Explainable Alzheimer's Diagnosis using 2D CNNs on 3D MRI brain: AXIAL Performance Enhancement

Politecnico di Milano
Davide Villani, Filippo Wang

Abstract

Diagnosing Alzheimer's Disease (AD) accurately from brain MRI scans is still a key challenge in medical AI; moreover, the lack of interpretability of Deep Learning models stops these systems from being used in real-world scenarios. In this work, we replicate the work of [LBF+24], which introduces a pipeline that combines 2D Convolutional Neural Networks (CNNs) with Explainable Artificial Intelligence (XAI) techniques to analyze 3D brain magnetic resonance images for the diagnosis of AD Using the publicly available "ADNI1: Complete 1Yr 1.5T dataset". After replicating the results of the authors, we applied methodological adjustments and hyperparameter tuning to improve the results of the original paper, achieving State-Of-The-Art metrics.

1 Dataset

The ADNI1: Complete 1Yr 1.5T dataset [RBD+21] comprises data from participants who underwent magnetic resonance imaging (MRI) scans at baseline (screening), 6 months , and 12 months using 1.5 Tesla MRI scanners. This standardized dataset was developed to promote consistency in data analysis and facilitate direct comparisons of various analysis methods

The dataset includes participants across three diagnostic categories:

- **Cognitively Normal (CN)**: 204 individuals 76.31 ± 5.22 years old
- **Mild Cognitive Impairment (MCI)**: 331 individuals 75.11 ± 6.92 years old, further divided into pMCI (*progressive* MCI) and sMCI (*stable* MCI)
- **Alzheimer's Disease (AD)**: 191 individuals 75.23 ± 7.02 years old

Each participant's scan consists of 3D images which are divided into 3 planes discretized into slices: **Sagittal Plane** containing on average 160–170 slices, **Axial Plane** on average 130–150 slices and **Coronal Plane** with 180-200.

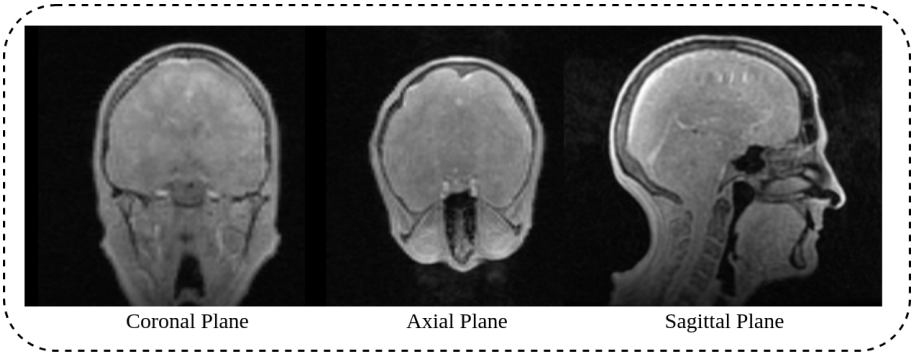


Figure 1: Example of Standard anatomical plane's slices before Pre-Processing

As we can see from Figure 1, raw MRI slices often contains many irrelevant structures, such as the skull and surrounding tissues, which are not directly related to the brain. To ensure accurate analysis, it is essential to exclude these non-brain regions. This is why a pre-processing stage is necessary to isolate and focus only on the brain.

2 Pipeline

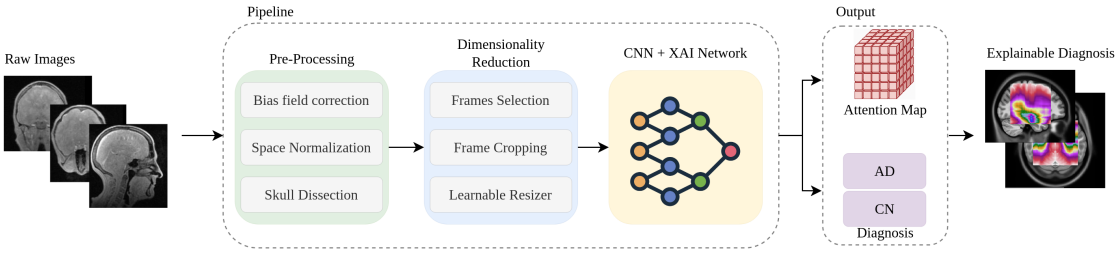


Figure 2: Pipeline highlighting the pre-processing steps

2.1 Pre-Processing

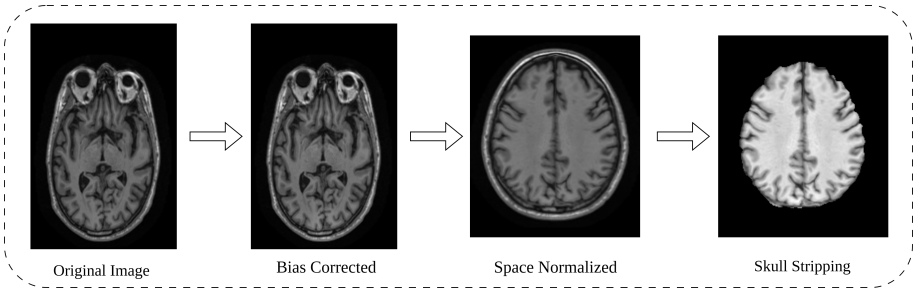


Figure 3: Example of a pre processing pipeline on a single slice

The pre-processing pipeline utilized in this study consists of three main stages designed to prepare the MRI images for further analysis:

- **Bias Field Correction**: To correct for common non-uniformities in MRI scans, we applied the N4ITK algorithm, a widely used method that helps improve image quality by reducing low-frequency intensity artifacts caused by magnetic field inhomogeneities.
- **Spatial Normalization**: The MRI scan of each subject was spatially aligned to a standard anatomical space, specifically the Montreal Neurological Institute (MNI) 152 template. This was achieved using the SyN (Symmetric Normalization) algorithm. The alignment was performed with respect to the ICBM 2009c nonlinear symmetric version of the MNI template, ensuring consistency across all images.
- **Skull Stripping**: Non-brain tissues such as the skull, scalp, and dura can interfere with analysis pipelines and distort measurements of brain structure. To eliminate these, we employed the Brain Extraction Tool (BET) implemented within the FSL software suite. This step is crucial for accurate brain morphometry and analysis.

The first two steps—bias field correction and affine registration—were carried out using the t1-linear pipeline provided by the Clinica platform.

2.2 Feature Extraction

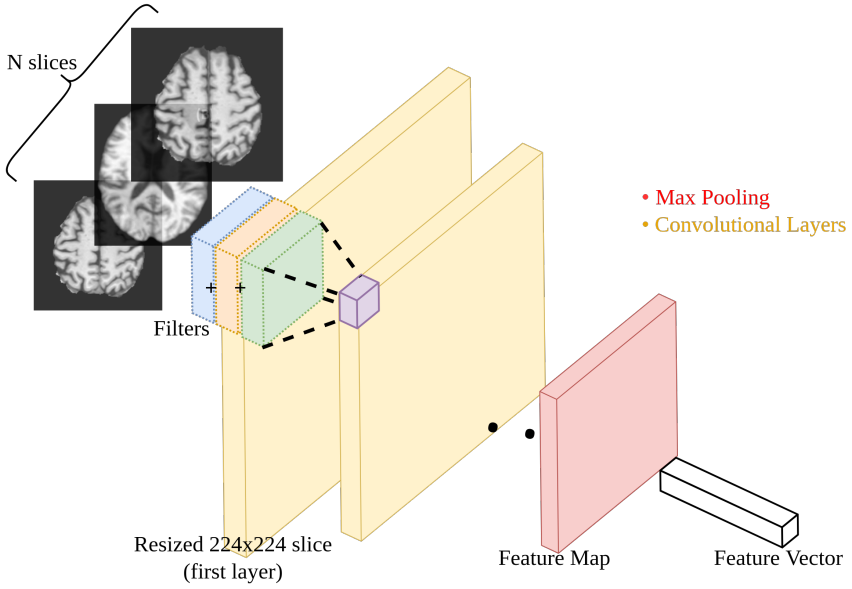


Figure 4: Convolutional Backbone with shared weights, this is applied to all the planes separately

After the pre-processing steps and the dimensionality reduction, what we are left with for each plane (*Sagittal, Coronal, Axial*) is a 3D volumetric brain scan represented as a tensor $X \in \mathbb{R}^{H\times W\times N}$, where H, W, and N denote the height, width, and depth (number of slices) of the pre-processed scans, respectively. We extract N axial 2D slices from this volume:

$$X = \{x_1, x_2, \ldots, x_N\}, \quad x_i \in \mathbb{R}^{H\times W}$$

Each slice x_i is a single-channel (grayscale) image which contrasts with standard 2D convolutional backbones such as *VGG 16*, pre-trained on the ImageNet dataset since they expect 3-channel RGB inputs. To adapt these models for grayscale images without redundantly replicating the input across channels, we modify the **first** convolutional layer. Under the assumption that the filters operate linearly and combine contributions additively across channels what we can do is to apply 3 filters to the same image just by adding them before and apply them only to our single-channel image (Figure 4).

