The 22nd South Tyrol Free Software Conference (SFSCON)

Data-driven knowledge discovery from Brain Imaging of Alzheimer's disease patients

Prof. Giuseppe Di Fatta

Director of the MSc Computing for Data Science

Computer Science and Artificial Intelligence Institute

Faculty of Engineering

Free University of Bozen-Bolzano, Italy

Giuseppe.DiFatta@unibz.it

NOI Techpark, Bolzano, Italy
10 November 2023

Outline

- Motivations
  - Dementia and Alzheimer's Disease (AD)
- MRI data-driven solutions for Alzheimer's disease classification
  - Data and data processing
  - Machine Learning models
    - outliers detection, feature selection, regression and classification
  - Explainability/Interpretability
- Conclusions

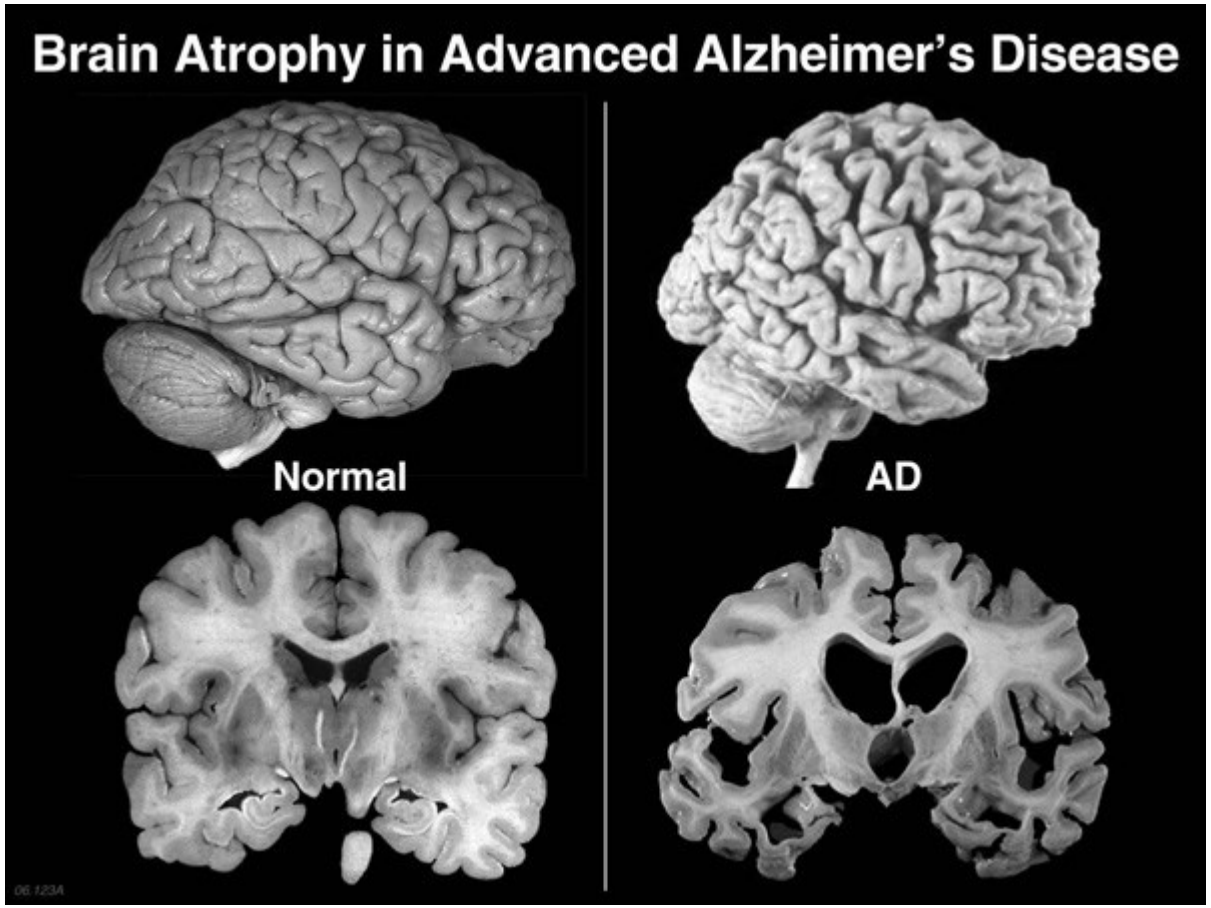
Dementia and Alzheimer's Disease

unibz

- Mental disorders are one of the five most costly conditions for medical spending.
  - Dementia is the most expensive condition for medical and social cost and, still, it receives only a fraction of the research funds w.r.t. other health conditions.

- Alzheimer's Disease (AD): 60-80% of dementia cases
  - Several national and international campaigns aimed at raising awareness and attention (e.g., www.worldalzmonth.org)
  - Currently **40+ million people** suffer from AD: **130-150 million by 2050**, one of the most significant global social economic health crises.
  - The exact cause of AD is unknown.
  - Importance of early and accurate diagnosis: quality of life can be significantly improved.

