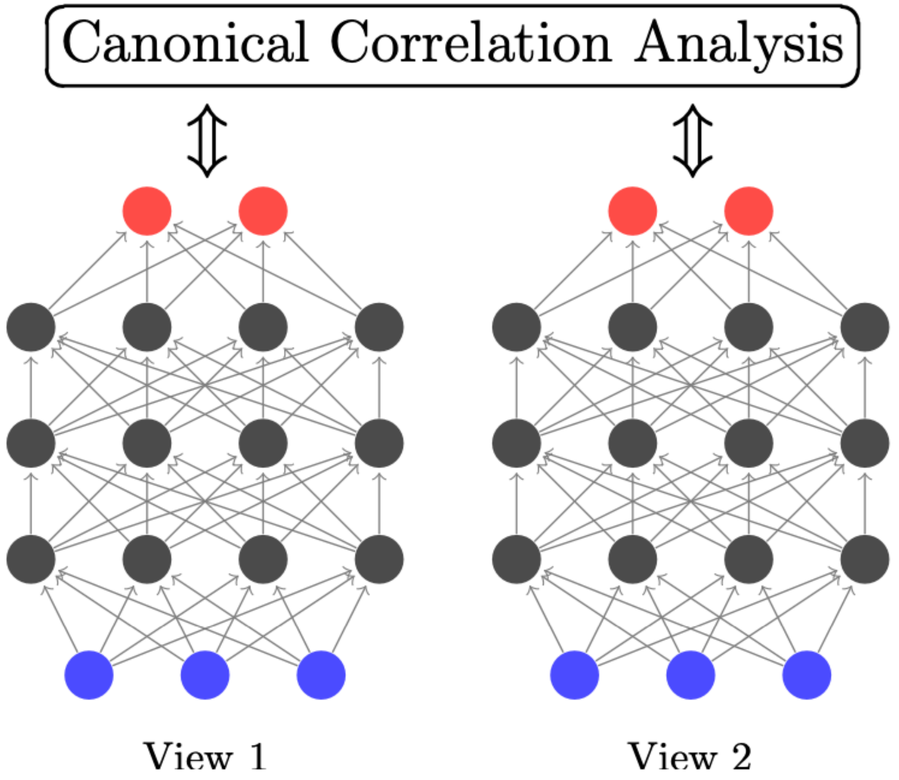
What is DCCA (2)

Normal CCA:

Find correlation between the blue data

Deep CCA:

Find correlation between the red (transformed) data



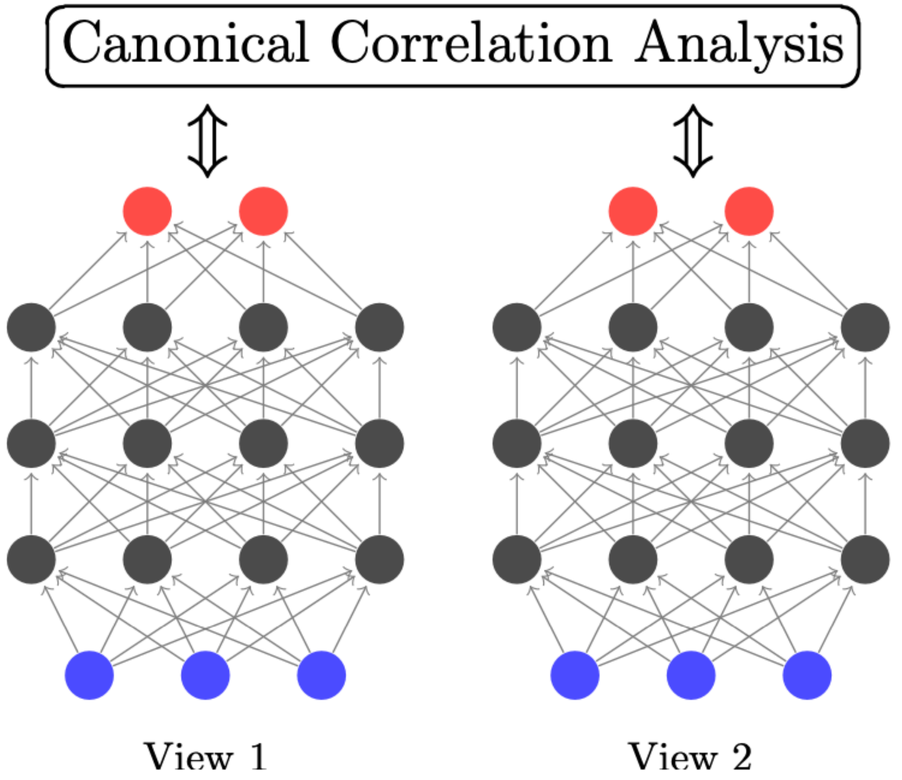
Using DCCA

2 Views:

- Imaging Data
- Genetic Data

Play with parameters: (Use GridSearchCV)