Each slice x_i is first resized to 224×224 pixels to match the expected input size of the backbone and normalized accordingly. It is then passed through the convolutional network $\mathcal{F}_\theta$, composed of convolutional layers and a global max-pooling operation, to extract a feature vector:

$$f_i = \text{AvgPool}(\mathcal{F}_\theta(x_i)) = \frac{1}{H'W'}\sum_{h=1}^{H'}\sum_{w=1}^{W'}\mathcal{F}_\theta(x_i)[h,w]$$

Here, d denotes the dimensionality of the output feature vector. Importantly, the same backbone $\mathcal{F}_\theta$ is used across all slices, meaning the weights θ are shared across the sequence—similar to weight-sharing in Recurrent Neural Networks (RNNs).Since the parameters θ are shared across all slices, the network updates them during backpropagation based on a *global loss* that accounts for all slices. This means that even though slices are independently encoded, their embeddings f_i contribute collectively to learning θ. Specifically, if $\mathcal{L}$ is the final loss function computed from a classifier output based on the aggregated feature vector, then the gradient w.r.t. θ is given by:

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{i=1}^{N}\frac{\partial \mathcal{L}}{\partial f_i}\cdot\frac{\partial f_i}{\partial \theta} = \sum_{i=1}^{N}\frac{\partial \mathcal{L}}{\partial \mathcal{F}_\theta(x_i)}\cdot\frac{\partial \mathcal{F}_\theta(x_i)}{\partial \theta}$$

This enforces that the shared encoder learns 2D representations optimized in the context of the entire 3D image. It is important to notice that when we talk about a 3D image we are referring to the 3D image of a specific plane, this is fundamental because we are still not taking into account the inter-plane dependencies but only the inter-slice dependencies which are computed intra-plane. As a result, we obtain a sequence of feature vectors:

$$F = \{f_1, f_2, \ldots, f_N\}, \quad f_i \in \mathbb{R}^d$$

This feature sequence compactly represents the original 3D volume and serves as a suitable input for the next module that capture *inter-slice dependencies*.

2.3 Attention XAI Fusion Module

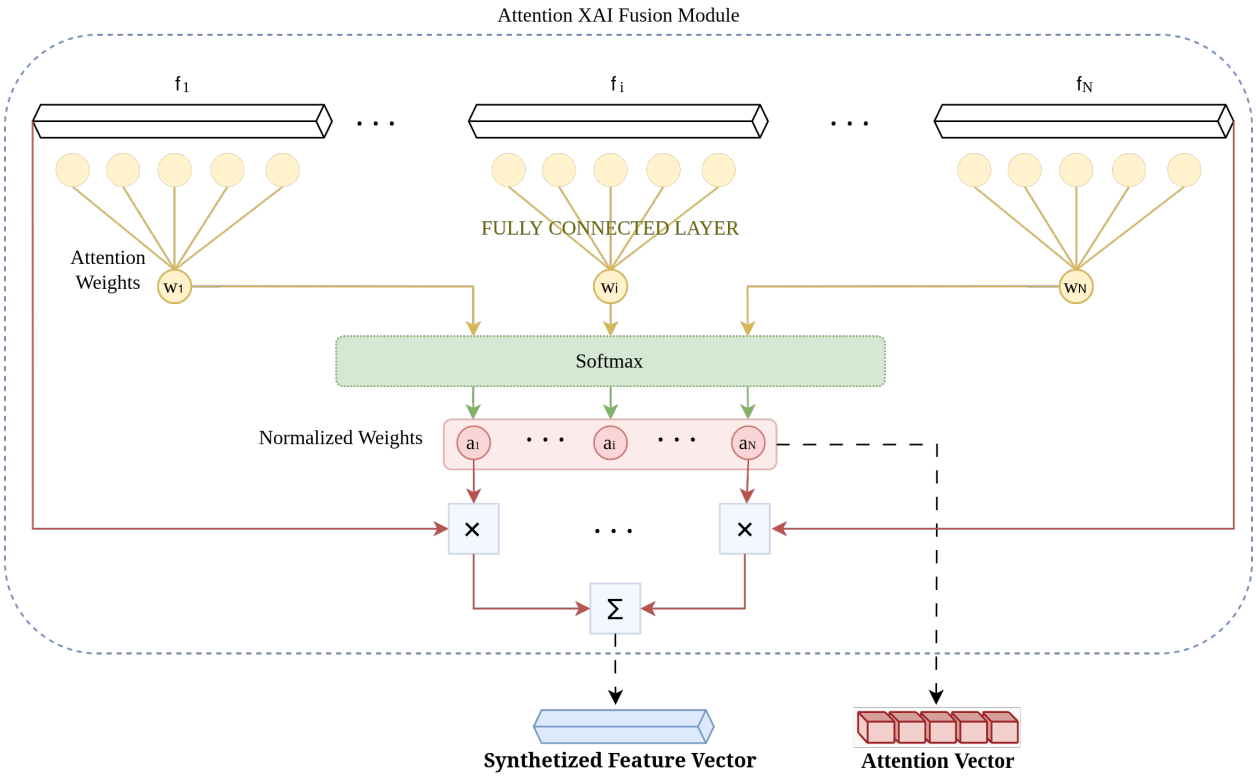


Figure 5: Attention XAI Fusion Module scheme applied to each plane slices

After extracting individual 2D slice features using a shared convolutional backbone, this module enables the network to learn *inter-slice dependencies* and capture *global 3D patterns*. To quantify slice importance, an attention mechanism is applied. Each feature vector f_i is passed through a lightweight fully connected (FC) layer to produce an attention score $w_i \in \mathbb{R}$:

$$w_i = \mathrm{FC}_{\mathrm{att}}(f_i)$$

This attention mechanism introduces only $d + 1$ parameters and remains computationally efficient. The attention scores $\{w_i\}$ are then normalized using a softmax function to obtain a probability distribution over slices:

$$\alpha_i = \frac{\exp(w_i)}{\sum_{j=1}^{N} \exp(w_j)}, \quad \sum_{i=1}^{N} \alpha_i = 1$$

Each normalized weight α_i quantifies the importance of slice x_i relative to the full volume. Finally, we compute a fused global feature representation by a weighted sum over the slice embeddings:

$$F = \sum_{i=1}^{N} \alpha_i f_i$$

This resulting vector $F \in \mathbb{R}^d$ encodes both spatial content and the attention-weighted contributions of all slices. It is then passed to the final classifier to predict the class label $\hat{y}$, completing the learning process. Importantly, the set of attention weights $\{\alpha_i\}$ offers interpretability, identifying which slices influenced the final prediction the most.

In summary, the output **for each plane slice** of this module is composed of :
A *Synthesized Feature Vector* F and an *Attention Vector* (Figure 5) , these will be used respectively in the Diagnosis and the Heat Attention Map Generation.

2.4 Diagnosis

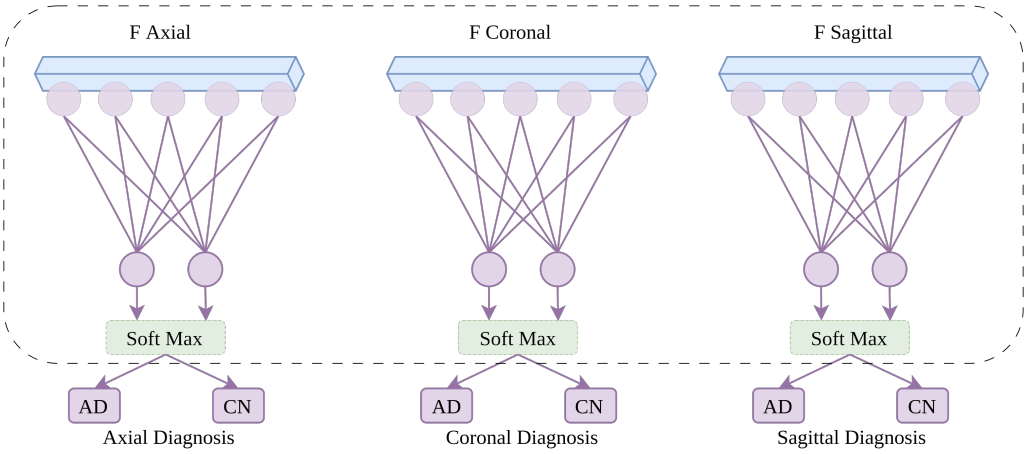


Figure 6: Diagnosis Module scheme applied to each plane

In the final *Diagnosis Module*, the synthesized feature vector $F \in \mathbb{R}^d$, which aggregates information from the entire 3D brain scan via the attention mechanism, is passed through a classification head for diagnosis prediction.

The head consists of a fully connected layer followed by a softmax activation to perform binary classification (e.g., Alzheimer's Disease (AD) vs. Cognitively Normal (CN)). Let FC_{head} be a Fully Connected layer. The output $\mathbf{z} \in \mathbb{R}^2$ is given by:

$$\mathbf{z} = \text{FC}_{\text{head}}(F)$$

These is then normalized using the softmax function to obtain the final class probabilities $\mathbf{D} = [D_1, D_2] \in [0, 1]^2$, where $D_1 + D_2 = 1$:

$$D_k = \frac{\exp(z_k)}{\sum_{j=1}^{2} \exp(z_j)}, \quad k = 1, 2$$

Here, D_1 and D_2 represent the probabilities assigned to the two diagnosis classes. The model predicts the label corresponding to the higher probability. Thus, this module produces a diagnosis based on the attention-weighted, aggregated feature representation of the entire scan.

