PAKDD 2025
Sydney 10-13 June

Unlocking Neural Transparency: Jacobian Maps for Explainable AI in Alzheimer's Detection

Yasmine Mustafa[1], Mohamed Elmahallawy[2], Thomas Tie Luo[3]

[1] Missouri University of Science and Technology

[2] Washington State University

[3] University of Kentucky

University of Kentucky

Presenter: Thomas Tie Luo

Introduction

Alzheimer’s Disease

- AD is a leading cause of dementia with rising global burden
- Progresses from Mild Cognitive Impairment (MCI) to severe functional loss
- Early detection is key to intervention

EARLY STAGE:

- Trouble remembering events
- Difficulty recalling names
- Frequently loses personal items

MIDDLE STAGE:

- Worsening memory loss
- Confusion about names and relationships
- Difficulty with daily tasks

LATE STAGE:

- Difficulty recognizing family members
- Wheelchair dependence
- Trouble eating and loss of bowl/bladder control
- Limited vocabulary & comprehension

Neuroimaging + machine learning have shown promise

Introduction

Role of Explainable Artificial Intelligence (XAI)

- Black-box nature of AI models causes skepticism in clinics
- Trust and adoption require transparent decision-making
- XAI provides insights into **why/how** a prediction is made

XAI categories:

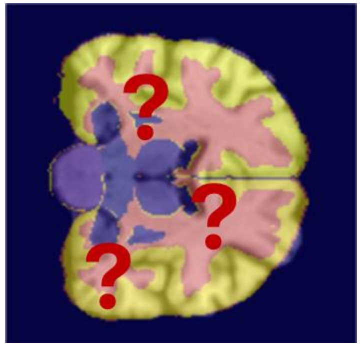
Pre-model / ante-hoc:	In-model / Intrinsic:	Post-model / post-hoc:
• Data or feature engineering before training the model (e.g., identifying key brain biomarkers).	• Incorporate model design or training mechanism into model itself	• Explaining predictions after the model runs (e.g., Grad-CAM).

Introduction

Challenges in Medical XAI

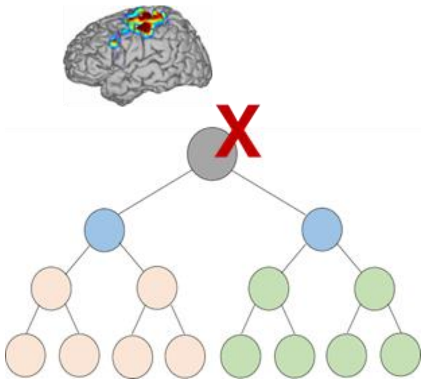
Post-hoc approaches

- Lack of ground truth to validate the explanation
- Lack of metrics for pathology alignment
- Post-hoc heatmaps (e.g., Grad-CAM) have limited reliability in brain scans (works better for natural images)



Intrinsic approaches

- While inherently interpretable (e.g., linear models or decision trees), they
- **Struggle to capture complex** patterns present in high-dimensional medical data



Our Method

Ante-hoc XAI with Jacobian Maps

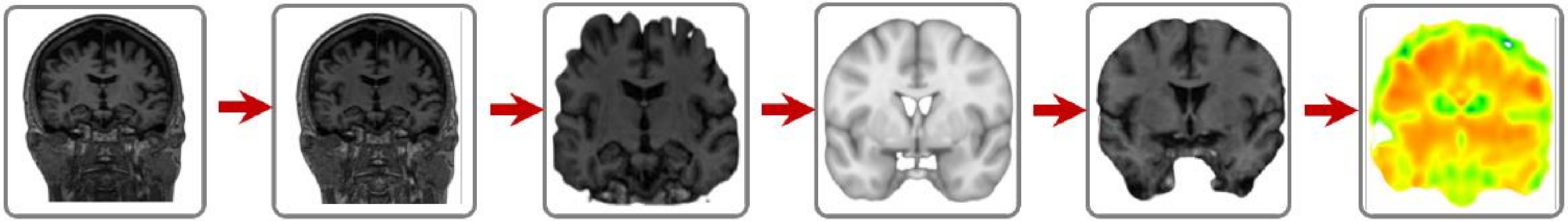
- Introducing Jacobian Maps (JMs) as an ante-hoc explainability tool for AD detection.

How it works (overview):

- Compute **Jacobian determinants**, which measure how much each **voxel** (3D pixel in a brain scan) **expands or shrinks** compared to a healthy brain.
  - This creates a subject-specific map of brain structural changes.
  - This map serves as a kind of **ground truth** that highlights the locations of brain changes.
- and we apply it **before** training medical AI models.

Method (details)

Transforming Brain Images into Jacobian Maps (JM)



(a) Raw (b) Bias field correction. (c) Brain extraction. (d) Register to MNI152. (e) Registered. (f) Jacobian map.

- Oasis dataset
- 3D images

OASIS
OPEN ACCESS SERIES OF IMAGING STUDIES

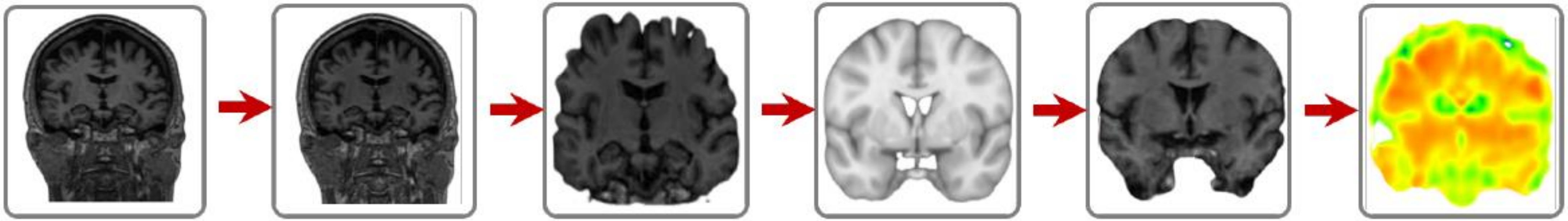
- Participants include **755 cognitively normal (CN)** adults and **622 patients at various stages of cognitive decline** with age between 42-95 yrs.
- Based on clinical dementia rating (CDR) scores:

CDR	Class
0	Normal
0.5	MCI
1	Mild
2	Moderate
3	Severe

Combined

Method (details)

Transforming Brain Images into Jacobian Maps (JM)



