Correct approach to multicentric data harmonization

P. Oliva, G. Serra, F. Mainas, B. Golosio, A. Retico

Multicentric datasets

- ML/DL are essential tools in data analysis in neuroimaging, due to the large number of features and the intrinsic multivariate nature of the problems.
- ML/DL algorithms need **datasets of appropriate size**, in order to be correctly trained.
- However, in this field, large datasets are often obtained by **collecting images from different centers**, thus bringing unavoidable bias in the analysis, due to differences in hardware and scanning protocols between different centers
- This ***site effect*** needs to be properly addressed to eliminate or reduce the associated bias

Harmonization of multicentric datasets

- The typical approach to **feature harmonization** on multicentric datasets is the COMBAT procedure:

$$y_{ijv}^{\text{ComBat}} = \frac{y_{ijv} - \widehat{\alpha}_v - \mathbf{X}_{ij}\widehat{\boldsymbol{\beta}}_v - \gamma_{iv}^*}{\delta_{iv}^*} + \widehat{\alpha}_v + \mathbf{X}_{ij}\widehat{\boldsymbol{\beta}}_v$$

Harmonization of multicentric datasets

- The typical approach to harmonization on multicentric datasets is the COMBAT procedure:

$$y_{ijv}^{\text{ComBat}} = \frac{y_{ijv} - \widehat{\alpha}_v - \mathbf{X}_{ij}\widehat{\boldsymbol{\beta}}_v - \gamma_{iv}^*}{\delta_{iv}^*} + \widehat{\alpha}_v + \mathbf{X}_{ij}\widehat{\boldsymbol{\beta}}_v$$

Harmonization of multicentric datasets

- The typical approach to harmonization on multicentric datasets is the COMBAT procedure:

$$y_{ijv}^{\text{ComBat}} = \frac{y_{ijv} - \widehat{\alpha}_v - \mathbf{X}_{ij}\widehat{\boldsymbol{\beta}}_v - \gamma_{iv}^*}{\delta_{iv}^*} + \widehat{\alpha}_v + \mathbf{X}_{ij}\widehat{\boldsymbol{\beta}}_v$$

Harmonization of multicentric datasets

- The typical approach to harmonization on multicentric datasets is the COMBAT procedure:

$$y_{ijv}^{\text{ComBat}} = \frac{y_{ijv} - \widehat{\alpha}_v - \mathbf{X}_{ij}\widehat{\boldsymbol{\beta}}_v - \gamma_{iv}^*}{\delta_{iv}^*} + \widehat{\alpha}_v + \mathbf{X}_{ij}\widehat{\boldsymbol{\beta}}_v$$

- The COMBAT harmonization is often performed on the complete dataset, as a ***preprocessing*** step.
- However, in a rigorous validation scheme, this is not a correct approach since there is a potential data leakage from any successively defined ***test set*** to the corresponding ***training set***.

ABIDE

The Autism Brain Imaging Data Exchange (ABIDE) initiative has aggregated **functional** and **structural** brain imaging data collected from laboratories around the world to accelerate our understanding of the neural bases of autism.
It was created through the aggregation of datasets independently collected across more than 24 international brain imaging laboratories and are being made available to investigators throughout the world, consistent with open science principles.

ABIDE I

- 17 international sites
- 1112 subjects,
  - 539 ASD
  - 573 typical developing (TD) controls
- ages 7-64 years
- ABIDE I preprocessed

ABIDE II

- 19 international sites (7 new members)
- 1114 subjects,
  - 521 ASD
  - 593 TD
- ages 5-64 years
- Preprocessed data are not available for ABIDE II

Features

We investigated both:

5995 functional connectivity features: processed using C-PAC, which is a configurable, open-source pipeline. Corrections steps: motion correction, slice timing correction, band-pass filtering, spatial smoothing, and registration.

Time series are extracted using the Harvard-Oxford atlas which outlines 110 ROIs. The (static) connectivity is evaluated using the Pearson correlation (with Fischer z-transform). PCA was also investigated on these data, nested in the CV.