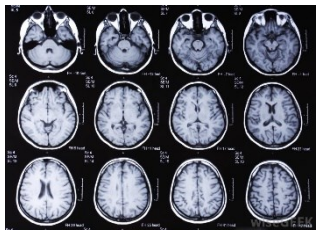
Alzheimer's Disease (AD)

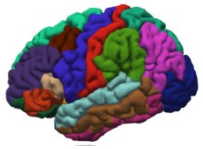


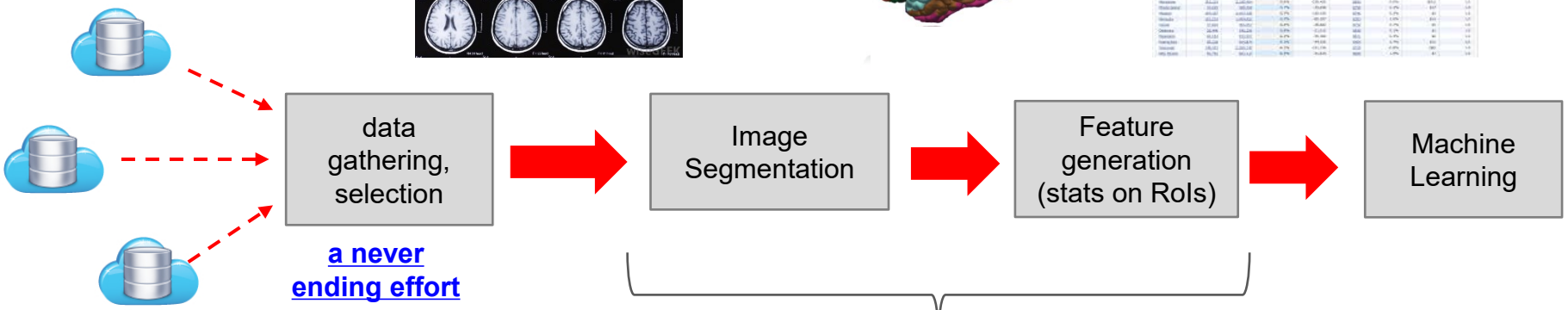
Data Workflow

unibz

Brain MRIs from
3170 subjects





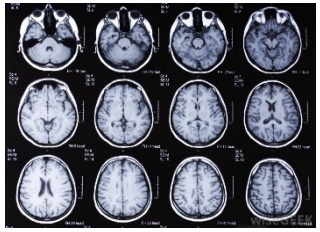


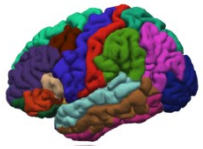
Pre-processing

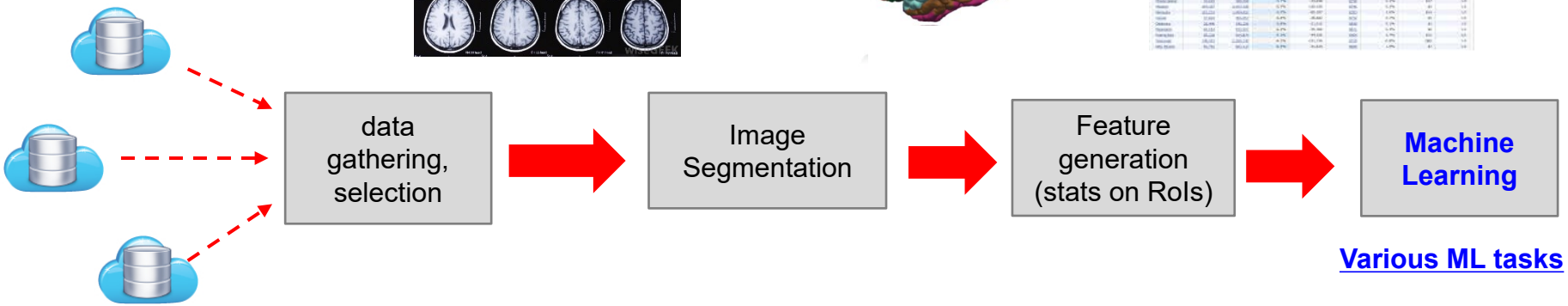
- A single MRI scan file typically takes 30-40 MB.
- After pre-processing the files generated for a single image take 350-500 MB.
- Pre-procesing of a single image takes ~8h ➔ several months or even years
- A **data parallel** approach used to speed up computation ➔ a few months

