UNIVERSITÀ DEGLI STUDI DI CASSINO E DEL LAZIO MERIDIONALE

CORSO DI DOTTORATO IN METODI, MODELLI E TECNOLOGIE PER L'INGEGNERIA

DIPARTIMENTO DI INGEGNERIA ELETTRICA E DELL'INFORMAZIONE

Generative and Explainable Deep Learning for Alzheimer's Disease Analysis

Gabriele Lozupone

gabriele.lozupone@unicas.it

In Partial Fulfillment of the Requirements for the Degree of

PHILOSOPHIAE DOCTOR in

Electrical and Information Engineering

15/04/2026

TUTOR
Prof. Alessandro Bria
Prof. Francesco Fontanella
Prof. Claudio De Stefano

COORDINATOR
Prof. Fabrizio Marignetti

UNIVERSITÀ DEGLI STUDI DI CASSINO E DEL LAZIO MERIDIONALE

CORSO DI DOTTORATO IN METODI, MODELLI E TECNOLOGIE PER L'INGEGNERIA

Date: **15/04/2026**

Author: **Gabriele Lozupone**

Title: **Generative and Explainable Deep Learning for Alzheimer's Disease Analysis**

Department: **DIPARTIMENTO DI INGEGNERIA ELETTRICA E DELL'INFORMAZIONE**

Degree: **PHILOSOPHIAE DOCTOR**

Gabriele Lozupone

Signature of Author

Summary

Alzheimer's disease (AD) is a growing clinical challenge that demands earlier and more reliable computational biomarkers. Although recent deep learning systems have reported promising results in AD analysis, their clinical translation remains constrained by three recurring limitations: insufficient interpretability, limited generalizability across cohorts, and the scarcity of labelled medical data. This thesis investigates these issues across two complementary diagnostic modalities, handwriting and neuroimaging, with the goal of developing methods that are clinically aligned, methodologically rigorous, and transferable.

The first modality considered is handwriting, explored as a preliminary and non-invasive source of diagnostic information for AD. A unified multi-task framework is developed to learn shared patterns across 25 heterogeneous writing tasks acquired from scanned paper sheets in a dataset specifically designed for diagnosis. The results show that a single model can support automated AD discrimination with promising accuracy and that transfer learning from large-scale natural image datasets remains effective even in this underexplored setting. At the same time, this contribution should be interpreted as a first step: future studies will require richer datasets that better control relevant confounders, such as arthritis and other age-related motor impairments, and longitudinal data that enable investigation of disease onset and progression rather than diagnosis alone.

The second modality is structural MRI, where the thesis addresses the core challenges more directly. First, the AXIAL framework tackles interpretability within a supervised classification setting through multi-plane attention fusion and both quantitative and qualitative explainability validation. Rather than relying on post-hoc visualisation methods, AXIAL embeds attention into the diagnostic pipeline itself and produces fold-consistent 3D attention maps that align with established AD-related neuroanatomical markers, including hippocampal atrophy, medial temporal lobe degeneration, and ventricular enlargement. AXIAL shows that interpretability and diagnostic performance are not mutu-

ally exclusive, and that parameter-efficient 2D/2.5D designs can recover clinically meaningful volumetric information under limited-data conditions.

However, AXIAL still depends on labelled cohorts and task-specific supervision. To address the remaining limitations of annotation scarcity and cross-dataset transfer, the thesis then moves from supervised classification to unsupervised representation learning. This progression culminates in the Latent Diffusion Autoencoder (LDAE), a generative foundation model that performs diffusion in a compressed latent space. By learning from unlabeled MRI data, LDAE decouples representation learning from downstream supervision and enables zero-shot transfer to unseen datasets. A single pre-training stage yields a reusable representation that supports multiple downstream tasks, including diagnosis, age prediction, temporal interpolation, and counterfactual generation, while remaining substantially more computationally efficient than voxel-space diffusion autoencoders.

Taken together, these contributions support three main claims. First, non-invasive modalities such as handwriting deserve further investigation as potentially useful sources of AD-related information, but stronger evidence will require better-controlled and longitudinal cohorts. Second, interpretability and methodological rigor are prerequisites for clinical trust, often more important than marginal improvements in benchmark performance. Third, diffusion-based unsupervised learning provides a practical route toward richer and more transferable representations when labelled data are limited, while also enabling generative capabilities beyond purely discriminative models. Overall, the thesis argues that AI systems for Alzheimer's disease should prioritize clinical alignment, transparency, and representation quality over isolated performance gains.

Contents

List of Figures

List of Tables

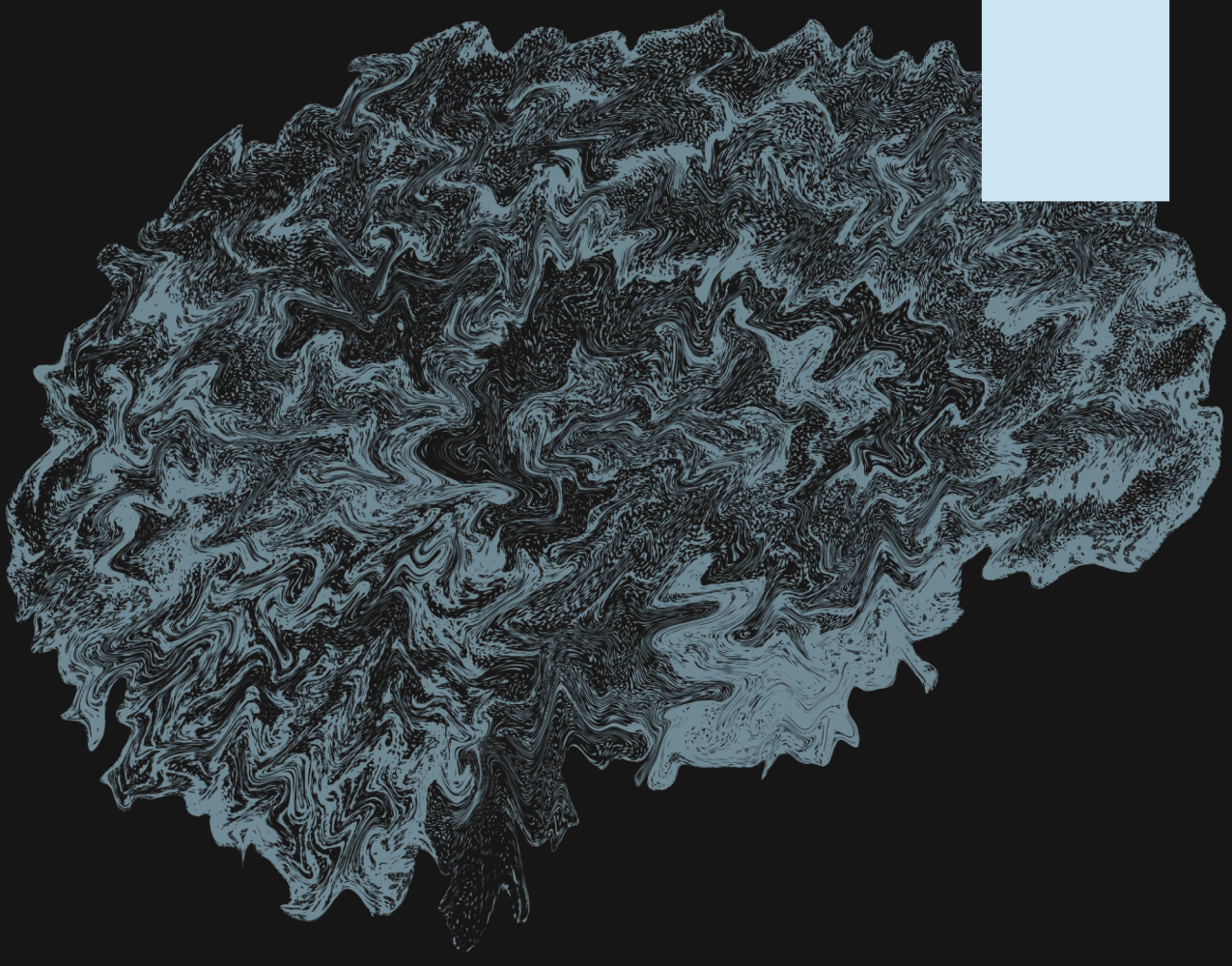
Acronyms

2D	Two-Dimensional.
3D	Three-Dimensional.
ACC	Accuracy.
AD	Alzheimer's Disease.
AdaGN	Adaptive Group-wise Normalization.
ADNI	Alzheimer's Disease Neuroimaging Initiative.
AE	Autoencoder.
AI	Artificial Intelligence.
AIBL	Australian Imaging, Biomarkers and Lifestyle.
ANTs	Advanced Normalization Tools.
ASHS	Automatic Segmentation of Hippocampal Subfields.
AUROC	Area Under the Receiver Operating Characteristic.
BIDS	Brain Imaging Data Structure.
BSI	Boundary Shift Integral.
CN	Cognitive Normal.
CNN	Convolutional Neural Network.
CSF	Cerebrospinal Fluid.
DAE	Diffusion Autoencoder.
DDIM	Denoising Diffusion Implicit Model.
DDPM	Denoising Diffusion Probabilistic Model.
DL	Deep Learning.
DLADS	Deep Learning-Aided Diagnostic Systems.
DPM	Diffusion Probabilistic Model.
DSM-5	Diagnostic and Statistical Manual of Mental Disorders, 5th ed..

EMA	Exponential Moving Average.
EMCI	Early Mild Cognitive Impairment.
FAB	Frontal Assessment Battery.
FC	Fully Connected.
FLAIR	Fluid-Attenuated Inversion Recovery.
FSL	FMRIB Software Library.
GAN	Generative Adversarial Network.
GFNet	Global Filter Network.
GradCAM	Gradient-weighted Class Activation Mapping.
IAM	IAM Handwriting Database.
KL	Kullback-Leibler.
LDA	Linear Discriminant Analysis.
LDAE	Latent Diffusion Autoencoder.
LDM	Latent Diffusion Model.
LMCI	Late Mild Cognitive Impairment.
LPIPS	Learned Perceptual Image Patch Similarity.
MAE	Mean Absolute Error.
MCC	Matthews Correlation Coefficient.
MCI	Mild Cognitive Impairment.
ML	Machine Learning.
MLP	Multi-Layer Perceptron.
MMSE	Mini-Mental State Examination.
MNI	Montreal Neurological Institute.
MoCA	Montreal Cognitive Assessment.
MP-RAGE	Magnetization Prepared Rapid Gradient Echo.
MRI	Magnetic Resonance Imaging.
MSE	Mean Squared Error.
MTA	Medial Temporal Atrophy.
MTL	Medial Temporal Lobe.
NIA-AA	National Institute on Aging-Alzheimer's Association.
OASIS	Open Access Series of Imaging Studies.
OCR	Optical Character Recognition.

PD	Parkinson's Disease.
PDAE	Pretrained Diffusion Autoencoder.
PET	Positron Emission Tomography.
pMCI	Progressive Mild Cognitive Impairment.
ResNet	Residual Network.
RMSE	Root Mean Square Error.
ROI	Region of Interest.
SEN	Sensitivity.
SGD	Stochastic Gradient Descent.
sMCI	Stable Mild Cognitive Impairment.
SNR	Signal-to-Noise Ratio.
SPE	Specificity.
SSIM	Structural Similarity Index Measure.
UNet	U-Net.
VAE	Variational Autoencoder.
VGG	Visual Geometry Group.
ViT	Vision Transformer.
XAI	Explainable Artificial Intelligence.

1



INTRODUCTION

Chapter 1

Introduction

Alzheimer's Disease (AD) is the most common cause of dementia, accounting for approximately 60–80% of dementia cases worldwide [4]. It represents one of the most pressing public health challenges of the 21st century. In 2020, more than 55 million individuals worldwide were living with dementia, a number projected to rise to over 139 million by 2050 due to population aging [113]. In the United States alone, an estimated 7.2 million individuals aged 65 years or older currently live with Alzheimer's dementia, with projections indicating a near doubling by 2060 in the absence of effective disease-modifying therapies [4]. AD is also a leading cause of mortality. In 2020 it was ranked as the seventh-leading cause of death in the United States, however recent data indicate that Alzheimer's will likely resume its place as the sixth-leading cause of death [4]. Unlike cardiovascular diseases and stroke, whose mortality rates have declined over recent decades, deaths attributed to AD increased by more than 142% between 2000 and 2019 [4]. It is estimated that one in three older adults dies with AD or another form of dementia, underscoring the profound clinical and societal burden of neurodegenerative disorders.

Neuroimaging plays a pivotal role in the diagnosis and study of AD. The disease is characterized by progressive structural and functional brain alterations, including cortical thinning, hippocampal atrophy, ventricular enlargement, and metabolic changes detectable through imaging modalities such as Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) [72, 79]. Structural MRI, in particular, provides non-invasive access to morphological biomarkers associated with disease progression and is widely used in both clinical and research settings. However, identifying imaging signs of early-stage AD remains challenging for human experts, motivating the growing interest in Artificial Intelligence (AI)–based approaches.

Over the past decade, deep learning has emerged as a dominant paradigm in medical image analysis and has been extensively applied to AD detection and prognosis using neuroimaging data [38]. Before its widespread adoption, most neuroimaging studies on AD relied on more conventional machine learning pipelines, in which morphometric, voxel-wise, or region-based measurements were extracted from MRI scans and then analyzed with standard classification algorithms [42, 54]. Convolutional Neural Network (CNN)s are well suited to this task because they can learn hierarchical spatial representations directly from raw imaging data. Compared with these standard feature-engineering approaches, deep learning models have often shown stronger performance in capturing subtle and spatially extended patterns of neurodegeneration. Numerous studies have shown that CNN-based models can achieve high accuracy in distinguishing AD patients from cognitively normal controls using structural MRI, with reported accuracies often exceeding 90% under controlled experimental conditions [79, 38]. In prognostic settings, deep learning models have also been applied to predict the conversion from Mild Cognitive Impairment (MCI) to Alzheimer's dementia, achieving accuracies in the range of 80–85% [79]. These results suggest that AI-based systems can approach, and in some cases surpass, expert-level performance in specific diagnostic tasks. In most cases, these results have been reported on large, well-curated research cohorts, most notably the Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI provides longitudinal MRI, PET, and clinical data for thousands of subjects across the cognitive aging spectrum and has become the de facto benchmark for AI-based AD research [106]. Other commonly used datasets include the Open Access Series of Imaging Studies (OASIS), which offers cross-sectional and longitudinal MRI data covering normal aging and dementia [101]. Models trained on these datasets frequently report high classification metrics, particularly for binary AD versus control tasks. Recent trends include the use of transfer learning, whereby deep neural networks pretrained on large-scale natural image datasets or related neuroimaging tasks are fine-tuned for AD classification [155]. Additionally, transformer-based architectures and hybrid CNN–transformer models are increasingly explored to capture long-range spatial dependencies in three-dimensional brain images. While these approaches often yield incremental performance gains, they also introduce increased model complexity and data requirements.

Despite positive results in research environments, the translation of AI-based AD imaging systems into routine clinical practice remains limited. To date, no deep learning system has been widely adopted as a standalone diagnostic tool for Alzheimer's disease in clinical neurology or radiology workflows [103]. One major concern is the discrepancy between reported research performance and real-world generalizability. Several recent reviews have highlighted that many AI models for AD imaging suffer from methodological limitations, including data leakage, improper cross-validation strategies, and lack of independent external validation [171, 183]. In some cases, models initially reported to

achieve accuracies above 90% experienced dramatic performance drops when evaluated under stricter experimental protocols or on independent cohorts. This suggests that a substantial portion of reported gains may reflect overfitting to specific datasets rather than genuine disease-related signal extraction. Moreover, fewer than 20% of published deep learning models for AD imaging have been validated on external datasets, and virtually none have undergone prospective clinical testing [183]. Differences in scanner hardware, acquisition protocols, population demographics, and diagnostic criteria across clinical centers further complicate generalization. As a result, clinicians remain cautious about integrating AI predictions into diagnostic decision-making, particularly for early-stage or borderline cases.

Another major barrier to clinical adoption of AI systems in Alzheimer's disease is the lack of interpretability. Most deep learning models operate as black boxes, providing predictions without transparent reasoning. In medical contexts, particularly for neurodegenerative diseases with complex and heterogeneous presentations, clinicians require explanations that align with established neurobiological knowledge [103]. Post-hoc explainability techniques, such as gradient-based saliency maps and class activation mapping, have been widely applied to visualize regions of interest contributing to model predictions [141]. While these methods often highlight anatomically relevant areas, including the hippocampus and medial temporal lobe, their clinical reliability and consistency remain debated [145, 161]. Importantly, most explainability studies rely on qualitative visual inspection rather than quantitative validation or expert consensus, limiting their translational value. In addition to interpretability, data-related challenges pose significant obstacles. Neuroimaging datasets for AD are relatively small compared to those typically required to train deep neural networks, increasing susceptibility to overfitting. Class imbalance, site-specific biases, and heterogeneous disease trajectories further complicate model development. Although recent studies show improved methodological rigor compared to earlier work, reproducibility and robustness remain central concerns [171].

As performance on standard AD classification benchmarks approaches saturation, research focus is shifting toward more clinically meaningful tasks, such as multi-class classification, differential diagnosis among dementia subtypes, and longitudinal disease modeling. These tasks present greater challenges and more closely reflect real clinical decision-making. Generative models, including Generative Adversarial Network (GAN), Variational Autoencoder (VAE), and Denoising Diffusion Probabilistic Model (DDPM), have gained increasing attention in AD research. These models have been employed for data augmentation, modality synthesis, and disease progression modeling [1]. Synthetic MRI or PET images improve classifier performance and mitigate class imbalance. Furthermore, generative frameworks enable the simulation of future brain states, offering an additional option for personalized prognosis. Unsupervised and self-supervised gen-

erative approaches have also been explored for anomaly detection, where models learn the distribution of healthy brain images and identify pathological deviations without explicit diagnostic labels [60]. These approaches are particularly appealing for early-stage detection, where labeled data are scarce and disease manifestations are subtle.

Generative models offer useful approaches to data scarcity, representation learning, and disease modeling, while explainable AI methods aim to bridge the gap between algorithmic predictions and clinical reasoning. Integrating these two paradigms, generative modeling and explainable deep learning, represents an important research direction toward trustworthy, clinically meaningful AI systems for AD analysis.

1.1 Clinical Overview of Alzheimer's Disease

AD is marked by the gradual and unrelenting degeneration of neurons throughout the brain. Clinically, it is defined by a steadily worsening cognitive impairment that eventually compromises independence in daily life [91]. The disease process typically unfolds over years: early in the course, changes are subtle and may not impair daily activities, but AD inexorably advances from mild forgetfulness to severe dementia. The underlying pathology, including neuron loss, synaptic dysfunction, and the accumulation of toxic proteins, relentlessly progresses, meaning that once neuronal damage occurs it cannot be reversed with current therapies. This biological irreversibility translates to the clinical reality that lost cognitive functions (memory, reasoning, etc.) cannot be fully regained [94, 4]. The clinical trajectory of AD is often conceptualized as a continuum: it begins with a long silent phase of brain changes (preclinical AD), then a symptomatic but mild stage (mild cognitive impairment), and ultimately culminates in dementia due to AD. Throughout this continuum, AD pathology advances from region to region in the brain, and symptoms worsen accordingly (as detailed in later sections). From a patient and caregiver perspective, AD's progressive nature means increasing assistance is required over time, from help with complex tasks like finances or medication, to eventually basic self-care, and the condition is uniformly fatal in the long run (typically 8-12 years after diagnosis, though with wide individual variability) [157]. AD's progressive, irreversible course is what makes it so devastating, but also what motivates intense research into early diagnostic markers and disease-slowing treatments.

For convenience, Table 1.1 summarizes the main stages of the AD continuum by linking their typical clinical manifestations with the brain regions and imaging findings most commonly involved. The temporal progression should be interpreted qualitatively, as the duration of each stage varies substantially across individuals. The table also highlights

Table 1.1: Schematic overview of the AD continuum, linking the main clinical stages with representative symptoms, functional impact, and the neuroanatomical regions or imaging correlates most commonly involved.

Stage	Dominant clinical features	Main regions / imaging correlates
Preclinical AD *Silent phase before overt symptoms*	No clear functional impairment; individuals may be clinically normal despite underlying pathology. Biomarker abnormalities may precede overt symptoms by years [152, 73].	Molecular biomarkers may become abnormal before overt structural damage, whereas MRI abnormalities may still be minimal or absent in this phase [70, 72].
MCI due to AD *Earliest clearly symptomatic phase*	Mild but measurable cognitive decline, typically with prominent episodic memory impairment, while basic daily functioning is largely preserved. Subtle fine-motor changes, including reduced fluency in drawing and handwriting tasks, may already emerge in this phase [2, 75, 120, 29, 179].	Early medial temporal involvement, especially hippocampal atrophy on structural MRI, with prognostic value for progression to dementia [139, 129, 28].
AD dementia *Clinically established disease*	Multidomain cognitive impairment involving memory, language, executive and visuospatial functions; neuropsychiatric symptoms become more common, and loss of independence becomes clinically evident. Fine-motor and handwriting impairments may also become more apparent as cognitive and motor planning deficits progress [73, 136, 22, 100, 27, 29].	More widespread medial temporal and temporoparietal atrophy, often accompanied by ventricular enlargement and progressive structural decline on MRI [139, 72, 46].

that subtle fine-motor and handwriting-related alterations can emerge during symptomatic stages, providing part of the clinical rationale for the handwriting-based biomarkers investigated in Chapter 2.

1.1.1 Clinical Continuum: From Cognitively Normal to MCI to AD Dementia

One of the crucial concepts in understanding AD is its clinical continuum, which spans from normal cognition through mild impairment to severe dementia. The clinical manifestation of AD typically follows a gradual progression through identifiable transitional stages rather than presenting as an acute onset condition. A typical pathway begins with a Cognitive Normal (CN) state in late life, followed by the development of Mild Cognitive Impairment (MCI), and finally Alzheimer's-type dementia [54]. MCI is defined by measurable cognitive decline (for example, noticeable memory loss) that is greater than expected for age, but not severe enough to significantly interfere with daily functioning. It is often considered the prodromal phase of AD when due to AD pathology. In fact,

MCI is often described as the earliest clinical manifestation of AD in individuals who later progress to dementia [118, 2]. In research settings, particularly in ADNI-based studies, amnestic MCI has also been subdivided into Early Mild Cognitive Impairment (EMCI) and Late Mild Cognitive Impairment (LMCI) to capture differences in memory impairment severity and progression risk [95]. EMCI denotes a milder level of objective memory impairment, whereas LMCI reflects more pronounced impairment and is associated with a less favorable clinical course [95]. Patients with amnestic MCI (where memory is chiefly affected) are at elevated risk of progressing to AD dementia, and this stage can be viewed as an early symptomatic phase of AD [118, 2]. Importantly, however, not everyone with MCI will progress to dementia. Some remain stable for years or even revert to normal, especially when the underlying cause is not AD [118]. Nevertheless, in clinical settings about 10-15% of MCI patients convert to dementia each year [43]. Over a 5-year period, roughly half of individuals diagnosed with MCI due to AD may worsen to an AD dementia, whereas others may decline more slowly or not at all. This variability makes it critical to determine which MCI patients have underlying AD pathology and are likely to progress (progressive MCI vs. stable MCI). Clinicians and researchers stratify MCI cases accordingly, since prognostic outlook and management differ: an MCI patient who is progressive (showing early AD changes on imaging or biomarkers) might be a candidate for trials or aggressive intervention, whereas a stable MCI patient might be monitored with conservative management. Identifying prognostic factors for progression is a major goal, for instance, the presence of AD biomarkers or faster hippocampal atrophy on MRI strongly predicts conversion from MCI to Alzheimer's dementia. In summary, AD can be viewed as a continuum where initially a person is cognitively normal (but possibly accumulating silent pathology), then develops mild cognitive symptoms (MCI), and finally meets criteria for dementia. Figure 1.1 schematically illustrates this continuum across four progressive stages: CN, EMCI, LMCI, and AD dementia. Notable structural changes include progressive hippocampal atrophy and ventricular enlargement. Clinically, this staging aligns with how symptoms and functional impairments progressively worsen over time.

1.1.2 Clinical Symptoms and Manifestations

The hallmark symptom of AD is memory loss, especially difficulty forming new memories and recalling recent information (episodic memory impairment) [75]. Often an early complaint is forgetting conversations or repeating questions. This memory deficit typically reflects damage in the medial temporal lobe (e.g. hippocampus), which is crucial for new learning [58]. As AD advances, other cognitive domains become affected, including executive functions (planning, problem-solving) [136, 142], language (word-finding diffi-

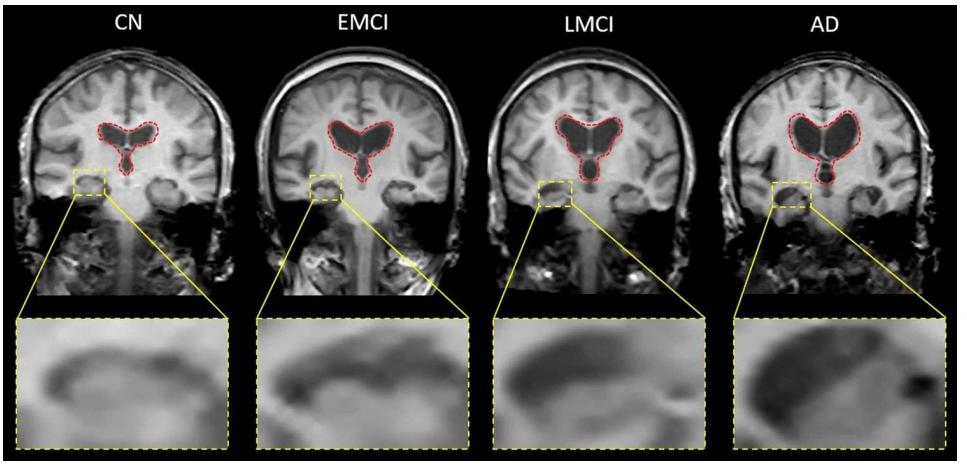


Figure 1.1: Coronal T1-weighted MRI slices across the AD continuum: CN, EMCI, LMCI, and AD. Yellow rectangles highlight hippocampal regions; red regions show lateral ventricles. Reprinted from [54].

culty, reduced vocabulary) [136], visuospatial skills (getting disoriented in familiar places, trouble copying figures) [128], and attention. In early stages, the cognitive deficits can be subtle, for instance, mild word-finding pauses or trouble multitasking, but in later stages they culminate in global impairment (in severe AD, patients may recognize neither close family nor their own surroundings) [136]. It is important to note that the pattern of cognitive decline in AD usually involves memory and orientation first (patients may become disoriented to date or place as the disease progresses) [136]. Over time, executive dysfunction (poor judgment, inability to plan or manage complex tasks) and language impairments (anomic aphasia, wherein naming common objects becomes difficult) also emerge as common features [136]. Each individual's symptom profile can vary, e.g. some patients present with prominent language problems (logopenic aphasia) or visuospatial disorientation (getting lost), but nearly all experience significant memory loss at some point. These cognitive symptoms correspond to the underlying neurodegeneration pattern, e.g. memory problems correlate with hippocampal damage [58].

1.1.3 Behavioral and Neuropsychiatric Symptoms

In addition to cognitive deficits, AD produces a range of behavioral and psychological changes, often referred to as neuropsychiatric symptoms. One of the most frequent is apathy, a loss of motivation and initiative. In fact, apathy is observed in roughly 60-70%

of AD patients and often appears early in the disease [33]. A person with AD may become indifferent to hobbies, require prompting to engage in activities, or sit idle for long periods. Notably, apathy can hide functional decline, as the individual may not initiate even simple tasks without cueing. Depression is another common symptom and can be difficult to disentangle from apathy; many patients with AD have apathy without classic depressive sadness [100]. Other behavioral changes include irritability, mood swings, anxiety, and in some cases disinhibition or agitation (especially in moderate stages) [22]. As AD reaches more advanced stages, psychotic symptoms such as delusions or hallucinations can occur (e.g. paranoid beliefs that someone is stealing items, or seeing people who aren't there) [22]. Personality often changes as well, a previously meticulous person might become careless or a gregarious person might withdraw socially [100]. These behavioral symptoms add to caregiver burden and are important clinical considerations. They also provide clues to clinicians; for example, early prominent personality change might suggest frontotemporal dementia rather than typical AD, whereas AD's behavioral changes tend to occur a bit later and be accompanied by the hallmark memory loss.

1.1.4 Motor and Fine-Motor Changes (Including Handwriting)

AD has traditionally been viewed primarily as a cognitive disorder, unlike Parkinson's Disease (PD) which is known for motor symptoms. Indeed, gross motor functions (walking, basic strength) tend to be preserved until relatively late in AD. However, research has shown that subtle motor deficits can occur earlier in the AD process than once believed [120]. Older clinical criteria expected motor signs only in advanced dementia, but it is known that certain motor changes, for example, reduced gait speed, decreased grip strength, or slight coordination difficulties, may be detectable in the MCI or early dementia stages and are even associated with higher risk of progression to dementia [120, 23]. In particular, fine-motor control (which requires complex coordination of cognitive planning with motor execution) can be impaired. Handwriting is an excellent example of a fine-motor skill that often deteriorates in AD and related disorders [29, 179]. In fact, handwriting and drawing tests have gained attention as sensitive markers of early cognitive impairment.

An example of a widely used clinical assessment is the Clock Drawing Test, which requires patients to draw a clock face with all numbers and set the hands to a specific time (typically 10 minutes past 11). This simple task engages multiple cognitive domains simultaneously: visuospatial organization, executive planning, semantic memory (knowledge of clock conventions), and fine-motor control [52]. Figure 1.2 illustrates typical clock drawings from cognitively normal individuals compared to those with AD or fron-

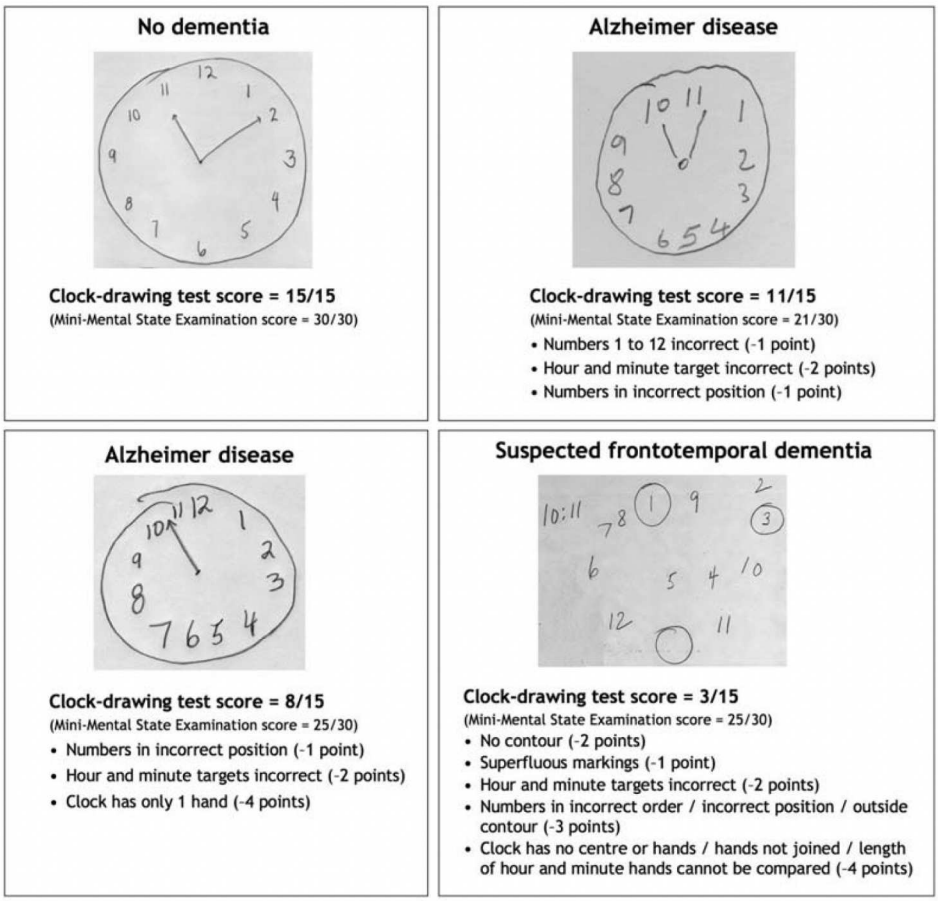


Figure 1.2: Clock Drawing Test examples comparing cognitively normal individuals with patients diagnosed with Alzheimer's disease or frontotemporal dementia. Patients were instructed to draw a clock face with all numbers and set the time to 10 minutes past 11. Reprinted from [44].

totemporal dementia, demonstrating characteristic impairments such as number crowding, spatial disorganization, incorrect time representation, and distorted clock shapes. The test's sensitivity to early cognitive decline makes it a valuable screening tool in clinical practice. Changes in handwriting in AD are multifaceted. Patients may show a slower writing speed, lower pen pressure, more hesitations, and irregular stroke size or spacing. Research indicates that AD patients produce handwriting that is less automatic, slower, and more variable than that of age-matched healthy individuals. For example, [29] found

that people with AD had significantly lower frequency and velocity of pen strokes when performing simple repetitive drawing movements, compared to cognitively normal controls. Other experiments have documented deficits in AD patients during various writing tasks: writing sequences of letters, writing single words and sentences, and even functional tasks like copying a check or writing a paragraph are performed with reduced fluency and more errors in AD [27]. These fine-motor impairments in writing are thought to arise from a combination of motor slowing and the cognitive strain of written language production. Clinically, one might notice an AD patient's handwriting becomes less legible, with letter omissions, tremulous lines, or spatial misalignments (e.g., drifting off a line). Such changes, while subtle, correlate with disease and have become the basis for newer diagnostic approaches, for instance using digitizing tablets or AI analysis of written scripts to detect AD-specific patterns [107]. It's important to highlight that handwriting is often among the first skills affected by neurodegenerative cognitive decline. This is because writing a sentence engages multiple domains (memory, language, visuospatial processing, and fine motor control), all of which can be compromised even in prodromal AD. In summary, while AD does not typically cause early tremors or stiffness (as PD does), it does subtly affect motor-function, especially fine movements, early on. Clinicians may not rely on motor signs for diagnosis in the way they do cognitive signs, but these motor changes (like shuffling gait or changes in handwriting) can reinforce suspicion of neurodegeneration and even serve as additional quantitative biomarkers for early detection.

1.1.5 Diagnostic Criteria and biomarkers in AD

Diagnosing AD is a multifactorial clinical process that combines cognitive assessment, medical evaluation, and increasingly, biomarker evidence. Traditionally, AD has been a diagnosis of exclusion, clinicians rule out other potential causes of dementia and look for a characteristic pattern of cognitive impairment [104, 2]. In a routine clinical setting, the workup for suspected AD includes: a thorough patient history (and collateral history from family), a physical and neurological exam, cognitive testing, laboratory tests to exclude metabolic or other reversible causes, and brain imaging. Standard cognitive screening tools such as the Mini-Mental State Examination (MMSE) or Montreal Cognitive Assessment (MoCA) are often used to document the degree of impairment [48, 104]. The current Diagnostic and Statistical Manual of Mental Disorders, 5th ed. (DSM-5) uses the term "Major Neurocognitive Disorder due to Alzheimer's Disease" for dementia due to AD, which requires evidence of significant cognitive decline in one or more domains (with memory often affected) that interferes with independence, a gradual onset and progressive course, and no alternate explanation (such as stroke, brain tumor, etc.) [39].

Likewise, the National Institute on Aging-Alzheimer's Association (NIA-AA) guidelines have outlined criteria for probable AD dementia, emphasizing insidious onset and documented progressive cognitive decline, either by history or serial cognitive testing [104, 2, 73]. These criteria also acknowledge the use of biomarkers (e.g. MRI scans) to increase diagnostic certainty. For the intermediate stage of MCI due to AD, the NIA-AA 2011 research criteria [2] proposed a framework requiring cognitive concern and impairment on tests, but with essentially normal daily function. The criteria stratified the likelihood that AD pathology is the cause based on biomarker evidence of amyloid and neurodegeneration. This resulted in categories such as "MCI due to AD, high likelihood" versus "intermediate" or "unlikely". In practice, for a general clinician, a case of MCI can be labeled "prodromal AD" if, an amyloid PET scan is positive or CSF markers show the AD pattern. If no biomarker testing is done, the clinician might simply diagnose "MCI" and note AD is another possible diagnosis.

Role of Neuroimaging and Structural Magnetic Resonance Imaging: Neuroimaging, particularly structural MRI, is a vital component of the diagnostic process for AD [73]. It is routinely performed in clinical dementia evaluations, serving both to exclude other potential causes of cognitive impairment (such as brain tumors, stroke, or normal-pressure hydrocephalus) and to identify patterns of brain atrophy consistent with AD [53]. The classic finding on MRI is hippocampal and medial temporal lobe atrophy, often disproportionate to normal aging [139]. Visual rating scales such as the Medial Temporal Atrophy (MTA) allow clinicians to grade hippocampal atrophy; high scores in amnestic patients strongly suggest underlying AD. Beyond its diagnostic utility, MRI-derived measures of atrophy have prognostic value: patients with MCI showing significant hippocampal or temporoparietal cortical atrophy are more likely to progress to AD dementia [28]. The central role of structural MRI as an AD biomarker is discussed in detail in Section 1.2.

1.2 Structural MRI as a Biomarker of AD

1.2.1 Why MRI is central among Alzheimer's Biomarkers

Structural MRI has become a primary AD biomarker in research and diagnosis due to several key advantages. First, MRI is entirely non-invasive and does not require ionizing radiation or tracer injections, unlike PET imaging or lumbar puncture for Cerebrospinal Fluid (CSF) biomarkers [187]. This makes MRI safe for repeat longitudinal scans, en-

abling the tracking of disease progression over time. Second, MRI scanners are widely available in clinical settings, making this modality accessible for routine dementia evaluations globally. MRI provides high-resolution, multi-planar images of brain anatomy with excellent soft-tissue contrast [187], allowing clinicians to directly visualize structural brain changes associated with neurodegeneration. Indeed, MRI can reveal characteristic atrophy patterns in the brain even at early stages of AD, before severe clinical symptoms emerge [90].

Another major reason MRI remains central is that it measures the downstream neurodegenerative component of AD pathology, i.e. brain or hippocampal atrophy, which closely correlates with clinical impairment. In the AT(N) biomarker framework, a scheme that classifies AD biomarkers into three groups, amyloid-β deposition (A), tau pathology (T), and neurodegeneration or neuronal injury (N), MRI-based measures of atrophy serve as the prototypical neurodegeneration marker, complementing amyloid and tau indicators [72]. While amyloid and tau PET or CSF tests detect AD's molecular hallmarks decades before symptoms appear, MRI biomarkers reflect the cumulative loss of neurons and synapses that correlates strongly with cognitive decline and symptom severity [72, 70]. For example, an abnormal MRI showing atrophy combined with positive amyloid markers predicts future cognitive decline far better than amyloid positivity alone, since visible neurodegeneration indicates that substantial neuronal damage has occurred, bridging the gap between silent pathology and clinical dementia [72, 70].

MRI's practical advantages also strengthen its central role. It is widely used in clinical dementia workups to both exclude other pathologies (e.g. stroke, tumor) and to identify AD-typical atrophy patterns [53, 73]. Updated diagnostic criteria now incorporate structural MRI findings: for instance, MRI evidence of disproportionate medial temporal lobe atrophy is included as a supportive biomarker that increases the confidence of an AD diagnosis [139]. Visual rating scales such as the Medial Temporal Atrophy (MTA) allow clinicians to grade hippocampal atrophy; high MTA scores in patients with memory problems strongly suggest underlying AD. More advanced volumetric analyses can measure hippocampal volume or cortical thickness precisely, correlating with disease progression and memory decline [70, 87]. While emerging biomarkers (such as tau PET or plasma tests) show promise, they remain expensive or less accessible. By contrast, structural MRI is part of standard care in neurology and can be acquired in most hospitals. Moreover, the extent of atrophy on MRI has prognostic implications: patients with MCI who exhibit significant hippocampal or temporoparietal cortical atrophy are more likely to progress to AD dementia in subsequent years [28]. In summary, MRI continues to be a central AD biomarker because it is widely available, repeatable, and directly visualizes the neurodegenerative changes that underlie cognitive symptoms, providing an essential complement to molecular biomarkers. Other biomarkers (amyloid PET, CSF tests, etc.)

are certainly valuable and provide specificity for AD pathology, but they are generally used alongside MRI rather than replacing it. MRI's unique strength is in capturing the end-result of the disease's pathological cascade, neuronal loss and brain atrophy, in a manner that is non-invasive and clinically interpretable. This central position of MRI in the biomarker hierarchy is likely to persist, even as newer techniques are introduced, because all patients ultimately manifest neurodegeneration that MRI can detect.

1.2.2 Structural MRI and T1-Weighted Imaging

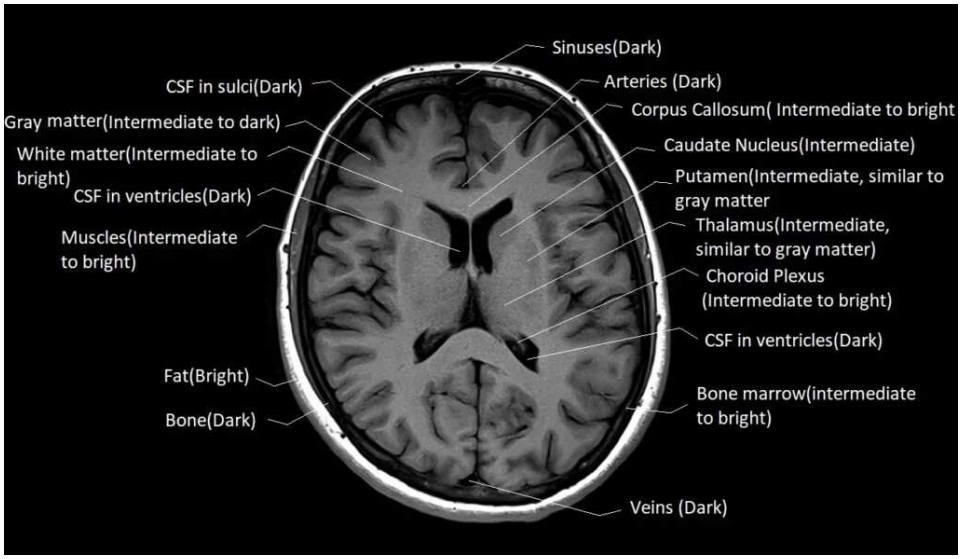


Figure 1.3: T1-weighted MRI of the brain showing characteristic tissue contrast: dark CSF, intermediate gray matter, and bright white matter. This contrast pattern enables precise delineation of brain structures critical for AD assessment. Image source: `www.mrimaster.com`.

Structural MRI can be acquired with different sequences, where a sequence is the specific acquisition protocol that determines how the signal is sampled and therefore which tissue characteristics are emphasized in the final image. The labels T1 and T2 refer to two different relaxation mechanisms of the MRI signal, and images are said to be T1-weighted or T2-weighted depending on which of these contrasts is predominantly emphasized. In brain imaging, T1-weighted scans are mainly used for anatomical detail and morphometric analysis, whereas T2-weighted and Fluid-Attenuated Inversion Recovery (FLAIR) images are more useful for highlighting fluid-sensitive contrast and white matter

lesions [187]. Acquisition also depends on scanner field strength, most commonly 1.5T or 3T; higher field strength generally provides higher signal-to-noise ratio and can support finer spatial resolution, although with technical trade-offs [147]. In large neuroimaging studies, structural scans are typically acquired as volumetric 3D datasets and can then be reformatted in sagittal, coronal, and axial planes for inspection and analysis [74].

T1-weighted MRI is the primary modality for structural brain imaging in AD for several reasons. Firstly, T1w scans provide excellent anatomical detail and tissue contrast: gray matter appears dark gray, white matter appears lighter, and Cerebrospinal Fluid registers as black (hypointense) [99]. This high gray-white matter contrast (illustrated in Fig. 1.3) facilitates clear delineation of cortical and subcortical structures. In practical terms, T1w images allow for precise visualization of the cortical ribbon, the shape and volume of hippocampus, the ventricular system, and other key neuroanatomical landmarks. Such images can be quantitatively analyzed to yield regional volume and cortical thickness measures with relatively standardized processing pipelines across studies.

As a result of these properties, high-resolution T1-weighted scans (often acquired as 3D volumetric sequences) have become the de facto standard for morphometric analysis in AD research. Large multi-center studies like the ADNI devoted substantial effort to optimizing a 3D T1w MRI protocol for consistent volumetric measurements [74]. In ADNI, a unified Magnetization Prepared Rapid Gradient Echo (MP-RAGE) sequence was selected and deployed across > 50 sites, with a target isotropic voxel size of approximately $1\ mm^3$ to capture fine anatomical details [74]. The emphasis on T1w MRI is so prominent that the ADNI MRI core explicitly stated the "most important image set was the T1-weighted 3D volumetric acquisition" for assessing brain morphometry [74]. Additional sequences (such as T2-weighted or Fluid-Attenuated Inversion Recovery (FLAIR) images) are typically acquired in studies to detect cerebrovascular lesions or other pathologies, but when it comes to measuring AD-related neurodegeneration (atrophy), the T1w scan is central. T1-weighted images are prone to automated analysis techniques (segmentation, surface mapping, etc.) and are easier to standardize across scanner platforms compared to multi-contrast or advanced sequences. This has enabled cross-cohort comparisons and pooling of data; for example, the ADNI T1w protocol and its quality control procedures (phantom scans, geometric corrections) have been adopted in other cohorts to harmonize structural MRI data. The focus on T1w MRI in this thesis thus ensures the use of the most widely validated and standardized structural imaging modality for Alzheimer's biomarker development.

