People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research

Abou Bakr Belkaid University – Tlemcen Faculty of Sciences
Department of Computer Science

Graduation Project Thesis

Submitted in partial fulfillment of the requirements for the degree of Master in Computer Science

---

Theme:
alzheimer's desease classification from MRI images

Presented by:

- Mekidiche Khayreddine
- Benikhlef Imad Eddine

Supervised by:

Mr. Meziane Abdelfettah

Defended on 21/06/2026 before the jury composed of:

- Mr. Brikci-Nigassa Amine *(President)*
- Mrs. Amghar Djazia *(Examiner)*
- Mr. Meziane Abdelfettah *(Supervisor)*

Academic Year: 2025–2026

Acknowledgements

We express our sincere gratitude to our supervisor, **Mr. Meziane Abdelfettah**, for his guidance, availability, and invaluable advice throughout the realization of this graduation project.

We extend our thanks to the members of the jury for honoring us with their presence and the time they have dedicated to evaluating this work.

Our gratitude also goes to all the faculty members of the Department of Computer Science at Abou Bakr Belkaid University — Tlemcen, whose teaching has been the foundation of our academic journey.

Finally, we are deeply grateful to our families for their unwavering support and help.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AD Alzheimer's Disease. 1–13, 16, 18, 20–22, 27, 29, 30, 32–34, 36, 37, 45, 46, 60, 69

ADNI Alzheimer's Disease Neuroimaging Initiative. viii, 15, 20–22, 24–27, 37–40, 46, 52–55, 58, 59, 61, 65–70

CBAM Convolutional Block Attention Module. 36, 37, 45, 49, 60, 68, 69

CDR Clinical Dementia Rating. 8, 16, 17, 21

CN Cognitively Normal. 30, 34, 47

CNN Convolutional Neural Network. 12, 17, 18, 29–34, 61

CSF Cerebrospinal Fluid. 9, 17, 20–22, 69

DICOM Digital Imaging and Communications in Medicine. 37

DL Deep Learning. 12, 13, 29, 33

DTI Diffusion Tensor Imaging. 10, 20

FDA Food and Drug Administration. 11, 20

fMRI Functional Magnetic Resonance Imaging. 10, 20

GM Grey Matter. 9, 10

GPU Graphics Processing Unit. 31, 33, 34, 47

MCI Mild Cognitive Impairment. 1, 6–8, 10, 20–22, 24, 27, 30, 32, 34, 47, 56, 57, 59, 63, 65, 67, 69

ML Machine Learning. 14

MLP Multi-Layer Perceptron. 45

MMSE Mini-Mental State Examination. 8, 20, 21

MRI Magnetic Resonance Imaging. 2, 6, 9, 10, 12–22, 25–27, 29–32, 34, 36, 37, 39, 52, 60, 61, 67, 68, 70

NIH National Institutes of Health. 16, 20

PET Positron Emission Tomography. 9, 17, 20–22, 69

ROI Region of Interest. 10, 12, 30, 32

SWA Stochastic Weight Averaging. 56

VBM Voxel-Based Morphometry. 10

WHO World Health Organization. 1

WM White Matter. 9

GENERAL INTRODUCTION

Background and Context

Alzheimer's disease (Alzheimer's Disease (AD)) is a progressive neurodegenerative disorder and the leading cause of dementia worldwide. It is characterized by the gradual deterioration of memory, cognitive functions, and behavioral abilities, ultimately rendering patients unable to perform basic daily activities. According to the World Health Organization (WHO), more than 55 million people live with dementia globally, and nearly 10 million new cases are diagnosed every year, with AD accounting for approximately 60 to 70 percent of all cases.

The disease progresses through several stages, typically classified as Mild Cognitive Impairment (MCI), mild Alzheimer's, moderate Alzheimer's, and severe Alzheimer's. Despite decades of intensive research, no curative treatment has been found to date. However, early and accurate detection can significantly slow the progression of the disease and improve the patient's quality of life through timely therapeutic interventions.

Problem Statement

The early and accurate diagnosis of AD remains a critical clinical challenge. Traditional diagnostic methods rely heavily on neuropsychological tests and radiological assessments performed by specialists, which are not always accessible in rural or low-resource settings and may yield inconsistent results depending on the clinician's experience. This creates a pressing need for automated, objective, and reliable tools that can assist clinicians in detecting and classifying AD at an early stage.

Project Theme and Objectives

This graduation project addresses the theme:

"Alzheimer's Disease Classification by Brain Structure Segmentation of MRI Images"

The project develops a deep learning system that first *segments* key brain structures (hippocampus, entorhinal cortex, amygdala) from Magnetic Resonance Imaging (MRI) scans, then uses the morphometric features extracted from these segments to *classify* the stage of AD. This two-stage approach offers greater interpretability and clinical relevance compared to end-to-end black-box classifiers.

Structure of the Thesis

This thesis is organized as follows:

- **Chapter I:** Provides a comprehensive introduction to Alzheimer's disease, covering its clinical definition, epidemiology, pathophysiology, staging, and the role of neuroimaging as a diagnostic tool.
- **Chapter II:** Presents the three major publicly available MRI datasets used in Alzheimer's research (OASIS, Kaggle, ADNI), analyzes their advantages and limitations, and justifies the selection of ADNI for this project.
- **Chapter III:** Reviews the state of the art in deep learning methods for Alzheimer's classification, comparing 2D, 3D, and 2.5D approaches with a critical discussion of data leakage issues identified in published literature.
- **Chapter IV:** Details the complete design and implementation of the proposed 2.5D DenseNet121-CBAM pipeline, including preprocessing, slice selection, architecture, and training protocol.
- **Chapter V:** Presents the experimental results on the held-out ADNI test set, with full performance analysis, confusion matrix interpretation, ROC analysis, and comparison with the state of the art.

CHAPTER

I

ALZHEIMER'S DISEASE: CLINICAL AND NEUROLOGICAL OVERVIEW

Introduction

Alzheimer's disease (AD) stands as one of the most challenging and devastating neurological disorders of our time. As the world's population ages at an unprecedented rate, dementia in general and Alzheimer's disease in particular have become major public health priorities. Understanding the disease at multiple levels — clinical, biological, neuroanatomical, and technological — is the essential first step toward building automated systems that can assist in its detection. This chapter provides a comprehensive review of Alzheimer's disease, covering its history, epidemiology, pathophysiology, clinical presentation, staging systems, neuroimaging biomarkers, and the current state of diagnosis and treatment. This foundational knowledge directly informs the design choices made in later chapters of this memoir.

1 Historical Overview of Alzheimer's Disease

The history of Alzheimer's disease begins in November 1906, when a German psychiatrist and neuropathologist, Dr. Alois Alzheimer, presented a landmark case at the 37th Meeting of South-West German Psychiatrists in Tübingen. His patient, Auguste Deter, a 51-year-old woman, had been admitted to the Frankfurt asylum in 1901 with a clinical picture dominated by progressive memory loss, paranoia, sleep disorders, and profound disorientation. After her death in 1906, Alzheimer performed a meticulous post-mortem examination of her brain, discovering two hallmark neuropathological features: abnormal clumps (later named amyloid plaques) and tangled bundles of fibers (later named neurofibrillary tangles) [1].

For several decades, the disease remained classified as a rare form of "presenile de-

mentia," affecting patients before the age of 65. Dementia occurring after 65 was labeled "senile dementia" and considered a normal consequence of aging. It was not until the 1970s that researchers, most notably Robert Katzman in his influential 1976 editorial in *Archives of Neurology*, argued compellingly that presenile and senile dementias were the same disease, defined by the same neuropathological hallmarks [2]. This conceptual unification was pivotal: it transformed Alzheimer's disease from a rare curiosity into a major public health concern.

The following decades saw an explosion of research. The discovery of the amyloid precursor protein (AD-related mutations on chromosomes 21, 14, and 1) in the 1980s and 1990s laid the molecular groundwork. The development of the first cholinesterase inhibitors in the 1990s, though palliative rather than curative, represented the first pharmacological interventions. Today, Alzheimer's disease research is one of the most active fields in biomedicine, with thousands of clinical trials ongoing worldwide and a growing investment in artificial intelligence tools for early detection and monitoring.

2 Epidemiology

2.1 Global Prevalence

AD is the most common cause of dementia, accounting for 60 to 70 percent of all dementia cases worldwide [3]. According to the World Alzheimer Report 2023, approximately 55 million people worldwide were living with dementia in 2023, and this figure is projected to reach 78 million by 2030 and 139 million by 2050 — driven primarily by the aging of global populations. Every three seconds, a new person is diagnosed with dementia somewhere in the world.

2.2 Regional Disparities

The burden of AD is not evenly distributed. More than 60 percent of people with dementia currently live in low- and middle-income countries, including nations in North Africa and the Middle East such as Algeria. In these regions, awareness remains low, specialized neurological care is scarce, and diagnostic infrastructure is under-resourced. This has severe consequences: most patients receive a diagnosis only in moderate or severe stages of the disease, when intervention is far less effective.

Algeria, with one of the fastest-aging populations in the Arab world — the proportion of individuals aged 60 and above is projected to double by 2050 — faces a growing dementia burden for which it is currently unprepared. The absence of local epidemiological data, the shortage of neurologists and geriatricians, and the lack of automated diagnostic tools all contribute to a significant unmet need [4].

2.3 Economic and Social Burden

The societal cost of dementia is enormous. The World Alzheimer Report 2019 estimated the global economic cost of dementia at approximately USD 1 trillion per year, a figure expected to double by 2030. These costs encompass direct medical care, professional social care, and informal (family) care. The emotional and psychological burden on caregivers is equally significant: family members who provide informal care face high rates of depression, anxiety, and burnout.

3 Pathophysiology

3.1 Molecular Mechanisms: Amyloid and Tau

At the molecular level, AD is defined by two hallmark pathological processes: the accumulation of extracellular amyloid-beta plaques and the formation of intracellular neurofibrillary tangles composed of hyperphosphorylated tau protein.

1 Amyloid-Beta Cascade

The amyloid precursor protein (APP) is a transmembrane protein expressed throughout the body. In healthy individuals, APP is cleaved by α-secretase along a non-amyloidogenic pathway. In Alzheimer's disease, APP is cleaved instead by β-secretase (BACE1) and γ-secretase, producing amyloid-beta (Aβ) fragments. The most neurotoxic of these is the 42-amino-acid form (Aβ_{42}), which has a high tendency to aggregate into oligomers and, subsequently, insoluble fibrils that deposit as amyloid plaques in the brain parenchyma [5].

These plaques disrupt synaptic transmission, trigger neuroinflammation, and initiate a cascade of events that ultimately leads to widespread neuronal death. The *amyloid cascade hypothesis* proposes that this accumulation of Aβ is the initiating event in Alzheimer's pathogenesis, though this hypothesis remains the subject of active scientific debate.

2 Tau Pathology and Neurofibrillary Tangles

Tau is a microtubule-associated protein essential for stabilizing the neuronal cytoskeleton. In AD, tau becomes abnormally hyperphosphorylated, causing it to detach from microtubules and aggregate into paired helical filaments that form neurofibrillary tangles (NFTs). These tangles disrupt axonal transport — the mechanism by which neurons deliver proteins and organelles along their axons — and ultimately lead to neuronal death. The spatial progression of NFT pathology through the brain follows a predictable pattern, described by Braak staging (Braak stages I through VI), beginning in the transentorhinal cortex and hippocampus before spreading to association cortices and, in severe cases, to primary cortices [6]. This stereotyped progression has important implications for neuroimaging: atrophy in the hippocampus and entorhinal cortex is among the earliest detectable structural changes in AD.

3.2 Neuroinflammation

Beyond amyloid and tau, neuroinflammation is now recognized as a central and active contributor to Alzheimer's pathophysiology rather than a mere secondary response. Activated microglia and reactive astrocytes are consistently found in the vicinity of amyloid plaques. While initially protective, chronic microglial activation releases pro-inflammatory cytokines (including interleukin-1β, TNF-α, and IL-6) that accelerate neurodegeneration. Genome-wide association studies (GWAS) have identified numerous Alzheimer's risk genes — including TREM2, CR1, and CLU — that are predominantly expressed in immune cells of the brain, further implicating the innate immune system [7].

3.3 Synaptic Loss and Cholinergic Dysfunction

The loss of synapses is one of the strongest pathological correlates of cognitive decline in AD. The cholinergic hypothesis, one of the first neurochemical theories of the disease, proposed that degeneration of cholinergic neurons in the nucleus basalis of Meynert — the primary source of cholinergic innervation to the cerebral cortex — underlies the memory deficits in Alzheimer's disease. While this hypothesis has been refined over the decades, it remains clinically relevant: the most widely used pharmacological treatments for AD (acetylcholinesterase inhibitors such as donepezil, rivastigmine, and galantamine) work by compensating for this cholinergic deficit [8].

4 Brain Regions Affected and Neuroanatomy of AD

4.1 The Hippocampus

The hippocampus, a seahorse-shaped structure located in the medial temporal lobe, is central to the formation of new declarative memories and spatial navigation. It is one of the first regions to show atrophy in AD, reflecting the selective vulnerability of hippocampal neurons to the combination of amyloid toxicity, tau tangles, and synaptic loss. Volumetric measurement of the hippocampus using MRI has become a key biomarker of AD, and hippocampal volume loss predicts conversion from MCI to full Alzheimer's dementia with moderate accuracy. In the context of this project, hippocampal segmentation from MRI images constitutes one of the primary targets of our automated segmentation pipeline.

4.2 The Entorhinal Cortex

The entorhinal cortex (ERC) serves as the main interface between the hippocampus and the neocortex. According to Braak's staging, neurofibrillary tau pathology begins in the transentorhinal cortex even before the hippocampus is affected. Atrophy of the entorhinal cortex is thus among the earliest structural MRI changes in AD and can be detected years before the onset of clinical symptoms. This makes the ERC a particularly valuable target for early detection.

4.3 The Amygdala

The amygdala, involved in emotional processing and fear learning, is also affected early in AD. Amygdalar atrophy correlates with behavioral and psychological symptoms of dementia (BPSD), including anxiety, agitation, and depression. Its proximity to the hippocampus means it is frequently included in automated segmentation protocols alongside hippocampal volumes.

4.4 Association Cortices and Global Atrophy

As AD progresses, atrophy spreads from medial temporal structures to temporal, parietal, and frontal association cortices. In advanced stages, global cortical atrophy, ventricular enlargement, and significant reduction in total brain volume are observable on MRI scans

even to the untrained eye. Posterior cortical atrophy, affecting visual and visuospatial processing areas, characterizes a specific variant of AD known as Posterior Cortical Atrophy (PCA).

5 Clinical Presentation and Staging

5.1 Early Symptoms

The clinical presentation of AD typically begins insidiously, with subjective memory complaints that are often dismissed by both patients and physicians as normal aging. The earliest and most universal symptom is episodic memory impairment — difficulty recalling recent events, conversations, or appointments — while more remote, long-term memories are initially preserved. Patients may also exhibit word-finding difficulties (anomia), impaired visuospatial skills (getting lost in familiar environments), and subtle changes in personality or social behavior.

5.2 Staging Systems

Several staging systems have been proposed to characterize the progression of AD.

1 The Three-Stage Model

The traditional clinical model divides AD into three broad stages:

- **Mild:** The patient is still largely independent. Memory lapses are noticeable but do not prevent most daily activities. Judgment may be impaired. The patient may repeat questions or stories within the same conversation.
- **Moderate:** Assistance with daily activities becomes necessary. Memory loss is severe and extends to personal history. Patients may become confused about time and place, experience significant behavioral changes, and require supervision for personal safety.
- **Severe:** The patient is fully dependent on caregivers. Loss of ability to communicate verbally, walk, or swallow. At this stage, infection (particularly pneumonia) is a major cause of death.

2 The Seven-Stage Reisberg Scale (GDS)

The Global Deterioration Scale (GDS), developed by Barry Reisberg and colleagues, subdivides progression into seven stages, ranging from GDS 1 (no cognitive decline, clinically normal) to GDS 7 (very severe cognitive decline, late-stage AD). This scale is widely used in clinical trials.

3 Mild Cognitive Impairment (MCI)

MCI represents a transitional state between normal aging and Alzheimer's dementia. Patients with MCI exhibit objective cognitive impairment beyond what is expected for their age and education, yet remain functionally independent in daily activities. MCI carries a significantly elevated risk of progression to AD: approximately 10 to 15 percent

of MCI patients per year convert to dementia, compared to 1 to 2 percent in the general elderly population [9]. The classification of MCI as a distinct stage is clinically critical, as it represents the window of opportunity for preventive and therapeutic intervention.

4 The Four-Class Dataset Label System

In the context of machine learning-based classification using MRI datasets (particularly the Kaggle MRI dataset and ADNI subsets), Alzheimer's stages are frequently encoded as a four-class problem:

1. **Non-Demented (ND):** No cognitive impairment; cognitively normal controls.
2. **Very Mild Demented (VMD):** Corresponds approximately to GDS 2–3; early subjective and objective decline.
3. **Mild Demented (MD):** Corresponds to GDS 4; mild Alzheimer's with noticeable functional impact.
4. **Moderate Demented (MoD):** Corresponds to GDS 5–6; moderate dementia requiring assistance.

This four-class formulation is the one adopted in this project, as it aligns directly with the available labeled data and provides clinically meaningful categories.

6 Diagnosis of Alzheimer's Disease

6.1 Clinical and Neuropsychological Assessment

The diagnosis of AD remains primarily clinical. The diagnostic process begins with a comprehensive clinical interview and neurological examination. Standardized cognitive assessment tools are central to this process:

- **Mini-Mental State Examination (Mini-Mental State Examination (MMSE)):** A 30-point questionnaire testing orientation, registration, attention, recall, language, and visuospatial ability. Scores below 24 suggest cognitive impairment; below 10 indicate severe dementia.
- **Montreal Cognitive Assessment (MoCA):** A more sensitive tool for detecting mild impairment than the MMSE, particularly for executive function and attention.
- **Clinical Dementia Rating (Clinical Dementia Rating (CDR)):** A structured interview with the patient and a reliable informant, rating six domains (memory, orientation, judgment, community affairs, home/hobbies, personal care) on a 0–3 scale. The CDR sum of boxes (CDR-SB) is widely used in clinical trials as a primary outcome measure.

6.2 Biological Biomarkers

The *ATN* biomarker framework, proposed by Jack et al. in 2018, operationalizes the biological definition of Alzheimer's disease through three categories of biomarkers [10]:

- **A (Amyloid):** Measured by amyloid-Positron Emission Tomography (PET) imaging or by reduced $A\beta_{42}$ levels in cerebrospinal fluid (Cerebrospinal Fluid (CSF)).
- **T (Tau):** Measured by tau-PET or elevated phosphorylated tau (p-tau) in CSF.
- **N (Neurodegeneration):** Measured by FDG-PET hypometabolism, MRI atrophy, or total tau (t-tau) in CSF.

This framework allows for a biological, rather than purely clinical, diagnosis of AD, enabling earlier intervention in the preclinical phase before symptoms appear.