221 structural features: processed with Freesurfer 6.0 with the recon-all pipeline, 186 cortical measures (volume, mean and standard deviation of the thickness, evaluated on the DKT altas, 62 regions), 26 volumes of sub-cortical structures and 9 global quantities

Data selections

We used two different data selections:

- **minSC:** males, 6-40yrs, successful reconstruction for both Freesurfer and C-PAC
- **fQC:** males, open eyes, FIQ>0, fd_mean_d<=3*MAD, successful reconstruction for both Freesurfer and C-PAC

	minSC	minSCas	fQC	fQCas
TD	344	265	320	247
ASD	344	255	298	226
Total	688	520	618	473

For each selection, we also reduced the investigation to **9-20yrs**, removing sites with less than 30 subjects (**as**).

Classification scheme

For the classification, we used a rbf SVM.

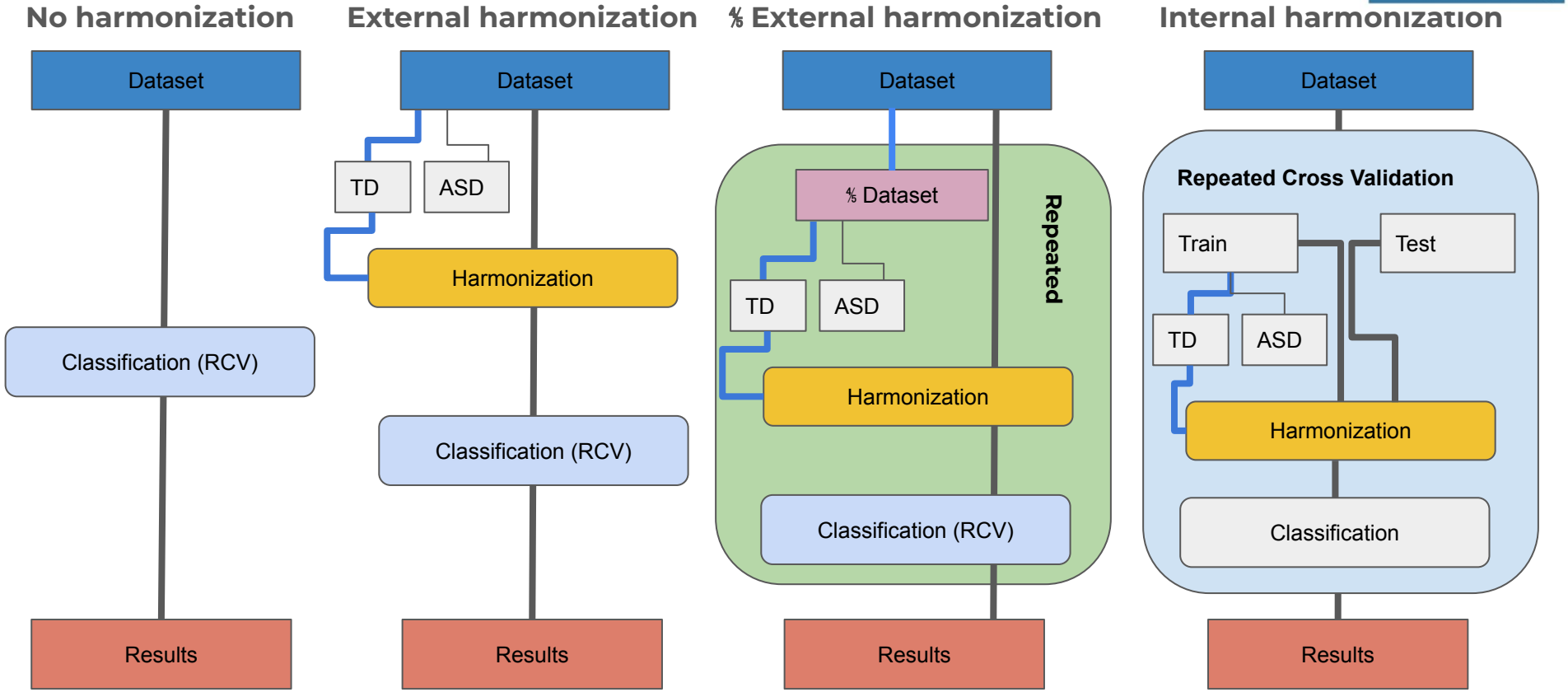
The used metric is the Area Under the ROC Curve (AUC).