Data Workflow

Brain MRIs from
3170 subjects



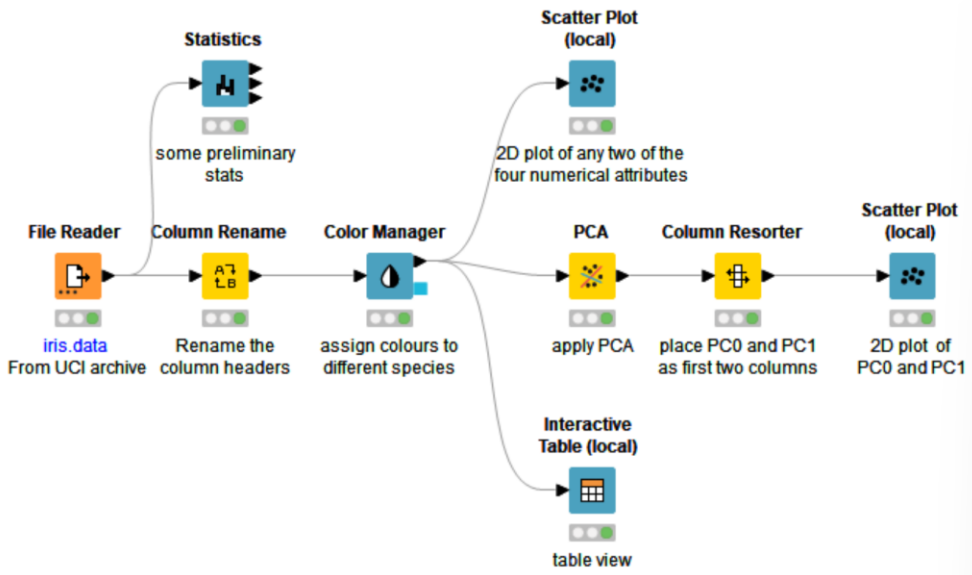




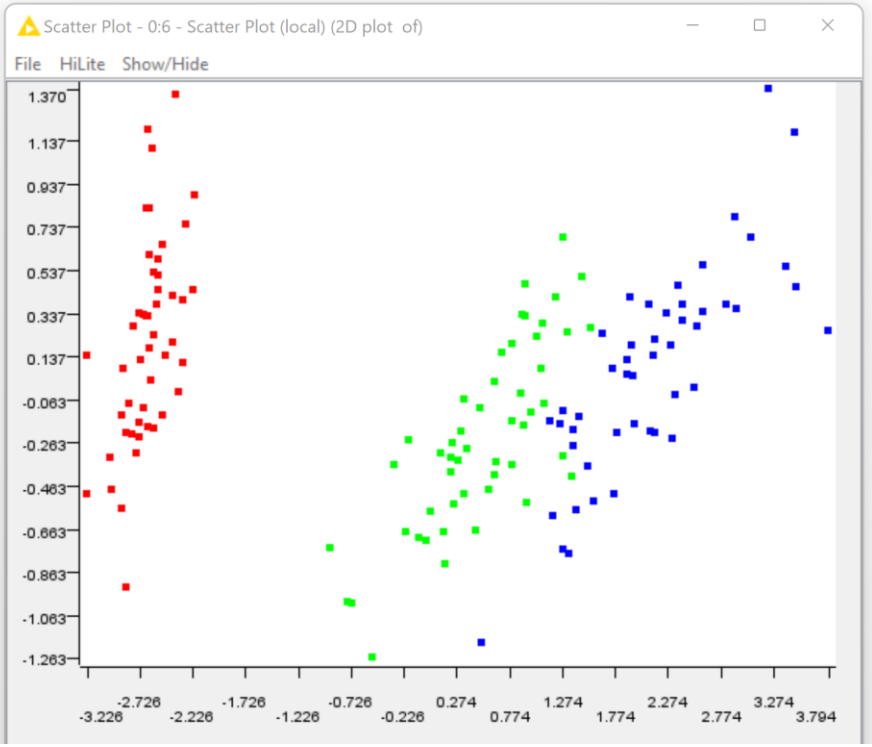
Machine Learning tasks

- Outlier Detection
- Feature selection
- Training predictive models
  - Regression for age estimation
  - Classification: binary and multiclass predictive models
- Testing the models: robust performance estimation
- Explainability and Interpretability

- KNIME is an **open-source** Data Science and Machine Learning Platform
  - codeless Machine Learning and Data Science
  - open-source and free (General Public License)
  - open and extendible platform



a KNIME data workflow
(an acyclic graph of computations)



Various Approaches since 2014

unibz

Binary (HC vs AD) and multiclass classification problems

- Input features:
  - No selection: all available (~400) numerical measurements from brain ROIs
  - Domain-specific feature selection: 16 hippocampal subfields
  - Automatic feature selection
    - Filter/Wrapper/Embedded methods

- Classification methods I – focus on predictive accuracy:
  1. Binary Region Classifier
  2. Probability Weight-based Classification
  3. Bayesian Classifier
  4. Support Vector Machine
  5. Deep Neural Networks

- Classification methods II – predictive accuracy and descriptive power:
  - brain age estimation

Predictive Task: Regression

unibz

- Input data: structural MRI scans (T1 weighting) of human brains
  - Healthy Control (HC) only or in combination with
  - Alzheimer's Disease (AD) (or a different condition)

- Regression task:
  - Estimate the subject's age from a structural MRI of the brain
    - e.g., Brain Age Gap Estimation (BrainAGE)
  - Challenges: data manipulation, accuracy, curse of dimensionality, confidence of models

Brain Age Gap Estimation

unibz

- Brain Age Gap Estimation (BrainAGE)
  - BrainAGE quantifies the difference between the actual anagraphic age of a subject and their age as predicted by ML models applied to neuroimaging data
  - It is considered a biomarker of brain health.
  - Typical method: age model of healthy subjects
    - Voxel-Based Morphometry (VBM) approach: PCA applied to 3,700 voxels to reduce the dimensionality of the input space
    - Relevance Vector Regression (RVR) and Support Vector Regression (SVR)

The “Apparent” Brain Age

unibz

- The **Apparent Brain Age (ABA)** for a specific neurodegenerative condition
  - goal-conditioned brain age estimation based on region-wise segmentation
  - inductive bias based on feature selection in order to improve the classification accuracy of the estimated brain age

