

Explainable AI and the brain

Characterizing heterogeneity in diverse clinical cohorts

Esten H. Leonardsen

24.10.24



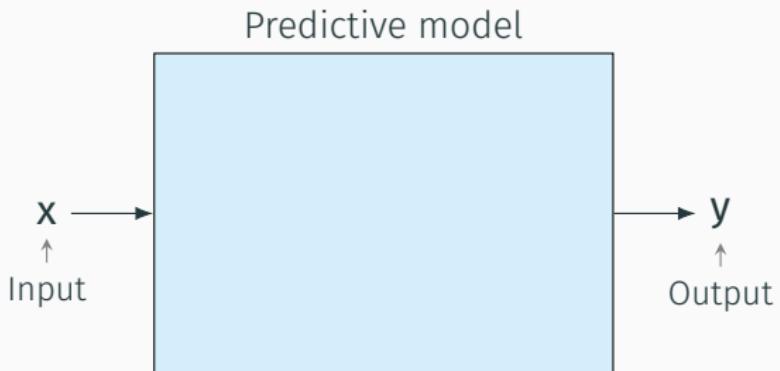
UNIVERSITETET
I OSLO

Outline

1. Theoretical background
 - a. Artificial neural networks
 - b. Explainable AI
2. Explainable AI and dementia
3. Explainable AI and brain age

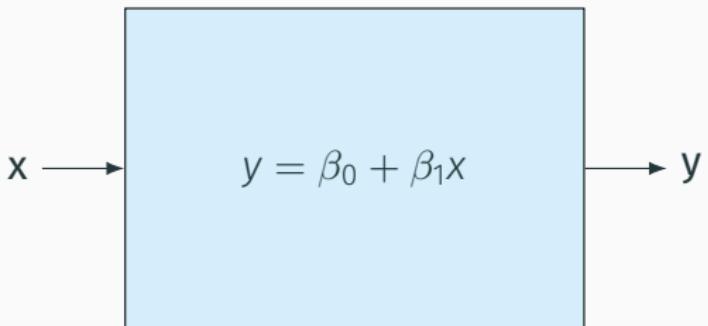


Artificial neural networks



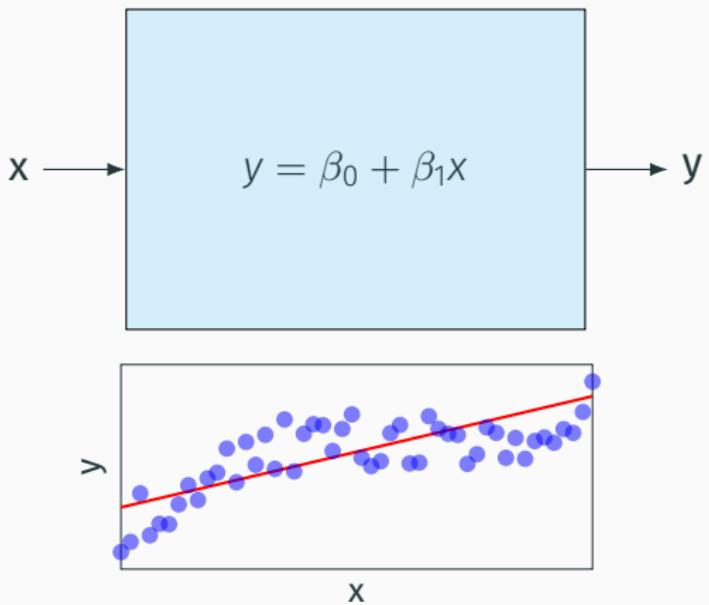
Artificial neural networks

Linear regression model

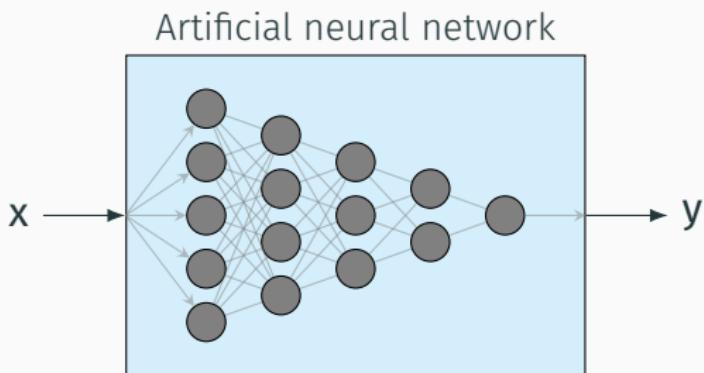


Artificial neural networks

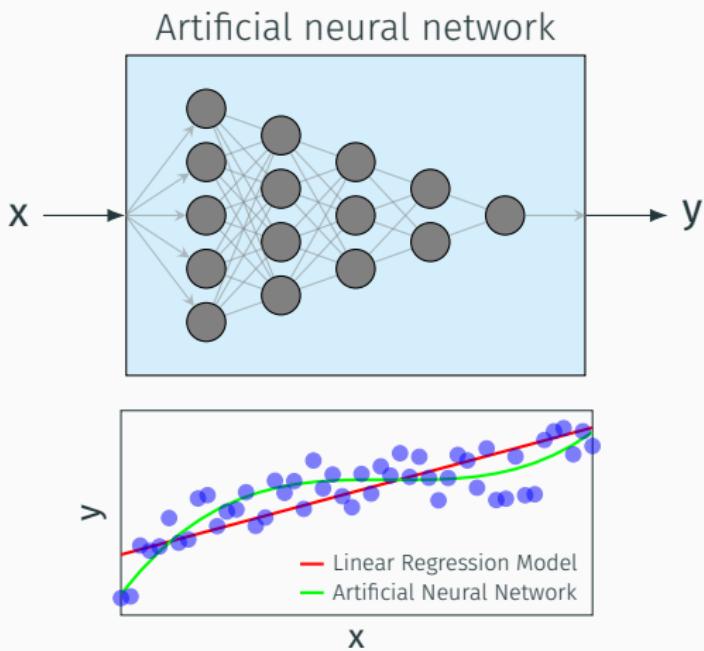
Linear regression model



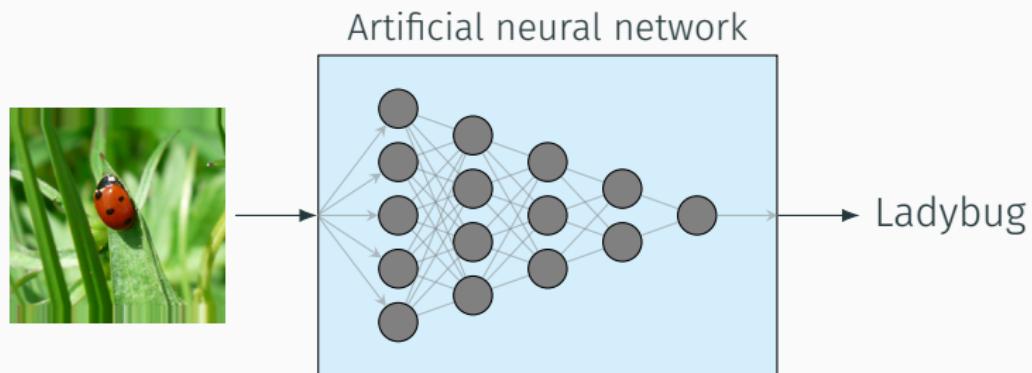
Artificial neural networks



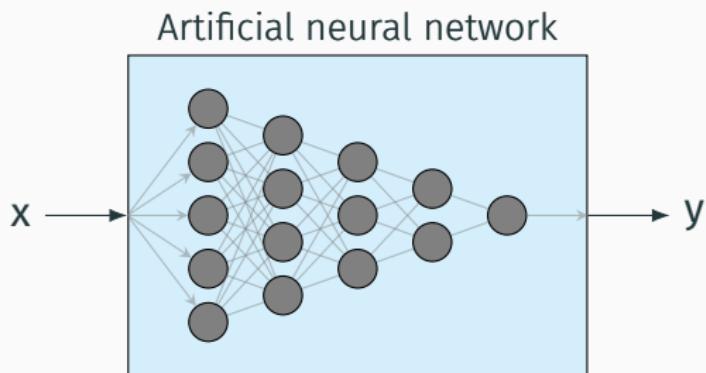
Artificial neural networks



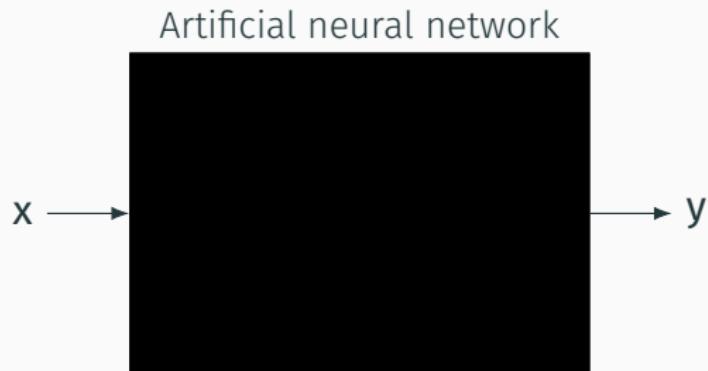
Artificial neural networks



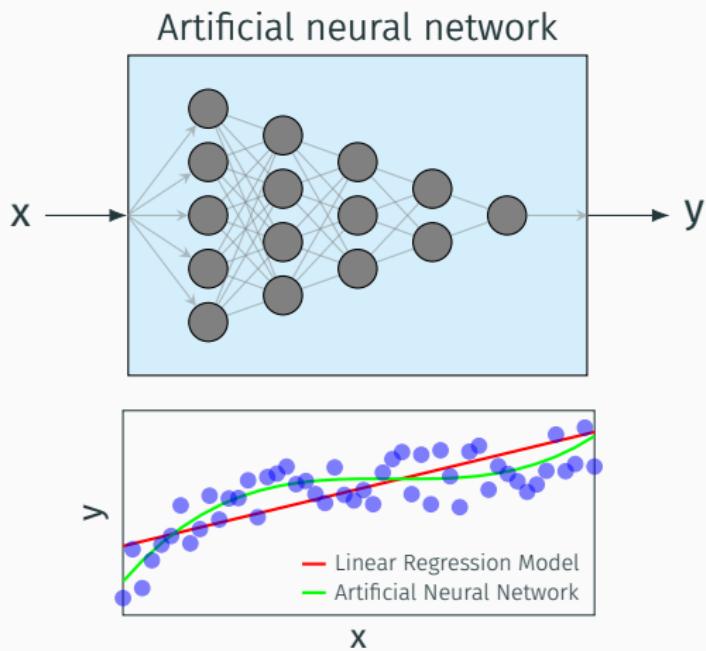
Artificial neural networks: Interpretability



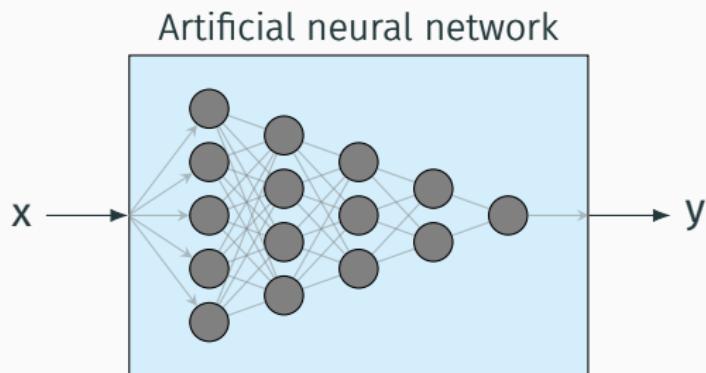
Artificial neural networks: Interpretability



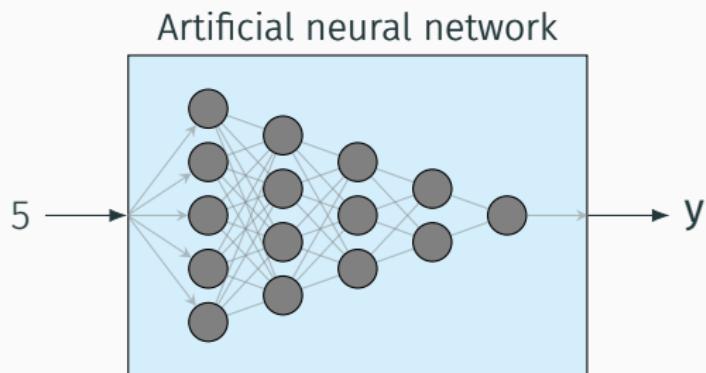
Artificial neural networks: Interpretability



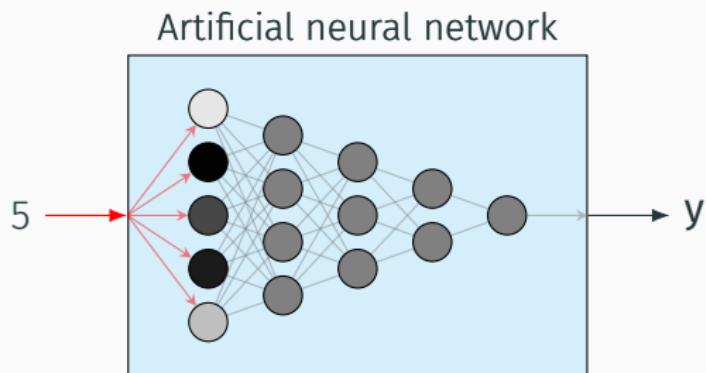
Artificial neural networks: Interpretability