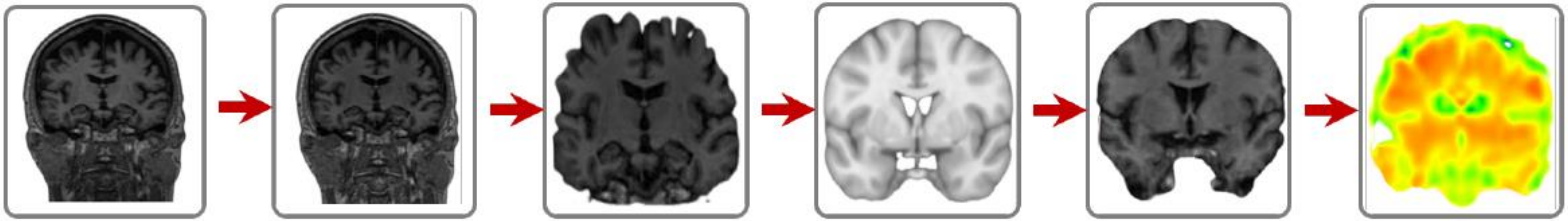
(a) Raw (b) Bias field correction. (c) Brain extraction. (d) Register to MNI152. (e) Registered. (f) Jacobian map.

(b) Bias Field Correction

- Corrects non-uniform intensity caused by magnetic field inhomogeneities in the scanner.
- Ensures that tissue intensity is consistent across the brain.
- Tool: FLIRT (FMRIB's Linear Image Registration Tool)

Method (details)

Transforming Brain Images into Jacobian Maps (JM)



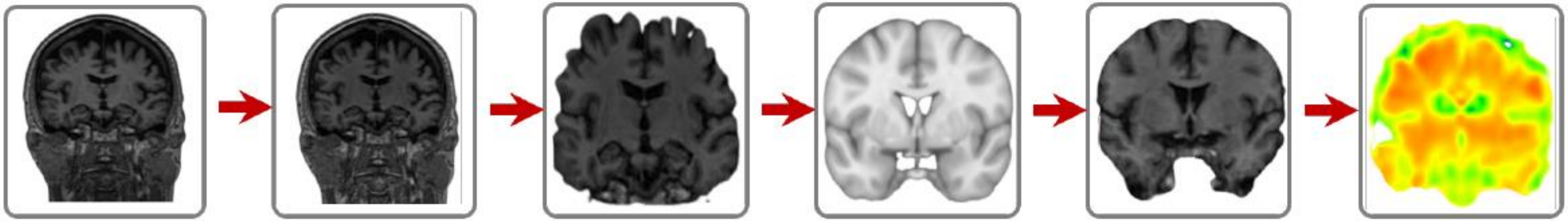
(a) Raw (b) Bias field correction. (c) Brain extraction. (d) Register to MNI152. (e) Registered. (f) Jacobian map.

(c) Brain Extraction

- Removes non-brain tissues (skull, skin, etc.) to isolate the brain.
- Reduces irrelevant variability and computation.
- Tool: BET (Brain Extraction Tool)

Method (details)

Transforming Brain Images into Jacobian Maps (JM)



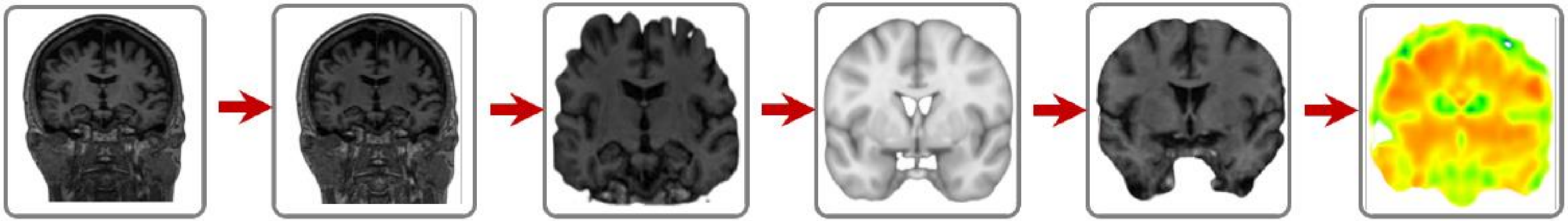
(a) Raw (b) Bias field correction. (c) Brain extraction. (d) Register to MNI152. (e) Registered. (f) Jacobian map.

(d) Registration to Template (MNI152)

- Align each brain to a common anatomical space to allow voxel-wise comparison.
- Uses **non-linear image registration** via Symmetric Normalization (SyN).
- Tool: ANTs (Advanced Normalization Tools)

Method (details)

Transforming Brain Images into Jacobian Maps (JM)



(a) Raw (b) Bias field correction. (c) Brain extraction. (d) Register to MNI152. (e) Registered. (f) Jacobian map.

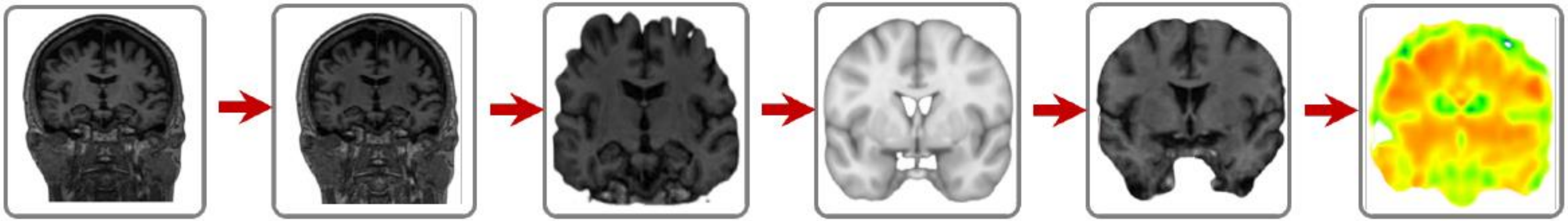
(e) Compute Deformation Vector Field

- After registration to MNI152, calculate how much each voxel has moved from the original brain.
- This deformation is expressed as a vector field:

$$v(x, y, z) = \phi(x, y, z) - (x, y, z)$$

where ϕ is the transformation function.

Transforming Brain Images into Jacobian Maps (JM)



(a) Raw (b) Bias field correction. (c) Brain extraction. (d) Register to MNI152. (e) Registered. (f) Jacobian map.