Objectives: simplest model with high accuracy, interpretability and specificity

- ✓ the design of a **data workflow** with a combination of **only linear models** to preserve the original input space semantics
- ✓ the definition of a novel **feature score** to measure the specific contribution of each selected morphological region to the classification prediction
- ✓ a rigorous and extensive **experimental comparative analysis** to validate the method: comparable or superior accuracy than SOTA ML methods
- ✓ demonstrate the **interpretability** of the classification method

- *A. Varzandian, M.A.S. Razo, M.R. Sanders, A. Atmakuru, G. Di Fatta, “**Classification-biased apparent brain age for the prediction of Alzheimer's disease**”, Frontiers in Neuroscience (581), 2021.*

ABA, Classification and Interpretability

unibz

1. Biased Forward Feature Selection (BFFS) as coarse-grained method: cross-validation on a correlation-based filter
2. Linear regression is applied to the selected features: LASSO regression with fine-grained feature selection method
3. Logistic regression based only on ABA and actual age
4. Interpretability with a measure of feature importance/score

$$aba = a_0 + \sum_{i=1}^{k} a_i \cdot f_i,$$

$$c_0 + c_1 \cdot age + c_2 \cdot aba < 0$$

$$\sum_{i=1}^{k} -\frac{c_2 \cdot a_i \cdot f_i}{c_0 + c_1 \cdot age + c_2 \cdot a_0} > 1$$

where $\{f_i\}$ are the k ROI features, a_i and c_j are model parameters

➔Feature Score:

$$s_i = -\frac{c_2 \cdot a_i \cdot f_i}{c_0 + c_1 \cdot age + c_2 \cdot a_0}$$

with $\sum_{i=1}^{k} s_i > 1$

Feature Selection (step 1)

unibz

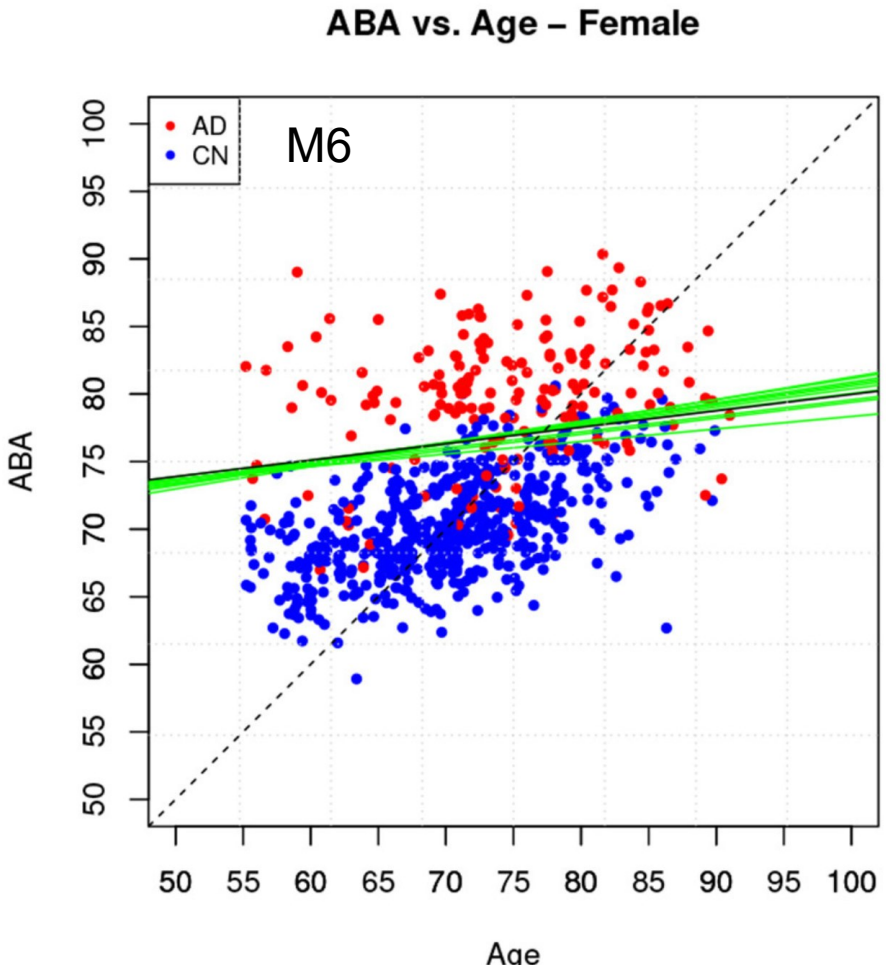
- results based on 1901 subjects (AD and CN)