- Output layer dimensions
- Number of Hidden Layers
- Hidden Layer Size
- Regularization
- Learning Rate
- Batching



DCCA Training Conclusions

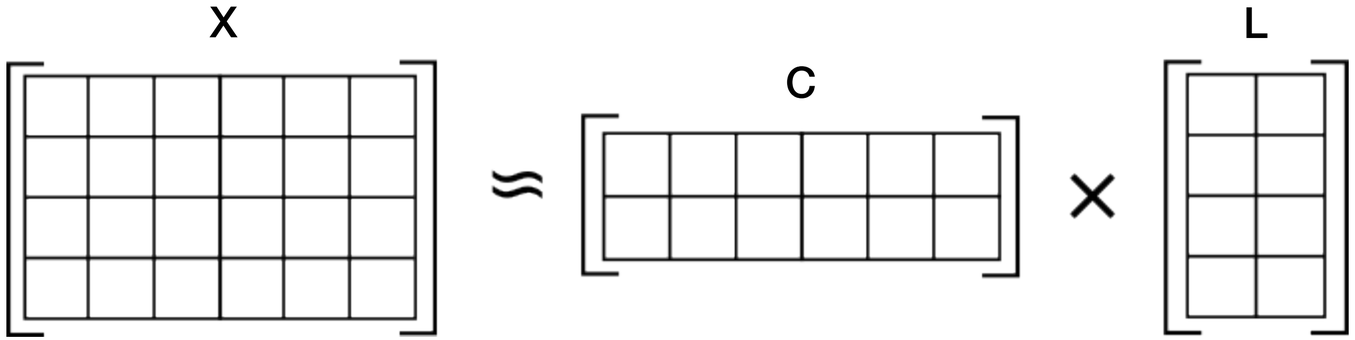
Parameter Action	*Correlation (Negative Loss)*
Output Dimension Size ↑	↑
Hidden Layer Size ↑	↑
Learning Rate	Medium to low LR is best
Batch Size	Stays basically the same
Regularization Parameter	Stays basically the same

Orthonormal Projective Non-Negative Matrix Factorization

What is OPNMF

- Orthonormal Projective Non-Negative Matrix Factorization
- Goal: dimensionality reduction (NMF)
- Approximation of the original array **X** with a multiplication of non-negative arrays **C,L**.
- OPNMF: $C^T C = I$
- Minimize the approximation error:

$$\begin{aligned} \min_{\mathbf{C}} \quad & \|\mathbf{X} - \mathbf{C}\mathbf{C}^T\mathbf{X}\|_F^2 \\ \text{subject to} \quad & \mathbf{C} \geq 0, \mathbf{C}^T\mathbf{C}=\mathbf{I}, \end{aligned}$$



A. Sotiras, S. M. Resnick and C. Davatzikos. *Finding imaging patterns of structural covariance via Non-Negative Matrix Factorization*. Neuroimage, 2015.