Artificial neural networks: Interpretability



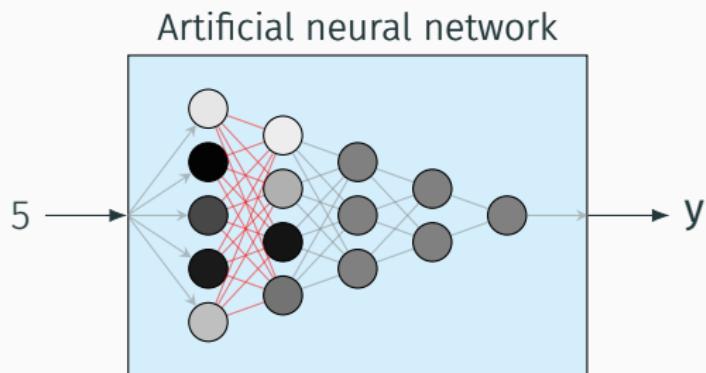
Artificial neural networks: Interpretability



$$n_j^i = f\left(\sum_{k=0}^n w_{jk}^i n_k^{i-1}\right)$$



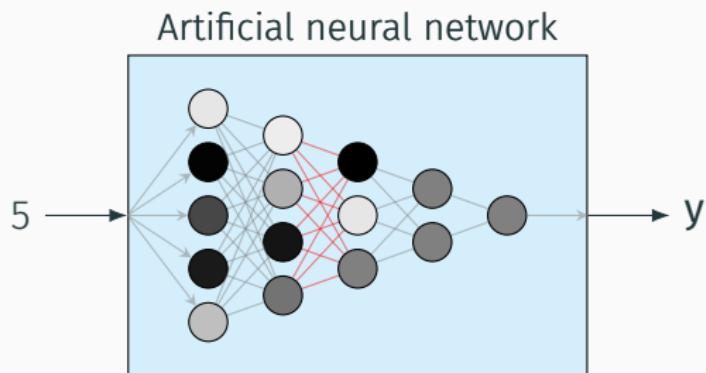
Artificial neural networks: Interpretability



$$n_j^i = f\left(\sum_{k=0}^n w_{jk}^i n_k^{i-1}\right)$$



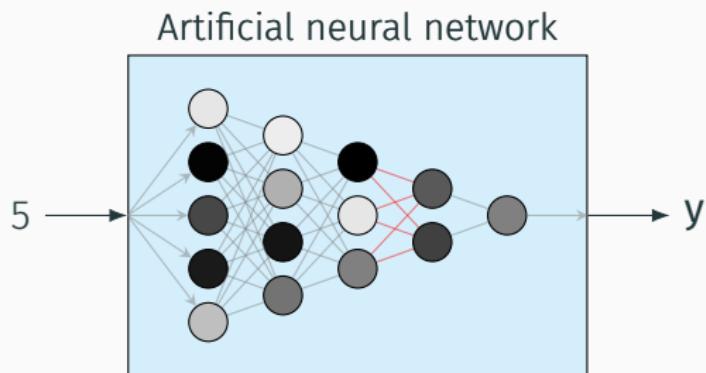
Artificial neural networks: Interpretability



$$n_j^i = f\left(\sum_{k=0}^n w_{jk}^i n_k^{i-1}\right)$$



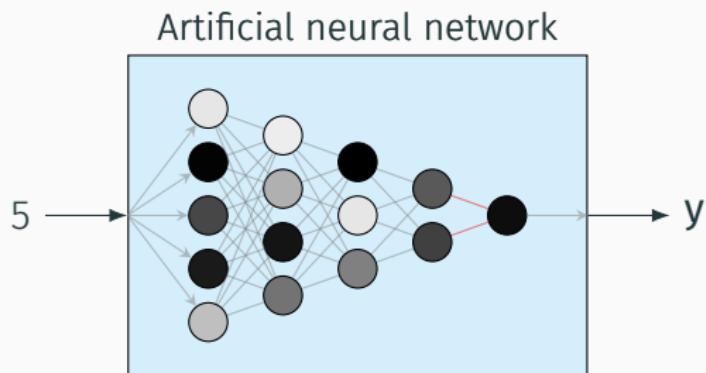
Artificial neural networks: Interpretability



$$n_j^i = f\left(\sum_{k=0}^n w_{jk}^i n_k^{i-1}\right)$$



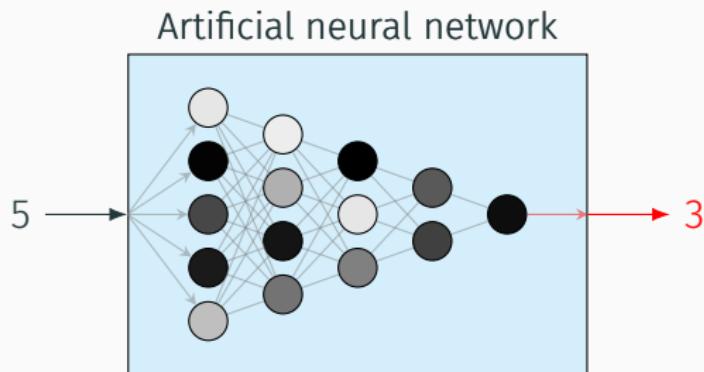
Artificial neural networks: Interpretability



$$n_j^i = f\left(\sum_{k=0}^n w_{jk}^i n_k^{i-1}\right)$$



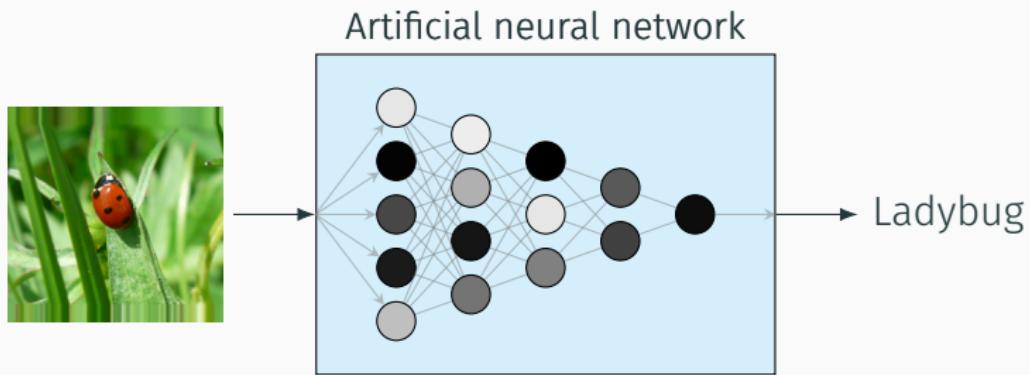
Artificial neural networks: Interpretability



$$n_j^i = f\left(\sum_{k=0}^n w_{jk}^i n_k^{i-1}\right)$$



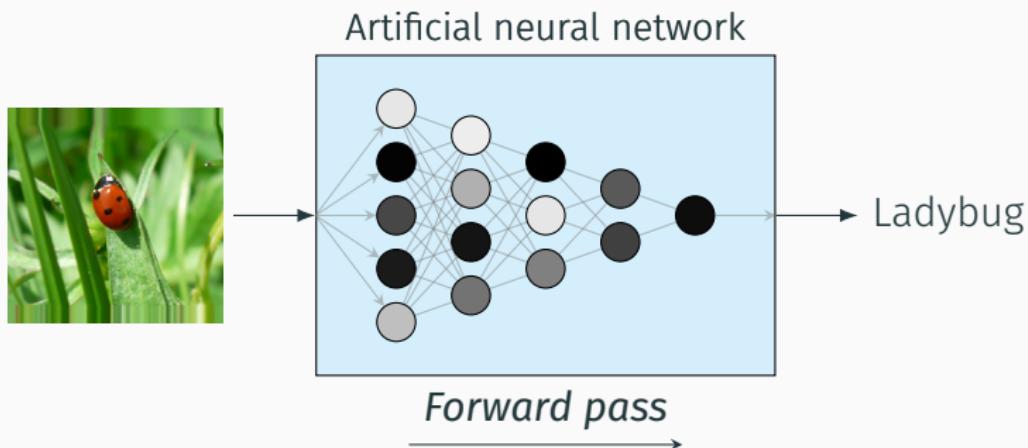
Artificial neural networks: Interpretability



$$n_j^i = f\left(\sum_{k=0}^n w_{jk}^i n_k^{i-1}\right)$$



Artificial neural networks: Interpretability



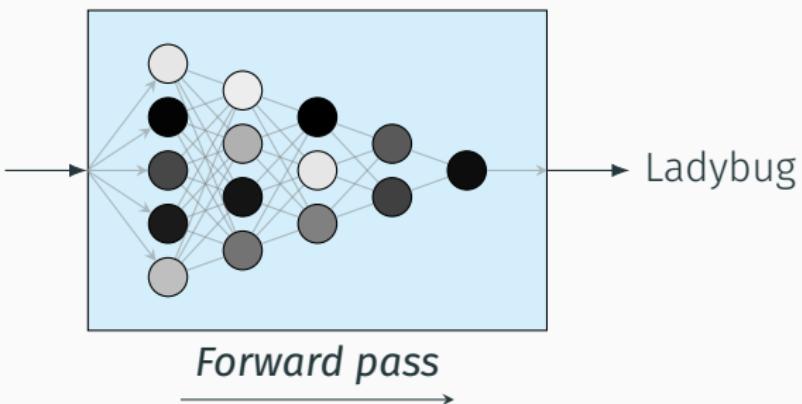
$$n_j^i = f\left(\sum_{k=0}^n w_{jk}^i n_k^{i-1}\right)$$



Artificial neural networks: Explainability



Artificial neural network



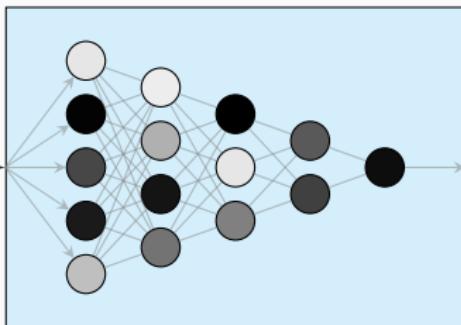
$$n_j^i = f\left(\sum_{k=0}^n w_{jk}^i n_k^{i-1}\right)$$



Artificial neural networks: Explainability



Artificial neural network



Ladybug

Backward pass



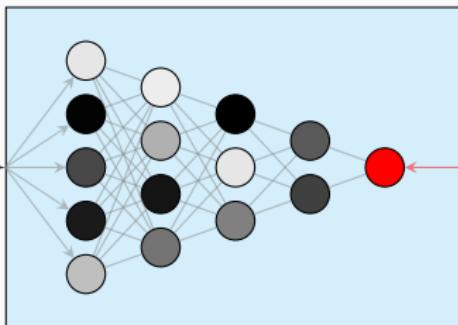
$$R_j^{i-1} = \sum_k \frac{n_j^{(i-1)} w_{jk}^{(i-1)}}{\sum_l n_l^{(i-1)} w_{lk}^{(i-1)}} R_k^i$$



Artificial neural networks: Explainability



Artificial neural network



Ladybug

Backward pass

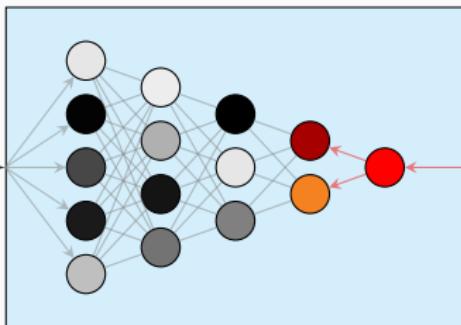
$$R_j^{i-1} = \sum_k \frac{n_j^{(i-1)} w_{jk}^{(i-1)}}{\sum_l n_l^{(i-1)} w_{lk}^{(i-1)}} R_k^i$$



Artificial neural networks: Explainability



Artificial neural network



Ladybug

Backward pass

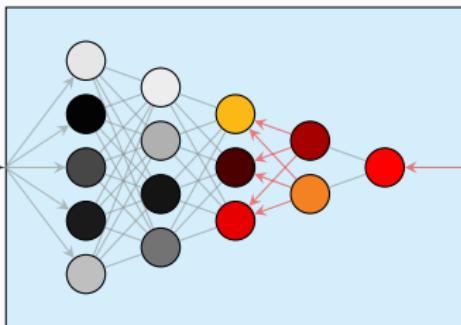
$$R_j^{i-1} = \sum_k \frac{n_j^{(i-1)} w_{jk}^{(i-1)}}{\sum_l n_l^{(i-1)} w_{lk}^{(i-1)}} R_k^i$$



Artificial neural networks: Explainability



Artificial neural network



Ladybug

Backward pass

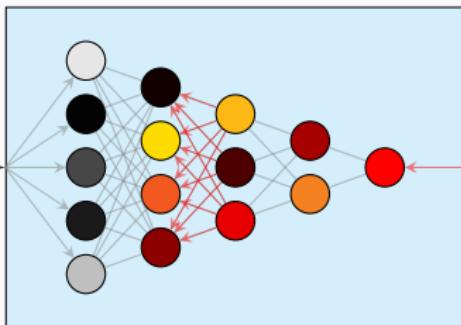
$$R_j^{i-1} = \sum_k \frac{n_j^{(i-1)} w_{jk}^{(i-1)}}{\sum_l n_l^{(i-1)} w_{lk}^{(i-1)}} R_k^i$$



Artificial neural networks: Explainability



Artificial neural network



Ladybug

Backward pass

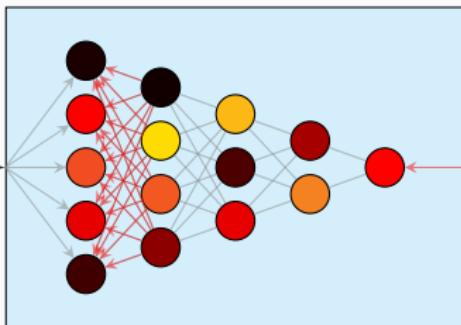
$$R_j^{i-1} = \sum_k \frac{n_j^{(i-1)} w_{jk}^{(i-1)}}{\sum_l n_l^{(i-1)} w_{lk}^{(i-1)}} R_k^i$$