2.5 Heat Map Generation

In order to make the model interpretable we need to generate an Attention Map which puts together the attention vectors coming from all three views, sagittal, axial and coronal, which we will refer to as *(s)*, *(a)* and *(c)*. As we saw in 2.3 one of the output of that Module is an Attention Vector, computed after applying a softmax to our Feature Vectors, this is done to each plane for each slice and what we end up with is N_a, N_s and N_c number of Attention Vectors. We integrate the attention weights α_s, α_c, and α_a into a unified 3D attention map. The 3D attention map A is constructed by applying the following operation to each voxel:

$$A[i, j, k] = \alpha_s[i] \cdot \alpha_c[j] \cdot \alpha_a[k] \tag{1}$$

To make interpretation and comparison easier, we normalize the entire 3D attention map into the range $[0, 1]$, ensuring that the highest values correspond to areas of high diagnostic relevance. We use min-max normalization:

$$A = \frac{A - \min(A)}{\max(A) - \min(A)} \tag{2}$$

A further step in order to achieve comparability across subjects, each MRI image, denoted by I, was normalized to the MNI 152 standard space using the transformation function f_{MNI152}, thus $I_{\text{norm}} = f_{\text{MNI152}}(I)$.
This step is crucial for aligning brain structures across different individuals to facilitate the identification of AD-specific biomarkers.
The resulting map highlights the brain regions most significantly associated with the diagnostic output of the network, offering insights into the pathological hallmarks of Alzheimer's Disease (AD) as learned by the model through its training process.

Heat Map

A binary heatmap, H_{binary}, was generated to isolate regions of significant structural patterns associated with Alzheimer's Disease (AD), utilizing a threshold θ set at the 99.9th percentile. The binary heatmap is defined as:

$$H_{\text{binary}}[i, j, k] = \begin{cases} 1, & \text{if } A[i, j, k] > \theta \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

For visualization purposes, the MRI data I_{norm} was augmented by overlaying H_{binary} to enhance the saliency of the regions implicated in AD:

$$I_{\text{XAI}} = I_{\text{norm}} + H_{\text{binary}} \times \delta \tag{4}$$

where δ is an amplification factor set to 10.

3 Results

3.1 Experimental Setup

Training was performed on the JEDI cluster at the Jülich Supercomputing Centre, utilizing NVIDIA GH200 GPUs.
The model was first tested using the best configuration specified by the authors of the original AXIAL paper and we then modified it to achieve better results.

Hyperparameter	Original Values	Final Values
num_epochs	100	30
batch_size	8	8
dropout	0.3	0.5
k_folds	5	8
num_slices	80	60
learning_rate	0.0001	0.00005
weight_decay	0.01	0.01
freeze_first_percentage	0.5	0.3
optimizer	AdamW	AdamW
patience(Early Stopping)	20	10

Table 1: Comparison of hyperparameter configurations

Before presenting the results, it is important to note that the tests were conducted on a relatively small dataset, comprising fewer than 400 patients. This limited dataset size was the main reason why our model—despite its complexity, with a backbone containing over 138 million parameters—was prone to overfitting on the training data.

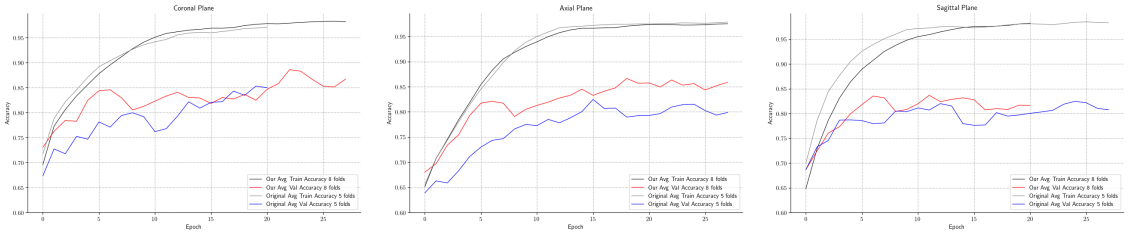
To address this issue, we implemented several modifications. First, we increased the number of folds in the cross-validation procedure, allowing the model to be trained on a larger portion of the data. This helped reduce overfitting and improved the loss on the validation set. Additionally, we reduced the spatial extent of the slices by selecting only the most informative ones. This not only decreased the likelihood of overfitting but also accelerated the training process—an essential consideration given that the model had to be trained within a 6-hour window due to cluster limitations.
It is really important to notice that when using cross validation it's fundamental to perform the data division in the correct step, since a wrong split will cause **data leakage** causing the model to have really high test accuracy scores only because the data subjects in validation and train get mixed up.

3.2 Experimental Results

In this section we talk about the results obtained with our tests and compare them to the optimal configuration described in the original paper, setting the same random seed for reproducibility.
Clinically, the most important sections used to predict Alzheimer are the coronal plane and the axial plane. Coronal slices, especially those aligned perpendicular to the long axis of the hippocampus, provide a clearer view of hippocampal atrophy. This orientation allows for better assessment of the cavities formed by hippocampal degeneration, which are indicative of AD.
All test results are reported as the average across all folds used in our best configuration. Specifically, while the original paper computed model accuracy by averaging over 5 folds, our results are based on an 8-fold cross-validation. To enable a clearer comparison, we also included box plots of the results. Given the relatively small size of the dataset, there can be significant variance in performance between different folds, which is an important factor to consider when interpreting the results.

