Workshop on Interpretability in Machine Intelligence in Medical Imaging Computing, **iMIMIC**
October 6th, 2024, Marrakech, Morocco

MICCAI 2024 Marrakesh MOROCCO

PIPNet3D: Interpretable Detection of Alzheimer in MRI Scans

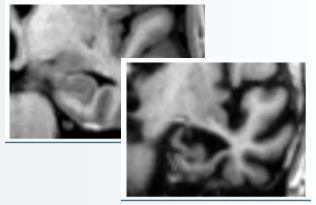
Lisa Anita De Santi, Jörg Schlötterer,
Michael Scheschenja, Joel Wessendorf , Meike Nauta,
Vincenzo Positano, Christin Seifert

Fondazione Monasterio
la ricerca che cura

Università di Pisa

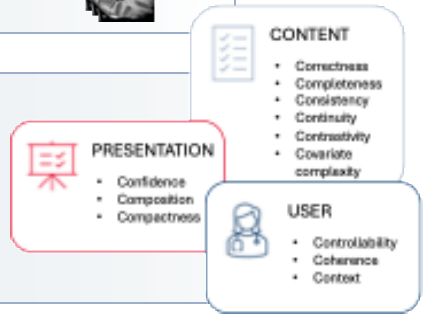
Dipartimento di Ingegneria dell'Informazione