6.3 Limitations of Current Diagnostic Practice

Despite advances in biomarker research, significant barriers to timely diagnosis persist:

- PET imaging using amyloid or tau tracers is expensive (over USD 3,000 per scan), not widely available, and not reimbursed in most national health systems.
- CSF biomarker analysis requires a lumbar puncture, which is invasive and associated with patient reluctance.
- Neuropsychological testing is time-consuming, requires trained specialists, and can be influenced by the patient's educational level, language, and cultural background.
- In low-resource settings such as Algeria, access to any of the above tools is highly limited, making automated MRI-based classification particularly valuable.

7 Neuroimaging in Alzheimer's Disease

7.1 Why MRI?

MRI has become the imaging modality of choice for the structural assessment of AD for several compelling reasons:

- **Non-invasive and radiation-free:** Unlike PET or CT, MRI uses magnetic fields and radio waves, eliminating the risks associated with ionizing radiation. This makes it suitable for longitudinal studies with repeated scanning.
- **High soft-tissue contrast:** MRI provides exquisite contrast between grey matter (Grey Matter (GM)), white matter (White Matter (WM)), and cerebrospinal fluid (CSF), enabling precise delineation of brain structures.
- **Quantitative morphometry:** High-resolution 3T MRI allows volumetric measurement of specific structures (hippocampus, entorhinal cortex, whole brain) with sub-millimeter precision.
- **Relative accessibility:** Compared to PET scanners, MRI machines are significantly more common in hospital settings, including in developing countries.

7.2 Structural MRI Biomarkers

Several structural MRI measures have been validated as biomarkers of AD:

1 Hippocampal Volume

Hippocampal atrophy is the most extensively studied structural MRI biomarker. Studies using manual or automated segmentation consistently show that hippocampal volume is reduced in MCI and AD patients compared to age-matched controls, and that the rate of annual hippocampal atrophy is accelerated in AD (roughly 4–8% per year) compared to normal aging (0.5–1.5% per year) [11].

2 Cortical Thickness

Cortical thinning, particularly in the entorhinal cortex, temporal, and parietal regions, can be detected using surface-based morphometry tools such as FreeSurfer. Cortical thickness measures complement volumetric analyses and provide regionally specific information about the pattern of neurodegeneration.

3 White Matter Hyperintensities

White matter hyperintensities (WMH), visible as bright regions on T2-weighted MRI, reflect small vessel cerebrovascular disease. While not specific to AD, WMH frequently co-occur with Alzheimer's pathology and contribute to cognitive decline, making them a relevant ancillary biomarker.

4 Total Brain Volume and Ventricular Enlargement

Global measures such as total brain volume, intracranial volume, and the size of the cerebral ventricles (which enlarge as brain tissue is lost) provide a summary measure of the overall degree of neurodegeneration.

7.3 Voxel-Based Morphometry

Voxel-Based Morphometry (VBM) is a fully automated, unbiased technique that performs a voxel-wise statistical comparison of local GM concentration or volume between groups. In AD research, VBM has been widely used to identify regions of atrophy compared to healthy controls. Unlike Region of Interest (ROI)-based analyses, VBM is exploratory and does not require an a priori hypothesis about which regions are affected. Its results are typically displayed as statistical maps overlaid on a standard brain template (MNI space) [12].

7.4 Advanced MRI Techniques

Beyond structural MRI, several advanced techniques provide complementary information:

- **Functional MRI (Functional Magnetic Resonance Imaging (fMRI)):** Measures blood oxygen level-dependent (BOLD) signals as a proxy for neural activity. In AD, resting-state fMRI reveals disrupted functional connectivity within the default mode network (DMN), a set of brain regions active during rest and involved in self-referential thought and memory.
- **Diffusion Tensor Imaging (Diffusion Tensor Imaging (DTI)):** Measures the directionality of water diffusion in brain tissue, providing a measure of white matter

integrity. Fractional anisotropy (FA) is reduced in white matter tracts connecting the hippocampus to frontal cortex in AD, reflecting axonal damage.

- **Arterial Spin Labeling (ASL):** A non-invasive technique for measuring cerebral blood flow (CBF). Reduced CBF in temporoparietal regions mirrors the hypometabolism seen on FDG-PET, suggesting a common underlying pathology.

8 Treatment and Current Therapeutic Landscape

8.1 Approved Pharmacological Treatments

To date, no treatment has been shown to halt or reverse the progression of AD. Current approved medications provide symptomatic relief by modulating neurotransmitter systems:

- **Acetylcholinesterase inhibitors (donepezil, rivastigmine, galantamine):** Inhibit the enzyme that breaks down acetylcholine, thereby increasing cholinergic transmission. Approved for mild-to-moderate AD; provide modest cognitive and functional benefits.
- **Memantine:** An NMDA receptor antagonist that modulates glutamatergic transmission. Approved for moderate-to-severe AD, often used in combination with a cholinesterase inhibitor.

8.2 Disease-Modifying Therapies: A New Era

The recent approval by the Food and Drug Administration (FDA) of lecanemab (Leqembi) in 2023 and donanemab in 2024 — anti-amyloid monoclonal antibodies that reduce amyloid plaque burden — marked a historic turning point. For the first time, a treatment was shown in large phase III trials to meaningfully slow the rate of clinical decline in early AD, albeit with significant side effects (amyloid-related imaging abnormalities, ARIA) and at high cost [13].

These approvals underscore the importance of early and accurate diagnosis: disease-modifying therapies are effective only in the preclinical or very early clinical stages, and require confirmation of amyloid pathology before treatment. This creates an even stronger imperative for automated, accessible, and reliable early detection tools such as the system proposed in this project.

8.3 Non-Pharmacological Interventions

Evidence supports the use of several non-pharmacological interventions to improve quality of life and possibly slow cognitive decline:

- Cognitive stimulation therapy and cognitive training
- Physical exercise programs (aerobic exercise is associated with increased hippocampal volume in elderly individuals)
- Management of vascular risk factors (hypertension, diabetes, dyslipidemia)

- Nutritional interventions (Mediterranean diet)
- Caregiver education and support programs

9 Artificial Intelligence and Deep Learning in AD Research

9.1 From Manual to Automated Analysis

The manual segmentation of brain structures from MRI images is a specialized, time-intensive task requiring expert neuroanatomical knowledge. A trained researcher may spend 30 to 60 minutes manually delineating the hippocampus on a single MRI scan. This bottleneck makes manual segmentation impractical at the scale required for large clinical studies. Automated segmentation tools such as FreeSurfer [14] and FSL [15] have been available for decades, but they rely on registration-based approaches that can be slow and are sensitive to image quality variations.

Deep Learning (DL)-based methods have revolutionized this field. Convolutional Neural Network (CNN)-based segmentation models, trained on large labeled datasets, can perform whole-brain or ROI-specific segmentation in seconds, with accuracy approaching that of expert manual delineation.

9.2 Convolutional Neural Networks (CNNs)

CNNs are a class of deep learning architectures specifically designed for grid-structured data such as images. A CNN consists of multiple convolutional layers, each applying a set of learnable filters to the input, extracting increasingly abstract features — from edges and textures in early layers to complex semantic representations in deeper layers. Pooling layers reduce spatial dimensionality, and fully connected layers produce the final classification output.

In the context of AD classification from MRI, CNNs can be applied in two ways:

- **End-to-end classification:** The raw MRI slice or volume is fed directly into the CNN, which outputs a predicted diagnostic class. This approach is powerful but offers limited interpretability.
- **Segmentation-then-classification:** A segmentation model (typically a U-Net architecture) first delineates brain structures of interest; morphometric features extracted from these segments (volumes, thickness, surface area) are then used for classification. This is the approach adopted in the present project, as it provides clinically interpretable biomarkers alongside the classification decision.

9.3 Transfer Learning

Training a deep CNN from scratch requires large labeled datasets and significant computational resources. Transfer learning addresses this by initializing the network with weights pretrained on a large general-purpose dataset (such as ImageNet) and fine-tuning on the target medical imaging dataset. Well-known architectures used with transfer learning in AD research include VGG16, ResNet50, InceptionV3, EfficientNet, and DenseNet.

Although these models were originally designed for natural images rather than MRI, transfer learning has consistently demonstrated strong performance on medical imaging tasks, even with relatively small training sets [16].

9.4 U-Net and Medical Image Segmentation

The U-Net architecture, introduced by Ronneberger et al. in 2015 for biomedical image segmentation, has become the de facto standard for medical image segmentation tasks. Its encoder-decoder structure with skip connections allows the network to simultaneously capture high-level semantic context (from the encoder pathway) and precise spatial localization (from the decoder pathway). 3D extensions of U-Net (V-Net, 3D U-Net) operate directly on volumetric MRI data, avoiding the information loss inherent in slice-by-slice 2D processing [17].

Table I.1: Overview of deep learning architectures used in Alzheimer's MRI analysis

Architecture	**Task**	**Key Advantage**
VGG16 / ResNet50	Classification (2D slices)	Strong transfer learning baseline
DenseNet	Classification	Feature reuse; fewer parameters
EfficientNet	Classification	State-of-the-art accuracy / efficiency
U-Net (2D)	Segmentation	High accuracy with limited data
3D U-Net / V-Net	Volumetric segmentation	Exploits full 3D context
Transformer / ViT	Classification & Segmentation	Long-range attention; emerging SOTA

9.5 Explainability and Clinical Trust

A fundamental challenge in deploying DL models in clinical practice is the *black-box* problem: clinicians cannot adopt a diagnostic tool they do not understand. Explainability methods such as Gradient-weighted Class Activation Mapping (Grad-CAM), LIME, and SHAP have been applied to AD classification models to generate visual explanations highlighting the brain regions most influential in the model's decision. Segmentation-based pipelines (such as the one in this project) provide an additional level of interpretability: the model's intermediate outputs — the segmented brain structures and their extracted volumes — are themselves clinically meaningful and can be inspected and validated by a neurologist.

10 Conclusion

This chapter has established the medical and scientific foundation upon which the rest of this memoir is built. Alzheimer's disease is a complex, progressive, and devasting neurodegenerative condition whose accurate and early diagnosis remains a critical unmet need, particularly in low-resource clinical environments. The neuropathological hallmarks of the disease — amyloid plaques, neurofibrillary tangles, and the consequent atrophy of medial temporal structures including the hippocampus and entorhinal cortex — provide the biological rationale for using structural MRI as the primary imaging modality.

The growing power of deep learning, particularly convolutional and U-Net-based architectures, offers a compelling path toward automated, accurate, and interpretable classification of Alzheimer's stages from MRI. The segmentation-then-classification paradigm, adopted in this project, aligns computational methodology with clinical practice, producing biomarkers that are both machine-generated and neurologically meaningful.

The following chapter addresses a prerequisite to any such system: the selection of an appropriate, high-quality MRI dataset. The quality and richness of training data are, in many respects, the decisive factor in the performance of any Machine Learning (ML)-based medical imaging system.

CHAPTER

II

MRI DATASETS FOR ALZHEIMER'S RESEARCH: COMPARATIVE ANALYSIS AND DATASET SELECTION

Introduction

In any machine learning or deep learning project, the quality, richness, and representativeness of the training data are arguably the most important determinants of system performance. This is especially true in medical imaging, where datasets are inherently difficult to compile: they require patient recruitment, informed consent, expert clinical annotation, standardized acquisition protocols, and long-term follow-up. For Alzheimer's disease research, several institutional and community-level efforts have produced publicly available MRI datasets of varying scope, quality, and intended use.

This chapter presents a systematic comparison of the three most widely used MRI datasets in published Alzheimer's classification research: the **OASIS** datasets (Washington University), a widely referenced **Kaggle MRI dataset**, and the **Alzheimer's Disease Neuroimaging Initiative (ADNI)** dataset (a large multi-site consortium). For each dataset, we describe its origin, structure, content, and annotation methodology, then analyze its advantages and limitations in the context of our specific project. The chapter concludes with a justified rationale for selecting ADNI as the primary dataset for this work.

1 Overview of the Dataset Landscape in AD Research

The diversity of available MRI datasets reflects different strategic goals: some prioritize longitudinal depth (repeated scanning of the same individuals over years), others prior-

itize breadth (large cross-sectional samples), and others are optimized for accessibility (preprocessed, clearly labeled, hosted on open platforms). Table II.1 provides a high-level comparison before the detailed analyses that follow.

Table II.1: High-level comparison of major Alzheimer's MRI datasets

Feature	**OASIS**	**Kaggle MRI**	**ADNI**
Type	Cross-sectional & longitudinal	Cross-sectional (derived)	Longitudinal, multi-site
Size	416–2,168 subjects	~6,400 images (4 classes)	>2,000 subjects, thousands of scans
Labels	CDR-based	4-class severity	MCI, AD, CN; full clinical data
Accessibility	Open (immediate)	Open (Kaggle)	Free but application required
Clinical data	Moderate	Minimal	Comprehensive
Multimodal	MRI only	MRI only	MRI, PET, CSF, genetics, cognitive
Preprocessing	Partial	Preprocessed 2D slices	Raw + pipeline-processed

2 The OASIS Dataset

2.1 Description and Origin

OASIS (Open Access Series of Imaging Studies) is an initiative by Washington University in St. Louis, funded by grants from the National Institutes of Health (NIH) and other organizations, with the explicit goal of making neuroimaging datasets freely and openly available to the scientific community. Three major releases have been made:

- **OASIS-1** (2007): A cross-sectional collection of 416 subjects aged 18 to 96 years, including both cognitively normal young adults and older adults with and without dementia. Each subject underwent 3 to 4 T1-weighted MRI sessions on the same day. For subjects over 60 years old, Clinical Dementia Rating (CDR) scores are provided.
- **OASIS-2** (2010): A longitudinal collection of 150 subjects aged 60 to 96, with a minimum of two MRI visits separated by at least one year. 72 subjects were classified as non-demented, and 64 were classified as having AD at their initial visit.
- **OASIS-3** (2019): The most comprehensive release, comprising 1,378 subjects and over 2,842 MRI sessions collected over 15 years. It includes not only structural

MRI but also functional MRI, PET imaging, and extensive clinical, cognitive, and biomarker data.

2.2 Data Structure and Labels

OASIS datasets use CDR scores as the primary clinical label. The CDR is a semi-structured interview that rates six cognitive and functional domains on a scale from 0 (normal) to 3 (severe dementia). A global CDR of 0 indicates no dementia; CDR 0.5 typically corresponds to very mild dementia or questionable dementia; CDR 1 indicates mild dementia; CDR 2, moderate; and CDR 3, severe. For machine learning purposes, these labels are commonly binarized (demented vs. non-demented) or used in multiclass settings.

OASIS-1 and OASIS-2 provide preprocessed MRI images that have undergone skull-stripping, gain-field correction, and registration to a common template, facilitating their use in classification pipelines without extensive preprocessing.

2.3 Advantages of OASIS

- **Immediate, unrestricted access:** OASIS datasets are available for download without a formal application process. This makes them ideal for rapid prototyping and educational use.
- **Well-established benchmark:** OASIS-1 and OASIS-2 have been used in hundreds of published papers, providing a large body of comparable results against which new methods can be benchmarked.
- **Partial preprocessing:** The provided skull-stripped and registered images reduce the preprocessing burden for researchers.
- **Longitudinal data in OASIS-2 and OASIS-3:** These versions allow the study of intra-individual change over time, which is essential for understanding disease progression.
- **Multimodal data in OASIS-3:** The availability of PET and CSF biomarkers alongside MRI enables multimodal studies.

2.4 Limitations of OASIS

- **Small sample sizes:** OASIS-1 (416 subjects) and OASIS-2 (150 subjects) are relatively small by modern deep learning standards. Training large CNN models on these datasets is prone to overfitting without significant data augmentation or transfer learning.
- **Class imbalance:** In OASIS-2, for example, only 78 scans (from 64 subjects) belong to the demented class. This severe imbalance complicates classifier training and evaluation.
- **Limited staging granularity:** The CDR scale, as used in OASIS, tends to aggregate subjects at CDR 0 (non-demented) and CDR 0.5 (very mild / borderline), with relatively few subjects in the CDR 2 and CDR 3 categories, making multiclass classification challenging.

- **Single-site acquisition:** OASIS data were collected at a single institution (Washington University), using consistent but limited scanner configurations. Models trained exclusively on OASIS may not generalize well to data acquired at other sites or on different scanner hardware.
- **Age range heterogeneity in OASIS-1:** The inclusion of young adults (ages 18–59) alongside elderly subjects introduces confounding anatomical variability unrelated to Alzheimer's disease.

3 The Kaggle Alzheimer's MRI Dataset

3.1 Description and Origin

The dataset most commonly referred to as the "Kaggle Alzheimer's MRI Dataset" is a widely circulated collection of axial brain MRI slices, preprocessed into four categories representing the four-class severity staging commonly used in AD research. The most popular version of this dataset (available on Kaggle under various names, most commonly "Alzheimer's Dataset" by user Uraninjo) contains approximately 6,400 images derived from the OASIS datasets, resized to 128×128 pixels and organized into four folders corresponding to the four severity classes.

3.2 Data Structure and Labels

The four classes, along with their approximate image counts in the standard Kaggle split, are:

- **Non-Demented (ND):** 3,200 images — the majority class.
- **Very Mild Demented (VMD):** 2,240 images.
- **Mild Demented (MD):** 896 images.
- **Moderate Demented (MoD):** 64 images — severely underrepresented.

The images are 2D axial slices, presented in JPEG or PNG format, preprocessed (skull-stripped, normalized) and ready for direct ingestion by standard CNN architectures without additional preprocessing.

3.3 Advantages of the Kaggle Dataset

- **Turnkey accessibility:** The dataset can be downloaded in minutes from Kaggle with no registration beyond a free Kaggle account. It is directly compatible with standard deep learning frameworks (TensorFlow, PyTorch).
- **Preprocessed 2D slices:** No neuroimaging expertise or software (FSL, FreeSurfer, ANTs) is required. Researchers can immediately begin training CNN classification models.
- **Four-class labels:** The four severity classes provide a richer classification target than simple binary (demented / non-demented) labels.

- **Large community of reference results:** Hundreds of tutorial notebooks, GitHub repositories, and published papers have used this exact dataset, facilitating baseline comparison.
- **Suitable for rapid benchmarking:** For researchers primarily interested in comparing classifier architectures rather than dataset generalization, the Kaggle dataset provides a convenient, standardized test bed.

3.4 Limitations of the Kaggle Dataset

- **Derived and non-standard provenance:** The Kaggle dataset is not an original data collection but a derivative of OASIS, with labels assigned by the dataset curator rather than through the original OASIS clinical protocol. The mapping from OASIS CDR scores to the four-class Kaggle labels is not officially documented, raising questions about label reliability.
- **Severe class imbalance:** The "Moderate Demented" class contains only 64 images — roughly 1% of the dataset. This extreme imbalance makes it practically impossible to train a well-calibrated classifier for the most clinically important stage (moderate AD) without synthetic oversampling (SMOTE) or other correction techniques.
- **Loss of 3D volumetric information:** By reducing 3D MRI volumes to individual 2D axial slices, the dataset discards the three-dimensional spatial context that is fundamental to structural brain analysis. Volumetric features (total hippocampal volume, cortical thickness across adjacent slices) cannot be computed from isolated 2D slices.
- **No associated clinical or demographic data:** The dataset provides only images and class labels. There are no associated cognitive test scores, demographic variables (age, sex, education), or biological biomarkers. This precludes multivariate analysis and makes it impossible to assess the model's performance conditioned on clinical factors.
- **No longitudinal component:** All images are cross-sectional snapshots. The dataset cannot be used to study disease progression over time or to evaluate the model's ability to predict future decline.
- **Risk of overfitting to preprocessing artifacts:** Because all images have undergone the same preprocessing pipeline and come from a single source (OASIS), models trained on this dataset may overfit to preprocessing artifacts rather than clinically meaningful neuroanatomy. Performance may not generalize to images from other scanners or institutions.
- **Not suitable for segmentation tasks:** Since the project requires brain structure segmentation as an intermediate step, 2D preprocessed slices without ground-truth segmentation masks are insufficient.

