FACULTY OF ENGINEERING AND ARCHITECTURE

GROUP OF ARTIFICIAL INTELLIGENCE AND SPARSE MODELLING

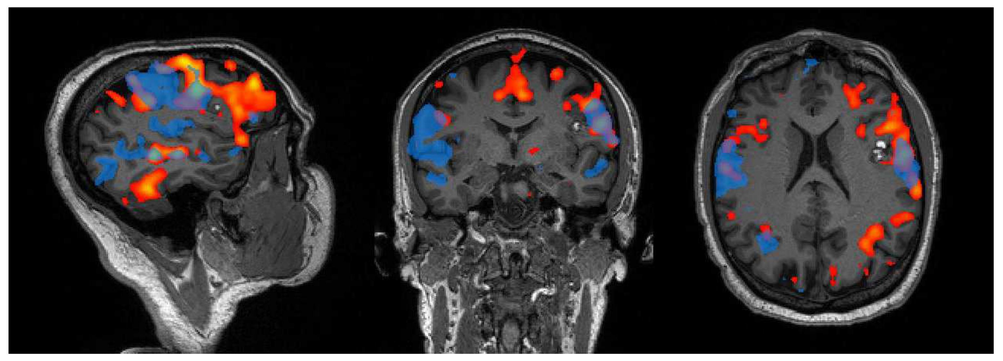
Nicolas Vercheval, Marin Benčevic, Dario Mužević, Irena Galić, Aleksandra Pizurica

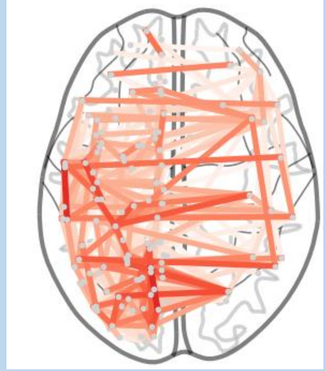
COUNTERFACTUAL FUNCTIONAL CONNECTOMES FOR NEUROLOGICAL CLASSIFIER SELECTION

Problem statement

Functional connectivity (FC): statistical relationship between the activations of different brain regions.

- functional connectome: the correlation matrix
- predictive power in detecting neurological disorders





Motivation

Counterfactual provide intuitive post-hoc classifier explanations:

- They fool the classifier by altering features relevant to their evaluation
- The produced counterfactuals are realistic and easy to interpret
- No available counterfactual method for FC classifiers

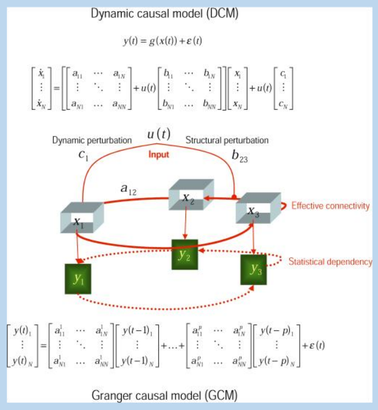
Counterfactual for model selection:

- Feedback loop with medical expert during model development
- Increase trust in the prediction

Challenges:

- Current counterfactual methods only work for specific model
- Some classifiers are not differentiable or not directly accessible
- Lack of unifying framework for comparing counterfactual explanations