Introduction



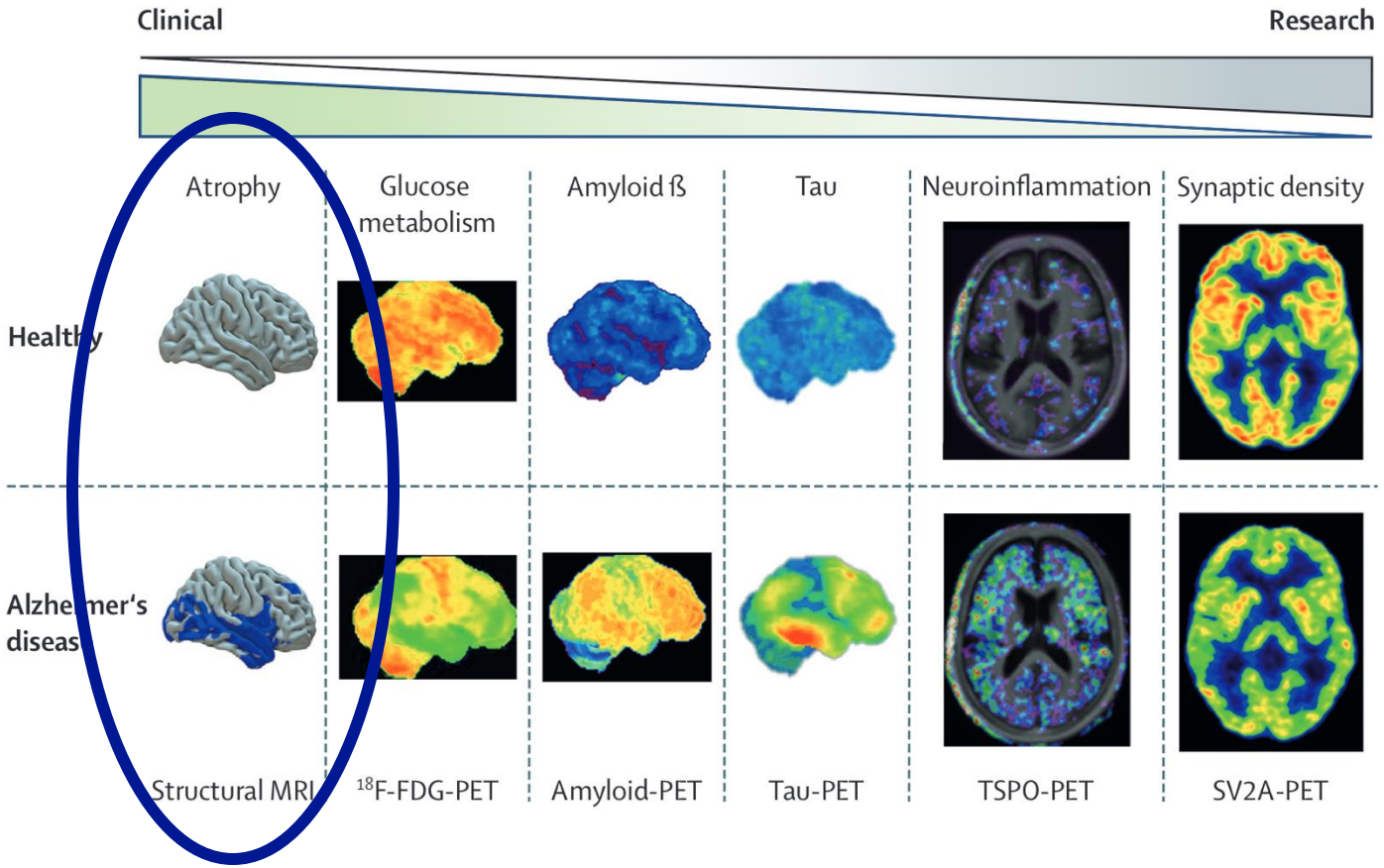
PIPNet3D

Prototype Evaluation



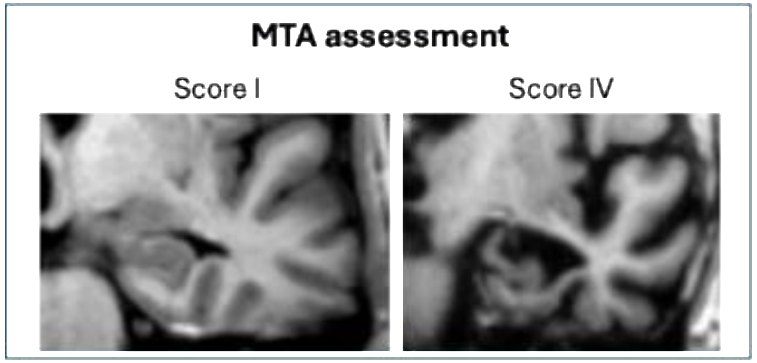
Conclusions

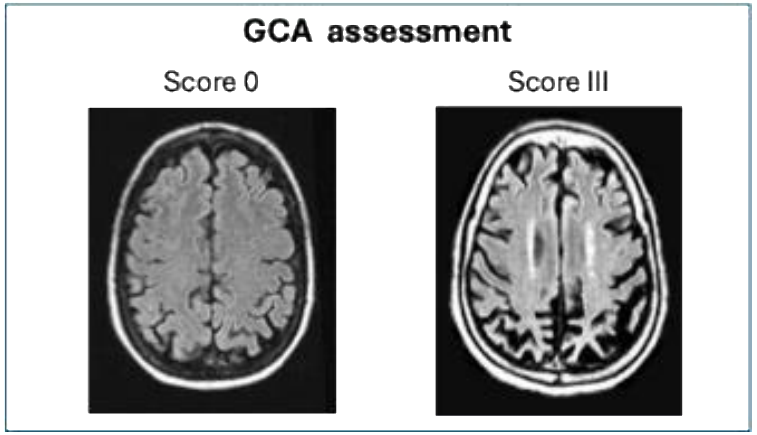
Alzheimer's Disease Diagnosis from sMRI

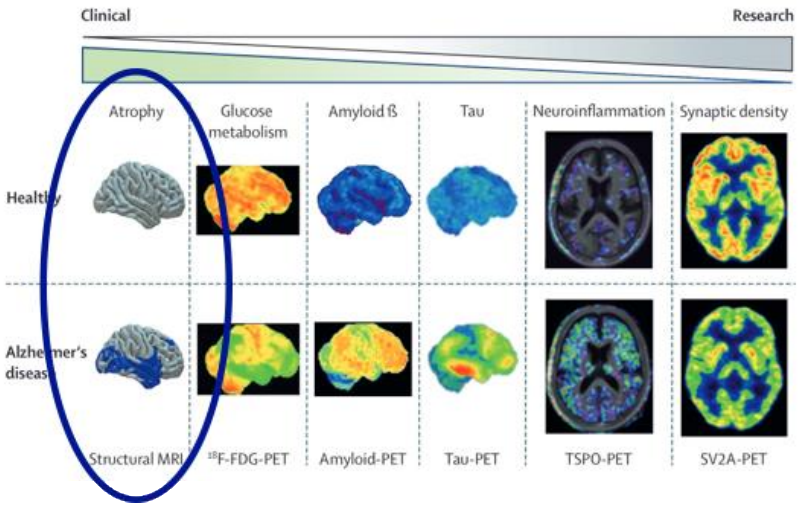


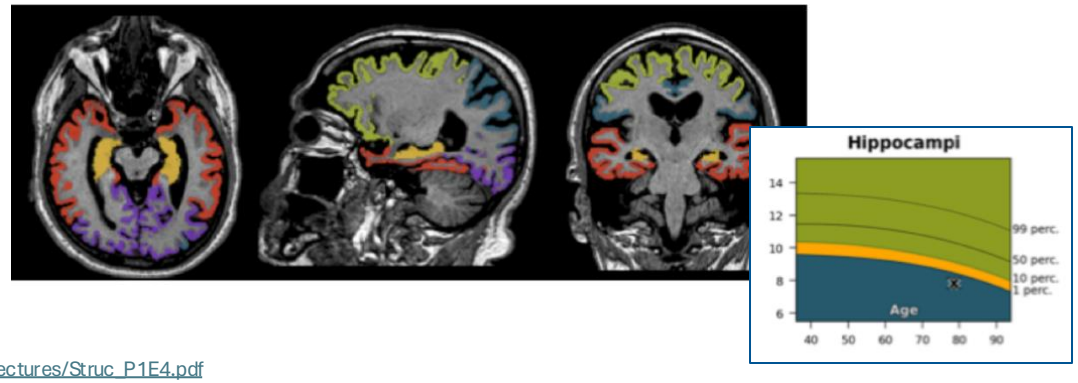
Source: https://doi.org/10.1016/S1474-4422(20)30314-8; https://alzimaging.com/ ; https://open.win.ox.ac.uk/pages/fslcourse/lectures/Struc_P1E4.pdf

Alzheimer’s Disease Diagnosis from sMRI









Source: https://doi.org/10.1016/S1474-4422(20)30314-8; https://alzimaging.com/ ; https://open.win.ox.ac.uk/pages/fslcourse/lectures/Struc_P1E4.pdf