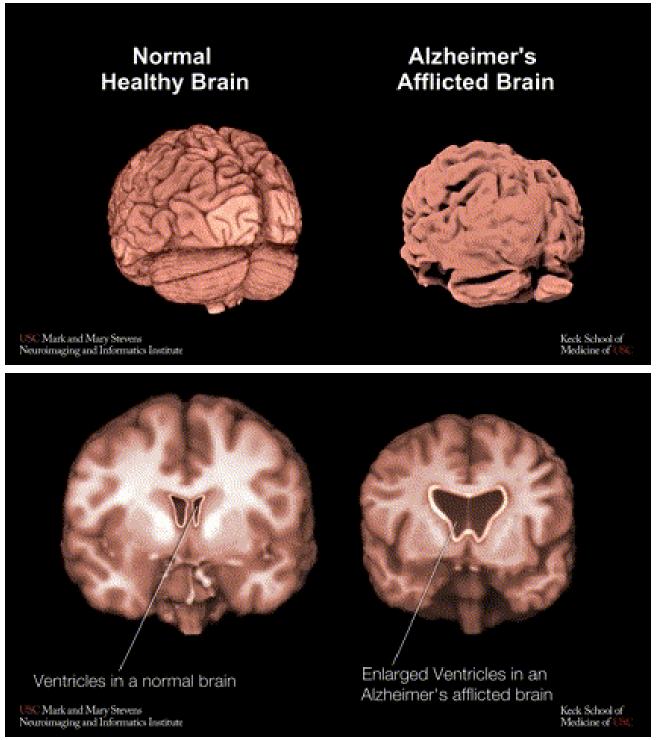
(f) Compute Jacobian Determinant

- A **Jacobian matrix** Is composed of the gradients of each deformation field $v(.)$, and it captures the stretching, compression, and shearing of the voxel.
- We compute the **determinant** of this matrix.
- Doing this for all the voxels result in the **Jacobian map.**

Method (details)

Computing Jacobian Maps

Brain Volume Changes Denoting Dementia



Deformation Field

$$v(x,y,z) = \phi(x,y,z) - (x,y,z)$$

Jacobian Matrix (J):

$$J(v) = \begin{pmatrix} \frac{\partial v_x}{\partial x} & \frac{\partial v_x}{\partial y} & \frac{\partial v_x}{\partial z} \\ \frac{\partial v_y}{\partial x} & \frac{\partial v_y}{\partial y} & \frac{\partial v_y}{\partial z} \\ \frac{\partial v_z}{\partial x} & \frac{\partial v_z}{\partial y} & \frac{\partial v_z}{\partial z} \end{pmatrix}$$

Jacobian Determinant:

$$\mathrm{JM}(x,y,z) = \det(J(v(x,y,z)))$$

Semantics:

Determinant of each voxel:
>1 → local **expansion**
=1 → no brain change
<1 → local **compression**

Jacobian Map

- Captures **subtle brain volume changes**
- Highlights **local brain morphometry**
- Provides informative representations for feature learning

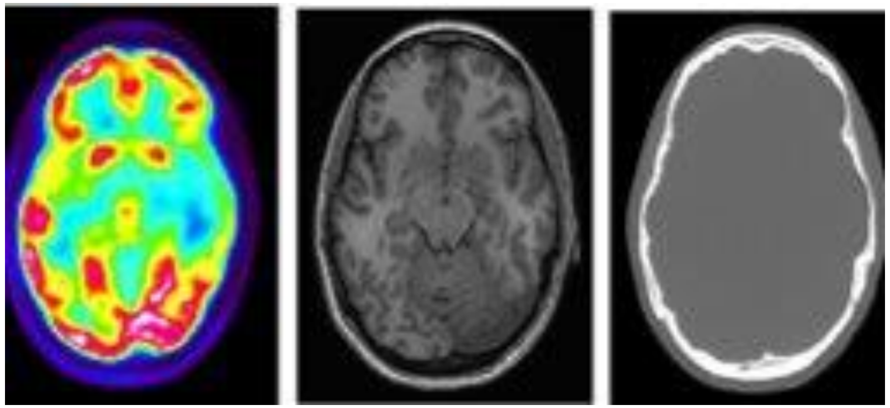
$$\begin{bmatrix} & \vdots & \\ \ldots & Det(J(v(x,y,z))) & \ldots \\ & \vdots & \end{bmatrix}_{\substack{x=1\ldots W \\ y=1\ldots H \\ z=1\ldots D}}$$

Skip if running short on time

Method

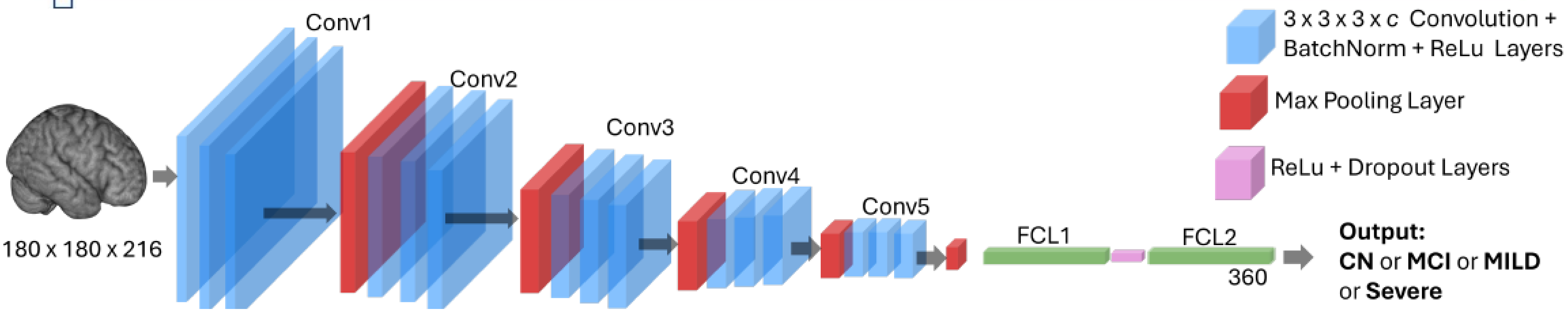
Key Advantages of JM for Ante-hoc XAI

- **No segmentation needed** → avoids label noise and complexity
- **Whole-brain coverage** → no patch-based sampling
- **Clinically intuitive** → visualizes structural atrophy directly
- **Quantitative** → preserves local volume change metrics
- **Generalizable** → works across MRI, PET, or CT if deformation fields are computed


PET MRI CT

Experiments

Model Architecture



- **Input: 3D Jacobian Maps** (or **standard registered MRI images** for comparison).
- Five 3D conv layers, each with a kernel of size 3×3×3.
- After each conv layer:
  - Batch Normalization: Stabilizes learning and accelerates convergence.
  - ReLU Activation: Introduces non-linearity.
  - Max-Pooling (at selected layers): Downsamples the spatial resolution.
- Output is flattened and passed through two Fully Connected (FC) layers.
  - The first FC layer uses Dropout for regularization.
  - The second FC layer outputs logits, normalized with Softmax to get class probabilities.