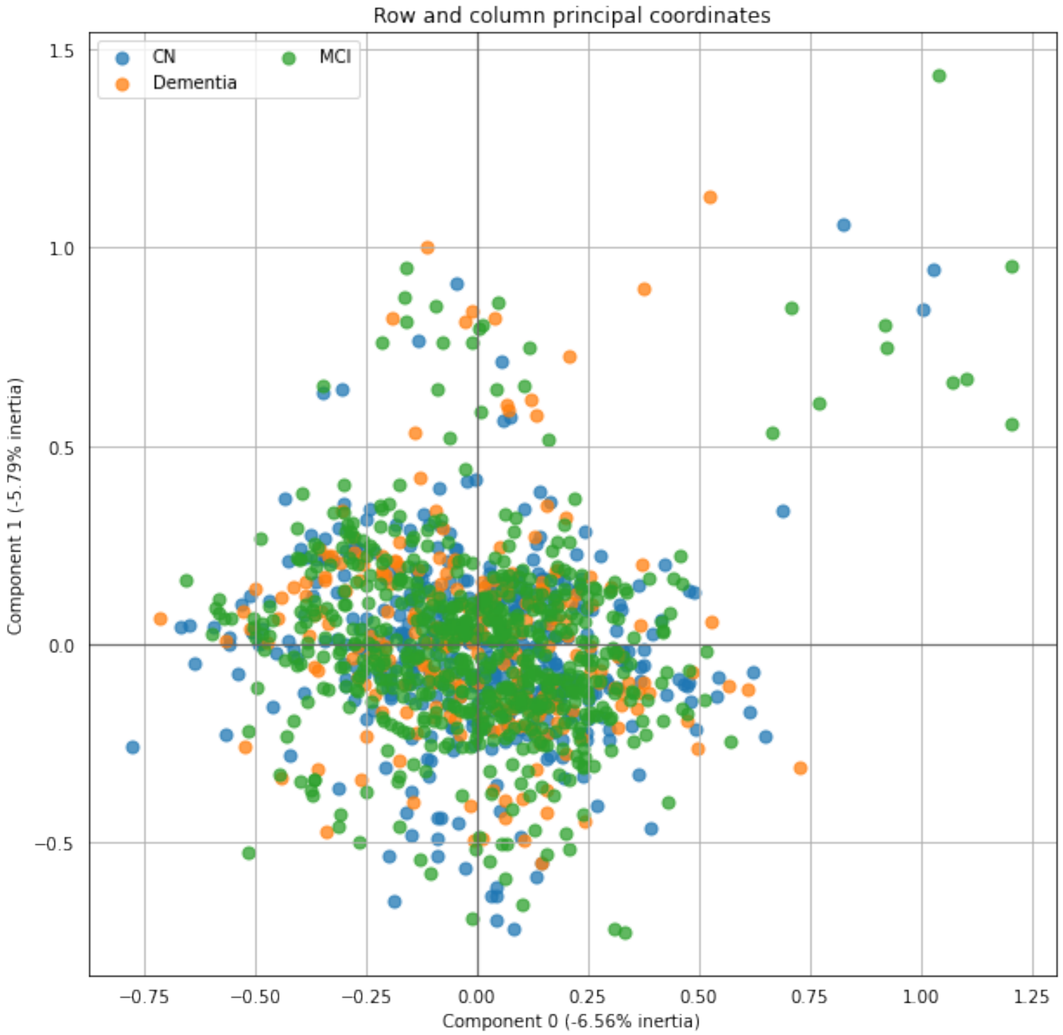
OPNMF - DCCA Training Conclusions

Parameter Action	*Correlation (Negative Loss)*
Output Dimension Size ↑	↑
Hidden Layer Size ↑	↑
Learning Rate	Medium to low LR is best
Batch Size	Stays basically the same
Regularization Parameter	Stays basically the same

Multiple Correspondence Analysis

What is MCA

- Multiple Correspondence Analysis
- Data Analysis Technique for categorical data (i.e. “A” / “B” / “C” or “True” / “False”)
- Similar to Principal Component Analysis
- Use on Genetic Data that have zero, one or two alleles



Brigitte Le Roux και Henry Rouanet.
Multiple correspondence analysis.
Sage Publi- cations, 2010.