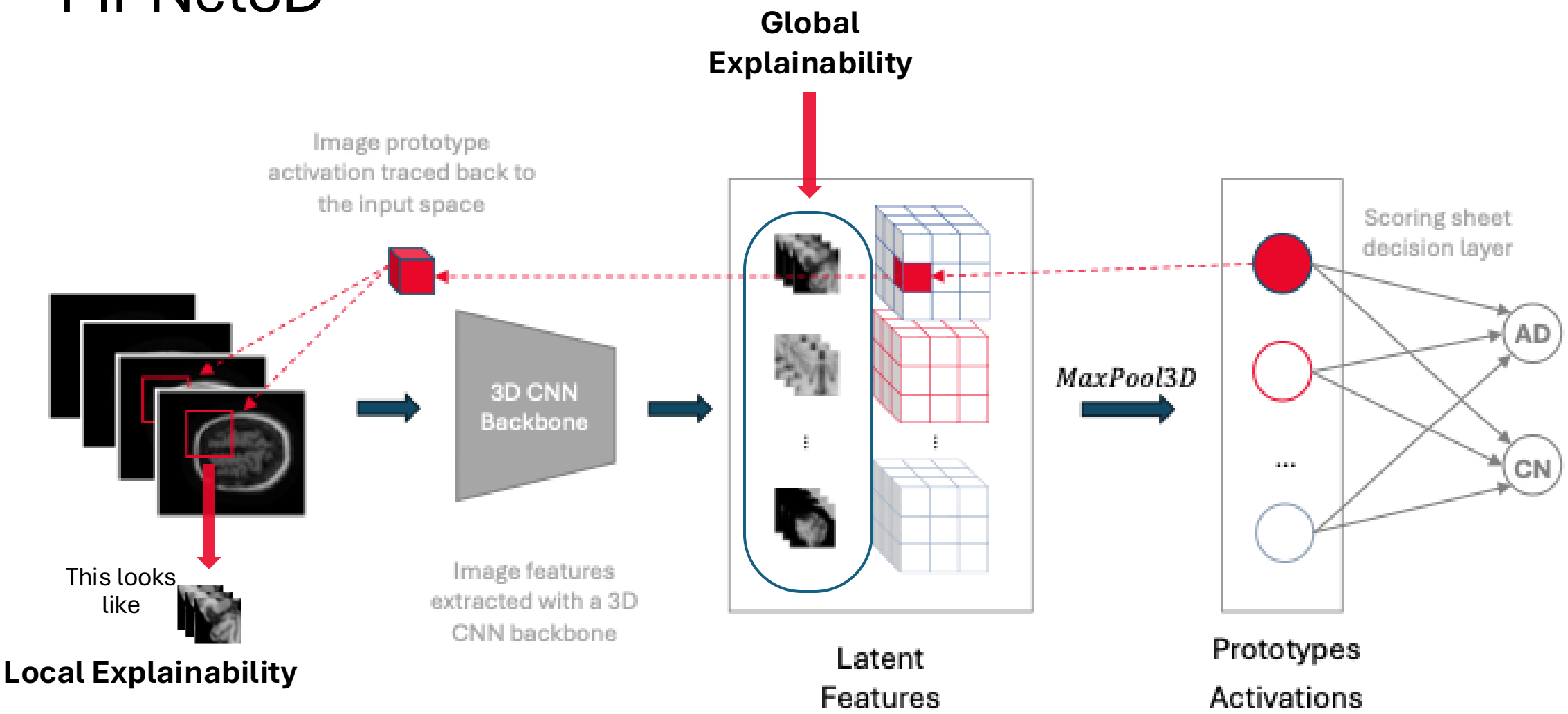
Alzheimer's Disease Diagnosis from sMRI

In summary:

- Performing an early, accurate, and objective **diagnosis** of **Alzheimer's Disease** is still an open challenge
- **sMRI image-biomarkers** are routinely employed to support clinical evaluation with semi-quantitative scales and automated software
  - Brain atrophy in Medial Temporal Lobe structures (e.g. entorhinal cortex, hippocampus), Limbic structures, Cortical region, white matter hyperintensities in Frontal lobes
- Current guidelines and practices present **limited detection capabilities**

➢ Exploit **Explainable Deep Learning** to support sMRI analysis and for potential image-biomarkers discovery

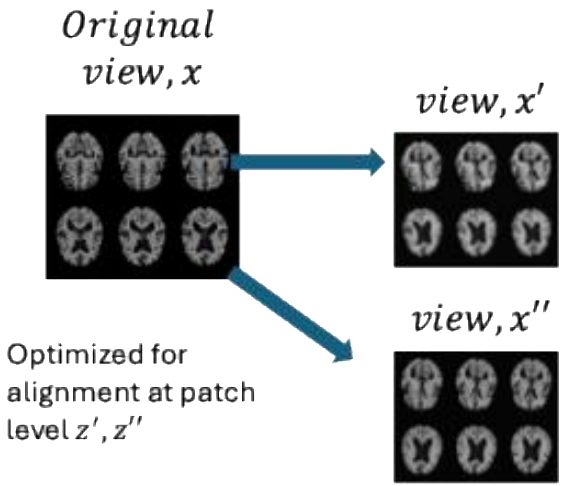
PIPNet3D


Global Explainability
Image prototype activation traced back to the input space
3D CNN Backbone
Image features extracted with a 3D CNN backbone
MaxPool3D
Scoring sheet decision layer
AD
CN
...
This looks like
Local Explainability
Latent Features
Prototypes Activations