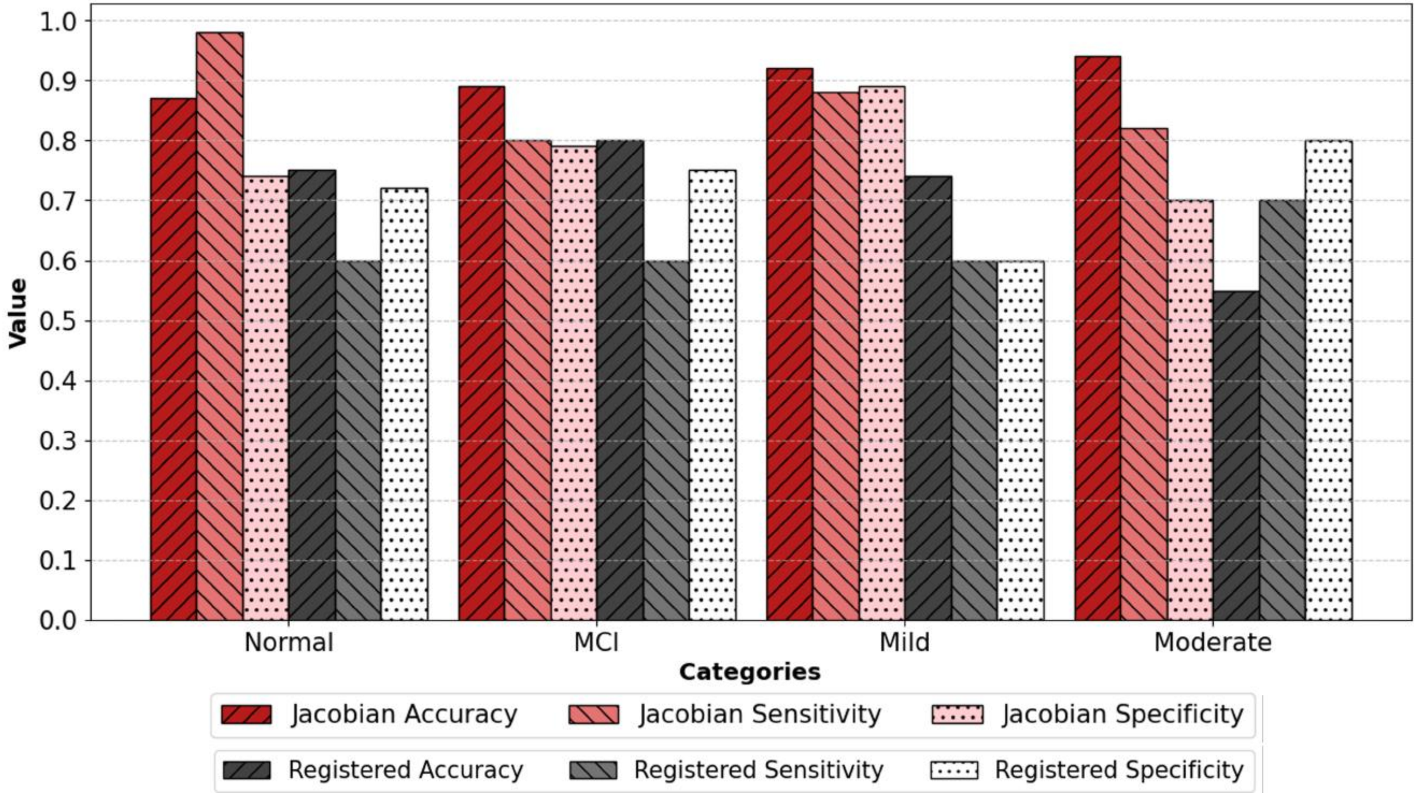
Training Details

- **Hyperparameters:**
  - Optimizer: Adam
  - Cross-Entropy loss
  - Learning Rate: 1e-4
  - Batch Size: 15

- **5-Fold Cross-Validation** is used to ensure robust evaluation:
  - Data is split into 5 subsets: 4 for training and 1 for validation in each fold.
  - 50 epochs per fold
  - **Early stopping** if validation loss does not improve (to prevent overfitting)

Experiments

Diagnostic Performance



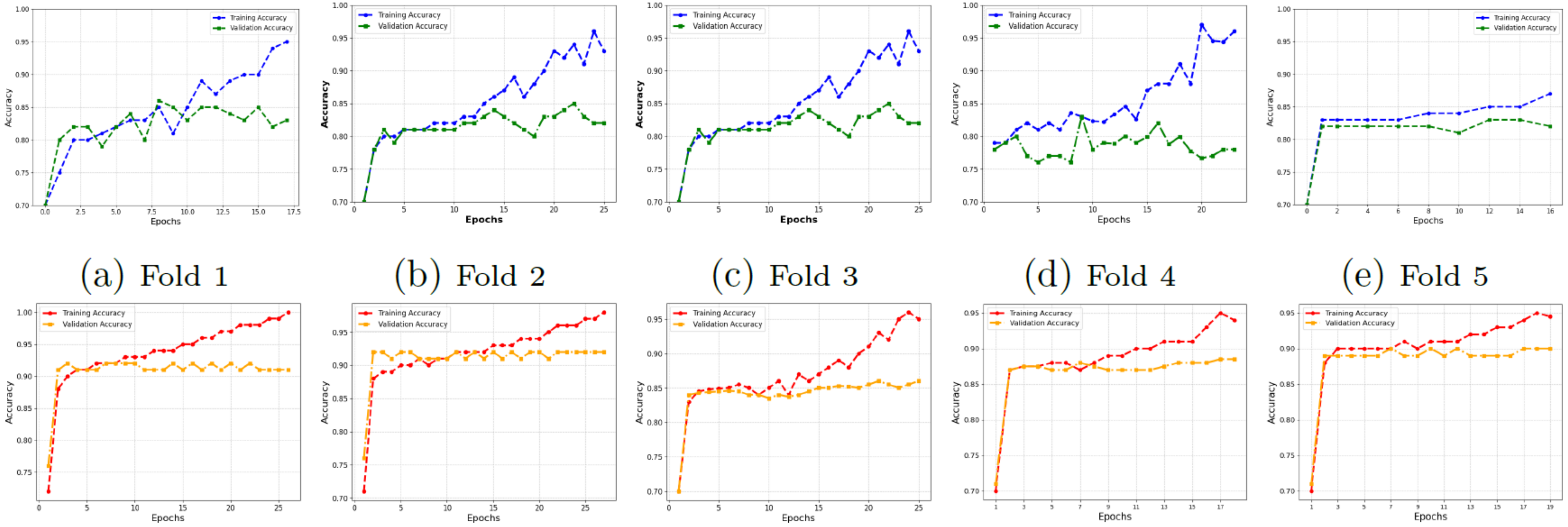
- **JM-base approach outperforms those using standard registered MRIs in all metrics.**
- JM contributes more discriminative information by capturing local volumetric brain changes.