We used a stratified k-fold (k=5) cross validation (CV) scheme.

The k-fold CV is repeated 50 times.

Results are reported as average and standard deviation of the AUC over the 50 repetition.

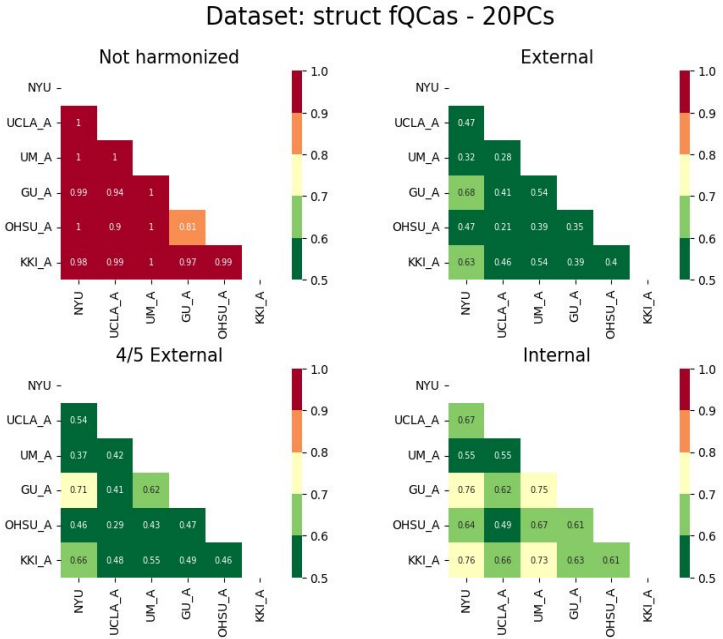
Harmonization approaches

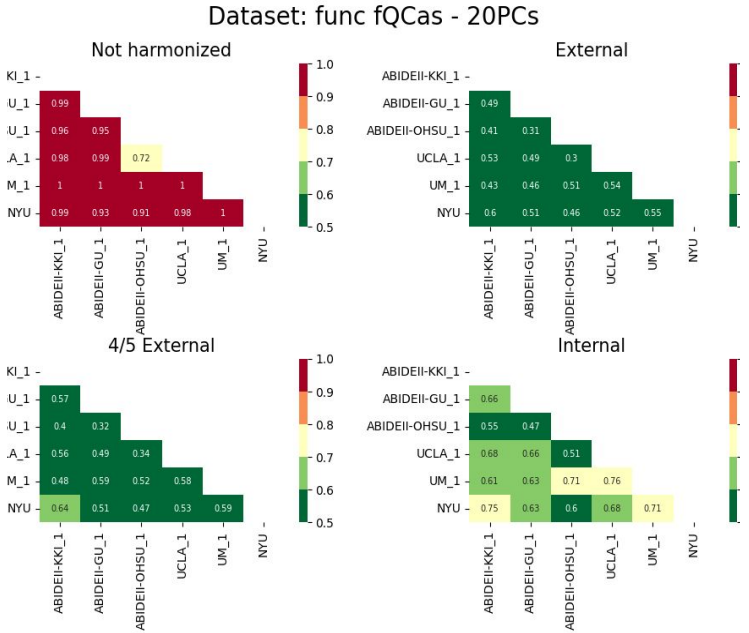