1.2.3 Neuroanatomical Signatures of AD

Medial Temporal Lobe (MTL) structures are among the earliest and most severely affected regions. In particular, the hippocampus and adjacent entorhinal cortex and parahippocampal gyrus undergo profound volume loss in AD [90]. Hippocampal atrophy is a hallmark of AD, reflecting the prominent degeneration of neurons that correlates with the hallmark episodic memory impairment. At autopsy, the hippocampus and entorhinal region are where neurofibrillary tangle pathology (tau) first accumulates, and this early vulnerability is mirrored by MRI-visible atrophy in these regions in the MCI stage of AD. Correspondingly, patients with AD show significantly smaller hippocampal volumes than age-matched controls, and those with prodromal AD (MCI due to AD) often already exhibit measurable MTL atrophy [90, 2]. MTA is so central to AD that it has been incorporated into diagnostic criteria and radiological scoring systems (discussed in Section 1.2.4).

Beyond the MTL, AD also targets specific neocortical association areas. Notably, the temporo-parietal cortex shows marked atrophy as the disease progresses. This includes the lateral temporal lobe, inferior parietal lobule, posterior cingulate, and precuneus regions; all parts of the so-called "default mode network" heavily implicated in AD. Atrophy in these regions underlies clinical deficits in language, visuo-spatial orientation, and higher-order associative functions that manifest in moderate to advanced AD [9, 53]. For example, degeneration of the parietal cortex contributes to impaired visuo-spatial abilities and disorientation, while lateral temporal atrophy may relate to language and semantic memory difficulties [174]. Ventricular enlargement is another MRI-visible consequence of AD neurodegeneration. The enlargement of the lateral ventricles, particularly the temporal horn adjacent to the hippocampus, is often striking on MRI of AD patients [110]. While ventricular size increases slightly with normal aging, AD accelerates this process; longitudinal studies have shown that AD patients have faster rates of ventricle expansion than healthy elders, and even MCI patients who progress to AD show greater ventricular growth than stable MCI. Thus, ventricular volume can serve as an additional and indirect global marker of neurodegeneration.

1.2.4 Quantitative and Qualitative MRI Biomarkers

MRI-based biomarkers of AD can be broadly divided into qualitative (visual) assessments and quantitative measurements. In clinical practice, radiologists often employ quick visual rating scales to judge the degree of atrophy in key regions, whereas research studies (and evolving clinical tools) use automated algorithms to derive volumetric or morpho-

metric measures. A classic example of a qualitative MRI biomarker is the Medial Temporal Atrophy (MTA) visual rating scale proposed by [139]. This simple 5-point scale (0–4) is based on visual inspection of a single coronal T1 slice through the MTL: it assesses the width of the choroid fissure, the width of the temporal horn of the lateral ventricle, and the height of the hippocampal formation (see Figure 1.4). A score of 0 indicates no atrophy, while a score of 4 indicates severe atrophy with markedly enlarged CSF spaces and a shrunken hippocampus. The Scheltens' MTA scale has proven highly practical and remains one of the most commonly used tools in both clinics and research to gauge AD-related atrophy. In fact, its diagnostic utility is such that an abnormal MTA score (e.g. ≥ 2 in younger patients or ≥ 3 in older patients) can significantly increase confidence that a patient's dementia is due to AD, as noted in clinical criteria [164]. Other visual rating scales exist as well. For example the Koedam scale [85] for posterior atrophy, which help qualitatively capture atrophy in regions beyond the hippocampus. Visual ratings have the advantage of being quick (they can be performed during routine scan review) and requiring no specialized software, thus effectively bridging neuroradiology expertise with standardized assessment.

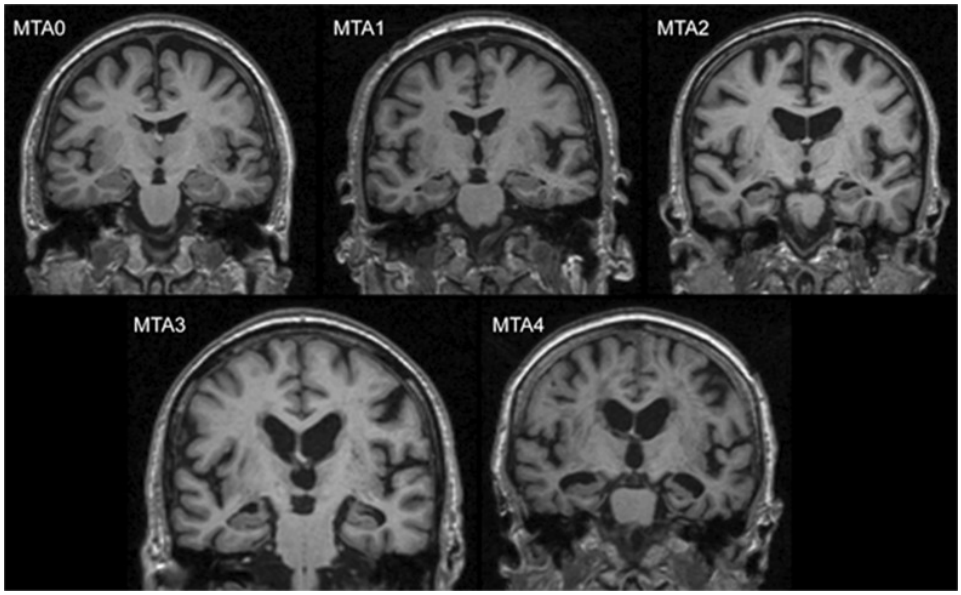


Figure 1.4: Progression of medial temporal lobe atrophy across MTA scores (0–4) on coronal MRI sections. Reprinted from [164].

On the quantitative side, MRI biomarker analysis provides a wealth of numeric measures. Foremost among these are volumetric measurements of specific brain structures.

For AD, volumetry of the hippocampus has been a focus for decades, manual tracings of the hippocampal formation on MRI were the gold standard in early research, consistently showing ∼10–30% volume reductions in AD patients compared to controls [71]. Nowadays, automated segmentation tools (e.g. FreeSurfer [47] or Automatic Segmentation of Hippocampal Subfields (ASHS) [163]) can compute hippocampal volumes (and volumes of many other regions) for a given scan. Similarly, whole-brain or ventricular volumes can be measured, or the rate of whole-brain atrophy can be quantified using techniques like the Boundary Shift Integral (BSI) (which calculates volume loss between two time points) [51]. Another widely used quantitative metric is cortical thickness. By reconstructing the cortical surface from a T1w MRI, one can measure the thickness of the gray matter layer at thousands of points across the cortex; AD is characteristically associated with thinning in the posterior cingulate, precuneus, and lateral temporo-parietal cortex, as noted earlier. These quantitative approaches offer fine-grained assessments and the ability to track change over time in a continuous manner (e.g. "mm^3 of hippocampal loss per year").

In summary, each approach has its strengths, and importantly they are complementary rather than mutually exclusive. Visual scales provide an immediate, clinically interpretable readout (e.g. "MTA score = 3" communicates marked medial temporal atrophy in one number), and their performance is not far from that of more labor-intensive quantitative methods: visual MTA correlates with hippocampal volume and, in some cohorts, discriminates AD from controls with accuracy comparable to volumetry [53, 8]. However, this does not mean that qualitative scales are sufficient on their own. First, MTA must be interpreted in an age-aware manner, because medial temporal atrophy also increases with normal aging and the practical cut-offs change across decades of life [164, 21]. Second, MTA is informative but not fully specific for AD: medial temporal atrophy can also be seen in other dementias, whereas some biomarker-positive AD patients, especially in early or atypical presentations, may still have age-appropriate MTA scores [53, 98]. Finally, a visual 0–4 rating inevitably compresses a complex and spatially distributed neurodegenerative pattern into a coarse ordinal score. For this reason, visual ratings are excellent rapid clinical tools, but volumetric measures remain preferable when continuous quantification is needed, such as longitudinal follow-up, clinical trials, or correlations with cognition and CSF biomarkers [53].

1.2.5 Methodological Challenges in MRI-Based Analysis

While MRI is a rich source of biomarkers, analyzing structural MRI data for AD is not without challenges. These challenges must be acknowledged, as they motivate the ad-

vanced methodological approaches introduced later in this work. Key issues include:

1. **High-dimensional, 3D data complexity:** Each MRI scan is a three-dimensional volume (typically with millions of voxels), and the pattern of AD-related changes can be subtle and spatially distributed. Traditional image analysis often reduced this complexity by examining a few slices or Region of Interest (ROI), but this risks missing the full extent of disease effects. Modern approaches aim to analyze the entire brain volume to capture the complex atrophy pattern, but this comes at the cost of requiring sophisticated 3D image processing and statistics to avoid false positives. The 3D nature of MRI also means that misalignment or deformation (due to head tilt, anatomy, etc.) can affect measurements, so images must be carefully spatially normalized for comparisons. Handling such high-dimensional data without oversimplification is a primary challenge [42, 131, 5].

2. **Limited labeled data and ground truth:** When developing computational models (for example, to predict diagnosis or to segment brain structures), a limitation is the availability of "ground truth" labels. Unlike some imaging domains, there is usually no voxel-level ground truth of "disease" in MRI – we infer AD by atrophy patterns rather than seeing a distinct lesion. Manually labeling brain regions (such as tracing the hippocampus) is labor-intensive, and while public datasets exist (e.g. ADNI provides many MRI scans), case-level labels such as "AD vs control" are typically derived from clinical characterization rather than direct pathology [119, 170]. Histopathological confirmation is generally available only in dedicated autopsy studies rather than routine imaging cohorts [174]. Thus, algorithms must often learn from limited and noisy labels. This scarcity of annotated data makes it challenging to train and validate reliable automated methods, especially those involving Machine Learning (ML) [170, 183].

3. **Intersubject variability:** Brain morphology varies widely across individuals due to age, sex, head size, genetics, and life experience ("cognitive reserve"). There is a substantial overlap between the range of some MRI measures in healthy elderly and those in AD patients. For instance, a cognitively normal 80-year-old might normally have some hippocampal atrophy or enlarged ventricles, whereas a younger AD patient might still have a relatively large brain volume. This variability can mask disease effects in cross-sectional analyses. It also complicates establishing universal cut-offs for biomarkers (what is "normal" vs "abnormal" at a given age) [53, 164, 21]. Deep learning approaches can help address this challenge by automatically learning complex, high-dimensional patterns that capture disease-specific features across diverse populations, potentially discovering subtle imaging signatures that transcend individual anatomical variability [42, 170].

4. **Scanner and site effects:** MRI images can differ not only because of biology but also due to technical factors. Different MRI scanners (or different field strengths, e.g. 1.5T vs 3T) produce images with varying contrast, resolution, and noise characteristics [147]. Even the same model scanner at two sites might be calibrated differently. Multi-center studies like ADNI confronted this by rigorously standardizing protocols and using phantom objects to calibrate scanners [106, 74, 176]. Despite such efforts, site-related variability can still confound analyses. An algorithm might inadvertently learn to distinguish scans from two hospitals rather than two patient groups if not properly controlled [170, 183]. While traditional approaches require extensive preprocessing standardization to harmonize images across sites, deep learning models may help by learning more stable representations from heterogeneous data, although this remains an active challenge rather than a solved problem [170].

5. **Partial volume effects and 3D resolution:** Structural MRI, even at 1 mm^3 resolution, has voxels that may contain a mix of tissue types at boundaries (gray matter, white matter, CSF). This partial voluming can introduce errors in measuring thin structures like the cerebral cortex or hippocampal subfields [158, 163]. Distinguishing true atrophy from imaging artifacts requires careful algorithm design. Additionally, standard structural MRI captures macroscopic atrophy rather than microscopic pathology, making it a relatively downstream marker of neurodegeneration in the AD cascade [70, 53]. This limits MRI sensitivity in detecting very early neurodegeneration, as macroscopic atrophy only becomes visible after more subtle biological changes have already begun. While higher-field MRI and advanced sequences are improving detection capabilities, this remains an inherent limitation of structural imaging [147].

These challenges set the stage for why advanced analysis techniques are needed. For instance, dealing with high-dimensional 3D data and intersubject variability calls for powerful statistical and ML methods to detect the "signal" of AD amidst noise. Likewise, the lack of abundant labeled training data motivates the use of data augmentation, transfer learning, or unsupervised approaches that do not require large annotated training sets. In the following chapters, some of these issues will be addressed by employing methods that can better handle variability, provide explainability aligned with MRI biomarkers, and support the training of large foundation models without annotated images.

1.3 AI for AD: Landscape and Open Gaps

This thesis explores three complementary paradigms in AI-based Alzheimer's disease analysis, each addressing specific limitations in current approaches while building toward increasingly general and clinically translatable frameworks. The evolution from traditional task-specific predictors to explainable models and ultimately toward foundation models represents a progression that mirrors the broader trajectory of AI in medical imaging.

1.3.1 From Feature Engineering to End-to-End Deep Learning

Early computer-aided diagnosis of AD relied heavily on manually engineered features and traditional ML classifiers. Building upon the quantitative and qualitative MRI biomarkers discussed in Section 1.2.4, domain experts would computationally extract neuroimaging measures such as hippocampal volume, cortical thickness, or ventricular shape descriptors from MRI scans, which were then fed into algorithms like support vector machines or logistic regression to distinguish AD from healthy aging [42, 59, 131]. These approaches essentially automated the extraction of features corresponding to established biomarkers like hippocampal atrophy or ventricular enlargement (see Section 1.2.3). While these methods demonstrated that structural MRI changes could aid in identifying AD, they depended heavily on domain knowledge and extensive preprocessing pipelines. The feature engineering process might miss subtle or complex patterns, potentially biasing the analysis toward known biomarkers and limiting the discovery of additional imaging signatures.

The rise of deep learning marked a transition toward automated feature learning directly from medical images. Rather than using predefined features, CNN-based models learn hierarchical representations at multiple levels of abstraction, capturing subtle anatomical patterns associated with AD without explicit human guidance [38, 79]. This transition enabled a broader analysis, minimizing human bias in feature selection and potentially uncovering patterns that expert-designed features might overlook. Initial CNN models for AD classification achieved competitive performance using structural MRI, often exceeding expectations under controlled experimental conditions on benchmark datasets like ADNI [79]. However, deep models also introduced new challenges, including the need for large labeled datasets, concerns about interpretability, and risks of overfitting, particularly given the relatively limited size of neuroimaging cohorts compared to natural image datasets [38, 103]. These challenges, along with issues of high-dimensional data complexity, intersubject variability, and scanner effects (as discussed in Section 1.2.5),

motivate the approaches explored in this thesis.

1.3.2 Complementary Biomarkers Beyond Neuroimaging: Deep Learning for Heterogeneous Handwriting Tasks

While neuroimaging remains central to AD diagnosis, complementary biomarkers extending beyond brain scans offer valuable diagnostic information, particularly when accessibility and cost are considerations. As discussed earlier in the context of motor and fine-motor changes in AD (see Section 1.1.4), handwriting emerges as a particularly relevant modality due to its sensitivity to the cognitive and motor impairments characteristic of neurodegenerative diseases [165, 80]. The documented deterioration in handwriting characteristics such as writing speed, pen pressure, stroke regularity, and spatial organization in AD patients motivates the exploration of computational approaches to detect these subtle patterns. Analysis of handwriting samples, whether captured digitally through tablets or as offline images scanned from paper, offers a non-invasive, accessible, and cost-effective screening tool that complements neuroimaging assessments.

Prior research in handwriting-based AD detection has predominantly relied on manual feature extraction followed by traditional ML classifiers or statistical analysis, typically evaluating each handwriting task independently with separate models [84, 126, 173, 20, 17]. While these approaches demonstrated the diagnostic potential of handwriting analysis, they remained limited by the need for manual feature engineering and failed to exploit possible synergies across diverse handwriting activities [20, 17, 15]. Deep learning approaches, particularly end-to-end models that automatically learn features directly from handwriting images, have been largely unexplored for AD handwriting-based detection [19, 18, 25]. More critically, no prior work has attempted to train a unified model capable of learning shared hierarchical representations across heterogeneous handwriting tasks ranging from graphic exercises (spirals, loops) to copying text, drawing clocks, and freeform writing [16, 20, 17, 15, 19, 18, 25]. Chapter 2 of this thesis addresses this gap by discussing a unified deep learning framework that learns cross-task representations from 25 diverse handwriting activities. By training a single model on the entire spectrum of tasks including clock drawing, signature, sentence copying, dictation, and memory-based writing, the framework captures common AD-related patterns that transcend individual task boundaries while also identifying task-specific contributions to diagnostic accuracy. This work conducts a broad architectural comparison of modern CNNs (ResNet [61], EfficientNet [155], ConvNeXt) and Vision Transformers [35], explores transfer learning from ImageNet to overcome limited medical data availability, and develops a multi-stage aggregation strategy that combines predictions across all tasks to achieve better diagnostic

performance. This unified approach demonstrates that shared representations learned across diverse handwriting modalities can capture generalizable AD signatures more effectively than isolated task-specific models.

1.3.3 Deep Learning for Neuroimaging: Balancing Context and Efficiency

When applied to structural MRI, deep learning methods face a fundamental tension between capturing full spatial information and maintaining computational feasibility. Approaches can be categorized into three families, each with distinct trade-offs. **3D CNNs** that process entire brain volumes can capture global patterns of neurodegeneration and preserve full three-dimensional context, potentially identifying distributed atrophy patterns across the brain [7, 45]. However, they are computationally expensive, memory-intensive, and suffer from a scarcity of large-scale pre-trained models, limiting opportunities for transfer learning. **Patch-based 3D approaches** reduce computational demands by extracting and analyzing smaller sub-volumes from MRI scans [127, 115], enabling more focused analysis of disease-relevant regions like the hippocampus while remaining tractable. Yet, by restricting the model's field of view, these methods may miss global patterns and inter-regional relationships crucial for understanding disease heterogeneity. **2D slice-based methods** treat individual MRI slices as standard images, enabling the use of well-established 2D CNN architectures and using pre-trained weights from ImageNet [81, 169]. This approach benefits from abundant pre-trained models and computational efficiency, but inherently loses three-dimensional spatial context by fragmenting the brain into independent slices.

Recent work has increasingly incorporated attention mechanisms and transformer architectures to address these limitations. Attention modules allow models to dynamically weight different spatial regions or slices, effectively learning which parts of the brain are most informative for diagnosis [78, 175]. Vision transformers, which process images as sequences of patches with self-attention, offer an alternative framework that can model long-range spatial dependencies [35]. Hybrid approaches that combine CNN feature extraction with transformer-based integration across slices or regions represent a useful middle ground, balancing computational efficiency with the ability to capture global context [65, 3].

1.3.4 The Interpretability Imperative: From Black Boxes to Clinically Meaningful Explanations

A critical barrier to clinical adoption of AI systems in AD diagnosis is the lack of interpretability. Most deep learning models function as black boxes, providing predictions without transparent reasoning [103]. In medical contexts, particularly for complex neurodegenerative diseases with heterogeneous presentations, clinicians require explanations that align with established neurobiological knowledge and support clinical decision-making. Post-hoc explainability techniques such as GradCAM [141] have been widely applied to visualize which brain regions contribute most to model predictions. While these methods often highlight anatomically relevant areas, their clinical reliability and consistency remain debated [145, 161]. Importantly, most explainability studies rely on qualitative visual inspection rather than quantitative validation against known biomarkers or expert consensus, limiting their translational value [145, 161].

Attention-based architectures offer an alternative path toward inherent and learnable interpretability. By learning to weight different spatial locations, slices, or feature channels, attention mechanisms provide built-in explanations derived from the model's own parameters rather than post-hoc approximations [77, 140]. When properly designed, attention maps can be more consistent, localized, and aligned with clinical expectations than gradient-based saliency methods. Moreover, quantifying the importance of specific anatomical regions in model predictions enables validation against established AD biomarkers such as hippocampal atrophy or temporal lobe changes, bridging the gap between algorithmic predictions and clinical reasoning. Chapter 3 explores this space comprehensively, proposing an explainable framework that analyzes MRI scans across multiple orthogonal slicing planes (sagittal, coronal, and axial). Using attention mechanisms, the model generates 3D attention maps that consistently highlight the neuroanatomical regions characteristic of AD neurodegeneration as described in Section 1.2.3, including the hippocampus, entorhinal cortex, and inferior lateral ventricles, while achieving high diagnostic accuracy. This approach provides both visual and quantitative evidence that the model focuses on clinically relevant features.

1.3.5 Methodological Rigor: Addressing Reproducibility and Generalization Challenges

Despite impressive reported performance, many AI models for AD imaging suffer from methodological limitations that compromise their reliability and generalizability. Recent systematic reviews have identified recurring issues including data leakage from improper

train-test splits, lack of independent external validation, and failure to account for site or scanner variability [171, 183]. In some cases, models initially reported to achieve accuracies above 90% experienced dramatic performance drops when evaluated under stricter protocols or on independent cohorts, suggesting that much of the reported performance reflected overfitting to dataset-specific characteristics rather than genuine disease signal extraction. Fewer than 20% of published deep learning models for AD imaging have been validated on external datasets, and virtually none have undergone prospective clinical testing.

These challenges stem from several sources. Neuroimaging datasets for AD are relatively small compared to those required to train deep neural networks without overfitting, particularly for complex architectures [171, 103]. Subject-level data leakage can occur when multiple scans or slices from the same individual appear in both training and test sets, allowing models to memorize patient-specific patterns rather than learning generalizable disease features [171, 183]. Differences in MRI acquisition protocols, scanner manufacturers, field strengths, and population demographics across clinical centers further complicate model generalization by introducing potential model-learned biases unrelated to diagnostic patterns [183, 74]. Addressing these issues requires careful experimental design including subject-level data splitting, rigorous cross-validation strategies, independent test sets, and external validation on datasets acquired under different conditions [171, 183]. Reproducibility also demands transparent reporting of preprocessing steps, hyperparameters, and making code and models publicly available when possible [171, 103]. Chapter 3 addresses these concerns by adopting a standardized ADNI collection of scans and employing a rigorous and standardized preprocessing pipeline [74]. The study also implements 5-fold cross-validation with strict subject-level data splitting to prevent leakage, validates the methodology through quantitative XAI metrics aligned with known AD biomarkers, and provides open-source code for reproducibility. Chapter 4 builds upon this foundation by reusing the same preprocessing pipeline across ADNI, AIBL, and OASIS-3 datasets, employing consistent 90-10 subject-level train-test splits, and critically evaluating zero-shot transferability by testing models pre-trained on ADNI directly on external cohorts through linear probe experiments without fine-tuning.

1.3.6 From Generative Models to Foundation Models: Unsupervised Representation Learning

Generative models such as GANs [55], VAEs [82], and DDPMs [64] have primarily been applied to AD research for data augmentation, cross-modality synthesis, and segmentation tasks [1]. While these applications have proven valuable, they have not fully ex-

plored the representation learning capabilities of generative models, particularly diffusion-based frameworks. Recent work on diffusion autoencoders [124] has demonstrated that diffusion models can learn compact, semantically meaningful representations that enable controllable image manipulation and attribute-specific modifications. However, applying these techniques to high-dimensional 3D MRI data remains computationally prohibitive. Latent Diffusion Model (LDM)s [133] address this challenge by performing diffusion in a compressed latent space, dramatically improving efficiency while maintaining generation quality.

The key advantage of unsupervised representation learning lies in its ability to use vast quantities of unlabeled neuroimaging data. Clinical repositories contain millions of brain MRI scans acquired for diverse purposes: routine health screenings, evaluation of unrelated neurological conditions, or monitoring of diseases other than AD. These scans typically lack diagnostic labels for AD, making them unsuitable for supervised learning. However, they capture fundamental aspects of brain anatomy, aging, and anatomical variability that are relevant across conditions. Unsupervised methods can extract this knowledge by learning representations that capture the underlying structure of brain imaging data without requiring task-specific annotations. A model pre-trained on such diverse, unlabeled data can subsequently be adapted to specific downstream tasks like AD diagnosis with minimal labeled examples, following the foundation model paradigm that has transformed natural language processing and computer vision.

This approach offers multiple advantages for AD neuroimaging. First, it amortizes the computational cost of training across multiple downstream applications. Second, it enables better generalization by learning from diverse data during pre-training, reducing overfitting to specific datasets or diagnostic criteria. Third, it facilitates transfer learning to data-scarce scenarios, such as rare diseases or small clinical cohorts, by using representations learned from larger unlabeled datasets. However, realizing this vision faces challenges: medical imaging datasets remain orders of magnitude smaller than those used to train foundation models in other domains, heterogeneity in scanner protocols and acquisition parameters complicates dataset aggregation, and the optimal architectures and pre-training objectives for 3D medical imaging foundation models remain open research questions.

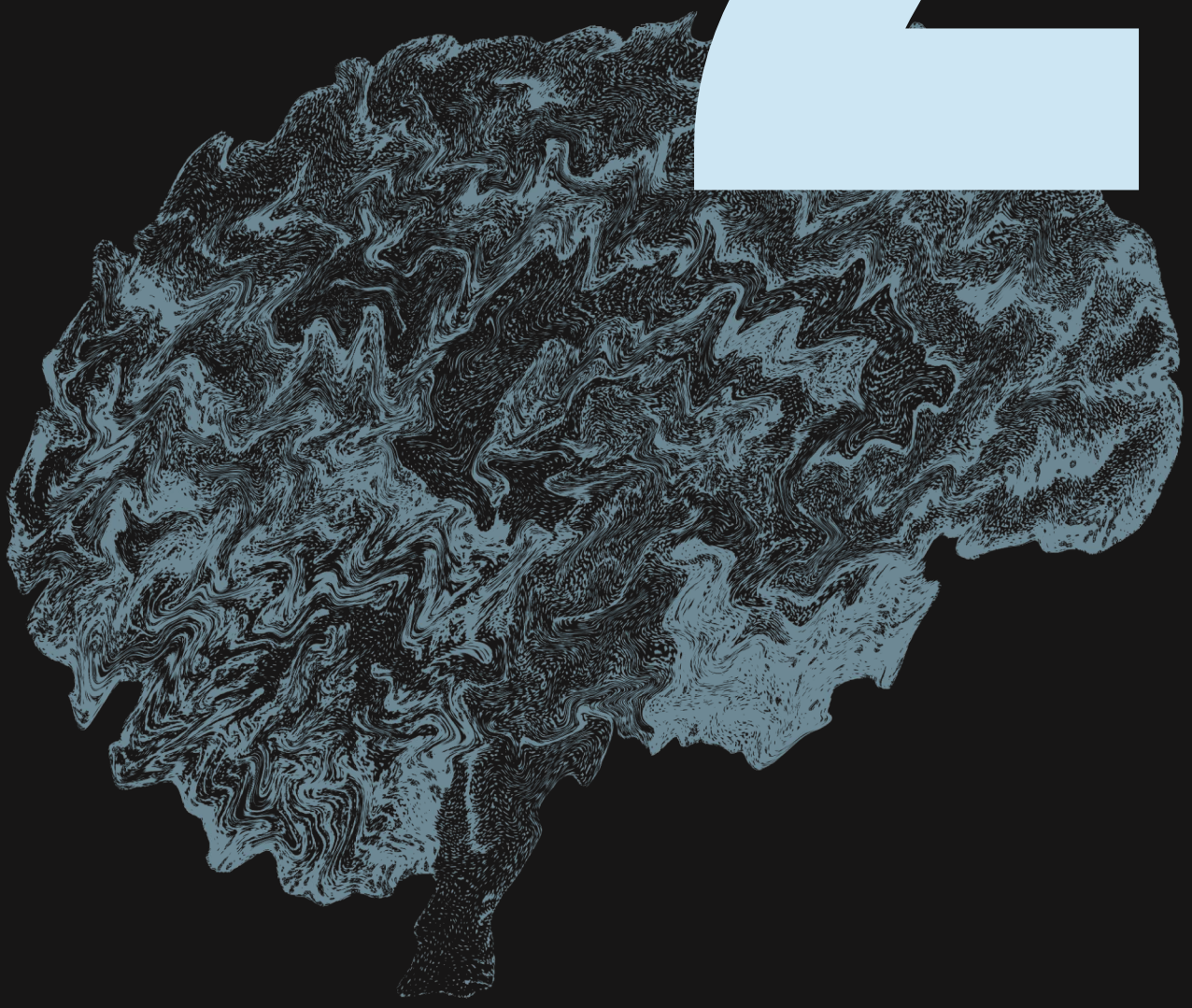
Chapter 4 addresses some of the mentioned challenges by introducing Latent Diffusion Autoencoder (LDAE), a framework that adapts diffusion autoencoders to operate in the latent space of a pre-trained autoencoder, enabling efficient unsupervised representation learning from 3D brain MRI at scale. Trained unsupervisedly on ADNI data, LDAE learns representations that transfer effectively to unseen datasets (AIBL and OASIS-3) and support multiple downstream tasks (diagnosis, age prediction, interpolation) through sim-

ple linear probing without model fine-tuning. Beyond zero-shot transferability, LDAE's semantic latent space enables controllable anatomical manipulation for interpretability (e.g., simulating subject-specific disease state changes) and generation of missing intermediate scans in longitudinal series through latent interpolation, providing both analytical and clinical utility. This multi-task capability and generative flexibility suggest that diffusion-based representation learning can serve as a foundation for more general neuroimaging analysis frameworks, moving beyond task-specific predictors toward models that capture fundamental aspects of brain structure and aging.

1.4 Thesis Structure and Contributions

The remainder of this thesis is organized as follows. Chapter 2 explores handwriting-based biomarkers as accessible screening tools, employing deep learning to aggregate predictions across diverse handwriting tasks. Chapter 3 develops an attention-based framework for neuroimaging that provides both high diagnostic accuracy and clinically interpretable 3D attention maps validated against established AD biomarkers. Chapter 4 introduces LDAE, an efficient unsupervised representation learning framework that demonstrates zero-shot transfer capabilities across datasets and enables applications such as disease progression simulation and longitudinal scan interpolation. Together, these contributions span the progression from task-specific predictors to explainable diagnostic models to general-purpose representation learning, reflecting the broader evolution of AI in medical imaging toward interpretable, reliable, and clinically deployable systems.

2



HANDWRITING-BASED BIOMARKERS FOR ALZHEIMER'S DISEASE

REFERENCE PAPER

Transformers and CNNs in Neurodiagnostics: Handwriting Analysis for Alzheimer's Diagnosis

AUTHORS

Gabriele Lozupone[1]
Emanuele Nardone[1]
Cesare Davide Pace[1]
Tiziana D'Alessandro[1]

[1]Department of Electrical and Information Engineering (DIEI), University of Cassino and Southern Lazio

published as: Lozupone, G., Nardone, E., Pace, C. D., and D'Alessandro, T. (2025). Transformers and CNNs in Neurodiagnostics: Handwriting Analysis for Alzheimer's Diagnosis. In International Conference on Pattern Recognition (pp. 447-463). Springer, Cham.

Chapter 2

Handwriting-Based Biomarkers

2.1 Introduction

As introduced in Section 1.3.2, handwriting analysis represents a promising biomarker for AD diagnosis. While Chapter 1 established the neurobiological foundations of AD and the documented motor and fine-motor impairments affecting handwriting (Section 1.1.4), this chapter presents a comprehensive deep learning framework that operationalizes handwriting as a practical diagnostic tool. The deterioration in handwriting characteristics provides measurable indicators of the cognitive and motor dysfunction characteristic of AD [165, 80, 88, 109].

Handwriting analysis offers several practical advantages over neuroimaging: it is non-invasive, substantially more cost-effective, and readily accessible even in resource-limited settings. Despite growing interest in computational handwriting analysis for neurodegenerative disease detection [26, 172], most research efforts have concentrated on Parkinson's Disease (PD), benefiting from established reference datasets [37]. For AD, research has typically relied on smaller, privately collected datasets and manual feature extraction methods followed by traditional ML classifiers [84, 126] or statistical analysis [173]. These approaches remain limited by manual feature engineering or by individual evaluation of handwriting tasks. This chapter addresses these limitations through an end-to-end deep learning approach that automatically learns multi-task hierarchical representations directly from offline handwriting images (scanned paper sheets rather than tablet-based digital traces).

Prior work on handwriting-based AD detection has followed a progression from traditional feature engineering to deep learning approaches. Earlier studies developed comprehensive handwriting protocols comprising multiple tasks designed to probe different cognitive, memory, and motor abilities affected by AD [16]. These initial approaches followed conventional methodologies: extracting manually engineered static and dynamic features from tablet-captured handwriting, then applying traditional ML classifiers to evaluate each task independently [20, 17]. Subsequent investigations explored genetic algorithm-based feature selection to combine predictions from task subsets [15], and examined CNN-based classification on synthetic images [19, 18]. More recent work demonstrated the diagnostic value of offline scanned images [25], which simplified data acquisition and enabled retrospective analysis of existing handwritten documents.

Despite these advances, all prior approaches shared a fundamental limitation: each handwriting task was evaluated independently by dedicated models, fragmenting the analysis and failing to leverage synergies across the diverse task spectrum. This chapter addresses this gap by proposing a unified deep learning framework that trains a single model across all 25 handwriting tasks simultaneously, ranging from graphic exercises (spirals, loops) to clock drawing, text copying, dictation, and memory-based writing. As outlined in Section 1.3.2, no prior work has attempted a comprehensive cross-task representation learning for handwriting-based AD detection. The hypothesis driving this unified approach is that shared hierarchical representations learned across heterogeneous handwriting activities can capture generalizable AD signatures that transcend individual task boundaries, while also enabling quantification of task-specific diagnostic contributions. By consolidating all tasks into a single training framework, the model can learn common AD-related patterns (such as motor control degradation, spatial disorganization, or cognitive load manifestations) that may remain obscured when tasks are analyzed in isolation.

This chapter makes three primary contributions to AD handwriting-based diagnosis:

1. **Comprehensive architectural evaluation**: A systematic comparison of modern CNN architectures (VGG, ResNet, EfficientNetV2, ConvNeXt) and Vision Transformers (ViT, Swin Transformer) identifies the most effective deep learning approaches for unified multi-task handwriting analysis.

2. **Transfer learning paradigm comparison**: Investigation of two distinct transfer learning strategies—task-domain transfer from ImageNet (general image classification) versus input-domain transfer from Optical Character Recognition (OCR) pre-training on handwritten text—testing the hypothesis that domain-specific pre-training on handwritten text (via TrOCR) might offer advantages over general-

purpose image pre-training for handwriting-based diagnosis.

3. **Multi-stage diagnostic aggregation framework**: Development of a three-stage experimental framework that evaluates individual task contributions, implements cross-validation with strict subject-level data splitting to prevent leakage, and aggregates predictions across all 25 tasks through majority voting to produce robust subject-level diagnoses.

The remainder of this chapter is structured as follows: Section 2.2 describes the 25-task handwriting protocol, dataset characteristics, and the CNN and transformer architectures evaluated, including the adaptation of TrOCR from OCR to classification. Section 2.3 details the three-stage experimental framework, cross-validation strategy, and evaluation metrics. Section 2.4 presents task-specific and subject-level diagnostic performance across all architectures. Finally, Sections 2.5 and 2.6 discuss implications for clinical translation, task informativeness, and future directions including the potential for longitudinal monitoring and retrospective analysis of historical documents.

2.2 Materials and Methods

This section outlines the materials required for proposed investigation, including the data and Deep Learning (DL) architectures employed. Section 2.2.1 delineates the data acquisition procedure and the processes used to obtain the offline images of handwriting samples. This image type was selected to represent handwriting since preliminary results of past activity [25] enhanced the advantages of using such a kind of data. Finally, Section 2.2.2 details the architectures evaluated.

2.2.1 Dataset

Data were acquired considering an experimental protocol [16] comprising 25 handwriting tasks, designed with the help of physicians to evaluate different abilities that could be affected by AD symptoms. The selected tasks belong to four categories: graphic, copy and reverse copy, memory, and dictation. Table 2.1 lists the performed tasks and the belonging category. The protocol was administered following strict recruiting criteria, with the support of the geriatric ward's Alzheimer unit at the "Federico II" hospital in Naples. In detail, participants underwent clinical assessment and standard cognitive evaluations, such as the Mini-Mental State Examination (MMSE) [48], Frontal Assessment Battery

Table 2.1: Description of handwriting tasks administered to participants. Categories: M = memory and dictation, G = graphic, C = copy.

#	Task Description	Category
1	Signature drawing	M
2	Join two points with a horizontal line, continuously for four times	G
3	Join two points with a vertical line, continuously for four times	G
4	Retrace a circle (6 cm of diameter) continuously for four times	G
5	Retrace a circle (3 cm of diameter) continuously for four times	G
6	Copy the letters 'l', 'm' and 'p'	C
7	Copy the letters on the adjacent rows	C
8	Write recursively a sequence of four lowercase letters 'l'	C
9	Write recursively a sequence of four lowercase cursive bigram 'le'	C
10	Copy the word "foglio"	C
11	Copy the word "foglio" above a line	C
12	Copy the word "mamma"	C
13	Copy the word "mamma" above a line	C
14	Memorize the words "telefono", "cane", and "negozio" and rewrite them	M
15	Copy in reverse the word "bottiglia"	C
16	Copy in reverse the word "casa"	C
17	Copy words (regular/non-regular/non-words) in boxes	C
18	Write the name of the object shown in a picture (a chair)	M
19	Copy the fields of a postal order	C
20	Write a simple sentence under dictation	M
21	Retrace a complex form	G
22	Copy a telephone number	C
23	Write a telephone number under dictation	M
24	Clock Drawing Test	G
25	Copy a paragraph	C

(FAB) [69], and Montreal Cognitive Assessment (MoCA) [108]. Participants in the study were carefully selected to ensure demographic and educational characteristics between the CN and AD groups were matched, as reported in Table 2.2. Individuals taking psychotropic drugs or substances that could influence cognitive abilities were excluded from both groups. The final dataset consisted of 180 individuals, with 90 diagnosed with AD and 90 CN.

Handwriting samples were acquired using a WACOM Bamboo Folio graphic tablet, which allowed participants to write on standard A4 paper sheets fastened to the tablet's surface. The sensor-equipped pen recorded spatial coordinates (x, y), pressure (z), and timestamps

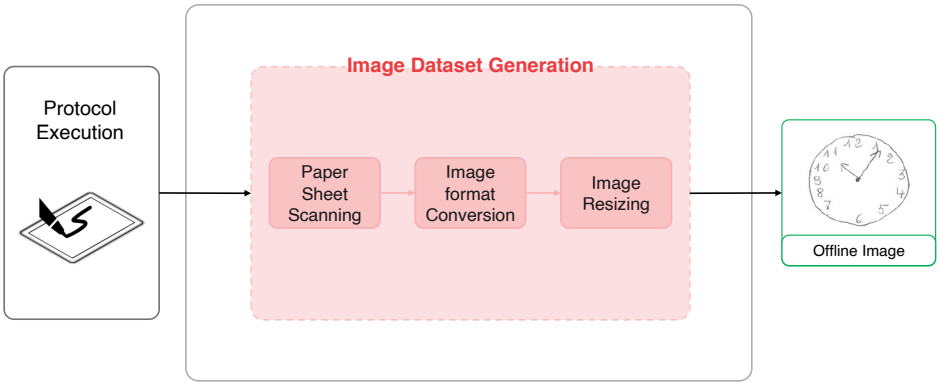


Figure 2.1: Workflow diagram of offline image dataset generation process.

Table 2.2: Demographic characteristics of study participants. Values represent mean (standard deviation).

Group	Age (years)	Education (years)	Women	Men
AD	71.5 ± 9.5	10.8 ± 5.1	46	44
CN	68.9 ± 12.0	12.9 ± 4.4	51	39

at $200Hz$, capturing in-air movements within $3cm$ from the tablet surface. However, for this study, only offline images obtained from the paper sheets were utilized, disregarding the dynamic aspects of handwriting. Figure 2.1 shows the generation process of offline images. After the protocol execution, the paper sheets were scanned and saved as *.tif* files, with each frame representing a task. Thus, each frame was extracted and underwent a segmentation algorithm to isolate the participant's handwritten trace and saved as a *.png* file. During this dataset construction stage, the segmented traces were normalized onto a 299×299 pixel canvas while ensuring the trace remained centred, minimizing information loss. This resolution refers to the offline dataset preprocessing only; a further resize to the model input size was applied later during training. The digitized images accurately reflect the participant's handwritten trace, with pixel values representing the natural grayscale shades of the ink on the paper, influenced by both the applied pressure and the dynamics of the movements. It is worth noting that each of the 180 participants executed each task only once, yielding a dataset of 4500 images, 180 for each of the 25 tasks.

2.2.2 ImageNet-Pretrained Networks

One of the investigations aims at comparing the ability of CNNs and ViTs to learn AD writing patterns. This section briefly discusses the architecture families chosen for the analysis and their peculiarities. All the architectures described in this section employed transfer learning from the ImageNet dataset [30]. The **VGG** architecture family, developed in [143], represents a significant advance in deep learning for image recognition tasks. Characterized by its simplicity and depth, the architecture employs 3x3 convolutions, evenly stacked, to increase the depth of the network without complicating its structure. The **ResNet** [61] family, inspired by VGG, introduced the residual block that enables learning the residual function for identity mapping that mitigates the vanishing gradient problem. The **EfficientNetV2** [155] architecture family pushes the boundaries of efficiency through innovative scaling strategies. By intelligently scaling model dimensions, they achieve remarkable accuracy and training speed enhancements, redefining the balance between performance and resource utilization. Transitioning from conventional convolutional designs, **ConvNeXt** introduces a paradigm shift inspired by Vision Transformers. In [97], the authors modified the ResNet architecture. They applied a macro design consisting of changes in the number of layers in each block. They patchified the input image with learnable convolution blocks characterized by stride increasing to simulate the ViT patch embedding [35]. These architectural choices show that purely convolutional architecture can compete with state-of-the-art Vision Transformers. The **ViT** [35] adapts the original Transformer architecture from Natural Language Processing to computer vision. The Transformer works with an input that consists of a sequence of words (or tokens). The authors of ViT generate a sequence by splitting the input image into non-overlapping patches with a fixed size of 16×16 pixels. Each patch is linearly projected into a fixed-sized space, and a class token is added to the sequence of embeddings. The class token enables the condensation of useful classification features through the self-attention mechanism and the direct connection with the head. The **Swin Transformer** family was proposed in [96], employing hierarchical structures and shifted window mechanisms. These architectural innovations enhance scalability and efficiency across a broad spectrum of vision tasks, demonstrating the potential of Transformer-based models in computer vision applications.

2.2.3 From OCR to Handwriting Alzheimer's Diagnosis

The chosen architecture is **TrOCR** presented in [93] and explicitly designed for OCR tasks. The motivation behind selecting TrOCR for this application stems not only from its performance, particularly on the Handwriting Dataset IAM Handwriting Database

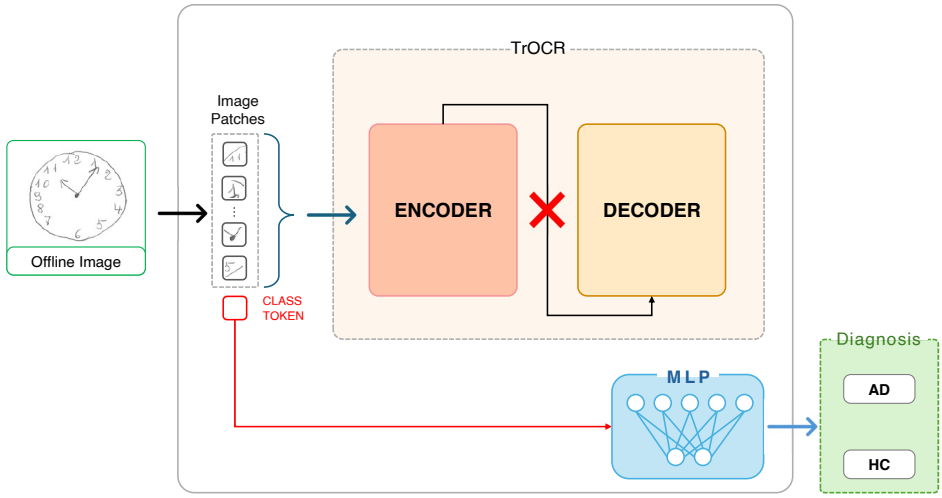


Figure 2.2: Modified TrOCR architecture scheme. Encoder retained for feature extraction, decoder removed, class token and MLP added for classification.

(IAM)[102], but also from an architectural consideration. TrOCR is composed of a vision Transformer encoder pretrained on handwritten text images and a Transformer decoder for autoregressive transcription. This makes it a relevant candidate for transfer learning in our setting because the encoder is expected to learn handwriting-aware visual representations, such as stroke shape, local character structure, spacing, and writing regularity, which are closer to our input domain than generic natural-image features. The **Transformer Encoder** is responsible for processing the input image. As in ViT, the encoder processes a sequence of tokens representing patch embeddings extracted from the handwritten text image and builds a global representation through self-attention. The **Transformer Decoder** instead uses the encoded representation to generate the output text one character or token at a time. This sequence-generation mechanism is essential for OCR, but it is not required for our task, where the objective is a single image-level decision rather than text transcription. As shown in Figure 2.2, the encoder is therefore retained as the feature extractor and converted to a ViT-like classifier by removing the decoder. A class token is added to the input token sequence, and its final representation is fed to an Multi-Layer Perceptron (MLP) classification head responsible for the diagnosis. In this way, the architecture preserves the handwriting-specialized encoder while replacing the transcription-oriented output module with a binary classification head.

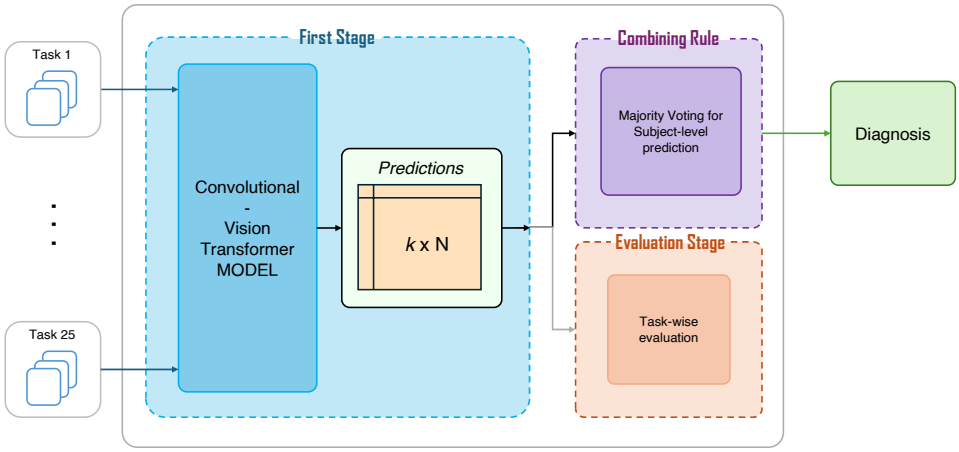


Figure 2.3: Overview of the Experimental Approach for Diagnosing AD Using DL Models on Handwriting Images. This figure illustrates the three-step process, starting with the training of various models using a 5-fold cross-validation method (First Stage), followed by the evaluation of task-specific accuracy (Evaluation Stage), and concluding with the aggregation of task-specific outcomes for individual diagnoses through the majority voting (Combining Rule).