Experiments

Fold-wise examination

No JM:

With JM:



(a) Fold 1 (b) Fold 2 (c) Fold 3 (d) Fold 4 (e) Fold 5

(f) Fold 1 (g) Fold 2 (h) Fold 3 (i) Fold 4 (j) Fold 5

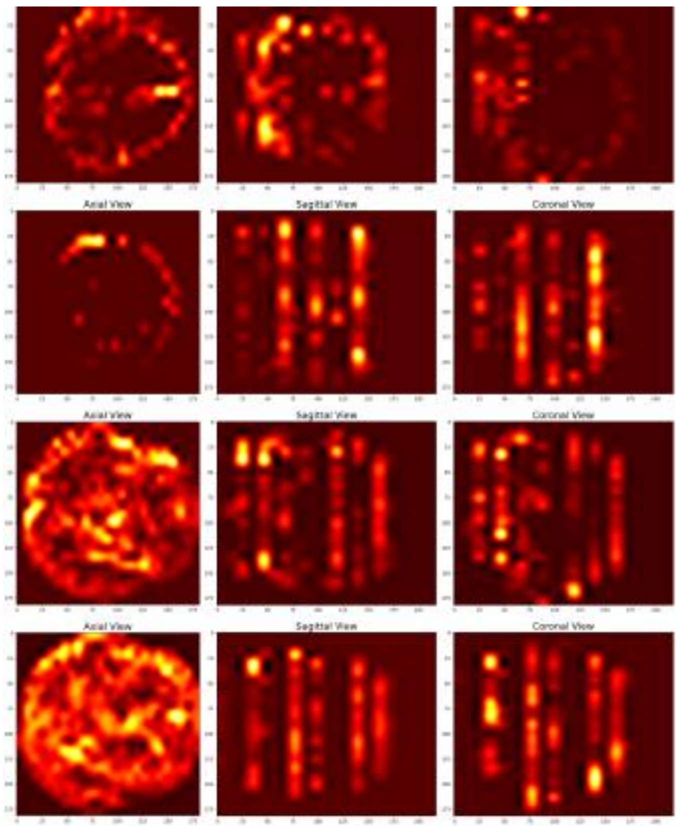
JM improves training stability:

- no JM: More fluctuation and less consistent validation accuracy.
- with JM: Smoother, stabler convergence with better validation performance.

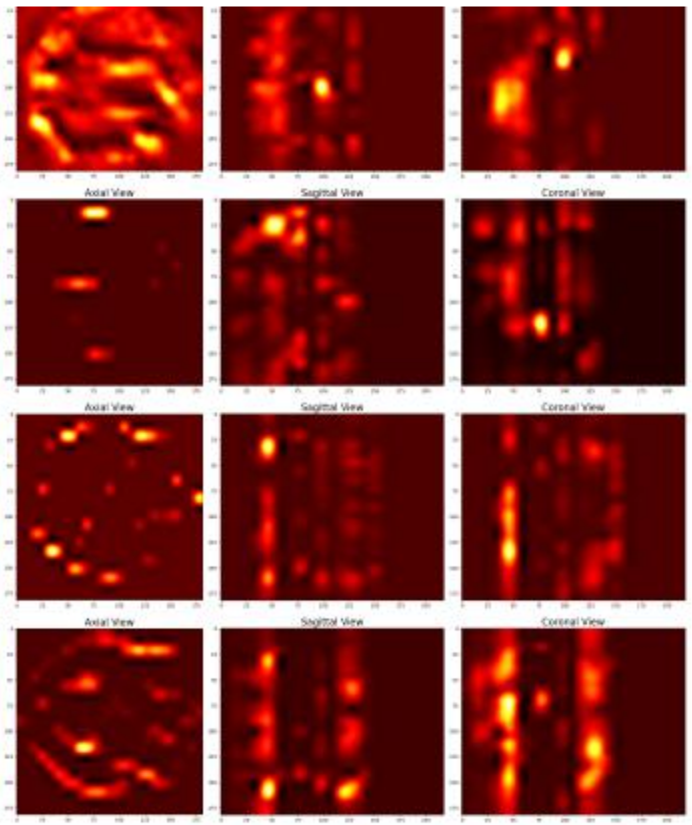
Experiments

Interpretations (Qualitative)

Grad-CAM is extended to 3D CNN to visualize the most influential regions.



No JM: broader, less localized/focused activations.