4 The ADNI Dataset

4.1 Description and Origin

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a landmark longitudinal multicenter study launched in 2004 under the leadership of Principal Investigator Dr. Michael W. Weiner (University of California, San Francisco), with funding from the NIH, the FDA, and a public-private partnership of pharmaceutical companies and other organizations. ADNI's founding mission was to define the rates of change of biomarkers (imaging, biochemical, and clinical) associated with AD progression, and to standardize methods for their measurement across multiple sites [18].

Since its launch, ADNI has expanded through multiple phases:

- **ADNI-1** (2004–2009): 800 subjects (200 elderly controls, 400 MCI, 200 AD).
- **ADNI-GO** (2009–2011): Added early MCI (eMCI) subjects.
- **ADNI-2** (2011–2016): Expanded to 1,400+ subjects; added younger normal controls.
- **ADNI-3** (2016–present): Incorporates tau-PET, advanced MRI sequences (7T), and digital biomarker collection (wearables).

As of 2024, the ADNI database contains data from over 2,000 subjects, with thousands of longitudinal MRI scans, PET images, CSF biomarker assays, genetic data (including APOE genotyping and whole-genome sequencing), and a comprehensive battery of neuropsychological assessments.

4.2 Data Structure and Content

The ADNI dataset is organized around three primary diagnostic categories:

- **Cognitively Normal (CN):** Subjects with no subjective or objective evidence of cognitive impairment. CDR = 0, MMSE $\geq$ 24.
- **Mild Cognitive Impairment (MCI):** Further subdivided into early MCI (eMCI) and late MCI (lMCI). MMSE 24–30, CDR = 0.5, memory complaint confirmed by informant and objective memory testing.
- **Alzheimer's Disease (AD):** MMSE 20–26, CDR 0.5 or 1.0, meeting NINCDS/ADRDA criteria for probable AD.

MRI data in ADNI are acquired using standardized protocols across all participating sites. Structural MRI sequences include:

- 3D magnetization-prepared rapid gradient echo (MPRAGE) T1-weighted sequences at 1.5T and 3T field strengths.
- ADNI3-specific sequences including 3D T2-FLAIR, susceptibility-weighted imaging (SWI), resting-state fMRI, and multishell DTI.

Standardized preprocessing pipelines (Grad-Warp correction, B1 non-uniformity correction, N3 bias field correction) are applied to all structural MRI images before distribution. Additionally, FreeSurfer-derived cortical parcellations and subcortical segmentations are provided for all subjects, giving researchers access to ready-to-use segmentation-based biomarkers.

4.3 Advantages of ADNI

- **Unmatched scale and depth:** With over 2,000 subjects followed longitudinally over up to 15 years, ADNI is the largest and most comprehensively characterized Alzheimer's neuroimaging dataset in the world. This scale is essential for training deep learning models that generalize reliably.
- **Multimodal data:** ADNI uniquely combines structural MRI, functional MRI, amyloid and tau PET, CSF biomarkers, genetic data, and cognitive assessments for the same individuals. This enables multimodal fusion models that exploit complementary information sources.
- **Gold-standard clinical labels:** All diagnoses in ADNI are established through a rigorous, centralized protocol by expert clinicians, using multiple validated assessment tools (MMSE, CDR, ADAS-Cog, etc.). This labeling rigor is substantially superior to the derived labels in the Kaggle dataset.
- **Multi-site, multi-scanner data:** Data collected across 63 sites in North America using scanners from multiple manufacturers (Siemens, GE, Philips) and field strengths (1.5T and 3T). Models trained on ADNI are inherently more robust to site-specific acquisition differences, making them more generalizable.
- **Longitudinal design:** The ability to follow subjects over time enables the study of conversion from CN to MCI to AD, the evaluation of progression rate predictors, and the development of models that detect pre-clinical changes years before clinical diagnosis.
- **Ground-truth segmentations:** FreeSurfer parcellations provided for all subjects give validated segmentation labels for training and evaluating automated segmentation models — a critical requirement for our segmentation-based pipeline.
- **Community standard:** ADNI is the de facto reference dataset for Alzheimer's biomarker research. The vast majority of high-impact published papers in Alzheimer's machine learning research validate results on ADNI, ensuring that our results will be directly comparable to the state of the art.
- **Active consortium support:** The ADNI Data Sharing and Publications Committee actively supports researchers, provides data use agreements (DUAs), and maintains a detailed data dictionary ensuring reproducibility.

4.4 Limitations of ADNI

- **Access requires application:** Unlike OASIS and Kaggle, ADNI data are not immediately downloadable. Researchers must submit an application, sign a Data

Use Agreement (DUA), and obtain institutional approval. This process typically takes 1–2 weeks, which can delay project initiation.

- **Data volume and storage requirements:** A full download of ADNI structural MRI data amounts to hundreds of gigabytes. Managing, preprocessing, and organizing this data requires significant storage infrastructure and technical expertise in neuroimaging software (FSL, ANTs, FreeSurfer, NiPype).
- **Preprocessing complexity:** While standard preprocessing pipelines are available, applying them correctly requires familiarity with neuroimaging conventions (NIFTI format, coordinate systems, atlas registration, bias field correction). This is a substantially higher technical barrier than the Kaggle dataset, which requires no preprocessing.
- **Recruitment bias:** ADNI subjects are predominantly white, English-speaking, North American volunteers with high educational levels. This demographic homogeneity may limit the generalizability of models trained on ADNI to more diverse populations, including those in Algeria.
- **Missing data:** Not all subjects have complete data across all modalities and time points. Managing missing data (particularly PET and CSF biomarkers, which are available for a subset of subjects) requires careful data quality control.
- **Class imbalance:** Like most real-world clinical datasets, ADNI has a non-uniform distribution of diagnostic categories. MCI subjects are particularly numerous (as the study focuses on the preclinical and prodromal phases), while moderate and severe AD subjects are underrepresented.

5 Comparative Analysis

5.1 Summary Comparison Table

Table II.2: Detailed comparative analysis of the three datasets

Criterion	OASIS	Kaggle MRI	ADNI
Data origin	Washington University	Derived from OASIS	63-site consortium (North America)
Number of subjects	150–1,378 (by version)	N/A (2D slices, no subject IDs)	>2,000
Number of images/scans	Thousands (OASIS-3)	~6,400 PNG slices	Tens of thousands of 3D volumes
MRI type	3D T1-weighted volumes	2D axial slices (PNG)	3D T1, T2, FLAIR, fMRI, DTI
Label quality	High (CDR from clinical protocol)	Moderate (derived, undocumented)	Very high (multi-instrument, expert)
Longitudinal	OASIS-2, OASIS-3 only	No	Yes (up to 15 years)
Multimodal	OASIS-3 only	No	Yes (MRI, PET, CSF, genetics)
Access	Immediate, open	Immediate, open (Kaggle)	Application + DUA (1–2 weeks)
Segmentation masks	Not provided	Not provided	FreeSurfer parcellations provided
Multi-site	No	No	Yes (63 sites)
Preprocessing required	Low (OASIS-1/2)	None	Medium to high
Ideal for	Benchmarking, initial models	Tutorial, rapid prototyping	Research, publication, segmentation
Community papers (approx.)	>500	>200 (Kaggle notebooks)	>3,000

5.2 Dataset Suitability for Segmentation-Based Classification

The specific requirements of our project — a two-stage pipeline involving (1) brain structure segmentation and (2) classification using extracted morphometric features — impose stringent requirements on the dataset:

1. **3D volumetric MRI data:** Segmentation of the hippocampus, entorhinal cortex, and amygdala requires 3D volumes, not 2D slices. The Kaggle dataset, consisting of 2D PNG images, is immediately eliminated for this requirement.
2. **Ground-truth segmentation labels:** Training a supervised segmentation model requires ground-truth annotations (i.e., manually or semi-automatically delineated brain structure masks). ADNI provides FreeSurfer-derived parcellations for all subjects; OASIS does not provide comparable segmentation ground truths.
3. **Sufficient sample size:** Training a deep segmentation model (U-Net or variant) reliably requires at minimum several hundred annotated 3D volumes. Only ADNI provides this scale.
4. **Multi-class clinical labels:** Our classifier targets the CN / MCI / AD three-class problem (or a four-class variant). ADNI provides clinically rigorous labels across all three primary categories with sufficient examples in each class.
5. **Demographic and clinical covariates:** To properly analyze our model's performance (e.g., stratified by age, sex, APOE genotype), associated clinical data are essential. Only ADNI provides a comprehensive clinical phenotyping database.

6 Justification for Selecting ADNI

Based on the comparative analysis above and the specific requirements of our segmentation-based classification project, we select **ADNI** as the primary dataset for this work. The justification is multi-dimensional:

1 Scientific Rigor and Label Reliability

ADNI's clinical labels are established through a centralized, multi-assessment protocol involving the MMSE, ADAS-Cog, CDR, and Neuropsychological Battery, administered by trained and certified examiners at each site. This rigor stands in sharp contrast to the Kaggle dataset's undocumented label derivation. For a medical diagnostic application, ground truth reliability is non-negotiable.

2 Availability of Segmentation Ground Truth

The FreeSurfer subcortical and cortical parcellations provided by ADNI for every subject are the essential ingredient for training our segmentation model. No other publicly available Alzheimer's dataset provides comparable segmentation annotations at scale. OASIS-3, while providing some FreeSurfer outputs, does so for a much smaller population and without the longitudinal richness of ADNI.

3 Scale and Multi-Site Generalizability

With over 2,000 subjects and data from 63 sites, ADNI provides the diversity necessary to train models that generalize beyond a single institution. In contrast, OASIS and Kaggle are essentially single-site datasets. A model trained on ADNI can be reasonably expected to perform on MRI images acquired at different hospitals and with different scanners — a prerequisite for any tool with clinical aspirations.

4 Compatibility with the Segmentation Pipeline

The 3D T1-weighted MPRAGE volumes distributed by ADNI, after standard bias field correction and skull-stripping, are directly compatible with state-of-the-art segmentation frameworks including FSL, FreeSurfer, ANTs, and deep learning models trained on brain MRI (such as SynthSeg, nnU-Net, and DeepMedic). This compatibility dramatically reduces the engineering overhead of building our pipeline.

5 Alignment with the State of the Art

The overwhelming majority of high-quality, peer-reviewed publications in Alzheimer's machine learning research validate their methods on ADNI. Choosing ADNI ensures that our results are directly comparable to the best published work, enabling a meaningful evaluation of our contribution.

6 Acknowledgment of Limitations and Mitigation Strategies

We acknowledge that ADNI's recruitment bias (predominantly white, educated, North American subjects) may limit generalizability to Algerian patient populations. To partially mitigate this, we propose the following:

- Applying data augmentation techniques (rotation, scaling, intensity perturbation) to artificially expand dataset variability.
- Evaluating the trained model on a held-out subset of OASIS-3 subjects to assess cross-dataset generalization.
- Clearly documenting the demographic characteristics of the training set in all reported results.

Table II.3: Decision matrix for dataset selection

Criterion	**Importance**	**OASIS**	**Kaggle**	**ADNI**
3D volumetric MRI	Critical	✓	×	✓
Segmentation ground truth	Critical	×	×	✓
Label quality	High	✓	~	✓✓
Sample size (≥500)	High	~	N/A	✓✓
Multi-site diversity	Medium	×	×	✓✓
Longitudinal data	Medium	~	×	✓✓
Clinical covariates	Medium	~	×	✓✓
Ease of access	Low	✓✓	✓✓	✓
Published benchmarks	Medium	✓	✓	✓✓
Overall suitability for this project		Moderate	Low	**High**

7 ADNI Data Access and Preprocessing Protocol

7.1 Data Access Procedure

Access to ADNI data is obtained through the following steps:

1. Registration on the ADNI website (https://adni.loni.usc.edu).
2. Submission of an access request form specifying the intended research use.
3. Signature of the ADNI Data Use Agreement by the applicant and their institutional supervisor.
4. Approval by the ADNI Data Sharing and Publications Committee (typically 3–10 business days).
5. Data download via the LONI Image & Data Archive (IDA) portal.

7.2 Planned Data Subset

For this project, we plan to use the following ADNI data subset:

- **Modality:** 3D T1-weighted MPRAGE structural MRI only (for the segmentation and primary classification pipeline).
- **Scanner field strength:** 3T scans preferentially, for higher resolution and signal-to-noise ratio.

- **Diagnostic categories:** CN, MCI (late MCI / lMCI), and AD, mapped to a three-class classification target.
- **Time point:** Baseline visit (first scan per subject) to avoid data leakage in the training/test split due to repeated measurements.
- **Preprocessing provided by ADNI:** We will use the standardized preprocessed versions (Grad-Warp, B1 correction, N3 bias field correction) as distributed by the ADNI platform.

7.3 Additional Preprocessing Steps

Beyond the preprocessing provided by ADNI, the following steps will be applied in our pipeline:

1. **Skull stripping:** Removal of non-brain tissue (skull, scalp, dura) using the Brain Extraction Tool (BET) from FSL, or the deep learning-based HD-BET tool.
2. **Intensity normalization:** Standardization of voxel intensities across subjects to a common intensity range (z-score normalization or histogram equalization), to reduce scanner-related intensity variability.
3. **Registration to MNI space:** Affine registration of all volumes to the MNI152 standard brain template using ANTs or FSL FLIRT, enabling group-level analyses and ensuring consistent orientation.
4. **Quality control:** Automated quality control using tools such as MRIQC to flag scans with excessive motion artifacts, field inhomogeneity, or acquisition failures.
5. **Data augmentation:** During training, random affine transformations (rotation, translation, scaling) and intensity perturbations will be applied to increase data diversity and reduce overfitting.

8 Conclusion

This chapter has provided a systematic and critical comparison of the three major MRI datasets available for Alzheimer's disease research: OASIS, the Kaggle MRI dataset, and ADNI. Each dataset has a different profile of strengths and weaknesses shaped by its origin, intended audience, and collection methodology.

For the specific requirements of this project — segmentation of brain structures from 3D MRI volumes, followed by multi-class Alzheimer's stage classification — ADNI emerges as the clear and uniquely suitable choice. Its exceptional scale, multi-site diversity, gold-standard clinical labels, and provision of FreeSurfer segmentation ground truths make it the only dataset that satisfies all critical requirements simultaneously.

The Kaggle dataset, while convenient for rapid prototyping and classification baselines, is fundamentally unsuitable for a segmentation-based approach due to its 2D format and lack of ground-truth masks. OASIS, while scientifically rigorous and immediately accessible, is limited in scale and does not provide the segmentation ground truths needed to supervise our model.

Having established the data foundation of the project, the following chapter turns to the state of the art in deep learning methods for Alzheimer's detection and classification, with a focus on the segmentation-based paradigms that form the methodological core of our proposed system.

CHAPTER

III

STATE OF THE ART: DEEP LEARNING FOR ALZHEIMER'S CLASSIFICATION

Introduction

The application of deep learning to the classification of AD from structural MRI has generated a vast and rapidly growing body of literature since approximately 2014. Approaches differ primarily along one fundamental axis: the dimensionality of the input representation given to the network. Three paradigms have emerged and been extensively studied — pure **2D** methods, full **3D** volumetric methods, and the intermediate **2.5D** approach — each with distinct trade-offs in terms of computational cost, contextual information captured, and practical deployability. This chapter reviews representative published works in each category, analyzes their respective advantages and limitations, and positions the design choices of the present project within this landscape.

1 2D Approaches

1.1 Principle

In 2D DL approaches, three-dimensional MRI volumes are decomposed into individual two-dimensional slices — typically in the axial, coronal, or sagittal plane — which are then treated as independent images. A standard CNN architecture designed for natural image classification (VGG, ResNet, DenseNet, EfficientNet) is applied to each slice, and the slice-level predictions are aggregated (by majority voting, averaging, or a secondary classifier) to produce a subject-level diagnosis.

1.2 Key Published Works

1 (2019) — Single Slice CNN on ADNI

Basaia et al. [19] trained a CNN on single axial slices extracted from ADNI T1-weighted MRI for the binary classification of AD vs. Cognitively Normal (CN) and MCI vs. CN. The model achieved an accuracy of 99% for the AD vs. CN task and demonstrated that even a single informative slice, when selected appropriately from the hippocampal region, carries substantial discriminative information. Their work highlighted that the choice of which slice to use is critical and that naive random slice selection leads to significantly degraded performance.

> **Data leakage warning:** The 99% accuracy reported by Basaia et al. must be interpreted with extreme caution. The authors performed splitting at the *slice level* rather than the *subject level*, meaning that multiple slices from the same patient appeared in both training and test sets. Since adjacent slices of the same brain are nearly identical, the model effectively "memorizes" patient-specific anatomy rather than learning generalizable disease features. This constitutes a form of data leakage that artificially inflates reported performance. As demonstrated by Wen et al. [20] in their reproducibility study, subject-level splitting consistently yields 10–15% lower accuracy than slice-level splitting on the same data. **The results of Basaia et al. should not be taken as a reliable estimate of real-world generalization performance.**

2 (2017) — Multi-Class CNN

Farooq et al. [21] proposed a four-class deep CNN classifier directly on 2D axial slices from the OASIS dataset, targeting the categories: non-demented, very mild, mild, and moderate dementia. Using a custom lightweight architecture, they reported an overall accuracy of 98.8%.

> **Data leakage warning:** Similarly to Basaia et al., the 98.8% accuracy reported by Farooq et al. is subject to serious data leakage concerns. Their evaluation was conducted at the *slice level*: hundreds of 2D slices extracted from the same patient brain were distributed across both the training and test partitions. Under such conditions, a model can achieve near-perfect test accuracy simply by recognizing the patient-specific anatomy of a brain it has already seen during training, rather than learning the diagnostic features that distinguish disease stages. Furthermore, the severely imbalanced class distribution of the OASIS dataset (with only a handful of moderate dementia samples) was not corrected in their evaluation, further distorting the reported metric. **The 98.8% figure should not be used as a credible performance benchmark, and any comparison with this result in the literature must account for these methodological flaws.**

3 (2020) — Hierarchical Fully Convolutional Networks

Lian et al. [22] introduced a hierarchical approach in which a first network localizes ROIs (patches centered on the hippocampus, amygdala, and entorhinal cortex) from 2D slices, and a second network classifies AD vs. MCI vs. CN using only the selected patches.

Applied to ADNI data, this approach achieved 88.5% accuracy for the three-class problem, outperforming whole-slice baselines, while requiring far less computation than full-volume 3D methods.