2.3 Experimental Approach

In order to evaluate DL models' ability in identifying AD from task-specific text images derived from handwriting work, a structured three-stage approach was adopted. The emphasis was placed on ensuring the comparability and reliability of the chosen convolutional and vision transformer models through a meticulous cross-validation process. Figure 2.3 presents an overview of the complete experimental framework, comprising three main phases: the First Stage, the Evaluation Stage, and the Combining Rule. Each of these phases is elaborated upon in detail within this section.

The First Stage involved training the selected DL models on the entire dataset, which included offline images of different subjects engaged in specific writing tasks designed to reveal the presence of AD. The goal was to refine the models to produce robust feature representations that accurately mirrored AD-related patterns in handwriting. To enable a fair comparison between the models and enhance their ability to generalize, a 5-fold cross-validation technique was adopted. This approach entailed dividing the dataset into five distinct segments, ensuring each segment acted as a test set in one iteration

and as part of the training/validation sets in the others. To prevent data leakage and preserve the integrity of our evaluation, all writing tasks from the same subjects were assigned exclusively to one of the training, validation, or test sets. Thus, no subject's data was shared across the sets. The data was subjected to an 80-20 split within each fold for training/validation and testing, respectively. The 80% portion allocated for training/validation was divided using an 80-20 ratio. Consequently, 64% of the total dataset was used for training purposes, 16% for validation, and 20% for testing. This structured division ensured that the generalization ability of each model was thoroughly assessed.

In the second phase, the models trained in the first stage were evaluated, emphasising task-specific predictions. A task-wise evaluation was established by structuring the predictions in a matrix format, with subjects delineated by rows and tasks by columns. This method facilitated the calculation of metrics for each task against the ground truth. Such an evaluation not only quantified the contributory value of individual tasks toward AD diagnosis but also allowed for an assessment of metric uniformity across tasks. From this analysis, two critical aspects were derived: the unique contribution of each task to the AD diagnostic process and potential insights into whether combining tasks could lead to more accurate and robust AD diagnoses.

The final step involved aggregating the results from the individual tasks based on the test set predictions to produce a final diagnosis for each subject. The aggregation of predictions across all tasks leveraged the complete information spectrum, thereby maximizing diagnostic accuracy. This integrated approach maximized comprehensiveness by utilizing all available subject data. Steps 2 and 3 relied exclusively on test set portions of each fold, thereby ensuring generalization to unseen data.

2.3.1 Experimental settings

We employed an AdamW optimizer with weight decay 10^{-2} and used a batch size of 32. During training, the images were resized to 224×224, which corresponds to the input resolution required by the pretrained models used in our experiments. Therefore, the 299×299 resolution described above refers to the standardized dataset preprocessing stage, whereas 224×224 is the final model-input size adopted for transfer learning. We chose a learning rate of 10^{-3} and trained all the models for 100 epochs, using an early stopping strategy with a patience of 15 epochs. The loss function used was the binary cross entropy.

This study uses Accuracy, Sensitivity, Specificity, and Matthews Correlation Coefficient (MCC) as performance metrics. Accuracy is a classifier's most common performance

measure, consisting of the percentage of correctly classified samples over the total. MCC [14] is a correlation coefficient between prediction and true label; it provides a more informative and truthful score than accuracy when evaluating binary classifications, allowing a more realistic interpretation of classifier performance, especially in the case of unbalanced datasets. Sensitivity and specificity serve as critical metrics for evaluating the diagnostic accuracy of our model, highlighting its ability to correctly identify AD cases and CN, respectively.

2.4 Results

In Table 2.3, we provide the results of the first stage of the experimental study. It can be observed that the architectures used reach an average Accuracy of approximately 74%, with the metrics calculated as the mean of a 5-fold cross-validation. Specifically, ConvNextSmall reaches 79.21% Accuracy and 82.32% Sensitivity. On the other hand, the VGG19 achieves 78.84% of Specificity, and ConvNextTiny obtains 57.88% of MCC. This achievement indicates that the architectures can discern the handwriting pattern of subjects with AD from those of CN.

Table 2.3: Performance comparison of CNN and Transformer models in the first stage. Values represent mean (standard deviation) across folds. Best results are highlighted in bold.

Model	**Accuracy (%)**	**Specificity (%)**	**Sensitivity (%)**	**MCC (%)**
VGG16	74.56 ± 4.35	74.61 ± 9.34	73.78 ± 5.76	48.59 ± 9.70
VGG19	75.48 ± 3.41	**78.84 ± 7.18**	71.83 ± 6.14	50.90 ± 7.34
ResNet50	70.74 ± 2.29	78.04 ± 8.35	64.08 ± 6.18	42.72 ± 5.61
ResNet101	71.45 ± 3.39	77.46 ± 6.80	65.35 ± 5.99	43.19 ± 7.35
EfficientNetV2M	72.50 ± 4.69	74.35 ± 9.01	70.28 ± 6.88	44.86 ± 10.00
EfficientNetV2S	73.05 ± 3.22	77.41 ± 9.46	68.52 ± 6.45	46.43 ± 7.62
ConvNextTiny	78.83 ± 2.65	77.96 ± 7.18	79.41 ± 8.57	**57.88 ± 5.32**
ConvNextSmall	**79.21 ± 3.39**	74.66 ± 8.38	**82.32 ± 7.47**	57.82 ± 7.64
ConvNextBase	78.03 ± 4.55	76.80 ± 10.88	78.01 ± 8.00	55.42 ± 10.12
VitB16	70.10 ± 3.82	73.85 ± 9.60	66.47 ± 9.53	40.88 ± 7.95
SwinV2B	77.61 ± 2.75	79.62 ± 5.99	75.70 ± 8.85	55.69 ± 5.23
SwinV2S	76.86 ± 2.95	74.66 ± 11.44	78.68 ± 9.18	54.36 ± 5.34
SwinV2T	77.13 ± 2.31	74.92 ± 5.61	78.44 ± 6.95	53.82 ± 5.00
TrOCR-Small-Hand	67.53 ± 2.73	76.29 ± 6.01	59.57 ± 5.22	36.36 ± 5.20
TrOCR-Base-Hand	66.02 ± 3.04	77.98 ± 10.85	55.47 ± 12.62	35.10 ± 5.32

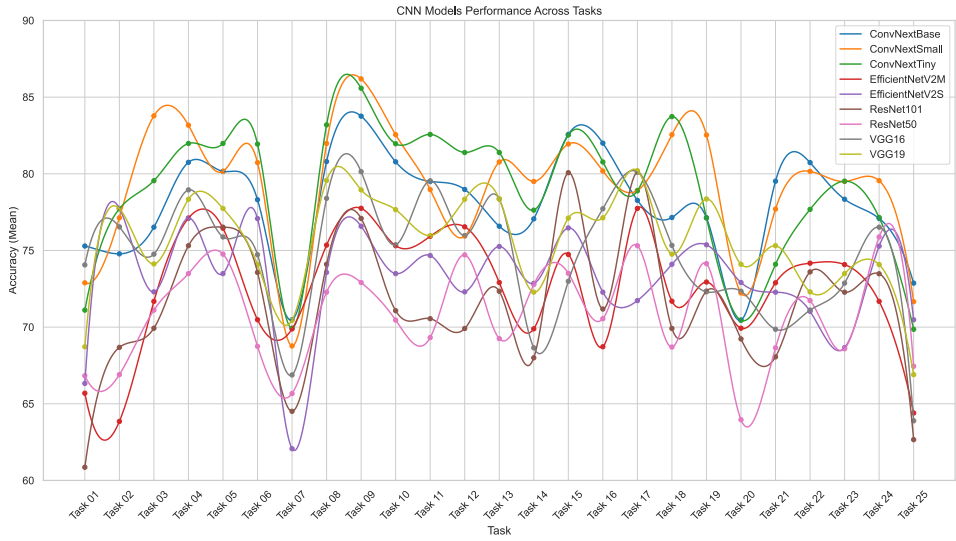


Figure 2.4: Accuracy comparison of CNN models across 25 tasks.

Figure 2.4 and Figure 2.5 provide an analysis of the performance across the pool of tasks for CNN and Vision Transformer architectures, respectively. We can observe that in both graphs, there is a large variability in performance in terms of accuracy depending on the task. For instance, Tasks 8 and 9, which have higher performance, independently of the family of the architecture used, have higher accuracy and demand higher levels of cognitive processing and motor coordination. Instead, Tasks 1, 7, 20, and 25, which require minimal cognitive effort and fine motor skills, demonstrated significantly lower performance. These likely offer fewer distinctive features for the architectures to learn from, leading to a diminished capacity to differentiate between AD and CN. Table 2.4 illustrates the performance achieved by the architecture for each designated task, highlighting a distinct prevalence of ConvNext over other architectures. From the results, we see that ConvNextSmall reached 86.19% accuracy and 72.20% of MCC in task 9, and 86.69% specificity in task 10, while ConvNextTiny reached 89.69% of sensitivity in task 14.

Table 2.5, obtained through majority voting rule, shows that ConvNextSmall has the best overall performance with an Accuracy of 87.99% and MCC of 76.75%, showcasing a remarkably high Sensitivity of 89.69%, indicating its superior ability to identify AD subjects. Its performance is balanced, excelling in both sensitivity and MCC. This suggests that it can effectively handle both AD and CN classifications. ResNet50 exhibits the highest Specificity of 92.06%, indicating a greater ability to identify CN subjects. However, its

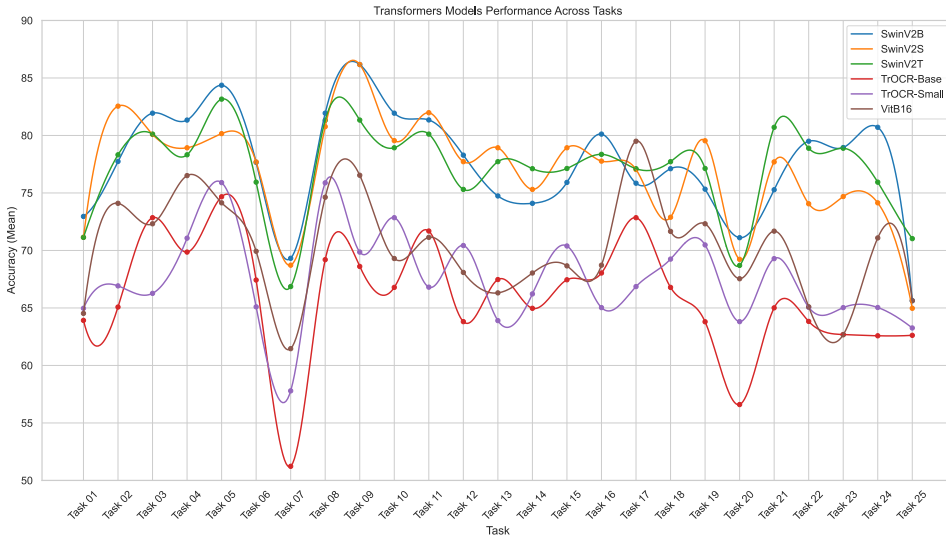


Figure 2.5: Accuracy comparison of Transformers models across 25 tasks.

lower Sensitivity 67.97%, and MCC 62.09% suggest that it may not be as effective at identifying AD subjects as other models. From the Vision Transformer architecture family, the SwinV2T performs well with an Accuracy of 85.54% and MCC of 71.53%. These models show a good balance between Sensitivity and Specificity, indicating their robustness in diagnosis but worse than ConvNextSmall. Variability across models is notable, with standard deviations indicating that some models (e.g., ConvNextBase, EfficientNetV2M) exhibit more variation in their performance metrics across folds. This variability might be due to differences in how models handle the intrinsic complexities and variations within the AD vs CN classification task. The TrOCR models show lower performance than other models in Table 2.5. Transfer learning from ImageNet has proven to be more effective. It can be assumed that for this specific analysis, a transfer of knowledge from task domains, e.g. classification task, is more valuable rather than input data domain, e.g. handwritten text images.

Table 2.6 presents the performance metrics for the best models, freezing a subset of the model's layers during training, which can assist in preventing overfitting and enhancing performance. ConvNextSmall achieves its peak accuracy of 84.44% when employing 25% freezing. However, this performance falls short of the 87.99% accuracy it attains without implementing any freezing technique. Conversely, the SwinV2T model shows improved performance with 50% freezing, achieving an accuracy of 86.15% and an MCC of 72.30%. This is higher than its non-frozen performance (85.54% accuracy and 71.53% MCC), indi-

Table 2.4: Top-performing model for each task ranked by accuracy. Values represent mean (standard deviation) across folds.

Task	Best Model	Accuracy (%)	Specificity (%)	Sensitivity (%)	MCC (%)
1	ConvNextBase	75.29 ± 7.57	72.67 ± 17.78	76.59 ± 11.94	51.02 ± 15.09
2	SwinV2S	82.55 ± 9.38	78.96 ± 10.22	85.83 ± 8.66	65.04 ± 18.50
3	ConvNextSmall	83.78 ± 6.43	80.44 ± 7.64	86.21 ± 9.11	67.23 ± 13.32
4	ConvNextSmall	83.17 ± 8.59	79.65 ± 14.71	85.91 ± 13.69	66.37 ± 17.83
5	SwinV2B	84.37 ± 7.44	83.31 ± 10.30	85.19 ± 8.34	68.90 ± 15.55
6	ConvNextTiny	81.94 ± 4.70	80.13 ± 10.35	82.89 ± 11.77	64.16 ± 9.56
7	ConvNextBase	70.48 ± 3.94	71.77 ± 16.21	67.43 ± 16.11	40.63 ± 9.15
8	ConvNextTiny	83.19 ± 6.68	86.10 ± 9.46	80.76 ± 13.41	67.35 ± 12.65
9	ConvNextSmall	**86.19 ± 6.47**	85.79 ± 5.39	86.32 ± 8.92	**72.20 ± 13.05**
10	ConvNextSmall	82.55 ± 3.81	**86.69 ± 6.20**	78.25 ± 8.98	65.45 ± 7.51
11	ConvNextTiny	82.57 ± 6.02	85.23 ± 10.31	79.58 ± 12.07	65.54 ± 12.10
12	ConvNextTiny	81.39 ± 8.11	83.45 ± 8.60	79.42 ± 21.25	64.52 ± 14.72
13	ConvNextTiny	81.39 ± 6.54	81.09 ± 6.23	81.94 ± 13.83	63.38 ± 12.61
14	ConvNextSmall	79.50 ± 5.05	66.51 ± 19.05	**89.69 ± 8.66**	59.63 ± 11.46
15	ConvNextBase	82.57 ± 7.70	74.54 ± 14.57	88.62 ± 8.97	64.83 ± 17.03
16	ConvNextBase	82.00 ± 7.68	78.31 ± 4.73	85.85 ± 13.13	64.12 ± 15.26
17	VGG19	80.20 ± 10.54	78.96 ± 17.90	82.24 ± 8.97	61.39 ± 20.57
18	ConvNextTiny	83.73 ± 3.50	78.83 ± 10.73	87.40 ± 2.96	67.00 ± 7.92
19	ConvNextSmall	82.53 ± 10.79	77.13 ± 14.89	86.75 ± 9.19	64.27 ± 23.15
20	VGG19	74.10 ± 3.40	84.35 ± 8.98	64.79 ± 6.95	50.21 ± 7.59
21	SwinV2T	80.71 ± 5.54	74.31 ± 4.33	85.76 ± 10.82	61.44 ± 10.82
22	ConvNextBase	80.73 ± 5.87	78.18 ± 18.05	81.37 ± 8.50	60.79 ± 13.18
23	ConvNextTiny	79.52 ± 3.94	75.65 ± 12.96	83.12 ± 8.86	59.82 ± 8.02
24	SwinV2B	80.71 ± 6.67	81.73 ± 10.91	80.12 ± 12.12	62.46 ± 12.61
25	ConvNextBase	72.87 ± 8.09	70.81 ± 15.21	73.85 ± 6.29	44.78 ± 17.83

cating that freezing half of the layers helps better adapt to the task. The TrOCR-Small-Hand model exhibits similar behaviour, showing a 6.63% accuracy improvement in 75% freezing setting. The VitB16 model performed better with 25% freezing, but this is still lower than its non-frozen setting (81.34% vs 77.70% accuracy), suggesting that this model benefits from having more trainable parameters for this specific task. Overall, these results demonstrate that the impact of freezing varies across different model architectures. While some models like SwinV2T and TrOCR-Small-Hand benefit from freezing, others like ConvNextSmall and VitB16 perform better when more layers are trainable. It's also worth noting that even with freezing, ConvNextSmall and SwinV2T remain the top-performing models, consistent with the initial results.

Table 2.5: Performance comparison of CNN and Transformer models using majority voting. Values represent mean (standard deviation) across folds. Best results are highlighted in bold.

Model	Accuracy (%)	Specificity (%)	Sensitivity (%)	MCC (%)
VGG16	82.57 ± 8.11	85.02 ± 11.98	79.69 ± 8.94	65.01 ± 17.04
VGG19	84.37 ± 3.35	91.99 ± 5.24	77.15 ± 9.23	70.09 ± 5.56
ResNet50	79.50 ± 2.39	**92.06 ± 7.45**	67.97 ± 8.41	62.09 ± 4.64
ResNet101	78.93 ± 7.61	90.16 ± 10.74	67.73 ± 11.99	59.84 ± 15.39
EfficientNetV2M	80.73 ± 8.89	86.63 ± 13.87	74.76 ± 10.46	62.56 ± 18.45
EfficientNetV2S	78.31 ± 5.89	87.73 ± 10.86	69.20 ± 9.48	58.55 ± 12.37
ConvNextTiny	87.38 ± 5.78	88.46 ± 10.19	86.59 ± 13.30	76.38 ± 10.70
ConvNextSmall	**87.99 ± 5.30**	85.15 ± 9.90	**89.69 ± 13.06**	**76.75 ± 10.63**
ConvNextBase	86.17 ± 8.65	87.37 ± 13.44	84.01 ± 12.28	72.60 ± 18.09
VitB16	77.70 ± 9.72	86.63 ± 10.46	69.42 ± 21.85	58.40 ± 16.42
SwinV2B	84.94 ± 5.07	89.63 ± 8.20	80.70 ± 11.98	71.36 ± 9.51
SwinV2S	83.74 ± 3.59	83.08 ± 12.53	84.30 ± 12.29	69.26 ± 7.26
SwinV2T	85.54 ± 1.18	86.89 ± 7.01	84.04 ± 6.86	71.53 ± 2.75
TrOCR-Small-Hand	71.66 ± 6.63	87.95 ± 8.26	57.23 ± 11.58	47.86 ± 11.38
TrOCR-Base-Hand	71.00 ± 11.54	90.16 ± 11.36	53.85 ± 20.86	48.17 ± 19.91

Table 2.6: Impact of layer freezing on model performance. Values represent mean (standard deviation) across folds. Best results for each model are highlighted in bold.

Model	Freezing (%)	Accuracy (%)	Specificity (%)	Sensitivity (%)	MCC (%)
ConvNextSmall	25	84.44 ± 9.30	82.06 ± 7.95	86.41 ± 13.11	**69.05 ± 19.04**
	50	82.55 ± 5.13	**89.48 ± 6.50**	76.27 ± 12.90	67.06 ± 9.13
	75	83.16 ± 4.82	83.84 ± 13.00	81.90 ± 11.53	67.08 ± 11.06
VitB16	25	81.34 ± 6.71	80.01 ± 10.94	82.04 ± 10.70	62.81 ± 14.39
	50	77.11 ± 5.60	88.17 ± 9.65	67.57 ± 13.57	57.76 ± 11.07
	75	77.13 ± 5.51	86.92 ± 11.19	68.40 ± 11.93	57.11 ± 10.60
SwinV2T	25	82.55 ± 6.41	84.94 ± 14.67	79.28 ± 17.37	67.22 ± 12.94
	50	**86.15 ± 5.60**	82.16 ± 14.88	**88.54 ± 8.45**	72.30 ± 11.91
	75	82.55 ± 9.23	81.34 ± 12.02	83.01 ± 12.76	65.21 ± 19.20
TrOCR-Small-Hand	25	77.13 ± 8.19	85.83 ± 16.42	68.02 ± 12.14	56.08 ± 17.52
	50	75.29 ± 7.03	86.94 ± 17.19	63.01 ± 12.62	53.44 ± 15.87
	75	78.29 ± 7.58	71.62 ± 7.92	84.36 ± 11.35	57.08 ± 16.40

2.5 Discussions

Our experimental results support the relevance of each stage of the proposed framework. Compared with prior handwriting-based AD studies, where tasks were typically analyzed separately and only later compared or combined [20, 17, 15, 19, 18], the unified setting adopted here suggests that aggregating information across tasks can improve robustness at the subject level. Subsequently, in the second stage, we aggregated results by task, unveiling the varying significance of different tasks for our research objectives. Lastly, the third stage integrated a majority voting rule applied to predictions from the same patient, supporting the idea that combining classifications across multiple tasks can enhance patient-level diagnostic accuracy. We evaluated and compared various architectures belonging to two distinct families: CNNs and Vision Transformers. Among the architecture families tested, ConvNext exhibited superior performance compared to other CNNs and Vision Transformers. Specifically, *ConvNextSmall* demonstrated better performance in Accuracy and Sensitivity, a crucial metric in the medical domain, indicating the proportion of correctly identified patients afflicted by a disease. Moreover, we explored the feasibility of transferring knowledge from OCR to AD recognition tasks. However, this approach achieved significantly lower performance compared to the best method, likely reflecting a domain mismatch between OCR-oriented pre-training and handwriting analysis for neurodiagnostic purposes. TrOCR is specifically designed to extract and interpret textual content from images [93], whereas AD-related handwriting alterations are more closely related to neuromotor and spatial abnormalities than to text transcription itself [165, 88, 109]. At the same time, this clinical interpretation should be made cautiously, especially in elderly cohorts, because handwriting degradation is not specific to cognitive decline. Age-related factors such as arthritis, reduced manual dexterity, tremor, visual impairment, or other motor comorbidities may also affect written traces and therefore act as confounders. In this study, these factors were not modeled explicitly and should be regarded as a limitation of the biomarker rather than implicitly attributing all performance differences only to AD-related cognitive impairment. From the second experiment, we underscored the variability in task performance within the handwriting protocol, independent of the architecture considered. Some tasks appear more informative than others, particularly those designed to engage multiple cognitive and motor functions, in line with the clinical rationale of the original protocol [16]. By contrast, simpler copy tasks that require less cognitive effort tended to obtain worse performance. This finding suggests that future refinements of the protocol could focus on the most discriminative tasks when combining predictions.

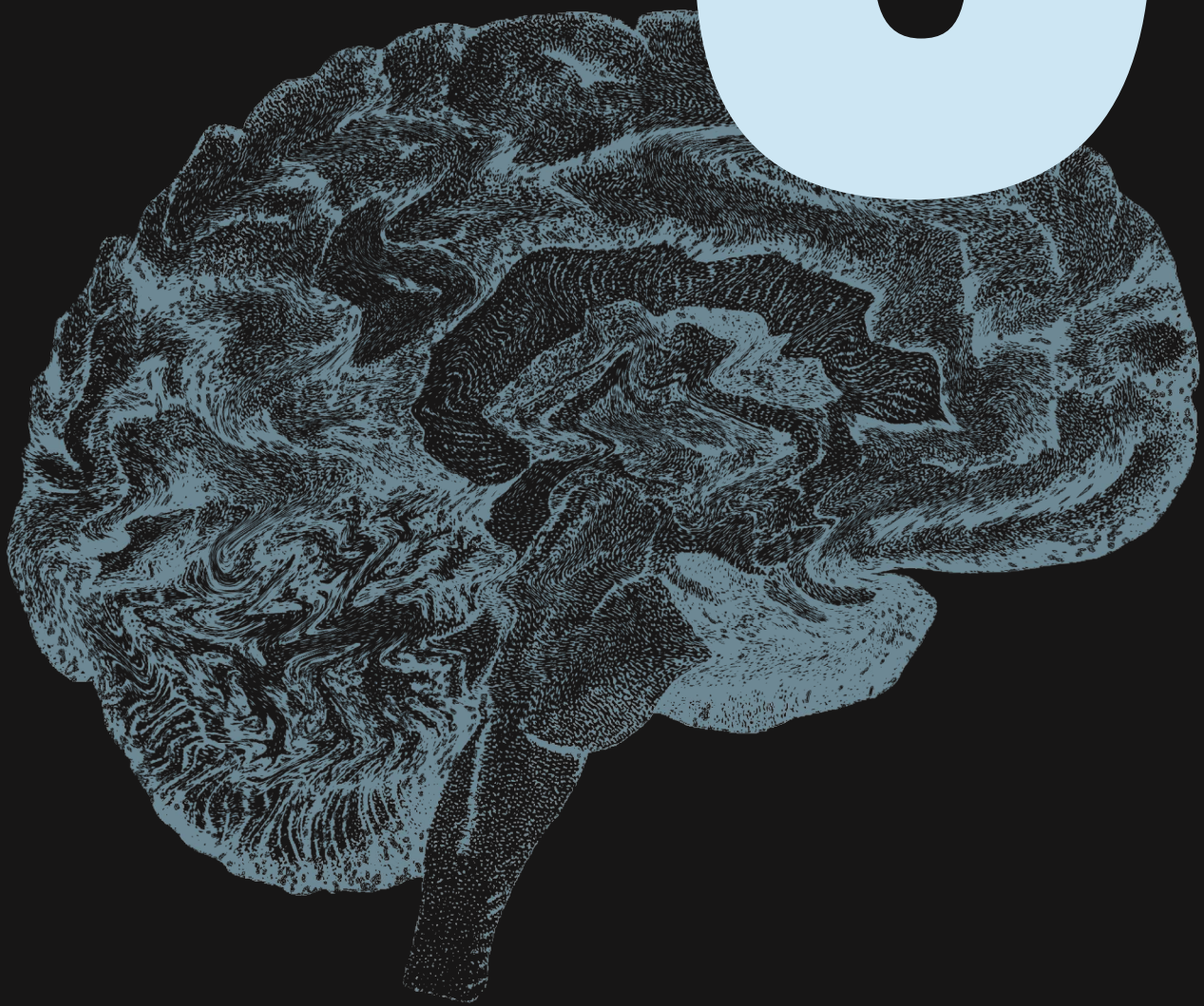
2.6 Conclusions

In this chapter, we presented a non-invasive and relatively low-cost approach that could support broader AD screening workflows through offline handwriting analysis [165, 25, 126]. By combining predictions across multiple tasks, the framework provides a subject-level assessment rather than an isolated task-wise evaluation.

Besides the encouraging results, several directions for further analysis remain open. In particular, because the system relies on offline data scanned from paper sheets, additional handwritten material could be incorporated relatively easily, which may be especially beneficial for more data-hungry architectures such as transformers. Moreover, collecting repeated handwriting samples would allow future work to evaluate whether this type of marker can also support longitudinal monitoring of cognitive-motor changes, a possibility already discussed in the handwriting literature [165, 173].

Finally, the use of scanned handwritten documents as the primary offline source also opens the possibility of retrospective analyses on previously collected material, consistent with prior work showing the feasibility of deep learning on offline handwriting images [25]. These broader applications remain to be validated on larger and more heterogeneous cohorts.

3



INTERPRETABLE ALZHEIMER'S DIAGNOSIS WITH DEEP LEARNING

REFERENCE PAPER

AXIAL: Attention-based eXplainability for Interpretable Alzheimer's Localized Diagnosis using 2D CNNs on 3D MRI brain scans

AUTHORS

Gabriele Lozupone[1,2]
Alessandro Bria[1]
Francesco Fontanella[1]
Frederick J.A. Meijer[3]
Claudio De Stefano[1]

[1]Department of Electrical and Information Engineering (DIEI), University of Cassino and Southern Lazio
[2]Diagnostic Image Analysis Group, Radboud University Medical Center
[3]Department of Medical Imaging, Radboud University Medical Center

Based on the arXiv preprint: arXiv:2407.02418, 2024

Chapter 3

Interpretable Alzheimer's Diagnosis with Deep Learning

3.1 Introduction

Having established the clinical foundations of AD and the central role of structural MRI as a non-invasive biomarker in Chapter 1, this chapter addresses a critical challenge in AI-based neuroimaging analysis: achieving high diagnostic accuracy while maintaining clinical interpretability with limited training data. As discussed in Section 1.3.3, existing DL approaches for MRI analysis face a fundamental trade-off between computational feasibility and spatial context preservation. 3D CNNs can capture global patterns of neurodegeneration but suffer from computational expense and limited transfer learning opportunities due to the scarcity of pre-trained models [7, 45, 83]. 3D patch-based methods reduce computational demands but restrict the model's field of view, potentially missing inter-regional relationships [127, 115]. Conversely, 2D slice-based methods enable efficient transfer learning from ImageNet-pretrained models and offer computational advantages, but inherently fragment the brain into independent slices, losing crucial inter-slice relationships that encode 3D spatial patterns [81, 169, 170].

Recent work has attempted to bridge this gap by incorporating attention mechanisms and transformer architectures to model inter-slice dependencies. The Attention Transformer model in [3] employs cross-attention between a central slice and surrounding slices to fuse volumetric information, while Conv-Swinformer [65] uses Swin Trans-

former blocks to establish semantic connections across planar features. However, as discussed in Section 1.3.4, these transformer-based approaches face overparameterization challenges when applied to limited medical datasets, increasing overfitting risk and undermining generalizability. Furthermore, most proposed methods rely on post-hoc explainability techniques such as GradCAM [141] that produce inconsistent saliency maps with limited clinical specificity, failing to reliably identify the neuroanatomical regions characteristic of AD described in Section 1.2.3 [145, 161, 166].

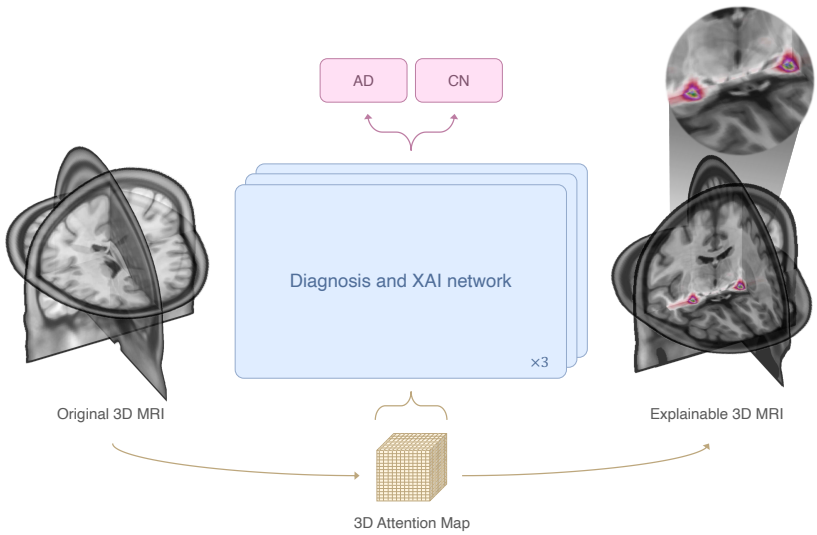


Figure 3.1: Schematic representation of the proposed diagnostic framework. The Diagnosis and XAI framework processes a 3D sMRI brain image to generate two key outputs: three diagnosis networks identifying the condition as either AD or CN from the three possible slicing axes and a corresponding 3D attention map that can be overlapped to the input image to visually highlight the brain regions the network focuses on to derive its diagnosis.

This chapter discuss **AXIAL** (**A**ttention-based e**X**plainability for **I**nterpretable **A**lzheimer's **L**ocalized diagnosis), a parameter-efficient framework that addresses these limitations through multi-plane attention fusion and learnable explainability. As illustrated in Fig. 3.1, AXIAL analyzes MRI scans across three orthogonal slicing planes (sagittal, coronal, and axial), training separate 2D CNN-based networks with attention mechanisms for each orientation. Unlike transformer-based approaches that require extensive parameterization for cross-attention or self-attention operations, AXIAL employs lightweight attention modules that learn to weight spatial regions within each plane independently, then synthesizes these plane-specific attention maps into a unified 3D voxel-level rep-

resentation. This design leverages the transfer learning advantages of 2D architectures while reconstructing volumetric context through attention fusion, maintaining computational efficiency suitable for limited-data medical scenarios.

The framework produces two key outputs: a diagnostic prediction and a 3D attention map that, when overlaid on the original scan, generates an explainable MRI highlighting diagnostically relevant brain regions. Critically, AXIAL introduces a quantitative evaluation approach for XAI that measures the alignment between model-generated attention maps and established neuroanatomical biomarkers of AD such as hippocampal atrophy, entorhinal cortex degeneration, and ventricular enlargement. This validation against known disease signatures, rather than relying solely on qualitative visual inspection, provides evidence that the model focuses on clinically meaningful features consistent with the pathophysiological mechanisms discussed in Chapter 1.

The specific contributions of this chapter are summarized as follows:

1. A novel slice-based classification framework with integrated attention-based XAI that highlights brain regions highly correlated with AD without sacrificing diagnostic performance, achieving competitive accuracy even with small standardized datasets through efficient transfer learning.
2. Investigation of a double transfer learning strategy for distinguishing between stable MCI (sMCI) and progressive MCI (pMCI), demonstrating how knowledge transfer from an AD vs CN task enhances model sensitivity to subtle morphological changes indicative of early-stage disease progression.
3. Comprehensive evaluation using a standardized subset of ADNI [176] with rigorous subject-level data splitting and reproducible preprocessing pipelines, addressing the methodological concerns regarding data leakage and reproducibility discussed in Section 1.3.5. The complete framework implementation is publicly available at `https://github.com/GabrieleLozupone/AXIAL.git`.
4. A quantitative approach to measure the importance of specific brain regions in model decision-making, enabling systematic validation of XAI outputs against medical knowledge about AD neuroanatomy and facilitating clinical trust in model predictions.

The remainder of this chapter is organized as follows: Section 3.2 reviews related work on DL-based AD analysis with particular focus on slice-based approaches and current XAI challenges. Section 3.3 describes the standardized ADNI dataset and preprocessing pipeline employed. Section 3.4 details the AXIAL framework architecture, including the

multi-plane diagnosis networks, attention fusion mechanism, and XAI approach. Section 3.5.1 presents experimental settings, diagnostic results across multiple classification tasks, and both qualitative and quantitative XAI validation. Section 3.6 discusses key findings, compares AXIAL to transformer-based alternatives, and examines limitations. Section 3.7 concludes with a summary of contributions and implications for clinical deployment. The attention-based representation learning developed in this chapter will later inform the unsupervised latent space analysis in Chapter 4, where diffusion models learn semantic representations without diagnostic labels.

3.2 Related work

As anticipated in Sections 1.3.3 and 3.1, DL-based analysis of MRI images to diagnose and predict AD can be divided into three broad categories: (i) approaches that use 3D CNNs to analyze the entire 3D volume, (ii) approaches that analyze 3D patches from the whole 3D volume, and (iii) approaches that classify 2D slices. The following subsections report recent research activities belonging to these categories. Furthermore, we discuss the current challenges and issues in DL-based AD analysis, focusing on reliability and interpretability.

3.2.1 3D CNNs

In [7], the authors propose a fully convolutional 3D CNN model replacing typical max-pooling operations with standard convolution layers and test the model to distinguish AD, pMCI and sMCI from the entire subject's brain MRI scan. The results presented in [45] show that an MRI-based 3D CNN outperforms other neuroimaging biomarkers of neurodegeneration in prodromal AD and also outperforms amyloid or tau pathology biomarkers. More sophisticated approaches introduce attention mechanisms into 3D CNNs, e.g., the study presented in [78] investigates a novel attention-based 3D ResNet architecture to diagnose AD and explore potential biological markers. Also, a novel Attention-based 3D Multi-scale CNN model was proposed in [175] to capture better and integrate multiple spatial-scale features of AD. These approaches are very promising and comprehensive since analyzing the entire volume allows the extraction of global information on a spatial scale, unlike techniques based on conventional 2D CNNs. However, they are computationally expensive and require a large dataset during the training phase to generalize well. Furthermore, given the scarcity of pre-trained models, this problem cannot be easily addressed with a transfer learning strategy [83].

3.2.2 3D Patch-based

One way to reduce the computational cost and the overfitting risk of 3D CNN-based approaches consists of reducing the model inputs. For this reason, several approaches have been proposed to analyze 3D patches extracted from MRI scans. In [127], the authors present a novel computationally efficient patch-level training strategy to train a 3D CNN. They developed a 3D CNN model for the AD classification task, using random subvolumes of MRI scans as training data. This model generates patient-specific disease probability maps that are used to train a Multi-Layer Perceptron (MLP) to distinguish between AD and CN subjects. On the other hand, in [115], the authors introduce a framework that utilizes a 3D CNN to extract local features from 3D patches in brain MRI scans, forming the basis for patch-level responses. These responses are then processed through a dual-branch approach, combining patch classification and location identification. These methods enhance computational efficiency by focusing on patches of the 3D volume rather than processing the entire volume. However, the capacity for recognizing and learning global-level patterns within the data is limited. Moreover, the encoding process for 3D patches relies on 3D CNNs and restricts the application of transfer learning strategies.

3.2.3 2D Slice-level

An alternative to 3D CNNs are the more widespread 2D CNNs. Given the 3D nature of the input, to use 2D CNNs, 2D images must be extracted from the MRI image. This process is commonly known as slicing. Slicing can be performed according to one of the three planes of the MRI brain scan: (i) axial, (ii) coronal, and (iii) sagittal [188].

In [169], the authors propose an eight-layer custom CNN with leaky rectified learning units to classify single-slice MRI images. Recently, the work presented in [10] investigates the ability of different CNN and Transformer-based models to classify single slices selected from MRI with a mechanism based on Shannon entropy. Although using 2D CNNs for straightforward classification of 2D images is a standard practice, this method is insufficient for making direct predictions about an individual patient. To achieve accurate subject-level predictions, it is necessary to integrate results from multiple-slice analyses.

The study presented in [81] introduces an ensemble learning architecture based on 2D CNNs, using a multi-model and multi-slice ensemble, in which the majority voting scheme is used to merge the multi-slice decisions of each model. In [86], the authors use a fusion

transformer block to merge the outputs of a ViT [36] and a GFNet [130], both pre-trained on ImageNet. The proposed approach allows them to correlate features extracted from spatial and frequency domains, but predictions are made at the slice level. Thus, they are combined using a majority voting approach to make a decision at the subject level. Classifying each slice independently and then using majority voting eliminates the opportunity to learn patterns and features that span across slices.

To overcome this limitation, [3] introduced the Attention Transformer model. This model comprises a VGG-16 as a Base Network and a Transformer Unit. The Base Network extracts feature representations from each brain MRI image slice. Unlike Transformer architecture that generates a new representation of the input sequence through self-attention, the Transformer Unit performs a cross-attention operation. This operation consists of considering as Query the feature map relating to the central slice of the 3D volume and as Key and Value the feature maps of the others. This method allows the fusion of information from the entire volume, capturing the interrelationships between the central and surrounding slices. Similarly, [65] proposed the Conv-Swinformer architecture consisting of a CNN and a Transformer encoder module. The CNN module summarizes the planar features of the MRI slices, and the Transformer module establishes semantic connections in 3D space for these planar features. In this case, the Transformer encoder comprises four Swin Transformer blocks that perform self-attention operations on multiple windows extracted from the feature maps sequence. However, the reliance on cross-attention or self-attention in Transformer-based models, as suggested in [3, 65], poses challenges in medical scenarios with limited datasets. The extensive parameterization inherent in these mechanisms increases the risk of overfitting when training on smaller and often imbalanced medical datasets. This undermines the model's generalizability to new data, crucial in medical diagnostics.

The overparameterization problem can be addressed using less complex self-attention variants to build lightweight architectures. In the study [167], the authors proposed a diagnosis network composed of two innovative modules in combination with basic ResNet blocks. The first is the slice-aware module designed to interpret important slices and regions. This module performs self-attention between coded sections in an optimized way to be more parameter efficient. The second is the slice-shift module, which allows joint inter- and intra-slice modelling to exchange information with neighbouring slices.

3.2.4 Challenges in DL-Based AD Analysis

Despite the potential of DL in AD diagnosis and prognosis using MRI images, several critical issues undermine the reliability and reproducibility of the findings. Furthermore,

the "black box" nature of DL impacts clinical confidence and transparency of model decisions.

This subsection provides an analysis of the challenges in the AD analysis in terms of reliability, reproducibility and XAI.

Reliability and Reproducibility: Recent reviews [170, 188] highlight several reliability and reproducibility issues in DL-aided AD MRI analysis:

- **Wrong Data Split**: Data leakage due to incorrect data splitting is recurring. When datasets are not split at the subject level, data from the same subject may appear in both training and test sets, leading to overestimated model performance. This problem is particularly pronounced in patch or slice-based approaches [170].

- **Late Split in Data Processing**: Data processing steps such as data augmentation and feature selection must be conducted post-data splitting to avoid contamination of the test set. If these steps involve test data, it biases the model, rendering the evaluation unreliable [170].

- **Biased Transfer Learning**: Transfer learning, while beneficial, can introduce bias, particularly when the source and target domains overlap. For instance, using a network pre-trained on an AD vs CN task for an MCI vs CN task can cause bias if the CN subjects overlap in both tasks' training or validation sets [170].

- **Absence of an Independent Test Set**: The integrity of model evaluation is compromised if the test set is used for any purpose other than final performance evaluation. Hyperparameter optimization must rely on a separate validation set. The absence of an independent test set in some studies leads to inflated performance metrics [170].

- **Differences in Diagnostic Criteria**: Variability in diagnostic criteria for AD across different studies creates inconsistencies in ground truth labeling. This variability complicates the comparison of results across studies and hampers the development of universally applicable models [188].

- **Lack of Reproducibility**: A significant challenge in the field is the non-availability of frameworks and models for public use. Without access to open-source code and detailed implementation procedures, reproducibility is hampered. This lack of transparency affects the ability to validate and build upon existing work [188].

Furthermore, [170] highlights that the prevalent use of the 2D slice approach in AD analysis is often associated with a series of methodological problems. In particular, the authors highlight that among the numerous studies using this approach, only a few have successfully addressed the abovementioned challenges. This fact underlines the need for rigorous methodological standards in the field.

In response to these challenges, the framework discussed in this chapter was carefully developed to avoid common weaknesses identified in previous studies. Section 3.5.1 details the data splitting for model selection and data augmentation used for this study. Section 3.4.2 outlines the double transfer learning strategy investigated. To ensure reproducibility, we provide an open-source repository with all necessary code and implementation procedures at `https://github.com/GabrieleLozupone/AXIAL.git`.

Challenges in current XAI approaches: Several studies have highlighted the difficulties that prevent using DLADS on medical images due to the current challenges of XAI methods [161, 145, 166]. XAI researchers often use self-intuition to determine a good explanation without validating with a medical professional. This missing validation is partially due to the inability to provide quantitative but only qualitative explanations. Qualitative explanations provided for small and specific subsets of subjects make a global evaluation and comparison of XAI techniques difficult. Furthermore, there is no specific correlation between the prediction and the associated brain region. Model-agnostic XAI techniques, such as GradCAM, often produce saliency maps that highlight extensive regions of the brain without sufficient specificity or consistency. Another critical limitation of GradCAM-like methods, mainly when applied to the analysis of 2D slices from MRI scans, is their focus on intra-slice features while neglecting the inter-slice relationships crucial for understanding the 3D brain regions related to AD. Aggregating these saliency maps to form a 3D visualization produces an overly generalized activation map that often highlights different and sparse areas.

As an alternative to traditional saliency-based methods, recent advancements have explored the use of attention-based models to enhance the interpretability of DL models. In this context, [77] proposed a trainable attention mechanism to highlight where and in what proportion the network paid attention to input images for classification. Attention amplifies relevant areas and suppresses irrelevant ones, improving interpretability. In [140], grid attention captures anatomical information in medical images, and attention coefficients were used to explain which areas of the image the network focused on. Recently [167], described in Section 3.2.3, introduced a slice-aware module to advance XAI in their methodology. Their variant of the attention mechanism enables the extraction of attention weights to determine the importance of each slice in the decision-making

process, offering a more precise understanding of model focus.

3.3 Materials

Table 3.1: Summary of participant demographics, mini-mental state examination (MMSE) and global clinical dementia rating (CDR) scores

Group	Subjects	Samples	Age	MMSE	CDR
CN	204	586	76.31 ± 5.22	29.11 ± 1.05	0.01 ± 0.16
AD	191	474	75.23 ± 7.34	22.46 ± 3.38	0.86 ± 0.43
pMCI	121	162	75.01 ± 6.68	26.47 ± 1.83	0.481 ± 0.09
sMCI	110	154	75.09 ± 7.02	27.80 ± 1.74	0.45 ± 0.17

The ADNI MRI Core in [176] has created standardized dataset collections comprising scans that met minimum quality control requirements to promote greater rigour in analysis and meaningful comparison of different algorithms. These standardized datasets within the ADNI archive allow researchers to download a complete and consistent set of images efficiently, thus facilitating comparative research into neurodegenerative diseases. These collections represent a small portion of the ADNI dataset and generally are not used in DL literature. However, we chose the "ADNI1: Complete 1Yr 1.5T" collection as it directly addresses the lack of reproducibility, thereby facilitating model performance comparisons. This collection of 1.5 Tesla MRI scans includes only 639 subjects of the 2.720 available in the ADNI study. All subjects have screening, along with 6 and 12-month follow-up scans. Furthermore, this small data subset is an excellent candidate for testing our method in the recurrent medical imaging data scarcity condition.