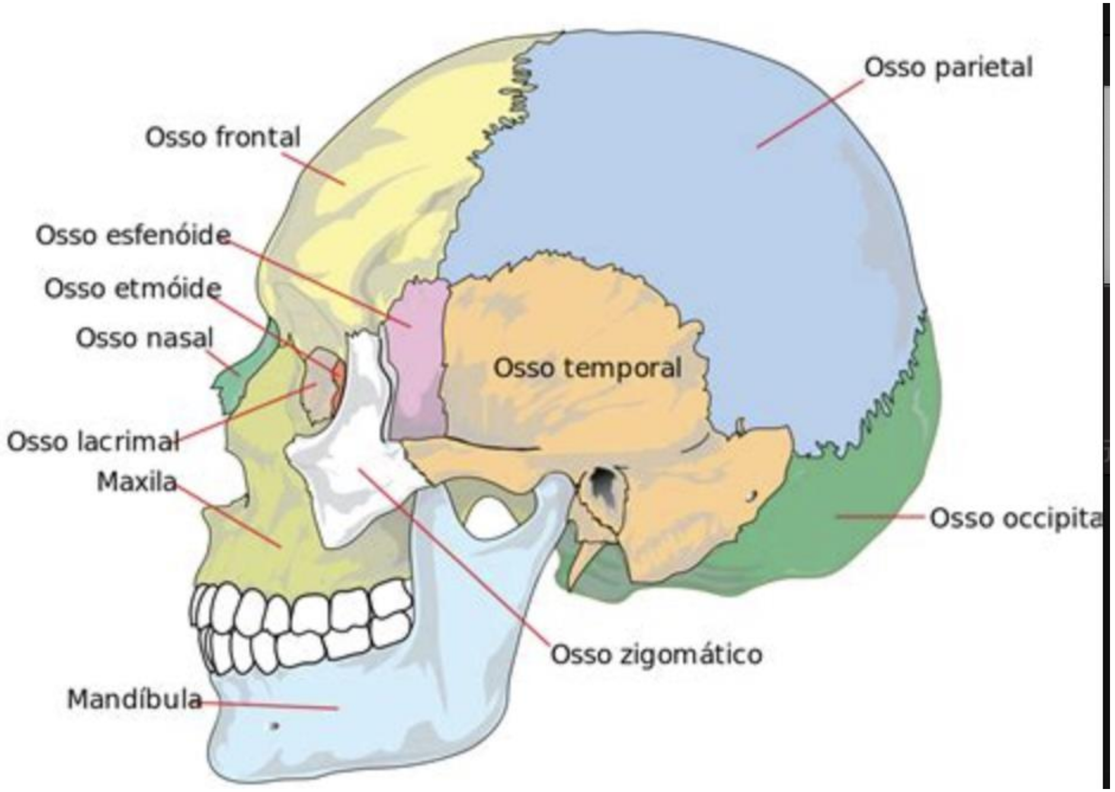
With JM: sharper focus on specific structural changes.

Experiments

Interpretations (Quantitative)

We apply the following steps to quantitatively measure the model interpretability:

1) Register Grad-CAM heatmaps to **MNI152** template.
2) Use the **Harvard-Oxford cortical atlas** to divide brain into anatomical regions.
3) Compute **average voxel intensity** within each region.
4) **Rank** regions by importance (activation level).



Experiments

Interpretations (Quantitative)

Avg. voxel intensity of each region

Table 1: Brain regions ranked by importance for each AD class based on heatmap intensity.

CN	MCI	MLD	SEV
Frontal-Temporal (2.71)	Frontal-Temporal (2.45)	Frontal-Temporal (2.76)	Frontal-Temporal (2.76)
Sub-lobar (2.51)	Temporal Lobe (1.94)	Temporal Lobe (2.33)	Frontal Lobe (2.28)
Temporal Lobe (2.43)	Frontal Lobe (1.89)	Frontal Lobe (2.28)	Parietal Lobe (1.77)
Limbic Lobe (2.40)	Sub-lobar (1.78)	Sub-lobar (2.20)	Limbic Lobe (2.01)
Frontal Lobe (2.37)	Background (1.73)	Occipital Lobe (2.02)	Occipital Lobe (2.02)
Midbrain (2.29)	Limbic Lobe (1.67)	Pons (1.82)	Pons (1.82)
Pons (2.26)	Occipital Lobe (1.59)	Posterior Lobe (1.91)	Posterior Lobe (1.91)
Background (2.08)	Anterior Lobe (1.53)	Background (1.73)	Background (1.73)
Parietal Lobe (2.07)	Medulla (1.37)	Anterior Lobe (1.53)	Medulla (1.43)
Posterior Lobe (1.98)	Midbrain (1.61)	Medulla (1.43)	Anterior Lobe (1.53)
Medulla (1.76)	Frontal-Temporal (2.45)	Parietal Lobe (1.77)	Midbrain (1.90)
Anterior Lobe (1.97)	Parietal Lobe (1.33)	Frontal Lobe (1.89)	Pons (1.82)

- Frontal-temporal region is a key area involved in memory, decision-making and language.
- Changes / degenerates early, even starting from CN and MCI.

Experiments

Interpretations (Quantitative)

Avg. voxel intensity of each region

Table 1: Brain regions ranked by importance for each AD class based on heatmap intensity.

CN	MCI	MLD	SEV
Frontal-Temporal (2.71)	Frontal-Temporal (2.45)	Frontal-Temporal (2.76)	Frontal-Temporal (2.76)
Sub-lobar (2.51)	Temporal Lobe (1.94)	Temporal Lobe (2.33)	Frontal Lobe (2.28)
Temporal Lobe (2.43)	Frontal Lobe (1.89)	Frontal Lobe (2.28)	Parietal Lobe (1.77)
Limbic Lobe (2.40)	Sub-lobar (1.78)	Sub-lobar (2.20)	Limbic Lobe (2.01)
Frontal Lobe (2.37)	Background (1.73)	Occipital Lobe (2.02)	Occipital Lobe (2.02)
Midbrain (2.29)	Limbic Lobe (1.67)	Pons (1.82)	Pons (1.82)
Pons (2.26)	Occipital Lobe (1.59)	Posterior Lobe (1.91)	Posterior Lobe (1.91)
Background (2.08)	Anterior Lobe (1.53)	Background (1.73)	Background (1.73)
Parietal Lobe (2.07)	Medulla (1.37)	Anterior Lobe (1.53)	Medulla (1.43)
Posterior Lobe (1.98)	Midbrain (1.61)	Medulla (1.43)	Anterior Lobe (1.53)
Medulla (1.76)	Frontal-Temporal (2.45)	Parietal Lobe (1.77)	Midbrain (1.90)
Anterior Lobe (1.97)	Parietal Lobe (1.33)	Frontal Lobe (1.89)	Pons (1.82)

- Rank of temporal lobe increases in MCI and MLD.
- Temporal lobe includes hippocampus (海马体) and entorhinal cortex (内嗅皮层), and is among the first regions to show atrophy, with **memory loss** being a key symptom.

Experiments

Interpretations (Quantitative)

Avg. voxel intensity of each region

Table 1: Brain regions ranked by importance for each AD class based on heatmap intensity.