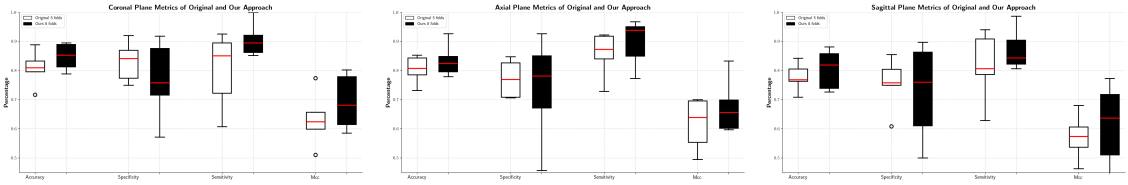
Plane	ACC		SPE		SEN		MCC	
	Ours	Original	Ours	Original	Ours	Original	Ours	Original
Coronal	**85.30%**	80.63%	78.21%	**83.23%**	**91.02%**	79.60%	**69.24%**	63.45%
Axial	**84.17%**	80.92%	75.33%	**77.42%**	**89.91%**	83.45%	**66.32%**	61.81%
Sagittal	**81.79%**	78.25%	73.21%	**74.65%**	**86.73%**	80.44%	**61.13%**	55.53%

Table 2: Average (across folds) performance comparison across different brain planes



(a) Coronal Plane (b) Axial Plane (c) Sagittal Plane

Figure 7: Accuracy curves for each anatomical plane on test and validation sets.



(a) Coronal Plane (b) Axial Plane (c) Sagittal Plane

Figure 8: Test metrics for each anatomical plane.

As shown in Table 2, our approach consistently outperforms the original model across all anatomical planes in terms of **accuracy**, **sensitivity**, and **MCC**. The **coronal plane** achieves the highest **accuracy** (85.30%) and excels in **sensitivity** (91.02%) and **MCC** (69.24%), reflecting its importance in identifying hippocampal atrophy one of the key structural indicators of Alzheimer's disease. The **axial plane** similarly shows strong results in sensitivity (89.91%) and MCC (66.32%), while the **sagittal plane**, though less clinically emphasized, still improves upon the baseline across all metrics.

It is worth noting that our model exhibits lower **specificity** compared to the original configuration. However, in the clinical context of Alzheimer's disease, this trade-off is acceptable—**higher sensitivity is generally more desirable**, as it reduces the risk of false negatives. Missing a diagnosis of Alzheimer's could delay intervention and treatment, which makes a more sensitive model particularly valuable for early detection.

The accuracy curves in Figure 7 demonstrate a stable convergence behavior. The box plots shown in figure 8 higher medians, particularly in the coronal and axial planes, indicating better performance across folds.

3.3 Explainability

In this section, we show how the attention weights are distributed across planes. These visualizations enable the identification of brain regions that are most influential in the model's predictions, thereby enhancing the interpretability and clinical relevance of our results.

During inference, a 3D attention map (saved as a .npy file) that encodes the learned importance values for each voxel. These raw attention maps are then normalized and center-padded to match the dimensions of the anatomical MRI template.

Figure 9 illustrates the mean 3D attention map generated by our model, averaged across subjects and overlaid on the MNI152 standard brain template. The map is derived by multiplicatively combining the normalized attention weights from the axial, coronal, and sagittal prediction networks. The localization of the highlighted areas suggests the model's capacity to pinpoint specific brain parts critical for its diagnostic output. The histograms showing the averaged attention distribution in Figure 10 give us a further insight on the importance density across planes.

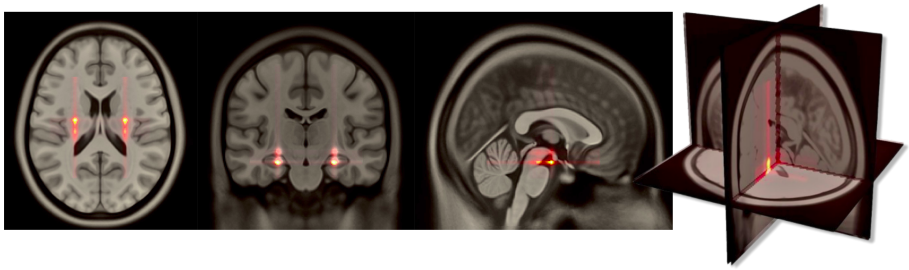


Figure 9: Visualization of mean 3D attention map from the entire dataset overlaid on the MNI152 template using Axial3D (VGG16).