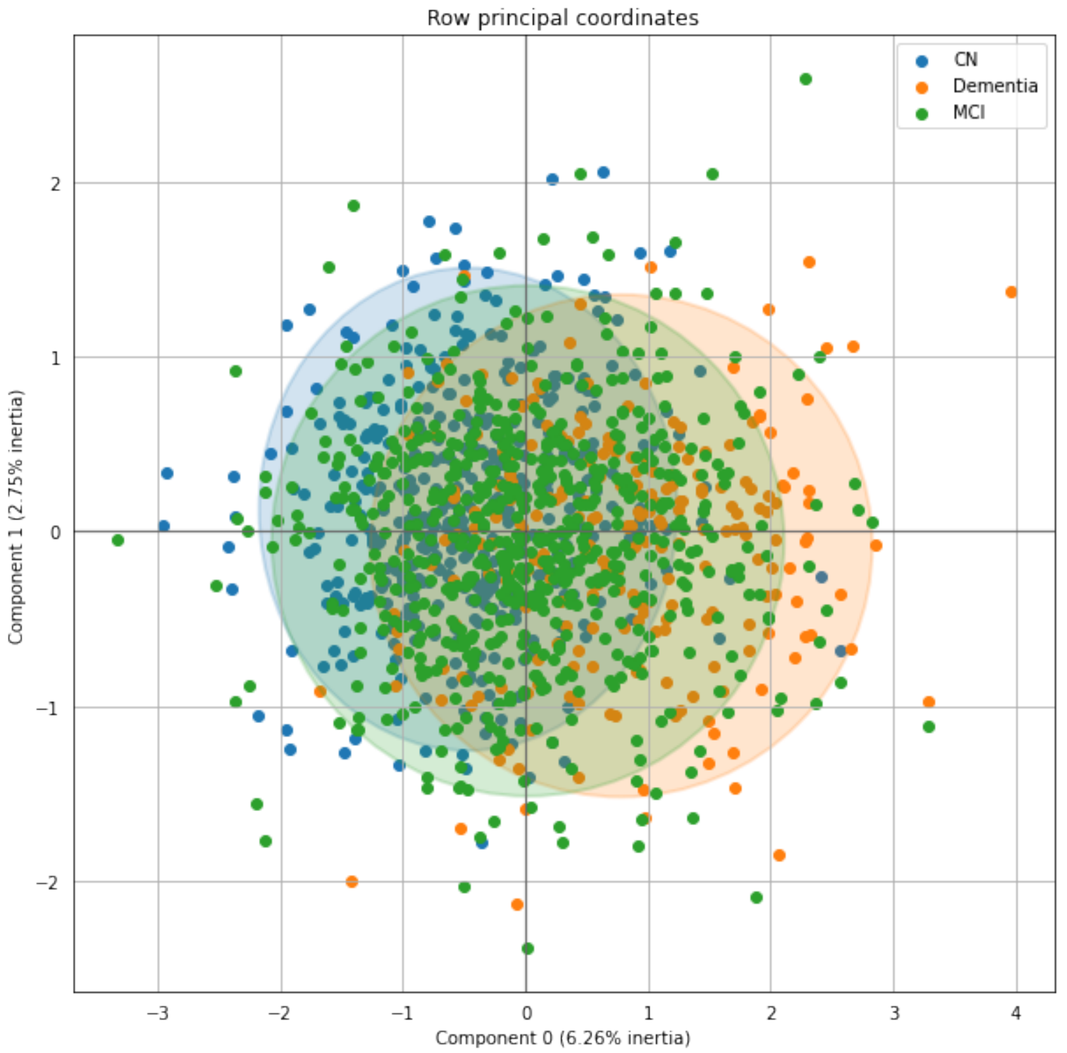
MCA - DCCA Training Conclusions

Parameter Action	*Correlation (Negative Loss)*
Output Dimension Size ↑	↑
Hidden Layer Size ↑	↑
Learning Rate	Medium to low LR is best
Batch Size	Stays basically the same
Regularization Parameter	Stays basically the same

Factor Analysis of Mixed Data

What is FAMD

- Factor Analysis of Mixed Data
- Data Analysis Technique for mixed data (i.e. categorical and numerical data)
- Essentially PCA for numerical & MCA for categorical
- Use on genetic data that take values from [“0”,“1”,“2”] and imaging data that take real values

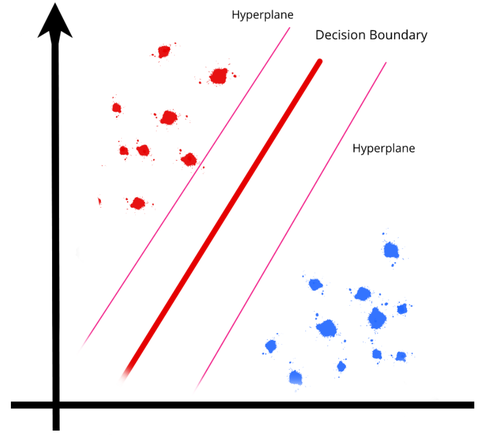


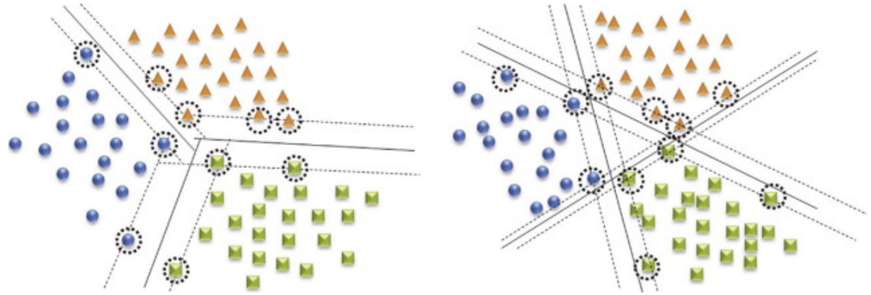
Jérôme Pagès. Chapter 3, Factorial Analysis of Mixed Data.” Multiple Factor Analysis by Example Using R. The R Series, 1:67–72, 2015.