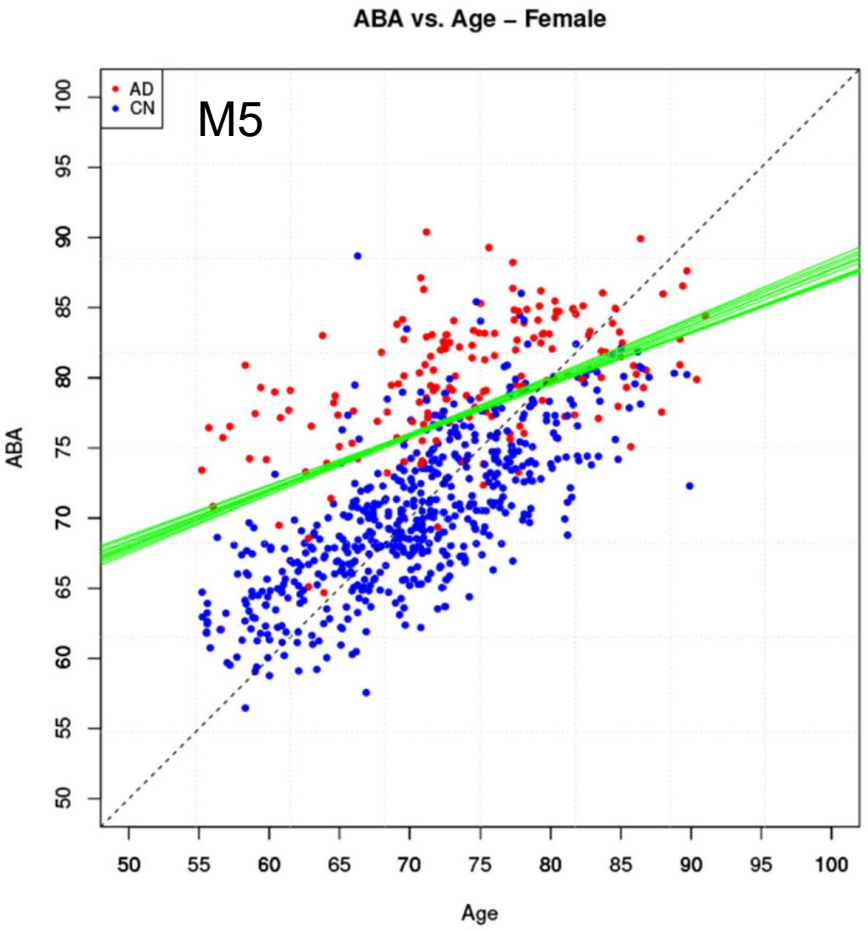
TABLE ROIs selected over all folds of all cross-validation trials for both genders.

	F&M			F			M		
ROI	**LH&RH**	**LH**	**RH**	**LH&RH**	**LH**	**RH**	**LH&RH**	**LH**	**RH**
Entorhinal_thickness	**76%**	**97%**	**54%**	**70%**	**96%**	44%	**81%**	**98%**	**64%**
Whole_hippocampus	**55%**	**100%**	9%	**56%**	**100%**	12%	**53%**	**100%**	6%
Middletemporal_thickness	**52%**	**61%**	43%	**55%**	34%	**76%**	49%	**88%**	10%
Subiculum	44%	39%	49%	**50%**	**66%**	34%	38%	12%	**64%**
CA1	35%	45%	24%	23%	32%	14%	46%	**58%**	34%
Molecular_layer_HP	34%	**67%**	1%	17%	34%	0%	**51%**	**100%**	2%
Amygdala	37%	16%	**58%**	22%	16%	28%	**52%**	16%	**88%**
Hippocampal_tail	25%	31%	19%	19%	20%	18%	31%	42%	20%
Presubiculum	25%	20%	30%	25%	12%	38%	25%	28%	22%
HATA	13%	9%	17%	20%	14%	26%	6%	4%	8%
GC_ML_DG	12%	16%	7%	4%	0%	8%	19%	32%	6%
Fusiform_thickness	11%	15%	7%	4%	4%	4%	18%	26%	10%
Bankssts_thickness	11%	17%	4%	18%	30%	6%	3%	4%	2%
Middletemporal_volume	11%	13%	8%	14%	18%	10%	7%	8%	6%
Inf_Lat_Vent	10%	13%	7%	0%	0%	0%	20%	26%	14%
Inferiortemporal_thickness	10%	8%	12%	9%	10%	8%	11%	6%	16%
CA3	9%	10%	8%	5%	4%	6%	13%	16%	10%
CA4	9%	6%	11%	10%	6%	14%	7%	6%	8%

Only features selected at least in 10% of the models either for the Left Hemisphere (LH) or the Right Hemisphere (RH) are included. ROIs are listed in decreasing order of the total frequency in all groups and both hemispheres. (Frequencies greater than or equal to 50% are in bold).

ABA Prediction and Classification (2 and 3)

