Dipartimento di Fisica "E. Fermi"
Università di Pisa

Explainable AI approaches to Study Autism Spectrum Disorders

Piernicola Oliva
Piernicola.oliva@unipi.it

Investigation of Autism Spectrum Disorder (ASD)

- ASD is a heterogeneous neurodevelopmental disorder characterized by:
  - reduced social skills or empathy
  - repetitive behaviors
  - speech difficulties
- The exact etiopathogenesis of ASD is not yet fully established
- Connected to early alteration of brain development
- Studied across multiple fronts: genetic, psychiatric, and neurological.

[American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 5th Edn. Arlington, VA: American Psychiatric Association (2013)]

Investigation of Autism Spectrum Disorder (ASD)

In Italy, the findings reveal an ASD prevalence rate of 1 in 77 children aged 7–9 years old, with a male to female prevalence ratio of 4.4:1

[Scattoni et al. Child and Adolescent Psychiatry and Mental Health (2023) 17:125]

Currently, there are **no reliable biomarkers** for autism spectrum disorder, either from a genetic or neuroimaging perspective.

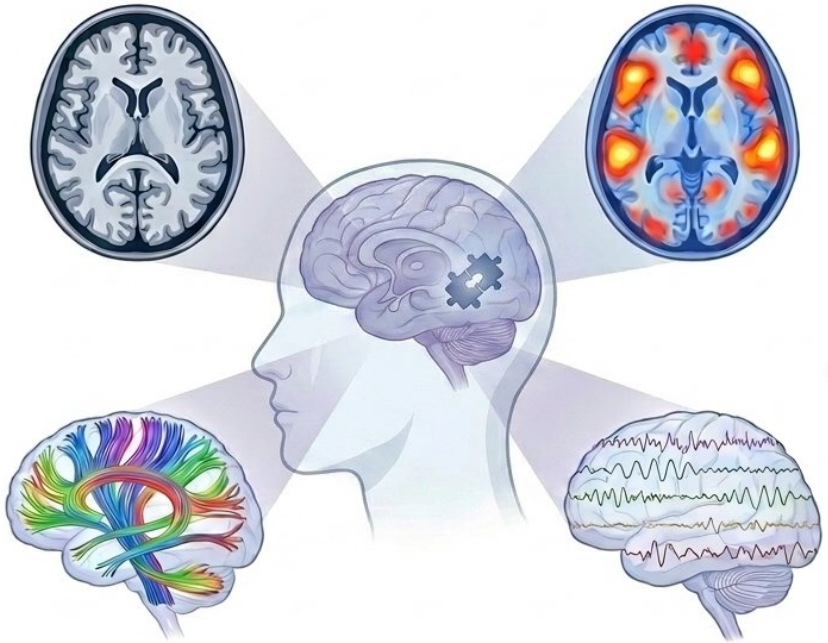
The diagnosis of autism is made through a comprehensive evaluation that includes **behavioral observation**, developmental history, and **standardized assessments** conducted by qualified specialists.

ASD and Neuroimaging

Several studies investigate neuroimaging data, in order to highlight brain signatures able to distinguish ASD subjects from controls.

Structural MRI
Produces high-resolution images of brain anatomy to measure structure (e.g., gray matter volume, lesions)

DTI
Maps white matter tracts by tracking the diffusion of water molecules along axonal pathways.



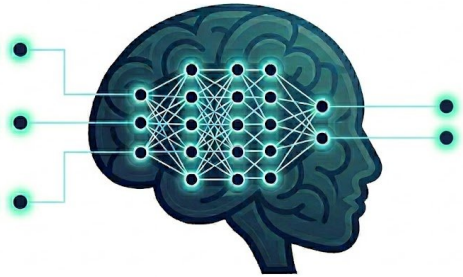
fMRI
Measures brain activity indirectly by detecting blood-oxygen-level–dependent (BOLD) changes linked to neural activity.

EEG
Records electrical brain activity via scalp electrodes with high temporal (millisecond) resolution.

[Andrews et al, In: Pratt J, Hall J, editors.
Biomarkers in Psychiatry. Current Topics in Behavioral Neurosciences. Springer, Cham (2018). 40.]

ASD with Machine Learning

- The study of ASD via neuroimaging data can really take advantage of machine learning analysis:
  - Large number of variables
  - Intrinsically multivariate problem
  - Main drawback: limited availability/access to data



- ML contribution to ASD study can be:

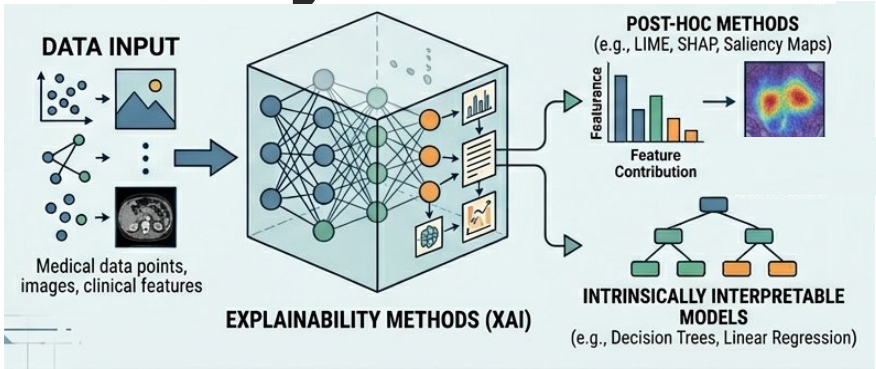
Early Detection
Young individuals for whom standardized test cannot be administered
- sMRI
- EEG?