We converted the dataset into the Brain Imaging Data Structure (BIDS) format [56]. An advantage of converting the ADNI dataset to a BIDS structure is the availability of specialized Python libraries for handling BIDS-structured data. These libraries, such as PyBIDS [181], provide powerful tools for querying, organizing, and processing neuroimaging data. To perform this conversion, we used Clinica [134, 137], an open-source software platform explicitly designed to make clinical neuroscience studies more accessible and reproducible.

During the conversion phase, the ADNI-to-BIDS converter provided in Clinica checks whether the data meets specific criteria, e.g. images that fail quality control are filtered out or, if there are multiple scans for a visit, the 'preferred scan' is selected according to specific decision rules [134]. Consequently, some of the MRIs in the original dataset

may be discarded at this stage. The "ADNI1: Complete 1Yr 1.5T" dataset contains 2,042 acquisitions from 639 subjects. During the conversion, 151 scans did not meet the criteria, so the resulting dataset was 1,891 samples without varying the number of subjects. Each subject had a maximum of 3 scans from baseline, 6-month and 12-month follow-up, and consequently, samples from the same patient with a different diagnosis are unlikely. Therefore, to define sMCI patients and pMCI patients, we used the complete clinical data from ADNI, which contains the diagnoses after several years of follow-up.

From the resulting dataset, we have extracted four classes:

- CN: 3D brain images of patients diagnosed as CN at the time of image acquisition.
- AD: 3D brain images of patients diagnosed with AD at the time of image acquisition.
- sMCI: 3D brain images of patients diagnosed with MCI at the time of image acquisition and who remain MCI over time.
- pMCI: 3D brain images of patients initially diagnosed with MCI at the time of scan but later diagnosed with AD 36 months post-acquisition.

The demographic information of the subjects in the "ADNI1: Complete 1Yr 1.5T" dataset is presented in Table 3.1, which provides essential context for our analysis and findings.

3.4 Methods

This section presents the explainable diagnostic pipeline developed in this study. The process starts with a pre-processing phase described in Fig. 3.2. The XAI diagnosis network illustrated in Fig. 3.3 processes the images as a series of 2D slices, producing diagnosis and attention weights to synthesize the slice's importance in decision-making. Finally, the attention weights of the slices across the three distinct views are combined to create a 3D attention map, as shown in Fig. 3.4.

In the following sections, we provide full details of each step.

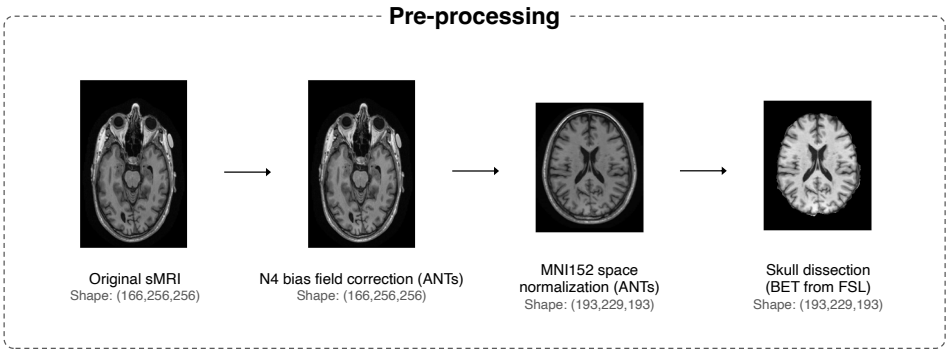


Figure 3.2: MRI data pre-processing pipeline: (i) Original sMRI, (ii) N4 bias field correction, (iii) MNI152 space normalization, and (iv) skull dissection.

3.4.1 MRI images pre-processing

As mentioned in Section 3.2.4, classification results are difficult to compare between studies due to their limited reproducibility. In the study [170], the authors highlight that variations in components such as participant selection and image pre-processing are critical aspects that directly influence this limitation. It is, therefore, essential to avoid ambiguity in both cases to facilitate the comparison between models. In Section 3.3, we described the procedures required to avoid ambiguity in participant selection. The pre-processing pipeline used comprises three steps:

1. Bias field correction using the N4ITK method [160];

2. Affine registration using the SyN algorithm [5] from ANTs [6] to align each image to the Montreal Neurological Institute (MNI) 152 space with the ICBM 2009c nonlinear symmetric template [50, 49];

3. Skull stripping to remove non brain-tissue from the 3D image using Brain Extraction Tool [146] from FSL [76].

We used the t1-linear pipeline of Clinica [134, 170] to perform the first two steps.

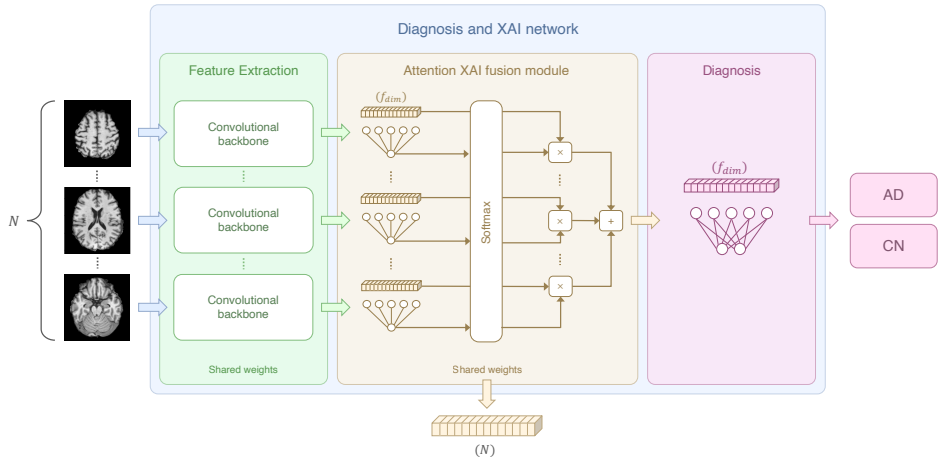


Figure 3.3: The proposed Diagnosis and XAI network: (i) Feature Extraction module, (ii) Attention XAI Fusion module, and (iii) Diagnosis module.

3.4.2 Diagnosis network

Radiologists typically review a series of 2D slice images when analyzing medical imaging data, even when examining 3D images. With this in mind, we have developed an efficient and explainable network for diagnosing AD that learns inter-slice relationships. This network, as illustrated in Fig. 3.3, comprises three key modules described in detail in this section.

Feature Extraction: This module employs a pre-trained 2D convolutional backbone to process a series of N 2D slices. The convolutional backbone, which can be any architecture such as VGG or ResNet, is pre-trained on the ImageNet dataset to handle cases of small dataset sizes in AD diagnosis tasks.

The original data are 3D images with a value for each voxel, so the resulting sliced 2D images are 1-channel images. Since pre-trained weights are for 3-channel images, we summed the pre-trained convolutional filters of the backbone's first layer. Since these filters operate linearly and sum their results across channels, we can sum all the filters over the input channel dimension in the first layer. This operation is equivalent to replicating the single-channel image three times, but it is more computationally efficient. The

extracted slices were resized to 224×224 and normalized.

Each convolutional backbone ends with a max-pooling operation that converts the feature map of each slice into a feature vector of dimension f_{dim}. This sequential input processing is achieved through sharing convolutional weights across the sequence, similar to a Recurrent Neural Network approach:

$$f_i = \text{MaxPool}(\text{Conv}(...(x_i)))$$

where x_i represents the i-th slice in the sequence, f_i is the resulting feature vector for that slice, and Conv denotes the convolutional operations performed by a convolutional architecture.

Double Transfer Learning: Models pre-trained on large datasets can help overcome challenges posed by limited data, but they may not be enough for complex tasks. As shown in Table 3.1, the data available for training a network to differentiate between AD and CN individuals is usually more than for distinguishing between sMCI and pMCI. Although both tasks are complex, the latter is considered more difficult. Therefore, it is more likely that transfer learning can distinguish AD from CN more effectively than recognizing the difference between sMCI and pMCI. Since doctors rely on their expertise to assess a patient's cognitive decline, leading to an AD diagnosis, we suggest fine-tuning a model pre-trained on the AD vs. CN task to the sMCI vs. pMCI task. Following our definition of classes in Section 3.3, the $AD \cup CN$ set is disjoint from $(sMCI \cup pMCI) \subset MCI$. As a result, adopting this strategy under these assumptions does not imply data leakage issues.

Attention XAI Fusion Module: This module enables the network to learn inter-slice dependencies and global 3D patterns through the Feature Extraction module. During backpropagation, the shared weights of the convolutional module are updated based on the entire image, thus enabling the learning of 2D patterns in relation to the importance of that slice at a global level. The Attention XAI Fusion module leverages a Fully Connected (FC) layer to assign importance to each section based on its feature vector. This introduces a parameter-efficient approach, with only $f_{\text{dim}} + 1$ learnable parameters:

$$w_i = \text{FC}_{attention}(f_i)$$

where w_i is the computed weight for slice i, and FC denotes the FC layer operation. Normalization of these weights is performed using a softmax function, ensuring they sum to one and represent the slice-importance distribution:

$$\alpha_i = softmax(w_i) = \frac{e^{w_i}}{\sum_{j=1}^{N} e^{w_j}}$$

where α_i is the normalized importance weight for the i-th slice. The module performs fusion by calculating a weighted sum of the feature vectors using their corresponding weights, generating a composite feature vector representing the entire brain image:

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

This module yields a feature vector F that synthesizes the brain image and a set of attention weights $\{\alpha_i\}$ highlighting the significance of each slice.

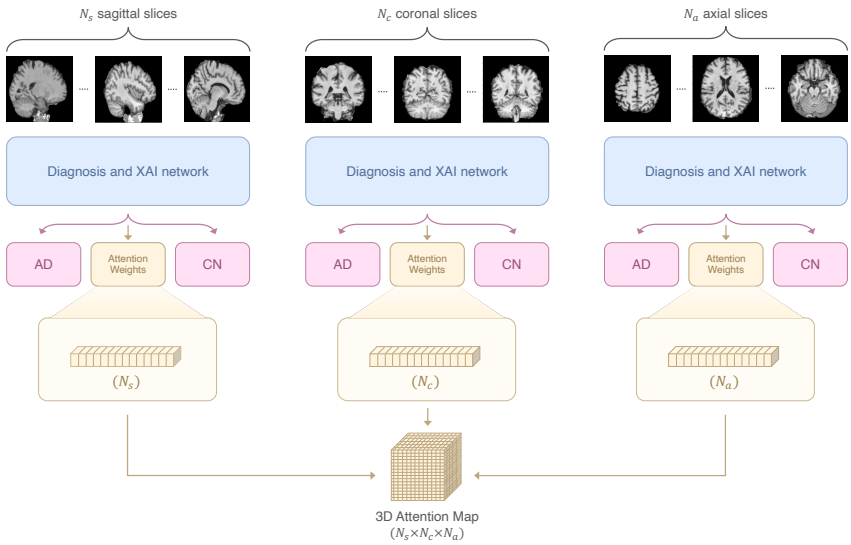


Figure 3.4: Representation of the proposed XAI attention-based approach. Three distinct networks process each slicing plane, and the slice attention weights from each plane are combined to produce a 3D attention map.

Diagnosis: In the final *Diagnosis* module, the synthesized feature vector F is fed into a head network comprising a FC layer with two output neurons for the binary classification task (e.g., AD vs. CN):

$$D = softmax(\mathrm{FC}_{\mathrm{head}}(F))$$

D represents the softmax output, providing a probability distribution over the diagnosis classes, and FC_{head} denotes the FC layer specific to the diagnosis task.

3.4.3 XAI Approach

Within our diagnostic framework for AD, we introduce an advanced XAI strategy to generate a 3D attention map, facilitating a deeper understanding of the neural network's diagnostic focus. As shown in Fig. 3.4, our approach harnesses attention weights across sagittal, coronal, and axial slicing planes, merging these into a comprehensive representation that highlights key brain regions indicative of AD.

Attention Weight Generation: The implemented XAI method yields attention weights for each of the three principal slicing planes of the brain: sagittal (s), coronal (c), and axial (a). We trained a dedicated Diagnosis and XAI Network for each plane to assess slice importance distribution from different views. The MRI brain image is sliced in the three planes, resulting in three sequences of 2D slices: N_s, N_c, and N_a slices for the sagittal, coronal, and axial planes.

Each network's Attention XAI module computes a set of attention weights α_s, α_c, and α_a for its corresponding plane. The computation is formalized as follows:

$$\alpha_p = \textit{softmax}(\text{FC}_{attention}(f_p))$$

where $p \in \{s, c, a\}$ represents each plane, f_p denotes the feature vectors derived from slices within that plane.

3D Attention Map Synthesis: 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 within the brain's imaging volume:

$$A[i, j, k] = \alpha_s[i] \cdot \alpha_c[j] \cdot \alpha_a[k]$$

For every voxel at coordinates $[i, j, k]$, its attention value is derived from the product of the corresponding sagittal, coronal, and axial attention weights. This process retains and amplifies the diagnostic importance perceived across all three anatomical planes, yielding a comprehensive 3D depiction of attention distribution.

To facilitate interpretation and comparison, we normalize the entire 3D attention map in the range [0, 1], ensuring that the highest values correspond to areas of high diagnostic

relevance. We used min-max normalization:

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

The resulting map highlights the brain regions most associated with the diagnostic output of the network, offering insights into the pathological hallmarks of AD as learned by the model through its training process.

Brain regions importance quantification: 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 $I_{norm} = f_{MNI152}(I)$. This step is crucial for aligning brain structures across different individuals to facilitate the identification of AD-specific biomarkers.

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

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

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

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

where δ is an amplification factor set to 10 to increase the prominence of the heatmap overlay.

Employing a template atlas, we determined Region of Interest (ROI) within the brain. For each ROI identified with a unique label r, we computed the overlap O_r as follows:

$$O_r = \sum_{i,j,k} (H_{\text{binary}}[i,j,k] > 0) \wedge (T_{atlas}[i,j,k] = r)$$

where T_{atlas} represents the template atlas data. The quantitative evaluation of each region's overlap involved several statistical measures calculated from the MRI data. The overlap volume of the heatmap in a given region is:

$$V_r = |O_r|$$

which can be useful to indicate predominant regions in decision process. The mean (μ_r) intensity of the heatmap within each identified region is given by

$$\mu_r = \frac{1}{V_r} \sum_{(i,j,k) \in O_r} A[i,j,k],$$

which represents the average activity level across the voxels of interest, allowing us to pinpoint regions with consistently high AD-related changes. The standard deviation (σ_r) of the heatmap intensities,

$$\sigma_r = \sqrt{\frac{1}{V_r - 1} \sum_{(i,j,k) \in O_r} (A[i,j,k] - \mu_r)^2},$$

provides insight into the variability of attention within the region, indicating the heterogeneity of AD impact across different brain areas. The maximum ($A_{max,r}$) and minimum ($A_{min,r}$) values of the heatmap,

$$A_{\text{max},r} = \max_{(i,j,k) \in O_r} A[i,j,k]$$

and

$$A_{\text{min},r} = \min_{(i,j,k) \in O_r} A[i,j,k],$$

respectively, highlighting the extremes in activation within the regions, shedding light on the most and least affected areas within the brain's AD-implicated regions. Lastly, the percentage of overlap (P_r) is calculated as

$$P_r = \frac{\sum_{i,j,k} H_{\text{binary}}[i,j,k]}{V_r}$$

to provide a single value that quantifies the portion of a specific brain region involved in the AD-related patterns identified by the network. This measure is a critical indicator of the importance of each region in distinguishing AD patients from CN individuals.

3.5 Experiment results

In this section, we first present the experimental setup developed to validate the method. Then, we present the results of the model in comparison to recent state-of-the-art attentional models on 2D slices. Finally, we present qualitative and quantitative results of our XAI approach compared to other existing XAI methods.

3.5.1 Experiment Setup

We implemented all considered methods in PyTorch [116] and trained them on an NVIDIA A100 80GB GPU. The pre-processed input 3D images (see Section 3.4.1) were converted to a sequence of N slices at runtime. The N slices were selected from the center by a slicing operation from the chosen plane. Specifically, given D, the dimension corresponding to the number of slices along the slicing direction, the selected slices belonged to the interval $[D/2 - N/2, D/2 + N/2]$. This slicing operation resulted in $N < D$ slices, thus reducing computational complexity with N considered as an optimization hyperparameter. All tested networks were trained with AdamW Optimizer with base learning rate 1×10^{-4} and weight decay 1×10^{-2}. An early stopping strategy selected the best model during the training phase on the validation set with patience 15. The training data augmentation strategy consisted of randomly flipping a 2D slice in the series with 0.3 probability. We implemented a 5-fold cross-validation technique to ensure robustness and generalizability of the results. We divided the dataset into five segments, ensuring each segment acted as a test set in one iteration and as part of the training/validation sets in the others. To prevent data leakage and preserve the integrity of our evaluation, we carefully assigned all 3D images from the same subjects exclusively to one of the training, validation, or test sets. In each fold, the 80% portion allocated for training/validation was further divided using an 80-20 ratio.

Performance Evaluation: We used ACC, SEN, SPE, and MCC as performance metrics. ACC is the most common performance measure for classification, and it consists of the percentage of correctly classified samples over the total. MCC [14] is a correlation coefficient between predictions and true labels; it provides a more informative and truthful score than accuracy when evaluating binary classifications, allowing a more realistic interpretation of classifier performance, especially in the case of unbalanced datasets. SEN and SPE serve as critical metrics for evaluating the diagnostic accuracy of our model, highlighting its ability to correctly identify positive and negative cases, respectively. In AD vs. CN task, the AD samples are in the positive class, and CN samples are in the negative one. In sMCI vs. pMCI task, the sMCI samples are in the positive class, and pMCI samples are in the negative one.

All the results obtained are averaged on the test sets from the 5-fold cross-validation.

3.5.2 Diagnostic results

This subsection focuses on our diagnosis network's results, summarizing the evaluations of various 2D convolutional backbones, the impact of different slicing planes, optimization of network parameters, the effectiveness of double transfer learning, and a comparison with state-of-the-art methods. Almost all experiments use the axial slicing plane due to its common use in clinical applications and because it generally yields better results.

2D backbone: As the first step, we analyzed the effectiveness of different convolutional backbones in the AD vs. CN task. To this aim, two sets of experiments were performed. First, we evaluated the performance of different backbones by considering them in the Feature Extraction Module of our network. Second, to allow an evaluation independent of our method, the best networks for each family were evaluated with a 2D Majority Voting approach, which is also considered one of the baseline methods. In the Majority Voting approach, we train 2D convolutional networks by individually labelling the slices. Consequently, the weights are updated from the slice-level error during training. Then, the predictions of the slices of a single 3D image are aggregated by choosing the most recurring one in the sequence as the final label.

Table 3.2: Performances varying backbone in Feature Extraction module in AD vs. CN task averaged over 5-fold cross-validation. Best results in **bold**, while second-best are underlined.

Backbones	ACC	SPE	SEN	MCC
VGG16	**0.829**	0.839	**0.816**	**0.655**
VGG19	0.819	**0.875**	0.749	0.633
ResNet34	0.770	0.843	0.681	0.534
ResNet50	0.804	0.860	0.734	0.602
ResNet101	0.785	0.822	0.739	0.563
EfficientNetV2S	0.764	0.805	0.713	0.521
EfficientNetV2M	0.764	0.833	0.679	0.520
DenseNet121	0.798	0.841	0.745	0.590

The first pool of experiments was conducted using a mini-batch of 16 and freezing the first 50% of the backbone's layers. The convolutional architectures chosen are: (i) VGG16 and VGG19 [144], (ii) ResNet34, ResNet50 and ResNet101 [62], (iii) EfficientNetV2 Small and EfficientNetV2 Medium [156] and (iv) DenseNet121 [67]. The results of this pool are shown in Table 3.2. In this first analysis, VGG16 had the highest accuracy (ACC: 0.829) and sensitivity (SEN: 0.816), suggesting it is the better-suited backbone for AD

patient identification. VGG19, while slightly less accurate, showed superior specificity (SPE: 0.875), indicating it is particularly adept at identifying CN cases. The ResNet series demonstrated that additional depth (as seen in ResNet101) does not equate to improved performance, with ResNet50 showing better accuracy and sensitivity. This could suggest diminishing returns with increased complexity for this task. EfficientNet models trailed behind their counterparts with identical accuracies (ACC: 0.764). DenseNet121, while providing good accuracy (ACC: 0.798), also did not meet the higher results obtained by other architectures. The MCC aligns with these findings, confirming VGG16 as the top-performing model in this case.

Based on the results obtained in Table 3.2, we selected the best backbone from each architecture family as candidates for the second pool of experiments. Since this pool comprises 2D rather than 3D images, we chose 32 as the mini-batch size. To achieve comparable results, the same percentage of backbone layers was frozen. As shown in Table 3.3, the performance trends of the architectures remain consistent with previous observations. VGG16 continued to lead in accuracy (ACC: 0.804) and MCC (MCC: 0.605). This result aligns with other works that design their approaches relying on a VGG as base network [3, 65]. Remarkably, our method outperforms the Majority Voting one across several metrics. The superiority can be seen from the average increase across all tested backbones of ACC by 1.6% and MCC by 3.08%. Notably, a good increase was observed with VGG16, with an ACC improvement of 2.5% and an MCC improvement of 5.0%.

Table 3.3: Performances varying backbone in Majority Voting approach in AD vs. CN task over 5-fold cross-validation. Best results in **bold**, while second-best are underlined.

Backbones	ACC	SPE	SEN	MCC
VGG16	**0.804**	**0.897**	0.688	**0.605**
ResNet50	0.798	0.873	**0.707**	0.592
EfficientNetV2S	0.759	0.839	0.664	0.516
DenseNet121	0.770	0.848	0.673	0.532

Slicing plane: Although the axial plane is the most commonly used in clinical applications, there could be better choices than this one in this classification scenario. We performed a slicing plane analysis after finding that VGG16 is the best backbone in the diagnostic task. For this set of experiments, we used 100 slices for each view, a batch size of 8, a learning rate 1e-4, and froze the first half of the backbone in training. Table 3.4 shows the performances averaged over the five cross-validation test sets varying the slicing plane. The results show that the axial plane achieved the best performance. This finding aligns with the clinical choice.

Table 3.4: Performances varying slicing plane on 100 slices with our approach in AD vs. CN task over 5-fold cross-validation. Best results in **bold**, while second-best are underlined.

Slicing plane	ACC	SPE	SEN	MCC
Axial	**0.839**	0.872	**0.799**	**0.674**
Coronal	0.816	**0.897**	0.715	0.628
Sagittal	0.820	0.860	0.772	0.636

Parameters Optimization: Since VGG16 was confirmed as the best backbone among those under investigation and the axial plane provided the best results, we selected this configuration for our Feature Extraction module. The second step consisted of optimizing other parameters: the number of slices extracted N, the batch size, and the percentage of backbone to freeze.

Table 3.5: Network Results using VGG16 as the backbone for Feature Extraction module in AD vs. CN task. Best results in **bold**, while second-best are underlined.

Num Slices	Batch Size	Freezing	ACC	SPE	SEN	MCC
70	8	50%	0.821	0.851	0.782	0.636
80	4	50%	0.825	0.867	0.774	0.646
	8	0%	0.818	0.861	0.764	0.630
		25%	0.840	0.892	0.776	0.677
		50%	**0.856**	**0.910**	0.792	**0.712**
		75%	0.799	0.877	0.707	0.603
	16	50%	0.829	0.839	**0.816**	0.655
90	8	50%	0.818	0.867	0.757	0.630
100	8	50%	0.839	0.872	0.799	0.674
120	8	50%	0.815	0.858	0.761	0.624

The results in Table 3.5 indicated that varying the number of slices and the percentage of first backbone layers frozen in the network impacted performances. The configuration with 70 slices, batch size of 8, and freezing 50% of the layers resulted in an ACC of 0.821 and MCC of 0.636. When the batch size was reduced to 4 under the same conditions, a slight improvement was observed with an ACC of 0.825 and MCC of 0.646. After increasing the number of slices to 80 and employing a batch size of 8, experiments with 0%, 25%, 50%, and 75% freezing were conducted. The highest accuracy was achieved at 50% freezing, yielding an ACC of 0.856 and MCC of 0.712, highlighting the best overall performance across all configurations. With a further increase to 90 slices, the performance

did not improve, maintaining an ACC of 0.818 with a 50% freezing and batch size of 8. Finally, the performance slightly fluctuated with 100 and 120 slices but generally did not surpass the optimal results obtained with 80 slices end 50% freezing.

Table 3.6: Double transfer learning for sMCI vs. pMCI task at 36 months. Best results in **bold**, while second-best are underlined.

Transfer Learning	Freezing	ACC	SPE	SEN	MCC
ImageNet	50%	0.4822	0.427	0.636	0.065
ImageNet + AD vs. CN	50%	0.703	**0.809**	0.4873	0.396
ImageNet + AD vs. CN	75%	**0.725**	0.763	**0.678**	**0.443**

Double Transfer Learning: The third step involved investigating the utility of transfer learning for predicting AD progression with the double transfer learning strategy proposed in Section 3.4.2. This approach leverages a backbone pre-trained on the ImageNet dataset, further fine-tuned through a subsequent task distinguishing AD from CN individuals. As mentioned in Section 3.4.2, it is crucial to note that the patient cohorts classified as AD or CN differ from those labeled as sMCI or pMCI, given that the latter are diagnosed as MCI. Consequently, this strategy does not cause data leakage problems.

First, we fine-tuned our model with VGG16 pre-trained on ImageNet in the Feature Extraction module on the entire AD vs. CN dataset. We froze the first 50% of the layers as this configuration provided the best results in Table 3.5. The final model pre-trained on the AD vs. CN task was selected using 10% of the dataset as a validation set. This step provides a model in which the first convolutional part is highly specialized in extracting fine-grained low-level feature representations and the rest in analyzing AD-related high-level patterns. Second, we further fine-tuned the AD pre-trained model on the sMCI vs. pMCI classification task at a 36-month forecast interval. The learning rate for this pool of experiments was reduced to 1×10^{-5} to fine-tune the pre-trained model weights, ensuring that the disease knowledge is preserved during the training process. Table 3.6 demonstrates the efficacy of this approach in this task. With baseline ImageNet transfer learning and 50% of the network layers frozen, the model achieved an ACC of 0.4822 and an MCC of 0.065. The MCC value of 0.065 indicates that the model could not find useful correlations to distinguish the two cases. Incorporating AD vs. CN knowledge with a 50% freezing resulted in a marked enhancement across metrics: an ACC of 0.703 and an MCC of 0.396. Advancing to a 75% freezing of the network layers under the double transfer learning paradigm, we observed a peak performance of an ACC of 0.725, SPE of 0.763, SEN of 0.678, and an MCC of 0.443.

Table 3.7: Comparison of performance against other approaches. Best results in **bold**, while second-best are underlined.

(a) AD vs. CN classification task

Networks	ACC	SPE	SEN	MCC
Attention Transformer	0.826±0.043	**0.914**±0.030	0.721±0.068	0.654±0.077
AwareNet Diagnosis	0.832±0.032	0.876±0.055	0.781±0.060	0.664±0.062
Ours	**0.856**±0.021	0.911±0.038	0.792±0.033	**0.712**±0.042
Majority Voting	0.804±0.048	0.898±0.039	0.691±0.090	0.608±0.088
Attention-Guided Voting	0.843±0.023	0.894±0.040	0.780±0.034	0.683±0.043
Majority Voting 3D	0.836±0.040	0.866±0.052	**0.801**±0.056	0.671±0.077

(b) sMCI vs. pMCI classification task

Networks	ACC	SPE	SEN	MCC
Attention Transformer	0.624±0.084	0.684±0.271	0.573±0.323	0.265±0.209
AwareNet Diagnosis	0.541±0.048	**0.788**±0.186	0.269±0.288	0.056±0.150
Ours	**0.724**±0.063	0.766±0.072	**0.678**±0.088	**0.443**±0.132
Majority Voting	0.613±0.048	0.609±0.105	0.641±0.099	0.249±0.097
Attention-Guided Voting	0.633±0.064	0.624±0.073	0.643±0.088	0.266±0.132
Majority Voting 3D	0.630±0.053	0.665±0.194	0.605±0.228	0.287±0.123

Comparison to state-of-the-art methods: We performed a comparison with current 2D slice attention-based state-of-the-art models. These methods include Attention Transformer [3] and AwareNet [167]. Attention Transformer uses a multi-head self-attention to perform a cross-attention operation resulting in a slice feature maps fusion. AwareNet represents the diagnosis network of a joint learning framework that uses a slice-aware module and a slice-shift module. The slice-aware module leverages the attention mechanism to interpret significant slices and regions. While we have utilized the official implementation of AwareNet, as it is available and provided by the authors, we have implemented the architecture for the Attention Transformer ourselves based on the paper details, as the authors did not provide the official implementation. For both methods, the hyperparameters were optimized to maximize performance. As in our case, Attention Transformer obtained the best results by freezing 50% of the base network pre-trained on ImageNet and a batch size of 8. The AwareNet network was trained from scratch with a batch size of 2. The network is initialized by the Kaiming method [63] and trained using the Adam optimization algorithm with $\beta_1 = 0.48$ and $\beta_2 = 0.999$. In both cases, the learning rate was set to 1×10^{-5} to prevent overfitting.

In the sMCI vs. pMCI task, we adopted the double transfer learning strategy proposed in this work for the Attention Transformer method. The first 75% of the Attention Transformer base network layers were frozen. Since the AwareNet is trained from scratch on the AD vs. CN task, we froze the first 50% of the backbone.

The results of our comparative analysis are summarized in Table 3.7, showing results for both (a) the AD vs. CN diagnosis task and (b) the sMCI vs. pMCI prediction task. To assess the effectiveness of our attentional module in enhancing diagnostic performance, we introduced three additional approaches: Majority Voting, Attention-Guided Majority Voting, and Majority Voting 3D. We detailed the Majority Voting approach implemented in Section 3.5.2. The Attention-Guided Majority Voting approach utilizes the network trained in the majority voting case. However, predictions are made during testing on a subset of slices identified as most informative by our attentional fusion module. Specifically, we identified a contiguous range where the attention values exceeded the 75th percentile. This range was calculated for the axial plane by averaging the attention weight distribution across the five validation sets of the fold cross-validation, resulting in the range [0, 25]. The results of this method validate the module's ability to select feature maps with high information content. Finally, the Majority Voting 3D approach replaces the attentional fusion module with an averaging module, where feature maps from individual slices are averaged to form a single feature map for making subject-level decisions. This approach further helps to evaluate the impact of slices' importance weighting compared to considering all slices as equally important.

Our model demonstrates effectiveness in disease diagnosis and prediction tasks, achieving (i) an accuracy (ACC) of 0.856 and a Matthews correlation coefficient (MCC) of 0.712 in the AD vs. CN comparison, and (ii) an ACC of 0.725 and MCC of 0.443 in the sMCI vs. pMCI task. While AwareNet shows promise in the diagnosis task with an ACC of 0.832 and an MCC of 0.659, it struggles in the disease prediction task, yielding an MCC of 0.039, likely due to insufficient data for training. On the other hand, Attention Transformer, leveraging double transfer learning, provides a more balanced alternative, with an MCC of 0.651 in the diagnosis task and an MCC of 0.238 in disease prediction. However, its increased parameterization from self-attention prevents it from bridging the gap with lightweight models like ours and AwareNet in the diagnosis task. Majority voting, lacking a mechanism to learn correlations between sections, underperforms compared to other methods. Incorporating subject-level error evaluation through the Majority Voting 3D approach the MCC improves from 0.605 to 0.667, even surpassing overfitting-prone approaches such as AwareNet and Attention Transformer. Furthermore, leveraging our attentional mechanism to identify informative slices while using networks pre-trained without considering slice relationships yields an MCC of 0.683. This result underscores our model's ability to select crucial slices for decision-making.

3.5.3 XAI results

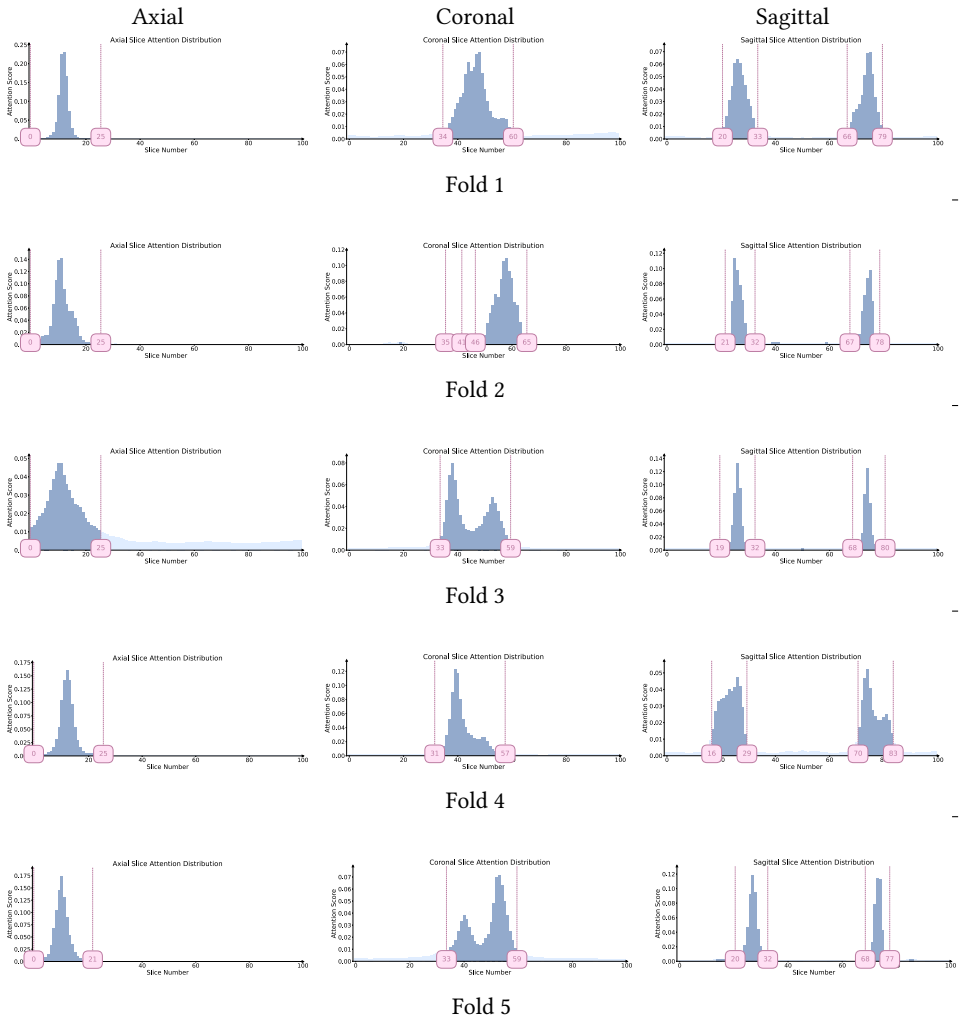


Figure 3.5: Average attention weight distributions generated by **our model** for each fold and each plane

This section analyzes the interpretability of our approach and those proposed in [167, 3]. Our XAI method described in Section 3.4.3 allowed us to produce a 3D attentional map starting from the attentional weight distributions of the axial, coronal, and sagittal planes. The authors of AwareNet [167] designed the slice-aware module of this network

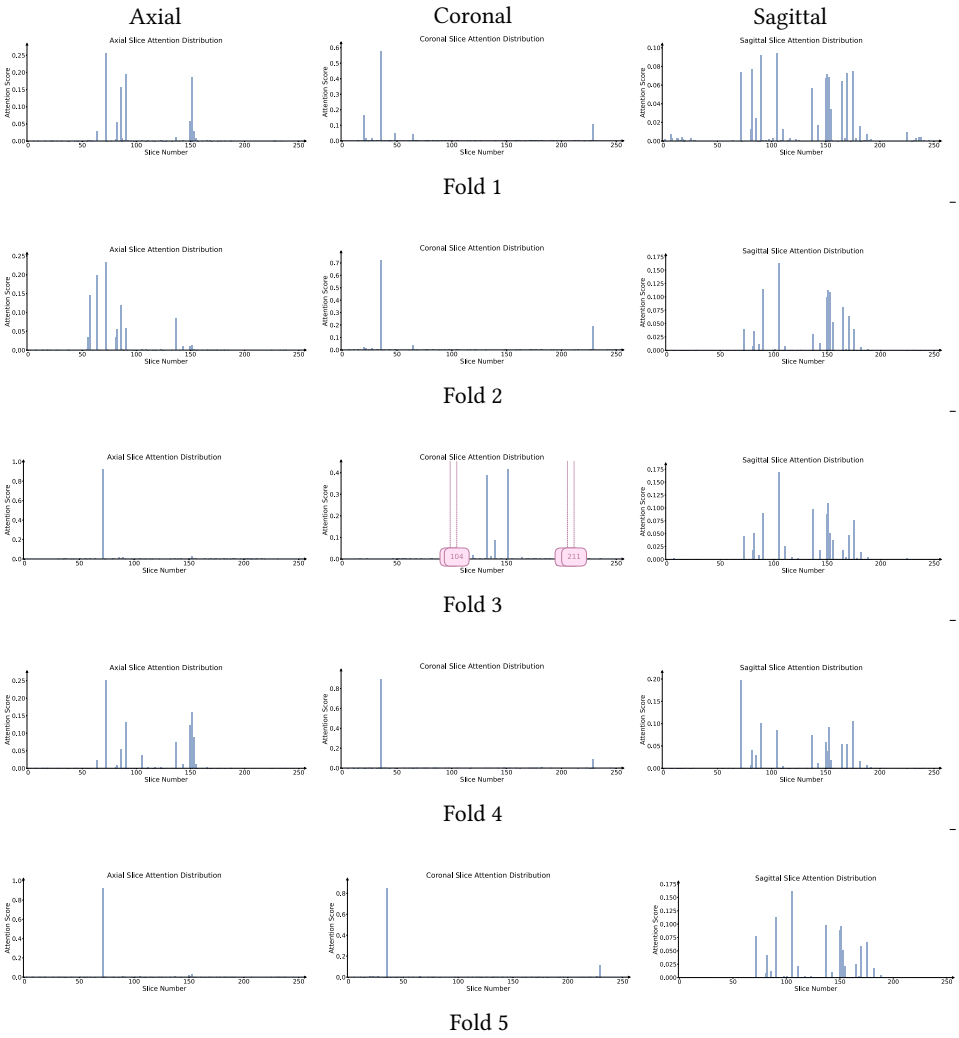


Figure 3.6: Average attention weight distributions generated by **AwareNet** for each fold and each plane

to extract, as in our case, a distribution of attentional weights capable of summarizing the importance of each slice in the decision-making process. As a result, our approach can produce a 3D map also using the model proposed in [167].

A recurring approach to interpreting decisions made by DL models is GradCAM [141].

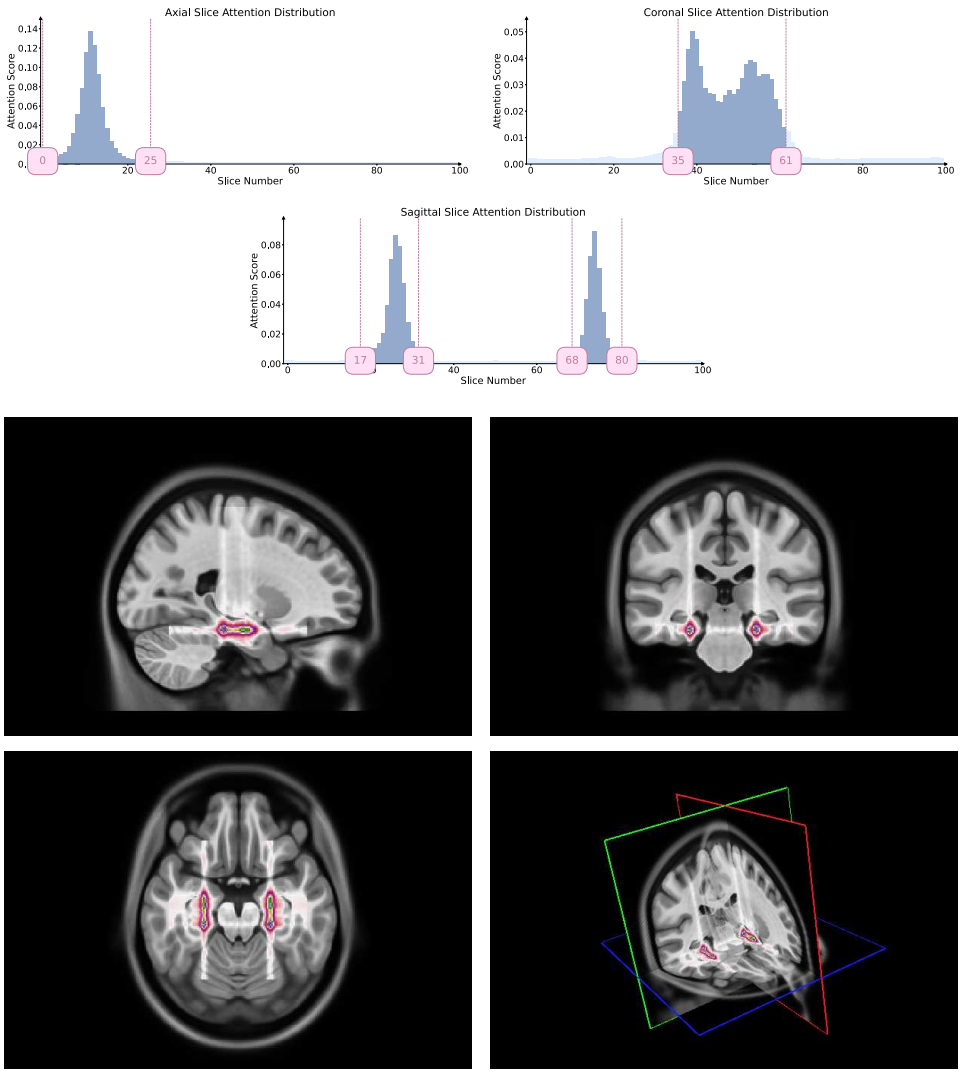


Figure 3.7: (**Top**) Attention distribution for each plane averaged on all the five test to provide entire dataset distributions; (**Bottom**) Visualization of mean 3D attention map of entire dataset overlapped to MNI152 template in Freeview to analyze the localization performance of our framework.

GradCAM is a technique for visualizing regions within an image that influence the classification decision of a convolutional neural network. It works by selecting a target con-

volutional layer, calculating the class score gradients relative to that layer's feature maps, averaging these gradients, and using them as weights to create a weighted sum of the feature maps. This sum is then passed through a ReLU function to produce a saliency map highlighting the areas that positively impact the class decision. However, GradCAM in this comparison scenario allows the production of 2D saliency maps, one for each slice, which deliver a qualitative but not quantitative result. To allow a fair comparison with other models, we have developed a variant of our XAI approach that allows us to generate 3D maps with GradCAM. The idea is to stack the produced 2D CAMs on top of each other to generate a 3D saliency map for each plane. Given A, S, and C, the 3D salient maps generated for the respective axial, sagittal, and coronal planes, the final 3D salient map is obtained pointwise as:

$$M[i, j, k] = S[i, j, k] \cdot C[i, j, k] \cdot A[i, j, k]$$

This 3D map can be used to quantitatively evaluate the impact of each brain area in the decision-making process with the metrics proposed in Section 3.4.3. This method also allows us to compare with Attention Transformer model proposed in [3], which is not designed to yield attentional weights for each slice directly. In our implementation, we applied GradCAM to the last layer of the convolutional backbone. This choice is motivated by the fact that the last convolutional layer retains high-level spatial information that is crucial for localizing salient regions in the input image.

In the following, we show the consistency of the attention weight distributions obtained from our model and AwareNet. Then, we present the results in qualitative and quantitative terms for each model using our XAI method both in the GradCAM case and with the attentional 3D maps.

Attention consistency analysis: Many studies focused on model interpretability by generating saliency maps for randomly chosen test subjects, enabling local qualitative analysis. However, this approach does not support the evaluation of interpretability consistency when the training and testing data vary. To examine this consistency in our approach, we considered the variability of attentional weight distributions across the five different folds used for cross-validation. Our evaluation leverages a clinical hypothesis specific to AD, which is believed to affect similar brain areas across different patients. For each fold, we calculated the weight distributions for each sample in the test set and derived an average distribution specific to that test set. By analyzing these average distributions across the five folds, we could assess whether our interpretability remained consistent, thereby supporting the generalizability of our findings across different subsets of data.

Fig. 3.5 presents our diagnosis network's average attentional weight distributions for each fold in the axial, coronal, and sagittal views. Upon examination of the histograms for each view, we observe a remarkable consistency in the distribution shapes across all five folds, indicating that our interpretability approach is stable despite the variation in the train/test set data. Specifically, the axial distributions reveal a consistent concentration of attentional weights around the initial slices. This trend suggests the model's recurrent focus on the brain's inferior regions, notably the areas where degenerative changes first manifest in AD, such as the *hippocampus*. In the coronal view, attentional weights are notably centered, indicating that the model consistently identifies the central part of the brain as more important. This central focus might correspond to the *medial temporal lobe*, including the *hippocampus* and the surrounding regions, further substantiating the axial findings. The sagittal view is the only bimodal distribution, suggesting that the model pinpointed symmetrical areas along this plane. We hypothesize that the network was focusing on the *hippocampus* since it adheres to all the constraints: situated in the inferior part of the brain, centrally located, and symmetrical. The consistency and specificity of these findings across multiple data folds strengthen the argument that our network could reliably identify specific brain regions as a critical biomarker for distinguishing between AD and CN subjects.

We also analyzed the consistencies of AwareNet distributions [167] to compare the robustness and interpretability of different attention mechanisms. However, the average distributions produced by this model, as seen in Fig. 3.6, present sparsely distributed peaks and do not allow the identification of predominant slice ranges. Furthermore, the distributions between the folds are inconsistent. These results indicate that on this dataset, AwareNet could not produce consistent attentional weights that are useful for contributing to the interpretability of its decisions.