Artificial neural networks: Explainability



Artificial neural network



Ladybug

Backward pass

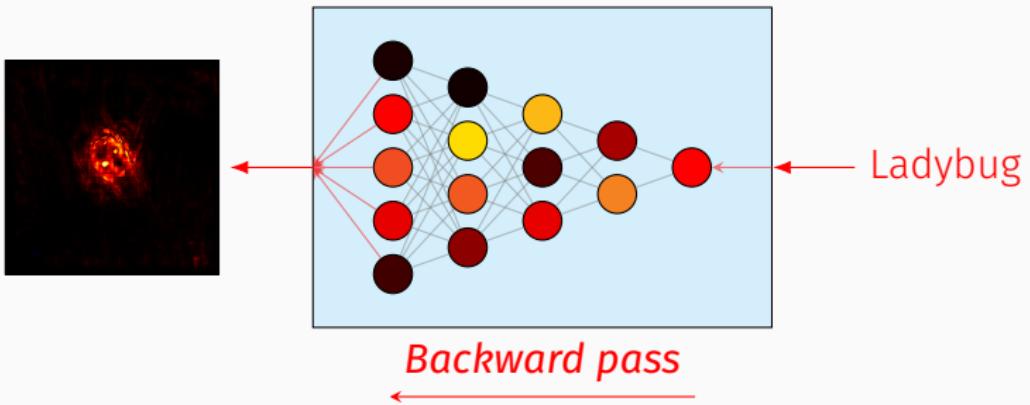
$$R_j^{i-1} = \sum_k \frac{n_j^{(i-1)} w_{jk}^{(i-1)}}{\sum_l n_l^{(i-1)} w_{lk}^{(i-1)}} R_k^i$$



Artificial neural networks: Explainability



Artificial neural network



$$R_j^{i-1} = \sum_k \frac{n_j^{(i-1)} w_{jk}^{(i-1)}}{\sum_l n_l^{(i-1)} w_{lk}^{(i-1)}} R_k^i$$





Reasons to use explainable AI:

- Sanity check models
- Building trust among users





Reasons to use explainable AI:

- Sanity check models
- Building trust among users
- Scientific discovery



Artificial neural networks: Explainability



Reasons to use explainable AI:

- Sanity check models
- Building trust among users
- Scientific discovery
- Characterize heterogeneity in groups we consider
(somewhat) homogeneous

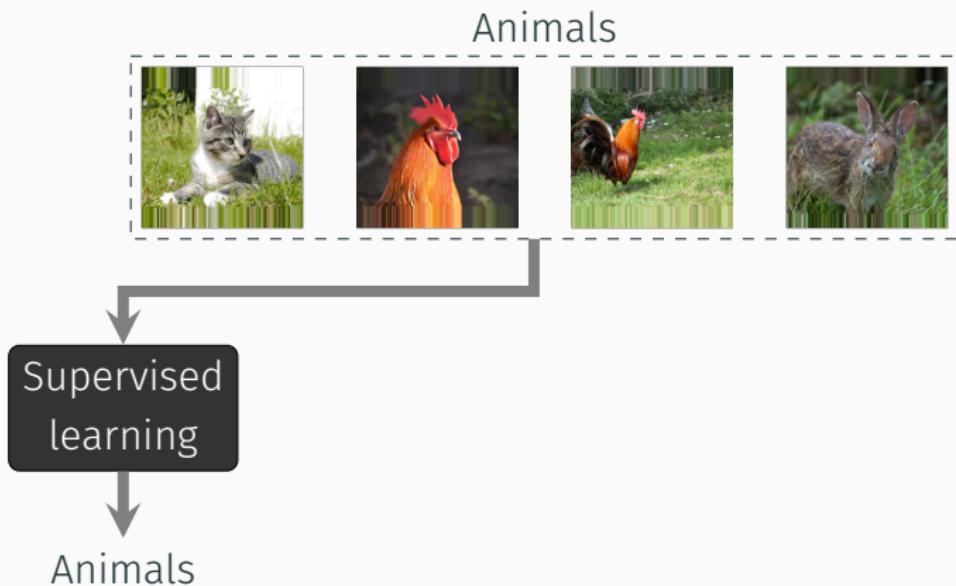


Explainable AI: The central idea

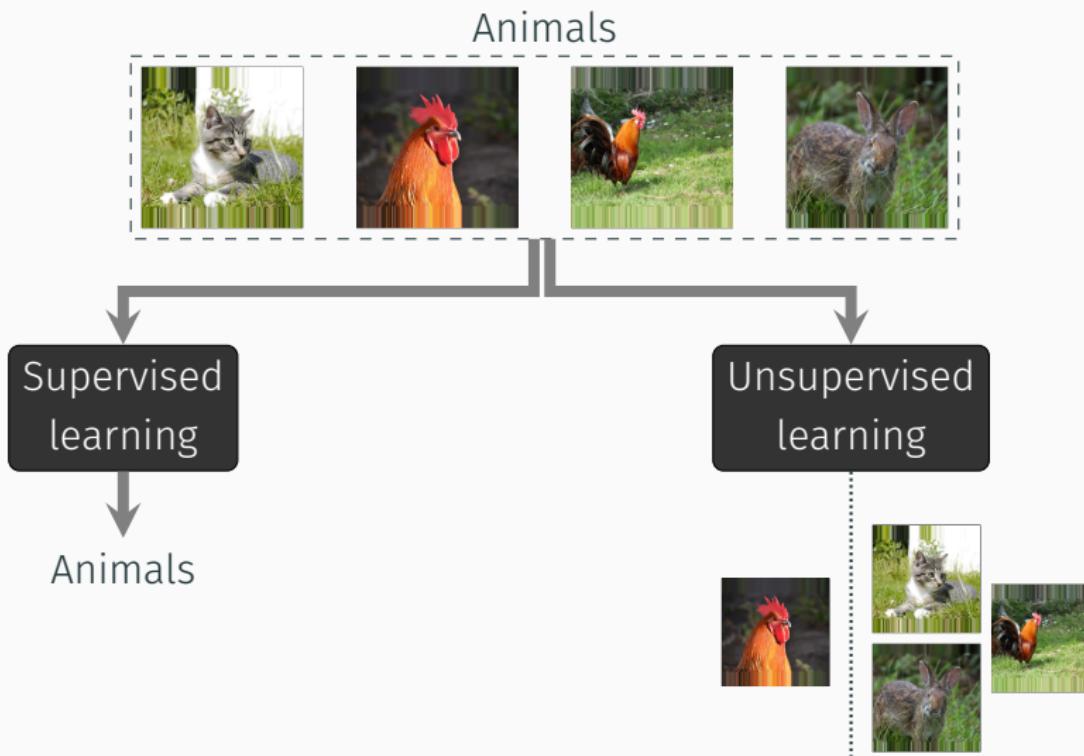
Animals



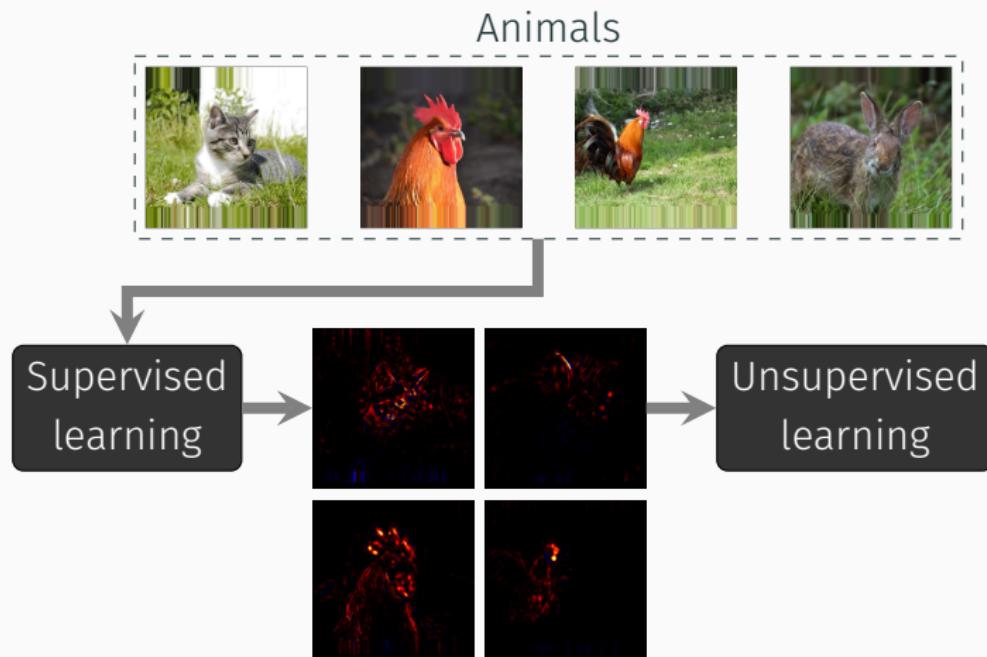
Explainable AI: The central idea



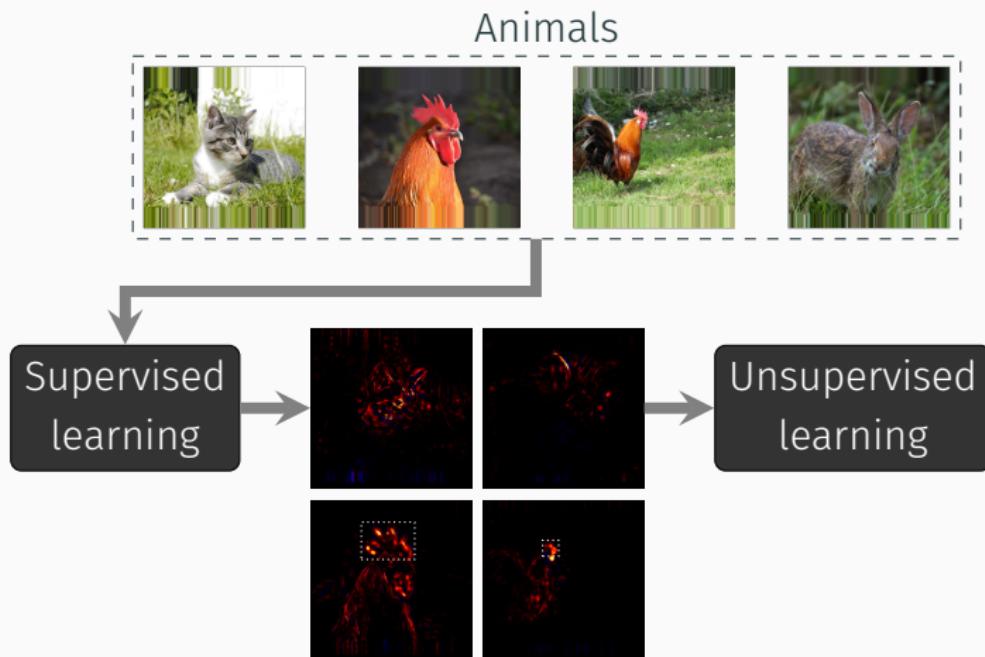
Explainable AI: The central idea



Explainable AI: The central idea



Explainable AI: The central idea



Explainable AI: Caveats



Predictive model



Bird



Explainable AI and the brain

Characterizing heterogeneity in diverse clinical cohorts



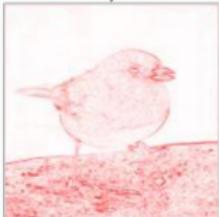
Explainable AI: Caveats



Predictive model



Bird



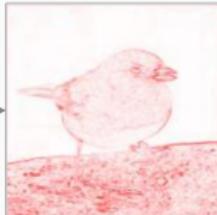
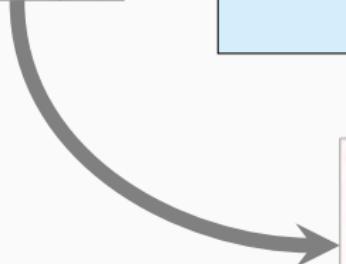
Explainable AI: Caveats