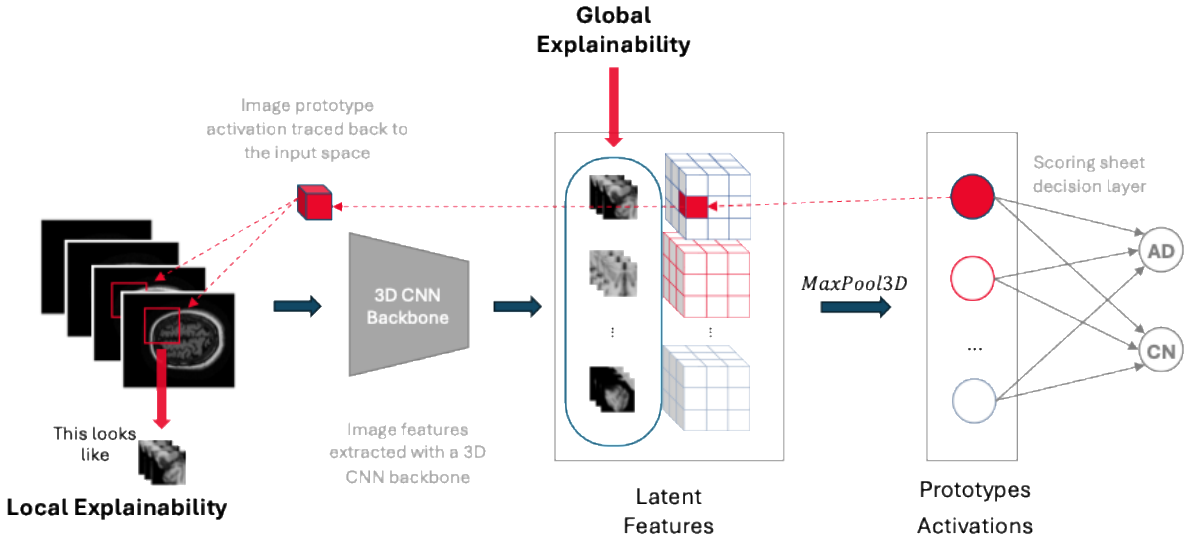
PIPNet3D

Learn **semantically meaningful** image prototypes

Self-supervised training paradigm to automatically learn image part-prototypes in line with human-concepts