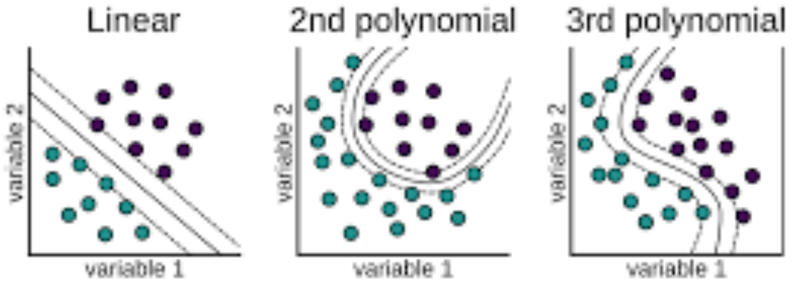
Classification

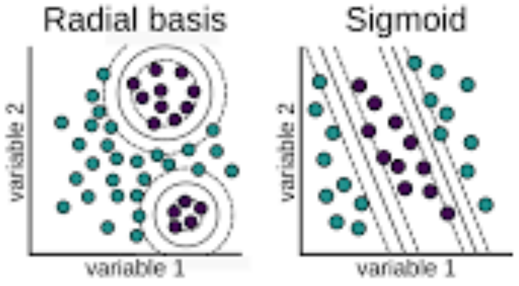
Classification using SVM

- Play around with:
  - Kernel (Linear, Poly, RBF)
  - Iterations
  - Coefficients and gamma params
  - Regularisation
- Use GridSearch
- Use Cross Validation (5 Folds)