Predictive model



Bird

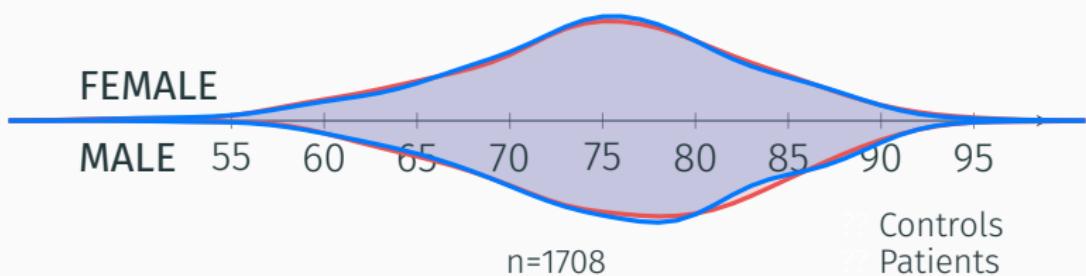


Explainable AI and dementia predictions

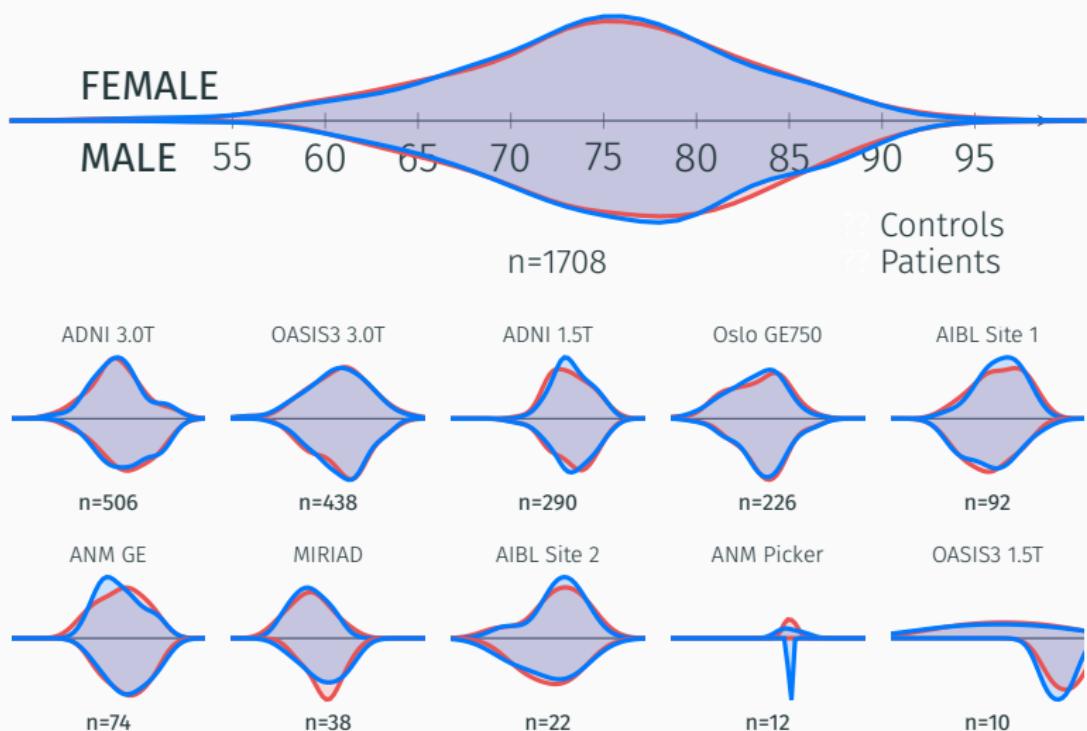


UNIVERSITETET
I OSLO

Explainable dementia classification: Dataset



Explainable dementia classification: Dataset

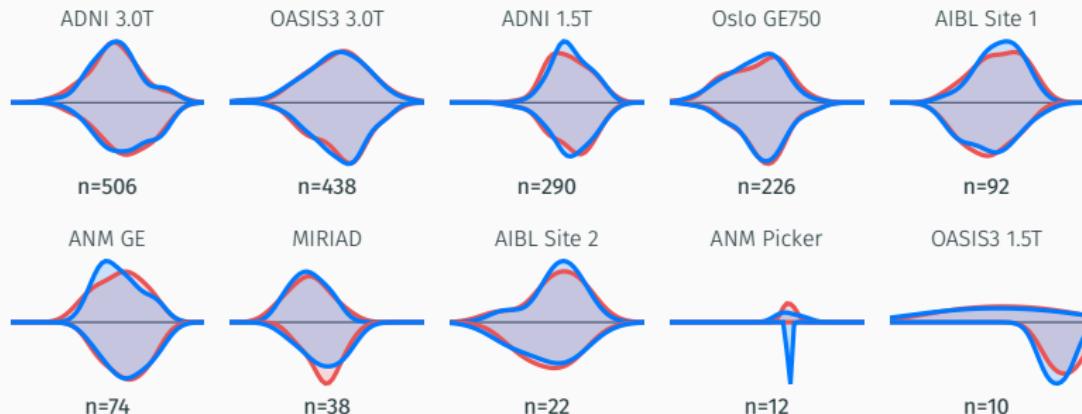


Explainable dementia classification: Dataset

ADNI:
Probable
Alzheimer's disease

OASIS3:
Dementia

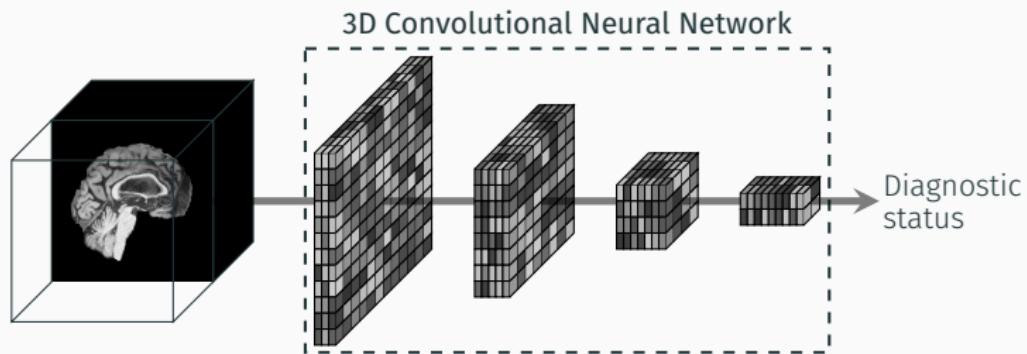
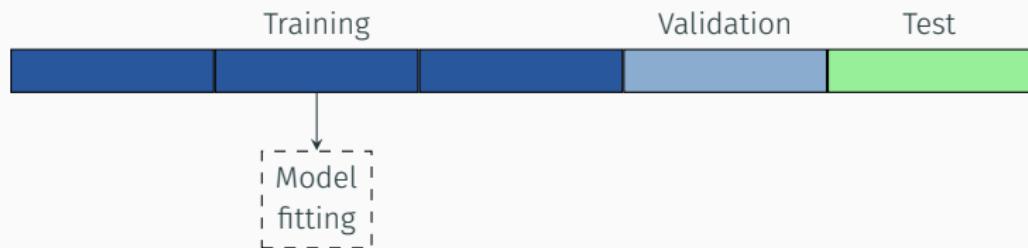
Oslo:
Alzheimer's disease,
Vascular dementia,
Unspecified dementia



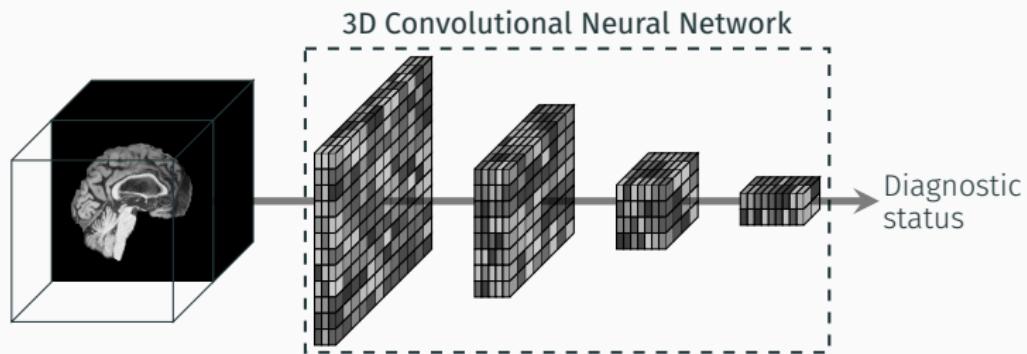
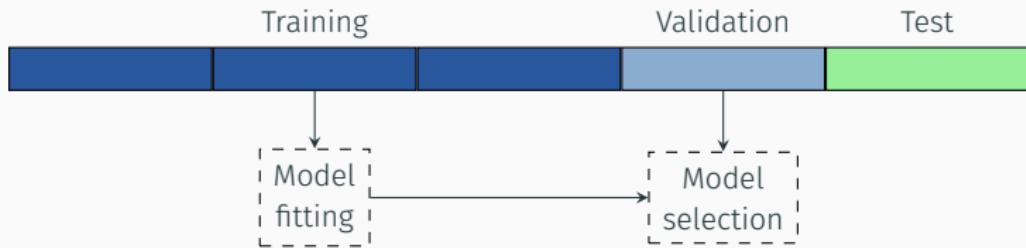
Explainable dementia classification: Modelling



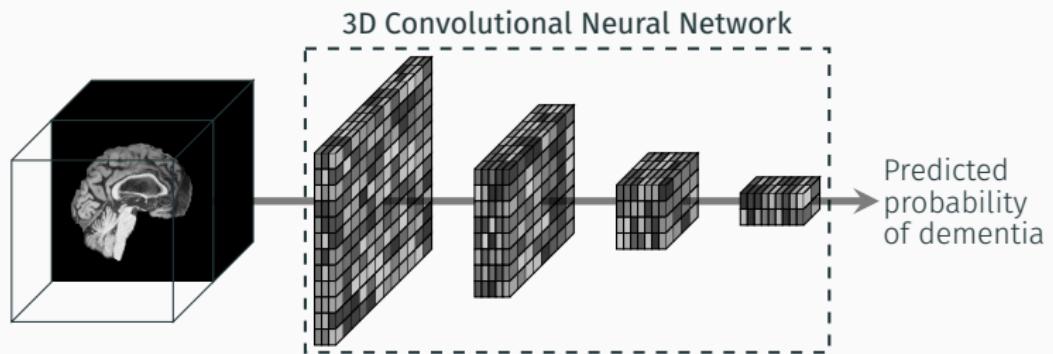
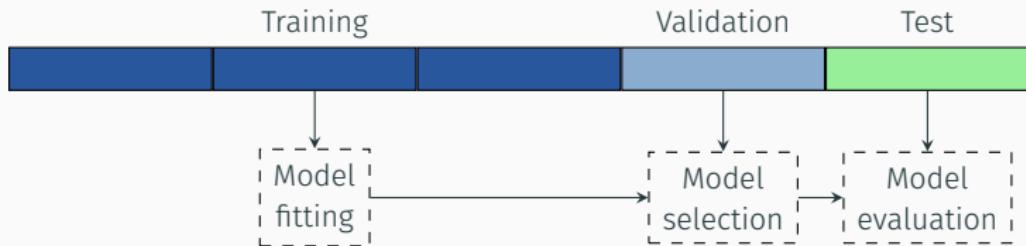
Explainable dementia classification: Modelling



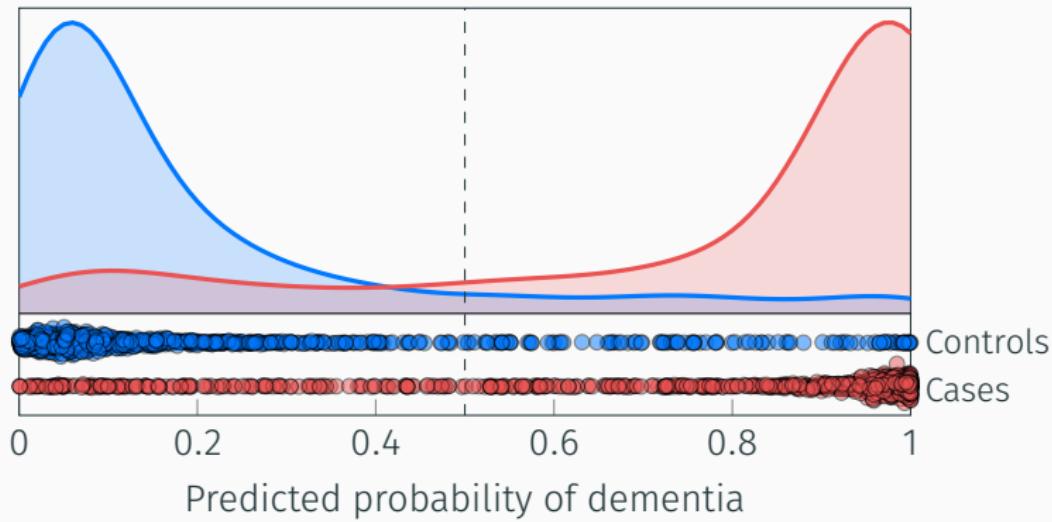
Explainable dementia classification: Modelling



Explainable dementia classification: Modelling



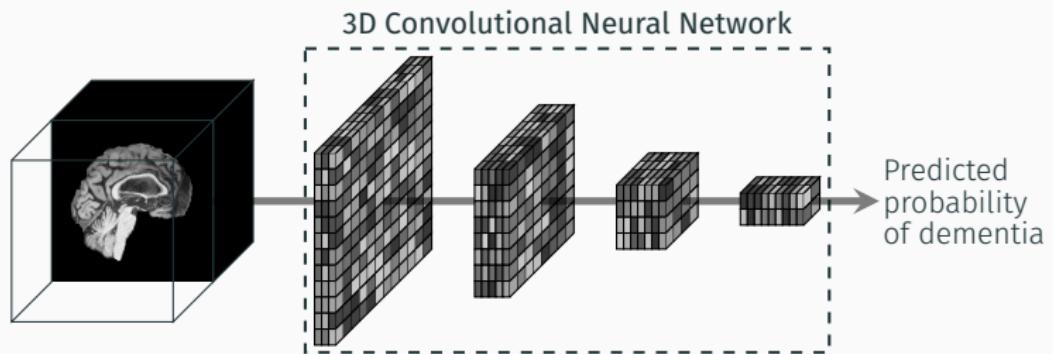
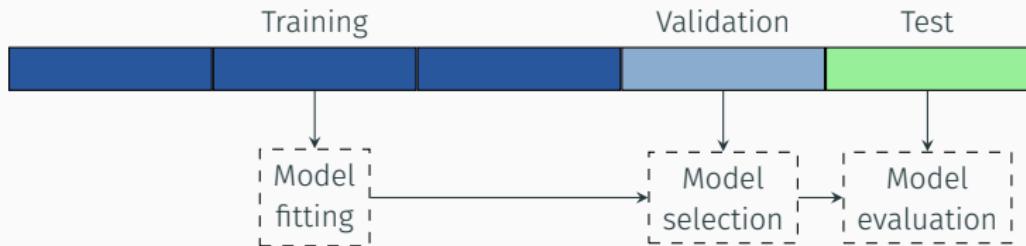
Explainable dementia classification: Modelling



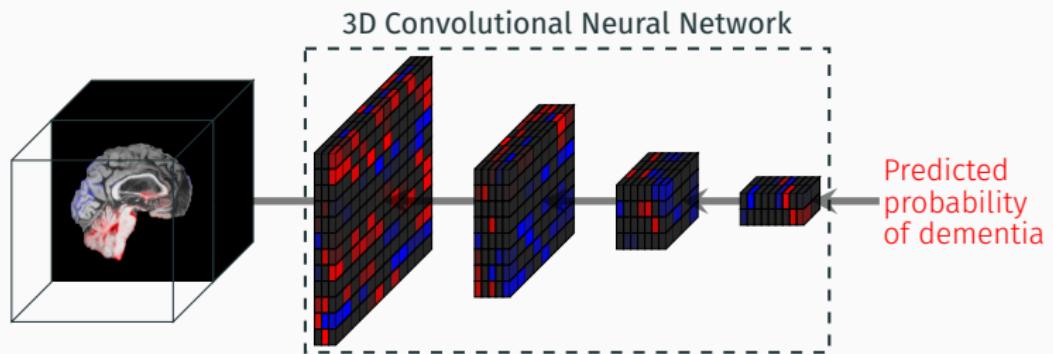
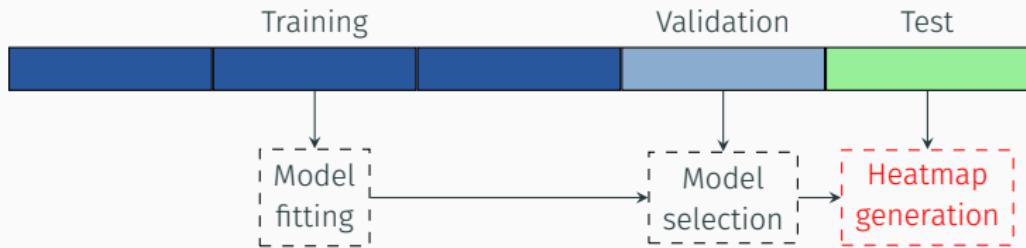
AUC=0.91, balanced accuracy=85%