Model development



Expert opinion

Use case

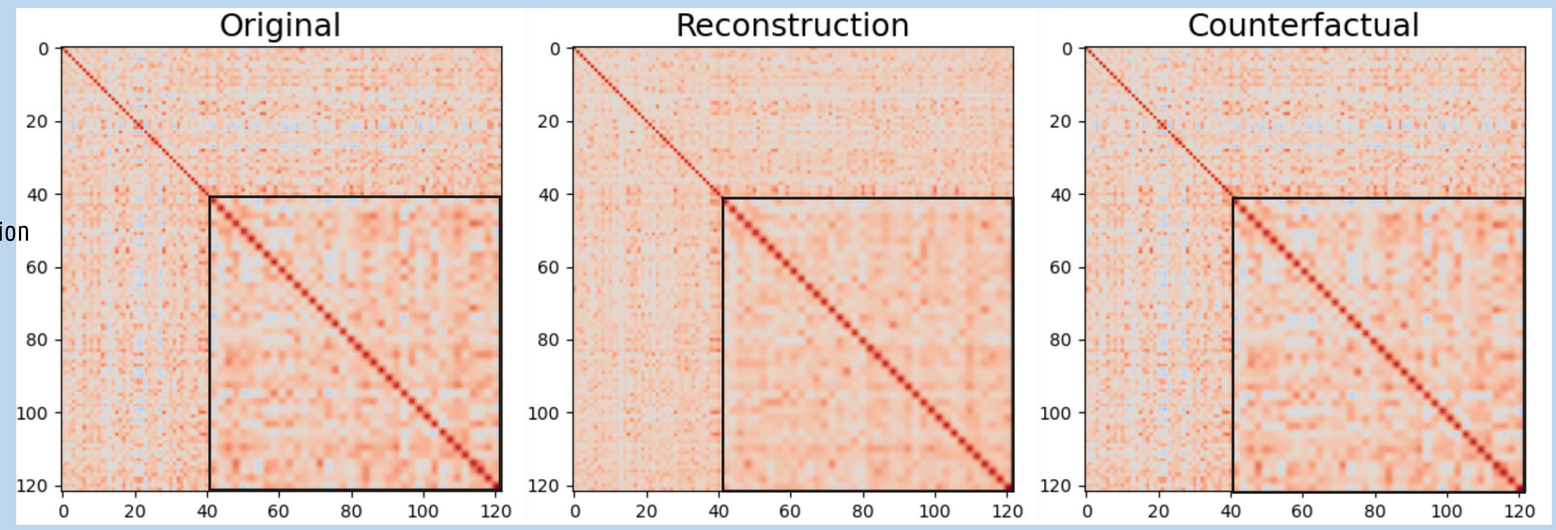
Proposed method

Builds on VAEX[1]:

- Hierarchical Variational autoencoder (HVAE)
- Condition HVAE on evaluation of the classifier
- Counterfactuals by conditioning the reconstruction with target evaluation

Active Noise Cancellation:

- Blurred generation create bias in the explanation
- Define noise as the difference between reconstruction and original
- Remove noise from the counterfactual



Experiments and results

Dataset: resting-state fMRI from ABIDE I

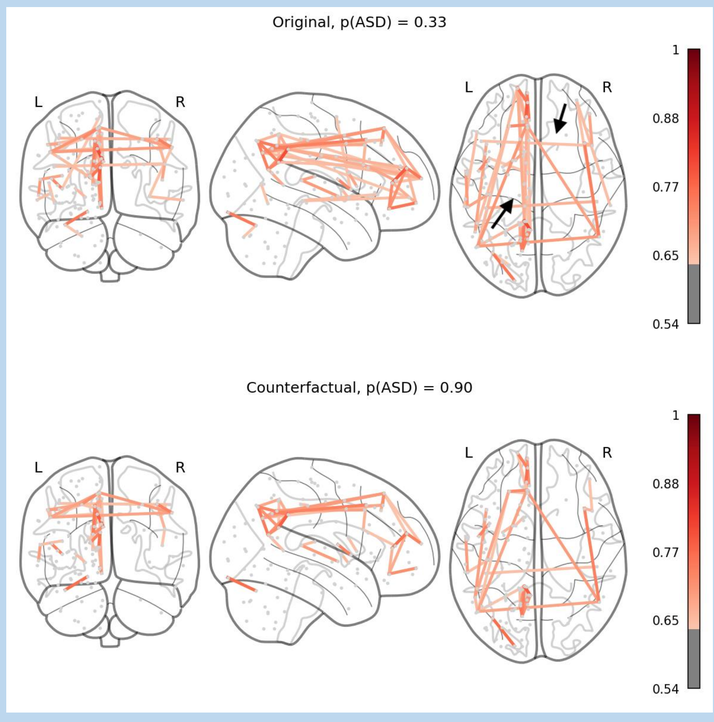
Label: Autistic Spectrum Disorder (ASD)

Classifiers:

- Support Vector machine (SVC)
- Multilayer Perceptron (MLP)
- MLP with autoencoder pretraining (MLP_{AE})

Metrics:

- Success Rate (SR) percentage of counterfactuals fooling the classifier
- Counterfactual Deviation (CD): shift in evaluation of the counterfactual
- Decision Boundary Success (DBS): perfect calibration when 0.5



Model	SR (%)	CD ($\times 10^{-2}$)	DBS ($\times 10^{-2}$)
SVC	100.0	45.9	63.6
MLP	100.0	54.9	47.1
MLP_{AE}	100.0	54.9	46.4

References:
[1] N. Vercheval, A. Pižurica, *Hierarchical variational autoencoders for visual counterfactuals.* ICIP, 2021.

Published in:
N. Vercheval, M. Benčevic, D. Mužević, I. Galić, A. Pizurica, *Counterfactual functional connectomes for neurological classifier.* Eusipco, 2023.

Contact:
<nicolas>.<vercheval>@ugent.be

GHENT UNIVERSITY