

Explainable artificial intelligence for automated amyloid-PET positivity classification

Sophie A. Martin¹, Lyduine Collij², Leonard Pieperhoff³, Ariane Bollack^{1,4}, Alle Meije Wink³, David Vallez-García^{5,6}, Frederik Barkhof^{1,3}, James H. Cole¹, on behalf of the AMYPAD consortium

¹UCL, UK
²Lund University, Sweden
³Amsterdam UMC, Netherlands

⁴GE Healthcare
⁵Barcelona-β Research Centre, Spain
⁶CIBER-BBN, Spain

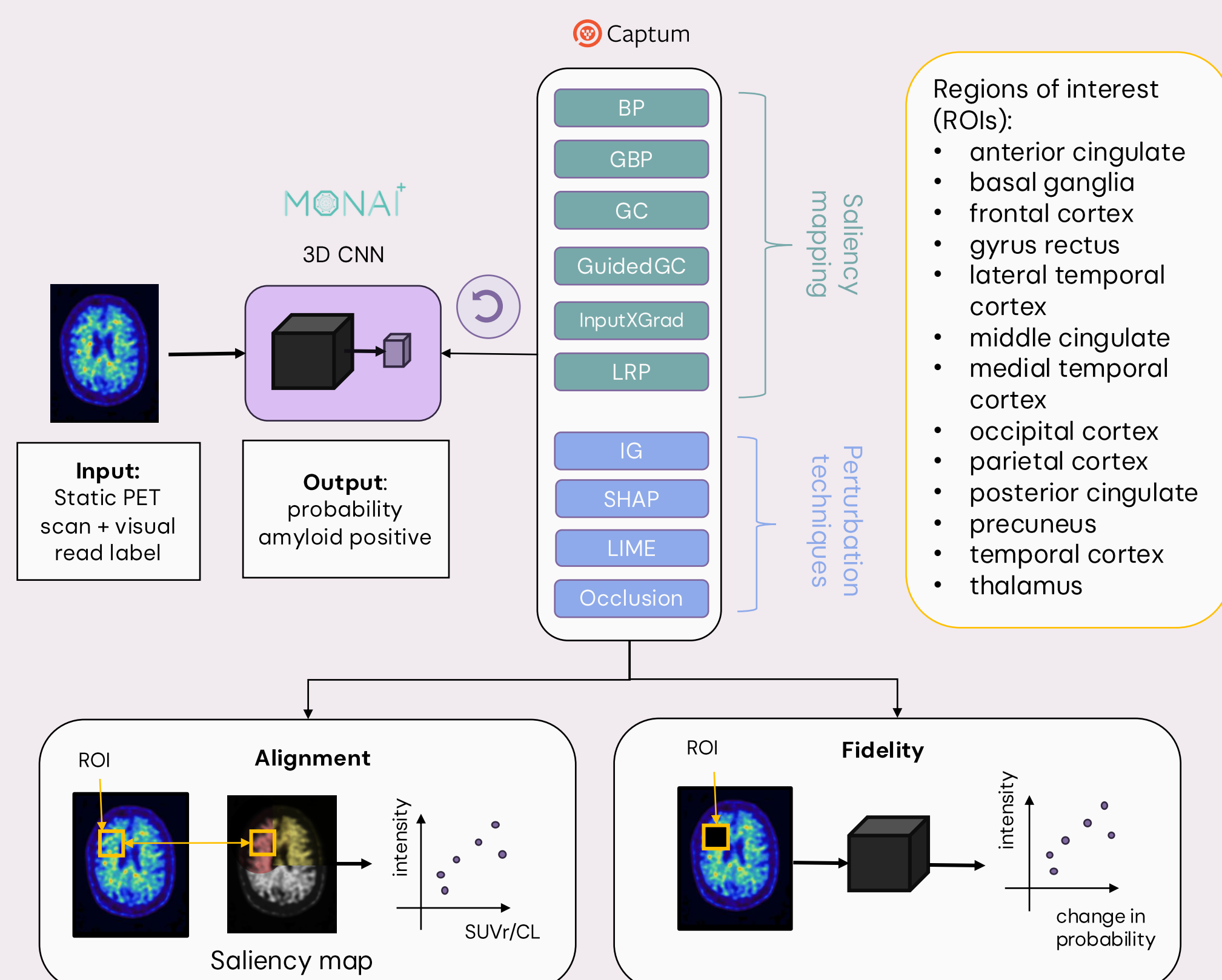
Motivation

- Positron emission tomography (PET) imaging can be used to detect amyloid plaques and support early diagnosis of Alzheimer's disease (AD).
- Artificial intelligence (AI)-based PET classification or quantification [1],[2] can address inter-rater subjectivity and facilitate rapid screening.
- Existing AI models lack interpretability.
- Explainable artificial intelligence (XAI) methods can help to interpret model outputs but produce variable results [3].
- Validating methods is challenging without a 'ground truth' reference and is often limited group-wise, structural imaging analyses [4].
- There remains a need for assessing the relevance of salient regions on a subject-level.