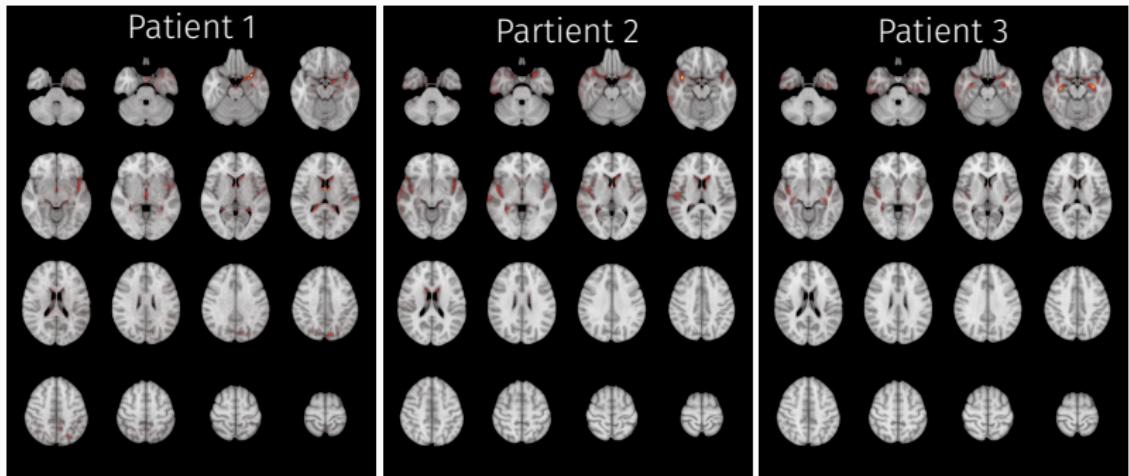
Explainable dementia classification: Modelling



Explainable dementia classification: Modelling

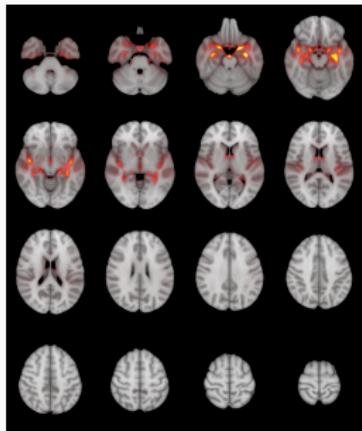


Explainable dementia classification: Modelling



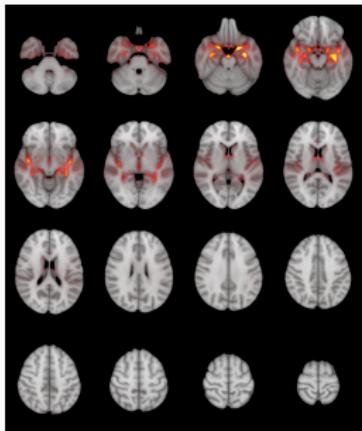
Explainable dementia classification: Validation

Explainable AI

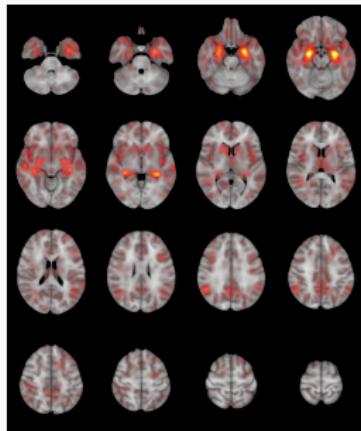


Explainable dementia classification: Validation

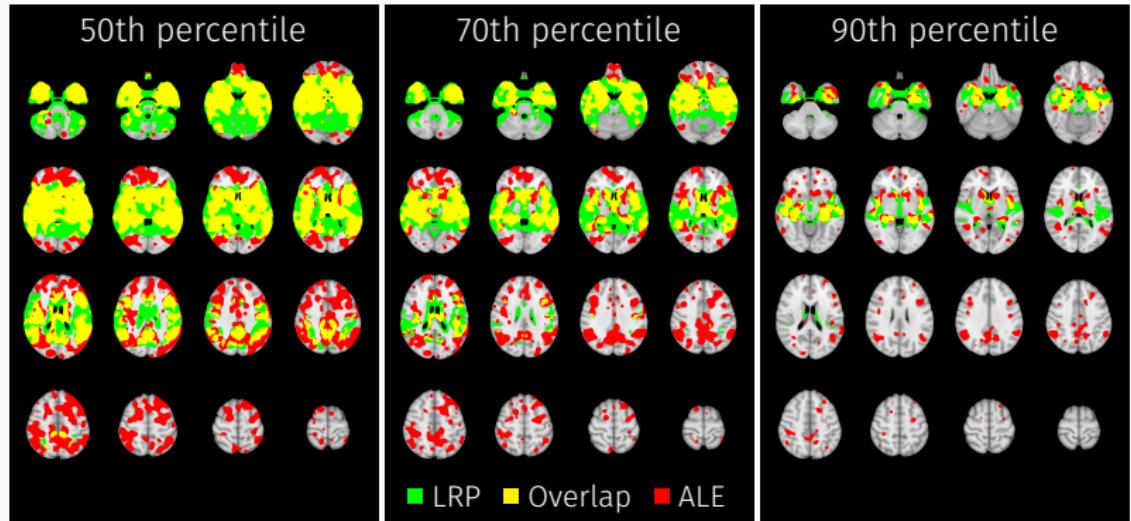
Explainable AI



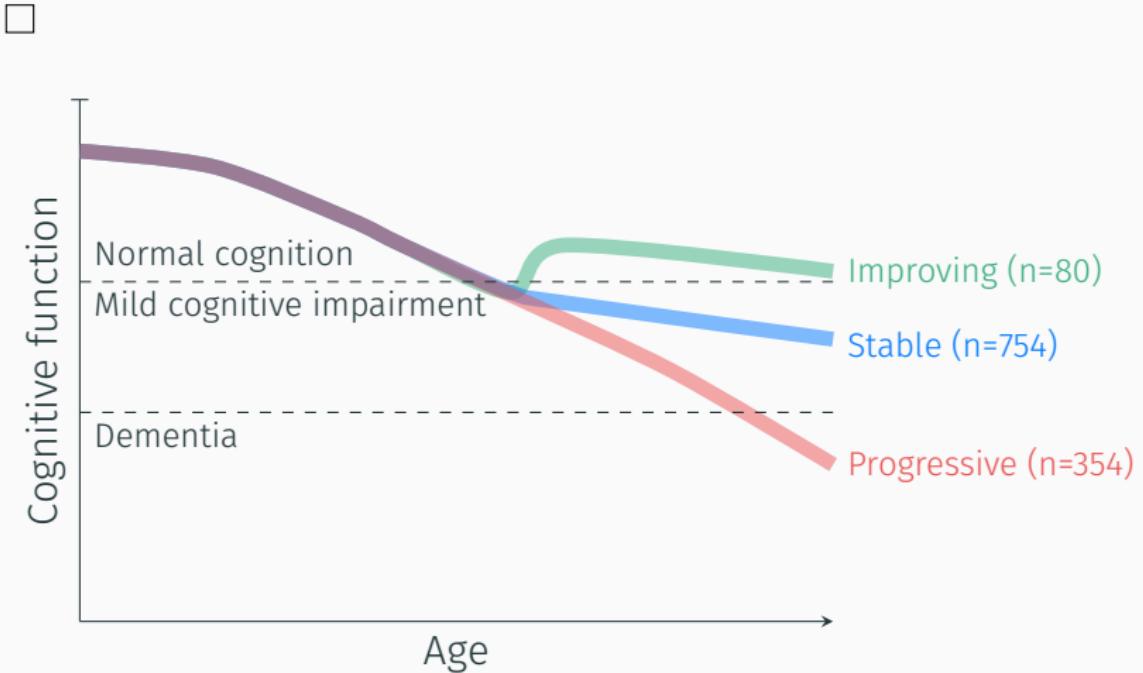
Human researchers



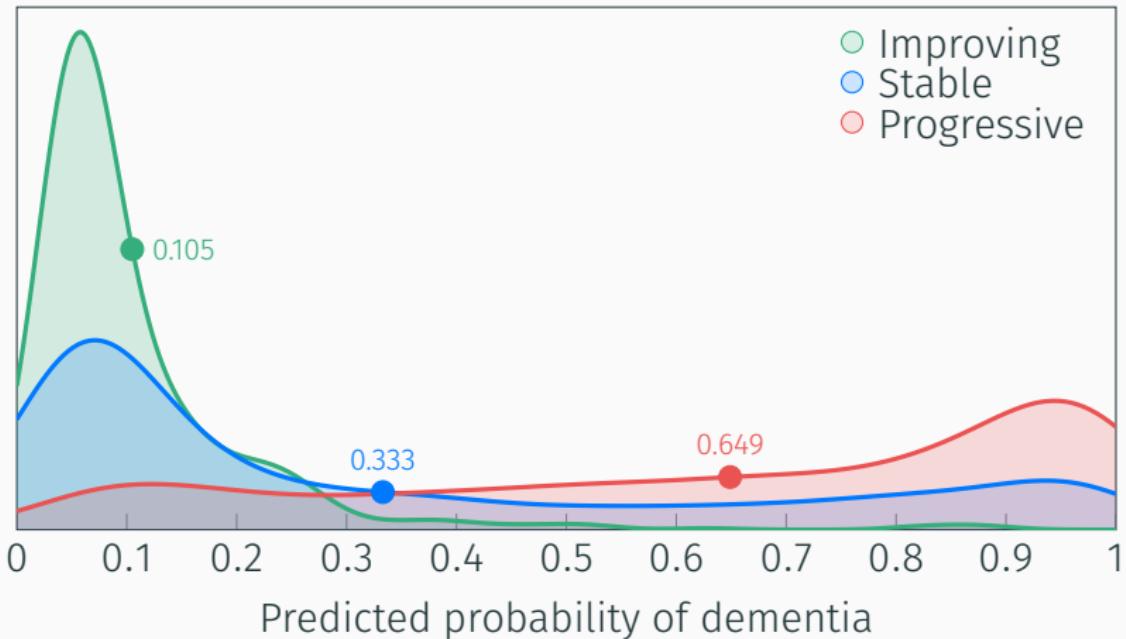
Explainable dementia classification: Validation



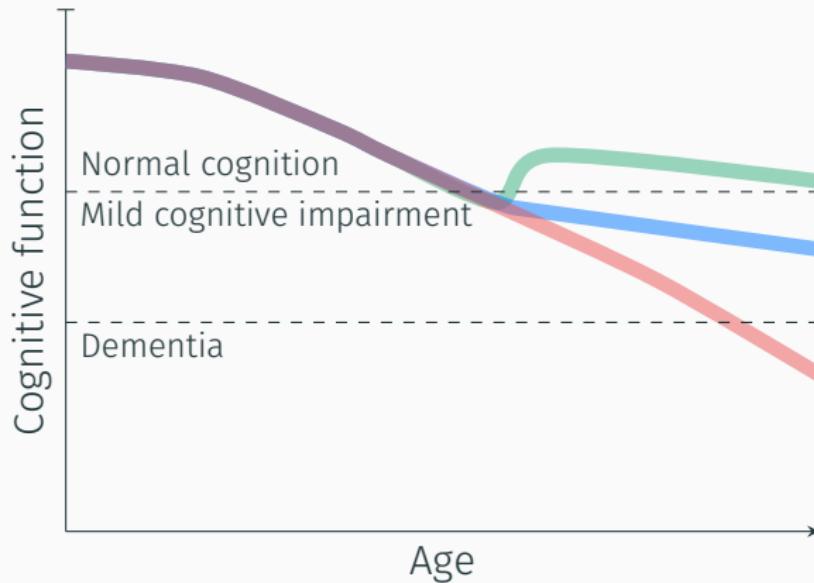
Explainable dementia classification: Application



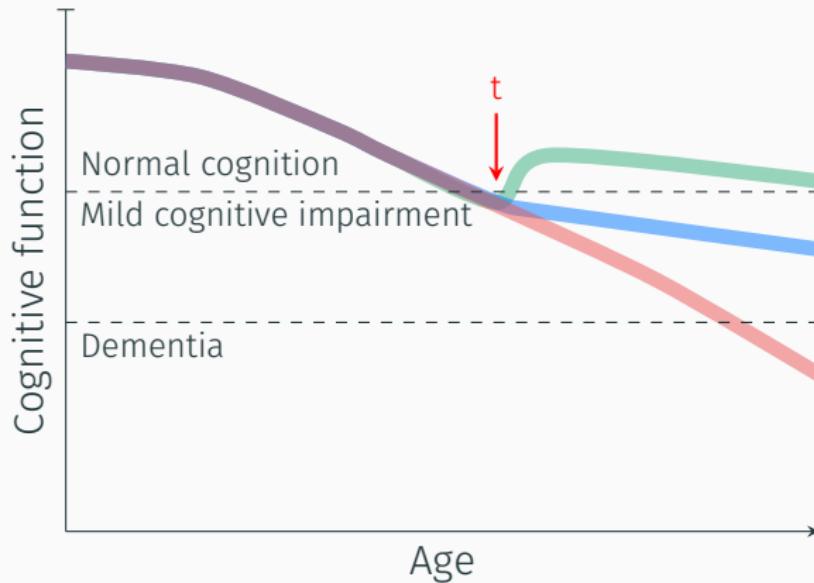
Explainable dementia classification: Application