3.5.4 Qualitative and quantitative results

This subsection examines the visual results and quantitative analysis concerning the brain areas emphasized by each model. It is important to note that the explainability analysis presented here was conducted solely for the AD vs. CN classification task. We decided not to generate explainability maps for the sMCI vs. pMCI classification task because the performance metrics for this classification problem were insufficient to ensure the reliability of the maps. Fig. 3.7, on the top, displays the attentional weight distributions across the three planes, averaged over all five folds of the cross-validation. This averaging provides a comprehensive view of the data distribution across all images in the dataset. Starting from the entire dataset distributions, the 3D attentional map was

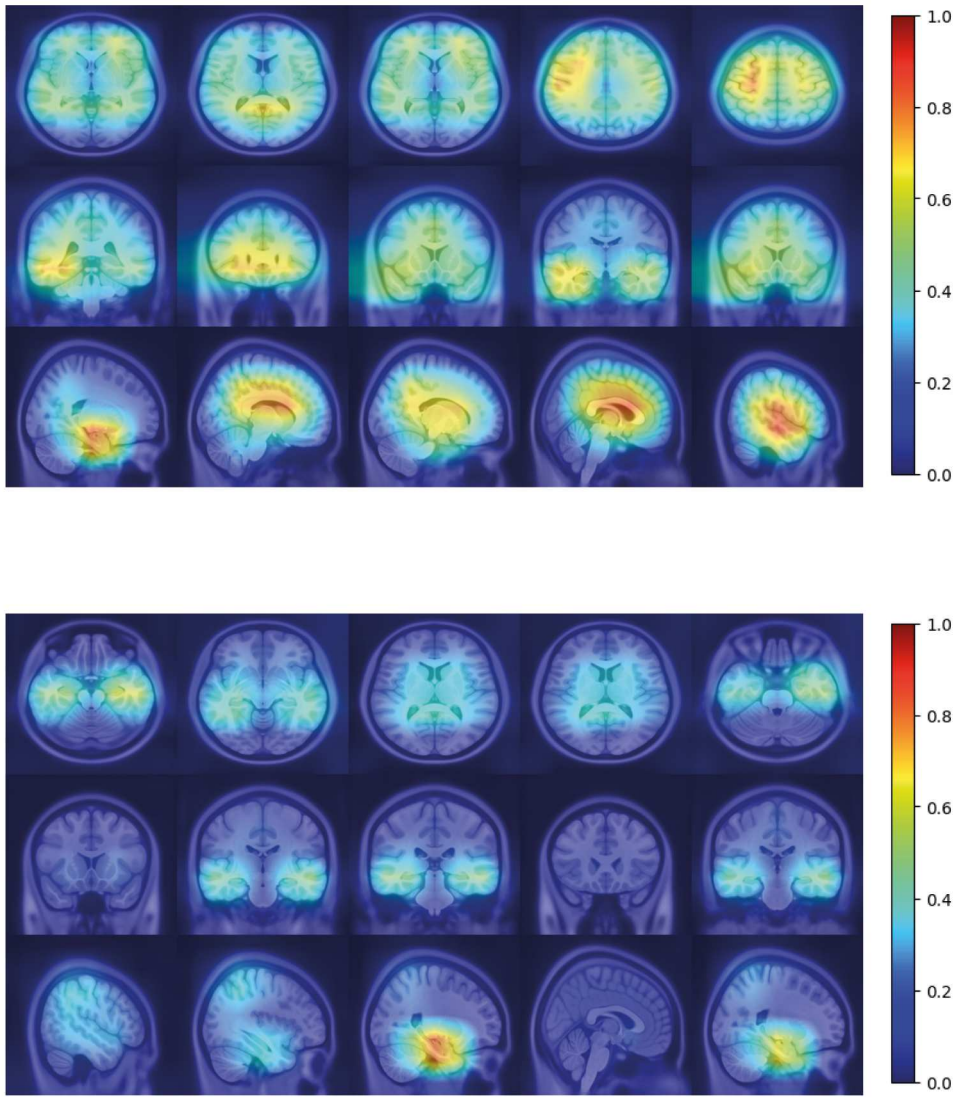


Figure 3.8: **GradCAM++** visualization of five slices for each plane selected randomly for **our model** (top) and **Attention Transformer** (bottom). The color bars show the scale of activation intensities.

created as detailed in Section 3.4.3. The averaged 3D map was enhanced by a factor of 10 and overlaid on the MNI152 template, which is representative of a typical patient's brain.

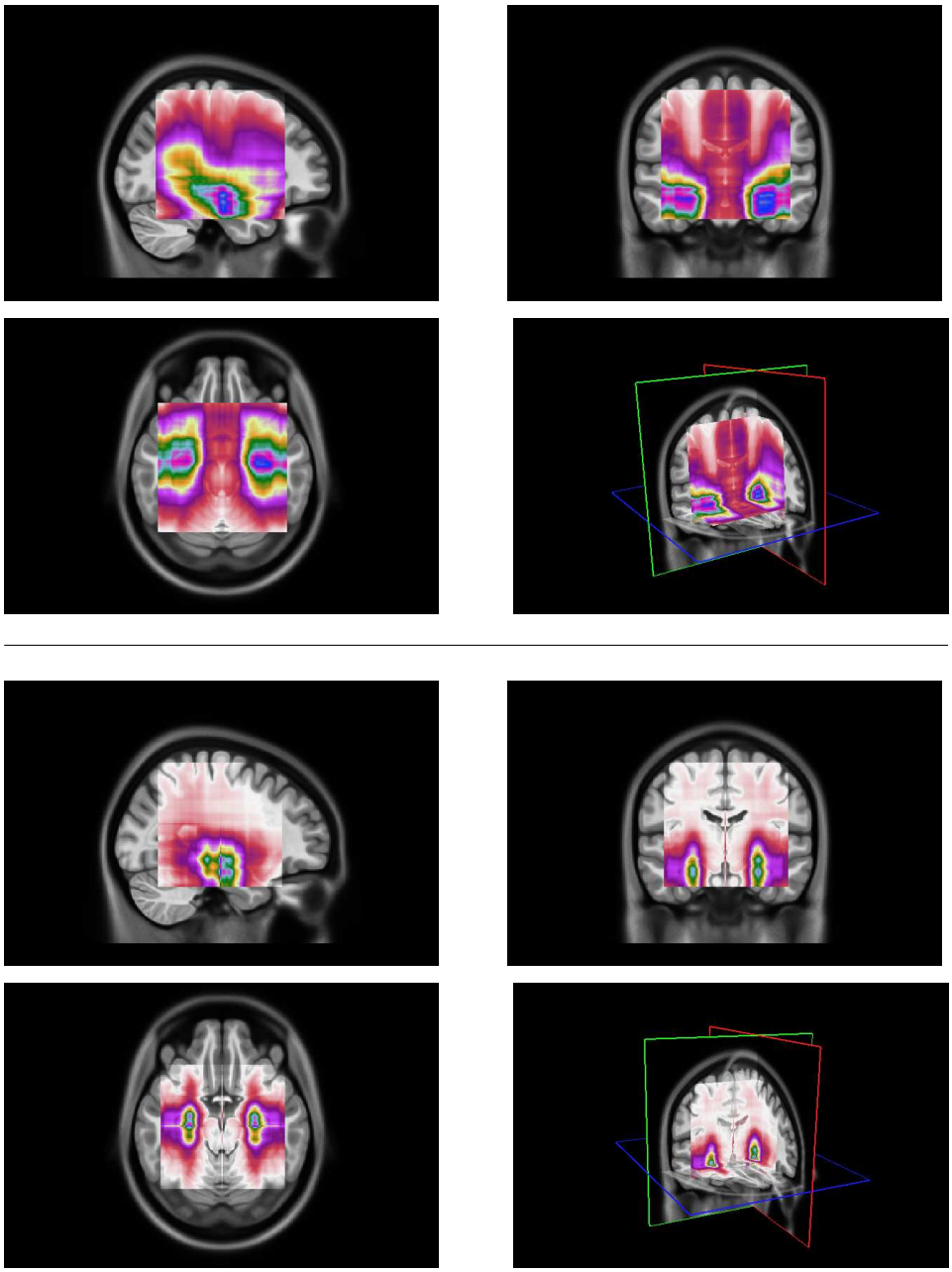


Figure 3.9: (**Top**) Visualization of mean 3D GradCAM++ map overlapped to MNI152 template with **our model**; (**Bottom**) Visualization of mean 3D GradCAM++ map overlapped to MNI152 template with **Attention Transformer**; Comparison visualizations captured in Freeview.

Combining this template with its corresponding atlas facilitates the identification of regions that, on average, received attention from the models. The bottom section of Fig. 3.7 shows the explainable MRI generated. The visual representation also indicates that the network targets the *medial temporal lobe region*, as suggested by the distributions. This result is confirmed by the quantitative analysis shown in Table 3.8, which reports the metrics for the 20 more extensive regions selected by our model and by AwareNet. As shown in Table 3.8(a), the three largest regions focused by our diagnosis model are the *hippocampus*, the *parahippocampus*, and the *amygdala*. In contrast, the 3D attentional map generated with AwareNet appears to focus on different regions. The first 3 regions highlighted are Cerebellum Gray Matter, Lateral Occipital, and Fusiform. The right part of the hippocampus appears only after them. From this result, it is also possible to note that with the same 99.9 percentile threshold for binarization, our model highlights a much more localized region. Specifically, our model selects 17 regions with a strong concentration in the top 14. On the contrary, AwareNet highlights 88 regions, 68 of which have been omitted in Table 3.8.

We also examined the interpretability of the Attention Transformer model proposed in [3]. For comparison, we created 2D saliency maps using the GradCAM++ algorithm [12] across all three views and all five test sets from the different folds. These maps were combined to create a unified average 3D saliency map as outlined in Section 3.5.3. This method was also applied to generate equivalent results from the saliency maps produced using our diagnostic model. As shown in Fig. 3.8, the 2D maps produced by our method are generally sparser compared to those from the Attention Transformer. The method introduced in [3] employs a cross-attention mechanism via a Multi-Head. Therefore, it is plausible that the Multi-Head allows it to generate 2D maps that align more meaningfully within the 3D context. This finding suggests that our approach may consider less contextual information from adjacent slices unless it is particularly relevant. In contrast, the cross-attention in the Attention Transformer might enable a more cohesive representation of the entire 3D space by considering both the local features within slices and their contextual interactions. This behavior is further clarified by creating 3D maps and overlaying them on the MNI152 template, similar to the attentional maps. As illustrated in Fig. 3.9 on the top, the 3D maps created using our model cover a broader and less concentrated area compared to those produced by the Attention Transformer, which are shown on the bottom. However, similar to the 3D attentional maps, both models predominantly focus on an area surrounding the hippocampus. As detailed in Table 3.9, both models identify key areas, such as the hippocampus and the amygdala. However, the emphasis on other regions varies markedly between the two. In the attention transformer model, there is a noticeable focus on the inferior lateral ventricles and the parahippocampal region, areas less emphasized by our model in this case. This result indicates that the Attention Transformer using cross-attention in combination with GradCAM can produce results

similar to those obtained by our method with a 3D attentional map. As seen in Tables 3.8(a) and 3.9(b), the first four areas on which our model focused with our approach are the same as those focused on by Attention Transformer with GradCAM. In contrast, our model with GradCAM shows broader involvement with regions such as the superior and middle temporal areas, which are not as prominent in the other cases.

3.6 Discussion

This section discusses the main findings of our study, including the prevalence of VGG architectures, the efficacy of double transfer learning, the effectiveness of our method against state-of-the-art ones, and the implications of our XAI in the context of medical imaging and diagnostics. Additionally, we thoroughly discuss the limitations and potential areas for further investigation in our study.

3.6.1 VGG for AD diagnosis

Different studies on AD have suggested using models from the VGG architecture family [65, 66, 105]. The selection of this family is supported by their effective performance in medical imaging, mainly when using transfer learning techniques [105]. While it is difficult to justify the clear superiority of this relatively simple architecture over more sophisticated ones such as EfficientNet and ResNet, its structure and pre-training stage may be the reason. The VGG models, for instance, VGG16, include a large number of parameters, approximately 138 million, with about 110 million dedicated to the classification layers and the remaining 28 million to convolutional layers. This structure suggests that most of the specific pattern recognition capabilities for class distinction learned from ImageNet likely reside in the classifier module. Therefore, the convolutional part may have acquired more general and transferable features than those in ResNet and EfficientNet, which rely on fully connected layers for classification.

3.6.2 Transfer domain knowledge for AD prediction

This study used a double transfer learning strategy to enhance model performance for the AD prediction task. The findings indicate that incorporating this additional domain-specific knowledge allows the model to more effectively differentiate subtle variations in brain morphology associated with early MCI progression. The utilization of such a strat-

egy addresses the challenges posed by limited training data, which is a common issue in medical imaging tasks due to privacy concerns and the difficulty of obtaining large annotated datasets. However, it is crucial to note that the model's performance, while promising, needs further improvement. The modest MCC values indicate that the model, although capable of identifying relevant patterns, may still struggle with generalization across diverse patient profiles or imaging conditions. This limitation underscores the need to further refine the model's architecture and training regimen, possibly by integrating richer datasets or applying more sophisticated image augmentation techniques to enhance its robustness and clinical applicability. We did not perform an XAI evaluation on this task due to the evident model's current explanatory limitations, underscored by its modest correlation metric. As the model's predictive accuracy and reliability improve, integrating advanced XAI techniques to provide deeper insights into the decision-making process will become essential. This advancement will not only increase transparency but could also contribute to the discovery of new structural brain markers that will help predict the disease well in advance.

3.6.3 Comparison with other attention-based methods

The evaluation of our proposed model against state-of-the-art attention-based techniques reveals several noteworthy insights, particularly in employing a more straightforward, lightweight architecture in medical imaging tasks constrained by limited data availability. Our model's performance suggests that a meticulously designed, less complex network can rival or surpass more intricate systems in diagnosis and explainability. The successful application of 2D CNNs with attention mechanisms in tasks traditionally reserved for 3D CNNs is particularly intriguing. It suggests that simpler 2D networks can effectively extract and utilize the spatial information necessary for accurate diagnosis with the proper architectural considerations and training strategies. This finding could impact the computational efficiency and accessibility of deep learning models for medical imaging, enabling their deployment in more diverse clinical environments with varying resource availability.

3.6.4 XAI Evaluation and Interpretability

Our XAI methods analysis has showcased both the potential and the challenges of interpreting DL models. Our findings show that the Attention Transformer model with GradCAM and our model, both with and without GradCAM, consistently highlighted regions such as the hippocampus, parahippocampal gyrus, amygdala and ventricles which

are well-documented in literature as being affected by AD [129, 162, 123, 46].

Our attention-based approach demonstrated results that align with clinical expectations, as evaluated by a radiologist on our team, with the algorithm ordering areas in a way that corresponds to diagnostic priorities and assigning comparable importance to the left and right portions of a brain region, which is consistent with clinical reasoning

A comprehensive assessment must incorporate clinical expertise to compare XAI methods and models. Neurologists and radiologists play a pivotal role in interpreting these XAI outputs, as their expertise in recognizing AD-specific biomarkers is crucial for validating the clinical relevance of the areas highlighted by the AI models. Even though we had the opportunity to validate the results with clinical experts, we know that involving clinical experts in the evaluation process presents its own challenges. It requires access to a panel of specialists willing to participate in such studies and a methodological framework that allows for averaging their insights to attain statistically significant conclusions. This process can be resource-intensive and difficult to implement across different studies, making it a less feasible option for consistent use. Metrics should be developed to simplify comparability across models and XAI methods in medical imaging. Such metrics could evaluate the relevance of highlighted regions within the context of a pathology. It should assess the localization and focus of maps and correlate these aspects with the known pathological features of the disease. By establishing these metrics, it could be easy to assess the comparison and validation of XAI approaches.

3.6.5 Limitations

In this section, we discuss several limitations identified in our study that warrant further exploration to enhance the robustness and applicability of our model. Firstly, the effectiveness of our method was assessed within a relatively limited dataset scenario, utilizing the ADNI1: Complete 1Yr 1.5T collection. While the model demonstrated promising results in this context, its performance against state-of-the-art models in larger datasets remains to be determined. Expanding the scope of testing to include a broader range of datasets with varying characteristics could help establish the scalability and general effectiveness of the proposed method.

Secondly, our XAI approach requires training three separate models, each tailored to one of the principal anatomical planes: axial, coronal, and sagittal. This requirement increases the computational demand and complexity of the training process since each model must be individually optimized and evaluated. Lastly, our approach lacks integration between the models corresponding to different anatomical planes. Currently, each

model is trained independently, without considering the inherent correlations between these planes. This segmentation of the training process potentially overlooks critical spatial information that could be utilized to enhance diagnostic accuracy and explainability. Future improvements could involve the development of integrated multi-view learning strategies, which could simultaneously process and cross-validate information across different planes, offering a more comprehensive understanding of the imaging data.

To evaluate the consistency of interpretability in our approach, we examined the variability of attentional weight distributions across the five different folds used for cross-validation. Our evaluation leverages a clinical hypothesis specific to AD, which is believed to affect similar brain areas across different patients [129]. However, we acknowledge that this assumption of spatial consistency may not hold for other diseases, e.g. cancer, where affected areas can vary significantly among patients. Therefore, this evaluation cannot be generalized a priori but is applicable in specific cases like AD, where the disease is supposed to impact similar regions in different patients.

3.7 Conclusions

In this chapter, we presented a method for AD diagnosis using pre-trained 2D CNNs to classify 3D volumes. Our approach integrates an attention mechanism to enhance the interpretability and accuracy of diagnosing AD and differentiating stable from progressive mild cognitive impairment. Our method outperformed baseline methods, achieving an MCC of 0.712 for distinguishing AD from CN subjects and 0.442 for sMCI from pMCI subjects. These results demonstrate the capability of our approach to classify these conditions effectively. A novel aspect of our work is the enhancement of model explainability. We successfully implemented voxel-level attention activation maps highlighting specific brain areas implicated in AD, such as the hippocampus, amygdala, parahippocampus, and inferior lateral ventricles. These regions are known to be crucial in AD pathology, making our model's outputs medically relevant. Our approach also includes a double transfer learning strategy that leverages pre-trained models to improve performance on limited datasets. This strategy utilizes knowledge transfer from the AD vs. CN task to enhance the model's sensitivity to subtle morphological changes associated with the progression of MCI, which are often hard to detect. In conclusion, our method advances diagnostic capabilities using 3D MRI scans for early AD detection and tries to address the critical need for explainability in medical imaging AI applications. By providing insights into the model's decision-making process, our approach helps bridge the gap between AI tools and clinical usability, making it a valuable asset for neurodegenerative disease research and potentially aiding in the clinical diagnosis and monitoring of AD progression.

Table 3.8: Importance measures of 20 largest brain regions computed with 3D Attention maps.

Brain Region	V_r	μ_r	σ_r	$A_{max,r}$	$A_{min,r}$	P_r
Hippocampus left	1562	0.136	0.139	0.762	0.028	0.333
Hippocampus right	1426	0.126	0.133	0.783	0.028	0.304
Parahippocampal left	688	0.129	0.137	0.884	0.028	0.254
Parahippocampal right	534	0.129	0.148	1.000	0.028	0.197
Amygdala left	480	0.097	0.092	0.620	0.028	0.291
Amygdala right	427	0.095	0.087	0.569	0.028	0.259
Inferior Lateral Ventricle right	232	0.113	0.129	0.677	0.028	0.219
Inferior Lateral Ventricle left	212	0.106	0.105	0.589	0.028	0.200
Cerebellum Gray Matter left	208	0.035	0.005	0.052	0.028	0.003
Lateral Orbitofrontal left	194	0.033	0.004	0.045	0.028	0.013
Fusiform right	184	0.045	0.015	0.107	0.028	0.014
Lateral Orbitofrontal right	140	0.034	0.004	0.046	0.028	0.009
Cerebellum Gray Matter right	119	0.034	0.005	0.054	0.028	0.002
Fusiform left	88	0.040	0.010	0.070	0.028	0.007
Entorhinal left	16	0.034	0.003	0.041	0.030	0.005
Ventral Diencephalon left	6	0.029	0.001	0.031	0.028	0.001
Entorhinal right	2	0.033	0.003	0.036	0.031	0.001
-	-	-	-	-	-	-
-	-	-	-	-	-	-
-	-	-	-	-	-	-

(a) Our model

Brain Region	V_r	μ_r	σ_r	$A_{max,r}$	$A_{min,r}$	P_r
Cerebellum Gray Matter - right	192	0.001	0.003	0.022	0.000	0.003
Lateral Occipital - right	173	0.003	0.008	0.067	0.000	0.008
Cerebellum Gray Matter - left	132	0.001	0.004	0.035	0.000	0.002
Lateral Occipital - left	116	0.002	0.011	0.106	0.000	0.005
Fusiform - left	93	0.001	0.002	0.009	0.000	0.007
Fusiform - right	74	0.001	0.001	0.007	0.000	0.005
Hippocampus - right	74	0.004	0.013	0.076	0.000	0.016
Lateral Orbitofrontal - right	70	0.000	0.001	0.003	0.000	0.005
Entorhinal - right	62	0.009	0.025	0.117	0.000	0.019
Ventral Diencephalon - right	62	0.001	0.003	0.014	0.000	0.010
Hippocampus - left	55	0.003	0.012	0.088	0.000	0.012
Brainstem - right	53	0.002	0.010	0.076	0.000	0.003
Lingual - right	48	0.000	0.001	0.003	0.000	0.004
Lateral Orbitofrontal - left	47	0.000	0.001	0.004	0.000	0.003
Parahippocampal - right	47	0.000	0.000	0.002	0.000	0.017
Cerebellum White Matter - right	36	0.000	0.000	0.002	0.000	0.003
Ventral Diencephalon - left	35	0.002	0.005	0.022	0.000	0.006
Entorhinal - left	35	0.011	0.042	0.186	0.000	0.011
Amygdala - right	35	0.000	0.000	0.002	0.000	0.021
Superior Frontal - right	27	0.000	0.000	0.000	0.000	0.001

(b) AwareNet

Table 3.9: Importance measures of 20 largest brain regions computed with 3D GradCAM maps.

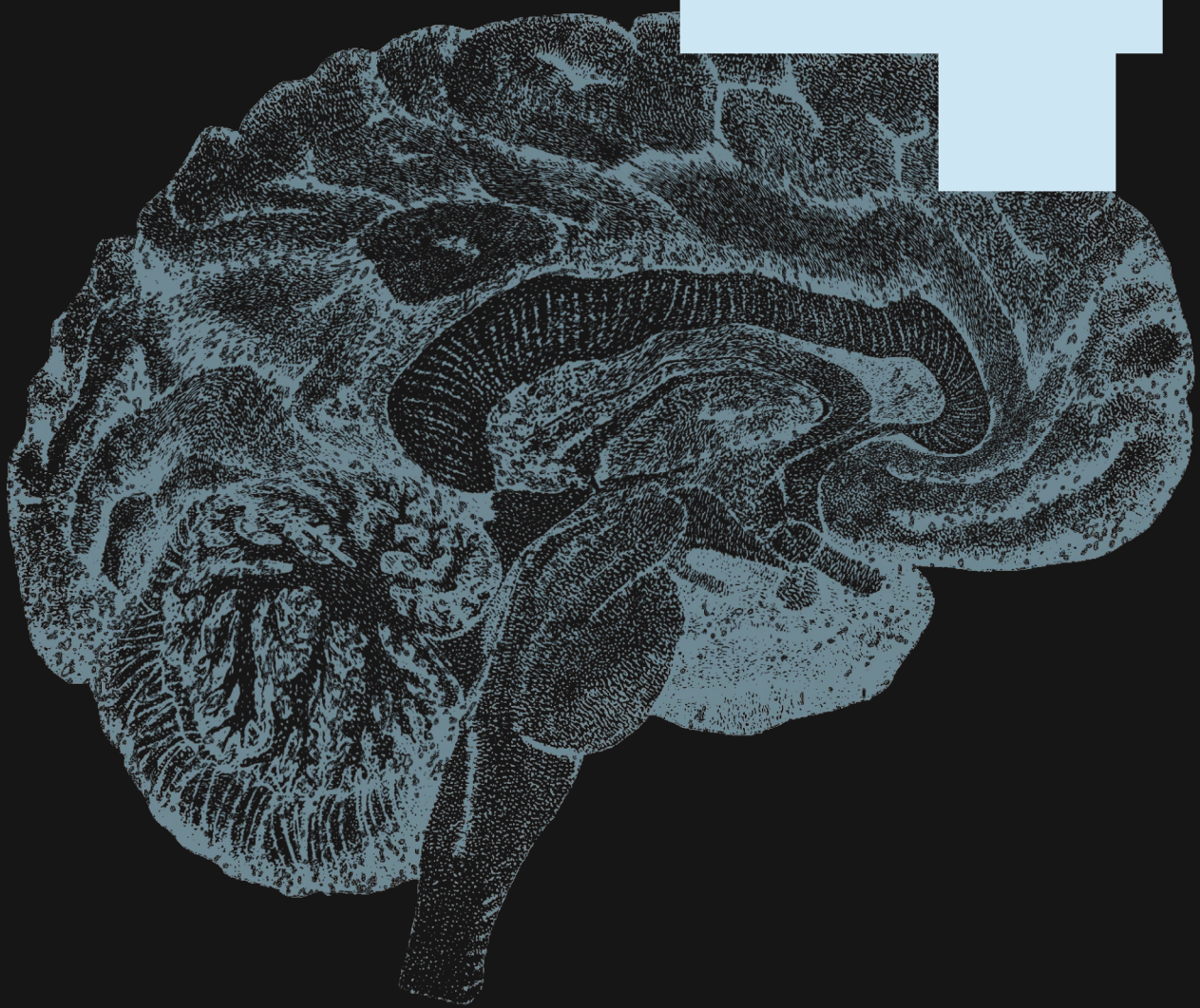
Brain Region	V_r	μ_r	σ_r	$M_{max,r}$	$M_{min,r}$	P_r
Hippocampus left	1722	0.874	0.057	0.992	0.773	0.367
Hippocampus right	965	0.835	0.041	0.963	0.773	0.206
Inferior Lateral Ventricle left	648	0.861	0.054	1.000	0.773	0.612
Inferior Lateral Ventricle right	435	0.830	0.038	0.944	0.773	0.411
Superior Temporal left	164	0.810	0.022	0.860	0.774	0.006
Amygdala right	146	0.795	0.015	0.844	0.774	0.088
Superior Temporal right	143	0.816	0.029	0.888	0.774	0.006
Amygdala left	118	0.795	0.021	0.867	0.773	0.072
Middle Temporal left	96	0.825	0.028	0.873	0.774	0.003
Middle Temporal right	95	0.823	0.028	0.887	0.774	0.003
Parahippocampal left	91	0.829	0.038	0.928	0.774	0.034
Entorhinal left	77	0.825	0.032	0.891	0.774	0.024
Fusiform left	65	0.814	0.033	0.903	0.774	0.005
Insula left	8	0.787	0.009	0.806	0.775	0.001
Insula right	7	0.787	0.014	0.809	0.776	0.001
Parahippocampal right	2	0.773	0.000	0.774	0.773	0.001
Ventral Diencephalon left	2	0.779	0.005	0.784	0.774	0.000
Inferior temporal left	1	0.775	0.000	0.775	0.775	0.000
Putamen left	1	0.776	0.000	0.776	0.776	0.000
-	-	-	-	-	-	-

(a) Our model

Brain Region	V_r	μ_r	σ_r	$M_{max,r}$	$M_{min,r}$	P_r
Hippocampus - left	1921	0.614	0.089	0.921	0.482	0.409
Hippocampus - right	1728	0.628	0.103	1.000	0.482	0.368
Amygdala - right	409	0.576	0.068	0.779	0.482	0.248
Inferior Lateral Ventricle - right	322	0.625	0.090	0.847	0.482	0.304
Inferior Lateral Ventricle - left	318	0.649	0.088	0.899	0.483	0.300
Amygdala - left	312	0.577	0.065	0.731	0.482	0.189
Parahippocampal - right	176	0.524	0.029	0.599	0.484	0.065
Parahippocampal - left	156	0.527	0.032	0.598	0.482	0.058
Entorhinal - left	132	0.545	0.041	0.651	0.482	0.041
Ventral Diencephalon - left	124	0.574	0.066	0.783	0.483	0.020
Entorhinal - right	106	0.541	0.045	0.663	0.482	0.033
Putamen - left	46	0.549	0.040	0.628	0.484	0.007
Ventral Diencephalon - right	38	0.537	0.050	0.687	0.483	0.006
Pallidum - left	23	0.519	0.025	0.583	0.485	0.014
Fusiform - left	5	0.502	0.013	0.522	0.487	0.000
Fusiform - right	2	0.484	0.000	0.484	0.483	0.000
Putamen - right	2	0.495	0.001	0.496	0.495	0.000
-	-	-	-	-	-	-
-	-	-	-	-	-	-
-	-	-	-	-	-	-

(b) Attention Transformer

4



DIFFUSION-BASED FOUNDATION MODELS FOR SEMANTIC LEARNING IN 3D BRAIN IMAGING

REFERENCE PAPER

Latent Diffusion Autoencoders: Toward Efficient and Meaningful Unsupervised Representation Learning in Medical Imaging - A Case Study on Alzheimer's Disease

AUTHORS

Gabriele Lozupone[1,2]
Alessandro Bria[1]
Francesco Fontanella[1]
Frederick J.A. Meijer[3]
Claudio De Stefano[1]
Henkjan Huisman[2]

[1]Department of Electrical and Information Engineering (DIEI), University of Cassino and Southern Lazio
[2]Diagnostic Image Analysis Group, Radboud University Medical Center
[3]Department of Medical Imaging, Radboud University Medical Center

Lozupone, G., Bria, A., Fontanella, F., Meijer, F. J. A., De Stefano, C., Huisman, H., et al. (2026). Latent Diffusion Autoencoders: Toward Efficient and Meaningful Unsupervised Representation Learning in Medical Imaging--A Case Study on Alzheimer's Disease. Medical Image Analysis, 103932.

Chapter 4

Diffusion-based Foundation Models for Semantic Learning in 3D Brain Imaging

4.1 Introduction

As introduced in Section 1.3.6, the progression toward foundation models represents a fundamental shift in medical imaging analysis: moving from task-specific predictors that require labeled data to general-purpose representation learners that extract knowledge from vast quantities of unlabeled scans. While Chapters 2 and 3 demonstrated the effectiveness of supervised deep learning for AD diagnosis using handwriting biomarkers and interpretable neuroimaging frameworks respectively, both approaches fundamentally depend on diagnostic labels and careful data splitting to prevent overfitting. This chapter operationalizes the foundation model paradigm by introducing Latent Diffusion Autoencoder (LDAE), a framework that learns semantically meaningful representations from unlabeled 3D brain MRI scans through unsupervised diffusion-based learning, subsequently transferring to multiple downstream tasks without task-specific fine-tuning.

The conceptual motivation for unsupervised representation learning in neuroimaging was established in Section 1.3.6: clinical repositories contain millions of brain MRI scans acquired for diverse purposes beyond AD diagnosis, capturing fundamental aspects of brain anatomy, aging, and anatomical variability without requiring disease-specific an-

notations. While supervised approaches like those in Chapter 3 addressed methodological rigor through careful subject-level data splitting and cross-validation strategies (Section 1.3.5), they remain constrained by the limited availability of labeled AD datasets. Unsupervised methods circumvent this limitation by learning representations from the underlying structure of brain imaging data itself, subsequently adapting to specific tasks with minimal labeled examples.

Recently, Diffusion Probabilistic Model (DPM)s (DPMs) have achieved remarkable performance in image synthesis and data distribution modeling [31, 111], with stable training processes that surpass deep generative models like GANs [55] and VAEs [82, 132]. As discussed in Section 1.3.6, while generative models have primarily been applied to AD research for data augmentation and cross-modality synthesis [1], their representation learning capabilities remain underexplored. The seminal work on Diffusion Autoencoder (DAE)s [124] demonstrated that diffusion models can learn compact, semantically meaningful representations enabling controllable image manipulation and attribute-specific modifications, surpassing GANs in feature disentanglement while preserving image quality and training stability. Traditional DPMs can encode an input image x_0 into a spatial latent variable x_T through reverse diffusion, but this representation lacks semantic properties such as disentanglement, compactness, and the ability to perform meaningful linear interpolation. In contrast, DAEs ensure representations are both compact and disentangled while facilitating semantic interpolation, making them strong candidates for structured semantic learning. Subsequent architectures explored different representation learning strategies, including DAEs from pretrained DPMs (Pretrained Diffusion Autoencoder (PDAE)) [186] and the SODA architecture [68] that introduced layer modulation to improve semantic attribute disentanglement.

However, applying diffusion-based representation learning to high-dimensional 3D MRI data remains computationally prohibitive. Current frameworks operating in native voxel space are infeasible for 3D data due to memory and processing requirements of running iterative diffusion processes on full volumes. In 3D brain MRI, latent diffusion approaches have primarily targeted unconditional and conditional generation [117, 121] or disease progression prediction using pre-computed anatomical features [182, 125]. Other recent works have extended diffusion models to medical imaging tasks with fundamentally different objectives than unsupervised representation learning. For example, Diff-UNet [178] integrates diffusion into a UNet backbone for supervised segmentation, framing it as discrete denoising to generate label maps—improving voxel-wise prediction rather than learning general-purpose latent spaces. Other approaches focus on conditional generation between imaging modalities [177], using cross-conditioned UNets to synthesize new images rather than extracting transferable representations. Large-scale models like 3D MedDiffusion [168] achieve high-quality controllable generation, while others pro-

pose task-specific frameworks such as Medical World Models for treatment-conditioned tumor evolution [180] or tumor synthesis across organs for segmentation improvement [13]. Although these methods showcase diffusion's versatility for generation and augmentation, they remain goal-driven generative frameworks rather than representation-learning systems. Overall, despite rapid expansion of diffusion models in medical imaging, a critical gap exists for frameworks making the semantic-rich learning of diffusion autoencoders tractable for 3D medical imaging.

Contributions This chapter introduces LDAE as an efficient framework for unsupervised and meaningful representation learning, positioning it as the culmination of the thesis progression established in Chapter 1. While Chapter 2 demonstrated multi-task learning across heterogeneous handwriting modalities and Chapter 3 introduced interpretable supervised learning with rigorous cross-validation methodologies, this chapter presents a foundation model approach enabling general-purpose representation learning without task-specific supervision. The proposed approach builds upon principles established in DAE [124], PDAE [186], and LDMs (LDMs) [133]. Unlike conventional diffusion-based autoencoders operating in original image space, LDAE applies the diffusion process in a compressed latent space, addressing the computational scalability challenge that Section 1.3.6 identified as a fundamental barrier to foundation models in medical imaging. Our contributions are:

- We introduce a novel and computationally efficient framework, LDAE, that for the first time enables the application of diffusion autoencoders to high-dimensional volumetric data by operating in a compressed latent space. This methodological advancement bridges the gap between semantic-rich diffusion models and scalable representation learning.
- We validate the hypothesis that LDAE enables unsupervised learning of clinically meaningful representations in the context of AD and ageing, as conceptualized in Section 1.3.6.
- We show that LDAE outperforms voxel-space DAEs in speed ($20\times$ faster inference) and reconstruction quality.
- We validate the versatility of LDAE representations across multiple tasks, including disease diagnosis, age prediction, counterfactual generation, and interpolation of intermediate scans in longitudinal series.

Our methodological innovation is a three-stage training strategy that successfully adapts DAE principles to latent space operation. This approach consists of: (i) training a feature-

aware Autoencoder (AE) with perceptual losses to compress high-dimensional MRI scans into a lower-dimensional latent space; (ii) pretraining a diffusion model on the compressed latent representations; and (iii) LDAE unsupervised representation learning with an encoder-decoder to fill the posterior mean gap, following the strategy introduced in PDAE [186]. This work explores diffusion autoencoders in medical imaging and their adaptation to operate in the latent space of LDMs, enabling computationally efficient unsupervised representation learning from large-scale 3D MRI data.

To validate the proposed LDAE framework, we employed longitudinal 3D brain MRI data from the ADNI database [119], building upon the standardized T1-weighted MRI protocol and preprocessing pipelines discussed in Section 1.2. As established in that section, ADNI provides a large and diverse collection of high-quality 3D scans with consistent acquisition protocols, making it an ideal platform for unsupervised training and subsequent evaluation. Following the thesis structure outlined in Section 1.3.6, we focus on AD not to address a specific clinical problem, but to demonstrate that the proposed framework can learn meaningful and transferable representations from large-scale unlabeled 3D medical data in an efficient manner. The extensively documented neuroanatomical patterns of AD described in Section 1.2.3—including hippocampal and medial temporal lobe atrophy, temporo-parietal cortical thinning, and ventricular enlargement—support both quantitative benchmarking on established downstream tasks (e.g., AD diagnosis, age prediction) and qualitative assessment of generative capabilities, including synthesis of missing scans in longitudinal studies and counterfactual explanations (i.e., visualizing plausible anatomical changes that would alter a model's prediction). To assess zero-shot transferability, a key capability distinguishing foundation models from task-specific approaches, we further evaluated LDAE on the AIBL [40] and OASIS-3 [89] datasets, addressing the cross-dataset generalization challenges discussed in Section 1.3.5. We used the model pretrained on ADNI to extract semantic representations from these unseen datasets and trained a simple linear classifier without model fine-tuning, comparing performance against fully supervised alternatives.

Our evaluation aims at validating two key hypotheses: (i) LDAE learns semantically rich representations that capture clinically relevant attributes related to AD and ageing, thus enabling unsupervised representation learning in latent space; and (ii) LDAE provides high-quality generation and reconstruction while significantly improving computational efficiency over conventional voxel-space DAEs. For hypothesis (ii), we compare LDAE to voxel-space DAEs and demonstrate significant efficiency improvements, achieving $20\times$ speedup in inference time while surpassing reconstruction quality. For hypothesis (i), we perform linear-probe evaluations on learned semantic embeddings. LDAE achieves strong performance in downstream tasks such as AD diagnosis (AUROC: 89.48%, ACC: 83.65%) and age prediction (MAE: 4.16 years, RMSE: 5.23 years), indicating that learned

latent codes encode clinically meaningful information. We further validate semantic interpretability through latent attribute manipulation, enabling alteration of brain MRI anatomical structures related to disease and age progression consistent with the neuroanatomical signatures described in Section 1.2.3. Semantic and stochastic interpolation experiments show LDAE's capacity to predict missing intermediate scans in longitudinal series, achieving robust performance even at longer temporal gaps.

While Section 4.1 established the conceptual foundation for foundation models in neuroimaging, the following sections provide the mathematical formulations and architectural details necessary for understanding LDAE's implementation. The remainder of this chapter is organized as follows. Section 4.2 provides technical background on DPMs, diffusion autoencoders, and the mathematical principles underlying our framework. Section 4.3 describes the multi-stage training approach and architectural choices. Section 4.4 presents experimental settings and evaluation protocols. Section 4.5 reports results on generation, manipulation, interpolation, and downstream tasks. Section 4.6 discusses findings and concludes the chapter.

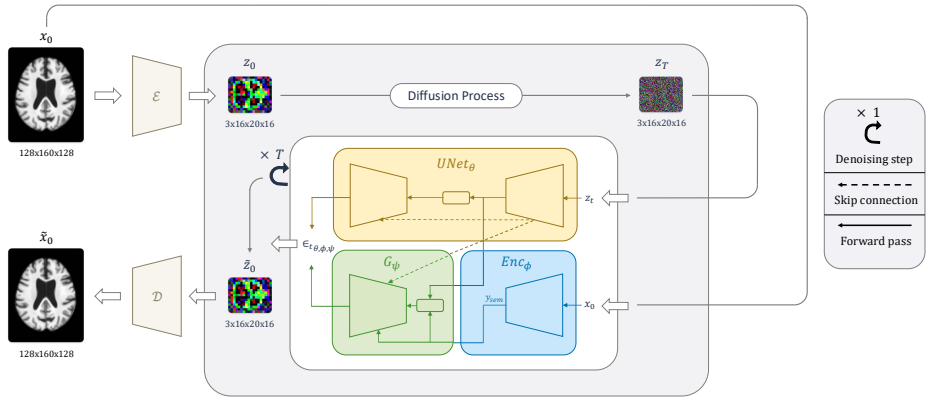


Figure 4.1: Overview of the proposed 3D LDAE framework for unsupervised representation learning in brain medical imaging. The framework consists of three key components: (i) a **compression model** ($\mathcal{E}$ and $\mathcal{D}$), which encodes high-dimensional MRI brain scans ($\mathbf{x}_0$) into a lower-dimensional latent representation ($\mathbf{z}_0$), facilitating efficient processing; (ii) a **Latent Diffusion Model (LDM)**, trained to learn the distribution of the compressed representations through a diffusion process, progressively transforming $\mathbf{z}_0$ into $\mathbf{z}_T$ and vice versa via a UNet-based denoising network (UNet$_\theta$); and (iii) the **semantic encoder-decoder model** (G_ψ and Enc_ϕ), which learns a meaningful latent representation ($\mathbf{y}_{sem}$) from the input scan and utilizes it to guide the reverse diffusion process via a gradient estimator (G_ψ).

4.2 Background

DPMs are generative models that model a target distribution by learning a denoising process at varying noise levels [148]. This concept is inspired by nonequilibrium thermodynamics, in which a physical system starts from a structured, low-entropy state that is gradually "diffused" or driven toward a more disordered, high-entropy equilibrium state over time. In principle, the system can be steered back toward a more ordered configuration, although this typically requires precise control and information about the underlying dynamics. In diffusion-based generative models, we begin with real data and then apply a stochastic "diffusion" of noise step-by-step. Each step slightly corrupts the data by adding Gaussian noise to arrive at a highly noisy, nearly featureless distribution that is mathematically close to a pure Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

4.2.1 Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (DDPMs), introduced in [64], model a forward diffusion process that gradually adds Gaussian noise to a data sample over a sequence of T steps. Starting from a clean data point $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, the process constructs a Markov chain $\mathbf{x}_0 \to \mathbf{x}_1 \to \cdots \to \mathbf{x}_T$, where each transition is defined as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}\right), \tag{4.1}$$

with a fixed variance schedule $\{\beta_t\}_{t=1}^T$, where $0 < \beta_t < 1$. It is convenient to define $\alpha_t = 1 - \beta_t$, and its cumulative product $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$. This formulation allows expressing the noisy sample $\mathbf{x}_t$ directly in terms of $\mathbf{x}_0$ without requiring intermediate steps. The marginal distribution becomes:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{4.2}$$

which shows that $\mathbf{x}_t$ is a weighted combination of the original signal and Gaussian noise. As T grows large, $\bar{\alpha}_T$ tends toward zero, so that $q(\mathbf{x}_T|\mathbf{x}_0) \approx \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Strictly speaking, exact convergence to $\mathcal{N}(\mathbf{0}, \mathbf{I})$ occurs only in the limit $T \to \infty$, but for sufficiently large T the distribution is practically indistinguishable from an isotropic standard Gaussian. A useful consequence of this formulation is the ability to reparameterize the sampling of $\mathbf{x}_t$ using the so-called *reparameterization trick* [64], expressing $\mathbf{x}_t$ as:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{4.3}$$

To enable data distribution learning and sample generation, the goal is to learn the reverse diffusion process $p(\mathbf{x}_{t-1}|\mathbf{x}_t)$, which reconstructs clean data from noise-corrupted inputs. As shown in [148], using a sufficiently large number of steps (typically $T = 1000$) allows the reverse process to be well-approximated by a Gaussian:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I}), \tag{4.4}$$

where the mean μ_θ is parameterized by a neural network (typically a UNet [64]) and σ_t is a known or learned variance. In practice, instead of directly predicting μ_θ, the model is trained to predict the noise $\boldsymbol{\epsilon}$ used in the forward process. This choice leads to a simplified and stable training objective, as the predicted mean can be reconstructed from the predicted noise using the known forward process parameters. Specifically, the model is trained to minimize the following mean squared error loss:

$$L_\epsilon = \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon},t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2\right], \tag{4.5}$$

where $\mathbf{x}_t$ is computed via Eq. 4.3. This objective corresponds to a simplified form of the variational lower bound on the data log-likelihood and has been widely adopted in modern DDPM frameworks [32, 112, 124, 149].

4.2.2 Denoising Diffusion Implicit Models

Denoising Diffusion Implicit Model (DDIM) proposed in [149] introduces a non-Markovian reverse process that, unlike standard DDPMs, enables faster inference with fewer steps. In particular, DDIM introduced the following novel joint inference distribution

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \cdot \frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0}{\sqrt{1 - \bar{\alpha}_t}}, \sigma_t^2\mathbf{I}\right), \tag{4.6}$$

and the following novel generative process:

$$\begin{aligned}\mathbf{x}_{t-1} =& \sqrt{\bar{\alpha}_{t-1}}\left(\frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_{\theta_t}(\mathbf{x}_t, t)}{\sqrt{\bar{\alpha}_t}}\right) + \\ &+ \left(\sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2}\right)\epsilon_{\theta_t}(\mathbf{x}_t, t) + \sigma_t\epsilon_t,\end{aligned} \tag{4.7}$$

while maintaining the DDPM marginal distribution (Eq. 4.2). Since the marginal is preserved, Eq. 4.7 can be used to obtain the latent variable x_{t-1} from x_t through ϵ_θ from

a pretrained DDPM, where $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$. Consequently, σ_t determines the degree of stochasticity in the generation process.