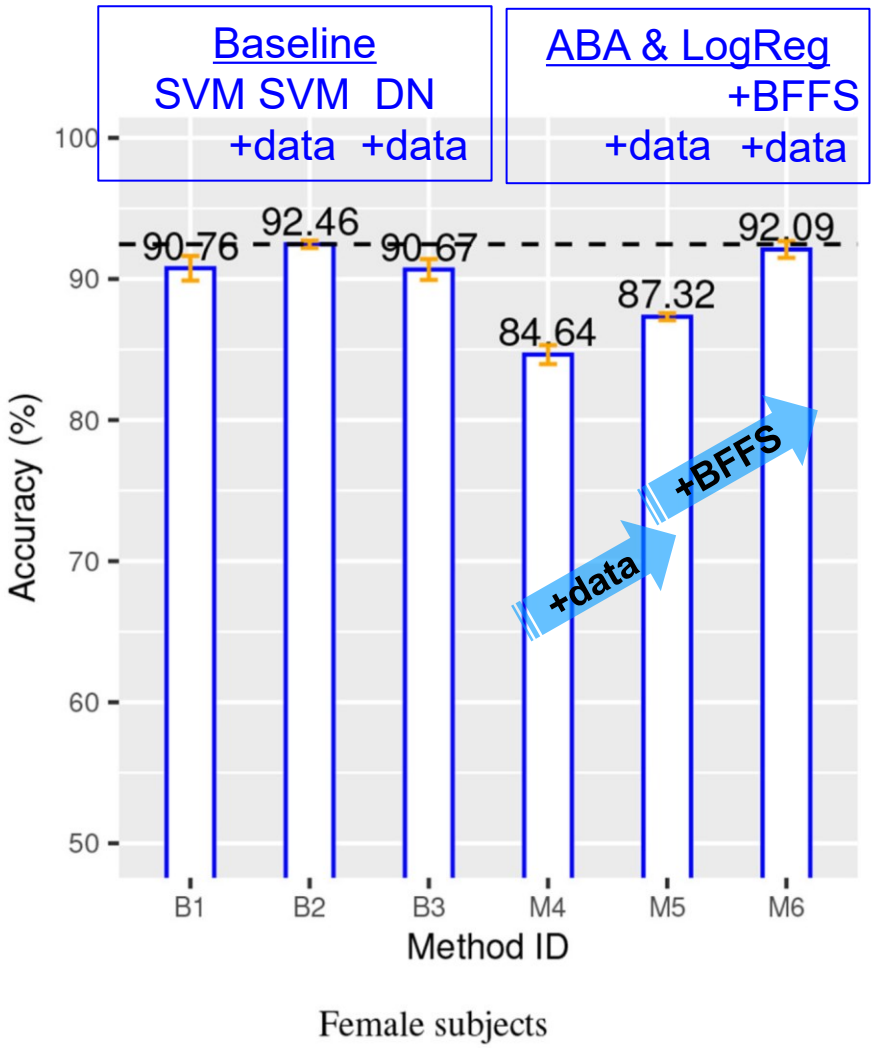




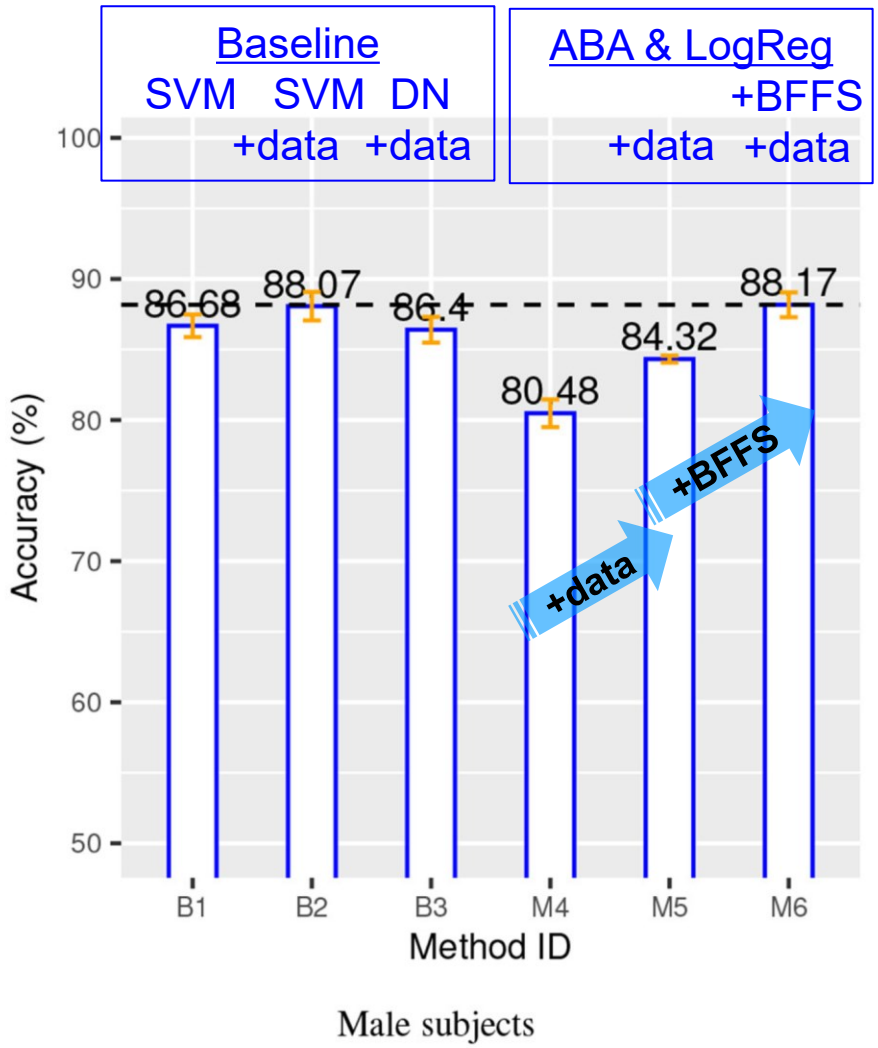
Method M6: Apparent Brain Age vs. chronological age for the female gender group. The plot on the left shows ABA vs. age; the plot on the right shows the ADS vs. age for the same data and with the equivalent boundary line. The ABA values are generated as test set predictions from the folds of a single cross-validation execution. The solid black line is the logistic regression decision boundary of the final model; the green lines are the boundaries in the individual folds of cross-validation; the dashed line is included as reference. The black boundary line is indicative and is obtained from the single final model trained on all data.