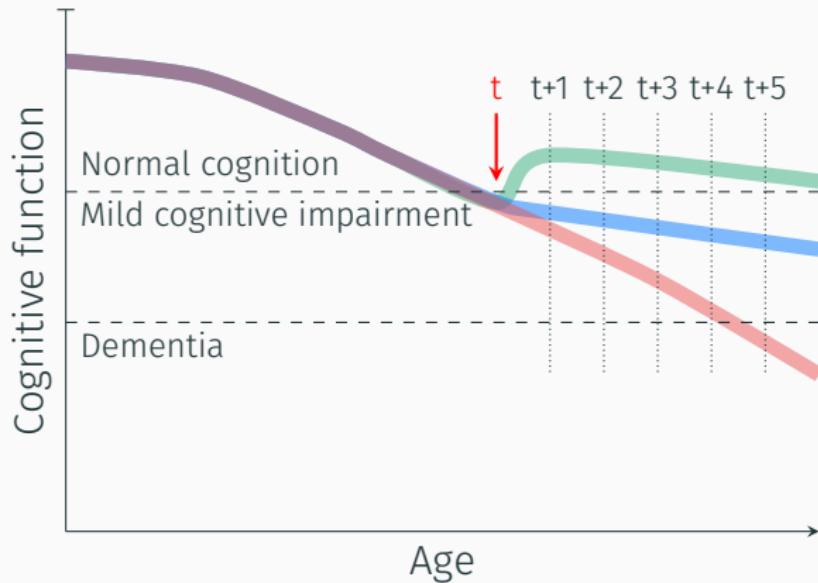
Explainable dementia classification: Application



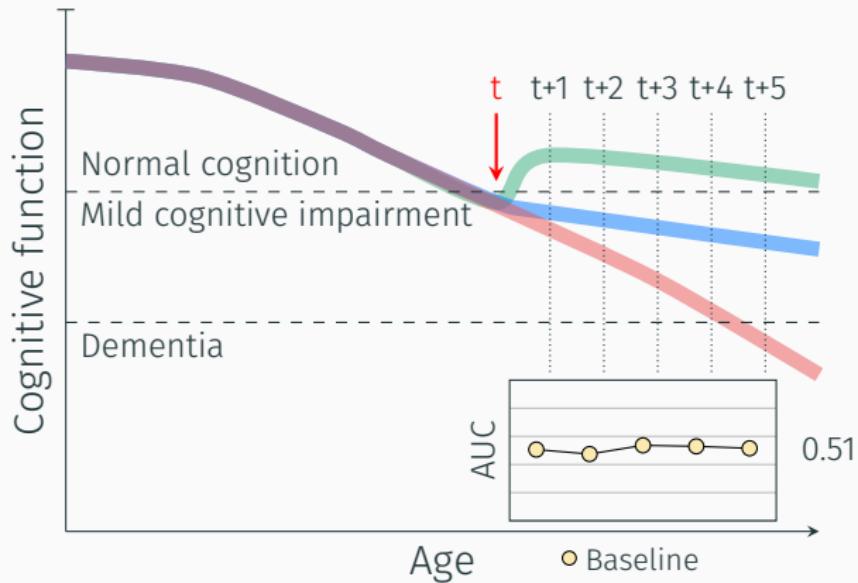
Explainable dementia classification: Application



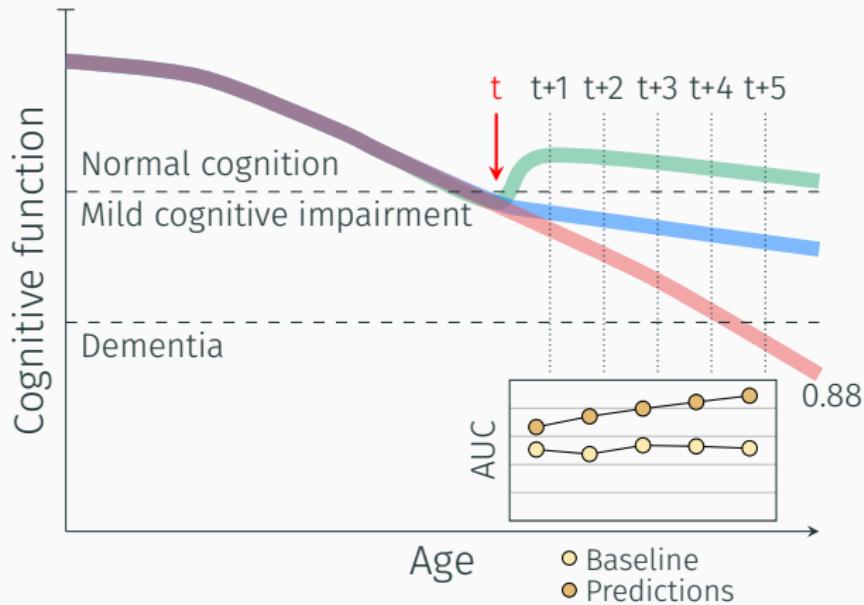
Explainable dementia classification: Application



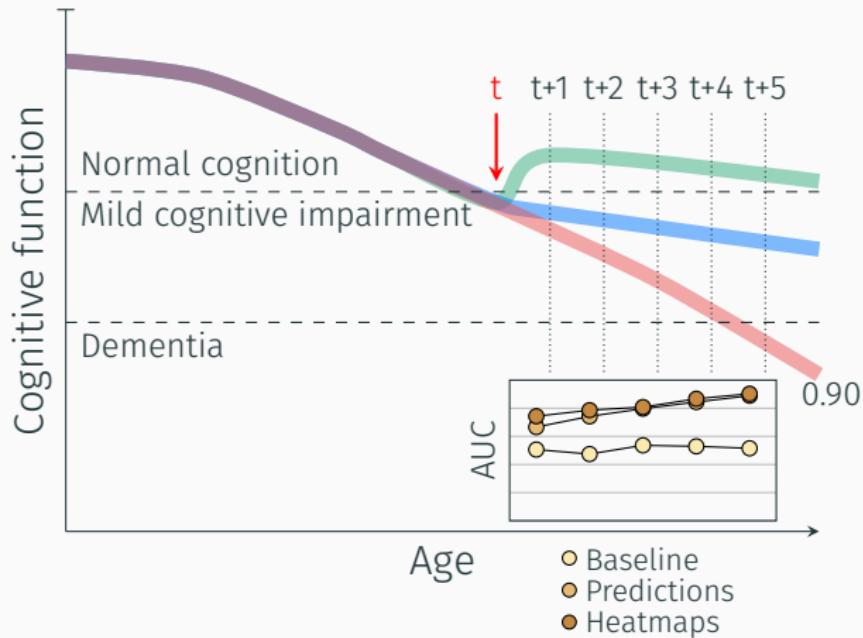
Explainable dementia classification: Application



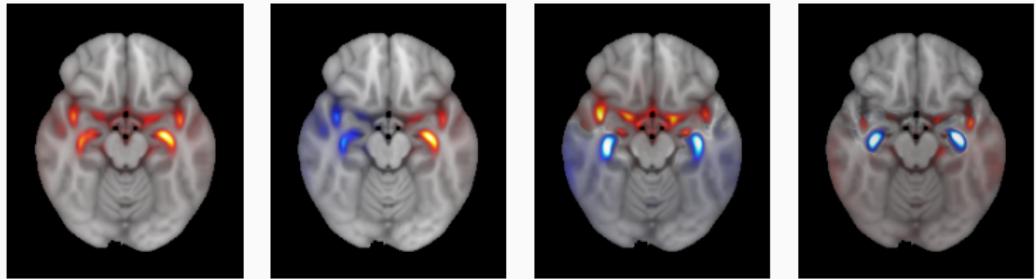
Explainable dementia classification: Application



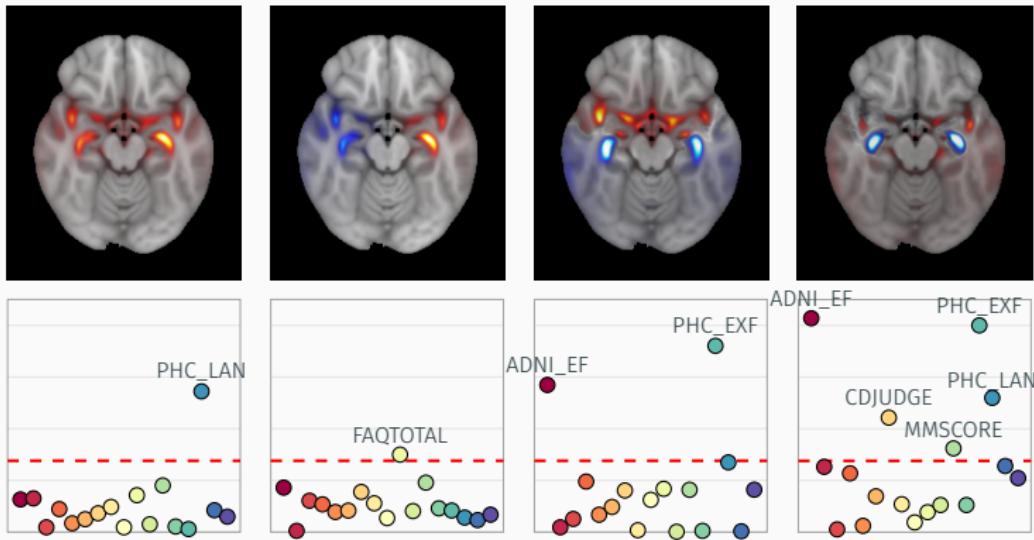
Explainable dementia classification: Application



Explainable dementia classification: Application



Explainable dementia classification: Application



Explainable dementia classification: Application



We used explainable AI to generate heatmaps characterizing the manifestation of dementia-related aberrations in individual brains

- Information in the heatmaps allowed us to predict progression from mild cognitive impairment to dementia with an AUC of 0.9 after 5 years
- Distinct patterns in the heatmaps were associated with decline in various cognitive domains