CN	MCI	MLD	SEV
Frontal-Temporal (2.71)	Frontal-Temporal (2.45)	Frontal-Temporal (2.76)	Frontal-Temporal (2.76)
Sub-lobar (2.51)	Temporal Lobe (1.94)	Temporal Lobe (2.33)	Frontal Lobe (2.28)
Temporal Lobe (2.43)	Frontal Lobe (1.89)	Frontal Lobe (2.28)	Parietal Lobe (1.77)
Limbic Lobe (2.40)	Sub-lobar (1.78)	Sub-lobar (2.20)	Limbic Lobe (2.01)
Frontal Lobe (2.37)	Background (1.73)	Occipital Lobe (2.02)	Occipital Lobe (2.02)
Midbrain (2.29)	Limbic Lobe (1.67)	Pons (1.82)	Pons (1.82)
Pons (2.26)	Occipital Lobe (1.59)	Posterior Lobe (1.91)	Posterior Lobe (1.91)
Background (2.08)	Anterior Lobe (1.53)	Background (1.73)	Background (1.73)
Parietal Lobe (2.07)	Medulla (1.37)	Anterior Lobe (1.53)	Medulla (1.43)
Posterior Lobe (1.98)	Midbrain (1.61)	Medulla (1.43)	Anterior Lobe (1.53)
Medulla (1.76)	Frontal-Temporal (2.45)	Parietal Lobe (1.77)	Midbrain (1.90)
Anterior Lobe (1.97)	Parietal Lobe (1.33)	Frontal Lobe (1.89)	Pons (1.82)

- In late AD, neurodegeneration becomes widespread.
- Frontal lobe handles planning, judgment, social behavior.
- Parietal (顶骨) lobe controls spatial orientation, attention.
- These areas are affected less during early stages but more in late stages.

Skip if running short on time

Experiments

Interpretations (Quantitative)

Table 1: Brain regions ranked by importance for each AD class based on heatmap intensity.

Avg. voxel intensity of each region

CN	MCI	MLD	SEV
Frontal-Temporal (2.71)	Frontal-Temporal (2.45)	Frontal-Temporal (2.76)	Frontal-Temporal (2.76)
Sub-lobar (2.51)	Temporal Lobe (1.94)	Temporal Lobe (2.33)	Frontal Lobe (2.28)
Temporal Lobe (2.43)	Frontal Lobe (1.89)	Frontal Lobe (2.28)	Parietal Lobe (1.77)
Limbic Lobe (2.40)	Sub-lobar (1.78)	Sub-lobar (2.20)	Limbic Lobe (2.01)
Frontal Lobe (2.37)	Background (1.73)	Occipital Lobe (2.02)	Occipital Lobe (2.02)
Midbrain (2.29)	Limbic Lobe (1.67)	Pons (1.82)	Pons (1.82)
Pons (2.26)	Occipital Lobe (1.59)	Posterior Lobe (1.91)	Posterior Lobe (1.91)
Background (2.08)	Anterior Lobe (1.53)	Background (1.73)	Background (1.73)
Parietal Lobe (2.07)	Medulla (1.37)	Anterior Lobe (1.53)	Medulla (1.43)
Posterior Lobe (1.98)	Midbrain (1.61)	Medulla (1.43)	Anterior Lobe (1.53)
Medulla (1.76)	Frontal-Temporal (2.45)	Parietal Lobe (1.77)	Midbrain (1.90)
Anterior Lobe (1.97)	Parietal Lobe (1.33)	Frontal Lobe (1.89)	Pons (1.82)

- Limbic system is crucial for emotion and memory.
- Appears to be involved more in early and late stages while less in intermediate stages

Skip if running short on time

Experiments

Interpretations (Quantitative) - **Summary**

- Frontal-Temporal regions are dominant across all classes → Key biomarkers in AD.
- As AD progresses, Temporal Lobe becomes more important in MCI and MLD stages.
- Parietal and Frontal Lobes gain importance in SEV → reflects widespread neurodegeneration.
- Sub-lobar and Limbic regions show varying importance, capturing non-linear disease progression.
- **Consistency with clinical evidence** is observed in our experiments.

Experiments

Extension to Multi-modal Setting

- **Why Multi-modal Imaging?**
  - Different imaging modalities capture complementary information:
    - **MRI:** Excellent **soft tissue** contrast—shows structural brain changes like atrophy.
    - **CT:** Captures bone and **dense tissue** differences; helps with structural localization and calcification, and can **fill in missing contrast** in certain brain areas.

- We extend our study by combining MRI + CT:
  - Concatenate MRI and CT images along the channel dimension.
  - The fused volume is treated as a single input to the 3D CNN.
  - The model learns from joint features from the very beginning of processing.

Experiments

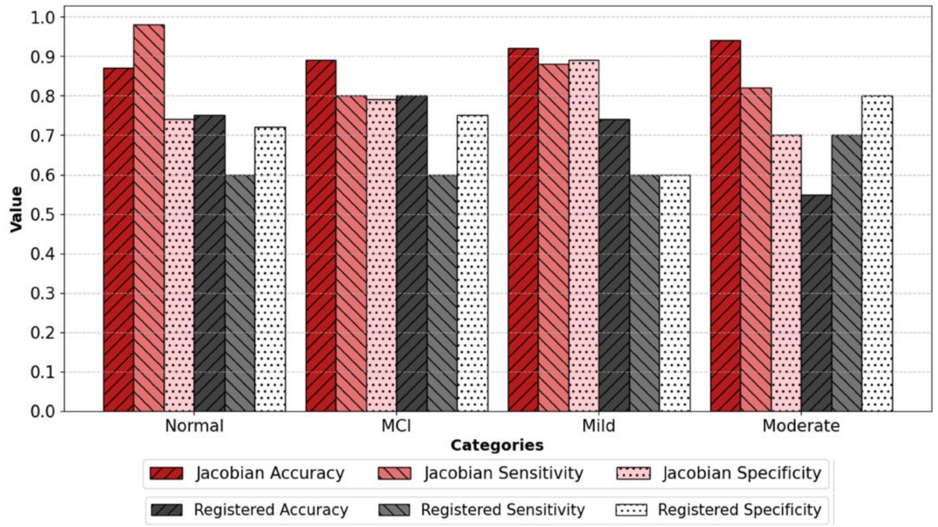
Extension to Multi-modal Setting

Results

	Accuracy				Precision				Recall			
	CN	MCI	MLD	SEV	CN	MCI	MLD	SEV	CN	MCI	MLD	SEV
REG	88.3	90.5	83.4	83.4	86.8	82.8	80	95.5	83.3	64.0	69.3	84.4
JM	95.2	96.3	90.2	90.2	92.8	100	83.33	98.6	94.96	89.6	78.6	90.2

- Across all stages (CN, MCI, MLD, SEV), the **use of Jacobian Maps still significantly improves all the metrics.**

- Compared to the unimodal (MRI) case, performance improves overall, especially in early stages (which is important)