A key property of DDIMs is that the generative process will be fully deterministic by choosing $\sigma_t = 0$ in Eq. 4.7. By doing so, we obtain a bijective, noise-free mapping that can be reversed by shifting the time index $t - 1 \rightarrow t + 1$. This allows DDIM to be used not only for generation (decoding from $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$) but also for *encoding* a data sample $\mathbf{x}_0$ into a corresponding latent representation $\mathbf{x}_T$ via repeated application of the same update rule in reverse. The inversion is possible because the network is trained to predict the noise component $\epsilon_\theta(\mathbf{x}_t, t)$ *at a specific time step* t. Consequently, the predicted noise remains valid regardless of whether the process runs forward or backward in time. By defining:

$$\mathbf{f}_\theta(\mathbf{x}_t, t) = \frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_{\theta_t}(\mathbf{x}_t, t)}{\sqrt{\bar{\alpha}_t}},$$

and inverting $\{Eq.\ 4.7\}_{\sigma_t=0}$ we can obtain the encoding step:

$$\mathbf{x}_{t+1} = \sqrt{\bar{\alpha}_{t+1}}\mathbf{f}_\theta(\mathbf{x}_t, t) + \sqrt{1 - \bar{\alpha}_{t+1}}\,\epsilon_\theta(\mathbf{x}_t, t), \tag{4.8}$$

In [124], it is shown that applying this encoding procedure on $\mathbf{x}_0$ to obtain the associated latent $\mathbf{x}_T$ allows DDIM to reconstruct $\mathbf{x}_0$ with high fidelity by generating from $\mathbf{x}_T$. However, since the resulting latent primarily captures low-level noise structure, it lacks high-level semantics, making tasks like semantically-smooth interpolation or attribute editing less effective in this latent space.

4.2.3 Diffusion Autoencoders

In order to obtain a semantically meaningful latent code, the authors of DAE [124] designed a conditional DDIM image decoder that approximates $p(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y}_{sem})$ in which $\mathbf{y}_{sem}$ is a non-spatial vector of dimension $d = 512$ learned from a **semantic encoder** that maps $\mathbf{x}_0$ into $\mathbf{y}_{sem} = Enc_\phi(\mathbf{x}_0)$. The encoder and the DDIM decoder are trained in conjunction by optimizing:

$$L_\epsilon = \sum_{t=1}^{T} \mathbb{E}_{\mathbf{x}_0, \epsilon_t} \left[\|\epsilon_\theta(\mathbf{x}_t, t, \mathbf{y}_{\text{sem}}) - \epsilon_t\|_2^2 \right]. \tag{4.9}$$

with respect to θ and ϕ by conditioning a UNet with the Enc_ϕ output using Adaptive Group-wise Normalization (AdaGN) layers as proposed in [31]. By training the two models simultaneously, the encoder Enc_ϕ is forced to learn as much information as possible to help the DDIM in the denoising process. The authors of PDAE [186] clarified this

behaviour by showing that there exists a gap between the posterior mean predicted by an unconditional DDPM ($\mu_\theta(\mathbf{x}_t, t)$) and the true one ($\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0)$). The posterior mean gap is caused by an information loss that, in theory, can be recovered by conditioning on some $\mathbf{y}$ that contains all information about $\mathbf{x}_0$. By letting $\mathbf{y} = \mathbf{y}_{sem}$ a learnable vector produced by an encoder, it will be forced to learn as much information as possible to fill the posterior mean gap and consequently as much information as possible from $\mathbf{x}_0$. Following these principles, it is possible to train a DAE from pretrained DDPMs, achieving better training efficiency and stability [186].

4.2.4 Classifier-Guided Sampling Method

The classifier-guided method allows to condition the generation of a DDPM towards some information, e.g. classes or prompts [148, 151]. It consists of training a classifier $p_\psi(\mathbf{y}|\mathbf{x}_t)$ on noisy images $\mathbf{x}_t$. The gradient $\nabla_{\mathbf{x}_t} \log p_\psi(\mathbf{y}|\mathbf{x}_t)$ points in the direction of increasing probability of $\mathbf{y}$ and can be leveraged to guide the generation towards samples correlated to information in $\mathbf{y}$. The conditional reverse process can be approximated using a Gaussian distribution, resembling the unconditional case, but with an adjusted mean:

$$p_{\theta,\psi}(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y}) \approx \mathcal{N}\Big(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t) + \sigma_t \cdot \nabla_{\mathbf{x}_t} \log p_\psi(\mathbf{y}|\mathbf{x}_t), \sigma_t\Big). \tag{4.10}$$

For DDIMs, a score-based conditioning trick [149, 150] can be applied to define a new function approximator for conditional sampling:

$$\hat{\epsilon}_{\theta,\psi}(\mathbf{x}_t, t) = \epsilon_\theta(\mathbf{x}_t, t) - \sqrt{1 - \bar{\alpha}_t} \cdot \nabla_{\mathbf{x}_t} \log p_\psi(\mathbf{y}|\mathbf{x}_t). \tag{4.11}$$

Based on this concept, the authors of [186] introduced an additional gradient estimator $G_\psi(\mathbf{x}_t, \mathbf{y}_{sem}, t)$ to approximate the classifier guidance term $\nabla_{\mathbf{x}_t} \log p(\mathbf{y}_{sem}|\mathbf{x}_t)$. This allows training a DAE from a pre-trained unconditional DDPM. The encoder processes the original image to produce a semantic representation $\mathbf{y}_{sem}$, which conditions the gradient estimator to approximate $\nabla_{\mathbf{x}_t} \log p(\mathbf{y}_{sem}|\mathbf{x}_t)$ and a frozen pre-trained DDPM is used to approximate $\epsilon_\theta(\mathbf{x}_t, t)$. This novel formulation enables a two-stage training strategy: (i) train an unconditional DDPM to learn the prior, and (ii) freeze the DDPM and train the encoder and gradient estimator to introduce semantic control. This decoupling leads to more stable and efficient training, as it eliminates the need for joint optimization, separating the concepts of generation and representation learning. More details will be discussed in Section 4.3.3 since we used the same concepts but on a compressed data representation as discussed in Section 4.3.1 and Section 4.3.2.

4.3 Methods

The overall scheme of the proposed framework is shown in Fig.4.1. This section provides a detailed description of the training stages and the network architecture choices. Since the framework is designed to learn two different types of latent representations from the original data, we will refer throughout the remainder of the manuscript to the latent space generated by the compression model as the *compressed space* (Section 4.3.1) and to the latent space produced by the semantic encoder as the *semantic space* (Section 4.3.3)

4.3.1 Compression Model

To make the training tractable for 3D data, we adopted the principles of LDMs, specifically the use of an AE to compress high-dimensional scans into a lower-dimensional latent space for efficient generation as in [125, 121]. Rather than training a compression model from scratch, we fine-tuned a model with weights pre-trained on the extensive UK Biobank dataset [153], by [121]. Training such a compression model from scratch would not only be time-consuming but would also require more data than the one available in ADNI. For this reason, we followed the same strategy as [125], who also adopted the pre-trained model from [121] and fine-tuned it for a Latent Diffusion model on the ADNI dataset. Both approaches followed the seminal paper on Stable Diffusion, [133], in which the authors found that the model requires a combination of multiple loss functions to achieve good reconstruction quality and learn a high-fidelity latent space. Although these losses do not act directly on the subsequent representation-learning stage, they shape the compact latent space in which the diffusion model subsequently operates. By preserving voxel-wise spatial correspondence with the input volume, the compressed representation allows the diffusion model to focus on clinically meaningful structure instead of low-level detail. Lacking this well-formed latent space, the diffusion process would be forced to run in the native voxel domain, an operation that is highly computationally expensive for 3D brain MRI.

Following [133], we employed an autoencoder architecture based on [41] as our perceptual compression model. This approach uses a combination of perceptual loss [185] and patch-based adversarial objective [34, 184], ensuring reconstructions remain within the image manifold while avoiding the blurriness characteristic of pixel-space losses. We applied Kullback-Leibler (KL) regularization to constrain the variance in the compressed space, similar to VAE [82, 132].

Given a 3D brain scan $\mathbf{x}_0 \in \mathbb{R}^{H \times W \times D \times 1}$, a compression encoder $\mathcal{E}$ encodes $\mathbf{x}_0$ into a

compressed representation $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0) \in \mathbb{R}^{h \times w \times d \times c}$, where $f = H/h = W/w = D/d$ is the downsampling factor. The decoder $\mathcal{D}$ reconstructs the image:

$$\tilde{\mathbf{x}}_0 = \mathcal{D}(\mathbf{z}_0) = \mathcal{D}(\mathcal{E}(\mathbf{x}_0)).$$

We train the compression model using a balanced combination of losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recons}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{perceptual}}\mathcal{L}_{\text{perceptual}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}, \tag{4.12}$$

where λ_{KL}, $\lambda_{\text{perceptual}}$, and λ_{adv} determine the relative importance of each term. Table 4.1 reports the values used in our implementation.

Table 4.1: AutoencoderKL architecture and training configuration

Parameter	**Value**
Input Resolution	$1 \times 128 \times 160 \times 128$
Latent Size	$3 \times 16 \times 20 \times 16$
Channels	[64, 128, 128, 128]
Residual Blocks per Level	2
Normalization	Group Normalization
KL Weight	1×10^{-7}
Adversarial Weight	0.025
Perceptual Weight	0.001
Perceptual Net	SqueezeNet (fake-3D ratio: 0.5)
Discriminator Layers / Channels	3 / 64
Optimizer	Adam
Learning Rate (Generator)	5×10^{-5}
Learning Rate (Discriminator)	1×10^{-4}
Batch Size (Effective)	8 (1 GPUs x 1, grad. accum. 8)
Training Time	20 epochs / $\sim$ 3 days
Pretrained Init	UK Biobank [121]
Hardware	1 $\times$ NVIDIA A100 80GB

Reconstruction Loss We use an ℓ_1 loss to measure pixel-wise differences between the original image $\mathbf{x}$ and the reconstruction $\tilde{\mathbf{x}}$:

$$\mathcal{L}_{\text{recons}} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|_1.$$

Table 4.2: Latent Diffusion Model pretraining configuration

Parameter	Value
Input Latent Shape	$3 \times 16 \times 20 \times 16$
Channels	[256, 512, 768]
Residual Blocks per Level	2
Attention Resolutions (factors)	[2, 4]
Dropout	0.1
Timesteps	1,000
Beta Schedule	Linear: $\beta_t \in [10^{-4}, 2 \times 10^{-2}]$
Optimizer	Adam
Learning Rate	2.5×10^{-5}
EMA Decay	0.999
Batch Size (Effective)	128 (2 GPUs × 64, grad. accum. = 2)
Training Time	500 epochs / $\sim$ 18 hours
Hardware	2 × NVIDIA A100 SXM4 40GB

KL Divergence Loss This term encourages the encoder's latent distribution to approximate a standard normal. In practice, we keep λ_{KL} sufficiently small (e.g., 10^{-7}) to prevent over-regularizing the latent space and losing high-frequency details. Let the latent representation have $L = h \times w \times d \times c$ total elements. Then the KL term is:

$$\mathcal{L}_{\text{KL}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{\ell=1}^{L} \frac{1}{2} \big(\mu_{i,\ell}^2 + \sigma_{i,\ell}^2 - \log(\sigma_{i,\ell}^2) - 1\big),$$

where $\mu_{i,\ell}$ and $\sigma_{i,\ell}$ are the mean and standard deviation for the ℓ-th latent component of the i-th sample.

Perceptual Loss We employ a feature-space perceptual loss [185]:

$$\mathcal{L}_{\text{perceptual}} = \sum_{l} \lambda_l \frac{1}{N_l} \|\phi_l(\mathbf{x}) - \phi_l(\tilde{\mathbf{x}})\|_2^2,$$

where $\phi_l(\cdot)$ denotes the activation of the l-th layer of a pretrained network, N_l is the number of features in layer l, and λ_l is the layer weight. We used a lightweight SqueezeNet pretrained on ImageNet in our implementation. For each anatomical plane (axial, sagittal, coronal), 50% of the slices are randomly selected. The loss is computed independently on the selected slices of each plane, then averaged first across slices within each plane and subsequently across the three planes, a strategy we refer to as fake-3D in Table 4.1.

Table 4.3: Representation learning stage configuration

Semantic Encoder	
Architecture	2.5D Attention-based Encoder
Slicing plane	Axial
Backbone	`ConvNeXt-Small`
Pretrained Init	ImageNet
Input Modality	$1 \times 128 \times 128 \times 160$
Input Sequence Length	128 slices
Multihead Attention	8 heads, attention dropout 0.1
Output Representation	Global, non-spatial vector $\mathbf{y}_{sem} \in \mathbb{R}^{768}$
Gradient Estimator G_ψ	
Architecture	Modified U-Net (same configuration of LDM)
Shared Components	Downsampling blocks from pretrained LDM
New Components	Middle, upsampling, and output blocks
Condition Injection	AdaGN
Weighting Scheme	SNR based, $\gamma = 0.1$
Training Configuration	
Optimizer	Adam
Learning Rate	2.5×10^{-5}
EMA Decay	0.999
Effective Batch Size	16 (2 GPUs × 2 × grad. accum. 4)
Training Duration	200 epochs / $\sim$2 days
Hardware	2 × NVIDIA A100 SXM4 40GB

Table 4.4: 3D CNN Semantic Encoder configuration (ablation experiment)

Parameter	**Value**
Input Shape	$1 \times 128 \times 160 \times 128$
Channels	[96, 96, 192, 192, 384]
Residual Blocks per Level	2
Attention Resolutions (factors)	[2, 4]
Dropout	0.1
Output Representation	Global, non-spatial vector $\mathbf{y}_{sem} \in \mathbb{R}^{768}$

Adversarial Loss To encourage sharper local detail and stable gradients, we use a patch-based Least Squares GAN objective [34, 184]. For the generator (autoencoder):

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{\mathbf{x}}\big[(D(\tilde{\mathbf{x}}) - 1)^2\big],$$

and for the discriminator:

$$\mathcal{L}_D = \frac{1}{2}\mathbb{E}_{\mathbf{x}}\big[(D(\mathbf{x}) - 1)^2\big] + \frac{1}{2}\mathbb{E}_{\mathbf{x}}\big[D(\tilde{\mathbf{x}})^2\big].$$

Here, $D(\cdot)$ indicates a patch discriminator that computes the loss on multiple local patches rather than the entire volume. The adopted patch-based adversarial loss forces the model to preserve local image textures. The discriminator network parameters used in our implementation are reported in Table 4.1.

4.3.2 Latent Diffusion Model - Pretraining

The trained perceptual compression model provides a compact representation of the input scan by a factor f along each dimension, leading to a volume with f^3 times fewer voxels. This model discards imperceptible high-frequency details while retaining crucial semantics embedded in low-frequency structures, making it an efficient yet expressive space for generative modelling. However, to ensure stability and improve the effectiveness of diffusion training, we normalize the learned latents as suggested in Appendix G. of [133]. Specifically, after encoding with $\mathcal{E}$, we rescale the latent representation to have unit variance across the first batch, ensuring a standardized latent space that facilitates diffusion learning.

In this stage, we train a DDPM without conditioning on the input image, resulting in a time-conditioned UNet operating in the compressed space. The latent lower bound is:

$$L_{LDM} = \sum_{t=1}^{T} \mathbb{E}_{\mathcal{E}(x),\epsilon_t}\left[\|\epsilon_\theta(\mathbf{z}_t, t) - \epsilon_t\|_2^2\right]$$

in which $\epsilon_{\mathbf{t}} \in \mathbb{R}^{h\times w\times d\times c} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{z_t} = \sqrt{\alpha_t}\mathbf{z_0} + \sqrt{1-\alpha_t}\epsilon_{\mathbf{t}}$ and $t \sim \mathcal{U}(0, T)$ during training with $T = 1000$. For time conditioning, the sampled timestep t is first encoded using sinusoidal positional encoding, then projected to produce a time embedding. This embedding is injected into the network via AdaGN [32] in each residual block, where it generates scale and shift parameters that modulate the normalized feature maps. The reverse DDIM process using Eq.4.7 will sample a new $\hat{\mathbf{z_0}}$ from the compressed latent distribution $p(\mathbf{z})$ that can be decoded to image space trough $\mathcal{D}$.

4.3.3 Representation Learning - LDAE

Now that we have pretrained the LDM we can employ a semantic encoder that maps $\mathbf{x}_0$ into $\mathbf{y}_{sem} = Enc_\phi(\mathbf{x}_0)$ and a gradient estimator for the compressed space guidance simulation $G_\psi(\mathbf{z}_t, \mathbf{y}_{sem}, t)$. Similarly to [186] the gradient estimator G_ψ is used to estimate $\nabla_{\mathbf{z}_t} \log p(\mathbf{y}_{sem}|\mathbf{z}_t)$ and the conditional decoder will approximate:

$$p_{\theta,\psi}(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{y}_{sem}) = \mathcal{N}\Big(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t) + \sigma_t \cdot G_\psi(\mathbf{z}_t, \mathbf{y}_{sem}, t), \sigma_t\Big). \tag{4.13}$$

LDAE is trained as a regular DPM by optimizing the following variational lower bound derived objective:

$$\begin{aligned} L_{LDAE}(\psi, \phi) = \mathbb{E}_{\mathbf{x}_0,t,\epsilon}\Bigg[\lambda_t \Bigg\| \epsilon - \epsilon_\theta(\mathbf{z}_t, t) \\ + \frac{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}{\beta_t} \cdot \sigma_t \cdot G_\psi(\mathbf{z}_t, Enc_\phi(\mathbf{x}_0), t)\Bigg\|^2\Bigg]. \end{aligned} \tag{4.14}$$

Note that: (i) the pretrained LDM parameters θ are frozen during this phase, consequently Eq. 4.14 optimizes only parameters ψ and ϕ; and (ii) the gradient estimator operates on the compressed latent distribution ($p(\mathbf{z})$) while the encoder process the original image ($\mathbf{x}_0$). This adopted optimization, similar to PDAE, forces the predicted mean shift $\sigma_t \cdot G_\psi(\mathbf{z}_t, Enc_\phi(\mathbf{x}_0), t)$ to fill the latent posterior mean gap $\tilde{\mu}_t(\mathbf{z}_t, \mathbf{z}_0) - \mu_\theta(\mathbf{z}_t, t)$ by learning as much information as possible $\mathbf{y}_{sem}$ from $\mathbf{x}_0$.

Following [186], we adopt their proposed weighting scheme of the diffusion loss (Eq. 4.14), which has been shown to improve training stability and representation learning. Instead of using a constant weighting factor ($\lambda_t = 1$), they suggest a signal-to-noise ratio (SNR)-based scheme:

$$\lambda_t = \left(\frac{1}{1 + \text{SNR}(t)}\right)^{1-\gamma} \cdot \left(\frac{\text{SNR}(t)}{1 + \text{SNR}(t)}\right)^{\gamma}, \tag{4.15}$$

where $\text{SNR}(t) = \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}$ and γ is a hyperparameter controlling the balance between early-stage and late-stage weighting. Following their recommendation, we set $\gamma = 0.1$ as it down-weights the loss for both very low and very high diffusion steps while encouraging the model to focus on learning richer representations in intermediate stages.

4.3.4 Encoder Design

To enable efficient training and convergence in unsupervised representation learning, we adopted an encoder design inspired by the diagnostic framework proposed in Chapter 3.

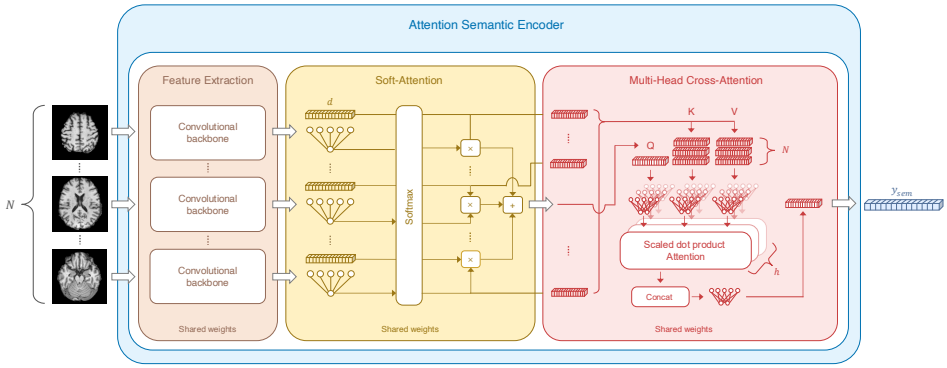


Figure 4.2: **Architecture of the Semantic Encoder used in the LDAE framework.** The input 3D brain MRI scan $x_0 \in \mathbb{R}^{H\times W\times D}$ is sliced along the axial plane into a sequence of 2D slices, each processed independently through a shared 2D CNN backbone (e.g., ConvNeXt-Small) to extract slice-level embeddings $e_i \in \mathbb{R}^d$. These embeddings are then aggregated via a two-stage attention mechanism: *SoftAttention* computes a global summary vector Q as a weighted mean over the sequence, and *CrossAttention* simulates self-attention by querying Q with the original embeddings E, yielding a final global non-spatial semantic vector $y_{\text{sem}} \in \mathbb{R}^d$ used to guide the reverse diffusion process.

In the previous chapter, we observed that 2D models improved convergence speed and performance in data scarcity scenarios. Our encoder leverages a 2D CNN for feature extraction while back-propagating the error signal at the volume level. Specifically, we considered axial slices as a sequential input of 2D images, each embedded by a 2D CNN. An overview of the new proposed encoder architecture is shown in Figure 4.2.

Formally, given a sequence of axial slices $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^{L}$, each slice is independently encoded using a shared 2D convolutional encoder f_{cnn}:

$$\mathbf{e}_i = f_{\text{cnn}}(\mathbf{x}_i), \quad \mathbf{e}_i \in \mathbb{R}^d$$

Stacking the embeddings yields the matrix:

$$\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_L] \in \mathbb{R}^{L\times d}$$

where L is the number of slices and d is the feature dimension. In order to aggregate the slice-level features into a single global representation $\mathbf{y}_{\text{sem}} \in \mathbb{R}^d$, we employ a two-stage attention mechanism consisting of soft-attention followed by multi-head cross-attention. We first compute a summary vector $\mathbf{Q} \in \mathbb{R}^{1\times d}$ as a weighted average of the slice em-

beddings:

$$\mathbf{Q} = \sum_{i=1}^{L} \alpha_i \mathbf{e}_i$$

where the attention weights $\alpha_i \in [0, 1]$ are computed as:

$$\alpha_i = \frac{\exp(w_i)}{\sum_{j=1}^{L} \exp(w_j)}, \quad w_i = \text{MLP}(\mathbf{e}_i)$$

with $w_i \in \mathbb{R}$ obtained from a learnable scalar projection (e.g., a shared MLP) applied to each $\mathbf{e}_i$. This allows the model to focus on the most informative slices in a data-driven manner. To refine the global summary $\mathbf{Q}$ with contextual information from the full sequence, we apply multi-head cross-attention using $\mathbf{Q}$ as the query and $\mathbf{E}$ as both key and value:

$$\mathbf{y}_{\text{sem}} = \text{MultiHeadAttention}(\mathbf{Q}, \mathbf{K} = \mathbf{E}, \mathbf{V} = \mathbf{E})$$

For each attention head $h \in \{1, \ldots, H\}$, the computation is:

$$\text{head}^{(h)} = \text{softmax}\left(\frac{(\mathbf{Q}\mathbf{W}_Q^{(h)})(\mathbf{E}\mathbf{W}_K^{(h)})^\top}{\sqrt{d_h}}\right)\mathbf{E}\mathbf{W}_V^{(h)} \in \mathbb{R}^{1\times d_h}$$

where $d_h = d/H$ and $\mathbf{W}_Q^{(h)}, \mathbf{W}_K^{(h)}, \mathbf{W}_V^{(h)} \in \mathbb{R}^{d\times d_h}$ are learned projection matrices. The final representation is obtained by concatenating all heads and applying a linear projection:

$$\mathbf{y}_{\text{sem}} = \text{Concat}(\text{head}^{(1)}, \ldots, \text{head}^{(H)})\mathbf{W}_O \in \mathbb{R}^{1\times d}$$

with $\mathbf{W}_O \in \mathbb{R}^{d\times d}$ a learned output projection.

This formulation yields a single global vector $\mathbf{y}_{\text{sem}}$, a context-aware summary that reflects the full slice sequence. This approach is compatible with standard 2D CNN backbones and supports the use of pretrained weights while enabling semantic aggregation across the volumetric axis.

4.3.5 Gradient-Estimator Design

The gradient estimator $G_\psi(\mathbf{z}_t, \mathbf{y}_{sem}, t)$ is implemented as a modified UNet architecture as proposed in [186]. It shares the same downsampling path and time embedding modules as the pre-trained LDM introduced in Section 4.3.2, enabling reuse of the learned representations from the unconditional model. However, to enable conditional guidance based on the semantic encoding $\mathbf{y}_{sem}$, we add a second, dedicated upsampling path. As a result, G_ψ consists of a shared encoder and two decoder branches:

1. The first branch (original) corresponds to the unconditional denoiser ϵ_θ, which was optimized during the LDM pretraining stage and remains fixed (i.e., weights are frozen) throughout the current training procedure.

2. The second branch is newly initialized, including its middle blocks, upsampling path, and output layers, and uses the same skip connections of the frozen encoder. This branch is trained from scratch using Eq. 4.14.

Following the conditioning strategy proposed in [186], we use AdaGN [32] to inject both timestep t and semantic condition $\mathbf{y}_{sem}$ into the new decoder. AdaGN applies a learned affine transformation to the normalized feature map $\mathbf{h}$:

$$\text{AdaGN}(\mathbf{h}, t, \mathbf{y}_{sem}) = \mathbf{y}^s_{sem}\,(\mathbf{t}_s \cdot \text{GroupNorm}(\mathbf{h}) + \mathbf{t}_b) + \mathbf{y}^b_{sem},$$

where $[\mathbf{y}^s_{sem}, \mathbf{y}^b_{sem}]$ and $[\mathbf{t}_s, \mathbf{t}_b]$ are scale and bias vectors obtained via linear projections of $\mathbf{y}_{sem}$ and t, respectively. This design allows the model to retain the low-level representations learned during the unconditional pretraining, while enabling the new branch to learn how to guide the denoising process using the high-level semantic codes.

4.3.6 Gradient Estimator as Stochastic Encoder and Conditional Decoder

While serving as an estimator for the denoising function $\hat{\epsilon}_{\theta,\psi}$, the gradient network $G_\psi(\mathbf{z}_t, \mathbf{y}_{sem}, t)$ assumes a dual role within the proposed LDAE framework: it operates both as a *stochastic encoder* and a *conditional decoder*. This dual capability, initially introduced in the DAE framework [124], is fundamental to supporting both accurate reconstruction and controllable semantic generation.

Stochastic Encoder via DDIM Inversion To encode a compressed sample $\mathbf{z_0} = \mathcal{E}(\mathbf{x_0})$ into its stochastic latent code $\mathbf{z}_T$, we employ a deterministic DDIM-like backward (Eq. 4.8) using the semantic code $\mathbf{y}_{sem} = \text{Enc}_\phi(\mathbf{x}_0)$ and the gradient estimator G_ψ:

$$\mathbf{z}_{t+1} = \sqrt{\bar{\alpha}_{t+1}} f_\theta(\mathbf{z}_t, t, \mathbf{y}_{sem}) + \sqrt{1-\bar{\alpha}_{t+1}}\,\hat{\epsilon}_{\theta,\psi}(\mathbf{z}_t, t, \mathbf{y}_{sem}), \tag{4.16}$$

where $\mathbf{f}_\theta$ is the DDIM predictor and ϵ_θ is approximated using G_ψ. Although the DDIM process is deterministic, the latent code $\mathbf{z}_T$ encodes the stochastic, subject-specific aspects of the input (e.g., anatomical variability, or subtle structural features) that are not captured by the semantic embedding $\mathbf{y}_{\text{sem}}$, which primarily represents global brain morphology. This approach is motivated by the DAE framework [124], in which semantic

content is extracted by the encoder, while the DDIM inversion reconstructs residual variability. Together, this enables near-exact reconstructions and controllable manipulation while preserving the original image characteristics.

Conditional Decoder for Reconstruction and Generation Conversely, the same gradient estimator can decode a noisy latent code $\mathbf{z}_T$ into a clean latent $\mathbf{z}_0$ via the conditional DDIM reverse (generative) process. This decoding can be performed in two modes:

1. **Reconstruction:** starting from the inferred $\mathbf{z}_T$ obtained via inversion and $\mathbf{y}_{sem}$, enabling near-exact reconstructions.
2. **Generation:** starting from pure noise $\mathbf{z}_T \sim \mathcal{N}(0, I)$ and guiding the process with $\mathbf{y}_{sem}$, producing samples that shares semantic attributes with x_0.

This dual capability allows the model to interpolate in latent space, perform counterfactual generation, and reconstruct missing intermediate timepoints, as explored in our semantic evaluation experiments (see Section 4.4).

4.3.7 Controlling Semantic Attributes via Latent Linear Directions

Once a semantic-rich encoder is trained (Sections 4.3.3–4.3.4), we can manipulate the latent representation $\mathbf{y}_{sem} \in \mathbb{R}^d$ to alter specific features in the reconstructed 3D scan as proposed in [124]. Concretely, suppose we train a linear classifier

$$\ell(\mathbf{y}_{sem}) = \mathbf{w}^\top \mathbf{y}_{sem} + b$$

to distinguish, for example, AD from CN participants. The set of points $\mathbf{y}_{sem}$ satisfying $\mathbf{w}^\top \mathbf{y}_{sem} + b = 0$ defines an hyperplane in $\mathbb{R}^d$. By definition, the weight vector $\mathbf{w}$ is orthogonal to this hyperplane, meaning that $\mathbf{w}^\top(\mathbf{y}_{sem,2} - \mathbf{y}_{sem,1}) = 0$ for any two points $\mathbf{y}_{sem,1}$ and $\mathbf{y}_{sem,2}$ on the decision boundary. Hence, $\mathbf{w}$ defines a principal direction in $\mathbf{y}_{sem}$-space along which movement will have the maximum effect on the classifier's output ($\ell(\mathbf{y}_{sem})$). Intuitively, translating $\mathbf{y}_{sem}$ in the direction of $\mathbf{w}$ (i.e., $\mathbf{y}_{sem} \leftarrow \mathbf{y}_{sem} + \alpha\mathbf{w}$) will "add" the corresponding AD-related features to the reconstructed scan, whereas moving in the opposite direction ($\alpha < 0$) will "subtract" them. Empirically, this directional manipulation can disentangle specific symptoms or morphological traits in the semantic representation, thereby giving control over generated 3D reconstructions.

4.3.8 Semantically meaningful interpolation

To assess the semantic smoothness of the learned semantic space, we performed interpolation using both the semantic and stochastic codes. This enables qualitative and quantitative evaluation of how well the latent space captures continuous and clinically meaningful transformations across brain MRI scans.

Following the DAE framework [124], we perform linear interpolation in the semantic space and spherical linear interpolation in the stochastic one. Given two input scans $\mathbf{x}_1$ and $\mathbf{x}_2$, their corresponding semantic and stochastic representations are denoted as $\mathbf{y}^1_{\text{sem}}, \mathbf{z}^1_T$ and $\mathbf{y}^2_{\text{sem}}, \mathbf{z}^2_T$, respectively. The interpolated latent representations at interpolation factor $t \in [0, 1]$ are computed as:

$$\text{LERP}(\mathbf{y}^1_{\text{sem}}, \mathbf{y}^2_{\text{sem}}; t) = (1-t) \cdot \mathbf{y}^1_{\text{sem}} + t \cdot \mathbf{y}^2_{\text{sem}} \tag{4.17}$$

$$\text{SLERP}(\mathbf{z}^1_T, \mathbf{z}^2_T; t) = \frac{\sin((1-t)\theta)}{\sin(\theta)}\mathbf{z}^1_T + \frac{\sin(t\theta)}{\sin(\theta)}\mathbf{z}^2_T \tag{4.18}$$

where the angle θ between $\mathbf{z}^1_T$ and $\mathbf{z}^2_T$ is given by:

$$\theta = \arccos\left(\frac{\langle \mathbf{z}^1_T, \mathbf{z}^2_T \rangle}{\|\mathbf{z}^1_T\| \cdot \|\mathbf{z}^2_T\|}\right) \tag{4.19}$$

The interpolated pair $(\mathbf{y}^{(t)}_{\text{sem}}, \mathbf{z}^{(t)}_T)$ is then decoded using the reconstruction procedure described in Section 4.3.6.

4.4 Experiments

4.4.1 Dataset and Preprocessing

We performed our experiments on the ADNI database considering all the stages of cognitive decline, ranging from CN to MCI and AD. The demographic information of the subjects in the dataset is summarized in Table 4.5, providing essential context for our analysis and findings.

BIDS Conversion As in Sec. 3.4.1, the dataset was first converted into the BIDS format [57] using the Clinica platform [135, 138]. During this phase, the Clinica `ADNI-to-BIDS`

Table 4.5: Population statistics across CN, AD, and MCI groups. The statistics include age, mini-mental state examination (MMSE) and global clinical dementia rating (CDR) scores.

Group	Subjects	Samples	Age	MMSE	CDR
CN	965	3673	75.43 ± 6.83	29.06 ± 1.20	0.03 ± 0.17
AD	748	2064	76.36 ± 7.59	21.79 ± 4.44	0.95 ± 0.52
MCI	1193	4603	74.72 ± 7.74	27.51 ± 2.24	0.46 ± 0.22

converter automatically applies a quality control check, selecting the preferred scan for each visit and discarding scans that fail predefined quality checks.

In addition to ADNI, we evaluated model transferability on two external datasets: OASIS-3 [89] and AIBL [40]. Both datasets were also converted to BIDS format using the dedicated Clinica tools `OASIS3-to-BIDS` and `AIBL-to-BIDS`, respectively. This standardization ensured consistent preprocessing and inference across datasets while ensuring compatibility with the pipelines developed for ADNI.

Preprocessing Pipeline The preprocessed dataset was prepared using a slightly modified version of the neuroimaging pipeline detailed in Sec. 3.4.1. The pipeline consisted of the following sequential steps:

1. **Bias Field Correction:** Intensity non-uniformities were corrected using the N4ITK algorithm [159].
2. **Skull Stripping:** Brain tissue was extracted using the deep learning-based brain extraction models proposed in [24] from ANTsPyNet package.
3. **Affine Registration:** Each brain-extracted volume was affinely aligned to the MNI152 ICBM 2009c nonlinear symmetric template [50, 49] using the SyN algorithm from the ANTs toolkit [5, 6].

Inference-Time Normalization We employed the MONAI framework [11] for batch preprocessing at inference time. The following transformations were applied:

1. **Voxel Spacing Normalization:** All images were resampled to a uniform voxel spacing of 1.5mm isotropic using B-spline interpolation.

2. **Spatial Resizing:** Volumes were resized using cropping or padding to a target shape (e.g., $128 \times 160 \times 128$).

3. **Intensity Normalization:** Intensities were rescaled to the $[0, 1]$ range using min-max normalization.

4.4.2 Experimental Setup

AutoencoderKL Training To enable efficient diffusion training in a compressed representation space, we fine-tuned a 3D perceptual autoencoder to compress full-resolution MRI scans into a compact latent representation. We used the `AutoencoderKL` implementation from the `MONAI Generative` [122] framework. The model was initialized from a checkpoint pre-trained on the UK Biobank dataset [154] and fine-tuned on volumes resampled to a resolution of $128 \times 160 \times 128$. The detailed training setup is reported in Table 4.1.

Latent Diffusion Model Pretraining Following the compression model training stage, we pretrained a DDPM directly in the compressed latent space. This stage models the distribution of latent representations $\mathbf{z}_0 \in \mathbb{R}^{3\times16\times20\times16}$ ($170\times$ smaller), significantly reducing the computational burden compared to operating in voxel space. The training follows the standard DDPM formulation with a linear noise schedule and uses a 3D UNet architecture trained to predict the added Gaussian noise in the latent space. The Exponential Moving Average (EMA) of model weights update was applied during training for improved stability and sample quality. A full list of architectural and training parameters is provided in Table 4.2.

Representation Learning via 2.5D Semantic Encoder To guide the reverse diffusion process with clinically meaningful features, we trained a semantic encoder network to extract a compact latent code $\mathbf{y}_{sem}$ from the original 3D brain volume $\mathbf{x}_0$. This latent code serves as a conditioning vector during denoising via the gradient estimator $G_\psi(\mathbf{z}_t, \mathbf{y}_{sem}, t)$, as detailed in Section 4.3.3. The encoder follows a 2.5D strategy, where axial slices are independently processed using a 2D CNN backbone, and then aggregated using a combination of soft attention and cross-attention mechanisms to produce a global, non-spatial embedding. We employed a pre-trained ConvNeXt-Small model as the slice-level feature extractor, adapted to single-channel inputs (grayscale MRI slices), with embedding dimension $d = 768$ and input sequence length $L = 128$. The grayscale adaptation consisted of summing the pre-trained convolutional filters of the backbone's

first layer, thus allowing the usage of pre-trained weights without adding extra channel dimensions. The semantic encoder and gradient estimator were optimized jointly to minimize the weighted diffusion loss described in Equation 4.14, using the Signal-to-Noise Ratio (SNR)-based weighting scheme proposed by [186]. The gradient estimator G_ψ was implemented as a modified UNet architecture, reusing the encoder and time embedding modules from the pre-trained unconditional LDM (see Table 4.3). A new decoder branch, composed of middle, upsampling and output blocks, was initialized with the same configuration of the LDM. A full list of architecture and training parameters for this stage is provided in Table 4.3. To assess the contribution of our 2.5D design, we conducted an ablation experiment replacing the semantic encoder with a 3D CNN variant, configured as detailed in Table 4.4.

Data Splits and Model Selection Strategy We adopted a consistent 90-10 train-test split at the subject level for all training stages. Within the 90% training partition, a random 1% subset was reserved for validation purposes. This minimal validation split was chosen to preserve maximal training data availability, as the main objective was image reconstruction rather than classification. Validation was conducted using image-level reconstruction metrics. For the *compression model* and the *representation learning stage*, model selection was based on the best Structural Similarity Index Measure (SSIM) score observed on the validation set. In the representation learning stage, we specifically monitored the SSIM between the original image and the reconstruction generated using the semantic embedding $\mathbf{y}_{sem}$ and pure noise as stochastic input. This metric served as a proxy for the quality and completeness of the semantic information extracted by the encoder. No validation-based checkpointing was used in the *latent diffusion model pretraining* stage. Instead, the model parameters at the final epoch were retained for downstream use, as unconditional diffusion sampling was the primary goal, and model weights updates were performed through an EMA strategy. For quantitative comparison, we evaluated reconstruction performance using MAE, SSIM, and the Learned Perceptual Image Patch Similarity (LPIPS) metric, which corresponds to the perceptual loss used in the compression model (Section 4.3.1).

Linear Probe Setup To evaluate the semantic quality of the learned representations, we performed linear probe experiments on two downstream tasks: AD diagnosis and age prediction. For this purpose, we trained linear models on top of the fixed semantic embeddings extracted from the trained LDAE encoder. Given an input 3D brain MRI scan $\mathbf{x}_0$, we first projected each image into the semantic space using the encoder $\text{Enc}_\phi(\mathbf{x}_0)$, obtaining a semantic vector $\mathbf{y}_{\text{sem}} \in \mathbb{R}^{768}$. These embeddings were computed for all samples in the dataset. The dataset was initially split at the subject level into 90% training

and 10% test sets. From the 90% training portion, we held out 1% for validation and used the remaining 89% for training. We further split this 89% into 70% training and 30% validation subsets for the linear probe experiments. The original 10% test set was kept as a fixed benchmark for final evaluation. For the AD vs. CN classification task, we trained a linear classifier consisting of a single fully connected layer with binary cross-entropy loss. We trained a linear regressor using Mean Squared Error (MSE) loss for the age prediction task. The classifier was trained using the Adam optimizer for 200 epochs with a 1×10^{-3} learning rate. The regressor was trained using Stochastic Gradient Descent (SGD) for 1000 epochs with the same learning rate. Note that all linear models were trained on the fixed semantic vectors without finetuning the encoder.

These linear probe experiments aim to quantify the extent to which the learned semantic codes $\mathbf{y}_{\text{sem}}$ encode clinically meaningful and linearly separable features relevant to disease classification and age estimation.

Table 4.6: Comparison of voxel-space DAE and latent-space LDAE in terms of input resolution, model size, training duration, and reconstruction efficiency.

Metric	**DAE**	**LDAE**
Input Resolution	$112 \times 128 \times 112$	$128 \times 160 \times 128$
Model Parameters	130M	920M
Training Duration	140 epochs / $\sim$1 week	500+200 epochs / $\sim$3 days
Hardware	2 × A100 40GB	2 × A100 40GB
Inference Time ($T = 100$)	$\sim$120 seconds	$\sim$6 seconds
SSIM ($\uparrow$) at $T = 100$	0.892	**0.962**
LPIPS ($\downarrow$) at $T = 100$	**0.038**	0.076

Interpolation for Missing Scan Generation Our LDAE framework posits a compact, semantically continuous latent space where proximity and direction correlate meaningfully with input data properties. Since the hypothesis is that this space can linearly separate subjects by clinical status (AD vs. CN) and age (Sec. 4.5.4), interpolation serves as a critical test of its continuity. We hypothesize that if the space meaningfully organizes subjects based on clinical state, it should also map the continuous progression between these states. Thus, the primary objective is to validate that moving between two points in this space corresponds to a smooth, anatomically plausible transformation in the image domain. Successful interpolation, yielding a realistic intermediate scan, provides strong evidence that the model learns fundamental generative factors of temporal brain changes, not just discrete states. The ability to generate high-fidelity intermediate scans has significant practical implications:

- **Missing Data Imputation in Clinical Trials**: Our model can generate plausible intermediate scans (e.g., a Year 2 scan between Year 1 and Year 3) for longitudinal studies, addressing participant dropout and missed appointments. This enhances statistical power and reduces bias by enabling inclusion of subjects requiring complete time-series data.

- **Dense Data Augmentation for Predictive Models**: Interpolating between existing scans can generate synthetic data at finer temporal resolutions (e.g., 3, 6, and 9 months). This denser dataset can train more robust predictive models, sensitive to subtle, short-term changes in disease onset or conversion.

To formally assess this, we simulated missing follow-up scan generation by interpolating between a subject's earlier and later visits. It is important to recall that all input scans are nonlinearly registered to a common anatomical template, a crucial preprocessing step that minimizes inter-scan variability from head positioning or acquisition. This allows the model to focus on capturing underlying neuroanatomical and pathological changes over time, which is the biological progression this experiment aims to reflect.

We conducted the interpolation experiment on a subset of the held-out test set. For each subject with at least three visits in the test set, we selected all valid $(start, target, end)$ triplets such that the target timepoint lies temporally between the start and end. For each triplet, we computed the interpolation factor α based on the temporal offset of the target with respect to the interval defined by start and end. We then interpolated linearly between the semantic codes ($\mathbf{y}_{\text{sem}}$) and spherically between the stochastic codes ($\mathbf{z}_T$) to generate the latent pair $(\mathbf{y}_{\text{sem}}^{(t)}, \mathbf{z}_T^{(t)})$ corresponding to the target intermediate timepoint t. This experiment was conducted on a subset of 30 subjects from the test set, yielding approximately 1400 valid triplet configurations. The number of possible triplets T for a subject with n sessions grows approximately as $T \propto \binom{n}{3}$, under the constraint that the target scan lies strictly between the start and end. This implies that even modest increases in subjects considered for the evaluation will increase significantly the number of generations to compute. The generated latent representation was decoded using the DDIM-based reverse process described in Section 4.3.8. The resulting image $\hat{\mathbf{x}}_0$ was compared to the ground truth scan $\mathbf{x}_0^{\text{target}}$ using SSIM and MSE. We report average SSIM and MSE metrics stratified by the temporal gap between start and end scans (time gap), the minimum distance of the predicted target from the endpoints (prediction gap), and the relative position of the target within the interval (normalized to $[0, 1]$). This allows us to assess how interpolation accuracy varies with temporal context.

4.5 Results

In this section, we present experimental results evaluating the proposed LDAE framework. The results are structured to progressively validate the components of the pipeline and support the two main hypotheses introduced in Section 2.1.

4.5.1 AutoencoderKL: Reconstruction Quality

We begin by evaluating the perceptual compression model based on `AutoencoderKL`. This autoencoder compresses each 3D brain MRI from $1 \times 128 \times 160 \times 128$ into a latent representation of size $3 \times 16 \times 20 \times 16$, reducing the volume by a factor of approximately $170\times$. Despite this compression, the model achieves high-fidelity reconstructions, with an SSIM of 0.962 and an MSE of 0.001 on the external test set (see Table 4.7). Qualitative examples are shown in Figure 4.3. These results highlight the model's capacity to retain high-frequency anatomical details in the compressed latent space.

It is important to note that this reconstruction accuracy establishes an upper bound for the performance of the subsequent latent diffusion models (LDDIM and LDAE), as the final decoded output always passes through the AutoencoderKL decoder.

4.5.2 Reconstruction Quality

To evaluate the reconstruction capability of our framework, we compared its performance against state-of-the-art latent diffusion models for brain MRI. It is important to note that models like [121] and [125], which we denote as LDDIM in Table 4.7, are predominantly focused on image generation. While they rely on latent diffusion for high-fidelity synthesis, their core aim is not unsupervised representation learning for downstream analytical tasks. Our comparison with LDDIM, therefore, specifically evaluates its reconstruction quality through DDIM inversion, providing a direct metric of generative fidelity achieved within a latent diffusion framework. Given LDAE's core objective of enabling efficient and semantically meaningful unsupervised representation learning in medical imaging, our primary comparative focus for both reconstruction quality and computational efficiency is with the 3D DAE operating directly in voxel space.

As shown in Figure 4.4 and Table 4.7, the proposed LDAEachieves reconstruction quality on par with AutoencoderKL and LDDIM. Figure 4.4 shows reconstruction results for two representative test subjects: a CN subject (top) and an AD subject (bottom). The first

column shows the original brain MRI scan, while the second column presents the reconstruction obtained using both the semantic embedding $y_{\text{sem}} = \text{Enc}_\phi(x_0)$ and the stochastic latent z_T obtained via DDIM inversion of the encoded compressed latent $z_0 = \mathcal{E}(x_0)$. Columns 3-5 show reconstructions obtained by keeping y_{sem} fixed while sampling different $z_T^{(i)} \sim \mathcal{N}(0, I)$. Despite the stochastic variation, the reconstructions retain global anatomical and disease-relevant structure, indicating that y_{sem} captures high-level semantics while z_T encodes low-level variability. In the AD subject, expected pathological traits (e.g., ventricular enlargement) are consistently preserved, whereas in the CN subject, normal cortical volume and structure remain stable across samples. For instance, at $T = 50$, LDAE obtains SSIM = 0.962, LPIPS = 0.076, and MSE = 0.001, matching the performance of both the LDDIM and the upper bound imposed by the AutoencoderKL. Since LDAE operates in the compressed latent space, it allows efficient model parameters scalability. Specifically, the full parameter count of $\sim$920M does not arise from a single oversized model but from the modular design of our framework. The diffusion UNet backbone follows the same configuration as prior works [121, 125] (see Table 4.2) and accounts for $\sim$434M parameters. The Gradient Estimator reuses the downsampling path of this UNet but introduces its own middle, upsampling, and output blocks, contributing an additional $\sim$384M parameters. Finally, other auxialiry components that enables representation learning (semantic encoder, conditioning layers, and output heads) account for $\sim$50M parameters. Importantly, the UNet backbone remains frozen during representation learning, and the trainable portion (Gradient Estimator + encoder) is closer in scale to other state-of-the-art diffusion models. This modular design is necessary to preserve the pretrained generative prior while enabling stable and tractable semantic representation learning. Despite this, LDAE remains more efficient (see Table 4.6).