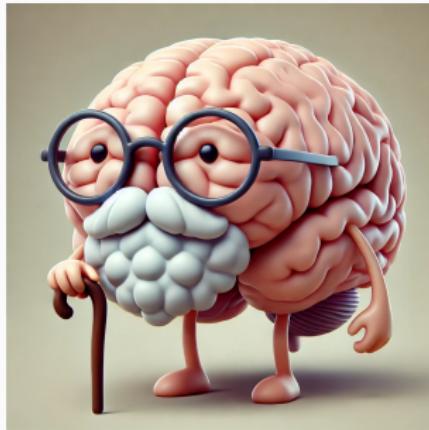


Explainable AI and brain age predictions



UNIVERSITETET
I OSLO

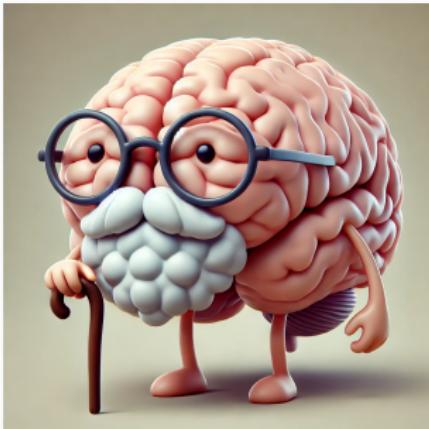
Explainable brain age: Motivation



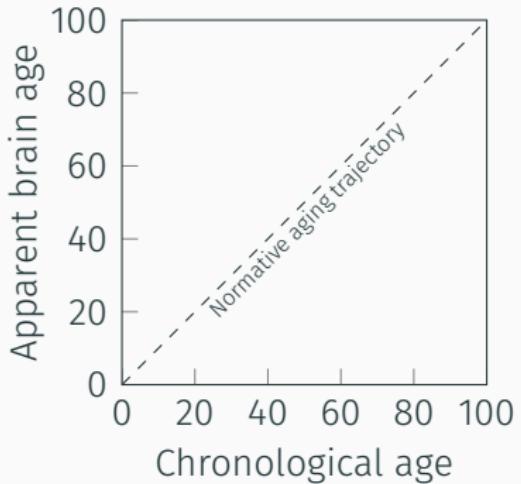
Generated by Dall-E 3



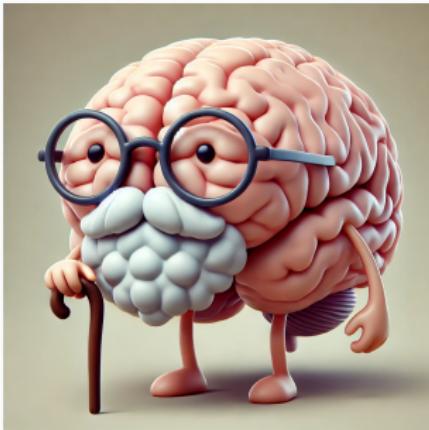
Explainable brain age: Motivation



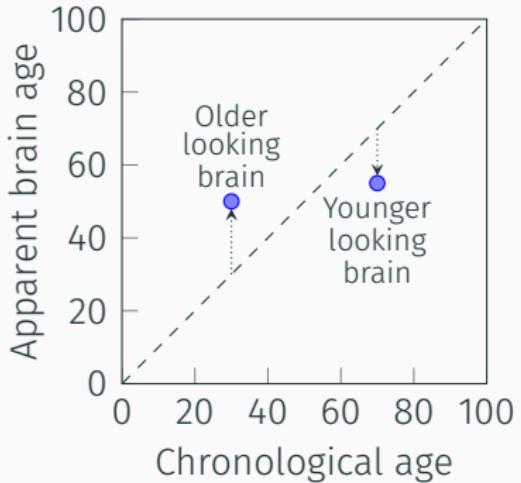
Generated by Dall-E 3



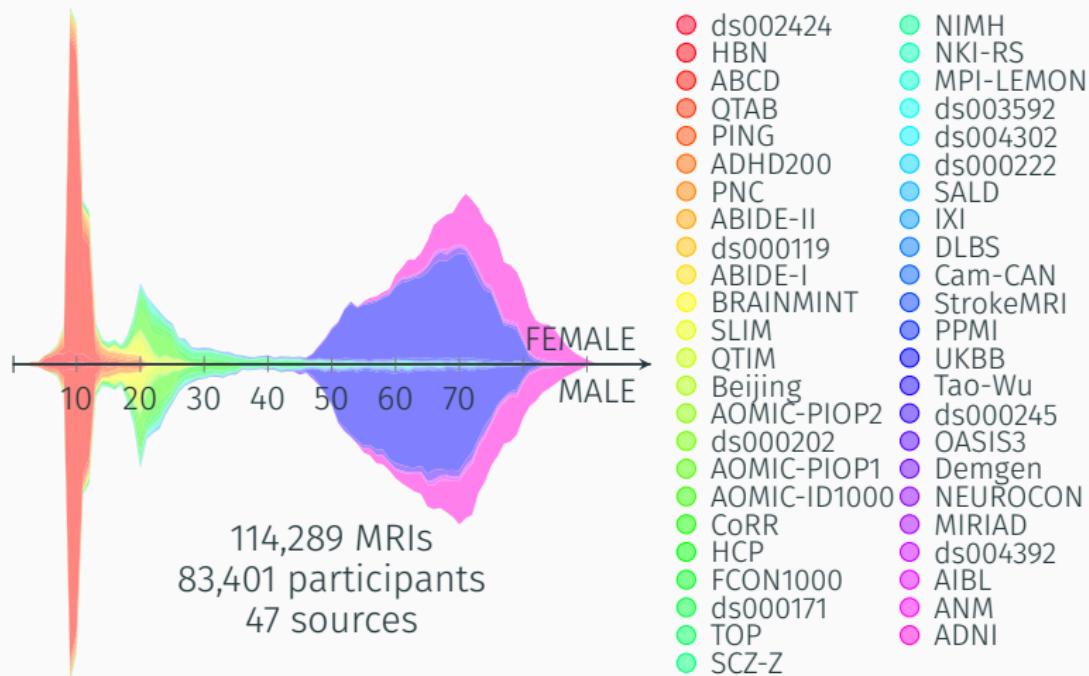
Explainable brain age: Motivation



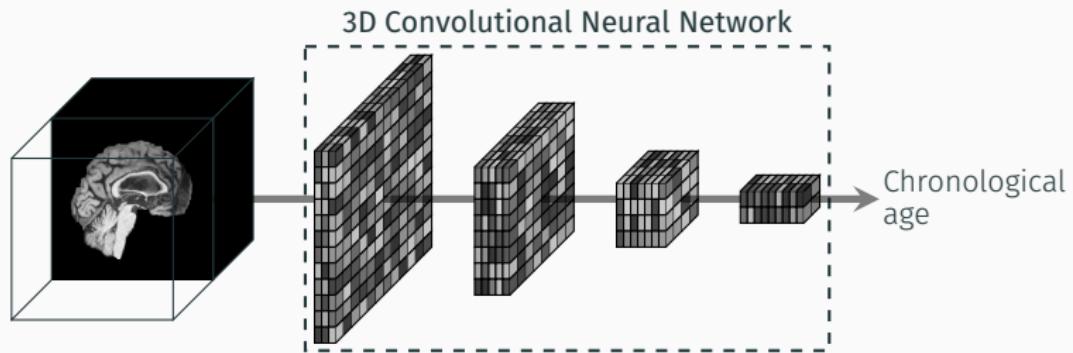
Generated by Dall-E 3



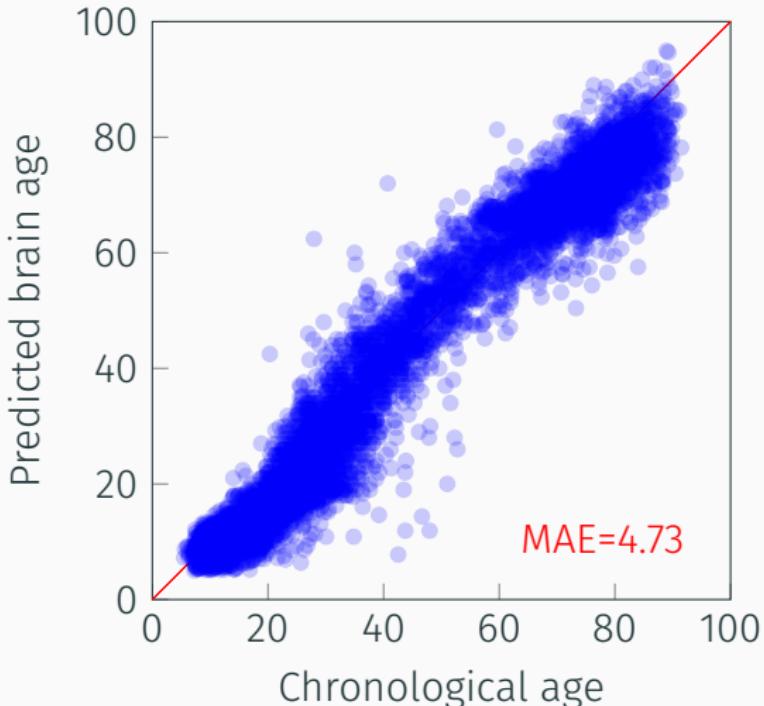
Explainable brain age: Motivation



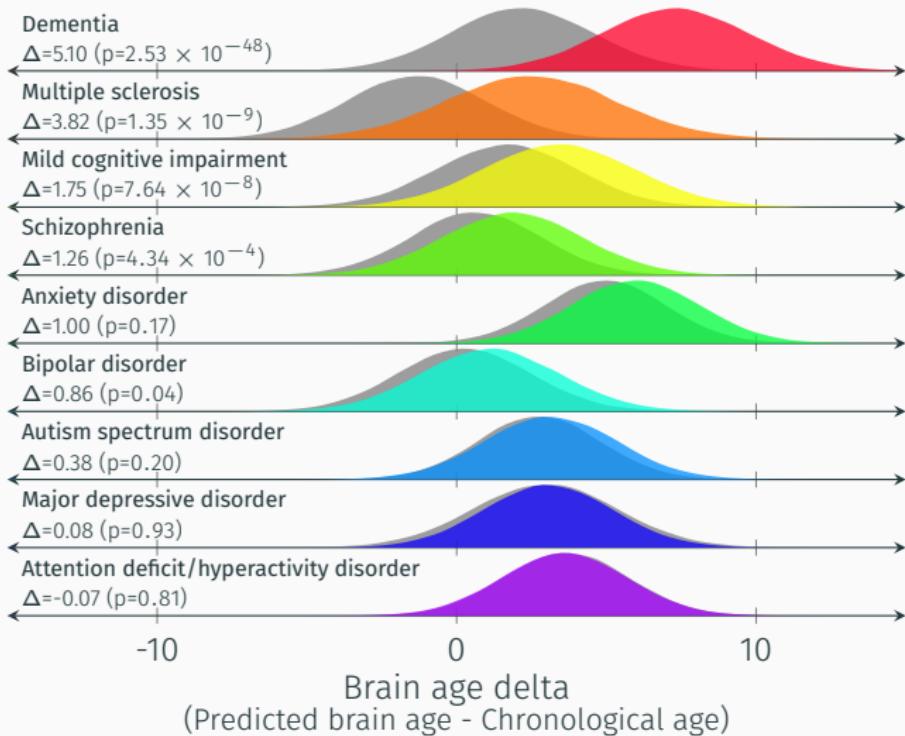
Explainable brain age: Motivation



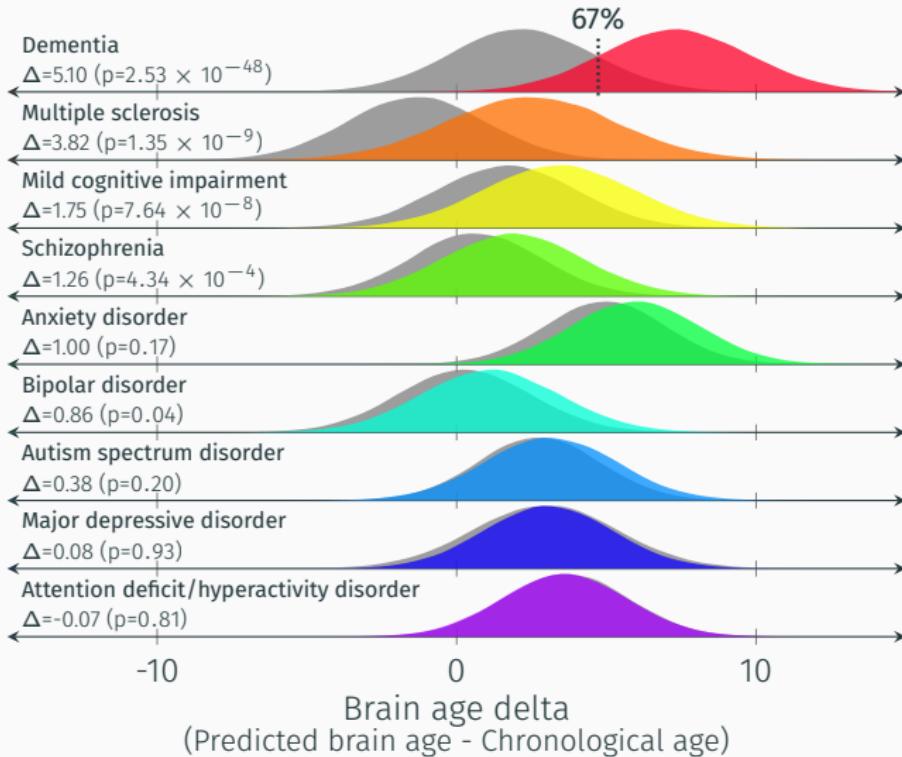
Explainable brain age: Motivation



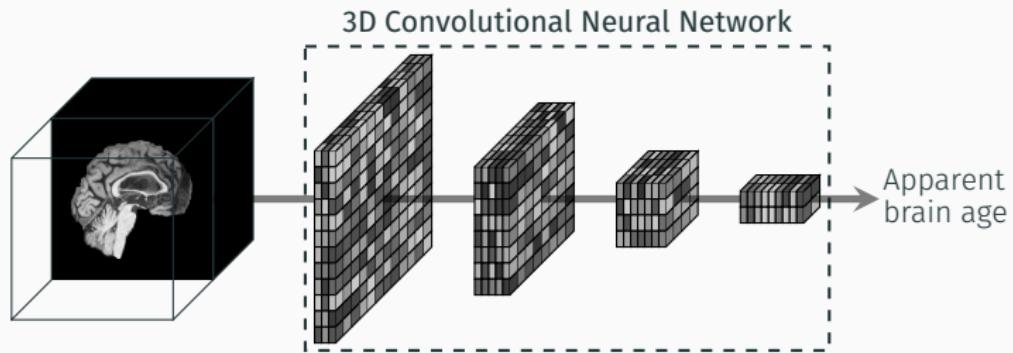
Explainable brain age: Motivation



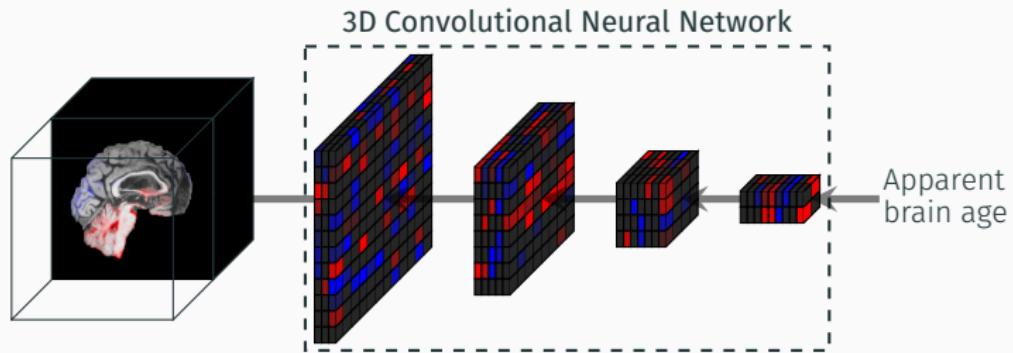
Explainable brain age: Motivation



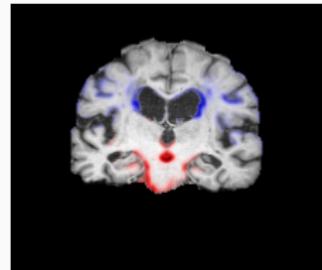
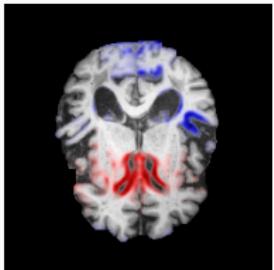
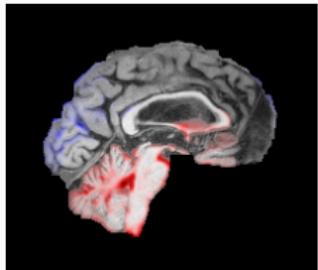
Explainable brain age: Methods