Etiology of Disorder
Contribution to identify the brain regions most involved in the disorder
- fMRI
- DTI

- We are mainly involved in investigating the etiology of the ASD
- In these studies ML is typically used to separate the ASD class from the neurotypical (TD) one

The need for explainability

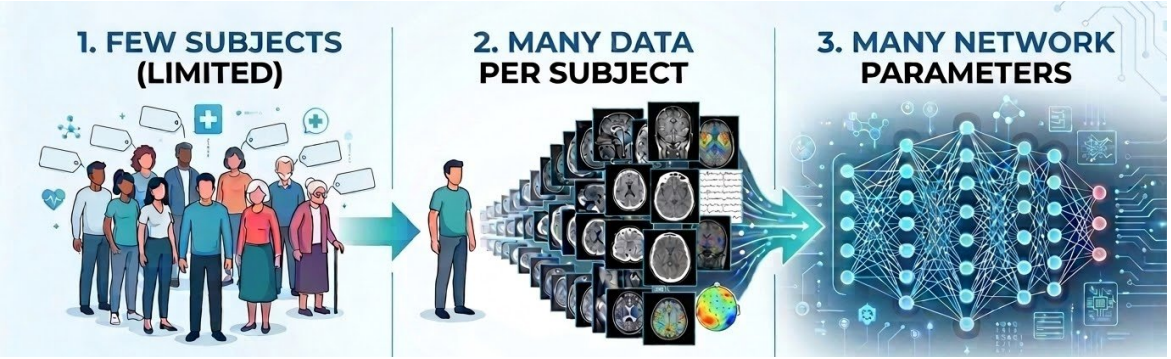
In the context of using machine learning to study the characteristics of the disorder, the features that are primarily responsible for the classification performance are of greater interest than the performance itself.



Intrinsically explainable models
Decision-making process is transparent and understandable by design.
These algorithms allow to directly see how inputs influence predictions.
(Decision trees, linear regression, linear Support Vector Machine, …)

SHAP (SHapley Additive exPlanations) is a unified framework for interpreting ML predictions. It uses principles from cooperative game theory to assign each feature a measure of contribution to a model's output.
SHAP helps explain complex "black box" models in transparent, human-understandable terms.

No Data, No Intelligence: The Neuroimaging Bottleneck



Limited availability of patient cohorts for specific/rare pathologies. (N=10-100)

Large amount of data collected for each patient. Scans (MRI, fMRI, DTI, EEG, PET) and clinical data.

Deep Learning models with millions/billions of parameters to learn complex patterns from data.

Multicenter datasets

The size of available datasets for neuroimaging in ASD is often limited and obtaining a large sample from a single clinical center can be challenging, leading to the collection of data from multiple clinical centers (multi-center datasets).

- Higher the number of subjects in the training set higher is the affidability of the model
- The study is less prone to overfitting
- Higher generability of the study

ABIDE is a large **public multicenter** dataset that provide researchers with neuroimaging data (**MRI**), along with demographic and clinical information.
It is composed by more of **2000 individuals**

- Differences in scanner type and protocols, acquisition method
- Analysis influenced by the batch effect

Site effect and harmonization

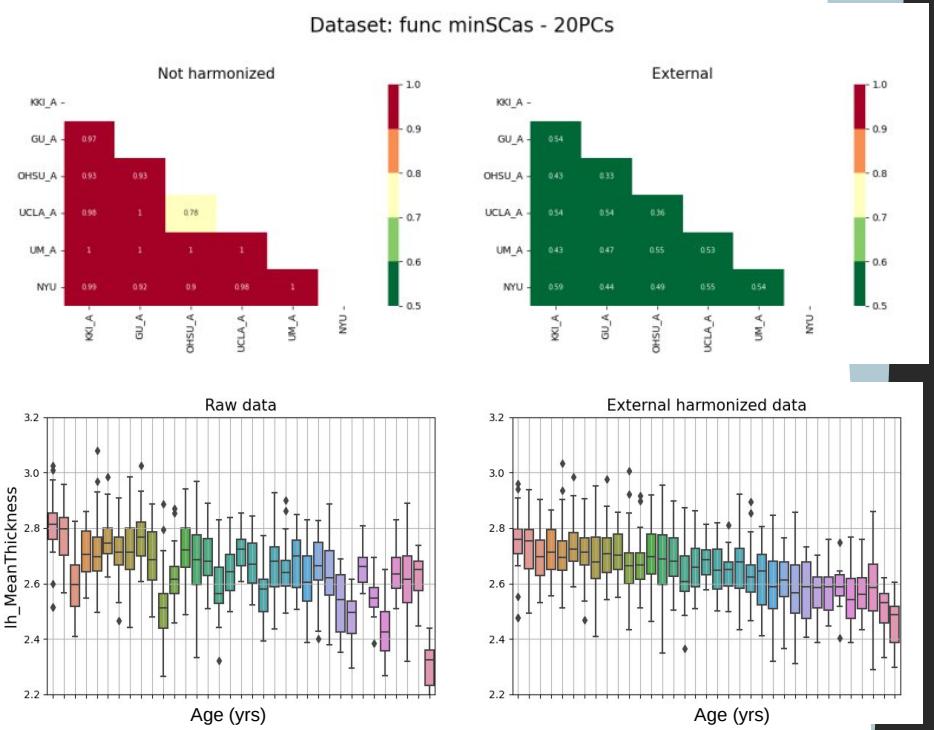
COMBAT harmonization:

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

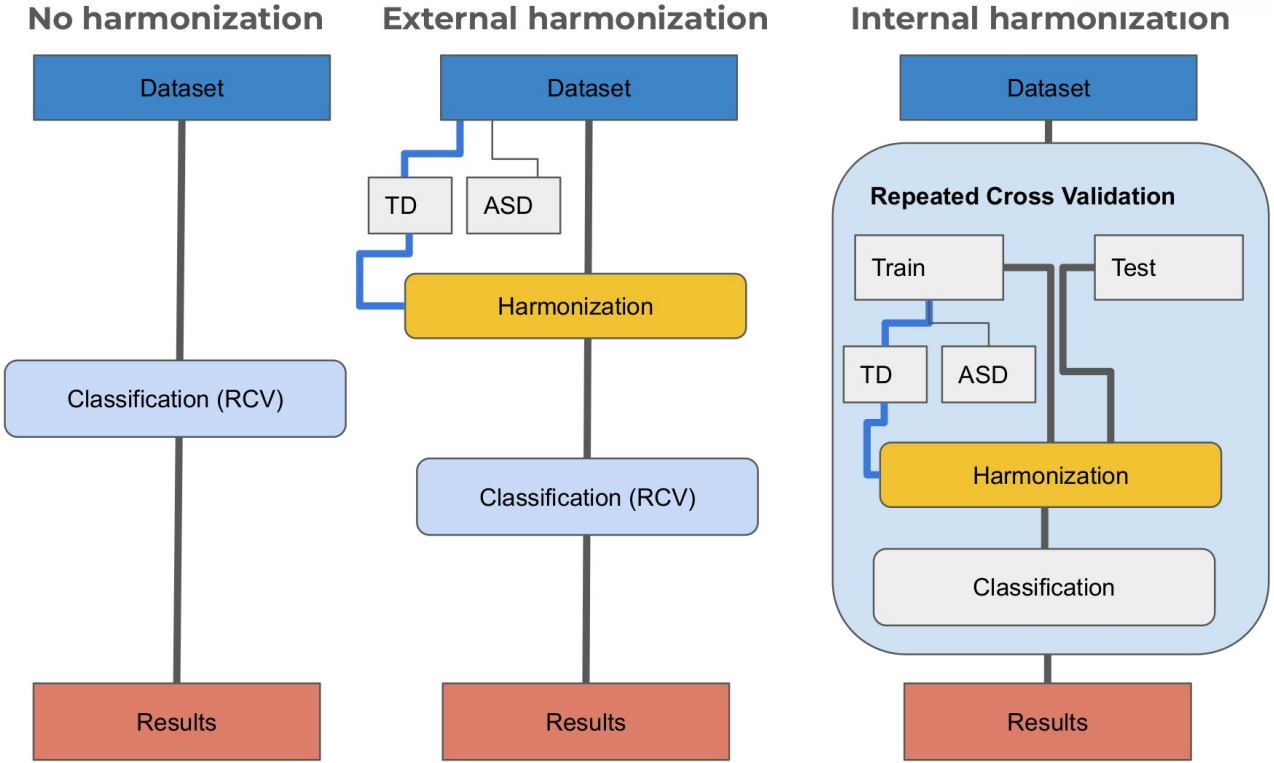
R. Pomponio et al., NeuroImage, 2020,https://doi.org/10.1016/j.neuroimage.2019.116450

COMBAT is often <u>performed on the whole dataset</u> as a preprocessing step.

Potential data leakage since the training and test sets are defined only after this step.