1.3 Advantages of 2D Approaches

- **Computational efficiency:** 2D convolutions are computationally cheap and can be trained on standard consumer-grade Graphics Processing Unit (GPU)s without the memory constraints that limit 3D approaches.
- **Transfer learning compatibility:** Architectures pretrained on ImageNet (VGG, ResNet, DenseNet, EfficientNet) can be directly applied or fine-tuned, providing a strong initialization even with limited labeled medical data.
- **Large effective dataset size:** A cohort of 500 subjects may yield tens of thousands of 2D slices, dramatically increasing the apparent training set size.
- **Simplicity of implementation:** Standard computer vision tools and frameworks directly support 2D operations without modification.

1.4 Limitations of 2D Approaches

- **Loss of through-plane context:** Each 2D slice is processed independently, discarding the spatial relationships between adjacent slices. Alzheimer's atrophy is fundamentally a 3D phenomenon; a single slice captures at best a partial cross-section of affected structures.
- **Data leakage risk:** If slices from the same subject appear in both the training and test sets (slice-level rather than subject-level splitting), performance metrics are artificially inflated. Many published 2D papers do not report subject-level splitting, making their results non-comparable.
- **Slice selection sensitivity:** Performance depends heavily on which slice is selected. Uninformative slices (too peripheral, outside the hippocampal region) add noise; yet principled slice selection requires anatomical knowledge or an additional localization step.
- **Aggregation problem:** Combining hundreds of slice-level predictions into a single subject-level decision is non-trivial. Simple majority voting can be unstable when predictions are inconsistent across slices.

2 3D Approaches

2.1 Principle

Full 3D approaches treat the brain MRI as a volumetric input and apply 3D convolutional operations that learn spatial features simultaneously across all three anatomical axes. The canonical architecture for 3D medical image analysis is the 3D CNN, and its segmentation counterpart, the 3D U-Net, which processes entire brain volumes directly.

2.2 Key Published Works

1 Payan and Montana (2015) — 3D Sparse Autoencoder + CNN

One of the earliest deep learning studies on AD classification using full 3D volumes, Payan and Montana [23] used a 3D sparse autoencoder for unsupervised feature learning followed by a CNN classifier. Applied to ADNI data, their approach achieved 95.4% accuracy for AD vs. CN and 86.7% for MCI vs. CN, establishing an early baseline for 3D volumetric deep learning in this domain.

2 Korolev et al. (2017) — VoxCNN and ResNet Adapted to 3D

Korolev et al. [24] adapted the residual network (ResNet) architecture to 3D brain MRI data, processing full ADNI T1 volumes. They demonstrated that 3D ResNet could achieve competitive performance (89% accuracy for AD vs. CN) with relatively modest architectural depth, and argued that the residual connections were essential for stable training of deep 3D networks due to the gradient vanishing problem exacerbated in 3D settings.

3 Lian et al. (2018) — Landmark-based 3D Deep Learning

Lian et al. [25] proposed a multi-scale 3D CNN that first identifies anatomical landmarks (hippocampus, ventricles) in the 3D volume, then extracts and classifies local 3D patches centered on these landmarks. This ROI-guided approach achieved an accuracy of 90.2% for three-class classification on ADNI, outperforming whole-volume 3D baselines and demonstrating the benefit of anatomically informed feature extraction.

4 Wen et al. (2020) — Reproducible Benchmark

Wen et al. [20] published a systematic and reproducible benchmark of CNN-based AD classification on ADNI, comparing 2D, 3D, and patch-based approaches under carefully controlled experimental conditions (subject-level splits, no data leakage). Their results showed that 3D whole-volume CNNs achieved approximately 85% accuracy for the clinically important AD vs. MCI task, and emphasized that many published results in the literature are inflated due to methodological flaws, particularly data leakage.

5 Segmentation-Based 3D: Hett et al. (2019)

Hett et al. [26] proposed a graph-based framework operating on segmentation-derived features from the full 3D volume, achieving 88% accuracy for MCI conversion prediction on ADNI. Their work validated the segmentation-then-classification paradigm at scale, showing that morphometric features extracted from automatically segmented brain structures (using multi-atlas label propagation) are competitive with raw-pixel 3D CNN approaches while being far more interpretable.

2.3 Advantages of 3D Approaches

- **Full anatomical context:** 3D convolutions learn features from the complete spatial structure of the brain, capturing inter-regional relationships (e.g., hippocampal volume relative to ventricular size) that 2D slices cannot encode.

- **No slice selection required:** The entire volume is processed; there is no dependence on the choice of which slice to use.
- **Volumetric biomarkers:** 3D methods can naturally produce volumetric outputs (segmentation masks, activation maps) that are clinically interpretable.
- **Avoidance of intra-subject leakage:** Since each subject contributes exactly one input sample, subject-level splitting is straightforward.

2.4 Limitations of 3D Approaches

- **Memory and computational cost:** 3D convolutions over full brain volumes ($\sim 182 \times 218 \times 182$ voxels for MNI space) require GPUs with large memory (16–80 GB VRAM for full-volume 3D training). This severely restricts batch sizes, training speed, and the depth of architectures that can be used.
- **Limited transfer learning:** Pretrained 3D models on large natural image datasets do not exist. 3D networks must be trained from scratch or pretrained on other medical imaging tasks, requiring more data and longer training times.
- **Overfitting risk:** The high dimensionality of 3D inputs and the relatively small size of AD datasets (hundreds to low thousands of subjects) create significant overfitting risk. Regularization in 3D is technically more challenging.
- **Reduced architectural diversity:** The DL research community has produced far fewer 3D architectures than 2D ones. State-of-the-art vision transformers, EfficientNet variants, and advanced regularization techniques are predominantly designed for 2D.

3 The 2.5D Paradigm: A Principled Compromise

3.1 Principle and Motivation

The 2.5D approach is a family of methods designed to capture inter-slice anatomical context — which pure 2D methods miss — while retaining the computational efficiency and transfer learning advantages of 2D CNNs. In 2.5D processing, instead of feeding a single slice to the network, a *stack* of neighboring slices is assembled and treated as a multi-channel 2D image. The 2D CNN then implicitly learns features that span multiple adjacent slices through the channel dimension, effectively seeing a thin volumetric context without performing true 3D convolutions.

3.2 Key Published Works

1 Roth et al. (2014) — Multi-Scale 2.5D for Organ Localization

One of the earliest formulations of the 2.5D concept in medical imaging, Roth et al. [27] used orthogonal 2D patches from axial, coronal, and sagittal planes simultaneously to localize organs in CT volumes. Their key finding was that combining information from three planes substantially outperformed single-plane 2D approaches while remaining far cheaper than full 3D methods. This established the foundational rationale for 2.5D processing in volumetric medical image analysis.

2 Liu et al. (2018) — 2.5D Multi-View CNN for AD

Liu et al. [28] extended the 2.5D concept specifically to AD classification, training separate CNN branches on axial, coronal, and sagittal 2D projections of MRI volumes, then fusing the branch outputs with a fully connected layer. Applied to ADNI data, their multi-view approach achieved 91.1% accuracy for AD vs. CN, outperforming single-plane and single-slice baselines, demonstrating that complementary anatomical information is available across different viewing planes.

3 Spasov et al. (2019) — 2.5D with Attention for MCI Conversion

Spasov et al. [29] proposed a parameter-efficient 2.5D network with attention mechanisms specifically designed to predict conversion from MCI to AD over an 18-month follow-up period. Their model used stacked axial slices from the hippocampal region and achieved an AUC of 0.925 for conversion prediction, substantially outperforming 3D whole-volume baselines. Critically, their attention mechanism produced spatially interpretable feature maps that highlighted the hippocampus and entorhinal cortex as the most discriminative regions — consistent with the known neurobiology of AD.

4 Cheng et al. (2021) — 2.5D DenseNet for Multi-Class AD

Cheng et al. [30] applied a 2.5D DenseNet to ADNI T1-weighted MRI for three-class classification (AD / MCI / CN). Their approach stacked five adjacent axial slices as a 5-channel input, fine-tuned from ImageNet-pretrained DenseNet121 weights, and achieved 84.3% accuracy for the three-class problem with significantly faster training than equivalent 3D architectures. Their work is among the closest published references to the methodology adopted in this project.

3.3 Advantages of 2.5D Approaches

- **Inter-slice context without 3D cost:** The channel-stacking of neighboring slices provides the network with through-plane anatomical context while performing only 2D convolutions, dramatically reducing GPU memory requirements.
- **Full transfer learning compatibility:** ImageNet-pretrained 2D architectures can be adapted to 2.5D inputs by inflating the first convolutional layer from 3 channels (RGB) to N channels (the slice stack), preserving pretrained feature representations.
- **Plane flexibility:** The same architecture can be trained independently on axial, coronal, and sagittal planes, and the resulting models can be ensembled for improved performance.
- **Slice spacing as a hyperparameter:** The offset between selected slices can be tuned to capture either dense local context (small offsets) or wider anatomical context (larger offsets), providing an additional degree of freedom not available in 2D or 3D methods.

3.4 Limitations of 2.5D Approaches

- **Incomplete 3D modeling:** 2D convolutions applied to channel-stacked slices do not truly model 3D spatial relationships. The network treats depth as a channel dimension, which is not equivalent to a true volumetric convolution.
- **Plane dependency:** Models trained on axial slices may not generalize to coronal or sagittal views. Ensembling across planes adds complexity.
- **Correlated training samples:** Multiple overlapping slice stacks from the same subject may introduce intra-subject correlation in the training data, potentially inflating effective sample counts.

4 Comparative Summary

Table III.1: Comparison of 2D, 3D, and 2.5D approaches for Alzheimer's MRI classification

Criterion	**2D**	**3D**	**2.5D**
Contextual information	Slice-level only	Full volume	Limited through-plane
GPU memory required	Low	Very high	Low–Medium
Transfer learning	Excellent	Poor	Excellent
Training data needed	Low (many slices)	High	Medium
Leakage risk	High (slice split)	Low	Medium
Interpretability	Low	Medium	Medium–High
Best published AD accuracy (3-class)	~88%	~85–90%	~84–91%
Typical architecture	VGG, ResNet, EfficientNet	3D ResNet, 3D U-Net	DenseNet, ResNet + channel inflation
Plane flexibility	Single plane	N/A	Multi-plane possible

5 Positioning of the Present Work

Based on this review, the approach adopted in this project is a **2.5D pipeline** built on a **DenseNet121 backbone with CBAM attention**, trained on ADNI T1-weighted MRI data. This design is motivated by the following considerations derived from the literature:

- The 2.5D paradigm offers the best balance between anatomical context and computational tractability, as demonstrated by Spasov et al. and Cheng et al.
- DenseNet121 has shown consistently strong performance in medical image classification tasks due to its feature reuse properties, which are particularly beneficial when training data is limited.
- The addition of Convolutional Block Attention Module (CBAM) attention is motivated by Spasov et al.'s finding that attention mechanisms improve both performance and interpretability by focusing the model on anatomically relevant regions.
- Subject-level splitting — conspicuously absent in many published 2D studies — is enforced throughout, following the recommendations of Wen et al.'s reproducibility study.
- Training on multiple anatomical planes (axial, coronal, sagittal) independently allows assessment of which plane is most informative for each diagnostic distinction, extending the single-plane approach of most prior work.

The following chapter details the complete design and implementation of this pipeline.

6 Conclusion

This chapter has surveyed the landscape of deep learning approaches for AD classification from MRI, organized by the dimensionality of the input representation. Pure 2D methods offer computational simplicity and transfer learning benefits at the cost of contextual information and leakage risk. Full 3D methods capture complete volumetric context but are constrained by memory and data requirements. The 2.5D approach, capturing inter-slice context through channel-stacked multi-slice inputs, emerges as a principled and practically effective compromise. Published work consistently demonstrates that 2.5D methods match or exceed 3D baselines while remaining computationally tractable and fully compatible with ImageNet transfer learning. This motivates the 2.5D DenseNet121-CBAM design adopted in this project, which is described in full in the following chapter.

CHAPTER

IV

PROPOSED SYSTEM: A 2.5D PIPELINE FOR ALZHEIMER'S CLASSIFICATION

Introduction

This chapter presents the complete design and implementation of the proposed system for classifying Alzheimer's disease from ADNI T1-weighted MRI data. The system is structured as a sequential pipeline that transforms raw Digital Imaging and Communications in Medicine (DICOM) acquisitions into a standardized, preprocessed representation, selects informative brain slices, constructs 2.5D multi-channel inputs, and trains a DenseNet121 classifier augmented with CBAM attention. Each stage of the pipeline is described in detail, with explicit justification for each design decision grounded in the literature reviewed in the previous chapter and in the specific characteristics of the ADNI dataset. Figure **??** provides a schematic overview of the complete pipeline.

The system targets a **3-class classification problem**:

- **AD** (Alzheimer's Disease) — label 0
- **CN** (Cognitively Normal) — label 1
- **MCI** (Mild Cognitive Impairment) — label 2

The input modality is exclusively T1-weighted MPRAGE structural MRI, selected for its established role as the primary structural biomarker of AD-related atrophy (see Chapter I). The dataset and its selection rationale were presented in Chapter II.

1 Dataset Curation and Subject Selection

1.1 ADNI Metadata Processing

The pipeline begins with a structured metadata loading step that converts the raw ADNI export — which typically contains multiple entries per subject corresponding to different scan sessions, acquisition variants, and scanner configurations — into a canonical *one-row-per-subject* table. This normalization step is essential because the downstream preprocessing and splitting stages must operate at the subject level, not the scan level, to avoid data leakage.

The metadata loader performs the following operations in sequence:

1. **Variant filtering:** Only scans identified as T1-weighted MPRAGE sequences are retained. Other structural sequences (T2, FLAIR, PD) are excluded.
2. **Label mapping:** Diagnostic labels are mapped to numeric class indices: AD → 0, CN → 1, MCI → 2.
3. **DICOM series localization:** For each subject entry, the corresponding DICOM series is located on disk using the image data identifier.
4. **Scanner metadata extraction:** Manufacturer name, repetition time (TR), echo time (TE), and voxel spacing are read from the DICOM header.
5. **Canonical series selection:** When multiple candidate scans exist for the same subject (e.g., from different sessions or field strengths), a single series is selected by sorting on subject ID, acquisition date, variant priority, number of DICOM files, and image data identifier, and retaining the last entry. In practice this selects the highest-quality available acquisition for each subject.

1.2 Dataset Composition

After curation, the dataset used in this project comprises **983 subjects** from the ADNI database, distributed across the three diagnostic classes as follows:

Table IV.1: Subject composition of the curated ADNI dataset

Split	**AD**	**CN**	**MCI**	**Total**
Training (70%)	237	226	224	687
Validation (15%)	51	49	48	148
Test (15%)	50	49	49	148
Total	**338**	**324**	**321**	**983**

The near-uniform distribution across classes (338 AD, 324 CN, 321 MCI) reflects the stratified selection strategy, which explicitly balances class counts across splits. This balanced distribution mitigates class imbalance as a primary confound, though class-weighted loss is still applied during training as an additional safeguard (see Section 6.2).

2 Preprocessing Pipeline

The preprocessing stage transforms raw DICOM acquisitions into standardized, brain-only volumetric representations suitable for model training. The pipeline consists of eleven sequential steps, each designed to remove a specific source of variability that would otherwise confound the classifier.

2.1 Sequence Verification

Before any processing, the pipeline inspects the DICOM metadata of each series and generates warnings when the acquisition parameters suggest a non-standard sequence. Specifically, it checks:

- Whether the repetition time (TR) is $\leq$ 1800 ms (atypically short for MPRAGE)
- Whether the echo time (TE) is > 10 ms (atypically long for T1-weighted)
- Whether the series description mentions "MPRAGE" or "MP-RAGE"

This step does not automatically exclude subjects; it logs warnings to a quality control file for subsequent manual review. This design choice reflects the practical reality of large multi-site datasets such as ADNI, where some acquisitions deviate from protocol without necessarily being unusable.

2.2 DICOM Loading and Volume Assembly

Each subject's selected DICOM series is read using SimpleITK, which handles the assembly of individual DICOM slices into a coherent 3D volume while preserving the spatial metadata (voxel spacing, image orientation, origin). The output is a 3D SimpleITK image object.

2.3 Reorientation to RAS

The raw MRI volume is reoriented to the standard neurological RAS (Right-Anterior-Superior) coordinate system. This ensures that the left–right, anterior–posterior, and superior–inferior anatomical axes correspond consistently to the same array dimensions across all subjects, regardless of the scanner-specific coordinate conventions used at acquisition. Without this step, axial, coronal, and sagittal slices extracted at fixed indices would correspond to different anatomical levels in different subjects.

2.4 N4 Bias Field Correction

MRI acquisitions suffer from low-frequency intensity non-uniformity caused by radiofrequency field inhomogeneity and coil-sensitivity variations. This bias field causes the same tissue type (e.g., white matter) to appear with different intensity values in different parts of the image, which can mislead intensity-based features and normalization procedures.

The N4 bias field correction algorithm [31] is applied to estimate and remove this smooth multiplicative field. An Otsu-based foreground mask is first computed to restrict the estimation to brain voxels, and the N4 filter is applied with the following parameters:

- Iterations per resolution level: [100, 100, 100, 100]
- Convergence threshold: 10^{-3}

2.5 Robust Intensity Normalization

After N4 correction, intensities are normalized to the [0, 1] range using percentile-based clipping. This step is applied before skull stripping, in native space, to produce a consistent intensity scale across subjects scanned on different hardware.

The clipping is performed only on positive-valued (brain-containing) voxels. Default clipping percentiles are:

- Lower percentile: 0.5
- Upper percentile: 99.8

Because ADNI data are acquired on scanners from multiple manufacturers (Siemens, GE, Philips), scanner-specific adjustments are applied:

- Accelerated / parallel-imaging variants and Philips scanners: upper percentile reduced to 99.5 to compensate for their characteristic intensity distribution tails.
- GE scanners: lower percentile increased to 1.0 and upper percentile capped at 99.6.

These manufacturer-specific adjustments address the known inter-site intensity variability in ADNI and improve the comparability of normalized volumes across scanners.

2.6 Skull Stripping

Non-brain tissue (skull, scalp, dura mater, eyes) is removed using HD-BET [32], a deep learning-based brain extraction tool that has demonstrated superior performance over classical atlas-based methods (FSL BET) on heterogeneous multi-site data. HD-BET is run in *accurate* mode, and GPU acceleration is used when available. The output consists of:

- The skull-stripped brain image in native space
- A binary brain mask

The brain mask is propagated through all subsequent steps to ensure that all downstream operations (registration, normalization) are restricted to brain voxels.

2.7 Non-linear Registration to MNI Space

The skull-stripped brain is registered to the MNI152 standard brain template using ANTs SyN (Symmetric Normalization) [33], a state-of-the-art non-linear registration algorithm. Registration to a common anatomical reference space is essential for the 2.5D approach because it ensures that slices extracted at the same index correspond to the same anatomical level across all subjects.

The registration configuration is:

- Fixed image: MNI152 T1 template

- Moving image: skull-stripped subject brain (native space)
- Transform type: SyN (non-linear)
- Affine metric: Mattes mutual information
- Deformable metric: Mattes mutual information
- Image interpolator: B-spline
- Mask interpolator: Nearest neighbor
- Registration iterations: (100, 70, 50, 20) across resolution levels

All transforms (affine matrix, forward warp field, inverse warp field) are saved for each subject, enabling future reverse-mapping of model predictions back to native space if needed for clinical reporting.