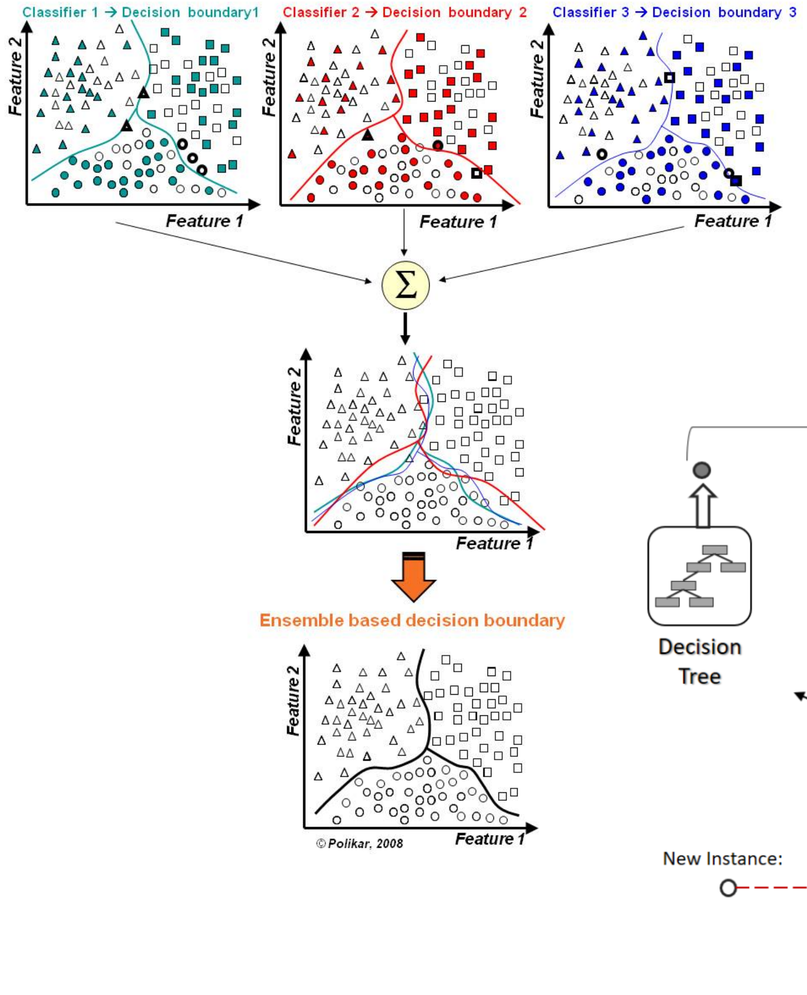


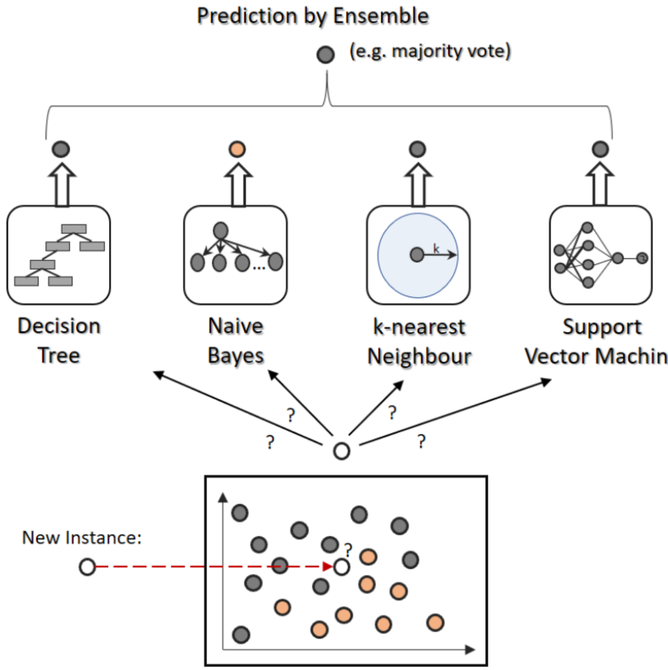




Classification using Ensemble Techniques

- Use the following ensemble algorithms:
  - **Bagging** (**B**ootstrap **Aggregating**)
  - **AdaBoost**
- Utilize the following base model classifiers:
  - **SVM**
  - **Decision Trees**
- Use GridSearch
- Use Cross Validation (5 Folds)





Metrics

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

$$F1\ score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad \Rightarrow \quad F1\ score = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

$$Balanced\ \ Accuracy = \frac{Sensitivity + Specificity}{2} = avg\{\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\}$$