Aims

We trained a deep learning model to predict amyloid positivity and validated XAI-generated saliency maps by:

- Assessing their agreement with subject-level regional distributions of amyloid deposition (alignment)
- Assessing the effect of removing salient regions on model probability (fidelity)
- Exploring whether subject-level heatmaps can be used to explain unequivocal cases



When predicting amyloid positivity, what information does a deep learning model use and which XAI method best captures its reasoning?

Methods

- 1316 baseline static amyloid PET scans from the AMYPAD Prognostic and Natural History Study (PNHS).
- Visual reads were performed by a panel of radiologists [5].
- All participants had a Clinical Dementia Rating score ≤ 1 .
- A 3D convolutional neural network (CNN) was trained to classify between visually amyloid positive and negative individuals.
- A random sample of 80% of the scans were used for model training and the remaining were used as a test set.

Feature	Sample characteristics		
	VR	n	mean $\pm$ std./%
Age	All	1316	67.8 $\pm$ 8.84
	Positive	273	71.7 $\pm$ 8.54
	Negative	1043	66.8 $\pm$ 8.63
Tracer (% FMT)	All	1316	56.0
	Positive	273	59.7
	Negative	1046	55.0
SUV _R _{gca}	All	1207	1.12 $\pm$ 0.21
	Positive	252	1.44 $\pm$ 0.24
	Negative	955	1.05 $\pm$ 0.09
CL _{gca}	All	1207	19.4 $\pm$ 27.9
	Positive	252	60.2 $\pm$ 32.6
	Negative	955	8.63 $\pm$ 12.1

Table 1. AMYPAD sample characteristics. For each feature, the number of participants n is given after participants with missing data were removed. Global cortical average SUVR and CL values are based on the LEAP [6] atlas using the whole cerebellum as a reference region. Abbreviations: VR = visual read, CDR = Clinical Dementia Rating, FMT = flutemetamol, SUVR = standardised uptake value ratio, CL = Centiloid, gca = global cortical average, std. = standard deviation

- The trained model achieved a test accuracy of 93.2% (89.1% balanced), with 87.5% precision and 81.7% recall.
- Per-tracer test accuracies were 96.7 and 98.4 for FBB and FMT respectively.

Test predictions			Tracer: florbetaben			Tracer: flutemetamol		
True label	0	1	0	1	0	1	1	
	196	7		81		2		115
1	11	49	5	17	6	32	32	
Predicted label			Predicted label			Predicted label		

Results

- Model explanations were generated using 10 XAI methods.
- Pearson's r was used to quantify the correlation between mean saliency and quantification measures in 13 pre-defined ROIs across all test subjects.
- For fidelity, mean saliency was correlated with the change in model probability when that region was occluded from the input PET volume.

