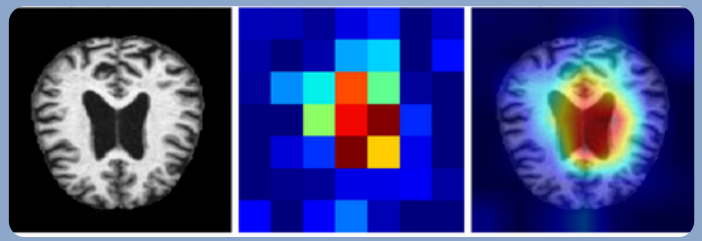
Can You Trust the Heatmap?

Validating Grad-CAM explanations in Alzheimer's MRI without pixel-level ground truth

Ing. Alexandra Čižmárová, Supervisor: **prof. Ing. Elena Zaitseva, PhD.,** Faculty of Management Science and Informatics, University of Žilina
Thesis: Analysis of Medical Data by XAI with a Grad-CAM method

The problem

- Deep networks reach high-level accuracy on medical images but stay **black boxes.**
- Grad-CAM heatmaps look persuasive and are routinely accepted without verification [2, 4].
- **No pixel-level annotations exist**, so a heatmap cannot be checked against a correct region.



Why Grad-CAM?

- **Post-hoc XAI method** - works on the trained network, no retraining or architecture change.
- **Needs only image-level labels** - the only supervision this dataset provides.
- **One backward pass per image**, so it scales to all 1,280 test images.
- Ruled out: LIME/SHAP need thousands of passes per image and segmentation needs masks we don't have.
- **Grad-CAM [1] is the field's default for CNN explanation** - which is precisely why its reliability deserves testing.

Model & data

- **ResNet50 with transfer learning**, 224×224×3 input, four diagnostic classes.
- Augmented Alzheimer MRI set (Kaggle); **60/20/20 train, validation, test split** - augmented images were only used for training.
- Both validation axes were run on all 1,280 held-out test images - no hand-picked examples.

Why it matters

AI is entering everyday radiology, Alzheimer's screening among the first uses, and EU rules require it to explain itself. **This protocol runs behind such a system and flags the scans where the explanation does not hold up.**

99.30%	2 axes	0
classifier accuracy - yet its explanations still fail	plausibility and faithfulness, applied together	pixel-level annotations required

A heatmap has no right answer so this protocol checks it against anatomy, and against the model itself.

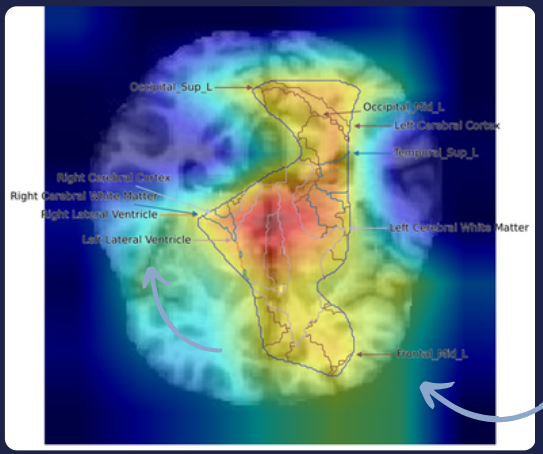
A two-axis framework for validating explanations without ground truth

Anatomical plausibility (qualitative screen)

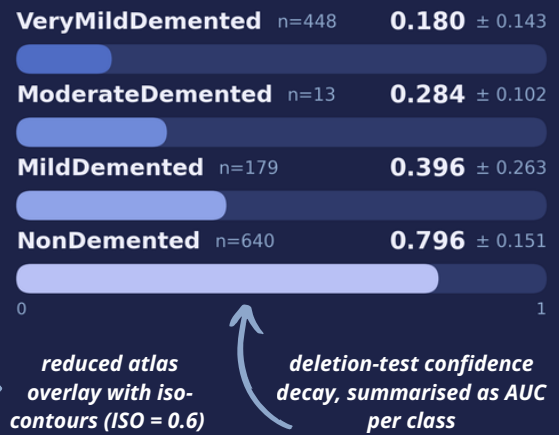
Does it look where it should?

- **Harvard–Oxford and AAL atlases [5] registered** to each MRI slice.
- Registration by **mutual information plus Canny edge overlap**.
- Heatmap **reduced to iso-contours**, then intersected with ranked named regions per image, read against structures implicated in AD.

Deliberately qualitative - quantifying plausibility requires the pixel-level ground truth that this work assumes is absent.



reduced atlas overlay with iso-contours (ISO = 0.6)



deletion-test confidence decay, summarised as AUC per class

Faithfulness (quantitative test)

Does it drive the prediction?

- **Top-ranked** 8×8 patches **deleted progressively** over 30 steps [3].
- Removed regions replaced by a blurred baseline.
- Drop in model **confidence summarised as an AUC**.

Lower AUC = the highlighted region genuinely mattered.

What the framework revealed

- **Explanation reliability** is not a fixed property of Grad-CAM - it **varies by class and by individual image**.
- Faithfulness did not track disease severity: VeryMildDemented explanations scored most faithful, NonDemented least.
- Recurring failure modes: **activation outside brain tissue and diffuse or near-empty heatmaps**: ModerateDemented produced an essentially blank map despite a confident prediction.
- The protocol needs no new annotation effort, so it **transfers to any weakly supervised imaging task**.

Conclusion

Plausibility and faithfulness together form **a practical, annotation-free protocol for auditing explanations in medical imaging**. Across 1,280 MRI scans it shows Grad-CAM's reliability is not constant - persuasiveness is not correctness.

Parts of this work are published in: Advanced Information Systems 10(2):79–86, 2026 (Scopus); Innovative Technologies and Scientific Solutions for Industries 1(35):17–27, 2026 (Scopus).
References: [1] Selvaraju et al., ICCV 2017; [2] Adebayo et al., NeurIPS 2018; [3] Petsiuk et al., BMVC 2018; [4] Arun et al., Radiology: AI 2021; [5] Tzourio-Mazoyer et al., NeuroImage 2002

ŽILINSKÁ UNIVERZITA V ŽILINE
Fakulta riadenia a informatiky