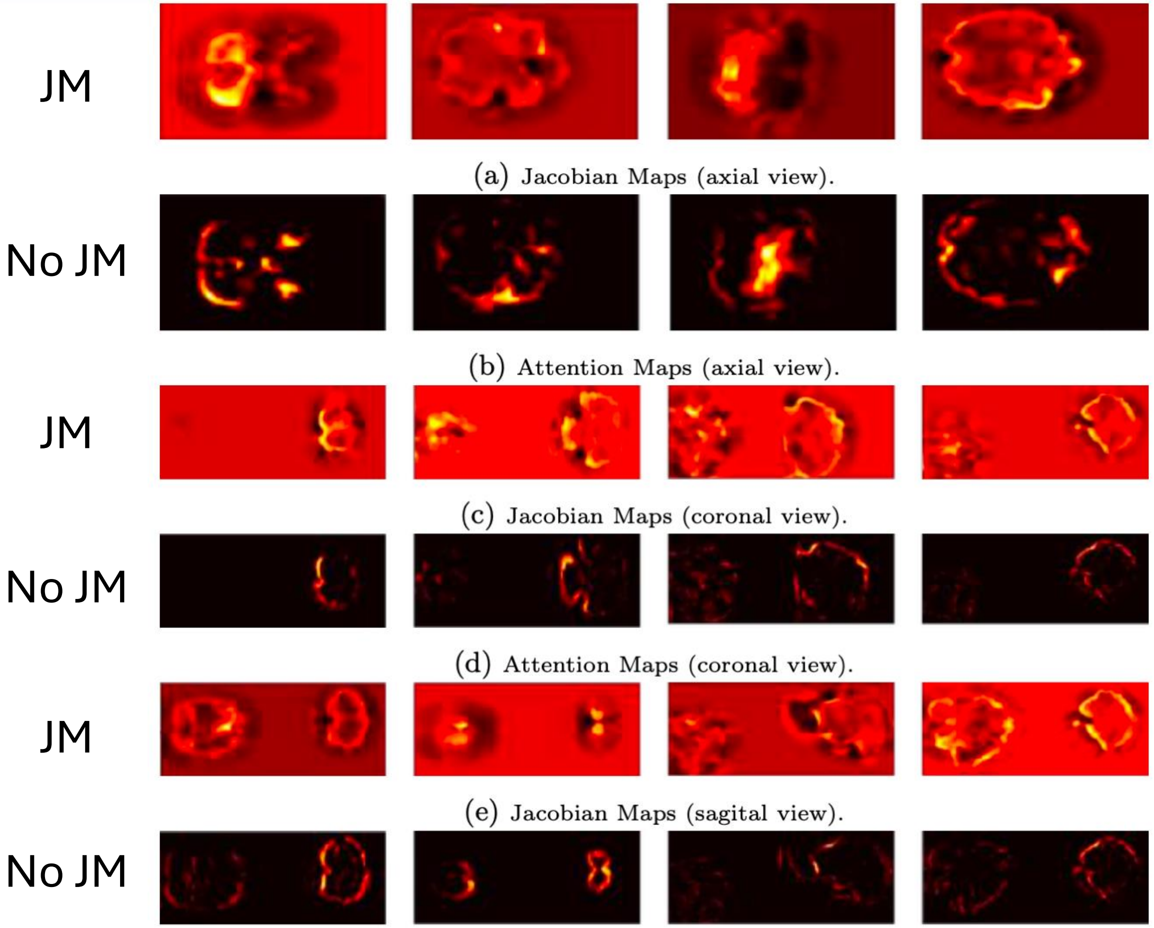


Experiments

Extension to Multi-modal Setting

Interpretability Results

Jacobian Maps **enhance interpretability further** by **elevating the intensity level** of AD-relevant regions; i.e., making the volumetric deformations more **pronounced.**



(a) Jacobian Maps (axial view).
(b) Attention Maps (axial view).
(c) Jacobian Maps (coronal view).
(d) Attention Maps (coronal view).
(e) Jacobian Maps (sagital view).
(f) Attention Maps (sagital view).

Conclusion

🔍 Clinical Challenge

- Alzheimer’s Disease (AD) is progressive (over multiple stages) and complex.
- Early, accurate diagnosis is critical — yet explainability is key to adoption as well.

💡 Our Contributions

- Introduced Jacobian Maps (JMs) as an **ante-hoc XAI** approach
  - Capturing subtle, localized brain deformations to enable more interpretable model
- Integrated JMs into a 3D CNN and provided visual + quantitative interpretability using JM + Grad-CAM.
- Extended to a multimodal (MRI + CT) setting.

📈 Key Outcomes

- Improved diagnostic performance across all AD stages.
- Enhanced interpretation consistent with clinical evidence (e.g., frontal-temporal lobes)
- Bridges the gap between deep learning and clinical interpretability.
- Scalable to other neurodegenerative diseases (e.g. Parkinson) and modalities (e.g., PET).


תודה
Dankie
Gracias
Спасибо
شكراً
Merci
Takk
Köszönjük
Terima kasih
Grazie
Dziękujemy
Dėkojame
Ďakujeme
Vielen Dank
Paldies
Kiitos
Täname teid
谢谢
Thank You
Tak
感謝您
Obrigado
Teşekkür Ederiz
감사합니다
Σας Ευχαριστούμ
ขอบคุณ
Bedankt
Děkujeme vám
ありがとうございます
Tack

Comparison with VBM

What is VBM:

- **Voxel Based Morphometry** (VBM) is a neuroimaging analysis technique that compares local brain anatomy across individuals or groups.
- It helps detect structural differences in brain tissues.

Key Difference:

- While VBM uses JMs *statistically* to compare anatomical differences across groups (e.g., Alzheimer's vs. control) as a modulation technique,
- We **repurpose JMs**:
  - As a *pre-model input* to directly train deep learning models, to enhance both performance and explainability.

Key novelties over VBM

1. Pre-model use of JMs in a deep learning pipeline

- Instead of using JMs after training for statistical analysis (as in VBM), we compute JMs and feed them as **input images** to a model.
- This is different from VBM's GLM or voxel-wise t-tests.

2. Explainability through 3D Grad-CAM on JMs

- We apply Grad-CAM on a JM-based 3D CNN to generate heatmaps.
- Because JMs already represent deformation, the Grad-CAM heatmaps are **more anatomically grounded** (i.e., the interpretability is *meaningfully tied to real morphometric change*).

3. Avoidance of segmentation and patch-based limitations

- VBM often depends on accurate segmentation and group-wise statistics.
- Our method bypasses those steps, offering **full-volume learning without patching or tissue segmentation**, which reduces preprocessing artifacts.