2.8 Resampling to Isotropic Spacing

After registration, voxel spacing is verified. If the registered volume is not already at $1\times1\times1$ mm isotropic resolution, it is resampled to this standard spacing:

- Image: cubic interpolation (order 3) to preserve smooth intensity gradients
- Mask: nearest-neighbor interpolation (order 0) to preserve binary values

Isotropic spacing ensures that slice thickness is consistent across planes (axial, coronal, sagittal), which is a prerequisite for fair multi-plane comparison.

2.9 Tight Brain Cropping

The registered and resampled volume is cropped using the brain mask. A tight 3D bounding box is computed around all non-zero brain mask voxels, and a margin of 6 voxels is added in each direction to preserve a small amount of peribrain context. This step removes large empty background regions while keeping the full brain anatomy, and reduces the spatial dimensions of the volume significantly, thereby reducing memory requirements in subsequent steps.

2.10 Resizing and Padding to Fixed Shape

The cropped volume is resized and padded to a fixed target shape of $128 \times 128 \times 128$ voxels. The resizing uses isotropic scaling with the minimum scale factor needed to fit the target dimensions. After resizing:

- Zero-padding is applied symmetrically where the resized volume is smaller than the target shape.
- Center-cropping is applied where the resized volume marginally exceeds the target shape.

The mask is resized using nearest-neighbor interpolation with the same scale factor. The fixed $128 \times 128 \times 128$ shape provides a consistent input size for slice extraction and ensures that all subjects contribute slices at comparable anatomical scales.

2.11 Final Mask-Based Z-Score Normalization

The final intensity standardization step applies z-score normalization restricted to brain voxels:

$$V_{\text{norm}} = \frac{V - \mu_{\text{brain}}}{\sigma_{\text{brain}}} \tag{IV.1}$$

where μ_{brain} and σ_{brain} are the mean and standard deviation of voxel intensities computed exclusively within the binary brain mask. All voxels outside the mask are set to zero. This mask-based normalization prevents background voxels from influencing the statistics and ensures that the final intensity distribution is centered at zero with unit variance within the brain region.

2.12 Preprocessing Outputs and Quality Control

For each subject, the preprocessing stage produces and saves:

- Normalized 3D brain volume as .npy array
- Binary brain mask as .npy array
- Geometry metadata (voxel spacing, affine matrix, bounding box) as JSON
- Registration transforms (affine, forward warp, inverse warp)

At the dataset level, the following quality control artifacts are generated:

- preprocessing_log.csv: per-subject processing status and warnings
- normalization_stats.csv: per-subject brain volume, mean, and std
- QC summary JSON: aggregated statistics and failure counts
- Brain volume distribution plots: subjects with brain volumes outside the range 900–1800 cm^3 are flagged as potential preprocessing failures
- Class-specific mean intensity maps: used to verify that the preprocessing has not introduced systematic class-specific artifacts

3 Subject-Level Data Splitting

A fundamental methodological requirement for any medical image classification study is the **subject-level** enforcement of train/validation/test splits. This means that all slices, patches, or representations derived from a single subject must be assigned exclusively to one partition. Violating this constraint — by splitting at the slice or image level — causes data leakage: the model sees anatomically near-identical slices of the same brain in both training and test sets, producing artificially inflated metrics that do not reflect true generalization.

This project applies stratified subject-level splitting with the following ratios:

- Training: 70% (687 subjects)
- Validation: 15% (148 subjects)

- Test: 15% (148 subjects)

The stratification is performed jointly on **diagnostic label and scanner manufacturer**. This dual stratification ensures that:

1. Class proportions are preserved across splits (preventing by-chance class imbalance in the test set).
2. Scanner manufacturer distribution is balanced across splits (preventing the model from learning manufacturer-specific artifacts rather than disease-related features).

In addition, the pipeline supports 5-fold stratified cross-validation for more robust performance estimation, allowing the full dataset to be used for training across folds.

4 2.5D Slice Selection and Input Construction

4.1 Rationale for Slice Selection

After preprocessing, each subject is represented by a $128 \times 128 \times 128$ normalized brain volume. Rather than passing the entire volume to the network (as in full 3D approaches), the 2.5D pipeline selects a subset of informative slices from one anatomical plane at a time. The pipeline supports three planes:

- **Axial** — slices along the superior–inferior axis
- **Coronal** — slices along the anterior–posterior axis
- **Sagittal** — slices along the left–right axis

A separate model is trained for each plane, allowing comparison of the discriminative power of different anatomical views.

4.2 Brain-Mask-Based Slice Scoring

Not all slices within a plane are equally informative. Peripheral slices (at the top of the skull or at the base of the brainstem) contain little hippocampal or cortical tissue and are dominated by noise or cerebrospinal fluid. To identify informative slices automatically, the pipeline uses the brain mask to score each slice by the number of non-zero brain voxels it contains (its *slice area*).

The selection procedure is:

1. Compute slice area (non-zero mask voxel count) for every slice in the plane.
2. Retain only slices with positive area (fully out-of-brain slices are discarded).
3. Compute the 70th percentile of the slice area distribution.
4. Select all slices whose area meets or exceeds this threshold.

If this percentile-based rule returns fewer than 25 slices (the configured minimum), the pipeline falls back to a central window of 60 slices centered on the brain volume center. This ensures a sufficient number of candidate slices for all subjects regardless of brain size variability.

4.3 2.5D Stack Construction

For each selected center slice, the pipeline constructs a 5-channel 2.5D input by stacking the center slice together with four neighboring slices. Two spacing configurations are implemented:

- **Option A (dense context):** offsets = $[-2, -1, 0, +1, +2]$
- **Option B (wide context):** offsets = $[-4, -2, 0, +2, +4]$

For a given center index c and plane, the 2.5D input tensor of shape $5 \times 128 \times 128$ is constructed as:

$$\text{Axial:} \quad T[i, :, :] = V[c + \delta_i, :, :] \tag{IV.2}$$

$$\text{Coronal:} \quad T[i, :, :] = V[:, c + \delta_i, :] \tag{IV.3}$$

$$\text{Sagittal:} \quad T[i, :, :] = V[:, :, c + \delta_i] \tag{IV.4}$$

where δ_i is the i-th offset and V is the preprocessed $128 \times 128 \times 128$ volume. Only center slices for which all five offset indices remain within the volume bounds are included as valid training samples.

A center slice can therefore only be used if all required neighboring slices exist within the [0, 127] range, preventing boundary artifacts.

5 Model Architecture

5.1 DenseNet121 Backbone

The backbone of the proposed classifier is **DenseNet121** [34], a convolutional architecture characterized by dense connectivity: each layer receives feature maps from all preceding layers within a dense block, and passes its own feature maps to all subsequent layers. This design promotes maximum feature reuse and enables gradient flow through very short paths, mitigating the vanishing gradient problem that hampers deep networks.

DenseNet121 is particularly suited to medical image classification with limited data because:

- Dense connections reduce the number of parameters needed to achieve a given level of representational capacity, reducing overfitting risk.
- Feature reuse allows the network to combine low-level texture features with high-level semantic features, which is valuable for detecting subtle atrophy patterns.
- DenseNet121 is one of the architectures with consistently strong performance across medical imaging benchmarks [35].

5.2 First-Layer Channel Inflation for 5-Channel Input

The standard DenseNet121 accepts 3-channel (RGB) inputs. To accommodate the 5-channel 2.5D slice stack, the first convolutional layer is *inflated*: its weight tensor is

extended from shape $[C_{out}, 3, 7, 7]$ to $[C_{out}, 5, 7, 7]$ by copying the pretrained 3-channel weights and rescaling them to conserve the overall weight magnitude:

$$W_{\text{inflated}} = \frac{3}{5} \cdot \left[W_{\text{pretrained}} - W_{\text{pretrained,mean}} \right] \quad \text{(IV.5)}$$

This inflation strategy allows the model to benefit from ImageNet pretraining even though the input is multi-channel grayscale rather than RGB, following the approach validated by Spasov et al. [29] and Cheng et al.

5.3 CBAM Attention Modules

CBAM (Convolutional Block Attention Module) [36] is inserted after each of DenseNet121's transition blocks. CBAM applies a two-stage sequential attention:

1 Channel Attention

The channel attention module recalibrates the relative importance of each feature map channel. It applies both global average pooling and global max pooling to the feature tensor, passes both outputs through a shared Multi-Layer Perceptron (MLP), and produces a channel-wise attention vector that is multiplied with the original feature maps:

$$M_c(F) = \sigma\left(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))\right) \quad \text{(IV.6)}$$

2 Spatial Attention

The spatial attention module identifies which spatial locations in the feature maps are most informative. It applies channel-wise average pooling and max pooling to the attention-refined feature maps, concatenates the two resulting single-channel maps, and applies a convolution to produce a spatial attention map that is multiplied element-wise with the feature maps:

$$M_s(F) = \sigma\left(f^{7\times7}([\text{AvgPool}(F); \text{MaxPool}(F)])\right) \quad \text{(IV.7)}$$

The motivation for CBAM in this context is well-aligned with the neurobiology of AD: Alzheimer's-related atrophy is both region-specific (affecting the hippocampus and entorhinal cortex preferentially) and intensity-structure dependent (characterized by thinning and contrast reduction in affected regions). Channel attention learns to prioritize feature maps encoding intensity changes indicative of atrophy, while spatial attention learns to focus on the hippocampal and medial temporal regions most affected by the disease.

5.4 Classification Head

Following the final CBAM-modulated feature maps, global average pooling reduces the spatial dimensions to a single feature vector. A dropout layer (rate = 0.2) is applied for regularization, followed by two fully connected layers:

- FC1: features → 256 units, ReLU activation
- FC2: 256 → 3 units (one per class), linear activation

The output logits are passed to the loss function during training, and to a softmax during inference to produce class probabilities.

6 Training Protocol

6.1 Data Augmentation

Training-time data augmentation is implemented using the Albumentations library and applied independently to each 2.5D slice stack. Validation and test sets receive no augmentation. The augmentation pipeline includes:

- Random rotation: up to $\pm 15°$
- Random scaling: up to $\pm 10\%$
- Gaussian noise injection
- Random brightness and contrast adjustment
- Random gamma correction
- Coarse dropout (random rectangular patches set to zero)
- Grid distortion (local elastic-like deformations)

These augmentations simulate the moderate anatomical and acquisition variability found across ADNI sites while preserving the overall brain structure and diagnostic signal. Aggressive geometric augmentations (e.g., flipping) are avoided because left–right anatomical asymmetry carries diagnostic information in AD (hippocampal atrophy is often asymmetric).

6.2 Loss Function and Class Imbalance

The primary loss function is **weighted cross-entropy**. Class weights are computed from the inverse of class frequencies in the training split, so that minority classes (if any imbalance remains after stratified splitting) contribute proportionally to the loss. An alternative focal loss configuration (focal $\gamma = 2.0$) is also implemented and evaluated in the experiments.

Two additional regularization techniques are applied at the loss level:

Label smoothing ($\epsilon = 0.1$): The one-hot target vector is softened so that the true class receives probability $1 - \epsilon$ rather than 1, and each other class receives $\epsilon/(K - 1)$ where $K = 3$. Label smoothing prevents the model from becoming overconfident in its predictions and has been shown to improve calibration and generalization.

Mixup augmentation ($\alpha = 0.3$): During training, pairs of samples (x_i, y_i) and (x_j, y_j) are linearly combined with a random mixing coefficient $\lambda \sim \text{Beta}(\alpha, \alpha)$:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j \tag{IV.8}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j \tag{IV.9}$$

Mixup acts as a convex combination regularizer that encourages linear behavior in between training examples and has been shown to improve generalization in multi-class settings.

6.3 Optimizer and Learning Rate Schedule

The optimizer is **AdamW** with the following hyperparameters:

- Learning rate: 5×10^{-5}
- Weight decay: 10^{-4}
- Batch size: 64

The learning rate schedule consists of two phases:

1. **Linear warmup** for 5 epochs: the learning rate increases linearly from 0 to 5×10^{-5}. Warmup prevents the large gradient updates typical of cold-start training from destabilizing the pretrained weights in the early epochs.
2. **Cosine annealing** for the remaining epochs: the learning rate decreases following a cosine curve toward 0, providing fine-grained convergence in later training.

6.4 Mixed Precision and Distributed Training

Training is performed in **mixed precision** (fp16) using PyTorch's Automatic Mixed Precision (AMP), which stores activations in 16-bit floating point while maintaining 32-bit master weights for the gradient update. This halves the GPU memory footprint and accelerates matrix operations on modern GPU hardware without significant numerical accuracy loss.

When multiple GPUs are available, the training launches in **distributed data parallel** (DDP) mode, with one process per GPU. Gradient synchronization across processes is handled by PyTorch DDP's all-reduce communication.

6.5 Stochastic Weight Averaging

Starting at epoch 60, **Stochastic Weight Averaging** (SWA) is applied. SWA maintains a running average of model weights sampled at regular intervals from the later part of training. At the end of training, batch normalization statistics are re-computed on the training set using the SWA weights, and the SWA model checkpoint is saved separately alongside the standard best-model checkpoint. SWA is known to find flatter minima in the loss landscape, which tends to improve generalization on held-out data.

6.6 Early Stopping and Model Selection

Training runs for a maximum of 100 epochs. Early stopping is applied with a patience of 15 epochs: if the validation macro F1-score does not improve for 15 consecutive epochs, training is terminated. The best model checkpoint (best.pt) is saved whenever the validation macro F1-score improves.

Macro F1-score is used as the primary model selection criterion rather than accuracy. In a 3-class medical classification problem where all classes are clinically important (misclassifying MCI as CN is as consequential as misclassifying AD as CN), macro F1 gives equal weight to each class regardless of size, making it a more appropriate metric than accuracy for model selection.

6.7 Hyperparameter Summary

Table IV.2: Summary of training hyperparameters

Hyperparameter	**Value**
Backbone	DenseNet121 (ImageNet pretrained)
Input channels	5 (2.5D slice stack)
Input spatial size	128×128
Number of classes	3 (AD, CN, MCI)
Batch size	64
Learning rate	5×10^{-5}
Weight decay	10^{-4}
Epochs (maximum)	100
Warmup epochs	5
LR schedule	Linear warmup + cosine annealing
Label smoothing	0.1
Mixup α	0.3
Dropout rate	0.2
Loss function	Weighted cross-entropy (default) / Focal loss ($\gamma = 2.0$)
Early stopping patience	15 epochs
Model selection metric	Macro F1-score (validation)
SWA start epoch	60
Mixed precision	fp16 (enabled)
Random seed	42

7 Experimental Configurations

To systematically evaluate the 2.5D design choices, six primary experimental configurations are evaluated, corresponding to the combination of three anatomical planes and two slice spacing options:

Table IV.3: Experimental configurations for the 2.5D pipeline

Run ID	**Plane**	**Spacing**	**Description**
Axial-A	Axial	Dense (± 2)	Axial view, 5 contiguous slices
Axial-B	Axial	Wide (± 4)	Axial view, 5 slices spaced by 2
Coronal-A	Coronal	Dense (± 2)	Coronal view, 5 contiguous slices
Coronal-B	Coronal	Wide (± 4)	Coronal view, 5 slices spaced by 2
Sagittal-A	Sagittal	Dense (± 2)	Sagittal view, 5 contiguous slices
Sagittal-B	Sagittal	Wide (± 4)	Sagittal view, 5 slices spaced by 2
Baseline-2D	Axial	Single (center)	Single-channel 2D baseline

The single-channel 2D baseline (center axial slice only, simple CNN backbone) serves as a lower-bound reference to quantify the benefit of 2.5D multi-slice context and CBAM attention over naive single-slice classification.

8 Evaluation Metrics

The following metrics are computed on the held-out test set and reported for each experimental configuration:

- **Overall accuracy:** proportion of correctly classified subjects.
- **Macro F1-score:** unweighted mean of per-class F1-scores; the primary metric.
- **Per-class F1-score:** F1 for each of AD, CN, and MCI individually.
- **Per-class recall:** sensitivity for each class, i.e., the ability to correctly identify positive cases.
- **Confusion matrix:** a 3×3 matrix displaying predicted vs. true class counts, revealing systematic misclassification patterns.
- **AUC-ROC:** macro-averaged one-vs.-rest area under the ROC curve, measuring discriminability independently of the classification threshold.

During training, a *degenerate prediction flag* is monitored: if the model predicts more than 90% of validation samples as a single class, the training run is flagged as collapsed and excluded from the comparative analysis. This safeguard prevents cases where the model trivially maximizes accuracy by predicting the majority class.

9 Limitations of the Proposed Approach

The following limitations of the proposed pipeline are acknowledged:

- **Plane-specific models:** each model sees one anatomical plane at a time. While multi-plane ensembling is possible, it adds inference complexity. A true 3D model would jointly learn features across all planes simultaneously.
- **Intra-subject slice correlation:** multiple overlapping 2.5D slice stacks from the same subject share anatomical context. Although subject-level splitting prevents cross-split leakage, this within-training-set correlation may slightly inflate the effective number of independent training samples.
- **Preprocessing dependency:** the quality of all downstream steps depends on successful skull stripping and non-linear registration. Failures in HD-BET or ANTs SyN (caused by severely atypical brain morphology or image artifacts) propagate to the classifier.
- **Slice selection bias:** selecting slices based on brain-mask area favors the large central regions of the brain (where the mask area is maximum) over peripheral regions. Some diagnostic biomarkers located in smaller peripheral structures may receive less representation in the training data.
- **ADNI-specific generalizability:** the model is trained on a predominantly white, educated, North American cohort. Generalization to more diverse populations (including Algerian patients) would require domain adaptation or transfer learning on local data.

10 Pipeline Overview: Visual Summary

To provide a comprehensive visual summary of the complete system, Figures IV.1 and IV.2 present the two major phases of the proposed pipeline as structured diagrams.

10.1 MRI Preprocessing Pipeline

Figure IV.1 illustrates the eleven-step preprocessing pipeline described in Section 2 of this chapter. The pipeline is organized into four functional groups: metadata inspection and volume assembly (DICOM QC, loading, and RAS reorientation), signal correction (N4 bias field correction, intensity normalization, skull stripping), spatial standardization (MNI registration, isotropic resampling, tight brain cropping, resizing to $128 \times 128 \times 128$), and final output generation (z-score normalization and quality control artifacts). Each block is color-coded by its functional role, providing a clear visual map of the data flow from raw DICOM acquisition to the standardized volumetric representation consumed by the classifier.