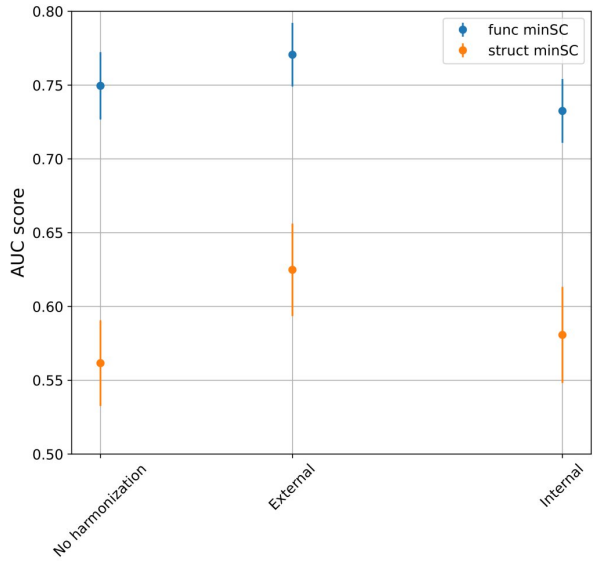
Harmonization approaches



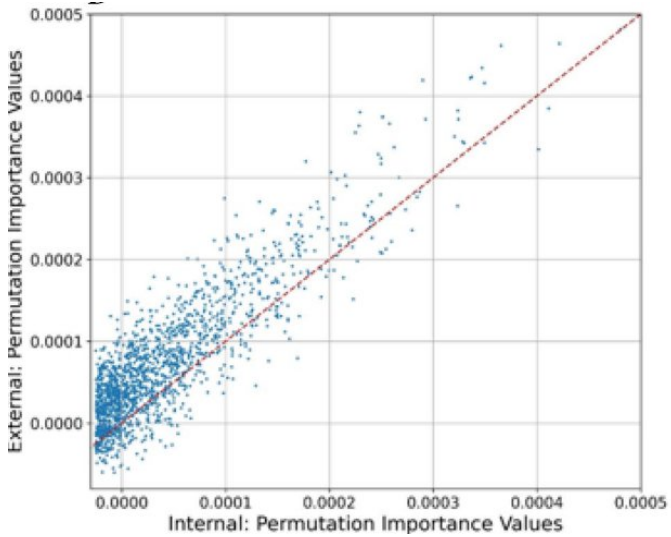
Serra *et al. Brain Informatics* (2023) 10:32
https://doi.org/10.1186/s40708-023-00210-x

Classification of ASD vs TD

Effect of the harmonization strategy



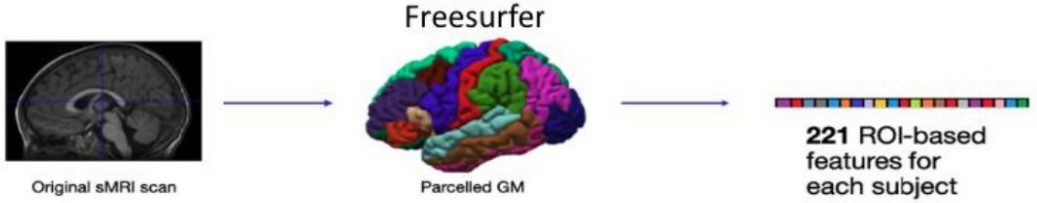
Feature Importance



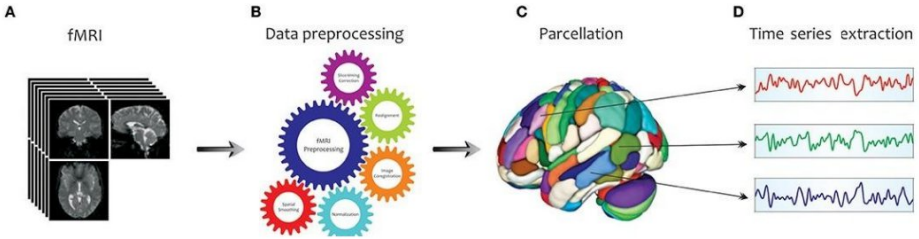
Serra *et al. Brain Informatics* (2023) 10:32
https://doi.org/10.1186/s40708-023-00210-x

MRI preprocessing pipelines

sMRI preprocessing

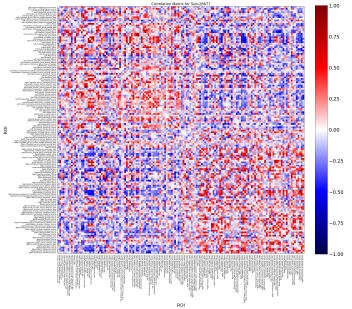


fMRI preprocessing



CPAC

$$r_{XY} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left(\sum_{i=1}^{N}(X_i - \bar{X})^2\right)\left(\sum_{i=1}^{N}(Y_i - \bar{Y})^2\right)}}.$$