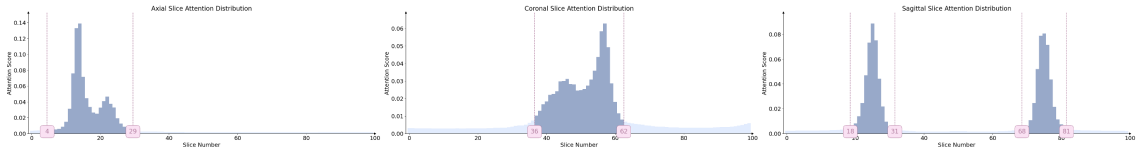


Figure 10: Attention distribution for each plane averaged on all the five test to provide entire dataset distributions

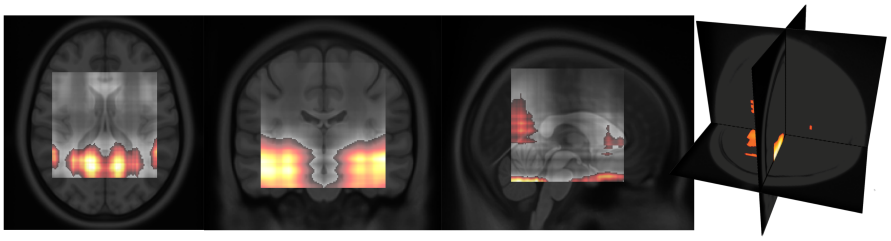


Figure 11: Visualization of mean 3D gradcam++ map from the entire dataset overlaid on the MNI152 template using Axial3D (VGG16).

Gradient-weighted Class Activation Mapping (GradCAM) is a famous visual explanation technique used in CNNs to highlight important regions in an input image that a model focuses on when making a prediction. As we can see in Figure 11, it struggles to pinpoint specific regions affecting the diagnosis outcome.

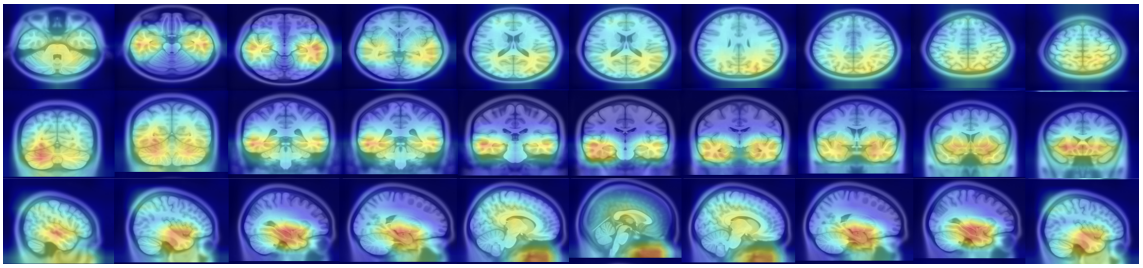


Figure 12: GradCAM++ visualization of 10 slices for each plane selected randomly

Table 3: Attention MAP Metrics by Brain Area

Brain Area	Volume	Attention				% of Region
		Mean	STD	Max	Min	
Hippocampus - right	1059	0.1020	0.1074	0.8053	0.0234	0.2256
Hippocampus - left	997	0.1006	0.1053	0.6655	0.0234	0.2124
Amygdala - left	399	0.1087	0.1094	0.7825	0.0234	0.2418
Lateral Orbitofrontal - left	345	0.0387	0.0139	0.0942	0.0235	0.0229
Amygdala - right	345	0.1076	0.1019	0.7106	0.0234	0.2091
Lateral Orbitofrontal - right	337	0.0382	0.0131	0.0855	0.0234	0.0224
Ventral Diencephalon - left	304	0.0680	0.0372	0.1726	0.0234	0.0498
Inferior Lateral Ventricle - right	297	0.1481	0.1792	0.9082	0.0234	0.2805
Inferior Lateral Ventricle - left	271	0.1578	0.1941	1.0000	0.0234	0.2559
Ventral Diencephalon - right	169	0.0625	0.0328	0.1504	0.0234	0.0277
Fusiform - right	155	0.0376	0.0092	0.0640	0.0234	0.0115
Parahippocampal - left	153	0.0520	0.0294	0.2061	0.0234	0.0565
Fusiform - left	136	0.0350	0.0105	0.0699	0.0234	0.0101
Parahippocampal - right	134	0.0599	0.0347	0.1790	0.0235	0.0495
Cerebellum Gray Matter - left	109	0.0355	0.0077	0.0479	0.0235	0.0015
Pallidum - left	88	0.0531	0.0321	0.1841	0.0239	0.0517
Cerebellum Gray Matter - right	70	0.0326	0.0067	0.0419	0.0234	0.0010
Pallidum - right	66	0.0542	0.0334	0.1672	0.0235	0.0388
Putamen - right	66	0.0639	0.0483	0.2168	0.0235	0.0100
Putamen - left	23	0.0415	0.0214	0.1112	0.0235	0.0035
Insula - right	11	0.0323	0.0080	0.0430	0.0261	0.0012
Superior Temporal - right	3	0.0251	0.0020	0.0279	0.0234	0.0001
Superior Temporal - left	3	0.0242	0.0002	0.0245	0.0240	0.0001