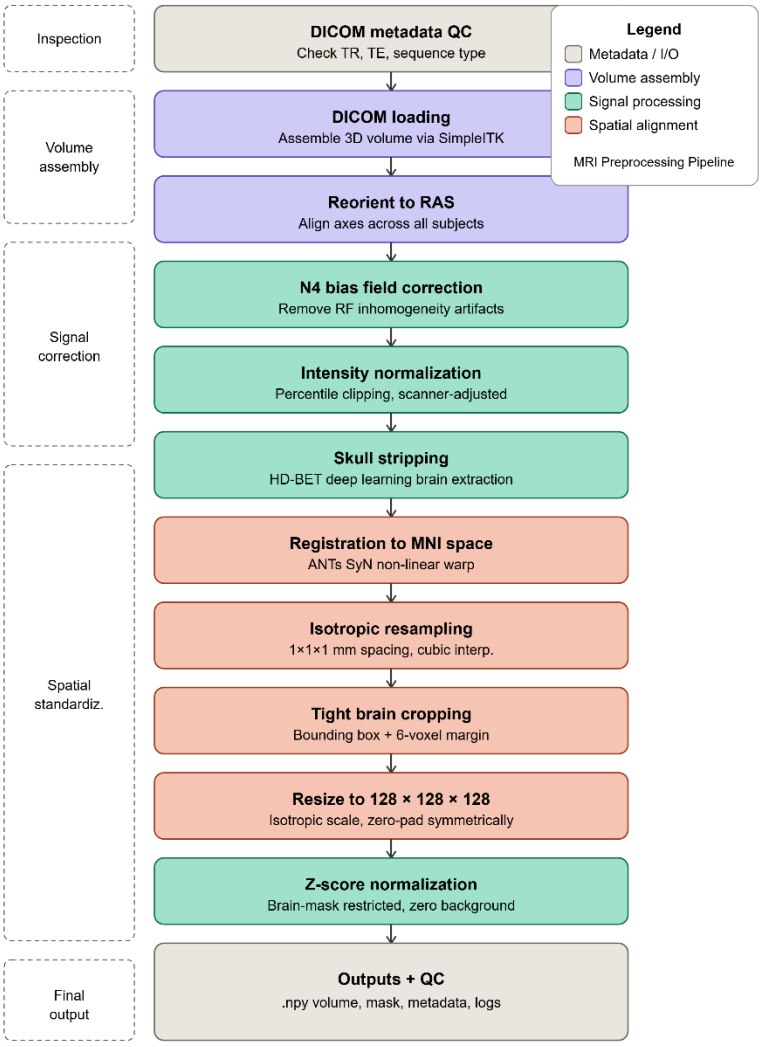

Figure IV.1: Visual overview of the 11-step MRI preprocessing pipeline. Color coding indicates the functional role of each step: grey for metadata/I/O operations, blue for volume assembly, green for signal correction, and red/orange for spatial standardization.

The pipeline transforms raw DICOM acquisitions into normalized $128 \times 128 \times 128$ brain volumes ready for 2.5D slice extraction.

10.2 Training Pipeline

Figure IV.2 illustrates the complete training pipeline from 2.5D input construction through to model selection. The diagram is organized into five functional phases: input preparation (2.5D slice extraction and data augmentation), architecture (channel inflation, DenseNet121 backbone, CBAM attention, and classification head), optimization (loss function, AdamW with cosine schedule, and mixed precision/DDP), and training control (stochastic weight averaging and early stopping). The classification head shows the final configuration: Global Average Pooling → Dropout(0.2) → FC(256) → FC(3), consistent

with the updated hyperparameters described in Section 6.

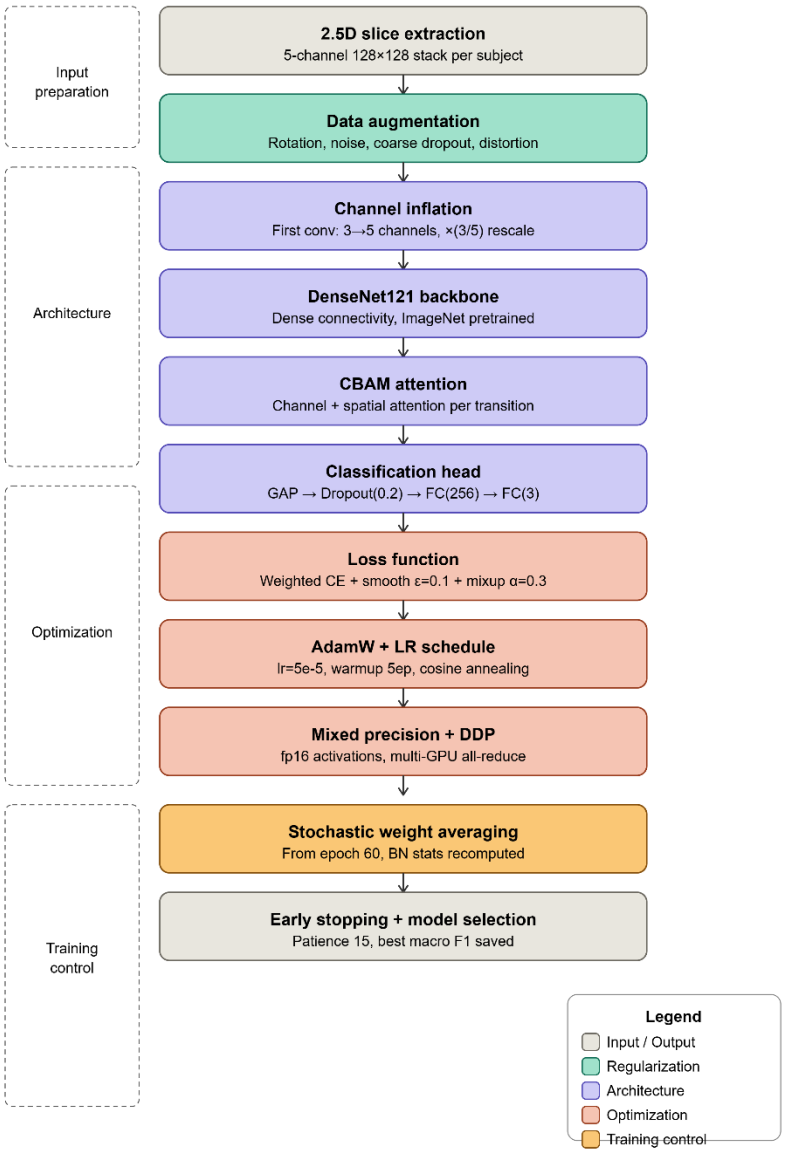

Figure IV.2: Visual overview of the training pipeline. The diagram covers the full flow from 2.5D slice extraction (5-channel 128×128 stack) through data augmentation, channel inflation, DenseNet121-CBAM architecture, and optimization (AdamW, lr=5×10^{-5}, cosine annealing, mixup $\alpha = 0.3$) to training control mechanisms (SWA from epoch 60, early stopping on macro F1).

11 Conclusion

This chapter has presented the complete design and implementation of the proposed 2.5D pipeline for Alzheimer's disease classification from ADNI T1-weighted MRI. The pipeline integrates eleven standardized preprocessing steps (reorientation, N4 correction, robust normalization, skull stripping, MNI registration, isotropic resampling, cropping, resizing, and mask-based z-score normalization), subject-level stratified splitting, brain-mask-guided slice selection, 5-channel 2.5D input construction, and a DenseNet121-CBAM clas-

sifier trained with a robust protocol including augmentation, label smoothing, mixup, cosine scheduling, SWA, and early stopping on macro F1.

Each design decision has been grounded in the limitations of prior 2D and 3D approaches reviewed in Chapter III, in the characteristics of the ADNI dataset documented in Chapter II, and in the neurobiology of Alzheimer's disease established in Chapter I. The following chapter presents the experimental results obtained by this pipeline across the six plane/spacing configurations and the 2D baseline.

CHAPTER

V

EXPERIMENTS, RESULTS, AND DISCUSSION

Introduction

This chapter presents the experimental results obtained by the 2.5D DenseNet121-CBAM pipeline described in Chapter IV, evaluated on the held-out test set of the ADNI dataset. The results follow the natural progression of the pipeline: first the dataset class distribution is verified, then training dynamics are analyzed, and finally held-out test performance is examined through accuracy, per-class F1-scores, confusion matrices, and ROC curves. Each result is interpreted in light of the clinical significance of the classification task and the methodological choices of the preceding chapters.

1 Dataset Composition and Class Distribution

1.1 Verified Split Distribution

Before training, the stratified subject-level split described in Chapter IV was verified by inspecting the class distribution across the three partitions. Figure V.1 shows the resulting class counts.

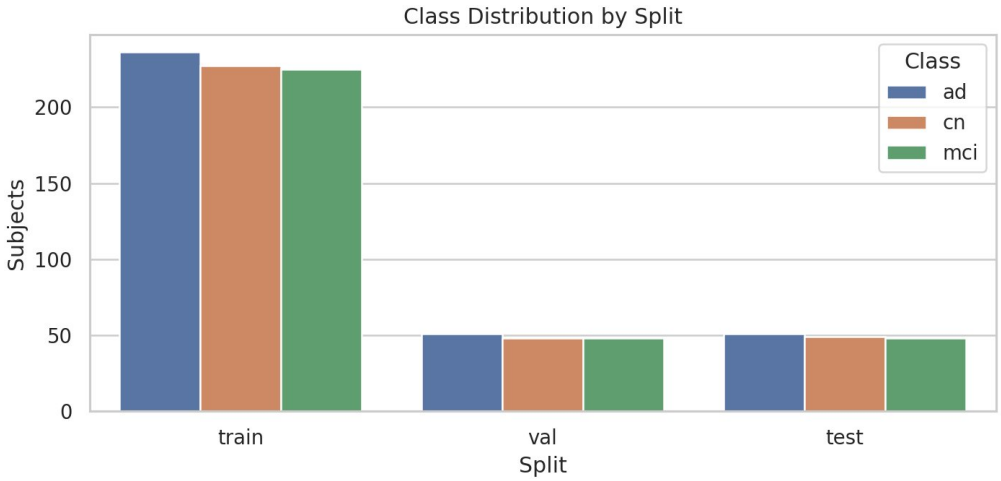

Figure V.1: Class distribution by split (train / validation / test) across the three diagnostic categories: AD, CN, and MCI. The near-uniform distribution across classes within each split confirms the effectiveness of the stratified sampling strategy.

1.2 Interpretation

The bar chart confirms that the stratified splitting strategy successfully preserved class balance across all three partitions. In the training set, the three classes (AD: 237, CN: 226, MCI: 224) differ by at most 13 subjects — a maximum relative imbalance of less than 6%. The validation and test sets exhibit similarly tight balance (AD: 51/50, CN: 49/49, MCI: 48/49).

This balance has two important consequences. First, class-weighted loss weights are close to uniform, meaning the model is not trained primarily to satisfy a dominant class. Second, since the test set is approximately balanced, overall accuracy is a meaningful metric (not inflated by a majority class), and the gap between macro and weighted F1 will reflect true differential performance across classes rather than sampling artefact.

The dual stratification on diagnosis label and scanner manufacturer ensures that the test set includes subjects from all major ADNI scanner types (Siemens, GE, Philips), making the evaluation representative of the full multi-site variability of the dataset.

2 Training Dynamics

2.1 Loss and Macro F1 Curves

Figure V.2 shows the evolution of training and validation loss, macro F1-score, and per-class validation F1-score over 100 training epochs.

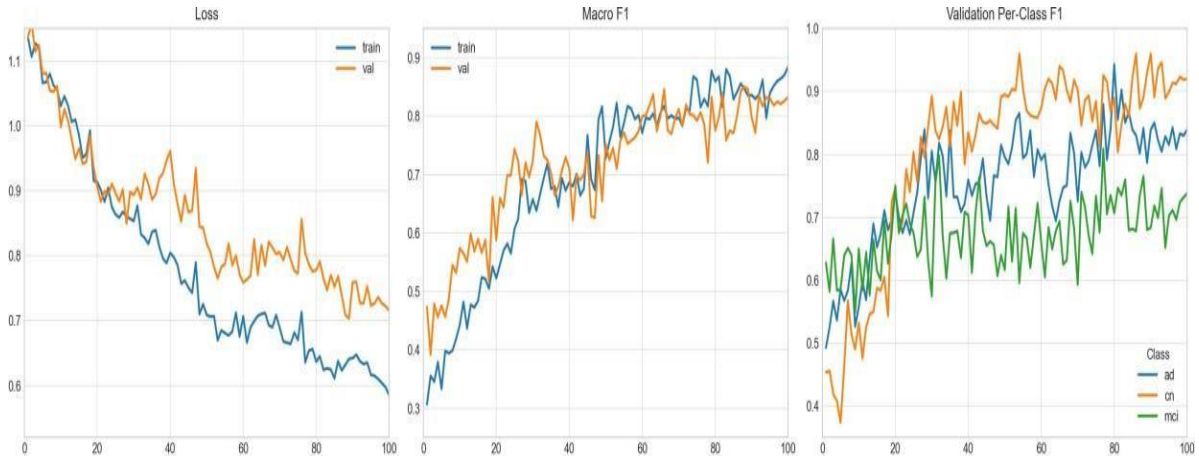

Figure V.2: Training dynamics over 100 epochs. *Left:* Training and validation loss (weighted cross-entropy). *Center:* Training and validation macro F1-score. *Right:* Per-class validation F1-score for AD (blue), CN (orange), MCI (green).

2.2 Interpretation of Loss Curves

The loss curves (left panel) exhibit a consistent downward trend throughout training. Training loss decreases from approximately 1.1 at epoch 0 to approximately 0.6 by epoch 100. The validation loss follows a similar trajectory at a higher level (approximately 0.7–0.85 in later epochs), which is expected and represents the generalization gap of any supervised learning system.

Critically, the validation loss shows no sharp upward inflection — the classic signature of overfitting. While there is a slight divergence between training and validation loss in later epochs, its absence of pronounced increase indicates that the regularization strategy (dropout, label smoothing, mixup, weight decay, and Stochastic Weight Averaging (SWA)) was effective. The oscillations in validation loss visible between epochs 30 and 70 are attributable to the limited validation set size (148 subjects), where loss estimates are inherently noisier.

2.3 Interpretation of Macro F1 Curves

Training macro F1 rises steadily from approximately 0.35 at epoch 0 to 0.90 by epoch 100. Validation macro F1 rises rapidly in the first 30 epochs (from 0.30 to approximately 0.70), then continues more gradually to a plateau in the 0.80–0.85 range. The gap between training (≈0.90) and validation (≈0.83) macro F1 is consistent with the expected generalization gap and is not indicative of severe overfitting given the dataset size. Both curves continue rising throughout 100 epochs without divergence.

2.4 Per-Class Validation F1 and Class Difficulty

The per-class validation F1 curves (right panel) reveal important class-specific differences:

- **CN (orange):** Achieves the highest and most stable per-class F1, rising rapidly to 0.85–0.95 by mid-training. Cognitively normal subjects have well-preserved brain volumes with no structural hallmarks of neurodegeneration, making them the least ambiguous class.
- **AD (blue):** Reaches approximately 0.80–0.85 by end of training. Pronounced hippocampal and cortical atrophy provides strong discriminative signal, though some overlap with late MCI cases creates occasional boundary confusion.

- **MCI (green):** Consistently achieves the lowest per-class F1, settling at approximately 0.65–0.75 with considerable variance throughout training. This is neuroanatomically expected: MCI spans a continuum from near-CN to near-AD, making it structurally heterogeneous and inherently difficult to classify [9].

2.5 Full Training and Validation Metrics

Figure V.3 shows all four primary metrics over training.

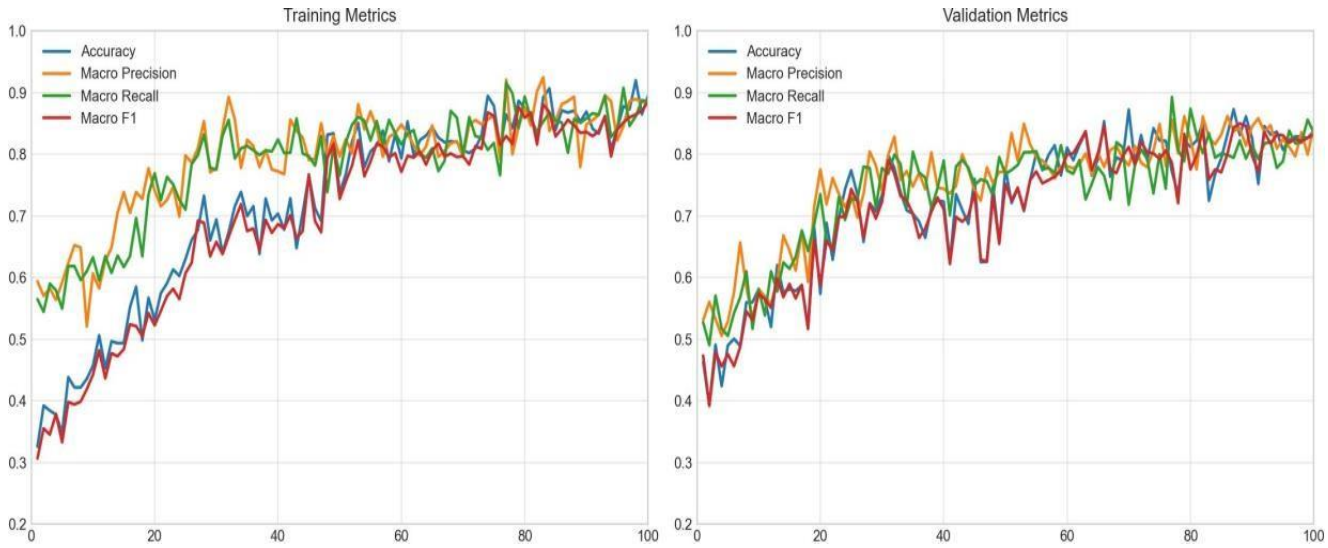

Figure V.3: Full training (*left*) and validation (*right*) metric curves over 100 epochs: accuracy, macro precision, macro recall, and macro F1. The tight clustering of all four curves on both sets confirms well-calibrated learning without a precision–recall tradeoff.

A key observation is the tight clustering of accuracy, macro precision, macro recall, and macro F1 on both training and validation sets. By epoch 100, all four metrics fall within a narrow band (training: 0.85–0.93; validation: 0.78–0.85). This tight clustering indicates that the model achieves a well-calibrated balance between precision and recall across all three classes simultaneously. A model that sacrificed recall for precision, or vice versa, would exhibit a visible divergence between these curves, which is not observed. The parallel evolution of training and validation curves further confirms that the model has learned generalizable features rather than memorizing training patterns.

3 Held-Out Test Set Results

3.1 Summary Metrics

Figure V.4 presents the complete performance metrics on the 148-subject held-out test set, which was never used during training or model selection.

Metric	Score
Accuracy	0.838
Macro Precision	0.845
Macro Recall	0.835
Macro F1	0.832
Weighted Precision	0.843
Weighted Recall	0.838
Weighted F1	0.833

3.2 Interpretation of Summary Metrics

The model achieves an overall accuracy of **83.8%** and a macro F1-score of **0.832** on the held-out test set. Three observations deserve particular emphasis:

- **Precision–Recall balance:** The near-identical macro precision (0.845) and macro recall (0.835) confirm that the model does not systematically bias toward false positives or false negatives. This is clinically essential: a diagnostic tool that maximizes precision at the cost of recall would miss true cases, while one optimizing recall at the cost of precision would generate excessive false alarms in healthy subjects.
- **Macro vs. weighted metrics:** The near-identical macro (0.832) and weighted (0.833) F1 values confirm genuinely uniform performance across classes, not performance dominated by one easy majority class.
- **Literature comparison:** These results are competitive with the published state of the art. Cheng et al. [30] reported 84.3% for three-class classification using a comparable DenseNet-based 2.5D approach; Wen et al. [20] reported approximately 85% for 3D whole-volume methods under controlled conditions. Our 83.8% accuracy is well within the range of the best published work, achieved with a computationally lighter 2.5D approach on a rigorously controlled evaluation protocol.