5995 connectivity features

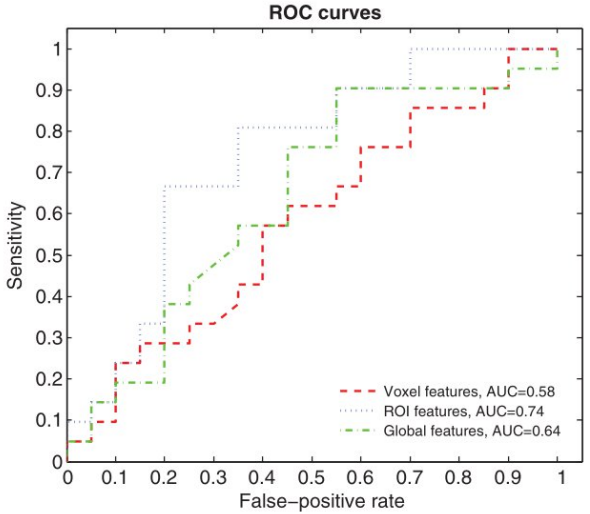
sMRI

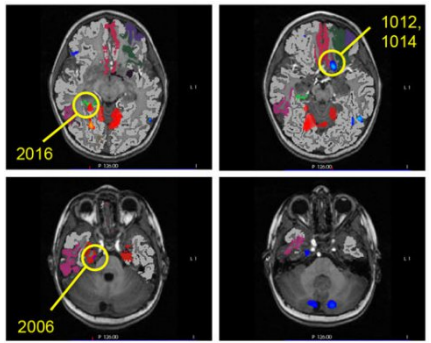
Dataset from IRCCS Stella Maris Foundation.

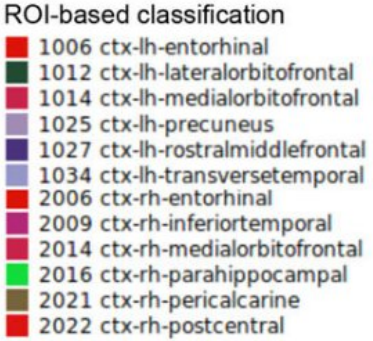
- 21 male children with a diagnosis of ASD
- 20 TD male children

Whole-Brain Feature Classification
Voxel-Based Feature Classification
ROI-Based Feature Classification

Linear SVM analysis



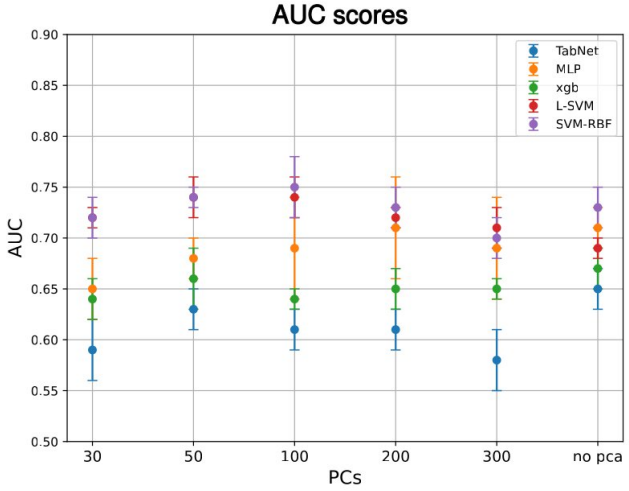




Gori et al, J Neuroimaging 2015;25:866-874. DOI: 10.1111/jon.12280

fMRI ML - DL

- Comparison of traditional ML and DL classifiers
- Functional connectivity features, evaluated on Harvard–Oxford atlas (110 regions) => 5995 features
- **Public multicenter dataset (ABIDE)**
- Evaluation of most relevant features