Classification Accuracy (3)

- results based on 1901 subjects (AD and CN)

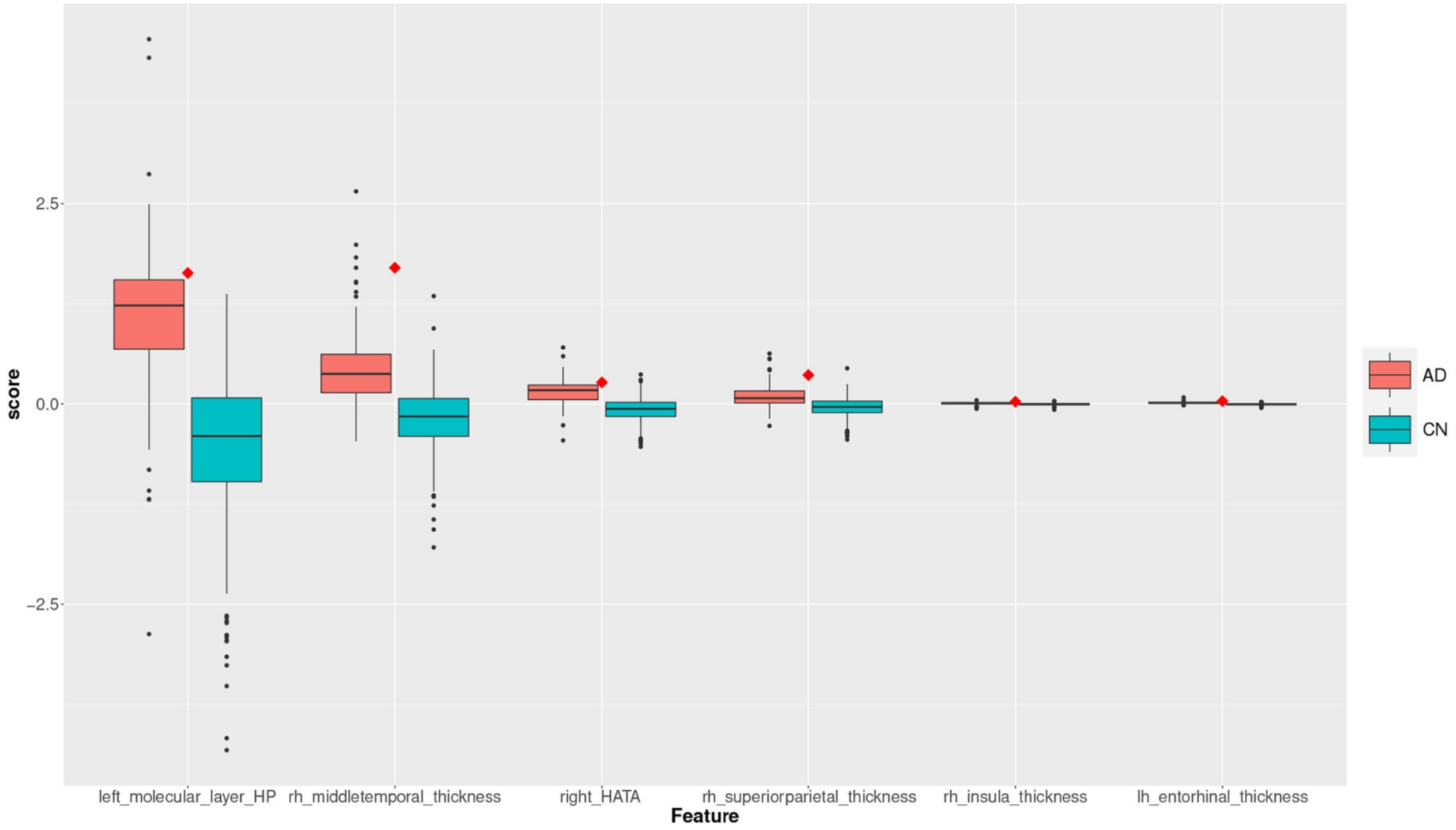


Female subjects

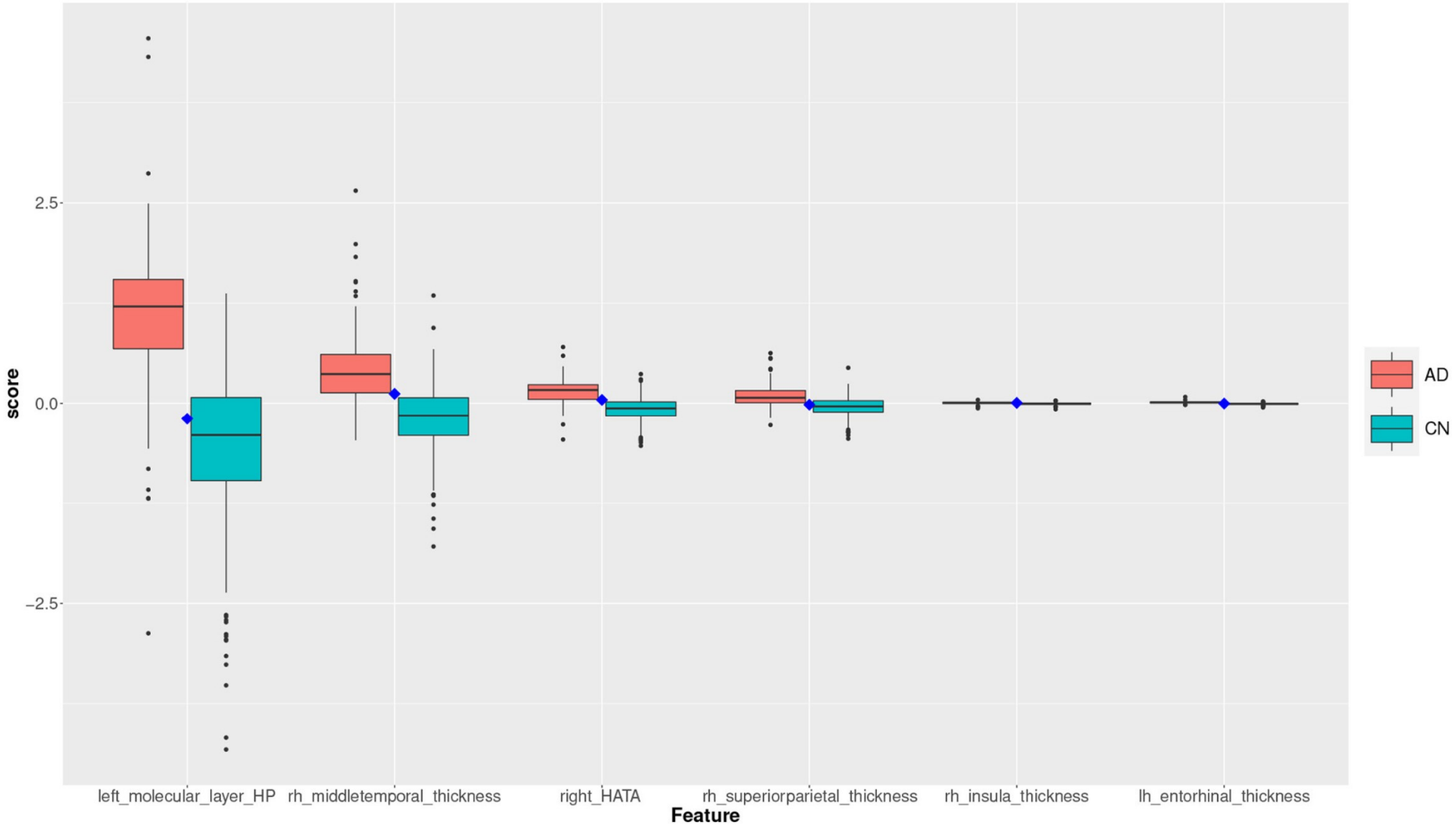


Male subjects

Feature Score (4): True Positive



Feature Score (4): True Negative



Conclusions

unibz

- With same pre-processing and features most classification methods are similar in terms of accuracy. However, black-box approaches do **lack descriptive capabilities**, especially in such high-dimensional spaces.
- A new candidate biomarker, the **Apparent Brain Age**, helps at improving predictive accuracy and can be useful on itself for its descriptive power.
- The new feature score helps to link age gaps to specific ROIs for a powerful combination of
  - **state-of-the-art accuracy**
  - **interpretability**
  - **specificity**

- Ongoing effort is directed to experimental analysis of multiclass problems and specificity.
- Data from other neurodegenerative diseases to extend the work further.

Relevant Publications

- A. Varzandian, M.A.S. Razo, M.R. Sanders, A. Atmakuru, G. Di Fatta, "**Classification-biased apparent brain age for the prediction of Alzheimer's disease**", Frontiers in Neuroscience (581), 2021.
- E.E. Bron et al., "**Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge**", NeuroImage, an Elsevier Journal of Brain Function, 1 May 2015, 111:562-79.