$$\mathcal{L}_A = -\frac{1}{HW}\sum_{h,w}\log(z'_{h,w,:}\cdot z''_{h,w,:})$$



$$o = \log((pw_c)^2 + 1)$$

Optimize for **compactness**

Loss function that optimizes for a sparse scoring sheet decision layer

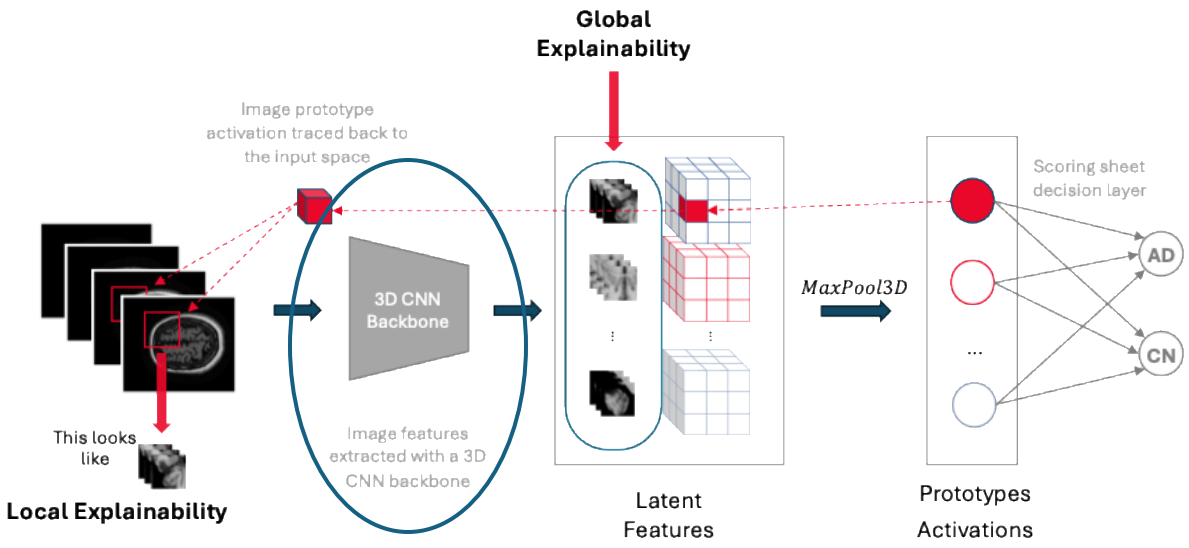
Handle **Out-of-Distribution** data

Abstain from decisions when there's no relevant detected prototypes

Original PIPNet, doi:10.1109/CVPR52729.2023.00269

PIPNet3D

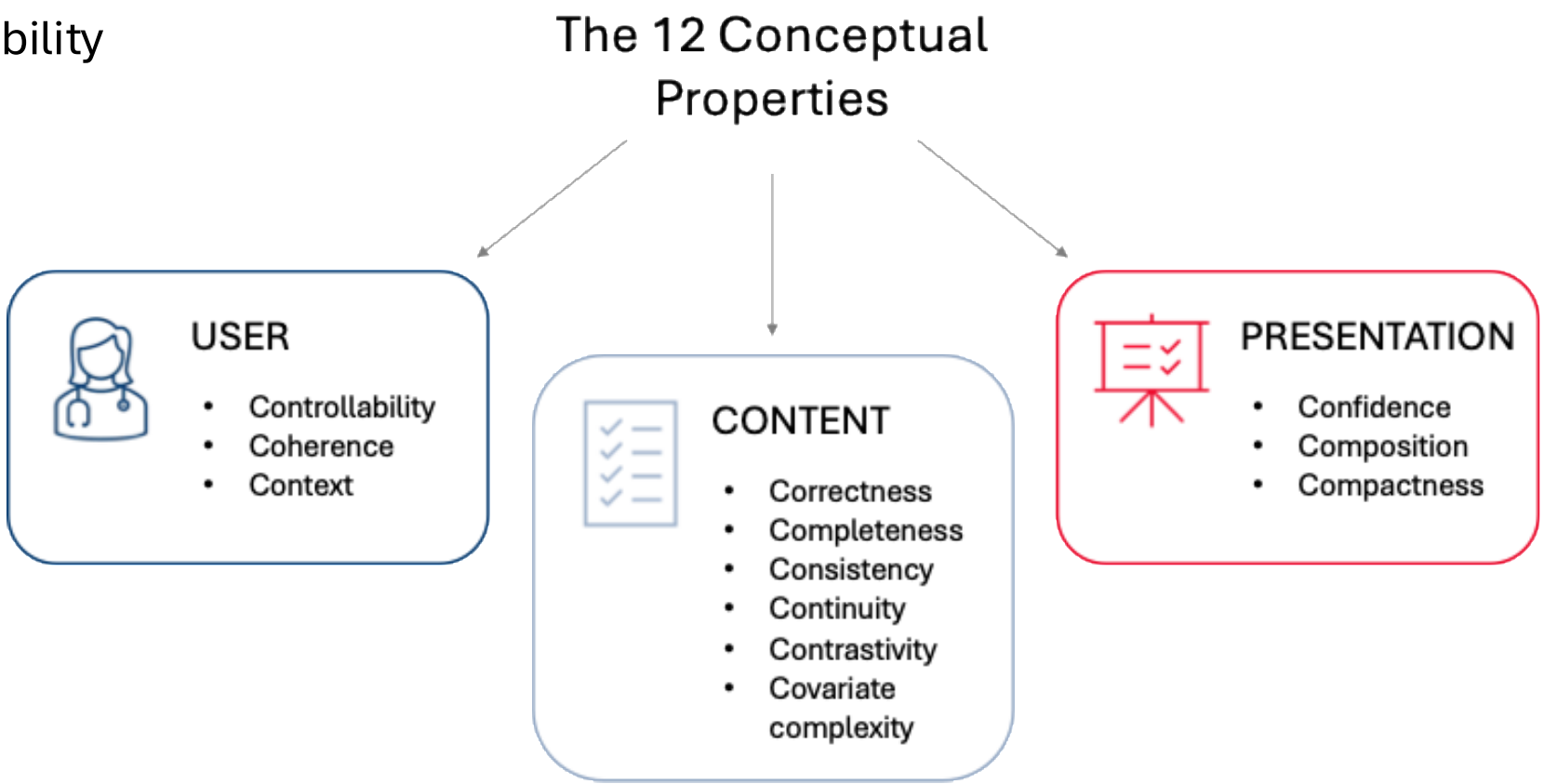
Data collection: *"ADNI1 Standardized Screening Data Collection for 1.5T"* sMRI from the Alzheimer's Disease Neuroimaging Initiative (ADNI): 307 Cognitively Normal (CN) and 243 Alzheimer's Disease (AD) different subjects



	Balanced Accuracy	Specificity	Sensitivity	F1
ResNet	80 ± 06	78 ± 15	82 ± 12	79 ± 07
PIPNet$_{RN}$	82 ± 02	88 ± 07	76 ± 09	79 ± 03
ConvNeXt	61 ± 07	67 ± 15	56 ± 24	54 ± 15
PIPNet$_{CN}$	69 ± 03	71 ± 07	68 ± 08	66 ± 04
After aligning with expert knowledge				
PIPNet$_{EK}$$^{\varnothing}$	**82** ± 02	88 ± 07	74 ± 12	78 ± 05
PIPNet$_{EK}$*	85	84	86	83

Prototype Evaluation

The Co-12 evaluation framework of explainability

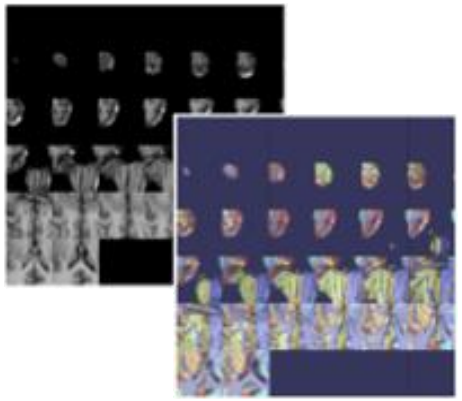


The Co-12: https://dl.acm.org/doi/10.1145/3583558

Prototype Evaluation

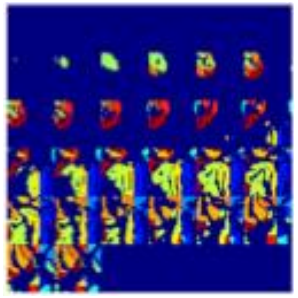
Functionally-Grounded Evaluation: Prototype Brain Entropy

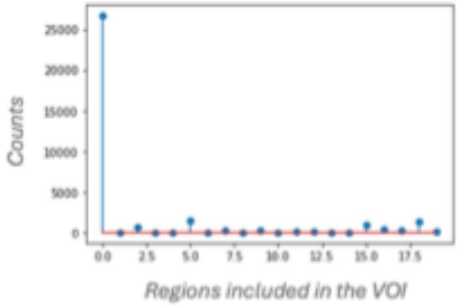
Assess purity in terms of brain regions included