Site distinguibility

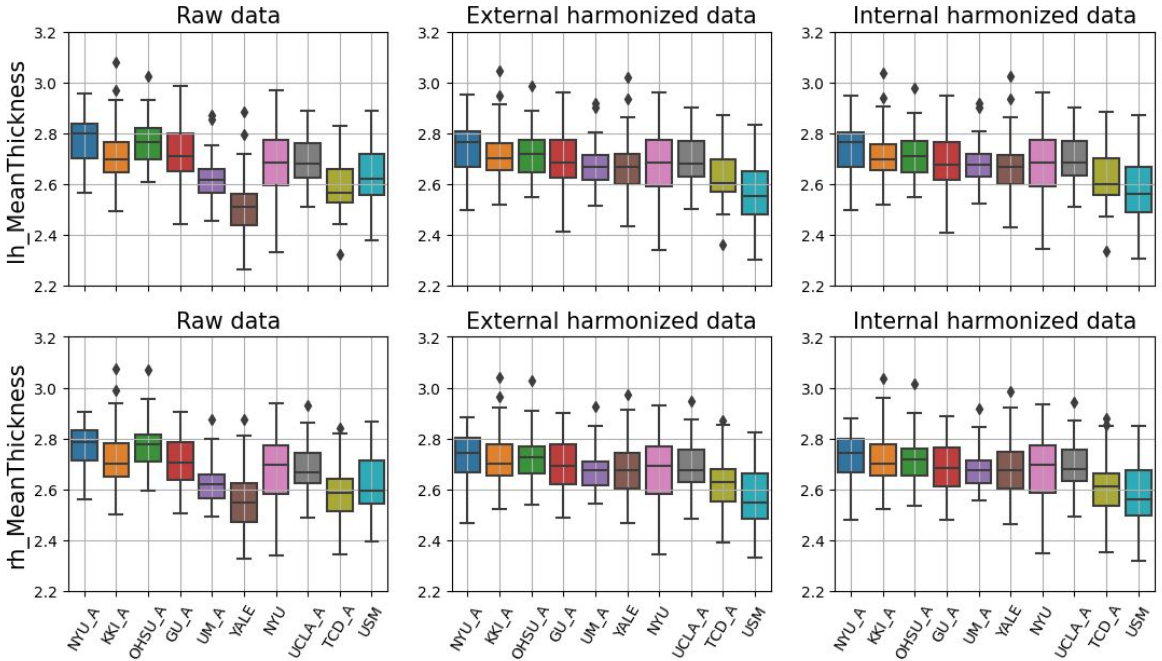
Only TDs, Site as label

In order to limit LSS-LES effects, only sites with more than **20 TDs** are considered.
Given the small size of the samples, we limited the analysis to the **first 20 PCs**.