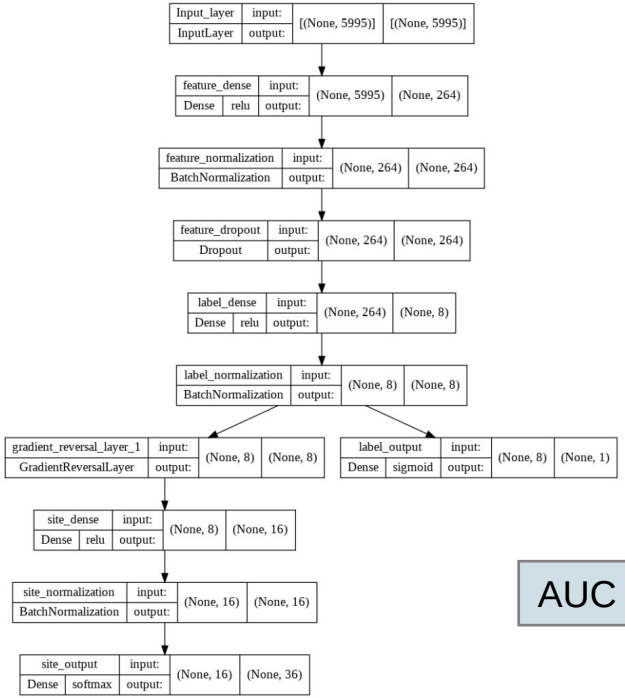


Occurrences	ROI	Anatomical Part	Mesulam
18	3102	L-Precuneous Cortex	Heteromodal
15	1002	L-Superior Temporal Gyrus; posterior division	Unimodal
15	501	R-Inferior Frontal Gyrus; pars triangularis	Heteromodal
14	1302	L-Middle Temporal Gyrus; temporo-occipital	Heteromodal
11	1101	R-Middle Temporal Gyrus; anterior division	Heteromodal
10	1301	R-Middle Temporal Gyrus; temporo-occipital	Heteromodal
8	4301	R- Parietal Operculum Cortex	Unimodal
8	3301	R-Frontal Orbital Cortex	Paralimbic
8	2702	L-Subcallosal Cortex	Paralimbic
8	1102	L-Middle Temporal Gyrus; anterior division	Heteromodal
7	3401	R-Parahippocampal Gyrus; anterior division	Paralimbic
7	2801	R-Paracingulate Gyrus	Heteromodal
7	2302	L-Lateral Occipital Cortex; inferior division	Paralimbic
7	1702	L-Postcentral Gyrus	Primary
6	2201	R-Lateral Occipital Cortex; superior division	Unimodal
6	401	R-Middle Frontal Gyrus	Heteromodal
5	4402	L-Planum Polare	Unimodal

- Functional features perform better than structural ones (AUC .75 vs .65)
- DL classifiers are not optimal for this type of data

Mainas et al, Appl. Sci. 2024, 14, 7632 https://doi.org/10.3390/app14177632

Data Harmonization with an Adversarial Learning

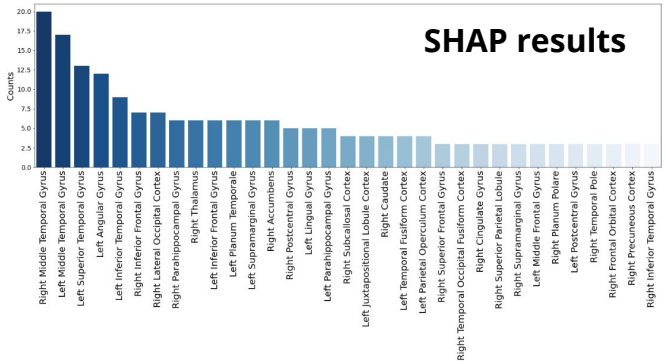


The network is forked into two DL branches:

- a TD vs ASD classifier
- a site classifier

The system is able to learn important features for classification, as well as to generalize this information from one site to another

AUC = 0.71 ± 0.01.



Campo et al, Appl. Sci. 2023, 13, 6486. https://doi.org/10.3390/app13116486

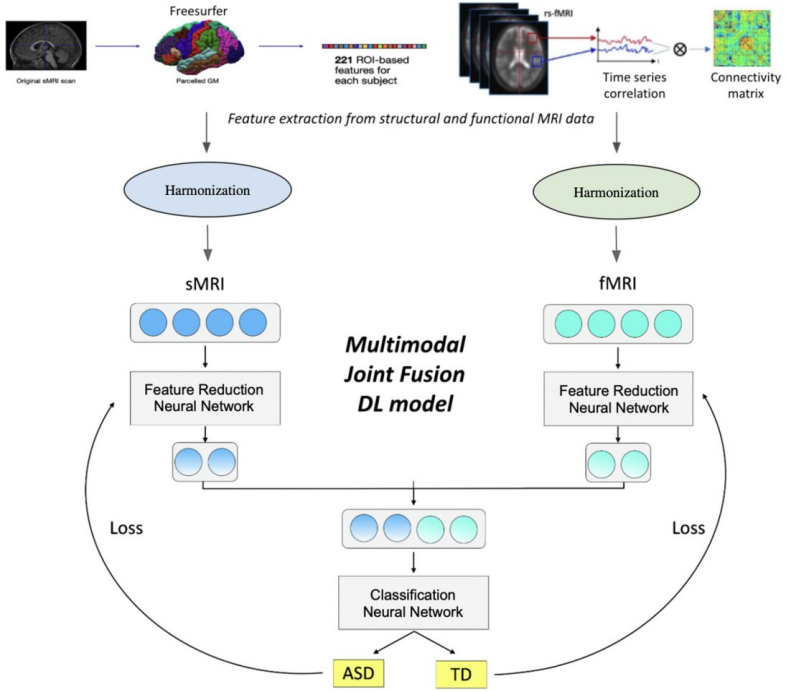
Multimodality

Combining multimodal brain imaging data provide more information for individual subjects by exploiting the complementarity of different imaging modalities.

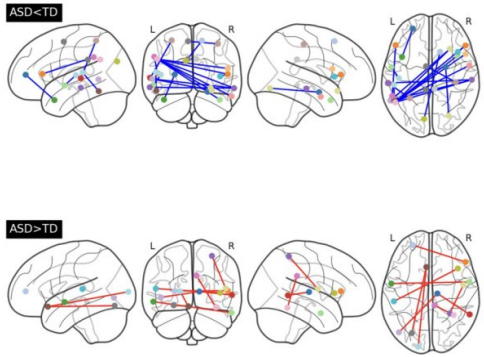
In the field of brain disorders research, fusion strategies are commonly utilized for both diagnosis and prediction purposes.