Annotate the image prototype with CerebrA atlas

Select the corresponding CerebrA atlas VOI





Compute the VOI histogram

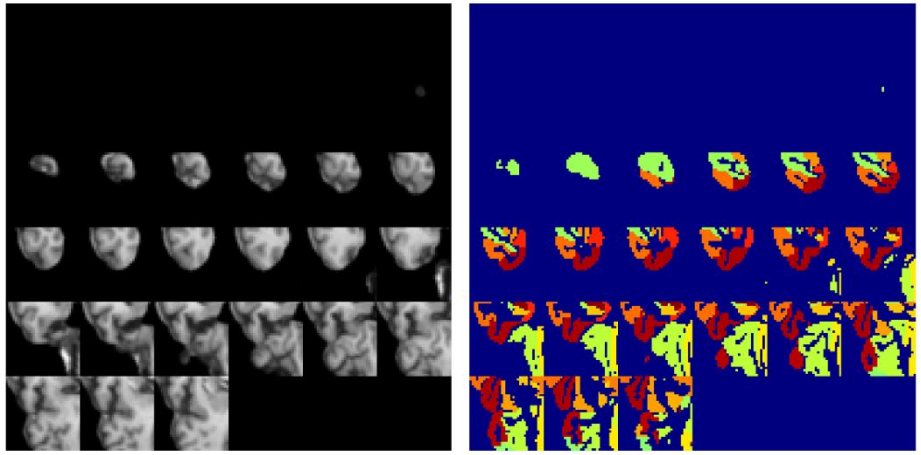
Compute Shannon Entropy H_p

Prototype Evaluation

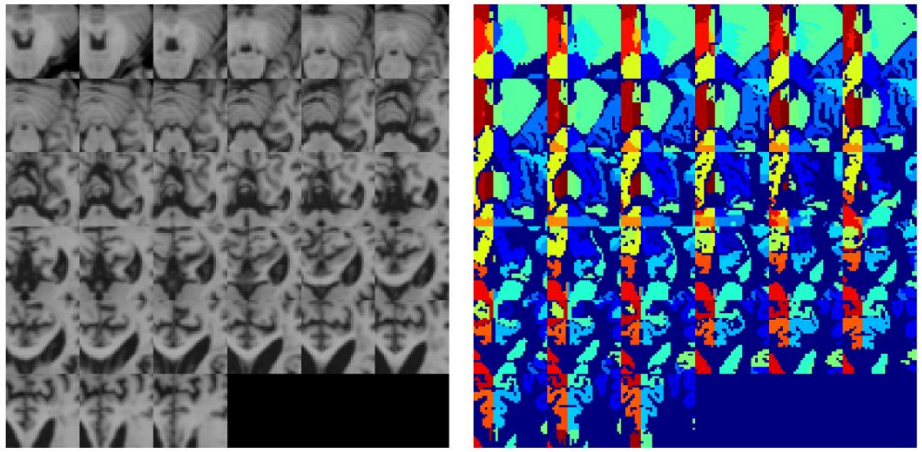
Functionally-Grounded Evaluation: Prototype Brain Entropy

Assess purity in terms of brain regions included

Prototype #1, low H_p



Prototype #2, high H_p



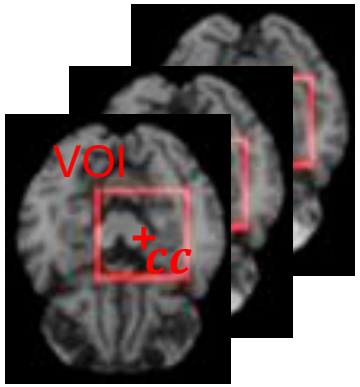
Prototype Evaluation

Functionally-Grounded Evaluation: Prototype Localization Consistency

Assess if the same prototype is detected in the same anatomical brain regions for different images

- Evaluate the part-prototype coordinate centre in every test image $VOI_{cc,p}|_{img}$
- Compute the average coordinate centre $\overline{VOI_{cc,p}}$
- Compute:

$$LC_p = \sum_{img} \frac{||VOI_{cc,p}|_{img} - \overline{VOI_{cc,p}}||}{l\sqrt{3}}$$



Prototype Evaluation

Functionally-Grounded Evaluation: Prototype Localization Consistency

Assess if the same prototype is detected in the same anatomical brain regions for different images

Prototype #1, localized into different images

Image #1

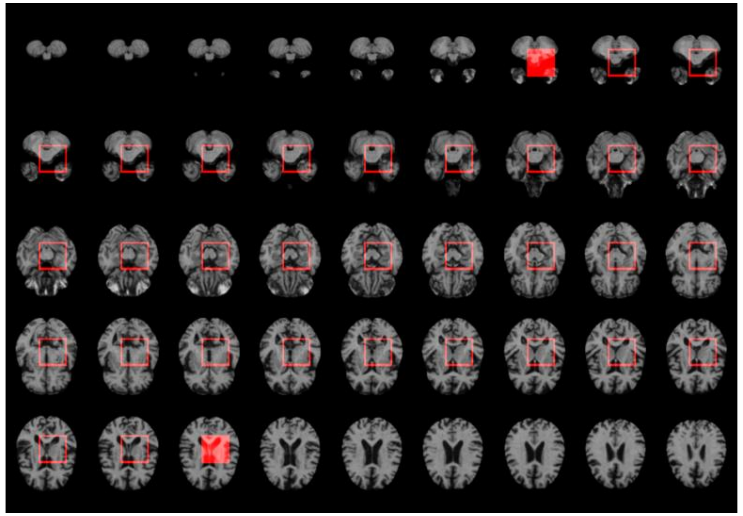
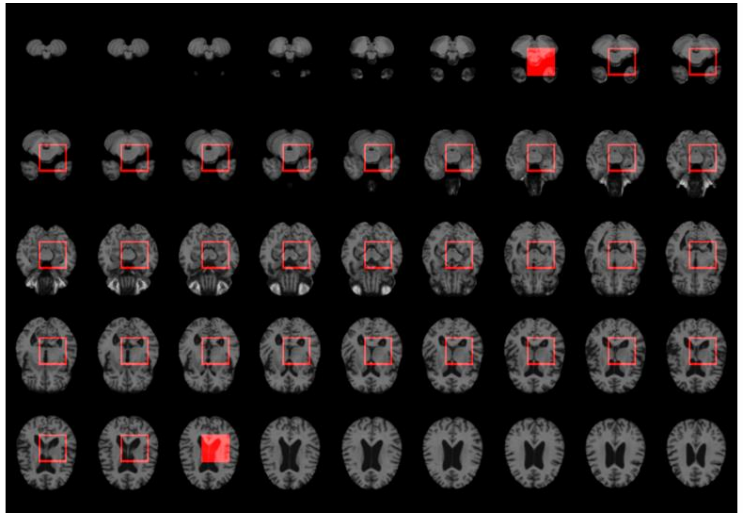