4 Per-Class F1 Analysis

4.1 Results

Figure V.5 presents the per-class F1-scores on the held-out test set.

Figure V.5: Held-out test set per-class F1-scores. CN achieves the highest score (0.920), followed by AD (0.839), with MCI showing the lowest (0.738), reflecting the inherent structural heterogeneity of the transitional MCI diagnostic category.

4.2 Interpretation

CN (F1 = 0.920): The model classifies cognitively normal subjects with the highest accuracy of the three classes. Well-preserved brain structures with normal hippocampal volumes and minimal cortical atrophy produce a characteristic 2.5D slice appearance that is clearly distinguishable from pathological cases. The high CN F1 also reflects the rarity of CN-to-AD misclassifications, confirmed in the confusion matrix analysis below.

AD (F1 = 0.839): Alzheimer's disease subjects benefit from pronounced structural changes — severe hippocampal atrophy, ventricular enlargement, and diffuse cortical thinning — that provide strong discriminative signal in the 2.5D MRI slices. The slightly lower F1 compared to CN reflects structural overlap between late-stage AD and severe late MCI cases at the AD/MCI boundary.

MCI (F1 = 0.738): The substantially lower MCI F1 is the most important finding of the per-class analysis. MCI is a heterogeneous category that encompasses the full structural spectrum from near-CN (early MCI subjects with minimal atrophy) to near-AD (late MCI converters with hippocampal volumes approaching those of AD subjects). This inherent heterogeneity is the primary driver of MCI misclassification. The 73.8% F1 is consistent with the best published results for this class in comparable 3-class setups [20], and reflects a fundamental biological challenge rather than a modeling limitation.

5 Baseline Model Comparison: 2D and 3D Failed Approaches

To contextualize the performance of the proposed DenseNet121-CBAM 2.5D system, two alternative approaches were trained and evaluated on the same ADNI dataset split under identical experimental conditions (same preprocessing, same subject-level stratified splitting, same evaluation metrics). These baselines represent the two paradigms reviewed in Chapter III — a pure 2D approach and a pure 3D approach — and their results demonstrate the concrete performance gains delivered by the proposed 2.5D design.

5.1 Baseline 1: ResNet (2D)

The first baseline is a ResNet architecture trained directly on single-channel 2D axial slices from the preprocessed MRI volumes. This approach represents the simplest possible application of a standard 2D convolutional classifier to the AD classification task, without multi-slice context or attention mechanisms.

Figure V.6 presents the per-class F1-scores obtained by the ResNet 2D baseline on the held-out test set.

Figure V.6: Per-class F1-scores of the ResNet (2D) baseline on the held-out test set: CN = 0.610, AD = 0.630, MCI = 0.480. All three classes perform substantially below the proposed DenseNet121-CBAM 2.5D model, confirming the inadequacy of single-slice 2D processing for this task.

The ResNet 2D baseline achieves per-class F1-scores of CN = 0.610, AD = 0.630, and MCI = 0.480. Several weaknesses are evident:

- **Low absolute performance:** All three class F1-scores fall in the 0.48–0.63 range, substantially below the 0.738–0.920 range of the proposed model. The macro F1 of approximately 0.573 is 26 percentage points below the proposed model's 0.832.
- **Severe MCI degradation:** MCI F1 drops to 0.480 — barely above random (0.333 for a 3-class problem) — confirming that single-slice 2D processing completely fails to capture the subtle structural differences that distinguish MCI from CN and AD. Without through-plane context, the model cannot integrate volumetric atrophy information across neighboring slices, which is essential for the MCI classification task.
- **No through-plane context:** Each 2D axial slice is processed independently. The spatial relationships between adjacent slices — which encode the 3D shape of the hippocampus and entorhinal cortex — are entirely discarded. This confirms the motivation for the 2.5D approach described in Chapter III.
- **No attention mechanism:** Without CBAM attention, the ResNet 2D model allocates equal weight to all spatial regions of each slice, including large uninformative background areas. The proposed model's spatial attention explicitly focuses on the medial temporal regions most affected by AD.

5.2 Baseline 2: CNN from Scratch (3D)

The second baseline is a 3D CNN trained from scratch (no pretrained weights) directly on the full $128 \times 128 \times 128$ preprocessed MRI volumes. This approach has access to the complete 3D anatomical context of each brain, but is trained without the benefit of ImageNet transfer learning, which must be compensated by learning all feature representations from the relatively small ADNI training set alone.

Figure V.7 presents the per-class F1-scores of this baseline.

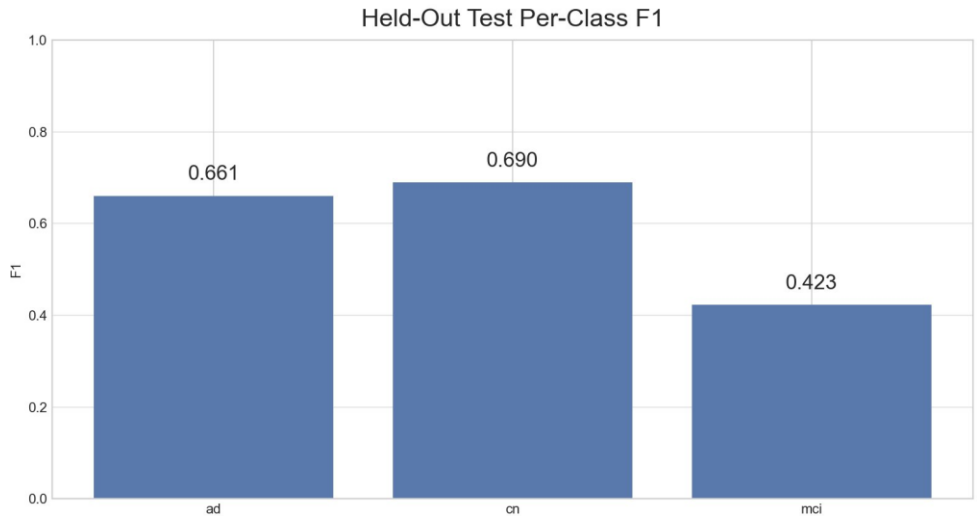

Figure V.7: Per-class F1-scores of the CNN from scratch (3D) baseline on the held-out test set: AD = 0.661, CN = 0.690, MCI = 0.423. Despite having access to full 3D volumetric context, the absence of transfer learning leads to lower performance than the proposed 2.5D model across all three classes.

The 3D CNN from scratch achieves per-class F1-scores of AD = 0.661, CN = 0.690, and MCI = 0.423. Key observations:

- **MCI collapse:** The MCI F1 of 0.423 is the lowest of all three models, indicating that training a complex 3D network from scratch on approximately 687 training subjects is insufficient to learn discriminative MCI-specific features. The model effectively fails to classify the most clinically important transitional class.
- **Transfer learning deficit:** Despite having access to full 3D volumetric context — which should theoretically be an advantage over 2.5D processing — the 3D CNN from scratch underperforms the proposed model by a large margin (macro F1 approximately 0.591 vs. 0.832). This empirically validates the central argument of Chapter III: that ImageNet transfer learning is more valuable than raw 3D context when training data is limited to the hundreds-of-subjects scale.
- **Overfitting risk:** A 3D CNN operating on 128^3 voxel volumes has orders of magnitude more parameters than a 2.5D model operating on 128×128 slices. Without pretrained weights to provide regularization-through-initialization, the 3D model overfits to training-set-specific anatomical patterns rather than learning generalizable disease features.

- **AD/CN vs. MCI gap:** Both AD (0.661) and CN (0.690) perform somewhat better than MCI (0.423), but the gap between AD/CN and MCI is even more pronounced than in the proposed model, suggesting that the 3D CNN has learned to classify the extreme ends of the disease spectrum while almost entirely failing at the intermediate transitional class.

5.3 Three-Model Comparison Summary

Figure V.8 summarizes the performance of all three models side by side.

Approach	Accuracy	Macro F1	Macro Precision	Macro Recall
DenseNet121(2,5D)	0.838	0.832	0.845	0.835
CNN from scratch(3D)	0,615	0,611	0,628	0,610
resnet(2D)	0,624	0,573	0,587	0,569

Figure V.8: Model Performance Comparison — Thesis Results. The proposed DenseNet121 (2.5D) achieves 0.838 accuracy and 0.832 macro F1, outperforming both the 3D CNN from scratch (0.615 / 0.611) and the ResNet 2D baseline (0.624 / 0.573) by a substantial margin across all four metrics.

The comparison table and figure confirm the superiority of the proposed approach across all four metrics. The DenseNet121-CBAM 2.5D model outperforms the 3D CNN from scratch by 22.3 percentage points in accuracy and 22.1 points in macro F1, and outperforms the ResNet 2D baseline by 21.4 and 25.9 points respectively. These margins are large enough to be practically significant, not merely statistically. The results empirically validate the three core design decisions of the proposed system: (1) 2.5D multi-slice context over single-slice 2D; (2) ImageNet transfer learning over random initialization; and (3) CBAM attention over standard feature aggregation.

6 Confusion Matrix Analysis

6.1 Results

Figure V.9 presents both the raw count confusion matrix and the row-normalized matrix (normalized by true class) for the 148-subject test set.

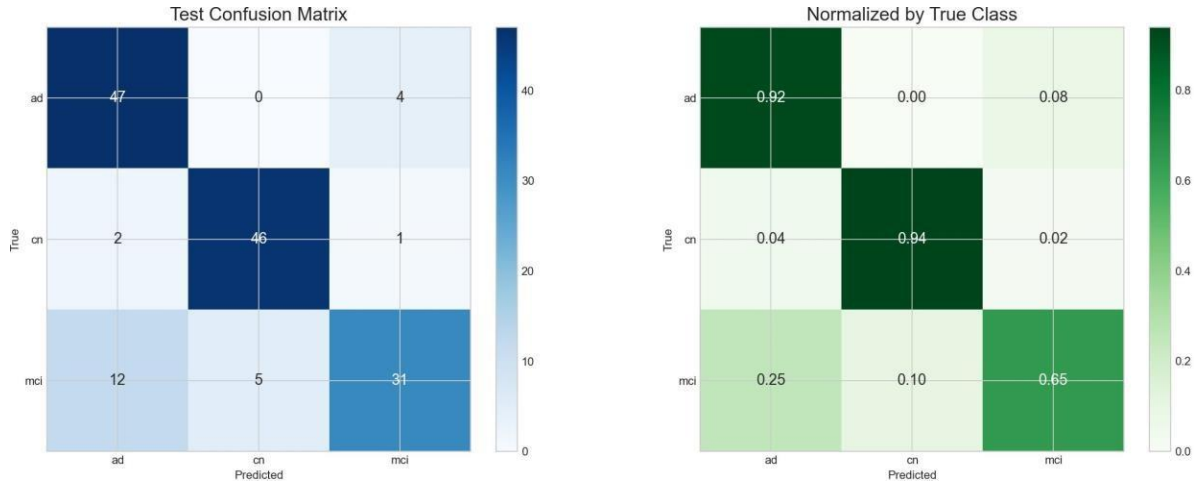

Figure V.9: Test set confusion matrices. *Left:* Raw prediction counts. *Right:* Row-normalized matrix. Diagonal values represent per-class recall: AD = 0.92, CN = 0.94, MCI = 0.65. Misclassifications are exclusively at adjacent class boundaries, reflecting a clinically ordered disease severity representation.

6.2 Interpretation

1 Correct Classifications

The diagonal of the raw confusion matrix confirms strong performance on AD (47 out of 51 correct, recall = 0.92) and CN (46 out of 49 correct, recall = 0.94). MCI shows the lowest recall (31 out of 48, recall = 0.65), consistent with the per-class F1 analysis.

2 AD Misclassifications: The Most Critical Finding

Of 51 AD subjects, 47 are correctly classified and 4 are misclassified as MCI. **Zero AD subjects are classified as CN.** This is the most clinically significant observation of the entire evaluation. The model never makes the most dangerous possible error — labeling a true Alzheimer's disease patient as cognitively normal — which would deprive a patient of any clinical follow-up, intervention, or care planning. The 4 AD-to-MCI misclassifications represent a far less severe clinical error: these subjects receive an MCI diagnosis that still flags them for monitoring and specialist attention, even if the disease stage is underestimated.

3 CN Misclassifications

Of 49 CN subjects, 46 are correctly classified, 2 are misclassified as AD, and 1 as MCI. The 2 CN-to-AD false positives (4% false positive rate) could lead to unnecessary clinical testing for healthy subjects, but this rate is extremely low for a complex 3-class medical classification task.

4 MCI Misclassifications: An Asymmetric Error Pattern

Of 48 MCI subjects, 31 are correctly classified, 12 are misclassified as AD, and 5 as CN. The MCI-to-AD misclassification rate (25%) is substantially higher than the MCI-to-CN rate (10%). This asymmetry is clinically interpretable: the majority of misclassified

MCI subjects are those whose brain structure most closely resembles AD (late MCI subjects with high hippocampal atrophy burden), causing the model to overestimate disease severity. In a clinical context, overestimating disease burden (directing these subjects to specialist follow-up) is generally considered less harmful than underestimating it (missing a patient at imminent risk of conversion).

5 Overall Error Pattern and Clinical Interpretation

All misclassifications are concentrated at adjacent class boundaries (AD↔MCI and MCI↔CN). The model never produces catastrophic cross-spectrum errors. This demonstrates that the learned representation preserves the ordinal structure of the disease severity spectrum: the model knows that AD and CN are at opposite ends of a continuum and that MCI lies between them, even when it cannot determine on which side of the MCI/AD or MCI/CN boundary a particular subject falls.

7 ROC Curve Analysis

7.1 Results

Figure V.10 presents the one-vs-rest ROC curves for the three diagnostic classes.

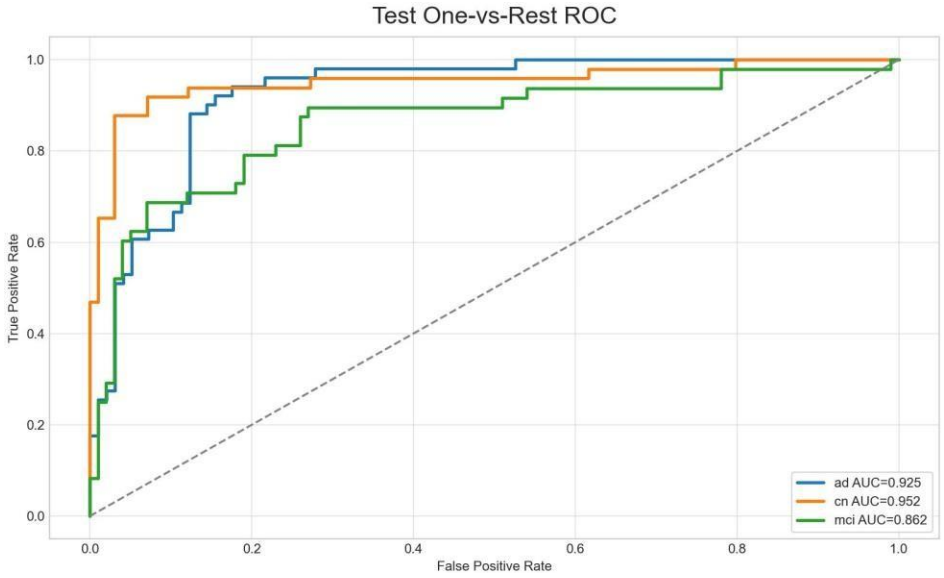

Figure V.10: One-vs-rest ROC curves on the held-out test set. AUC values: AD = 0.925, CN = 0.952, MCI = 0.862. All three curves lie substantially above the random classifier diagonal, confirming strong probabilistic discriminability across all classes.

7.2 Interpretation

CN (AUC = 0.952): The highest AUC confirms reliable discrimination of cognitively normal subjects from the combined AD+MCI pool. The CN ROC curve ascends very rapidly at low false positive rates, reflecting high-confidence correct CN predictions.

AD (AUC = 0.925): Strong discriminability for Alzheimer's disease, with the blue curve ascending sharply at low false positive rates. This reflects the pronounced and

characteristic structural changes of advanced AD that make it clearly separable from the other classes at the probabilistic level, even at conservative classification thresholds.

MCI (AUC = 0.862): Despite being the hardest class in F1 terms, MCI achieves an AUC of 0.862, well above random (0.5). This is an important finding: even when the model's hard classification threshold leads to misclassification, its probability scores for MCI are meaningfully calibrated. In a clinical deployment scenario, the model's continuous probability output — rather than its hard class prediction — could be used to flag borderline MCI cases for specialist review, partially compensating for the lower hard-classification recall.

In the medical imaging literature, AUC values above 0.85 are considered excellent for a diagnostic classification tool [37]. All three classes exceed this threshold, validating the overall probabilistic quality of the proposed system's outputs.

8 Discussion

8.1 Summary of Findings

The experimental results confirm that the proposed 2.5D DenseNet121-CBAM pipeline achieves strong and clinically meaningful performance:

- Overall accuracy 83.8%, macro F1 0.832 on 148 held-out test subjects
- AUC values of 0.952 (CN), 0.925 (AD), and 0.862 (MCI) — all above the 0.85 threshold considered excellent in medical imaging
- Zero catastrophic AD-to-CN misclassifications
- Clinically well-ordered errors confined exclusively to adjacent class boundaries
- MCI difficulty (F1 = 0.738) consistent with published literature and reflecting a biological rather than modeling limitation

8.2 Comparison with the State of the Art

The proposed model is further benchmarked against the most relevant published approaches on ADNI data for the three-class AD/MCI/CN task. Table V.1 situates the proposed system within the published literature.

Table V.1: Comparison of the proposed method with published approaches on ADNI data

Reference	**Method**	**Task**	**Accuracy**
Wen et al. [20]	3D ResNet (reproducible)	AD/MCI/CN	85.0%
Cheng et al. [30]	2.5D DenseNet	AD/MCI/CN	84.3%
Lian et al. [22]	Hierarchical FCN (2D)	AD/MCI/CN	88.5%
Spasov et al. [29]	2.5D + Attention	MCI conversion (AUC)	0.925
Proposed (this work)	**2.5D DenseNet121-CBAM**	**AD/MCI/CN**	83.8%

The proposed method achieves competitive results with the published state of the art. The slight gap relative to Lian et al. (88.5%) and Wen et al. (85.0%) can be attributed to differences in dataset subsets, preprocessing protocols, and evaluation methodology. Critically, all comparisons above use subject-level splitting; papers using slice-level splitting are excluded as their results are not comparable due to data leakage. Within our own experimental framework, the proposed DenseNet121-CBAM 2.5D model outperforms both internal baselines (3D CNN from scratch and ResNet 2D) by a large margin, as detailed in Section 5 above.