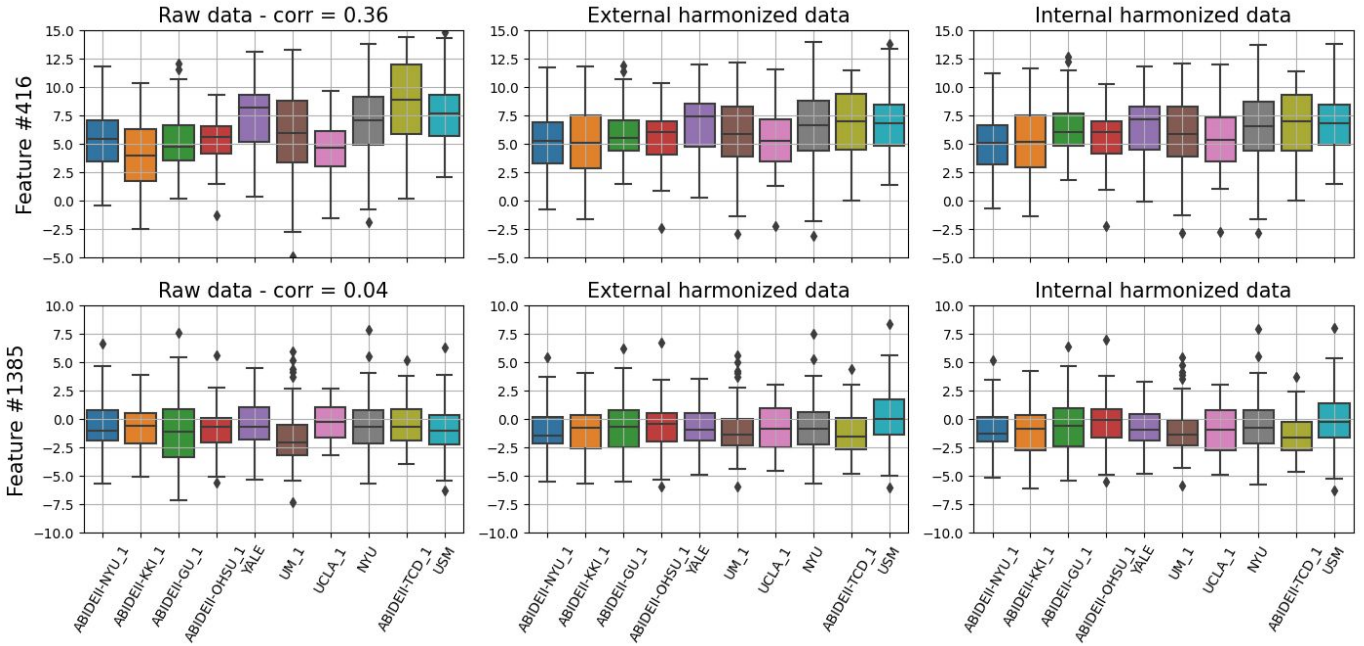


Age dependence



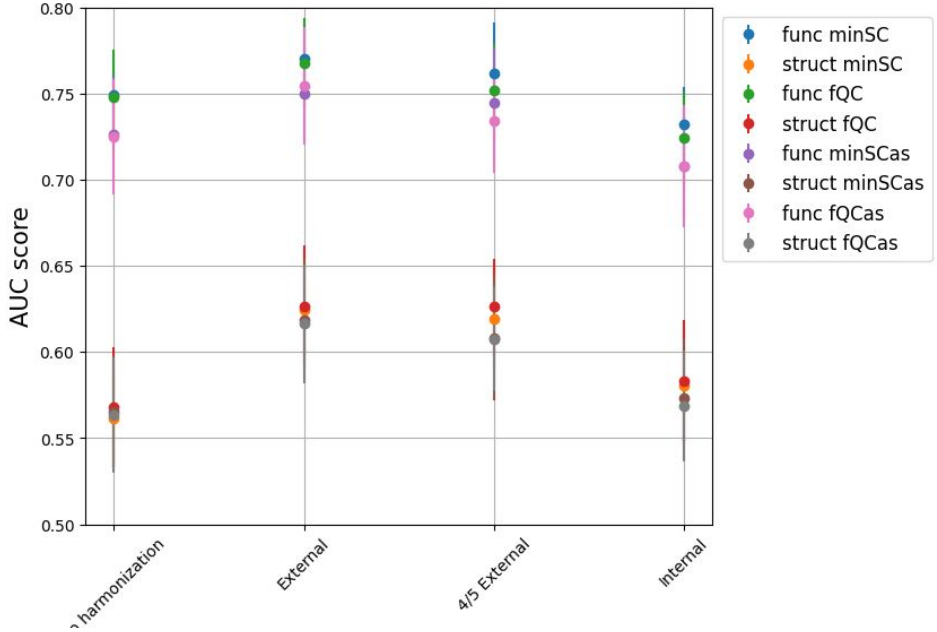
Structural features

Age dependence

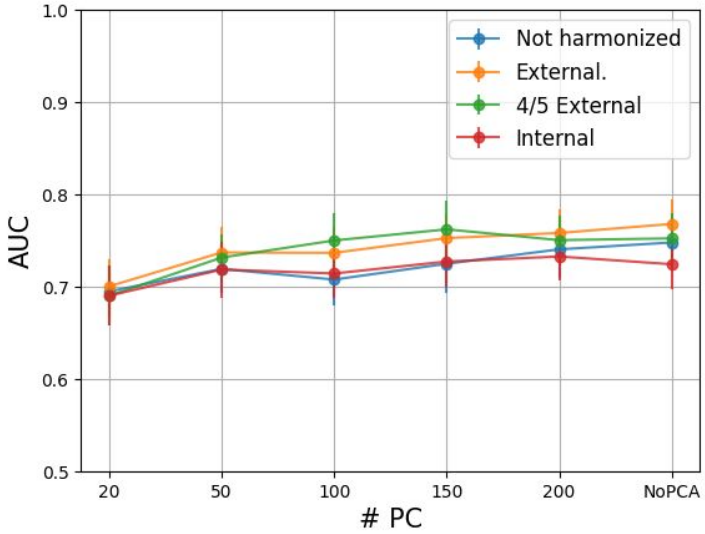


Connectivity features

Classification results



No PCA



Dependence on number of PCs (functional features)

Conclusions

- **The harmonization strategy affects the classification results**.
- If the **whole dataset** is used to harmonize, the **performances are higher**
    - These higher performances can be observed for **both structural and functional features**
    - They can be observed also if the size of the sample used to estimate the harmonization parameters is the same