Image #2

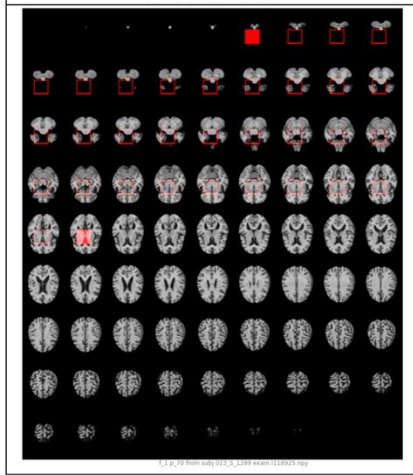
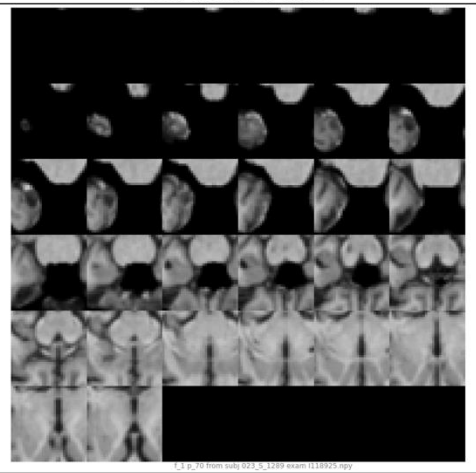


Prototype Evaluation

Expert Evaluation

Assess prototype coherency w.r.t. domain knowledge

All the prototype from the 5-fold ResNet18 was analysed by two radiologists from Marburg University Hospital using a survey with a 6 pts Likert scale

VOI Location	VOI detailed view	AI decision
		Alzheimer's Disease

1. This VOI is located in a clinically relevant region for diagnosing Alzheimer's Disease.
   Strongly disagree ☐ - ☐ - ☐ - ☐ - ☐ - ☐ Strongly agree

2. This VOI shows a brain pattern that exhibit pathologies
   Strongly disagree ☐ - ☐ - ☐ - ☐ - ☐ - ☐ Strongly agree

3. The decision of the AI model for this VOI is correct (NOT Cognitively Normal).
   Strongly disagree ☐ - ☐ - ☐ - ☐ - ☐ - ☐ Strongly agree

→ Localization Coherence

→ Pattern Coherence

→ Classification Coherence

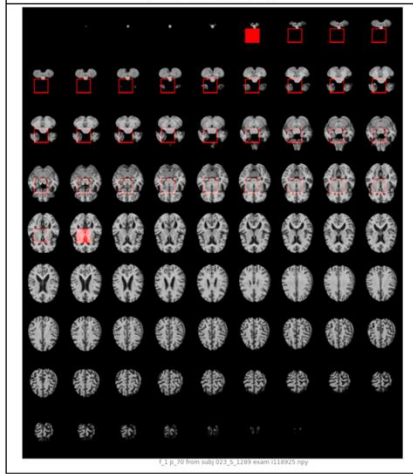
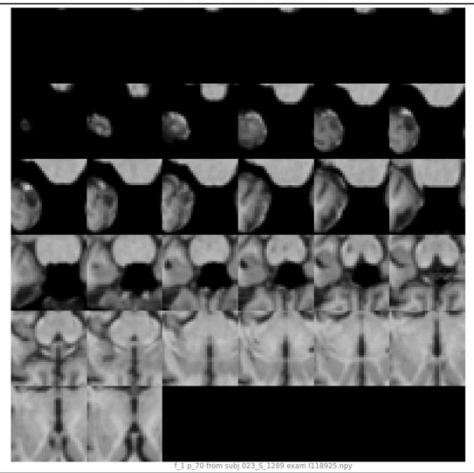
Prototype Evaluation

Expert Evaluation

Assess prototype coherency w.r.t. domain knowledge

All the prototype from the 5-fold ResNet18 was analysed by two radiologists from Marburg University Hospital using a survey with a 6 pts Likert scale

- Assess the inter-user agreement with the Interclass Correlation Coefficient (ICC)
- For every prototype, compute the average Localization, Pattern and Classification Coherency
- Consider as coherent a score > 3.5

VOI Location	VOI detailed view	AI decision
		Alzheimer's Disease

1. This VOI is located in a clinically relevant region for diagnosing Alzheimer's Disease.
   Strongly disagree ☐ - ☐ - ☐ - ☐ - ☐ - ☐ Strongly agree
2. This VOI shows a brain pattern that exhibit pathologies
   Strongly disagree ☐ - ☐ - ☐ - ☐ - ☐ - ☐ Strongly agree
3. The decision of the AI model for this VOI is correct (NOT Cognitively Normal).
   Strongly disagree ☐ - ☐ - ☐ - ☐ - ☐ - ☐ Strongly agree

ICC: 1) 0.80; 2) 0.76; 3) 0.85

Prototype Evaluation

PIPNet3D under the Co-12 framework

	Criterion	
Content	Correctness	Classification correct by design. Visualization with upsampling Chen et al. (2019) might not be fully correct Gautam et al. (2023), and is direction of future work.
	Completeness	Complete by design; the full model behavior (linear layer connections is explained). Representation learning (backbone) as blackbox with interpretable output (prototypes).
	Consistency	Consistent by design (no random components in the forward pass computations), but may be subject to standard numerical computation differences. We assessed the difference in prototype's localization when detected in different subjects (cf. LC_p).
	Continuity	Optimized by the contrastive learning setup of PIPNet, but not explicitly evaluated.
	Contrastivity	Output-sensitivity implemented by-design.
	Cov. complexity	Prototype purity assessed with anatomical brain atlas annotation (cf. H_p).
Presentation	Compactness	Evaluated by *global* and *local explanation size*.
	Composition	Showing both, overview and detail; localization of the found prototype in the brain, and its detail view.
	Confidence	Diagnosis prediction reported together with the prototypes' similarity score which constitutes the Alzheimer's fingerprint of the subject and the OoD detection confidence.
User	Context	We evaluated prototypes with domain experts in a human-grounded evaluation.
	Coherence	User study with radiologists evaluating whether the prototypes align with experts' visual assessment w.r.t. their localization, detected pattern, and diagnostic decision (cf. *Localization Coherence, Pattern Coherence, Classification Coherence*).
	Controllability	We use results collected from the coherence analysis to suppress the prototypes that are not in line with the expert evaluation.