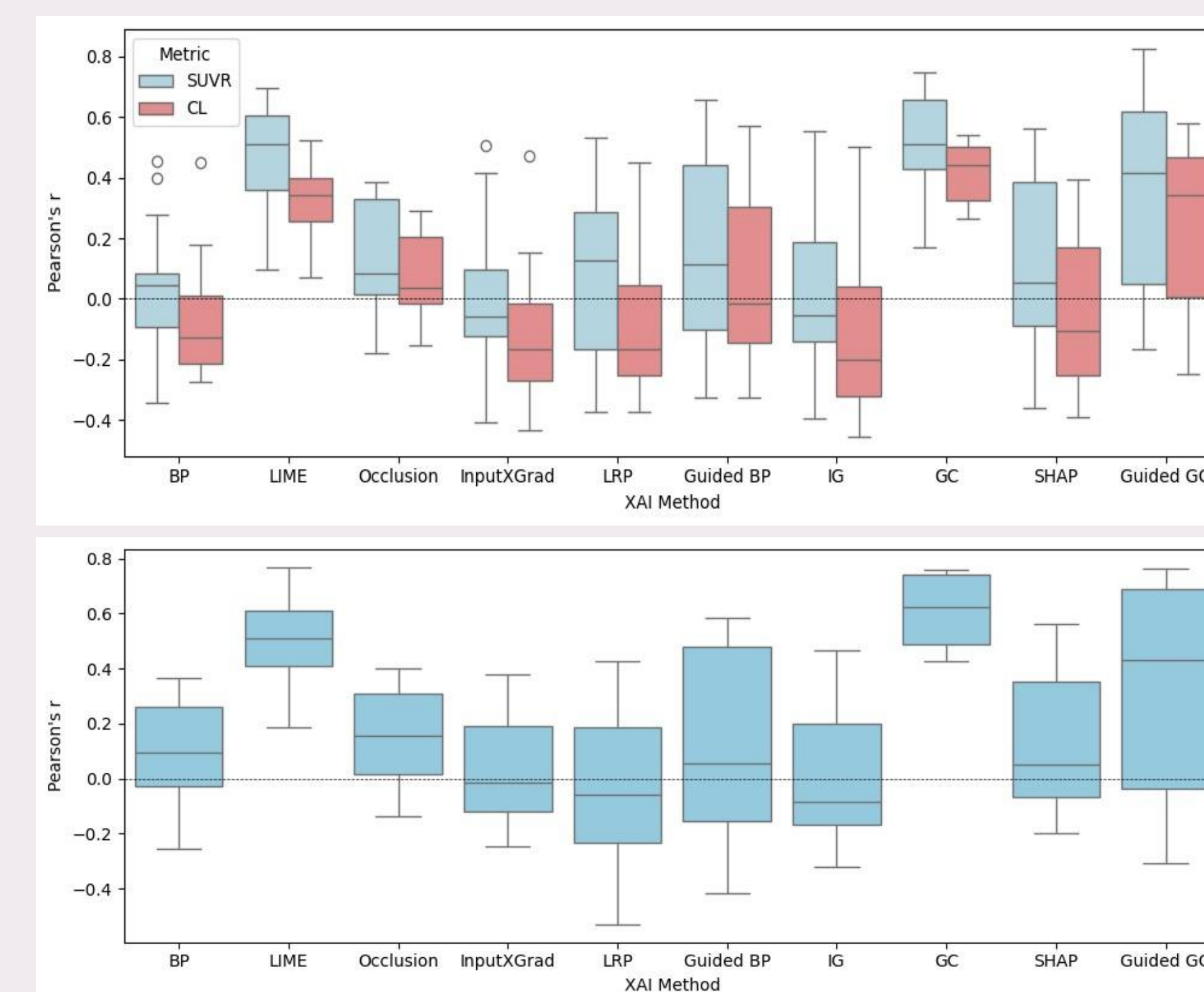


Figure 1. Box plots summarising the Pearson's r correlations for (top) alignment and (bottom) fidelity across 13 LEAP-defined ROIs.

Mean LIME and GradCAM saliences produced the highest correlations with SUVR, CL and change in model probability across the 13 ROIs.

- Subject-level explanations reveal distinct patterns of the regions that were most salient to the model's prediction.

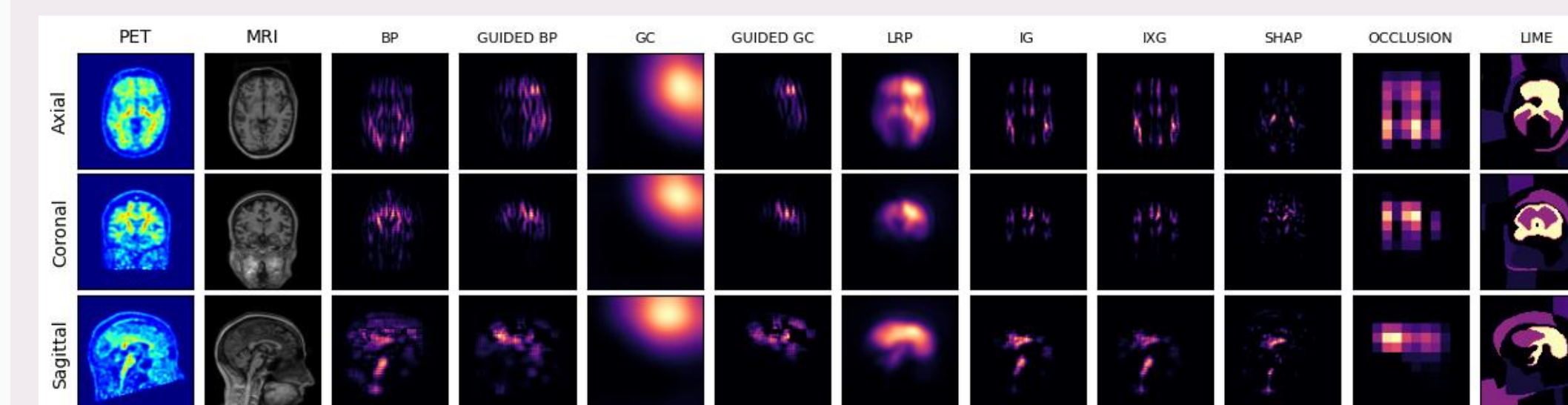


Figure 2. Mid-stack slices from a correctly classified test subject (visual read positive). Left to right: PET scan, MRI scan, saliency maps from 10 XAI methods.

References

- Fan et al., *Radiology: AI*, 2024.
- Kang et al., *Journal of Nuclear Medicine*, 2023.
- Martin et al., *Human Brain Mapping*, 2026.
- Wang et al., *NeuroImage*, 2023.
- Collij et al., *Frontiers in Neurology*, 2023.
- Wolz et al., *NeuroImage*, 2010.

Working on explainable artificial intelligence and/or amyloid PET classification for dementia applications?

Let's talk! sophie.a.martin@ucl.ac.uk