Combining data from different modalities allows the extraction of more comprehensive and complementary information. This, in turn, leads to the development of better performing models compared to those relying on a single data modality.

Deep learning based joint fusion approach



Type of model	AUC	Accuracy
Structural model	0.66 ± 0.05	0.75 ± 0.08
Functional model	0.76 ± 0.04	0.83 ± 0.12
Joint fusion model	0.78 ± 0.04	0.85 ± 0.12



Saponaro *et al. Brain Informatics* (2024) 11:2
https://doi.org/10.1186/s40708-023-00217-4

Deep learning based joint fusion approach

SHAP results:

Brain Regions (Measurement)			Cohen's d
Right Postcentral Gyrus	-	Left Juxtapositional Lobule Cortex	-
Right Thalamus	-	Right Middle Temporal Gyrus	+
Right Middle Temporal Gyrus (posterior division)	-	Left Angular Gyrus	-
Right Inferior Frontal Gyrus (pars triangularis)	-	Right Frontal Operculum Cortex	-
Right Temporal Pole	-	Left Angular Gyrus	-
Right Middle Temporal Gyrus (posterior division)	-	Left Middle Temporal Gyrus	-
Right Frontal Orbital Cortex	-	Left Middle Temporal Gyrus	-
Left Occipital Pole	-	Left Subcallosal Cortex	+
Right Superior Frontal Gyrus	-	Left Cuneal Cortex	-
Left Occipital Fusiform Gyrus	-	Left Subcallosal Cortex	+
Left Inferior Temporal Gyrus (temporooccipital part)	-	Left Planum Temporale	-
Right Frontal Orbital Cortex	-	Left Angular Gyrus	-
Superior Temporal (ThickAvg)			+
Left Middle Temporal Gyrus (posterior division)	-	Right Middle Temporal Gyrus (posterior division)	-

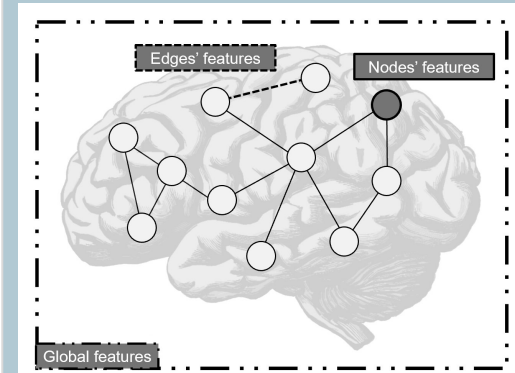
Saponaro *et al. Brain Informatics* (2024) 11:2
https://doi.org/10.1186/s40708-023-00217-4

Graph Neural Network based analysis for functional connectome

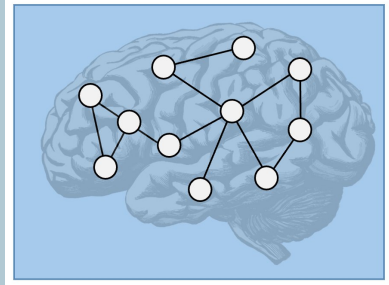
Feature Extraction

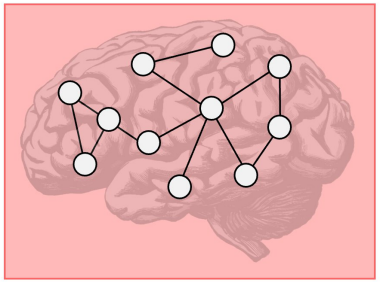
- Morphological features extraction using FreeSurfer
- Time Series Extraction using CPAC
- Other data such as age, sex, acquisition parameters…

Organize feature in graph structure inspired by Functional Connectome



Whole-graph binary classification between ASD and TD





Motisi et al., A Graph Neural Network Approach to Study the Human Connectome, Brain Informatics 2026, accepted

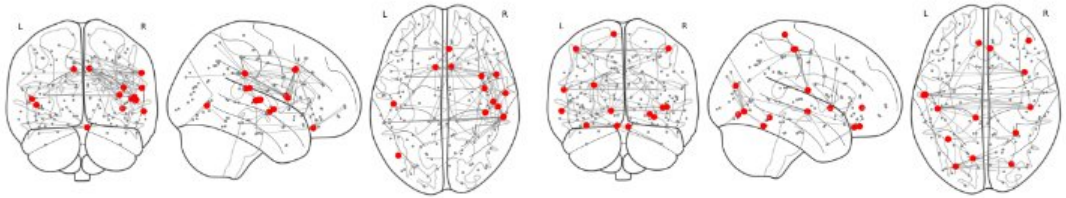
GNN: Results

	AUC [%]
Linear Regression	67±3
Decision Tree	57±2
Linear SVM	66±3
Random Forest	63±3
Proposed model	**72.2±1.8**