Explainable brain age: Methods



Explainable brain age: Methods

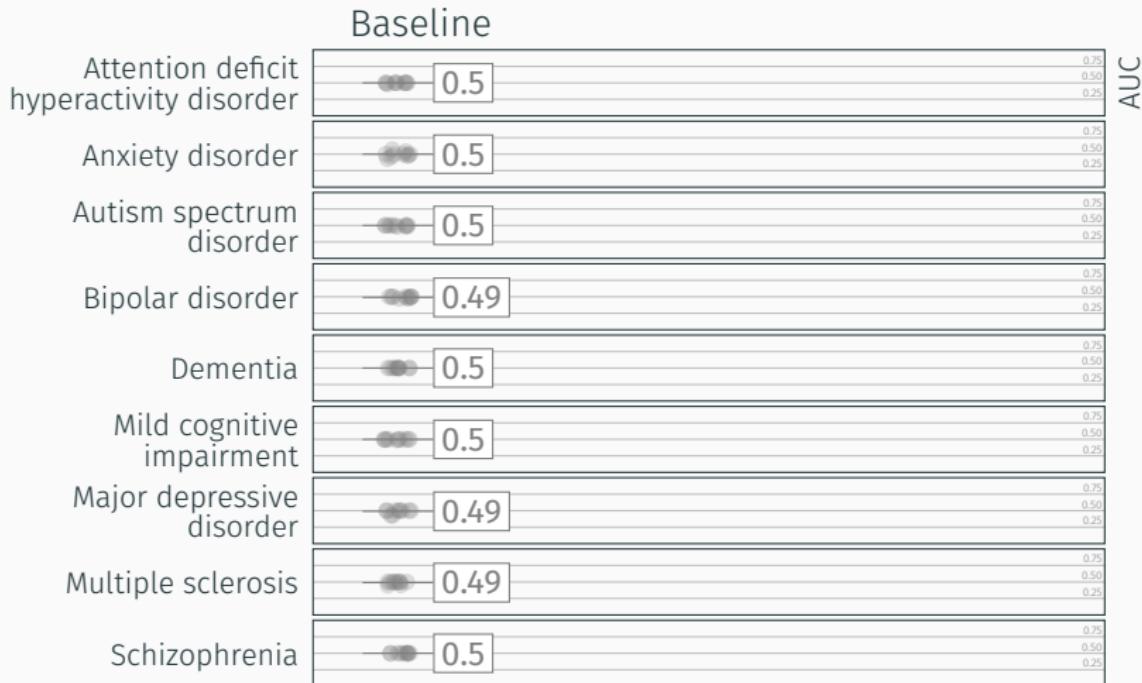


Younger
appearing

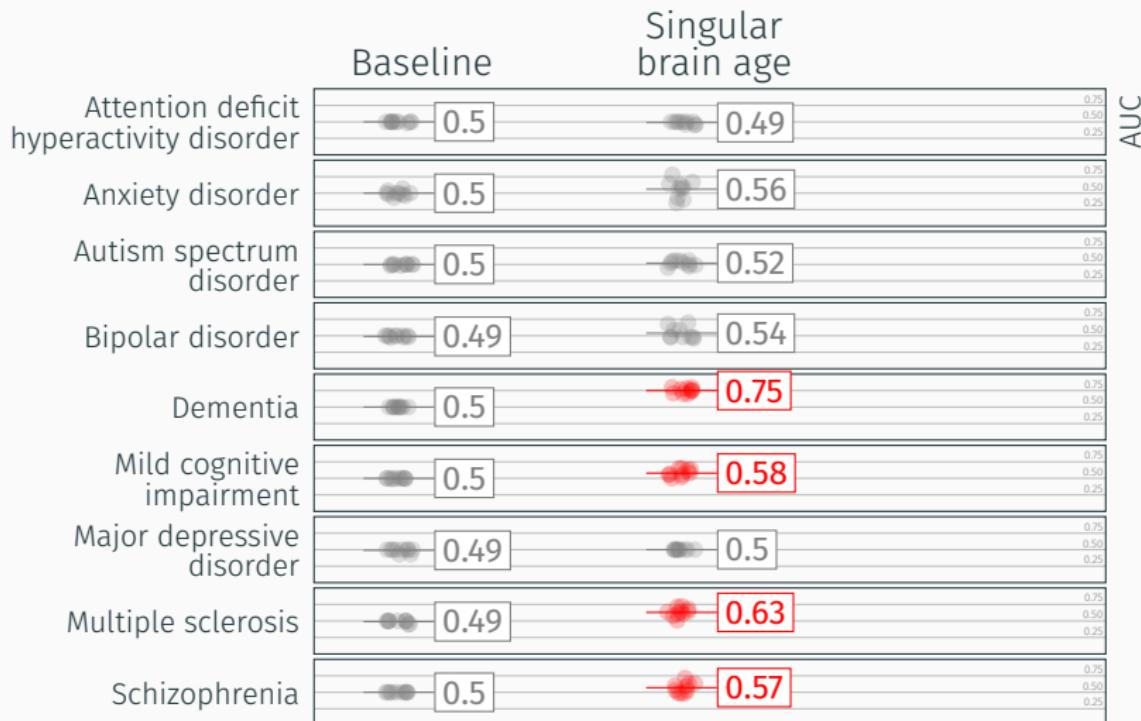
Older
appearing



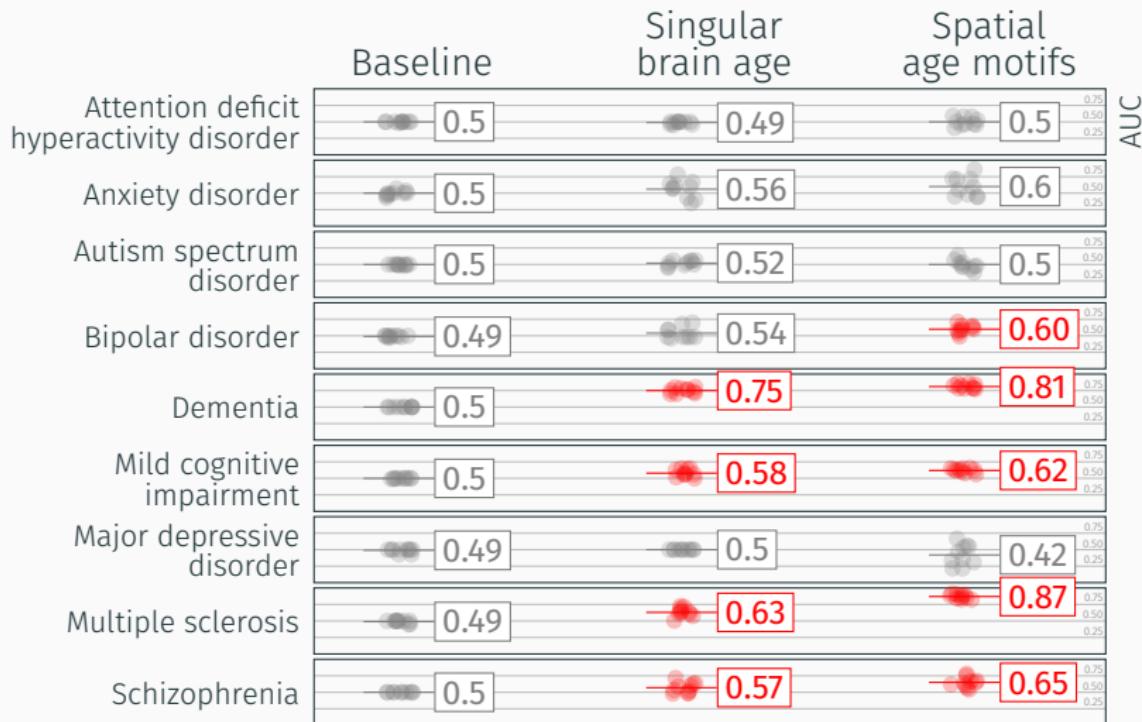
Explainable brain age: Results



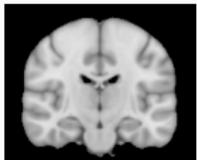
Explainable brain age: Results



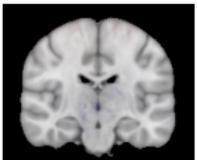
Explainable brain age: Results



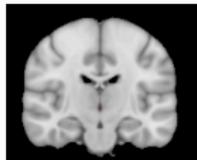
Explainable brain age: Results



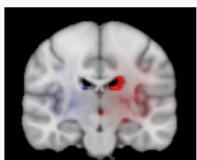
Attention deficit
hyperactivity disorder



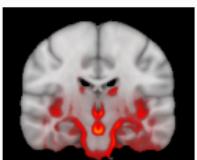
Anxiety
disorder



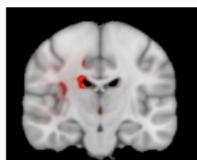
Autism spectrum
disorder



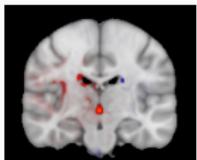
Bipolar
disorder



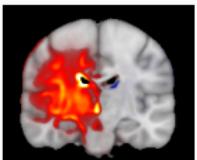
Dementia



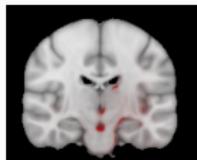
Schizophrenia



Major depressive
disorder



Multiple
sclerosis



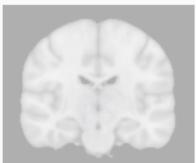
Mild cognitive
impairment



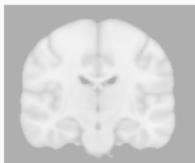
Explainable brain age: Results



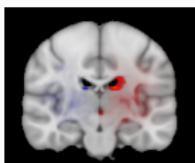
Attention deficit
hyperactivity disorder



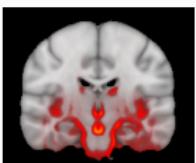
Anxiety
disorder



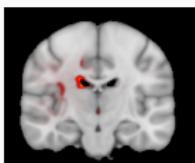
Autism spectrum
disorder



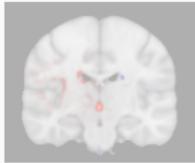
Bipolar
disorder



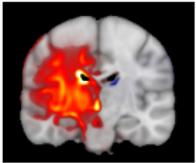
Dementia



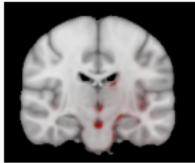
Schizophrenia



Major depressive
disorder



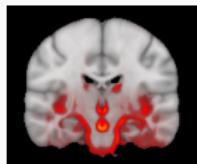
Multiple
sclerosis



Mild cognitive
impairment



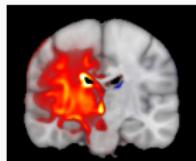
Explainable brain age: Results



Dementia

- Temporal pole
- Left amygdala
- Parahippocampal gyrus

Explainable brain age: Results

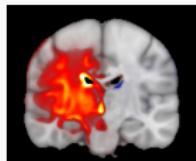


Multiple
sclerosis

- Right pallidum
- Right thalamus
- Right putamen



Explainable brain age: Results

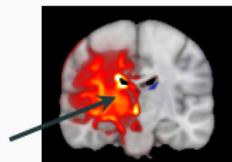


Multiple
sclerosis

- Right pallidum
- Right thalamus
- Right putamen



Explainable brain age: Results

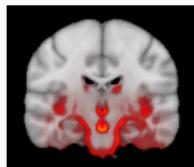


Multiple
sclerosis

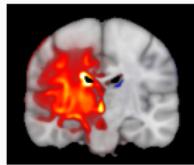
- Right pallidum
- Right thalamus
- Right putamen



Explainable brain age: Results



Dementia

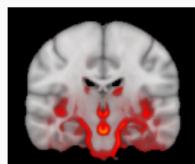


Multiple
sclerosis



Explainable brain age: Results

AUC



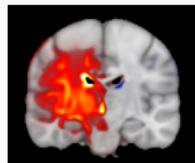
Dementia

Singular
brain age

0.75

Spatial
age motifs

0.81



Multiple
sclerosis

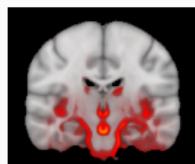
0.63

0.87



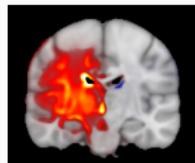
Explainable brain age: Results

AUC



Dementia

Singular brain age	Spatial age motifs
0.75	0.81
0.06	

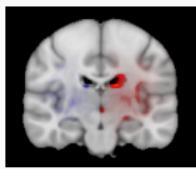


Multiple
sclerosis

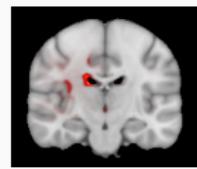
0.63	0.87
0.24	



Explainable brain age: Results



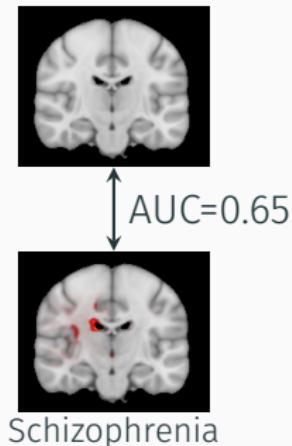
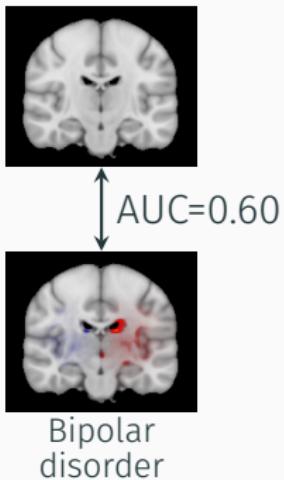
Bipolar
disorder



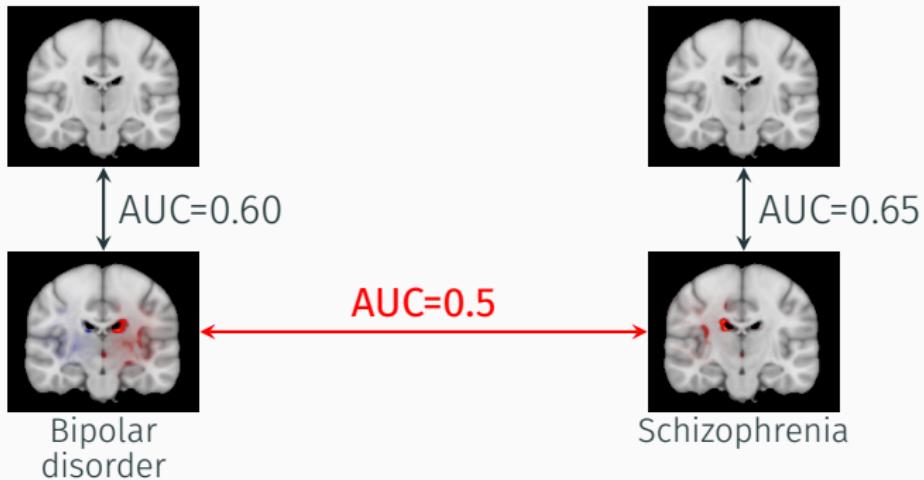
Schizophrenia



Explainable brain age: Results



Explainable brain age: Results



Explainable brain age: Results

We used explainable AI to generate heatmaps characterizing how different regions of the brain looked younger or older than expected in individual participants

- Spatial information about deviant aging allowed us improve case-control classification in multiple disorders
- Triangulation between average heatmaps in different disorders revealed both common and disorder-specific patterns
- **Results exposed weaknesses of the methodology, emphasizing the importance of rigorous validation**