- Potential **data leakage** is possible for these approaches are, since the training set is not completely blind to the test set
- We think that the ***internal* approach is hence the correct one**, since the test set is never used before the test step
- It would be interesting to investigate the **effect of different approaches in feature importance** (work in progress…)

Combining Structural and Functional MRI-based brain features with a Deep Learning Joint Fusion approach

Sara Saponaro, Francesca Lizzi, Giacomo Serra, Lorenzo Marini, Alessia Giuliano, Pernicola Oliva, Alessandra Retico

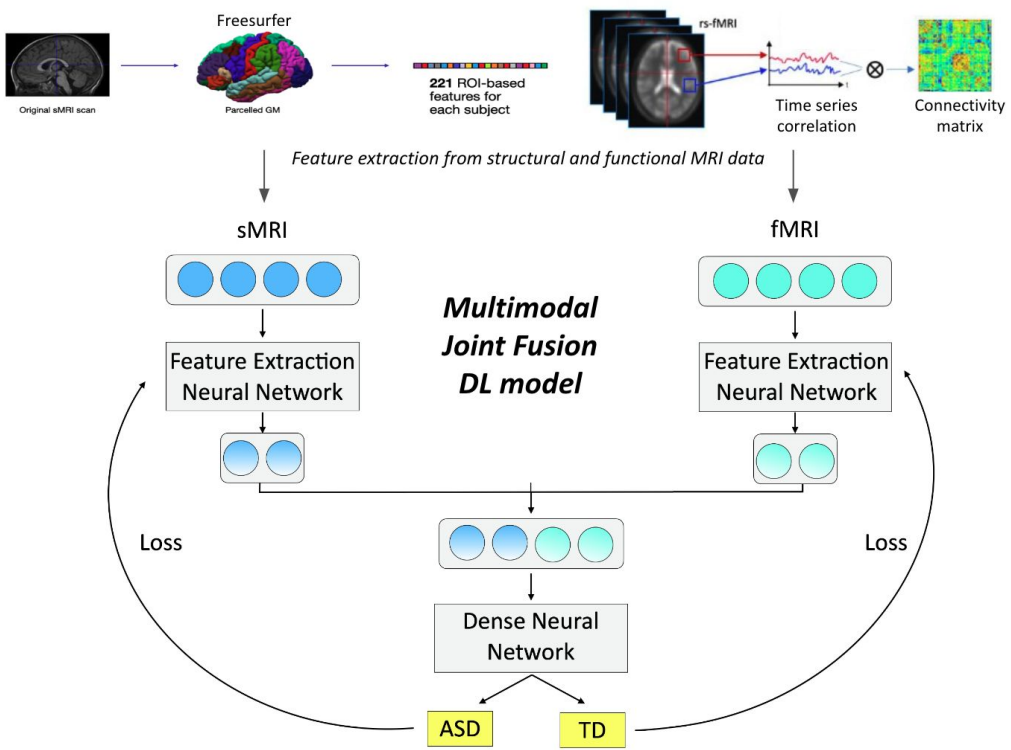
This work has been carried out within the **next_AIM** project and within the **FAIR-AIM** project funded by Tuscany Government (POR FSE 2014-2020).