Table 4: GradCam metrics by Brain Area

Brain Area	Volume	Cam				% of Region
		Mean	STD	Max	Min	
Hippocampus - left	1084	0.8250	0.0482	0.9662	0.7529	0.2309
Hippocampus - right	473	0.7944	0.0347	0.8973	0.7529	0.1008
Superior Temporal - right	395	0.8465	0.0656	1.0000	0.7532	0.0156
Inferior Lateral Ventricle - left	306	0.7910	0.0254	0.8749	0.7530	0.2890
Superior Temporal - left	252	0.8128	0.0472	0.9443	0.7533	0.0099
Inferior Lateral Ventricle - right	181	0.7902	0.0262	0.8557	0.7529	0.1709
Middle Temporal - left	162	0.8421	0.0523	0.9468	0.7535	0.0057
Middle Temporal - right	157	0.8552	0.0479	0.9827	0.7576	0.0055
Inferior temporal - left	130	0.8067	0.0368	0.9023	0.7531	0.0080
Amygdala - right	110	0.7669	0.0116	0.8369	0.7529	0.0667
Fusiform - left	104	0.8013	0.0334	0.8932	0.7530	0.0077
Insula - right	73	0.8059	0.0414	0.9102	0.7529	0.0077
Putamen - right	54	0.8021	0.0298	0.8753	0.7563	0.0081
Entorhinal - right	8	0.7590	0.0043	0.7675	0.7547	0.0025
Ventral Diencephalon - right	7	0.7650	0.0048	0.7717	0.7579	0.0011
Parahippocampal - left	5	0.7657	0.0099	0.7827	0.7541	0.0018
Ventral Diencephalon - left	5	0.7732	0.0072	0.7837	0.7613	0.0008
Putamen - left	4	0.7590	0.0048	0.7670	0.7547	0.0006
Amygdala - left	3	0.7586	0.0066	0.7679	0.7532	0.0018
Inferior temporal - right	3	0.7573	0.0043	0.7633	0.7532	0.0002
Insula - left	2	0.7558	0.0010	0.7568	0.7548	0.0002
Fusiform - right	1	0.7601	0.0000	0.7601	0.7601	0.0001

4 Conclusion

In this work, we replicated the results of *"AXIAL: Attention-based eXplainability for Interpretable Alzheimer's Localized Diagnosis using 2D CNNs on 3D MRI brain scans"*. We did not manage to replicate the same results by using the same configuration and random seed provided by the author; this could be due to some struggles we had during the initial dataset preparation. The *Alzheimer's Disease Neuroimaging Initiative (ADNI)* platform is going through some important changes, leading to the deprecation of the current version of the *Clinica* software that enables the data preprocessing of the raw dataset.

Anyhow, we were able to showcase **improved diagnostic metrics** through refined hyperparameter tuning and methodological adjustments.

Key achievements include **superior accuracy**, **sensitivity**, and **MCC** values compared to the original *AXIAL* paper, especially in the clinically vital **coronal** and **axial** planes. For instance, the coronal plane accuracy reached **85.30%** with **91.02% sensitivity**. These improvements were realized despite the challenges of a relatively small dataset and model complexity, which were managed by strategies like **increased cross-validation folds** and **optimized slice selection**.

We preferred a trade-off of *lower specificity* for *higher sensitivity*, since it's clinically preferable to minimize **false negatives** in the early detection of Alzheimer's Disease.

We have integrated **explainable AI (XAI)**, generating **3D heat maps** that highlight *AD-indicative brain regions* like the **hippocampus** and **amygdala**. We believe that this aspect might be crucial for *clinical relevance* and *model transparency*. In our future work, we could potentially expand on these methodologies to **larger datasets** and improve the data preprocessing pipeline by selecting slices based on their **importance**, instead of slicing a block from the middle of the brain.

References

[LBF+24] Gabriele Lozupone, Alessandro Bria, Francesco Fontanella, Frederick JA Meijer, and Claudio De Stefano. Axial: Attention-based explainability for interpretable alzheimer's localized diagnosis using 2d cnns on 3d mri brain scans. *arXiv preprint arXiv:2407.02418*, 2024.

[RBD+21] Alexandre Routier, Ninon Burgos, Mauricio Díaz, Michael Bacci, Simona Bottani, Omar El-Rifai, Sabrina Fontanella, Pietro Gori, Jérémy Guillon, Alexis Guyot, Ravi Hassanaly, Thomas Jacquemont, Pascal Lu, Arnaud Marcoux, Tristan Moreau, Jorge Samper-González, Marc Teichmann, Elina Thibeau-Sutre, Ghislain Vaillant, Junhao Wen, Adam Wild, Marie-Odile Habert, Stanley Durrleman, and Olivier Colliot. Clinica: An open-source software platform for reproducible clinical neuroscience studies. *Frontiers in Neuroinformatics*, 15, 2021.