CONTENT

- Correctness
- Completeness
- Consistency
- Continuity
- Contrastivity
- Covariate complexity

PRESENTATION

- Confidence
- Composition
- Compactness

USER

- Controllability
- Coherence
- Context

Prototype Evaluation

		M1	M2	M3	M4	M5	Average
Functional Evaluation	Global size ↓	10	11	5	5	4	7.0
	for CN	5	4	3	3	3	3.6
	for AD	5	7	2	2	1	3.4
	Local size ↓	5.4	5.2	2.7	3.0	2.4	3.8
	Sparsity ↑	0.990	0.989	0.995	0.995	0.996	0.993
	LC_p ↓	0.004	0.021	0.008	0.018	0.030	0.016
	for CN	0.009	0.003	0.000	0.015	0.030	0.017
	for AD	0.000	0.016	0.020	0.022	0.050	0.022
	H_p ↓	2.5	3.1	3.4	3.1	3.4	3.1
	for CN	2.8	3.3	3.5	3.1	2.9	3.1
	for AD	2.3	2.9	3.3	3.1	3.8	3.1
Users	Localization Coherence ↑	0.90	0.60	0.60	0.80	0.50	0.70 (3.5)
	for CN	0.80	0.25	0.67	0.67	0.33	0.54
	for AD	1.00	0.86	0.50	1.00	1.00	0.87
	Pattern Coherence ↑	1.00	0.90	0.80	0.80	0.80	0.90 (4.5)
	for CN	1.00	1.00	1.00	1.00	0.67	0.93
	for AD	1.00	0.86	0.50	0.50	1.00	0.77
	Classification Coherence ↑	1.00	0.90	0.80	1.00	0.80	0.90 (4.5)
	for CN	1.00	1.00	1.00	1.00	0.67	0.93
	for AD	1.00	0.86	0.50	1.00	1.00	0.87

Reduced n° of class-specific prototypes

Part-prototypes consistently located in the same brain regions (small LC_p)

Lower Localization Coherence for CN prototypes but with a Coherent Pattern

Worse AD Pattern Coherence for fold with lower AD Recall

	Precision		Recall	
	CN	AD	CN	AD
M1	0.88	0.81	0.84	0.86
M2	0.85	0.78	0.82	0.82
M3	0.81	0.84	0.89	0.73
M4	0.77	1	1	0.62
M5	0.81	0.78	0.84	0.75

Prototype Evaluation

USERS	Localization Coherence ↑	0.90	0.60	0.60	0.80	0.50	0.70 (3.5)
	for CN	0.80	0.25	0.67	0.67	0.33	0.54
	for AD	1.00	0.86	0.50	1.00	1.00	0.87
	Pattern Coherence ↑	1.00	0.90	0.80	0.80	0.80	0.90 (4.5)
	for CN	1.00	1.00	1.00	1.00	0.67	0.93
	for AD	1.00	0.86	0.50	0.50	1.00	0.77
	Classification Coherence ↑	1.00	0.90	0.80	1.00	0.80	0.90 (4.5)
	for CN	1.00	1.00	1.00	1.00	0.67	0.93
	for AD	1.00	0.86	0.50	1.00	1.00	0.87

Removing clinically irrelevant prototypes improved the model's compactness without impacting performances

	Balanced Accuracy	Specificity	Sensitivity	F1
ResNet	80 ± 06	78 ± 15	82 ± 12	79 ± 07
PIPNet_{RN}	82 ± 02	88 ± 07	76 ± 09	79 ± 03
ConvNeXt	61 ± 07	67 ± 15	56 ± 24	54 ± 15
PIPNet_{CN}	69 ± 03	71 ± 07	68 ± 08	66 ± 04
AFTER ALIGNING WITH EXPERT KNOWLEDGE				
$\text{PIPNet}_{EK}{}^{\varnothing}$	**82** ± 02	88 ± 07	74 ± 12	78 ± 05
$\text{PIPNet}_{EK}{}^{*}$	85	84	86	83

Conclusions

PIPNet3D is an interpretable part-prototype 3D classifier

PIPNet3D **performs equally** well to its corresponding **black-box** baseline to AD diagnosis from sMRI with a **reduced number** of part-prototype

We proposed two **novel metrics** for **functionally grounded evaluations**:

- ➢ *Prototype Brain Entropy* and *Prototype Localization Consistency*

We proposed a **domain-experts** generalizable **evaluation setup** for Coherency *(Localization, Pattern, Classification)* and observed:

i) PIPNet3D **aligns well** with domain knowledge

ii) Removing **clinically irrelevant prototypes** improved the model's compactness **without impacting performances**

Future Work:

- Include **intermediate** level of cognitive impairment (E-MCI, MCI, L-MCI)
- Integrate clinical information, e.g. Patient's **Age**
  - ➢ PIMPNet

PIMPNet

MICCAI 2024 Marrakesh MOROCCO

Thank You!
Questions?

Eng. Lisa Anita De Santi, PhD Student,
Department of Information Engineering, University of Pisa
Fondazione G.Monasterio – Bioengineering Unit

lisa.desanti@phd.unipi.it
desanti@monasterio.it

PIPNet3D GitHub

Fondazione Monasterio la ricerca che cura

Università di Pisa

Dipartimento di Ingegneria dell'Informazione