

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

Yasmine Mustafa¹, Mohamed Elmahallawy², Thomas Tie Luo³

- ¹ Missouri University of Science and Technology
- ² Washington State University

³ 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:



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.

Transforming Brain Images into Jacobian Maps (JM)



Transforming Brain Images into Jacobian Maps (JM)



(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)

Transforming Brain Images into Jacobian Maps (JM)



(c) Brain Extraction

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

Transforming Brain Images into Jacobian Maps (JM)



(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)

Transforming Brain Images into Jacobian Maps (JM)



(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)



(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.

Computing Jacobian Maps

Brain Volume Changes Denoting Dementia



Deformation Field

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

Jacobian Matrix (J):

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

Jacobian Determinant:

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

Jacobian Map

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

Semantics:

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

$$\begin{bmatrix} \vdots \\ \dots Det(J(v(x, y, z)) \dots \\ \vdots \end{bmatrix} x = 1 \dots W$$

$$y = 1 \dots H$$

$$z = 1 \dots D$$





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



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.



Fold-wise examination

JM improves training stability:

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

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.

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).



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

CN	MCI	$ \mathbf{MLD} $	MOD	
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.

Avg. voxel

intensity of

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

CN	MCI	MLD	MOD		
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.

Avg. voxel

intensity of

 \mathbf{CN} MCI MLD MOD 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)Medulla (1.43)Anterior Lobe (1.53)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)|

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

- 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.

Avg. voxel

intensity of



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

CN	MCI	MLD	MOD			
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

Avg. voxel

intensity of





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.

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.

Extension to Multi-modal Setting

Results

	Accuracy			Precision			Recall					
	CN	MCI	MLD	MOD	CN	MCI	MLD	MOD	CN	MCI	MLD	MOD
REG	88.3	90.5	83.4	83.4	86.8	82.8	80	95.5	83.3	64.0	69.3	84.4
$\mathbf{J}\mathbf{M}$	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)



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.



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.
- Anticipated to be extensible to other neurodegenerative diseases (e.g. Parkinson, epilepsy/seizures, autism, ADHD) and modalities (e.g., PET).

