Alzheimer's Diagnosis using 2D CNNs on 3D MRI brain scans and Explainable Artificial Intelligence

Politecnico di Milano
Davide Villani, Filippo Wang

Abstract

Diagnosing Alzheimer's Disease (AD) accurately and interpretably from brain MRI scans is still a key challenge in medical AI. In this work, we propose 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". We begin with the preprocessing of the MRI scans, followed by decomposition of the 3D brain volumes into 2D slices along the sagittal, axial and coronal planes. For each anatomical axis, we apply a 2D CNN to extract meaningful features from the slices, which are then aggregated into feature vectors. These vectors are passed through a fully connected neural network with softmax activation to derive axis-specific attention weights. Subsequently, these weights are used to compute both an attention-based 3D matrix of the brain and a composite weighted feature vector that captures the most informative characteristics across all three views. This final representation is passed through another fully connected layer and a softmax classifier to produce the final diagnosis. Using attention mechanisms and XAI, our approach not only achieves high diagnostic accuracy, but also provides interpretable insights into the regions of the brain most indicative of Alzheimer's disease. / ... /

1 Dataset

The ADNI1: Complete 1Yr 1.5T dataset 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, also 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.

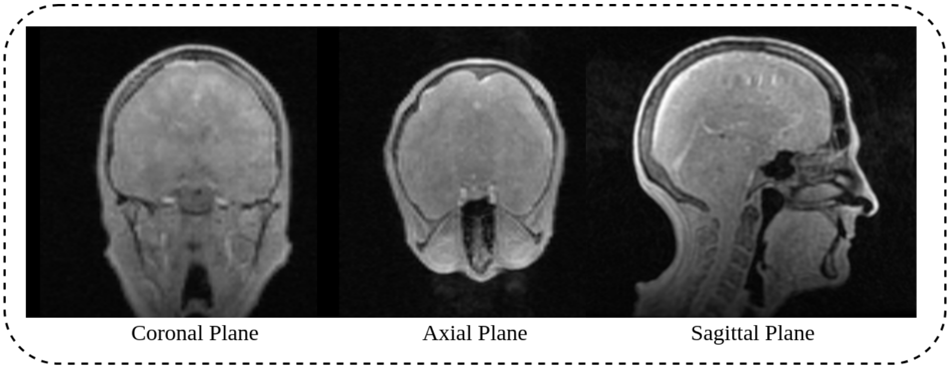

Figure 1: Example of Standard anatomical plane's slices before Pre-Processing

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

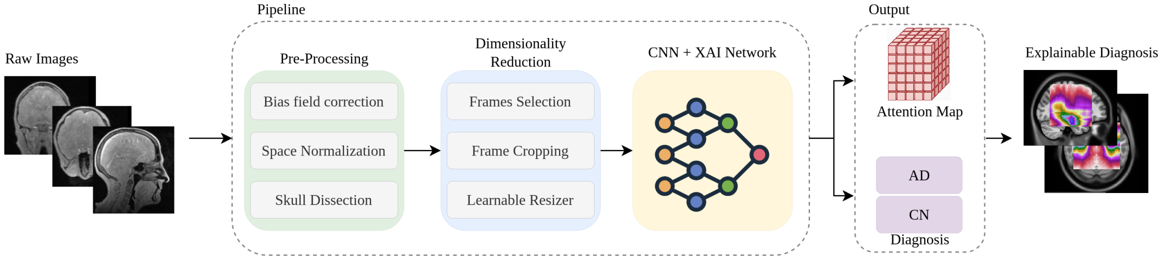

Figure 2: Pipeline highlighting the pre-processing steps

2.1 Pre-Processing

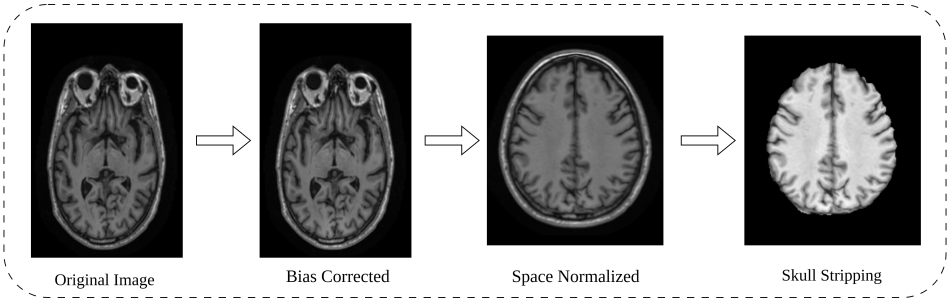

Figure 3: Example of a pre processing pipeline on a single slice

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 Dimensionality Reduction

/ ... /

2.3 Feature Extraction

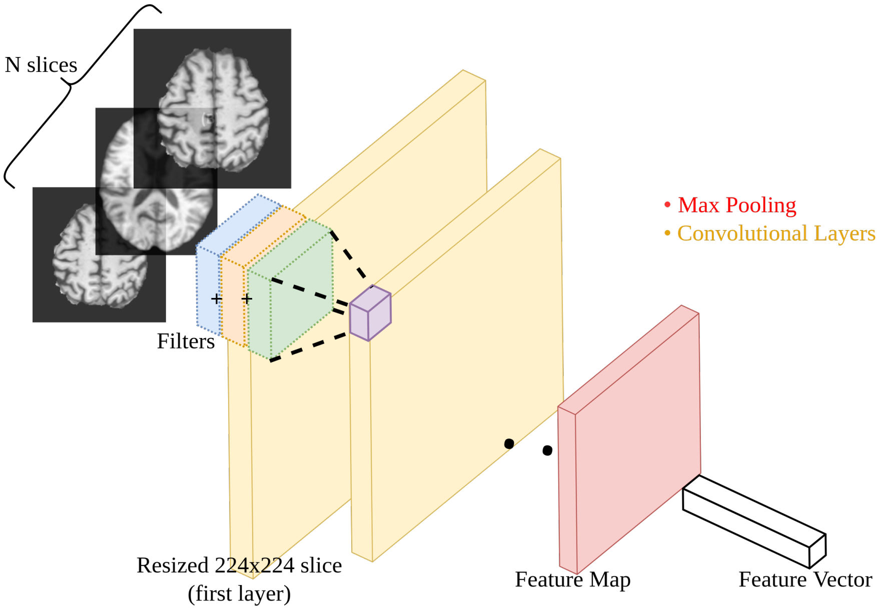

Figure 4: Convolutional Backbone with shared weights, this is applied to all the planes separately

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:

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

Each slice x_i is a single-channel (grayscale) image which contrasts with standard 2D convolutional backbones such as *VGG* or *ResNet*, 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:

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

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

2.4 Attention XAI Fusion Module

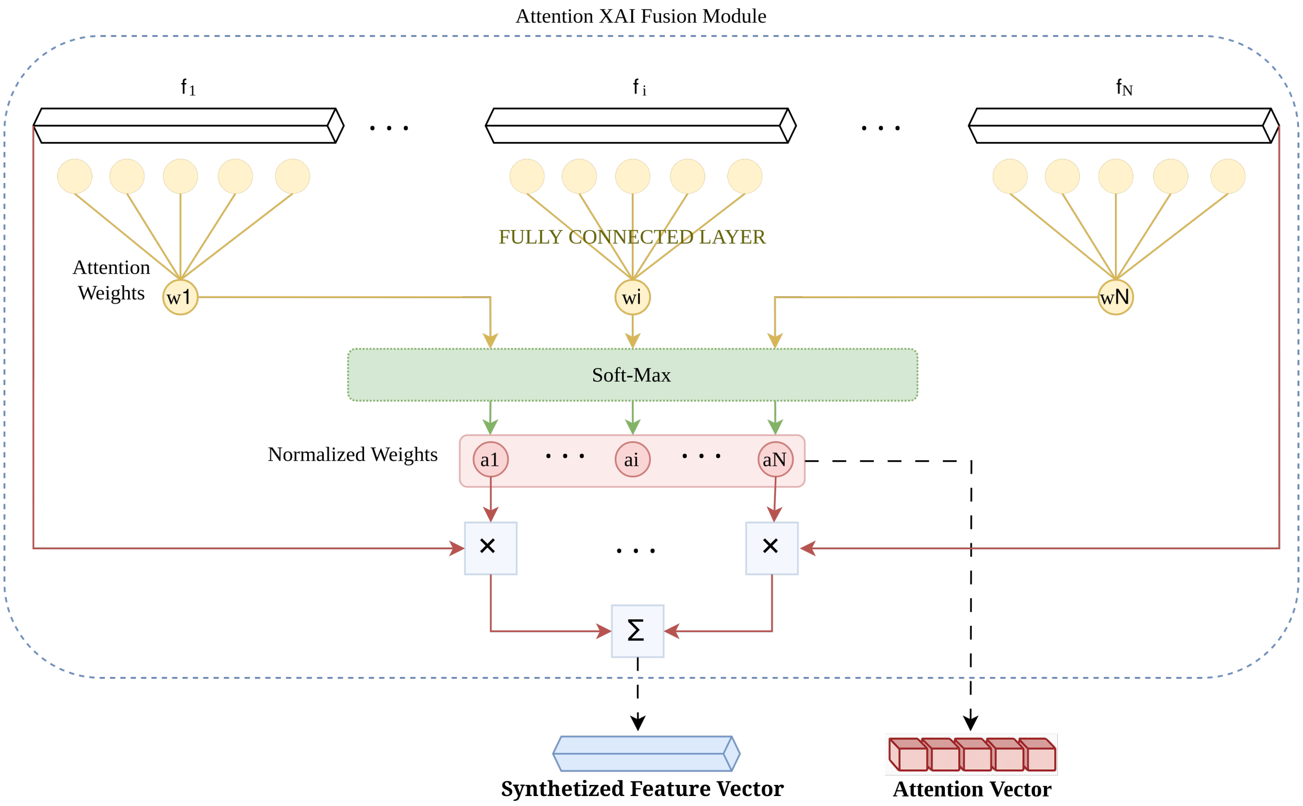

Figure 5: Attention XAI Fusion Module scheme applied to each plane slices

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 $w_i \in \mathbb{R}$:

$$w_i = \mathrm{FC}_{\mathrm{att}}(f_i) = \mathbf{w}^\top f_i + b$$

where $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are learnable parameters. 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.

2.5 Diagnosis

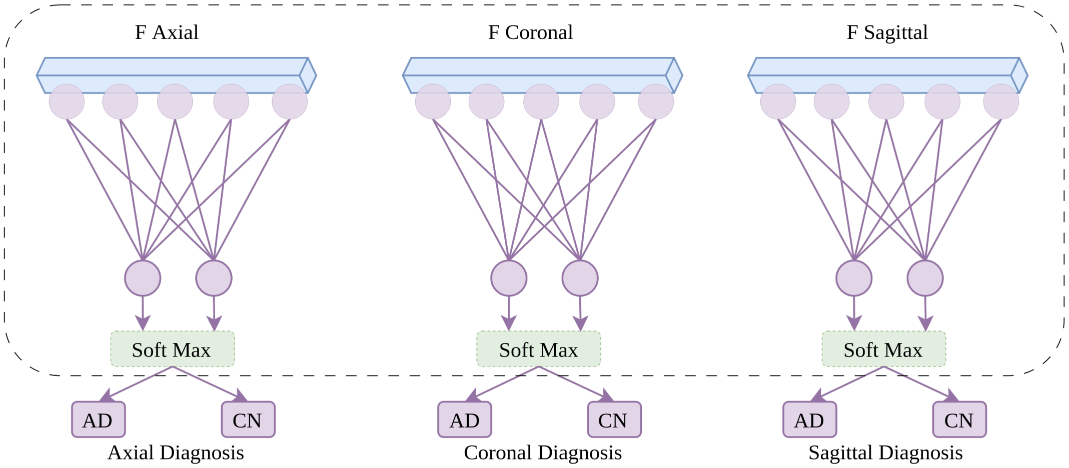

Figure 6: Diagnosis Module scheme applied to each plane

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 linear layer with weights $\mathbf{W} \in \mathbb{R}^{2\times d}$ and bias $\mathbf{b} \in \mathbb{R}^2$. The output $\mathbf{z} \in \mathbb{R}^2$ is given by:

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

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.

2.6 Heat Map Generation

3 Results

References