Background: The integration of the information encoded in multiparametric MRI images can enhance the performance of machine-learning classifiers.

AIM: In this study, we show how the combination of structural and functional MRI improves the performances of a deep learning (DL) model trained to discriminate subjects with Autism Spectrum Disorders (ASD) with respect to typically developing controls (TD).

Material and Methods:

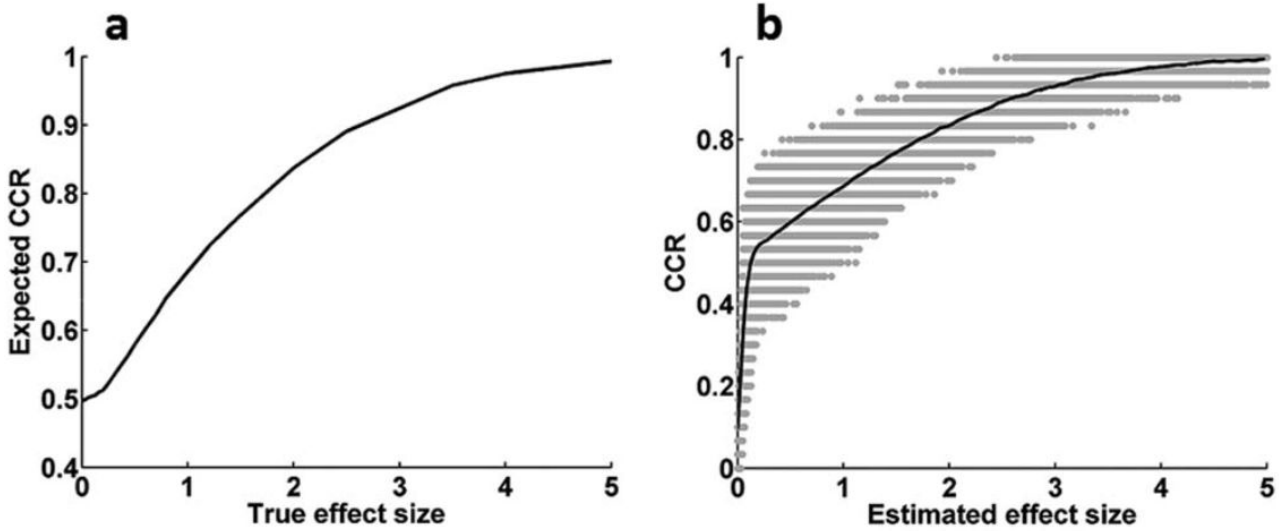
- Structural and functional MRI brain scans publicly available within the ABIDE I and II data collections were used
- 1383 male subjects with age between 5 and 40 years, including 680 subjects with ASD and 703 TD from 35 different acquisition sites were considered
- Due to the multisite nature of the dataset, the Freesurfer structural features and the functional connectivity measures were harmonized using the NeuroHarmonize package.
- The ASD vs. TD classification was carried out with a DL model with the **joint fusion approach**.
- The performance was evaluated by computing the Area under the Receiver Operating Characteristic (ROC) curve (AUC) within a 5-fold cross-validation.



Results:

- An AUC of **0.60±0.03** is obtained in the ASD vs. TD discrimination when only structural features are considered.
- AUC of **0.70±0.02** if only functional features are considered.
- Finally, the joint fusion approach leads to an AUC of **0.75±0.01**.

Below-chance level classification rates in low sample size/low effect size data



Jamalabadi, H., Alizadeh, S., Schönauer, M., Leibold, C. and Gais, S. (2016), Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. Hum. Brain Mapp., 37: 1842-1855. https://doi.org/10.1002/hbm.23140