For consistency and fairness, LDAE was trained at the same input resolution ($128 \times 160 \times 128$) as the latent diffusion baselines AutoencoderKL and LDDIM. In contrast, the DAE baseline could only be trained at $112 \times 128 \times 112$ due to memory constraints. On NVIDIA A100-SXM4-40GB GPUs, even with mixed precision training, the DAE fit only a batch size of 1 per GPU; to simulate a larger effective batch we relied on gradient accumulation. Under these conditions, the DAE required $\sim$1 week of training on 2 GPUs. LDAE, by comparison, was trained for 200 epochs over 2 days on the same GPU configuration ($2 \times$ A100-SXM4-40GB GPUs).

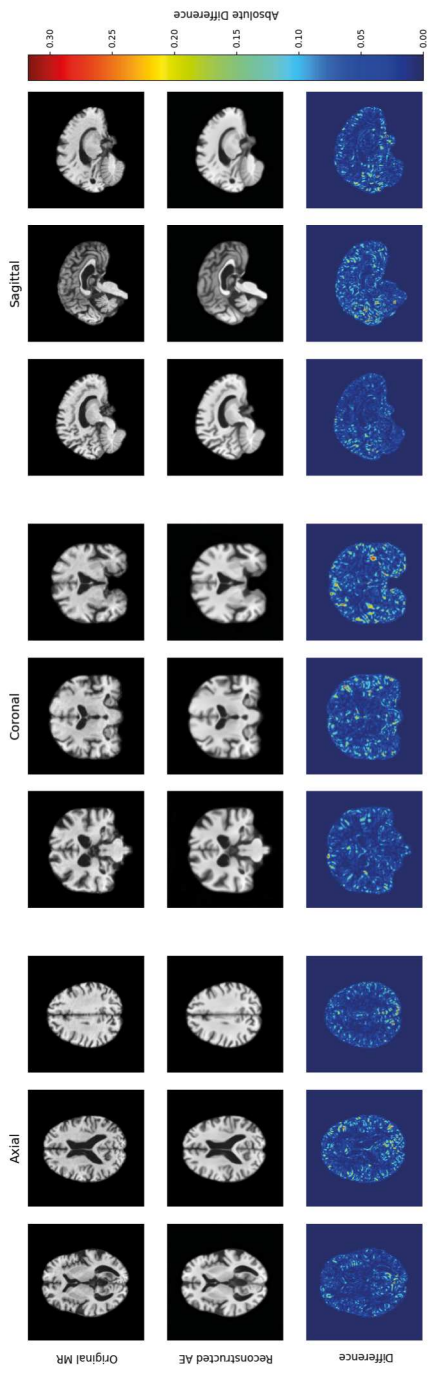


Figure 4.3: Qualitative reconstructions from AutoencoderKL. Top row: original scan slices. Middle row: reconstructed outputs from compressed latent codes. Bottom row: reconstruction error. The reconstructions preserve global and local anatomical features despite the $170\times$ compression.

At inference time, the efficiency gap is even more pronounced: LDAE requires approximately 6 seconds per reconstruction at $T = 100$ steps on an A100 GPU, while the full-resolution DAE takes approximately 2 minutes per scan due to the lack of compression and increased I/O overhead. These comparative results suggest that the DAE's parameter count may be insufficient to model the data distribution in voxel space effectively. However, during training, the UNet configuration was chosen to maximize the GPU memory usage with just one training batch, preventing us from testing higher-capacity variants to validate this hypothesis.

These results validate our second hypothesis: LDAE enables high-fidelity 3D brain MRI reconstruction with full semantic controllability and significantly improved computational efficiency, both during training and inference.

4.5.3 Semantic Guidance Enables Reconstruction from Pure Noise

To assess the semantic richness of the learned representation, we perform reconstructions using only the semantic code $y_{\text{sem}} = \text{Enc}_\phi(\mathbf{x}_0)$ and a randomly sampled stochastic code $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$. This setup evaluates whether the semantic code alone can guide the reverse diffusion process to generate anatomically plausible MRI that contains structure information of $\mathbf{x}_0$.

Figure 4.4 shows qualitative reconstructions for two randomly selected subjects: a CN individual (top) and an AD patient (bottom). For both cases, reconstructions are generated by fixing the semantic code $\mathbf{y}_{\text{sem}} = \text{Enc}_\phi(\mathbf{x}_0)$ and sampling multiple stochastic codes $\mathbf{z}_T$. Across samples, the reconstructions preserve global brain morphology, indicating that the semantic code captures the high-level anatomical and pathological attributes of the subject, while the stochastic component contributes only to low-level variability. In the CN subject, cortical thickness and brain volume are preserved across stochastic samples. In contrast, the AD subject reconstructions consistently display expected atrophic patterns, such as enlarged ventricles and reduced hippocampal volume, despite variation in image details. These observations suggest that $\mathbf{y}_{\text{sem}}$ encodes disease-relevant structural information supporting the first hypothesis. We quantitatively validate this observation in Table 4.7, where reconstructions generated solely from $\mathbf{y}_{sem}$ (without DDIM encoding) achieve an SSIM of 0.872 when compared to the original images. This behaviour is consistent with prior findings in DAE [124], where the semantic representation governs identity and structure, and the stochastic latent controls fine-grained appearance.

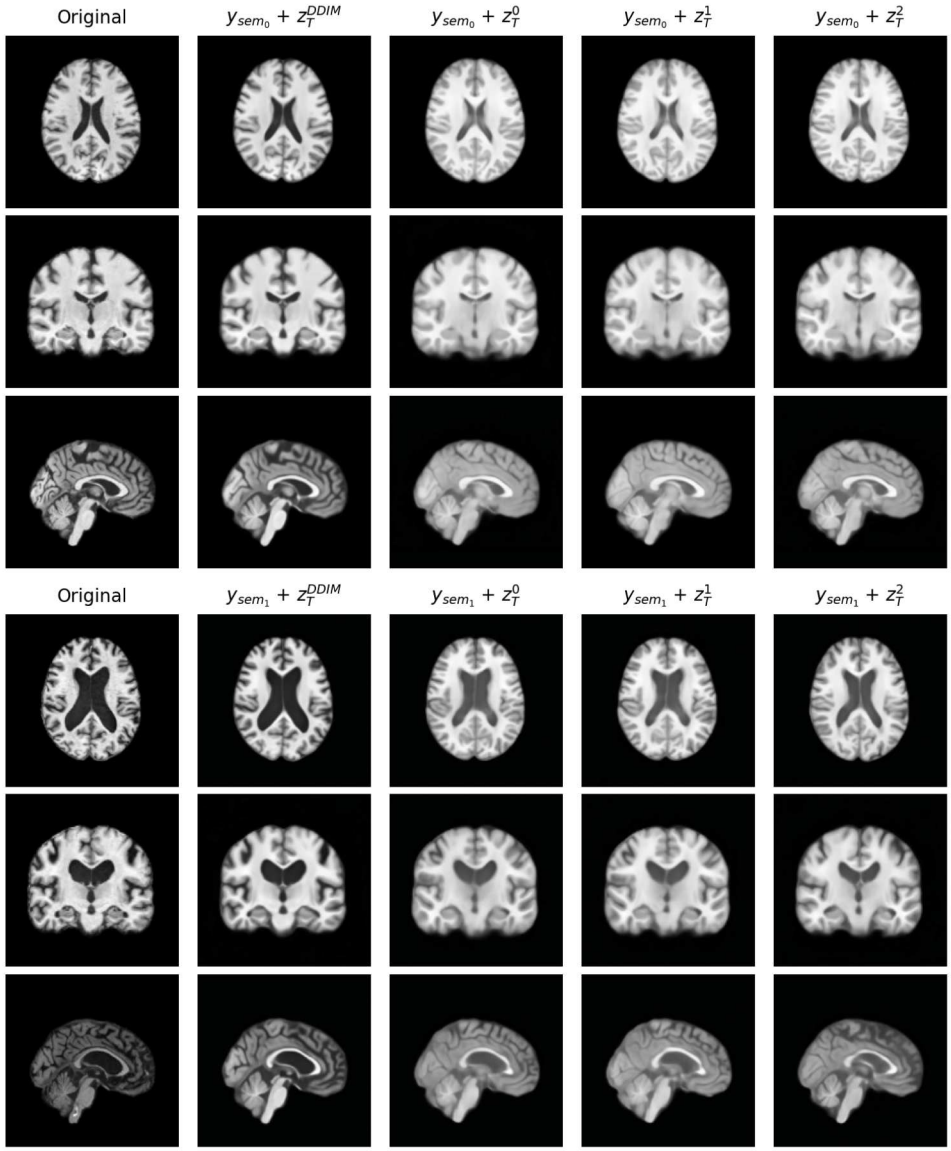


Figure 4.4: Reconstruction from semantic code and stochastic latent for CN (top) and AD (bottom) subjects.

4.5.4 Linear Probe Evaluation on Alzheimer's Disease Classification and Age Prediction

To quantitatively assess the semantic quality of the representations learned by our LDAE framework, we conducted linear probe experiments on two downstream tasks: (i) AD vs. CN classification, and (ii) age prediction. For this purpose, we trained a linear classifier (for AD vs. CN) and a linear regressor (for age) on top of the semantic vectors ($\mathbf{y}_{sem}$) extracted from the pre-trained LDAE encoder. As baselines, we also evaluated (i) a baseline DAE trained in the original voxel space with joint optimization of encoder and diffusion decoder and (ii) a fully supervised semantic encoder trained end-to-end with access to ground-truth diagnostic labels, and (iii) the AutoencoderKL trained for the compression task (Sec. 4.3.1). This model is the same perceptual autoencoder used in recent brain latent diffusion frameworks such as [121] and [125], in which only the encoder component of the AutoencoderKL is capable of performing unsupervised representation learning. As a member of the VAE family, AutoencoderKL represents a strong baseline for unsupervised compression and latent space learning in medical imaging. Given its design for learning a disentangled latent space with a KL divergence regularizer, it is inherently well suited for linear probing to assess the quality of its learned representations.

As shown in Table 4.8, the LDAE embeddings achieved strong performance on both tasks, with 83.65% accuracy and 89.48% AUROC on the AD classification task, and a Mean Absolute Error (MAE) of 4.16 years and RMSE of 5.23 years on the age prediction task. In comparison, while the 3D VAE exhibits a solid MCC score (0.5856), its performance, particularly in the age prediction task, indicates that its learned representations are less robust than LDAE's for linear probing-based regression. This suggests that while a VAE can compress data effectively and its latent space is prone to linear probing, it struggles to capture the complex, discriminative features necessary for high-fidelity clinical downstream tasks. To assess the contribution of the proposed 2.5D semantic encoder, we conducted an ablation study by replacing it with a 3D CNN variant. Despite the comparable parameter count, this configuration failed to achieve stable convergence and yielded lower performance in downstream linear-probe tasks such as AD classification and age prediction. We attribute this to the lack of strong inductive priors and the absence of large-scale pre-trained weights, which limited its ability to learn meaningful semantic representations in our unsupervised setting. These results further confirm that the proposed 2.5D attention-based encoder is not only computationally efficient but also essential for achieving stable training and superior downstream performance within our framework. To further assess whether the learned semantic space $\mathbf{y}_{sem}$ captures disease-relevant information, we applied Linear Discriminant Analysis (LDA) to project the 768-dimensional embeddings onto a single discriminative axis. Figure 4.5 shows kernel den-

sity estimates for the LDA projections of AD and CN subjects in both the training and test sets. Despite the encoder being trained without diagnostic supervision, the semantic embeddings demonstrate clear class separation, confirming that $\mathbf{y}_{sem}$ encodes clinically meaningful features in a linearly discriminative manner.

These findings support Hypothesis 1, demonstrating that the semantic encoder learns semantically rich representations. Moreover, the effectiveness of such representations on both classification and regression tasks confirms their generality across clinically relevant phenotypes.

4.5.5 Zero-Shot Transferability of Semantic Representations to Unseen Datasets

To assess the generalization capability of the proposed LDAE framework, we extend our evaluation to two external cohorts, AIBL and OASIS-3, on which the model was never trained. Importantly, we operate in a strict zero-shot setting: no fine-tuning is performed on the encoder. Instead, we extract semantic embeddings ($\mathbf{y}_{sem}$) from each subject and train a linear classifier for the AD vs. CN discrimination task.

As shown in Table 4.9, the LDAE embeddings yield competitive results. On AIBL, our approach achieves 87.9% accuracy and 0.8757 AUROC, outperforming a fully supervised encoder in terms of F1-score (0.7040 vs. 0.6042) and MCC (0.6311 vs. 0.5438). A similar trend is observed on OASIS-3, where LDAE reaches 76.7% accuracy and 0.8101 AUROC with balanced sensitivity and specificity. It is worth noting that these results are obtained using a simple linear head trained on top of inferenced features, without requiring computationally intensive 3D training, or task-specific optimization. In contrast, fully supervised training on 3D medical images typically demands large memory footprints, and prolonged training schedules due to the high dimensionality of volumetric data. Our results demonstrate that LDAE provides a strong alternative: by learning semantically rich and transferable features in a fully unsupervised manner, it supports rapid adaptation to unseen datasets with minimal computational and training supervision cost. This property makes it especially valuable in medical imaging domains where data heterogeneity and annotation scarcity are common.

Table 4.7: Comparison of **SSIM**, **LPIPS**, and **MSE** for various models at different sampling steps **T**.

Model	Latent dim	SSIM (↑)				LPIPS (↓)				MSE (↓)			
		T=10	T=20	T=50	T=100	T=10	T=20	T=50	T=100	T=10	T=20	T=50	T=100
AutoencoderKL (@14M)	5,120	0.962				0.075				0.001			
LDDIM (@486M)													
a) Encoded $\mathbf{x}_T$	5,120	0.959	0.962	0.962	**0.962**	0.077	0.075	0.075	0.075	**0.001**	**0.001**	**0.001**	**0.001**
LDAE (@920M)													
a) No encoded $\mathbf{x}_T$, from $\mathbf{y}_{sem}$	768	0.872	0.871	0.870	0.869	0.156	0.155	0.155	0.155	0.005	0.005	0.005	0.005
b) Encoded $\mathbf{x}_T$	5,888	0.953	0.960	0.962	**0.962**	0.081	0.077	0.076	0.075	**0.001**	**0.001**	**0.001**	**0.001**
DAE (@130M)													
a) No encoded $\mathbf{x}_T$, from $\mathbf{y}_{sem}$	768	0.272	0.287	0.281	0.283	0.460	0.256	0.182	0.170	0.017	0.016	0.016	0.016
b) Encoded $\mathbf{x}_T$	1,605,632	0.216	0.397	0.684	0.892	0.433	0.132	0.038	**0.030**	0.007	0.002	**0.001**	**0.001**

Table 4.8: Linear probe evaluation using semantic representations learned by the encoders of DAE and LDAE compared to the same encoder trained with supervision. Metrics are reported separately for Alzheimer's disease classification (left) and age prediction (right).

Model	AD vs. CN						Age Prediction	
	Accuracy	Sensitivity	Specificity	F1-score	MCC	AUROC	MAE ↓	RMSE ↓
LDAE (Ours)	0.8365	0.7040	**0.9102**	0.7548	0.6369	0.8948	**4.16**	5.23
DAE (Baseline)	0.7468	0.5605	0.8504	0.6127	0.4312	0.7055	4.93	6.11
3D VAE (Baseline)	0.7949	0.7731	0.8341	0.7440	0.5856	0.8978	13.67	17.79
Supervised Encoder	**0.8464**	**0.8053**	0.8743	**0.8088**	**0.6806**	**0.9067**	4.34	**4.63**
LDAE (3D CNN)	0.7420	0.5067	0.8728	0.5840	0.4132	0.7938	4.88	6.12

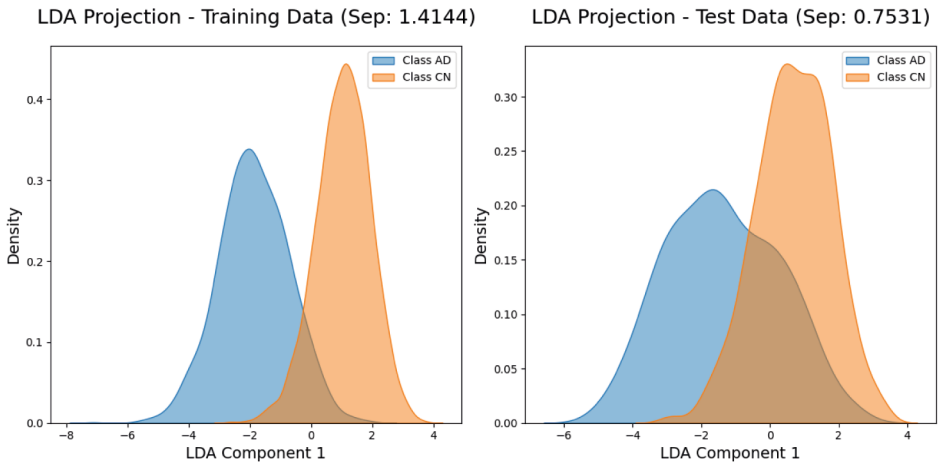


Figure 4.5: LDA projection of semantic representations ($\mathbf{y}_{sem}$) from the LDAE encoder, shown as class-wise density estimates along the first LDA component. Left: training data; right: test data. Although the encoder was not trained with diagnostic labels, the resulting separation between AD and CN classes indicates that the learned semantic space captures disease-relevant features.

4.5.6 Semantic Manipulation via Latent Directions

To qualitatively evaluate the interpretability and controllability of the learned semantic space, we conducted a semantic manipulation experiment following the strategy discussed in Section 4.3.7. Once a linear classifier is trained to distinguish between AD and CN subjects using the semantic representations $\mathbf{y}_{\text{sem}}$, its weight vector $\mathbf{w}$ defines the principal direction along which the most discriminative semantic variation lies. Since the classifier is trained with label 0 assigned to AD and 1 to CN, moving along $-\alpha\mathbf{w}$ amplifies AD-related features (manipulating CN towards AD). In contrast, moving in the opposite direction $\alpha\mathbf{w}$ reduces them (manipulating AD towards CN).

Figure 4.6 shows the effect of different manipulation strength, we progressively increased α from 0.0 to 5.0 for an AD subject. The anatomical changes become increasingly pronounced as α grows, particularly in hippocampal and ventricular regions, highlighting a smooth and meaningful trajectory in the latent space. Figure 4.7 shows two representative examples: in the first case (top), an AD subject is manipulated along the $\alpha\mathbf{w}$ direction towards a CN-like representation, and in the second (bottom), a CN subject is manipulated along the $-\alpha\mathbf{w}$ direction towards an AD-like representation. Both manip-

Table 4.9: Evaluation of zero-shot transferability on unseen datasets (AIBL and OASIS-3). We compare the performance of semantic representations learned by LDAE against a fully supervised encoder trained on ADNI. A simple linear classifier is trained on top of frozen features for the AD vs. CN task.

Dataset / Model	Accuracy	Sensitivity	Specificity	F1-score	MCC	AUROC
AIBL						
LDAE (Ours)	**0.8791**	**0.7586**	**0.9073**	**0.7040**	**0.6311**	0.8757
Supervised Encoder	0.8762	0.7250	0.8989	0.6042	0.5438	**0.9097**
OASIS-3						
LDAE (Ours)	0.7671	**0.6465**	0.7864	0.4339	0.3351	0.8101
Supervised Encoder	**0.8243**	0.5253	**0.8722**	**0.4522**	**0.3548**	**0.8213**

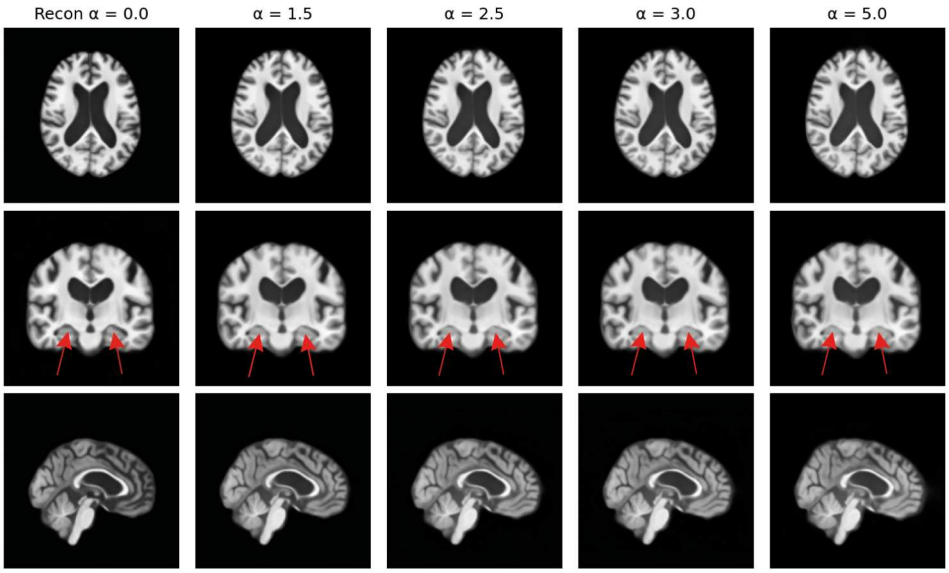


Figure 4.6: The manipulation coefficient α ranges from 0.0 (original reconstruction) to 5.0. With increasing α, AD-related changes are progressively attenuated. In particular, hippocampal and medial temporal lobe recovery is evident (red arrows), while overall age-related cortical atrophy remains preserved across all manipulation strengths. Cortical folding patterns and the subject's global anatomical footprint remain stable, confirming preservation of subject identity and age-related features despite disease-specific modifications.

ulations use a scaling factor $\alpha = 1.5$. In the top panel, a subject originally diagnosed with AD (Original AD) is manipulated toward a cognitively normal (CN)-like brain (Ma-

nipulated CN). Compared to the original scan, the manipulated image exhibits reduced hippocampal and medial temporal lobe atrophy (red arrows), while global age-related cortical atrophy remains comparable, ensuring preservation of age characteristics. Cortical folding patterns remain unchanged, confirming subject identity preservation. In the bottom panel, a subject originally diagnosed as CN (Original CN) is manipulated toward an AD-like brain (Manipulated AD). Here, the manipulated image shows increased hippocampal and medial temporal atrophy (red arrows) while retaining global cortical atrophy patterns and subject-specific cortical architecture. Difference maps (bottom row for each panel) highlight localized structural modifications, with warmer colors indicating greater voxel-wise absolute changes. These results demonstrate the ability of the proposed diffusion-based approach to bidirectionally and selectively modulate AD-related neuroanatomical signatures while preserving subject identity and age-related features.

4.5.7 Interpolation for Missing Scan Generation

To quantitatively assess interpolation accuracy, we compare the generated scan $\hat{\mathbf{x}}_0^{\text{target}}$ against the autoencoder reconstructed scan $\mathcal{D}(\mathcal{E}(\mathbf{x}_0^{\text{target}}))$ using SSIM and MSE. This ensures the evaluation is performed entirely within the autoencoder's output space, isolating the quality of interpolation in the compressed latent spaces from the reconstruction upperbound imposed by the compression model . The difficulty of this task is evaluated based on the *prediction gap*, defined as the minimum temporal distance between the target scan and its two neighbors: $\min(\text{target} - \text{start}, \text{end} - \text{target})$. A smaller prediction gap implies a target temporally close to either neighbor (easier), while a larger gap makes accurate interpolation more challenging. We report the results over approximately 1400 triplet configurations across 30 test subjects. As shown in Figure 4.8a and Figure 4.8b, the LDAE maintains strong generation performance even for wider prediction gaps. Notably, SSIM remains above 0.93 and MSE below 0.004 even at larger temporal gaps (up to 24 months), highlighting the semantic smoothness and temporal awareness encoded in the learned representations. In addition, Figure 4.10 provides qualitative examples of interpolated scans across different temporal gaps. The generated images appear visually coherent and anatomically plausible, preserving the subject identity and global brain structure. These results further support the hypothesis that LDAE captures a semantically meaningful representation and potentially a temporal progression trend.

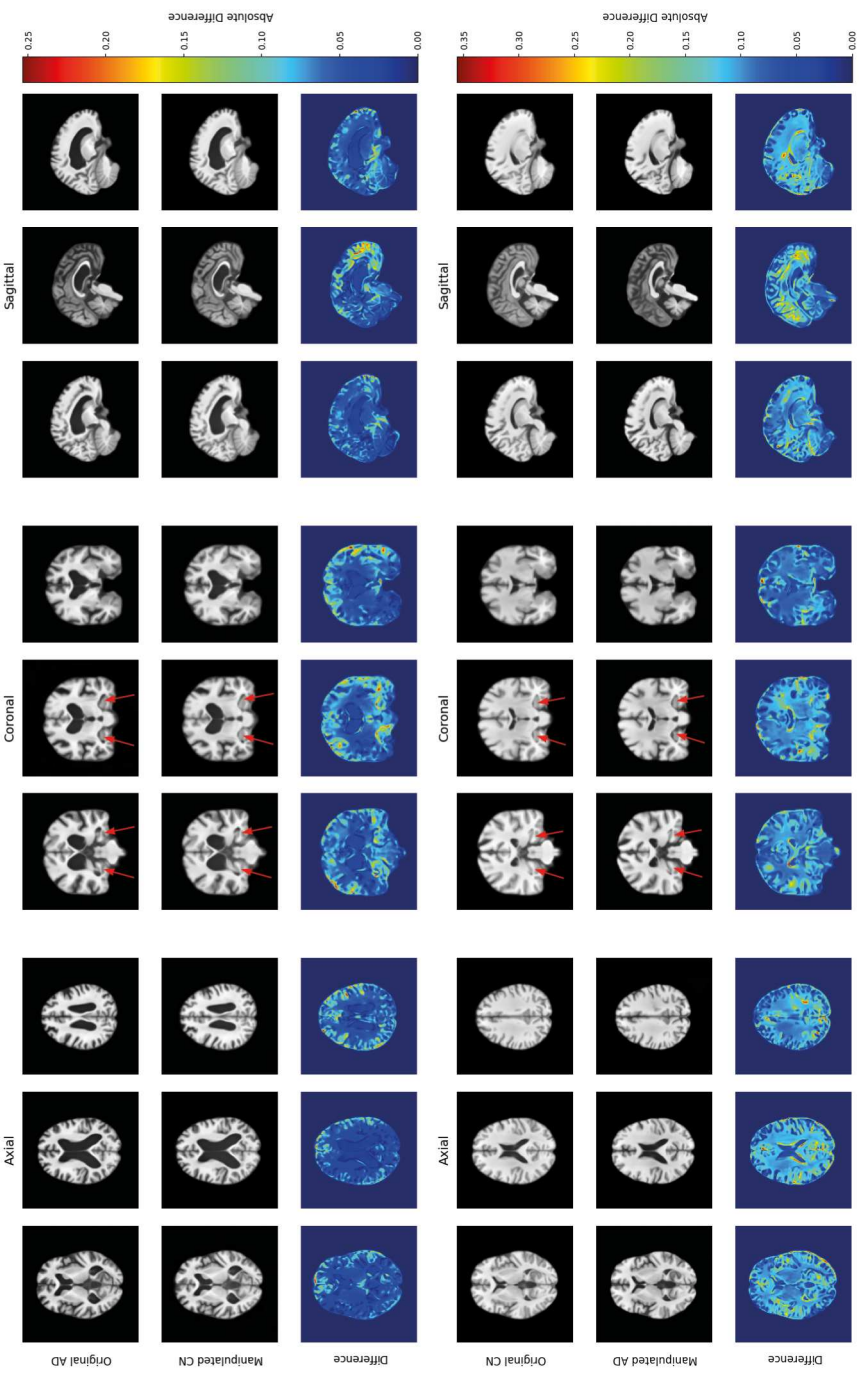
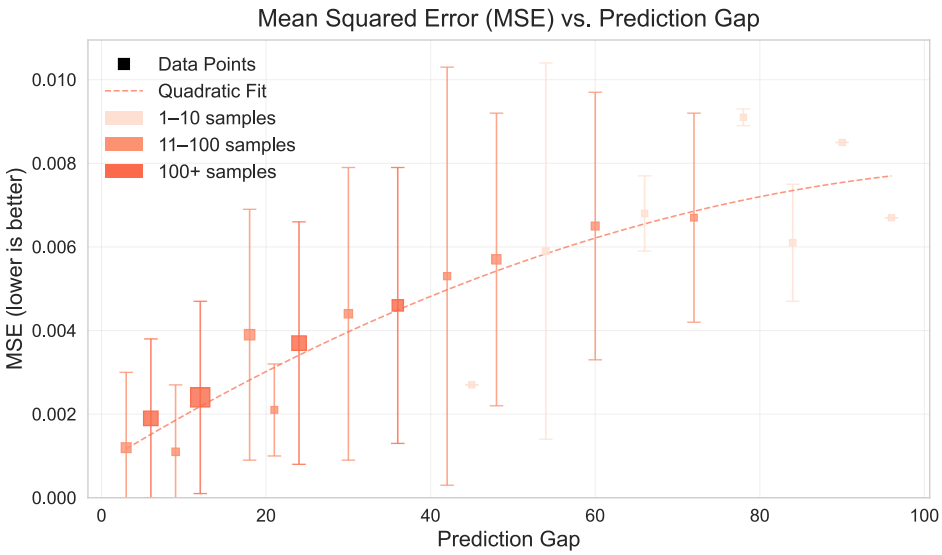
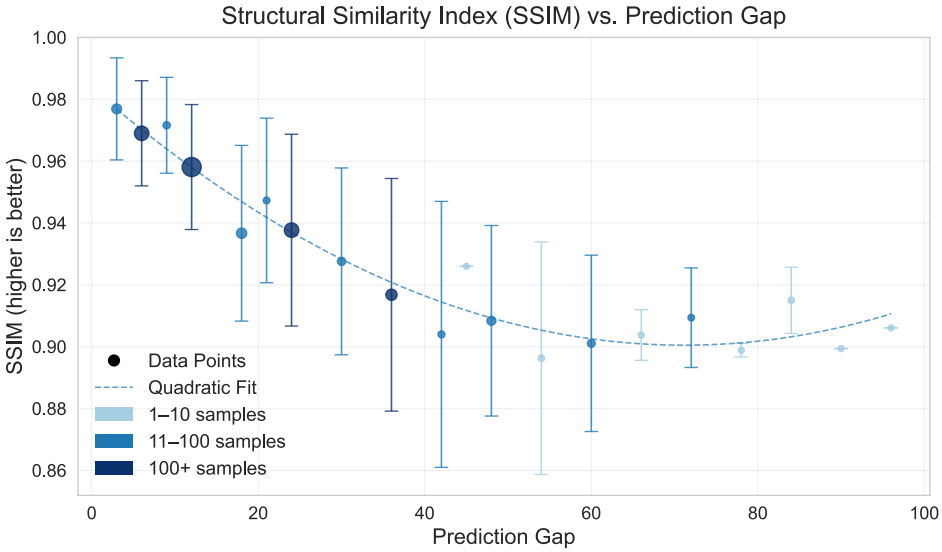


Figure 4.7: Bidirectional semantic manipulation of AD-related neuroanatomical features. The model selectively modifies disease-related characteristics while preserving subject identity and age-related features.

Interpolation in Stable vs. Progressive MCI: To examine the model's sensitivity to clinically meaningful disease progression, we conducted a targeted analysis on subjects with distinct clinical trajectories. We performed the interpolation experiment on two subgroups from our test set: sMCI, defined as subjects diagnosed with MCI at baseline who remained MCI at the 36-month follow-up, and pMCI, defined as subjects who were MCI at baseline and converted to AD by the 36-month follow-up. This latter group represents a more challenging scenario due to the significant and accelerated brain atrophy associated with disease conversion. As shown in Figure 4.9a and Figure 4.9b, the model's performance remains robust even in this demanding context. Notably, for pMCI subjects, the average SSIM remains high (close to 0.91) even when interpolating over a 36-month gap. This is meaningful because the interpolation factor is derived from the actual acquisition times of the two observed visits, and interpolation is performed directly in the latent codes rather than in image space. If the latent space did not encode a temporally coherent disease manifold, a simple time-aligned interpolation would not be able to reconstruct held-out intermediate scans with such fidelity, especially in subjects undergoing rapid and spatially heterogeneous atrophy. The qualitative example in Figure 4.10 illustrates precisely this behaviour: the interpolated volumes between two real visits follow the same anatomical progression trend observed in the true longitudinal scans, reproducing regionally specific changes such as ventricular enlargement and hippocampal shrinkage. This result strongly suggests that the interpolation within the latent space is capturing clinically meaningful temporal continuity between two disease states.

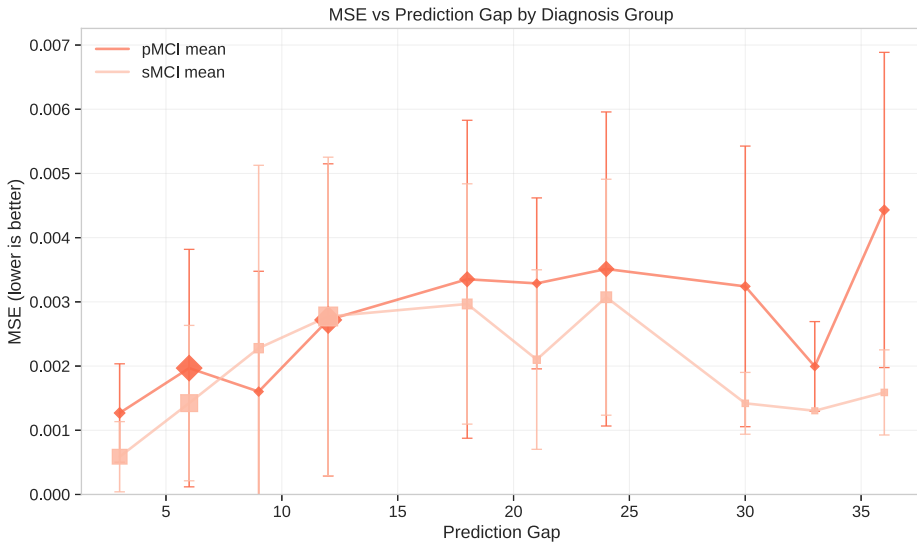


(a) Mean Squared Error (MSE) as a function of prediction gap. Larger gaps correspond to more difficult interpolation tasks.

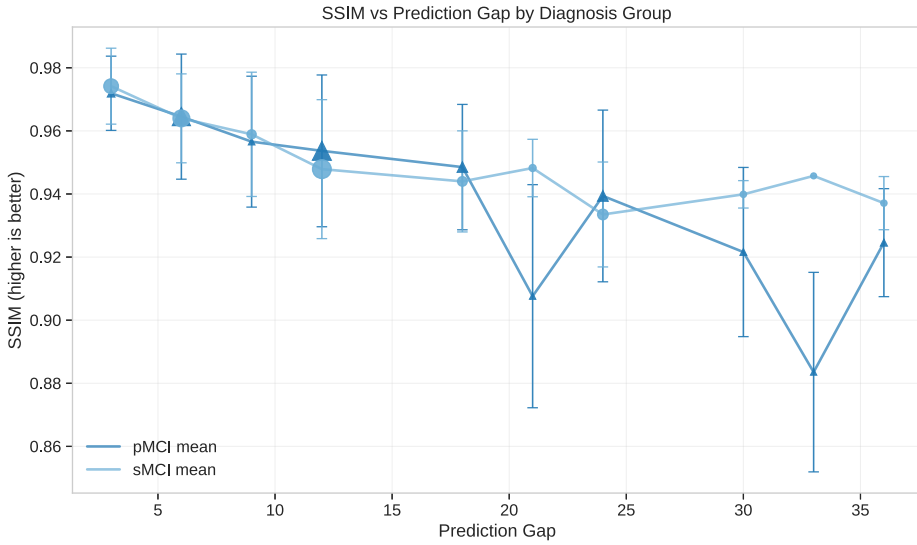


(b) Structural Similarity Index (SSIM) as a function of prediction gap. The model shows robust perceptual consistency across all temporal ranges.

Figure 4.8: Interpolation metrics across prediction gaps.



(a) Mean Squared Error (MSE) as a function of prediction gap for sMCI and pMCI subjects. Larger gaps correspond to more difficult interpolation tasks.



(b) Structural Similarity Index (SSIM) as a function of prediction gap for sMCI and pMCI subjects. The model shows robust perceptual consistency across all temporal ranges.

Figure 4.9: Interpolation metrics by prediction gap for sMCI and pMCI subjects.

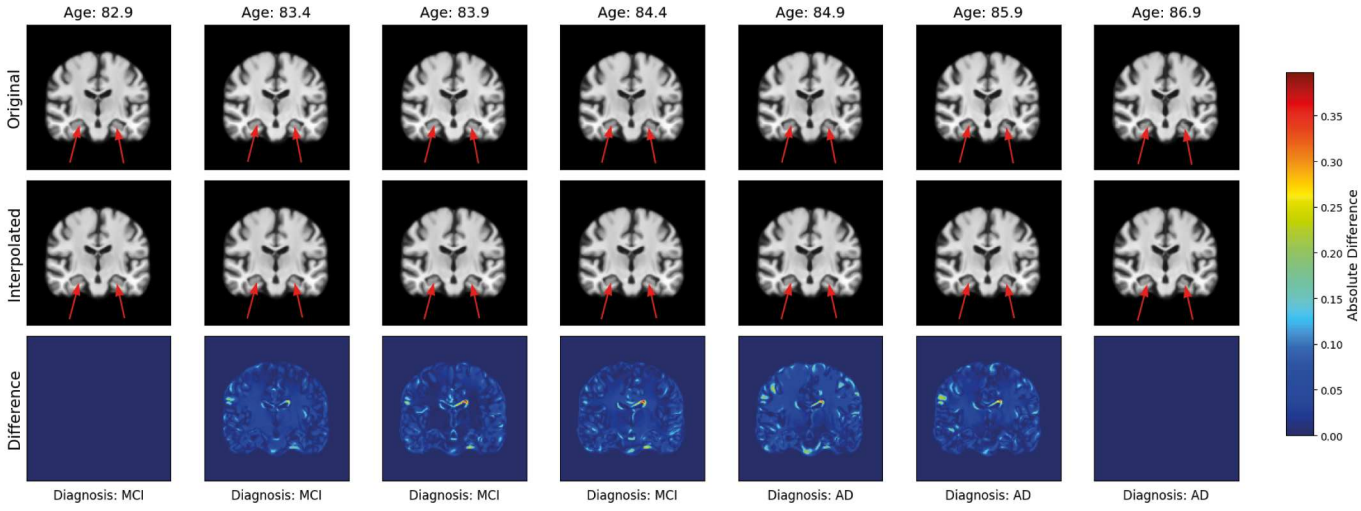


Figure 4.10: Qualitative example of latent interpolation for missing scan generation on a single subject diagnosed as pMCI with seven longitudinal scans acquired at 0, 6, 12, 18, 24, 36, and 48 months. The red arrows highlight the hippocampus region, where progressive atrophy is clearly visible, evolving from MCI to AD. This progression is consistently captured by our model in the latent space, with intermediate scans (e.g., 6 months, $\alpha = 0.125$, and 24 months, $\alpha = 0.5$) synthesized via linear interpolation in the semantic space and spherical interpolation in the stochastic space.

4.6 Discussion and conclusion

In this study, we introduced LDAE, a novel diffusion-based architecture specifically tailored for efficient and meaningful unsupervised representation learning, with a focused application on brain MRI scans related to AD.

Semantic Representation Learning The results indicate that LDAE captures semantic representations associated with structural brain variation related to AD and aging, consistent with the broader potential of diffusion-based representation learning [124]. The linear-probe evaluations quantitatively support this interpretation. Although the generative results were qualitatively reviewed by a clinical expert, broader clinical validation remains necessary, in line with recent literature emphasizing the need for interpretable and clinically grounded evaluation of AI systems in dementia imaging [103, 183]. Moreover, systematic multi-attribute disentanglement analyses would be important to ensure robust and interpretable attribute manipulation, given potential correlations between disease-related and age-related factors.

Generative Capabilities Reconstruction quality is inherently limited by the perceptual compression autoencoder's fidelity. It will be crucial to achieve a lossless reconstruction on a full resolution brain scan in order to use LDAE for missing scan generation and manipulation in clinical practice. Additionally, the current interpolation methods (LERP and SLERP) assume linearity in brain evolution trajectories. Incorporating advanced regression or forecasting models trained explicitly within the semantic space could better approximate realistic brain aging progressions.

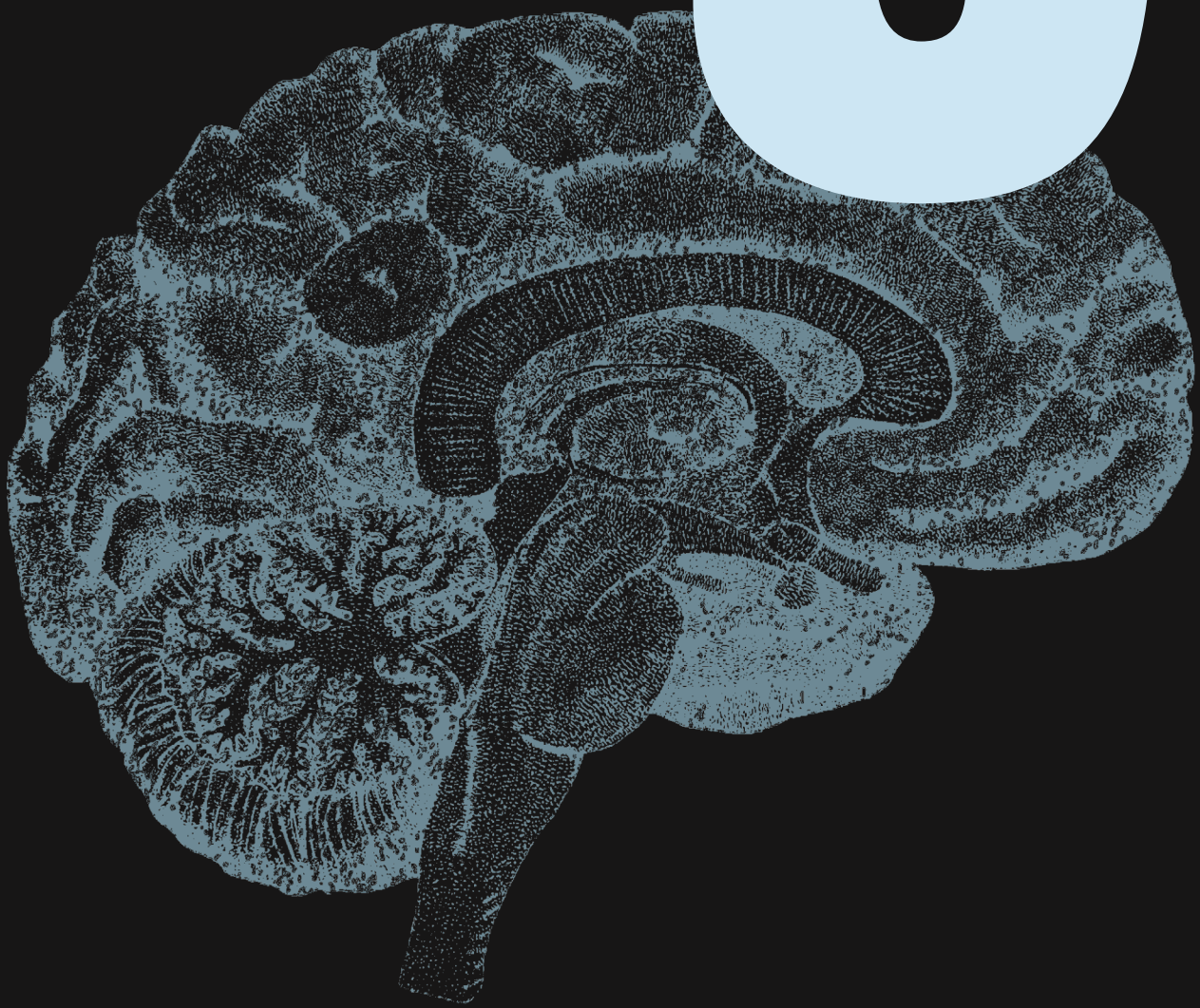
Potential Clinical Impact In longitudinal studies, LDAEs could be used to retrospectively capture and model subtle temporal changes of anatomical structures on medical imaging; specifically, the model can generate plausible intermediate scans between two observed timepoints (see Fig.4.10). This retrospective interpolation can assist in imputing missing visits or constructing denser synthetic trajectories for research purposes, which is relevant when evaluating the progression of cerebral atrophy in neurodegenerative diseases (e.g. Alzheimer or Parkinsonian disorders). Furthermore, it has potential clinical implications: by creating denser temporal trajectories, clinicians may achieve earlier and more accurate disease staging, provide more personalized prognostic assessments, and determine treatment timing more effectively. In clinical trials, generating intermediate scans can help preserve statistical power in smaller datasets, mitigate bias arising

from patient dropout, and uncover nuanced progression patterns that sparse sampling would otherwise obscure [92, 114]. Additionally, by generating realistic synthetic data that preserves the underlying distribution of the original dataset, LDAEs can help train ML models more effectively. This is particularly beneficial in rare disease studies, where obtaining large datasets is often infeasible. Another significant application lies in enhancing explainability through semantic latent manipulation. By isolating and controlling specific features within the latent space, researchers and clinicians can better understand the relationships between imaging patterns and disease characteristics. Moreover, the interpolation experiments suggest that the latent space learned by LDAE is semantically continuous in the temporal direction, which may enable future work to build explicit progression models on top of this representation.