ID	Anatomical ROI Name	Percentile
30	Left Precuneus	
9	Left Posterior Dorsal Cingulate Gyrus	99%
72	Left Subparietal Sulcus	
84	Right Posterior Dorsal Cingulate Gyrus	
34	Left Superior Temporal Gyrus (Lateral Aspect)	
4	Left Subcentral Gyrus and Sulci	
12	Left Opercular Part of Inferior Frontal Gyrus	
25	Left Angular Gyrus	95%
7	Left Middle Anterior Cingulate Gyrus and Sulcus	
100	Right Angular Gyrus	
29	Left Precentral Gyrus	
26	Left Supramarginal Gyrus	
66	Left Parieto-occipital Sulcus	
79	Right Subcentral Gyrus and Sulci	
105	Right Precuneus	90%

Subject ID: 51251 (TD)

Subject ID: 28787 (ASD)

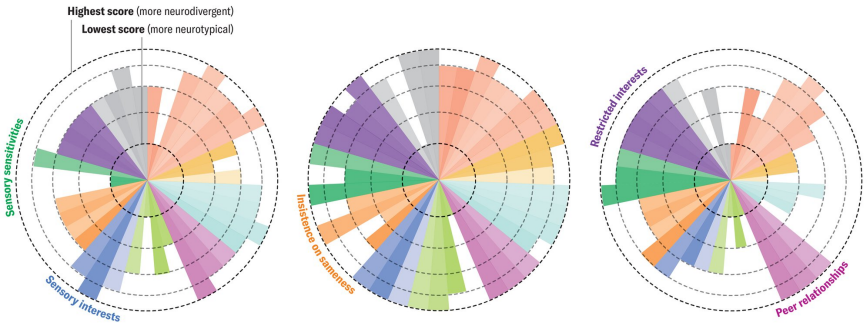


Motisi et al., A Graph Neural Network Approach to Study the Human Connectome, Brain Informatics 2026, accepted

Conclusion

- Explainable AI is able to separate TD from ASD
- Most relevant features are in agreement with clinical findings
- Multimodal (and multiscale) approaches are promising
- In order to unlock full power of ML methods, large, clean and cured datasets are essential
- Future research lines:

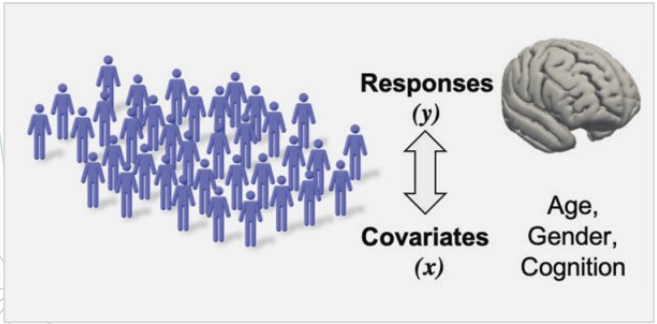
Decode the autism *spectrum*

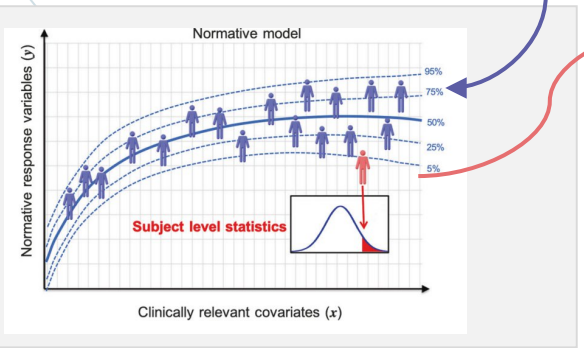


Tackle the gender bias in ASD studies

Thank you!

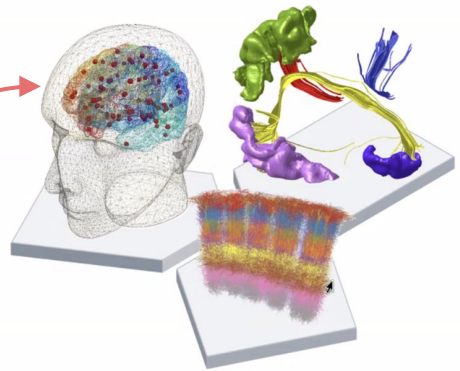
Beyond classification





Normative modelling provides a statistical framework to define the typical range of *brain connectivity or activity* in healthy individuals.

An **anomaly detection** model (e.g., a VAE) trained on healthy subjects can be applied to patient data to *identify brain regions and connections* that deviate from the normative pattern and are implicated in the disorder.



Whole-brain simulations provide a mechanistic validation of the deviations identified by anomaly detection.
Whole-brain modelling aims to reproduce brain dynamics and functionality using *computational models* at different spatial and temporal scales.

Marquand A.F. *et al.*; **Conceptualizing mental disorders as deviations from normative functioning**; 2019; doi: 10.1038/s41380-019-0441-1
Nguyen H.H. *et al.*; **Variational Autoencoder for Anomaly Detection: A Comparative Study;** 2024; doi: 10.48550
Griffiths J.D. *et al.*; **Whole-Brain Modelling: Past, Present, and Future**; 2022; doi: 10.1007/978-3-030-89439-9_13