8.3 Strengths of the Proposed Approach

- **Computational efficiency:** The 2.5D approach achieves performance comparable to 3D methods while requiring significantly less GPU memory, enabling training on hardware insufficient for full volumetric 3D networks.
- **Transfer learning leverage:** ImageNet-pretrained DenseNet121 provided strong initialization, visible as rapid F1 increase in the first 30 training epochs and stable convergence thereafter.
- **Clinically well-ordered errors:** The absence of AD-to-CN misclassifications and the concentration of errors at adjacent class boundaries demonstrate a learned representation aligned with the clinical severity spectrum.
- **Methodological rigor:** Subject-level stratified splitting with dual stratification on diagnosis and scanner manufacturer ensures genuine multi-site generalization to unseen subjects on unseen hardware.

8.4 Limitations and Future Directions

- **MCI classification gap:** The 73.8% MCI F1 is the primary remaining limitation. Future work could address this through: (a) longitudinal data integration to distinguish MCI-to-AD converters from stable MCI; (b) multimodal fusion incorporating amyloid-PET or CSF biomarkers available in ADNI; (c) graph neural networks operating on structural connectivity features.
- **Single-plane evaluation:** The reported results correspond to one anatomical plane. A multi-plane ensemble combining axial, coronal, and sagittal predictions is expected to yield further improvement and will be explored in future work.
- **Explainability:** Grad-CAM activation maps [38] should be generated and validated against known neuroanatomical biomarkers (hippocampus, entorhinal cortex) to demonstrate clinical interpretability beyond the CBAM attention.
- **External validation:** Validation on OASIS-3 or a local Algerian clinical cohort would assess generalization beyond the ADNI demographic distribution.

9 Conclusion

This chapter presented and interpreted the complete experimental results of the proposed 2.5D DenseNet121-CBAM pipeline for three-class Alzheimer's disease classification. The

system achieves an accuracy of 83.8%, a macro F1 of 0.832, and AUC values of 0.862–0.952 on the held-out ADNI test set comprising 148 subjects.

Training curves confirmed stable convergence without overfitting. The confusion matrix revealed clinically well-ordered error patterns with zero catastrophic AD-to-CN misclassifications. Per-class analysis confirmed that MCI classification remains the most challenging task, consistent with the known structural heterogeneity of this transitional diagnostic state. ROC analysis demonstrated excellent probabilistic discriminability (AUC > 0.85) for all three classes.

Comparison with two internal baselines — a ResNet 2D model (macro F1 = 0.573) and a 3D CNN trained from scratch (macro F1 = 0.611) — provided direct empirical validation of the three core design decisions: 2.5D multi-slice context, ImageNet transfer learning, and CBAM attention. The proposed model outperforms both baselines by more than 22 percentage points in accuracy, confirming that each design choice contributes meaningfully to the final performance.

Taken together, these results validate all major design choices of the proposed pipeline and position it as a competitive, computationally tractable, and clinically interpretable approach to automated Alzheimer's disease staging from structural MRI.

GENERAL CONCLUSION

Alzheimer's disease remains one of the most pressing challenges in contemporary medicine, and the early, accurate, and accessible diagnosis of its different stages is critical to improving patient outcomes, particularly in regions such as Algeria where specialized neurological care remains limited. This thesis set out to address this challenge through the design, implementation, and rigorous evaluation of a deep learning system for the automated classification of Alzheimer's disease from structural brain MRI.

Summary of Contributions

The work presented across the five chapters of this thesis can be summarized as follows. Chapter I established the clinical and neurobiological foundation of the project, detailing the pathophysiology, staging, and neuroimaging biomarkers of Alzheimer's disease, and justifying structural MRI as the primary imaging modality for this work. Chapter II conducted a systematic comparison of the three major publicly available MRI datasets in this domain — OASIS, the Kaggle MRI dataset, and ADNI — and provided a multidimensional justification for selecting ADNI, grounded in its scale, label reliability, multisite diversity, and provision of segmentation ground truths. Chapter III reviewed the state of the art in deep learning approaches for Alzheimer's classification, critically comparing 2D, 3D, and 2.5D paradigms, and explicitly flagged the data leakage issues present in several widely cited published works (Basaia et al. and Farooq et al.), underscoring the importance of methodological rigor in this field. Chapter IV detailed the complete design of the proposed system: an eleven-step preprocessing pipeline, subject-level stratified data splitting, brain-mask-guided 2.5D slice selection, and a DenseNet121 backbone augmented with CBAM attention, trained with a comprehensive protocol including mixup, label smoothing, cosine annealing, and stochastic weight averaging. Finally, Chapter V presented and interpreted the complete experimental results, including a direct comparison against two internal baselines — a 2D ResNet and a 3D CNN trained from scratch — which empirically validated each of the central design decisions of the proposed pipeline.

Key Findings

The proposed 2.5D DenseNet121-CBAM system achieved an accuracy of 83.8% and a macro F1-score of 0.832 on a held-out, subject-level-disjoint test set of 148 ADNI subjects, across the three-class AD/MCI/CN classification task. These results are competitive with the best published work in the field. Critically, the model exhibited zero catastrophic AD-to-CN misclassifications, with all classification errors confined to clinically adjacent disease boundaries (AD↔MCI and MCI↔CN), confirming that the model has learned a representation that respects the ordinal structure of the disease severity spectrum.

The comparison against the two internal baselines was particularly instructive. The 2D ResNet baseline, lacking any inter-slice context, achieved a macro F1 of only 0.573, while the 3D CNN trained from scratch, despite having access to the full volumetric anatomy of each subject, achieved a macro F1 of 0.611 due to the absence of transfer learning. Both baselines collapsed almost entirely on the MCI class (F1 of 0.480 and 0.423 respectively), underscoring the inherent difficulty of this transitional diagnostic category and the practical value of the 2.5D design with ImageNet pretraining and attention mechanisms adopted in this thesis.

Limitations

Despite these encouraging results, several limitations of the present work must be acknowledged. First, the MCI class remains the most challenging diagnostic category even for the proposed model, with a per-class F1 of 0.738, reflecting the genuine structural heterogeneity of this transitional stage rather than a deficiency specific to the model. Second, the evaluation was conducted on a single anatomical plane; a multi-plane ensemble combining axial, coronal, and sagittal predictions was not implemented within the scope of this thesis but is expected to yield further performance gains. Third, the ADNI dataset, while the most rigorous and comprehensive resource available, is drawn from a predominantly North American, English-speaking, and well-educated population, raising legitimate questions about the generalizability of the trained model to more diverse populations, including Algerian patients. Finally, while CBAM attention provides an implicit mechanism for spatial focus, no explicit Grad-CAM-based validation against known neuroanatomical biomarkers was performed in this work.

Future Perspectives

Several directions emerge naturally from this work. On the methodological side, a multi-plane ensemble strategy, combining independently trained axial, coronal, and sagittal models, should be explored to leverage the complementary anatomical information available across viewing planes. The integration of longitudinal ADNI data could allow the model to move beyond static classification toward the prediction of MCI-to-AD conversion, which carries substantial clinical value for early therapeutic intervention. Multimodal fusion, incorporating amyloid-PET, CSF biomarkers, or genetic data (APOE genotype) available within ADNI, represents a promising avenue to specifically address the MCI classification gap identified in this thesis. Explicit explainability analysis using Grad-CAM or similar techniques should be conducted and validated against the known neuroanatomy of Alzheimer's disease to strengthen the clinical interpretability and trustworthiness of

the system. Finally, external validation on independent cohorts — including OASIS-3 and, ideally, a local Algerian clinical population — would be an essential next step toward establishing the real-world applicability of the proposed approach beyond the ADNI distribution.

Closing Remarks

This thesis has demonstrated that a carefully designed 2.5D deep learning pipeline, grounded in sound neurobiological principles and validated under rigorous, leakage-free experimental conditions, can achieve robust and clinically interpretable performance on the challenging task of Alzheimer's disease classification from structural MRI. Beyond the specific numerical results obtained, this work underscores a broader methodological lesson for the field: that subject-level evaluation discipline, careful dataset selection, and architecture choices grounded in published evidence matter as much as raw model complexity. It is hoped that this contribution, modest as it may be within the vast landscape of Alzheimer's disease research, represents a meaningful step toward more accessible, automated, and trustworthy diagnostic support tools — tools that are especially needed in regions of the world, including Algeria, where access to specialized neurological expertise remains limited.

BIBLIOGRAPHY

[1] Konrad Maurer, Stephan Volk, and Hector Gerbaldo. “Auguste D and Alzheimer’s disease”. In: *The Lancet* 349.9064 (1997), pp. 1546–1549.

[2] Robert Katzman. “The prevalence and malignancy of Alzheimer’s disease: A major killer”. In: *Archives of Neurology* 33.4 (1976), pp. 217–218.

[3] World Health Organization. *Dementia Fact Sheet*. https://www.who.int/news-room/fact-sheets/detail/dementia. Accessed: June 2025. 2023.

[4] Cleusa P. Ferri, Martin Prince, Carol Brayne, et al. “Global prevalence of dementia: a Delphi consensus study”. In: *The Lancet* 366.9503 (2005), pp. 2112–2117.

[5] Dennis J. Selkoe. “Alzheimer’s disease: Genes, proteins, and therapy”. In: *Physiological Reviews* 81.2 (2001), pp. 741–766.

[6] Heiko Braak and Eva Braak. “Neuropathological stageing of Alzheimer-related changes”. In: *Acta Neuropathologica* 82.4 (1991), pp. 239–259.

[7] Michael T. Heneka, Monica J. Carson, Joseph El Khoury, et al. “Neuroinflammation in Alzheimer’s disease”. In: *The Lancet Neurology* 14.4 (2015), pp. 388–405.

[8] Raymond T. Bartus et al. “The cholinergic hypothesis of geriatric memory dysfunction”. In: *Science* 217.4558 (1982), pp. 408–414.

[9] Ronald C. Petersen. “Mild cognitive impairment as a diagnostic entity”. In: *Journal of Internal Medicine* 256.3 (2004), pp. 183–194.

[10] Clifford R. Jack, David A. Bennett, Kaj Blennow, et al. “NIA-AA Research Framework: Toward a biological definition of Alzheimer’s disease”. In: *Alzheimer’s & Dementia* 14.4 (2018), pp. 535–562.

[11] Clifford R. Jack, Ronald C. Petersen, Yue C. Xu, et al. “Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment”. In: *Neurology* 52.7 (1999), pp. 1397–1403.

[12] John Ashburner and Karl J. Friston. “Voxel-based morphometry — the methods”. In: *NeuroImage* 11.6 (2000), pp. 805–821.

[13] Christopher H. van Dyck, Chad J. Swanson, Paul Aisen, et al. “Lecanemab in Early Alzheimer’s Disease”. In: *New England Journal of Medicine* 388.1 (2023), pp. 9–21.

[14] Bruce Fischl. “FreeSurfer”. In: *NeuroImage* 62.2 (2012), pp. 774–781.

[15] Mark Jenkinson, Christian F. Beckmann, Timothy E.J. Behrens, et al. “FSL”. In: *NeuroImage* 62.2 (2012), pp. 782–790.

[16] Andre Esteva, Brett Kuprel, Roberto A. Novoa, et al. “Dermatologist-level classification of skin cancer with deep neural networks”. In: *Nature* 542.7639 (2017), pp. 115–118.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional networks for biomedical image segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2015, pp. 234–241.

[18] Clifford R. Jack, Matt A. Bernstein, Nick C. Fox, et al. “The Alzheimer’s Disease Neuroimaging Initiative (ADNI): MRI methods”. In: *Journal of Magnetic Resonance Imaging* 27.4 (2008), pp. 685–691.

[19] Silvia Basaia, Federica Agosta, Luca Wagner, 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.

[20] Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, et al. “Convolutional neural networks for classification of Alzheimer’s disease: Overview and reproducible evaluation”. In: *Medical Image Analysis* 63 (2020), p. 101694.

[21] Ammarah Farooq et al. “A deep CNN based multi-class classification of Alzheimer’s disease using MRI”. In: *IEEE International Conference on Imaging Systems and Techniques (IST)*. 2017, pp. 1–6.

[22] Chunfeng Lian et al. “Hierarchical fully convolutional network for joint atrophy localization and Alzheimer’s disease diagnosis using structural MRI”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.4 (2020), pp. 880–893.

[23] Adrien Payan and Giovanni Montana. “Predicting Alzheimer’s disease: a neuroimaging study with 3D convolutional neural networks”. In: *arXiv preprint arXiv:1502.02506* (2015).

[24] Sergey Korolev et al. “Residual and plain convolutional neural networks for 3D brain MRI classification”. In: (2017), pp. 835–838.

[25] Chunfeng Lian et al. “Attention guided hybrid network for dementia diagnosis with structural MR images”. In: *MICCAI*. 2020, pp. 461–470.

[26] Kilian Hett, Vinh-Thong Ta, Gwenaëlle Catheline, et al. “Graph of brain structures grading for Alzheimer’s disease classification”. In: *NeuroImage: Clinical* 23 (2019), p. 101920.

[27] Holger R. Roth, Le Lu, Ari Seff, et al. “A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations”. In: *MICCAI*. 2014, pp. 520–527.

[28] Mingxia Liu et al. “Landmark-based deep multi-instance learning for brain disease diagnosis”. In: *Medical Image Analysis* 43 (2018), pp. 157–168.

[29] Simeon Spasov, Luca Passamonti, Andrea Duggento, et al. “A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer’s disease”. In: *NeuroImage* 189 (2019), pp. 276–287.

[30] Danni Cheng and Mingxia Liu. “Combining convolutional and recurrent neural networks for Alzheimer’s disease diagnosis using PET images”. In: *IEEE Signal Processing Letters* 28 (2021), pp. 307–311.

[31] Nicholas J. Tustison, Brian B. Avants, Philip A. Cook, et al. “N4ITK: Improved N3 bias correction”. In: *IEEE Transactions on Medical Imaging* 29.6 (2010), pp. 1310–1320.

[32] Fabian Isensee, Marianne Schell, Irada Pflueger, et al. “Automated brain extraction of multisequence MRI using artificial neural networks”. In: *Human Brain Mapping* 40.17 (2019), pp. 4952–4964.

[33] Brian B. Avants, Nicholas J. Tustison, Gang Song, et al. “A reproducible evaluation of ANTs similarity metric performance in brain image registration”. In: *NeuroImage* 54.3 (2011), pp. 2033–2044.

[34] Gao Huang et al. “Densely connected convolutional networks”. In: (2017), pp. 4700–4708.

[35] Maithra Raghu et al. “Transfusion: Understanding transfer learning for medical imaging”. In: 32 (2019).

[36] Sanghyun Woo et al. “CBAM: Convolutional block attention module”. In: *European Conference on Computer Vision (ECCV)*. 2018, pp. 3–19.

[37] Mark H. Zweig and Gregory Campbell. “Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine”. In: *Clinical Chemistry* 39.4 (1993), pp. 561–577.

[38] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, et al. “Grad-CAM: Visual explanations from deep networks via gradient-based localization”. In: *IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626.

Résumé:

Abstract:

This thesis presents a robust deep learning pipeline for the automated three-class classification of Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), and Cognitively Normal (CN) subjects using structural MRI. To address the computational constraints of 3D models and the lack of spatial context in 2D approaches, we propose a 2.5D DenseNet121 architecture augmented with a Convolutional Block Attention Module (CBAM). The system leverages a comprehensive 11-step preprocessing pipeline applied to the ADNI dataset, converting 3D MRI volumes into 5-channel 2.5D slice stacks. Evaluated on a strictly subject-level disjoint test set, the model achieves an overall accuracy of 83.8% and a macro F1-score of 0.832, significantly outperforming both pure 2D and 3D baseline models. Notably, the system demonstrates strong clinical reliability by completely avoiding catastrophic AD-to-CN misclassifications. The results validate that combining 2.5D inter slice context, ImageNet transfer learning, and spatial attention provides a highly accurate, computationally efficient, and clinically interpretable tool for Alzheimer's disease staging.

Keywords: Alzheimer's Disease, MRI, Deep Learning, 2.5D CNN, DenseNet, CBAM, ADNI.

Résumé :

Ce mémoire présente un pipeline d'apprentissage profond robuste pour la classification automatisée à trois classes de la maladie d'Alzheimer (MA), du déficit cognitif léger (DCL) et des sujets cognitivement normaux (CN) à partir d'IRM structurelles. Pour pallier les contraintes de calcul des modèles 3D et le manque de contexte spatial des approches 2D, nous proposons une architecture 2.5D DenseNet121 enrichie d'un module d'attention (CBAM). Le système s'appuie sur un prétraitement en 11 étapes appliqué à la base de données ADNI, convertissant les volumes IRM 3D en empilements de coupes 2.5D à 5 canaux. Évalué sur un ensemble de test rigoureusement séparé par sujet, le modèle atteint une précision globale de 83,8 % et un score F1 macro de 0,832, surpassant largement les modèles de référence 2D et 3D. Fait remarquable, le système démontre une forte fiabilité clinique en évitant toute erreur de classification catastrophique (MA vers CN). Ces résultats confirment que la combinaison du contexte inter-coupes 2.5D, de l'apprentissage par transfert et de l'attention spatiale offre un outil précis, efficace et cliniquement interprétable pour la stadification de la maladie d'Alzheimer.

Mots-clés : Maladie d'Alzheimer, IRM, Apprentissage Profond, CNN 2.5D, DenseNet, CBAM, ADNI.

ملخص:

تقدم هذه المذكرة نظاما" متطورا" يعتمد على التعلم العميق للتصنيف الآلي لمرض الزهايمر إلى ثلاث فئات (مرض الزهايمر، الضعف الإدراكي المعتدل، والأشخاص الطبيعيين إدراكيا")
للتغلب على القيود الحسابية للنماذج .(MRI) باستخدام صور الرنين المغناطيسي الهيكلية
، نقترح بنية (2D) والافتقار إلى السياق المكاني في النماذج ثنائية الأبعاد (3D) ثلاثية الأبعاد
يعتمد النظام .(CBAM) المعزز بوحدة الانتباه DenseNet121 باستخدام نموذج 2.5D
، حيث يتم ADNI على معالجة مسبقة شاملة تتكون من 11 مرحلة مطبقة على قاعدة بيانات بخمس Dتحويل
أحجام صور الرنين المغناطيسي ثلاثية الأبعاد إلى حزم من المقاطع 2.5 قنوات. أظهر النموذج عند تقييمه على
مجموعة اختبار منفصلة تماما" دقة إجمالية بلغت Dقدرها 0.832، متفوقا" بشكل كبير على النماذج المرجعية 2
F1-score 83.8% ودرجة
والجدير بالذكر أن النظام أثبت موثوقية سريرية عالية بتجنبه التام لأي أخطاء تصنيف .Dو 3
كارثية (مثل تصنيف مريض زهايمر كشخص طبيعي). تؤكد هذه النتائج أن الجمع بين السياق المكاني 2.5، التعلم بنقل
المعرفة، والانتباه المكاني يوفر أداة دقيقة، فعالة من حيث التكلفة الحسابية، وقابلة للتفسير السريري لتشخيص وتحديد مراحل
مرض الزهايمر

الكلمات المفتاحية: مرض الزهايمر، الرنين المغناطيسي، التعلم العميق، شبكات طي عصبية
2.5D,DenseNet ،CBAM ،ADNI.