General limitations and Future Work The success of LDAE frameworks in medical imaging depends on rigorous validations across diverse populations, imaging scanners, and disease phenotypes. Variability in imaging protocols, demographic factors, and disease presentations can significantly impact model performance. Therefore, comprehensive cross-validation is essential to ensure the generalizability and robustness of these frameworks in real-world clinical settings.

Foundation Model capability These findings position LDAE as a promising framework for scalable and interpretable medical imaging applications and as a potential Foundation Model for 3D medical image analysis. Future work should explore its transferability to other tasks and modalities, investigate domain adaptation strategies, and benchmark its performance in diverse clinical settings to validate its generalization capacity and pretraining utility.

5



CONCLUSIONS

Chapter 5

Conclusions

This thesis addressed three key challenges that are widely recognized as limiting the clinical impact of AI-based systems for AD analysis: the dependence on task-specific labeled data, the lack of model interpretability, and the difficulty of learning robust representations from high-dimensional neuroimaging data under realistic data and validation constraints [171, 103, 183]. The work presented across Chapters 2–4 follows a progression from supervised multi-task classification to interpretable diagnosis and, finally, to unsupervised foundation model learning. Each chapter builds on the previous one, and together they outline a path toward AI systems that are not only accurate but also interpretable, efficient, and broadly applicable to neurological disease analysis.

Chapter 2 explored handwriting as a complementary biomarker for AD screening. As discussed in Chapter 1, AD affects fine-motor control and handwriting characteristics, and tasks such as writing, drawing, and copying engage multiple cognitive domains as visuospatial organization, executive planning, semantic memory, and motor coordination, that are progressively impaired by the disease. The main contribution of this chapter was demonstrating that AD diagnosis through handwriting analysis is feasible with high accuracy, reaching 87.99% when training a unified model across all 25 handwriting tasks simultaneously. This result represents a significant advancement over prior work, where each task was evaluated independently with dedicated models, and suggests that shared AD-related patterns such as motor control degradation, spatial disorganization, or cognitive load effects can be effectively captured when tasks are analyzed jointly rather than in isolation. The diagnostic performance achieved with ConvNextSmall outperformed both classical CNNs and Vision Transformers. The chapter also compared two transfer learning strategies: pre-training on ImageNet (a general classification task) versus pre-

training on handwritten text recognition (OCR via TrOCR). Interestingly, ImageNet pretraining proved more effective. This can be explained by the fact that TrOCR models are designed to extract and interpret text content from images, whereas the features relevant for AD detection are related to neuromotor characteristics such as stroke regularity, pen pressure variations, and spatial organization, which are better captured by classification-oriented representations. The analysis of task-specific performance further showed that tasks requiring higher cognitive effort and motor coordination were consistently more informative for diagnosis, independently of the architecture used. This finding suggests the potential for protocol refinement, focusing on the most discriminative tasks to enhance clinical efficiency. At the same time, handwriting should not be interpreted as a disease-specific marker in isolation, particularly in older adults, because reduced manual dexterity, arthritis, tremor, and other age-related motor comorbidities may also influence task execution independently of cognitive decline. Moreover, these results establish a foundation for future longitudinal studies investigating whether handwriting-based analysis could serve prognostic purposes, such as predicting conversion from MCI to AD dementia.

Chapter 3 moved to structural MRI, which has become central to AI research in AD diagnosis and prognosis due to its favorable balance between the richness of morphological information it provides and the relatively non-invasive nature of acquisition. As discussed in Chapter 1, several prior approaches reported accuracies exceeding 90% on ADNI or its subsets, yet many of these studies lacked methodological rigor in evaluation protocols, making fair comparison across methods difficult. The main contribution of this chapter was the comprehensive investigation conducted using a standardized ADNI collection explicitly designed to address reproducibility issues and serve as a benchmark for model comparison. By adopting this collection along with strict subject-level data splitting, reproducible preprocessing pipelines, and open-source implementation, the chapter established a rigorous evaluation framework that directly addresses the reproducibility concerns highlighted in recent reviews.

The AXIAL framework proposed in this chapter combines the computational efficiency of 2D slice-based processing with the spatial context preservation of 3D approaches. AXIAL uses pre-trained 2D CNNs to extract features from individual slices across three orthogonal planes (axial, coronal, and sagittal), then employs lightweight attention modules to weight the contribution of each slice and fuse the information into a unified 3D representation. A key finding was that the attention-based explainability mechanism demonstrated remarkable consistency: the model consistently focused on the hippocampus, parahippocampal gyrus, amygdala, and inferior lateral ventricles—regions well documented in literature as being affected early and severely in AD—across different experimental runs and different cross-validation folds. This consistency is in contrast to other

attention-based approaches which often produce diffuse and inconsistent saliency maps. The quantitative validation of these attention maps against atlas-defined regions, based on overlap volumes and intensities, moves beyond the qualitative visual inspection that characterizes most XAI studies in the field. Beyond explainability, the proposed approach achieved superior diagnostic performance compared to the tested alternatives, particularly on the challenging sMCI vs. pMCI task. A double transfer learning strategy, which first trains on the AD vs. CN task (where more data and clearer patterns are available) before fine-tuning on the prognostic task, significantly improved the model's sensitivity to subtle morphological changes, bringing the MCC from near-random levels to 44.3%.

Chapter 4 represented the final step in the thesis progression, moving from supervised approaches that require diagnostic labels to unsupervised representation learning. The motivation behind this shift is straightforward: clinical repositories contain millions of brain MRI scans acquired for various purposes beyond AD diagnosis, but these scans typically lack diagnostic labels and are therefore inaccessible to supervised methods. Unsupervised approaches can learn representations from the underlying structure of brain imaging data itself, and then adapt to specific downstream tasks with minimal labeled examples. However, applying diffusion-based representation learning to 3D MRI data has been computationally infeasible, as diffusion autoencoders require iterative denoising processes that are too expensive when applied to full brain volumes. The LDAE framework introduced in this chapter solved this problem through a three-stage training strategy: first, a perceptual autoencoder compresses brain MRI volumes into a much smaller latent space (170-fold volume reduction); then, an unconditional diffusion model is pretrained on these compressed representations; and finally, a semantic encoder coupled with a gradient estimator learns compact, meaningful embeddings from unlabeled data. This approach achieved a 20-times speedup in inference time compared to voxel-space diffusion autoencoders while improving reconstruction quality.

The learned representations were shown to encode clinically meaningful information, even though no diagnostic labels were used during training. Linear probes trained on the frozen semantic embeddings achieved strong performance in AD classification and age prediction, confirming that the latent space captures morphological changes related to neurodegeneration and aging. The zero-shot transfer experiments on AIBL and OASIS-3 showed that these representations generalize to unseen datasets acquired under different conditions and in different populations, without any fine-tuning. Notably, a simple linear classifier on LDAE embeddings outperformed a fully supervised encoder on external datasets in terms of F1-score and MCC, suggesting that unsupervised pre-training on diverse data produces representations that are more robust to distributional differences than features learned through supervised training on a single cohort. Beyond classification, LDAE's generative capabilities enabled semantic manipulation of brain scans—

transforming an AD brain toward a cognitively normal appearance, or vice versa—while preserving subject identity and age-related characteristics. These manipulations consistently affected the hippocampal and ventricular regions most associated with AD, providing an additional form of interpretability grounded in the generative process. The interpolation experiments demonstrated that the latent space captures temporally coherent progression patterns, generating plausible intermediate brain states between longitudinal visits with high fidelity even at temporal gaps of 24 months. The analysis of progressive versus stable MCI subgroups further showed that the model captures the accelerated atrophy pattern characteristic of subjects converting to dementia, indicating that the learned representation encodes a temporally aware disease progression.

Looking across the three chapters, several common themes emerge that connect the individual contributions into a coherent body of work. The first is the central role of transfer learning in addressing data scarcity, which is a persistent challenge in medical imaging. In Chapter 2, ImageNet pre-training enabled effective feature extraction from a handwriting dataset of only 174 subjects. In Chapter 3, the same strategy supported slice-level feature extraction from a small standardized MRI collection, and double transfer learning extended this to the data-scarce prognostic task. In Chapter 4, transfer learning operated at multiple levels: ImageNet pre-training in the semantic encoder, UK Biobank pre-training in the compression autoencoder, and unsupervised pre-training on ADNI enabling zero-shot transfer to external cohorts. This progressive reliance on increasingly sophisticated transfer strategies reflects a broader insight: when labeled data are scarce, the ability to reuse knowledge across tasks and datasets is essential for achieving robust generalization.

The second common theme is a consistent preference for 2D or 2.5D processing over full 3D computation. This was not a compromise but a deliberate design choice motivated by the availability of pre-trained weights and the limited size of medical imaging datasets. Chapter 2 processed handwriting as standard 2D images. Chapter 3 showed that 2D slice-based processing with lightweight attention mechanisms could match or outperform more complex 3D and transformer-based alternatives on small datasets. Chapter 4 adopted a 2.5D semantic encoder that processes axial slices through a shared 2D backbone before aggregating them with attention—a design validated through ablation studies showing that a full 3D CNN encoder failed to converge stably. This finding has practical implications: well-designed 2D or 2.5D architectures that leverage pre-trained weights can outperform more complex 3D alternatives in data-limited medical settings.

A third theme is the increasing commitment to methodological rigor across chapters. Chapter 2 implemented subject-level data splitting to prevent the data leakage that has inflated reported performance in prior handwriting analysis work. Chapter 3 adopted a

standardized dataset, reproducible preprocessing pipelines, open-source code, and quantitative XAI validation against neuroanatomical atlases. Chapter 4 extended the evaluation to zero-shot cross-dataset transfer, which is the most demanding test of generalizability, and included ablation studies to validate key architectural decisions. This progression directly responds to the reproducibility concerns raised in recent reviews of AI for AD, where many reported results appear to reflect dataset-specific overfitting rather than genuine disease-related signal extraction.

Attention mechanisms also played an important role across Chapters 3 and 4, serving a dual purpose. On one hand, they improved model performance by adaptively weighting the most informative features or slices. On the other hand, they provided interpretable explanations of the model's decisions. In Chapter 3, the attention distributions consistently converged on the brain's inferior regions where the medial temporal lobe structures reside, supporting the clinical interpretation of the results. In Chapter 4, the two-stage attention mechanism in the semantic encoder identified the most informative slices for encoding brain morphology, while the semantic latent space enabled a complementary form of interpretability through controllable generation. This dual utility of attention—both as a computational tool and as an interpretability mechanism—makes attention-based architectures particularly suitable for medical imaging, where transparency is not optional but a prerequisite for clinical deployment.

Despite the contributions of this thesis, several limitations should be acknowledged, as they point to important directions for future work. The datasets used, while carefully selected and rigorously processed, remain relatively small. Chapter 2 relied on 174 subjects, Chapter 3 on a standardized subset of 639 subjects, and even the full ADNI cohort used in Chapter 4 is much smaller than the datasets used to train foundation models in other domains. Scaling to millions of unlabeled brain scans from diverse clinical repositories, while addressing challenges such as scanner variability, protocol differences, and demographic diversity, is a natural next step. The temporal modeling in Chapter 4, based on linear and spherical interpolation in the latent space, assumes smooth trajectories of brain change. While the results showed good fidelity even at long temporal gaps, incorporating explicit non-linear progression models could better capture the complex dynamics of neurodegeneration. Similarly, the three separate plane-specific models in Chapter 3 could be integrated into a unified multi-view architecture that captures cross-plane correlations during training rather than combining them afterward.

An important open question concerns clinical validation. While the quantitative XAI analysis in Chapter 3 and the expert-reviewed manipulations in Chapter 4 provide encouraging evidence of clinical alignment, none of the proposed systems has been tested prospectively in clinical workflows. Bridging this gap will require close collaboration

between AI researchers and clinical practitioners, addressing not only performance evaluation but also interface design, uncertainty communication, and integration with electronic health records. Furthermore, the potential of combining handwriting-based screening with neuroimaging-based diagnosis and foundation model-derived representations has not been explored in this thesis, yet such a multi-modal approach could enable a more comprehensive assessment spanning behavioral biomarkers, structural brain changes, and learned disease representations.

From a translational perspective, moving these methods toward clinical or regulatory validation would require substantially more than retrospective accuracy estimates. Future studies should assess robustness across sites, scanners, acquisition protocols, and populations, while also documenting calibration, failure modes, and reproducibility under prospectively defined evaluation criteria. In parallel, the generative components discussed in Chapter 4 raise ethical considerations that are particularly relevant in medical imaging. Although controllable generation and counterfactual manipulation can support interpretability and hypothesis exploration, they also require careful governance to avoid overinterpretation of synthetic outputs, unintended amplification of dataset biases, or misleading clinical use outside validated settings. For this reason, methodological advances in generative modeling should be accompanied by transparent reporting, expert oversight, and clear boundaries between exploratory research use and clinically actionable evidence.

In summary, the methodologies developed in this thesis contribute to a broader shift in medical AI: moving from narrow, task-specific tools toward general-purpose systems that are interpretable and can learn from large quantities of unlabeled clinical data. The handwriting framework in Chapter 2 demonstrates that accessible, non-invasive biomarkers can complement neuroimaging for AD screening. The AXIAL framework in Chapter 3 shows that diagnostic accuracy does not have to come at the cost of interpretability, and that quantitative validation against clinical knowledge can help build the trust necessary for adoption. The LDAE framework in Chapter 4 demonstrates that unsupervised diffusion-based learning can extract clinically meaningful representations from 3D brain MRI at scale, enabling zero-shot transfer to unseen datasets, counterfactual explanations, and temporal modeling within a single architecture. Together, these contributions address some of the fundamental barriers—data scarcity, lack of transparency, computational cost, and limited generalizability—that have so far prevented AI systems from achieving meaningful clinical impact in AD analysis. The path forward will require scaling these approaches to larger and more diverse datasets, integrating complementary biomarker modalities, and, most importantly, validating them in prospective clinical settings where the ultimate measure of success is improved patient outcomes.

Bibliography

[1] Md Minul Alam and Shahram Latifi. "Early Detection of Alzheimer's Disease Using Generative Models: A Review of GANs and Diffusion Models in Medical Imaging". In: *Algorithms* 18.7 (2025), p. 434.

[2] Marilyn S Albert et al. "The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease". In: *Alzheimer's & dementia* 7.3 (2011), pp. 270–279.

[3] Fatih Altay et al. "Preclinical Stage Alzheimer's Disease Detection Using Magnetic Resonance Image Scans". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 17. 2021, pp. 15088–15097.

[4] Alzheimer's Association. "2025 Alzheimer's disease facts and figures". In: *Alzheimer's & Dementia* 21.4 (2025), e70235. DOI: https://doi.org/10.1002/alz.70235.

[5] Brian B Avants et al. "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain". In: *Medical image analysis* 12.1 (2008), pp. 26–41.

[6] Brian B Avants et al. "The Insight ToolKit image registration framework". In: *Frontiers in neuroinformatics* 8 (2014), p. 44.

[7] Silvia Basaia et al. "Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks". In: *NeuroImage: Clinical* 21 (2019), p. 101645.

[8] Lorena Bresciani et al. "Visual assessment of medial temporal atrophy on MR films in Alzheimer's disease: comparison with volumetry". In: *Aging Clinical and Experimental Research* 17.1 (2005), pp. 8–13. DOI: 10.1007/BF03337714.

[9] Randy L Buckner et al. "Molecular, structural, and functional characterization of Alzheimer's disease: evidence for a relationship between default activity, amyloid, and memory". In: *Journal of neuroscience* 25.34 (2005), pp. 7709–7717.

[10] Pierluigi Carcagni et al. “Convolution Neural Networks and Self-Attention Learners for Alzheimer Dementia Diagnosis from Brain MRI”. In: *Sensors* 23.3 (2023), p. 1694.

[11] M Jorge Cardoso et al. “Monai: An open-source framework for deep learning in healthcare”. In: *arXiv preprint arXiv:2211.02701* (2022).

[12] Aditya Chattopadhay et al. “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks”. In: *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2018, pp. 839–847.

[13] Qi Chen et al. “Towards generalizable tumor synthesis”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 11147–11158.

[14] Davide Chicco and Giuseppe Jurman. “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation”. In: *BMC genomics* 21 (2020), pp. 1–13.

[15] N. D. Cilia et al. “Using Genetic Algorithms for the Prediction of Cognitive Impairments”. In: *Lecture notes in computer science*. Vol. 12104. Springer, 2020, pp. 479–493.

[16] N.D. Cilia et al. “An Experimental Protocol to Support Cognitive Impairment Diagnosis by using Handwriting Analysis”. In: *Procedia Computer Science* 141 (2018), pp. 466–471.

[17] N.D. Cilia et al. “Using handwriting features to characterize cognitive impairment”. In: Springer, 2019.

[18] Nicole D Cilia et al. “Deep transfer learning algorithms applied to synthetic drawing images as a tool for supporting Alzheimer’s disease prediction”. In: *Machine Vision and Applications* 33.3 (2022), p. 49.

[19] Nicole D. Cilia et al. “From Online Handwriting to Synthetic Images for Alzheimer’s Disease Detection Using a Deep Transfer Learning Approach”. In: *IEEE Journal of Biomedical and Health Informatics* 25.12 (2021), pp. 4243–4254. DOI: 10.1109/JBHI.2021.3101982.

[20] Nicole Dalia Cilia et al. “Handwriting Analysis to Support Alzheimer’s Disease Diagnosis: A Preliminary Study”. In: *Computer Analysis of Images and Patterns*. Ed. by Mario Vento and Gennaro Percannella. Cham: Springer International Publishing, 2019, pp. 143–151.

[21] Jules J. Claus et al. “Practical use of visual medial temporal lobe atrophy cut-off scores in Alzheimer’s disease: Validation in a large memory clinic population”. In: *European Radiology* 27.8 (2017), pp. 3147–3155. DOI: 10.1007/s00330-016-4726-3.

[22] Nancy Cloak and Yasir Al Khalili. “Behavioral and psychological symptoms in dementia”. In: (2019).

[23] Mengzhao Cui et al. “Grip strength and the risk of cognitive decline and dementia: a systematic review and meta-analysis of longitudinal cohort studies”. In: *Frontiers in aging neuroscience* 13 (2021), p. 625551.

[24] Nicholas C Cullen and Brian B Avants. “Convolutional neural networks for rapid and simultaneous brain extraction and tissue segmentation”. In: *Brain Morphometry* (2018), pp. 13–34.

[25] Nicole Dalia Cilia et al. “Offline handwriting image analysis to predict Alzheimer’s disease via deep learning”. In: *2022 26th International Conference on Pattern Recognition (ICPR)*. 2022, pp. 2807–2813. DOI: 10.1109/ICPR56361.2022.9956359.

[26] C. De Stefano et al. “Handwriting analysis to support neurodegenerative diseases diagnosis: a review”. In: *Pattern Recognition Letters* 121 (2018), pp. 37–45.

[27] Claudio De Stefano et al. “Handwriting analysis to support neurodegenerative diseases diagnosis: A review”. In: *Pattern Recognition Letters* 121 (2019), pp. 37–45.

[28] Charles DeCarli et al. “Qualitative estimates of medial temporal atrophy as a predictor of progression from mild cognitive impairment to dementia”. In: *Archives of Neurology* 64.1 (2007), pp. 108–115.

[29] Margarete Delazer, Laura Zamarian, and Atbin Djamshidian. “Handwriting in Alzheimer’s disease”. In: *Journal of Alzheimer’s Disease* 82.2 (2021), pp. 727–735.

[30] J. Deng et al. “ImageNet: A large-scale hierarchical image database.” In: *CVPR*. IEEE Computer Society, 2009, pp. 248–255.

[31] Prafulla Dhariwal and Alexander Nichol. “Diffusion models beat gans on image synthesis”. In: *Advances in neural information processing systems* 34 (2021), pp. 8780–8794.

[32] Prafulla Dhariwal and Alexander Nichol. “Diffusion models beat gans on image synthesis”. In: *Advances in neural information processing systems* 34 (2021), pp. 8780–8794.

[33] Helena Dolphin et al. “An update on apathy in Alzheimer’s disease”. In: *Geriatrics* 8.4 (2023), p. 75.

[34] Alexey Dosovitskiy and Thomas Brox. “Generating images with perceptual similarity metrics based on deep networks”. In: *Advances in neural information processing systems* 29 (2016).

[35] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV].

[36] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[37] Peter Drotár et al. "Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease". In: *Artificial Intelligence in Medicine* 67 (2016), pp. 39–46. ISSN: 0933-3657. DOI: https://doi.org/10.1016/j.artmed.2016.01.004.

[38] Mr Amir Ebrahimighahnavieh, Suhuai Luo, and Raymond Chiong. "Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review". In: *Computer methods and programs in biomedicine* 187 (2020), p. 105242.

[39] FIFTH EDITION. "Diagnostic and statistical manual of mental disorders". In: *American psychiatric association, Washington, DC* (1980), pp. 205–224.

[40] Kathryn A Ellis et al. "The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease". In: *International psychogeriatrics* 21.4 (2009), pp. 672–687.

[41] Patrick Esser, Robin Rombach, and Bjorn Ommer. "Taming transformers for high-resolution image synthesis". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 12873–12883.

[42] Farshad Falahati, Eric Westman, and Andrew Simmons. "Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging". In: *Journal of Alzheimer's disease* 41.3 (2014), pp. 685–708.

[43] Sarah Tomaszewski Farias et al. "Progression of mild cognitive impairment to dementia in clinic-vs community-based cohorts". In: *Archives of neurology* 66.9 (2009), pp. 1151–1157.

[44] Howard H Feldman et al. "Diagnosis and treatment of dementia: 2. Diagnosis". In: *Cmaj* 178.7 (2008), pp. 825–836.

[45] Xinyang Feng et al. "A deep learning MRI approach outperforms other biomarkers of prodromal Alzheimer's disease". In: *Alzheimer's Research & Therapy* 14.1 (2022), p. 45.

[46] Luca Ferrarini et al. "Shape differences of the brain ventricles in Alzheimer's disease". In: *Neuroimage* 32.3 (2006), pp. 1060–1069.

[47] Bruce Fischl. "FreeSurfer". In: *Neuroimage* 62.2 (2012), pp. 774–781.

[48] M. F. Folstein, S. E. Folstein, and P. R. McHugh. "'Mini-mental state': A practical method for grading the cognitive state of patients for the clinician". In: *J. Psychiatric Res.* 12.3 (1975), pp. 189–198.

[49] Vladimir Fonov et al. "Unbiased average age-appropriate atlases for pediatric studies". In: *Neuroimage* 54.1 (2011), pp. 313–327.

[50] Vladimir S Fonov et al. "Unbiased nonlinear average age-appropriate brain templates from birth to adulthood". In: *NeuroImage* 47 (2009), S102.

[51] Peter A Freeborough and Nick C Fox. "The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI". In: *IEEE transactions on medical imaging* 16.5 (2002), pp. 623–629.

[52] Morris Freedman. *Clock drawing: A neuropsychological analysis*. Oxford University Press, 1994.

[53] Giovanni B Frisoni et al. "The clinical use of structural MRI in Alzheimer disease". In: *Nature Reviews Neurology* 6.2 (2010), pp. 67–77.

[54] Helia Givian, Jean-Paul Calbimonte, and for the Alzheimer's Disease Neuroimaging Initiative. "Early diagnosis of Alzheimer's disease and mild cognitive impairment using MRI analysis and machine learning algorithms". In: *Discover Applied Sciences* 7.1 (2024), p. 27.

[55] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[56] Krzysztof J Gorgolewski et al. "The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments". In: *Scientific data* 3.1 (2016), pp. 1–9.

[57] Krzysztof J Gorgolewski et al. "The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments". In: *Scientific data* 3.1 (2016), pp. 1–9.

[58] Angela Guarino et al. "Executive functions in Alzheimer disease: A systematic review". In: *Frontiers in aging neuroscience* 10 (2019), p. 437.

[59] Sven Haller, Karl O Lovblad, and Panteleimon Giannakopoulos. "Principles of classification analyses in mild cognitive impairment (MCI) and Alzheimer disease". In: *Journal of Alzheimer's Disease* 26.s3 (2011), pp. 389–394.

[60] Changhee Han et al. "MADGAN: Unsupervised medical anomaly detection GAN using multiple adjacent brain MRI slice reconstruction". In: *BMC bioinformatics* 22.Suppl 2 (2021), p. 31.

[61] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[62] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[63] Kaiming He et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[64] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[65] Zhentao Hu et al. "Conv-Swinformer: Integration of CNN and shift window attention for Alzheimer's disease classification". In: *Computers in Biology and Medicine* 164 (2023), p. 107304.

[66] Zhentao Hu et al. "VGG-TSwinformer: Transformer-based deep learning model for early Alzheimer's disease prediction". In: *Computer Methods and Programs in Biomedicine* 229 (2023), p. 107291.

[67] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[68] Drew A Hudson et al. "Soda: Bottleneck diffusion models for representation learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 23115–23127.

[69] A. Iavarone et al. "The frontal assessment battery (FAB): Normative data from an Italian sample and performances of patients with Alzheimer's disease and frontotemporal dementia". In: *Funct Neurol.* 19 (July 2004), pp. 191–195.

[70] Clifford R Jack et al. "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade". In: *The Lancet Neurology* 9.1 (2010), pp. 119–128.

[71] Clifford R Jack Jr et al. "MR-based hippocampal volumetry in the diagnosis of Alzheimer's disease". In: *Neurology* 42.1 (1992), pp. 183–183.

[72] Clifford R Jack Jr et al. "NIA-AA research framework: toward a biological definition of Alzheimer's disease". In: *Alzheimer's & dementia* 14.4 (2018), pp. 535–562.

[73] Clifford R Jack Jr et al. "Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's Association Workgroup". In: *Alzheimer's & Dementia* 20.8 (2024), pp. 5143–5169.

[74] Clifford R Jack Jr et al. "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods". In: *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* 27.4 (2008), pp. 685–691.

[75] Holger Jahn. "Memory loss in Alzheimer's disease". In: *Dialogues in clinical neuroscience* 15.4 (2013), pp. 445–454.

[76] Mark Jenkinson et al. "Fsl". In: *Neuroimage* 62.2 (2012), pp. 782–790.

[77] Saumya Jetley et al. "Learn to pay attention". In: *arXiv preprint arXiv:1804.02391* (2018).

[78] Dan Jin et al. "Attention-based 3D convolutional network for Alzheimer's disease diagnosis and biomarkers exploration". In: *2019 IEEE 16Th international symposium on biomedical imaging (ISBI 2019)*. IEEE. 2019, pp. 1047–1051.

[79] Taeho Jo, Kwangsik Nho, and Andrew J Saykin. "Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data". In: *Frontiers in aging neuroscience* 11 (2019), p. 220.

[80] Eric R. Kandel, J. H. Schwartz, and Thomas M. Jessell. *Principles of Neural Science*. 4th. McGraw-Hill Medical, July 2000.

[81] Wenjie Kang et al. "Multi-model and multi-slice ensemble learning architecture based on 2D convolutional neural networks for Alzheimer's disease diagnosis". In: *Computers in Biology and Medicine* 136 (2021), p. 104678.

[82] Diederik P Kingma. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[83] Marco Klaiber et al. "A systematic literature review on transfer learning for 3d-CNNs". In: *2021 international joint conference on neural networks (IJCNN)*. IEEE. 2021, pp. 1–10.

[84] Masatomo Kobayashi et al. "Automated Early Detection of Alzheimer's Disease by Capturing Impairments in Multiple Cognitive Domains with Multiple Drawing Tasks". In: *Journal of Alzheimer's Disease* 88 (June 2022), pp. 1–15. DOI: 10.3233/JAD-215714.

[85] Esther LGE Koedam et al. "Visual assessment of posterior atrophy development of a MRI rating scale". In: *European radiology* 21.12 (2011), pp. 2618–2625.

[86] Rafsanjany Kushol et al. "Addformer: Alzheimer's disease detection from structural mri using fusion transformer". In: *2022 IEEE 19th International Symposium On Biomedical Imaging (ISBI)*. IEEE. 2022, pp. 1–5.

[87] Renaud La Joie et al. "Region-specific hierarchy between atrophy, hypometabolism, and β-amyloid (Aβ) load in Alzheimer's disease dementia". In: *Journal of Neuroscience* 32.46 (2012), pp. 16265–16273.

[88] Jany Lambert et al. "Central and Peripheral Agraphia in Alzheimer's Disease: From the Case of Auguste D. to a Cognitive Neuropsychology Approach". In: *Cortex* 43.7 (2007), pp. 935–951. ISSN: 0010-9452.

[89] Pamela J LaMontagne et al. "OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease". In: *medrxiv* (2019), pp. 2019–12.

[90] Christian Ledig et al. "Structural brain imaging in Alzheimer's disease and mild cognitive impairment: biomarker analysis and shared morphometry database". In: *Scientific reports* 8.1 (2018), p. 11258.

[91] Antoine Leuzy et al. “Considerations in the clinical use of amyloid PET and CSF biomarkers for Alzheimer’s disease”. In: *Alzheimer’s & Dementia* 21.3 (2025), e14528.

[92] Chunyan Li et al. “Data-driven spatiotemporal modeling reveals personalized trajectories of cortical atrophy in Alzheimer’s disease”. In: *arXiv preprint arXiv:2511.08847* (2025).

[93] Minghao Li et al. *TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models*. 2022. arXiv: 2109.10282 [cs.CL].

[94] Z Li et al. “Global Burden of Dementia Death from 1990 to 2019, with Projections to 2050: An Analysis of 2019 Global Burden of Disease Study”. In: *The Journal of Prevention of Alzheimer’s Disease* (2024), pp. 1–9.

[95] Szu-Ying Lin et al. “The Clinical Course of Early and Late Mild Cognitive Impairment”. In: *Frontiers in Neurology* 13 (2022), p. 685636. DOI: 10.3389/fneur.2022.685636.

[96] Ze Liu et al. “Swin Transformer V2: Scaling Up Capacity and Resolution”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 11999–12009. DOI: 10.1109/CVPR52688.2022.01170.

[97] Zhuang Liu et al. “A ConvNet for the 2020s”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 11966–11976. DOI: 10.1109/CVPR52688.2022.01167.

[98] Flavia Loreto et al. “Visual atrophy rating scales and amyloid PET status in an Alzheimer’s disease clinical cohort”. In: *Annals of Clinical and Translational Neurology* 10.4 (2023), pp. 619–631. DOI: 10.1002/acn3.51749.

[99] Sara Lorio et al. “Neurobiological origin of spurious brain morphological changes: A quantitative MRI study”. In: *Human brain mapping* 37.5 (2016), pp. 1801–1815.

[100] Constantine G Lyketsos et al. *Neuropsychiatric symptoms in Alzheimer’s disease*. 2011.

[101] Daniel S Marcus et al. “Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults”. In: *Journal of cognitive neuroscience* 19.9 (2007), pp. 1498–1507.

[102] U.-V. Marti and H. Bunke. “The IAM-database: an English sentence database for offline handwriting recognition”. In: *International Journal on Document Analysis and Recognition* 5.1 (Nov. 2002), pp. 39–46. ISSN: 1433-2833. DOI: 10.1007/s100320200071.

[103] Sophie A Martin et al. “Interpretable machine learning for dementia: a systematic review”. In: *Alzheimer’s & Dementia* 19.5 (2023), pp. 2135–2149.

[104] Guy M McKhann et al. "The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease". In: *Alzheimer's & dementia* 7.3 (2011), pp. 263–269.

[105] Atif Mehmood et al. "A transfer learning approach for early diagnosis of Alzheimer's disease on MRI images". In: *Neuroscience* 460 (2021), pp. 43–52.

[106] Susanne G Mueller et al. "The Alzheimer's disease neuroimaging initiative". In: *Neuroimaging Clinics* 15.4 (2005), pp. 869–877.

[107] Emanuele Nardone et al. "Handwriting strokes as biomarkers for Alzheimer's disease prediction: A novel machine learning approach". In: *Computers in Biology and Medicine* 190 (2025), p. 110039.

[108] Z. S. Nasreddine et al. "The Montreal cognitive assessment, MoCA: A brief screening tool for mild cognitive impairment". In: *Journal of the American Geriatrics Society* 53.4 (Apr. 2005), pp. 695–699.

[109] J. Neils-Strunjas et al. "Dysgraphia in Alzheimer's disease: a review for clinical and research purposes". In: *J Speech Lang Hear Res* 49.6 (2006), pp. 1313–30.

[110] Sean M Nestor et al. "Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database". In: *Brain* 131.9 (2008), pp. 2443–2454.

[111] Alex Nichol et al. "Glide: Towards photorealistic image generation and editing with text-guided diffusion models". In: *arXiv preprint arXiv:2112.10741* (2021).

[112] Alexander Quinn Nichol and Prafulla Dhariwal. "Improved denoising diffusion probabilistic models". In: *International conference on machine learning*. PMLR. 2021, pp. 8162–8171.

[113] Emma Nichols et al. "Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019". In: *The Lancet Public Health* 7.2 (2022), e105–e125.

[114] Julie Ottoy et al. "Recent advances in neuroimaging of Alzheimer's disease and related dementias". In: *Alzheimer's & Dementia* 21.9 (2025), e70648.

[115] Changhyun Park, Wonsik Jung, and Heung-Il Suk. "Deep joint learning of pathological region localization and Alzheimer's disease diagnosis". In: *Scientific reports* 13.1 (2023), p. 11664.

[116] Adam Paszke et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32 (2019).

[117] Wei Peng et al. "Generating realistic brain mris via a conditional diffusion probabilistic model". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 14–24.

[118] Ronald C Petersen et al. "Mild cognitive impairment: ten years later". In: *Archives of Neurology* 66.12 (2009), pp. 1447–1455. DOI: 10.1001/archneurol.2009.266.

[119] Ronald Carl Petersen et al. "Alzheimer's disease Neuroimaging Initiative (ADNI) clinical characterization". In: *Neurology* 74.3 (2010), pp. 201–209.

[120] AF Pettersson, E Olsson, and L-O Wahlund. "Motor function in subjects with mild cognitive impairment and early Alzheimer's disease". In: *Dementia and geriatric cognitive disorders* 19.5-6 (2005), pp. 299–304.

[121] Walter HL Pinaya et al. "Brain imaging generation with latent diffusion models". In: *MICCAI Workshop on Deep Generative Models*. Springer. 2022, pp. 117–126.

[122] Walter HL Pinaya et al. "Generative ai for medical imaging: extending the monai framework". In: *arXiv preprint arXiv:2307.15208* (2023).

[123] Stéphane P Poulin et al. "Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity". In: *Psychiatry Research: Neuroimaging* 194.1 (2011), pp. 7–13.

[124] Konpat Preechakul et al. "Diffusion autoencoders: Toward a meaningful and decodable representation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10619–10629.

[125] Lemuel Puglisi, Daniel C Alexander, and Daniele Ravì. "Enhancing spatiotemporal disease progression models via latent diffusion and prior knowledge". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 173–183.

[126] Hengnian Qi et al. "A study of auxiliary screening for Alzheimer's disease based on handwriting characteristics". In: *Frontiers in Aging Neuroscience* 15 (2023). ISSN: 1663-4365. DOI: 10.3389/fnagi.2023.1117250.

[127] Shangran Qiu et al. "Development and validation of an interpretable deep learning framework for Alzheimer's disease classification". In: *Brain* 143.6 (2020), pp. 1920–1933.

[128] Natália Bezerra Mota Quental, Sonia Maria Dozzi Brucki, and Orlando Francisco Amodeo Bueno. "Visuospatial function in early Alzheimer's disease—the use of the Visual Object and Space Perception (VOSP) battery". In: *PloS one* 8.7 (2013), e68398.

[129] Y Lakshmisha Rao et al. "Hippocampus and its involvement in Alzheimer's disease: a review". In: *3 Biotech* 12.2 (2022), p. 55.

[130] Yongming Rao et al. "Global filter networks for image classification". In: *Advances in neural information processing systems* 34 (2021), pp. 980–993.

[131] Saima Rathore et al. "A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages". In: *NeuroImage* 155 (2017), pp. 530–548.

[132] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. "Stochastic backpropagation and approximate inference in deep generative models". In: *International conference on machine learning*. PMLR. 2014, pp. 1278–1286.

[133] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

[134] Alexandre Routier et al. "Clinica: An open-source software platform for reproducible clinical neuroscience studies". In: *Frontiers in Neuroinformatics* 15 (2021), p. 689675.

[135] Alexandre Routier et al. "Clinica: An open-source software platform for reproducible clinical neuroscience studies". In: *Frontiers in neuroinformatics* 15 (2021), p. 689675.

[136] Saeid Safiri et al. "Alzheimer's disease: a comprehensive review of epidemiology, risk factors, symptoms diagnosis, management, caregiving, advanced treatments and associated challenges". In: *Frontiers in Medicine* 11 (2024), p. 1474043.

[137] Jorge Samper-González et al. "Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data". In: *NeuroImage* 183 (2018), pp. 504–521.

[138] Jorge Samper-González et al. "Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data". In: *NeuroImage* 183 (2018), pp. 504–521.

[139] Philip Scheltens et al. "Atrophy of medial temporal lobes on MRI in" probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates." In: *Journal of Neurology, Neurosurgery & Psychiatry* 55.10 (1992), pp. 967–972.

[140] Jo Schlemper et al. "Attention gated networks: Learning to leverage salient regions in medical images". In: *Medical image analysis* 53 (2019), pp. 197–207.

[141] Ramprasaath R Selvaraju et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[142] Priscilla Chantal Duarte Silva et al. "Executive functions in Alzheimer's disease: A systematic review". In: *Journal of Alzheimer's Disease Reports* 6.1 (2022), pp. 81–99.

[143] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *CoRR* abs/1409.1556 (2014). arXiv: 1409.1556.

[144] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *CoRR* abs/1409.1556 (2014). arXiv: 1409.1556. URL: http://arxiv.org/abs/1409.1556.

[145] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. "Explainable deep learning models in medical image analysis". In: *Journal of imaging* 6.6 (2020), p. 52.

[146] Stephen M Smith. "Fast robust automated brain extraction". In: *Human brain mapping* 17.3 (2002), pp. 143–155.

[147] Brian J. Soher, Brian M. Dale, and Elmar M. Merkle. "A review of MR physics: 3T versus 1.5T". In: *Magnetic Resonance Imaging Clinics of North America* 15.3 (2007), pp. 277–290. DOI: 10.1016/j.mric.2007.06.002.

[148] Jascha Sohl-Dickstein et al. "Deep unsupervised learning using nonequilibrium thermodynamics". In: *International conference on machine learning*. PMLR. 2015, pp. 2256–2265.

[149] Jiaming Song, Chenlin Meng, and Stefano Ermon. "Denoising diffusion implicit models". In: *arXiv preprint arXiv:2010.02502* (2020).

[150] Yang Song and Stefano Ermon. "Generative modeling by estimating gradients of the data distribution". In: *Advances in neural information processing systems* 32 (2019).

[151] Yang Song et al. "Score-based generative modeling through stochastic differential equations". In: *arXiv preprint arXiv:2011.13456* (2020).

[152] Reisa A Sperling et al. "Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease". In: *Alzheimer's & dementia* 7.3 (2011), pp. 280–292.

[153] Cathie Sudlow et al. "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age". In: *PLoS medicine* 12.3 (2015), e1001779.

[154] Cathie Sudlow et al. "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age". In: *PLoS medicine* 12.3 (2015), e1001779.

[155] Mingxing Tan and Quoc Le. "Efficientnetv2: Smaller models and faster training". In: *International conference on machine learning*. PMLR. 2021, pp. 10096–10106.

[156] Mingxing Tan and Quoc V. Le. "EfficientNetV2: Smaller Models and Faster Training". In: *arXiv preprint arXiv:2104.00298* (2021).

[157] Stephen Todd et al. "Survival in dementia and predictors of mortality: a review". In: *International journal of geriatric psychiatry* 28.11 (2013), pp. 1109–1124.

[158] Jussi Tohka. "Partial volume effect modeling for segmentation and tissue classification of brain magnetic resonance images: A review". In: *World Journal of Radiology* 6.11 (2014), pp. 855–864. DOI: 10.4329/wjr.v6.i11.855.

[159] Nicholas J Tustison et al. "N4ITK: improved N3 bias correction". In: *IEEE transactions on medical imaging* 29.6 (2010), pp. 1310–1320.

[160] Nicholas J. Tustison et al. "N4ITK: Improved N3 Bias Correction". In: *IEEE Transactions on Medical Imaging* 29.6 (2010), pp. 1310–1320. DOI: 10.1109/TMI.2010.2046908.

[161] Bas HM Van der Velden et al. "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis". In: *Medical Image Analysis* 79 (2022), p. 102470.

[162] Gary W Van Hoesen et al. "The parahippocampal gyrus in Alzheimer's disease: clinical and preclinical neuroanatomical correlates". In: *Annals of the New York Academy of Sciences* 911.1 (2000), pp. 254–274.

[163] Koen Van Leemput et al. "Automated segmentation of hippocampal subfields from ultra-high resolution in vivo MRI". In: *Hippocampus* 19.6 (2009), pp. 549–557.

[164] Vilma Velickaite et al. "Medial temporal lobe atrophy ratings in a large 75-year-old population-based cohort: gender-corrected and education-corrected normative data". In: *European radiology* 28.4 (2018), pp. 1739–1747.

[165] G. Vessio. "Dynamic Handwriting Analysis for Neurodegenerative Disease Assessment: A Literary Review." In: *Applied Sciences* 9(21):4666 (2019). DOI: 10.3390/app9214666.

[166] Vimbi Viswan et al. "Explainable artificial intelligence in Alzheimer's disease classification: A systematic review". In: *Cognitive Computation* 16.1 (2024), pp. 1–44.

[167] Chenhui Wang et al. "Joint learning framework of cross-modal synthesis and diagnosis for alzheimer's disease by mining underlying shared modality information". In: *Medical Image Analysis* 91 (2024), p. 103032.

[168] Haoshen Wang et al. "3D MedDiffusion: A 3D Medical Latent Diffusion Model for Controllable and High-quality Medical Image Generation". In: *IEEE Transactions on Medical Imaging* (2025).

[169] Shui-Hua Wang et al. "Classification of Alzheimer's disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling". In: *Journal of medical systems* 42 (2018), pp. 1–11.

[170] Junhao Wen et al. "Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation". In: *Medical image analysis* 63 (2020), p. 101694.

[171] Junhao Wen et al. "Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation". In: *Medical image analysis* 63 (2020), p. 101694.

[172] P. Werner, S. Rosenblum, and A. Korczyn. "Handwriting Process Variables Discriminating Mild Alzheimer's Disease and Mild Cognitive Impairment". In: *Journal of Gerontology: PSYCHOLOGICAL SCIENCES* 61.4 (2006), pp. 228–36.

[173] Perla Werner et al. “Handwriting Process Variables Discriminating Mild Alzheimer’s Disease and Mild Cognitive Impairment”. In: *The journals of gerontology. Series B, Psychological sciences and social sciences* 61 (Aug. 2006), P228–36. DOI: 10.1093/geronb/61.4.P228.

[174] JL Whitwell et al. “MRI correlates of neurofibrillary tangle pathology at autopsy: a voxel-based morphometry study”. In: *Neurology* 71.10 (2008), pp. 743–749.

[175] Yuanchen Wu et al. “An Attention-Based 3D CNN With Multi-Scale Integration Block for Alzheimer’s Disease Classification”. In: *IEEE Journal of Biomedical and Health Informatics* 26.11 (2022), pp. 5665–5673.

[176] Bradley T Wyman et al. “Standardization of analysis sets for reporting results from ADNI MRI data”. In: *Alzheimer’s & Dementia* 9.3 (2013), pp. 332–337.

[177] Zhaohu Xing et al. “Cross-conditioned diffusion model for medical image to image translation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 201–211.

[178] Zhaohu Xing et al. “Diff-UNet: A diffusion embedded network for robust 3D medical image segmentation”. In: *Medical Image Analysis* (2025), p. 103654.

[179] Jin H Yan et al. “Alzheimer’s disease and mild cognitive impairment deteriorate fine movement control”. In: *Journal of Psychiatric Research* 42.14 (2008), pp. 1203–1212.

[180] Yijun Yang et al. “Medical world model: Generative simulation of tumor evolution for treatment planning”. In: *arXiv preprint arXiv:2506.02327* (2025).

[181] Tal Yarkoni et al. “PyBIDS: Python tools for BIDS datasets”. In: *Journal of Open Source Software* 4.40 (2019), p. 1294. DOI: 10.21105/joss.01294. URL: https://doi.org/10.21105/joss.01294.

[182] Jee Seok Yoon et al. “Sadm: Sequence-aware diffusion model for longitudinal medical image generation”. In: *International Conference on Information Processing in Medical Imaging*. Springer. 2023, pp. 388–400.

[183] Vanessa M Young et al. “Data Leakage in Deep Learning for Alzheimer’s Disease Diagnosis: A Scoping Review of Methodological Rigor and Performance Inflation”. In: *Diagnostics* 15.18 (2025), p. 2348.

[184] Jiahui Yu et al. “Vector-quantized image modeling with improved vqgan”. In: *arXiv preprint arXiv:2110.04627* (2021).

[185] Richard Zhang et al. “The unreasonable effectiveness of deep features as a perceptual metric”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.

[186] Zijian Zhang, Zhou Zhao, and Zhijie Lin. “Unsupervised representation learning from pre-trained diffusion probabilistic models”. In: *Advances in neural information processing systems* 35 (2022), pp. 22117–22130.

[187] Guohua Zhao et al. “Research on magnetic resonance imaging in diagnosis of Alzheimer’s disease”. In: *European Journal of Medical Research* 29.1 (2024), p. 632.

[188] Qinghua Zhou et al. “A Survey of Deep Learning for Alzheimer’s Disease”. In: *Machine Learning and Knowledge Extraction* 5.2 (2023), pp. 611–668.