Results

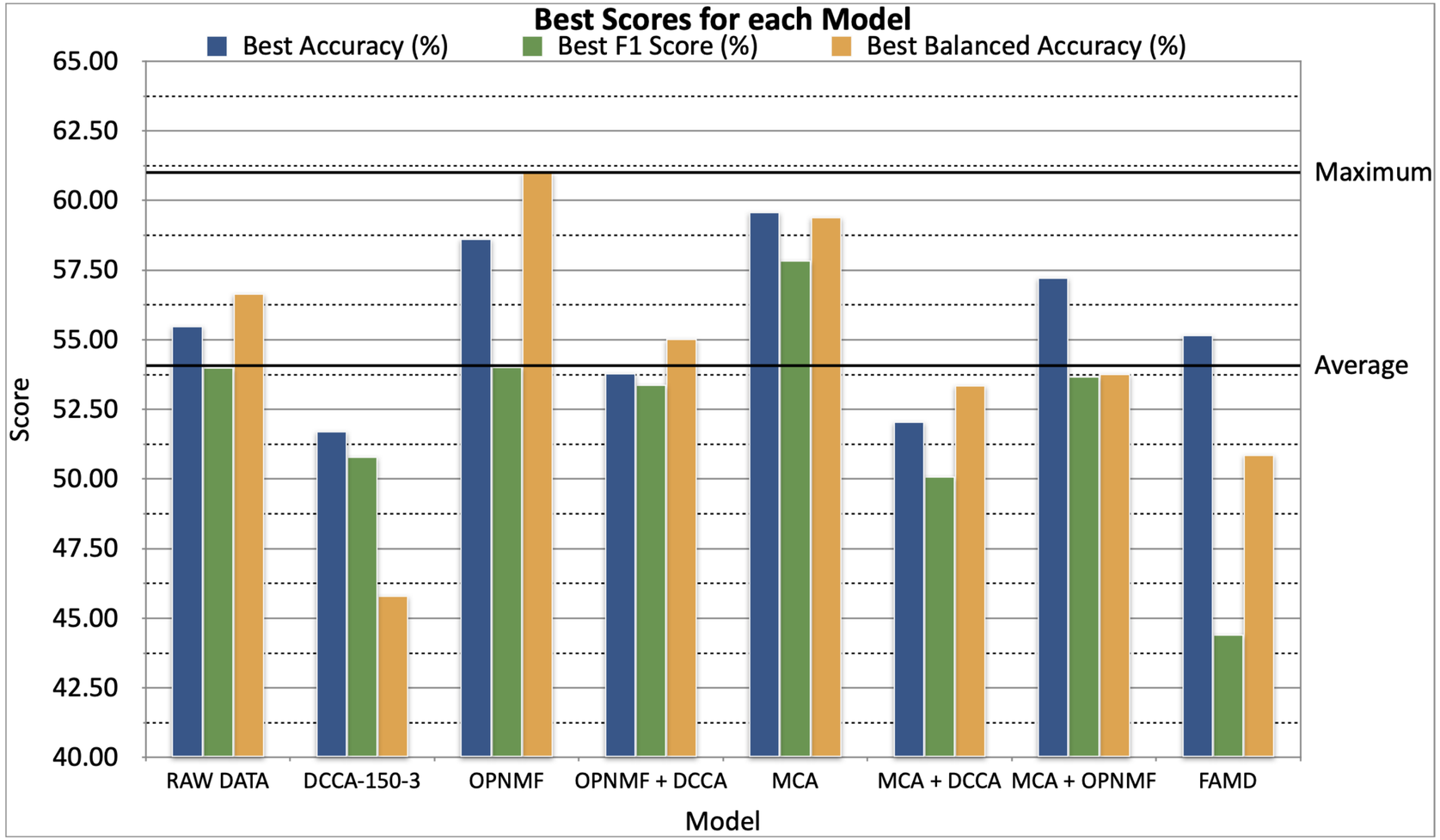
Results

Best scores for each model:

Model	Best Accuracy (%)	Best F1 Score (%)	Best Balanced Accuracy (%)	Notes
RAW DATA	55.48	54.00	56.65	145 ROIs (Scaled) and 54 SNPs (Balanced). Both AdaBoost DT.
DCCA-150-3	51.72	50.80	45.82	Output Dimension 150, 3 Hidden Layers, no scaling or balancing. Imaging Linear.
OPNMF	58.62	54.02	61.01	30 Imaging Components (After OPNMF) Balanced only. Imaging Bagging SVM.
OPNMF + DCCA-150-3	53.79	53.38	55.03	30 Imaging Components (After OPNMF) and 54 SNPs, then DCCA, then scaled and balanced. Both AdaBoost SVM.
MCA	59.59	57.85	59.41	145 ROIs (Scaled), 10 Genetic components, Balanced only. Both Poly SVM.
MCA + DCCA-150-3	52.05	50.10	53.37	145 ROIs and 10 Genetic components (After MCA), then DCCA, then scaled and balanced. Imaging Bagging SVM.
MCA + OPNMF	57.24	53.68	53.77	30 Imaging Components (After OPNMF) and 10 Genetic components (After MCA), Balanced only. Both AdaBoost SVM.
FAMD	55.17	44.42	50.87	10 Components, no scaling, no balancing. Both Poly / RBF SVM.

Results

Picking the best scores for each model:

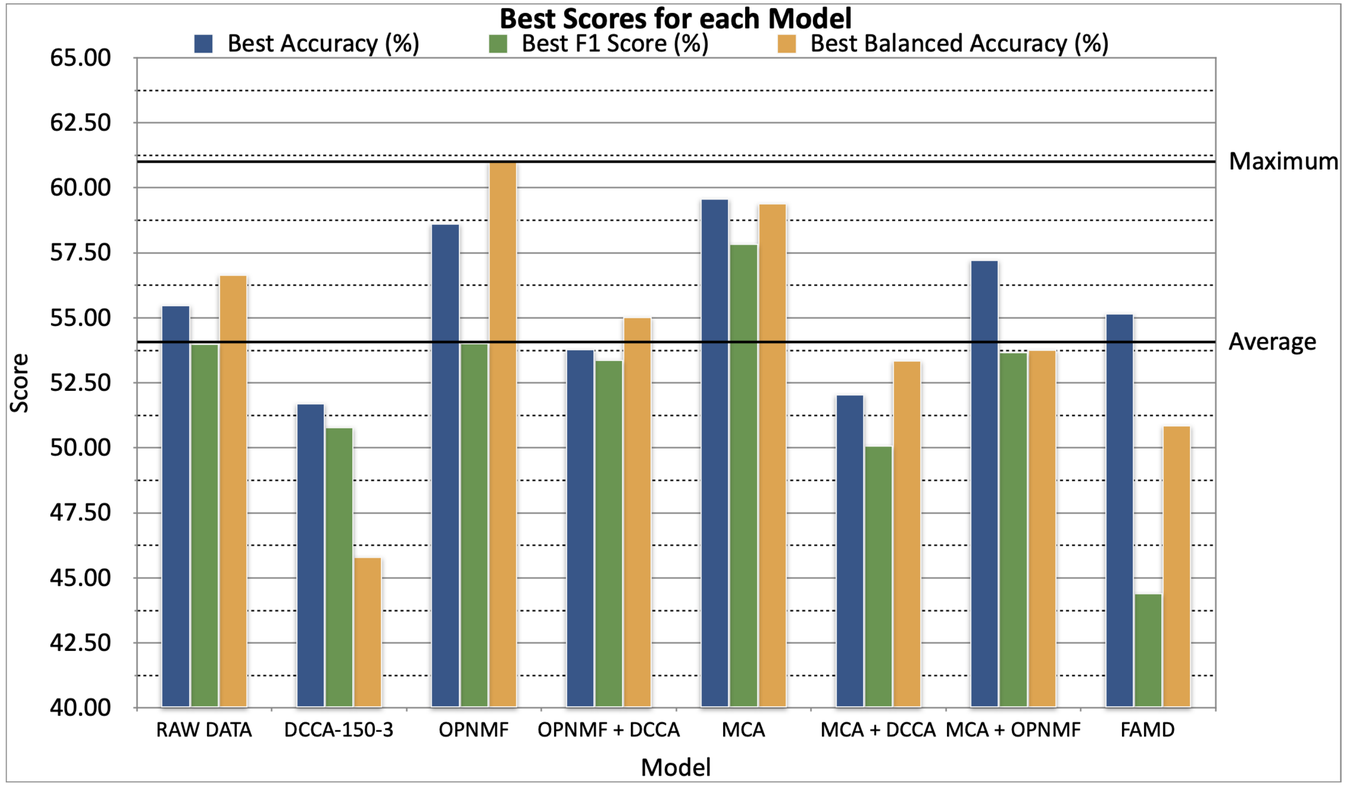


Conclusions

Conclusions (1)

- OPNMF on Imaging-only and MCA on Imaging & Genetic are the best performing models

- Best models use either Imaging & Genetic views or only Imaging. Genetic-only is worse.



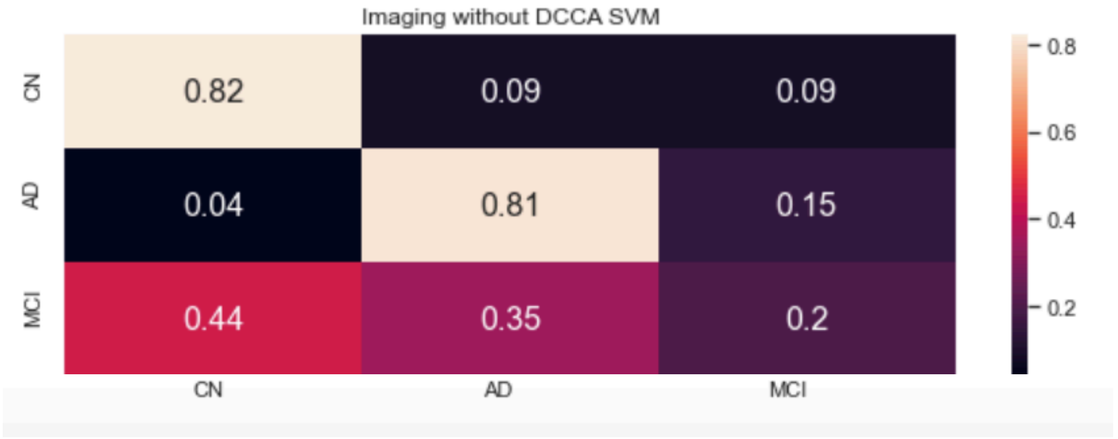
Conclusions (2)

- In most cases the polynomial and RBF kernels outperform the linear SVM kernel (single classifier)
- DCCA increases linear correlation between views.
- Ensemble Classifiers are better than single-classifier models.
- Bagging is better than AdaBoost

Model	Best Accuracy (%)	Best F1 Score (%)	Best Balanced Accuracy (%)	Notes
RAW DATA	55.48	54.00	56.65	145 ROIs (Scaled) and 54 SNPs (Balanced). Both AdaBoost DT.
DCCA-150-3	51.72	50.80	45.82	Output Dimension 150, 3 Hidden Layers, no scaling or balancing. Imaging Linear.
OPNMF	58.62	54.02	61.01	30 Imaging Components (After OPNMF) Balanced only. Imaging Bagging SVM.
OPNMF + DCCA-150-3	53.79	53.38	55.03	30 Imaging Components (After OPNMF) and 54 SNPs, then DCCA, then scaled and balanced. Both AdaBoost SVM.
MCA	59.59	57.85	59.41	145 ROIs (Scaled), 10 Genetic components, Balanced only. Both Poly SVM.
MCA + DCCA-150-3	52.05	50.10	53.37	145 ROIs and 10 Genetic components (After MCA), then DCCA, then scaled and balanced. Imaging Bagging SVM.
MCA + OPNMF	57.24	53.68	53.77	30 Imaging Components (After OPNMF) and 10 Genetic components (After MCA), Balanced only. Both AdaBoost SVM.
FAMD	55.17	44.42	50.87	10 Components, no scaling, no balancing. Both Poly / RBF SVM.

Conclusions (3)

AD vs CN classification is good (Accuracy = 93%, same as published works with similar models and data)



Thank you