- A. Spedding, G. Di Fatta and J.D. Saddy, "**An LDA and Probability-based Classifier for the Diagnosis of Alzheimer's Disease from Structural MRI**", Workshop on High Performance Bioinformatics and Biomedicine (HiBB), the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Nov. 9-12, 2015, Washington D.C., pp. 1404 - 1411.
- A. Spedding, G. Di Fatta and M. Cannataro, "**A Genetic Algorithm for the Selection of Structural MRI Features for Classification of Mild Cognitive Impairment and Alzheimer's Disease**", WS on Machine Learning in decision making for biomedical applications, IEEE Int.l Conf. on Bioinformatics and Biomedicine, Nov. 9-12, 2015, Washington D.C., pp. 1566-1571.
- A. Sarica, G. Di Fatta, G. Smith, M. Cannataro, D. Saddy, "**Advanced Feature Selection in Multinominal Dementia Classification from Structural MRI Data**", in the Proc. of the Workshop on Computer-Aided Diagnosis of Dementia based on structural MRI data (CADDementia), Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2014, Boston, MA, USA
- A. Sarica, G. Di Fatta, M. Cannataro, "**K-Surfer: A KNIME extension for the management and analysis of human brain MRI FreeSurfer/FSL Data**", Int.l Conf. on Brain Informatics and Health, 11-14 August 2014, Warsaw, Poland, Springer LNCS V.8609, 2014, pp. 481-492.
- M. Berthold, N. Cebron, F. Dill, G. Di Fatta, T. Gabriel, F. Georg, T. Meinl, P. Ohl, C. Sieb, B. Wiswedel, "**KNIME: the Konstanz Information Miner**", Proc. 4th Annual Industrial Simulation Conference (ISC), Palermo, Italy, June 5-7, 2006, pp.58-61.