Comparative evaluation of MRI radiomic models for Alzheimer's disease prognosis

University of Turku
Department of Computing
Master of Science (Tech) Thesis
Biomedical Engineering and Health Technology
June 2026
Nea Kontturi

Supervisors:
Adj. Prof. Harri Merisaari
Tuukka Panula

UNIVERSITY OF TURKU
Department of Computing

Nea Kontturi: Comparative evaluation of MRI radiomic models for Alzheimer's disease prognosis

Master of Science (Tech) Thesis, 65 p., 3 app. p.
Biomedical Engineering and Health Technology
June 2026

Alzheimer's disease (AD) is a progressive neurodegenerative disorder in which early detection is critical for targeted intervention. This thesis evaluates the prognostic value of structural magnetic resonance imaging (sMRI) volumetry, high-dimensional radiomics, and cognitive assessments, independently and in combination, for predicting the progression from mild cognitive impairment (MCI) to AD approximately two years after baseline.

Participants were classified as stable MCI (sMCI) and progressive MCI (pMCI) based on a two-year diagnostic follow-up. T1-weighted MRI scans and cognitive assessments from the Alzheimer's Disease Neuroimaging Initiative (ADNI) were analyzed. Automated segmentation and radiomic extraction pipelines were developed to extract volumetric and radiomic features. Explainable machine learning models based on logistic regression with elastic net regularization were trained using nested cross-validation and evaluated on an independent hold-out test set.

The results demonstrated that both sMRI-derived biomarkers and cognitive assessments contain valuable prognostic information. Evaluated on the independent test set, the region of interest (ROI) radiomic model achieved the highest predictive performance among the sMRI models (AUC=0.79). However, radiomic models did not demonstrate statistically significant improvements over the conventional volumetric model (AUC=0.73, DeLong's test p=0.087). While the best performance was achieved by a multimodal model combining ROI radiomic features and cognitive assessments (AUC=0.82), the model combining standard volumetric features with cognitive assessments (AUC=0.80) was identified as the optimal, most clinically practical solution.

The findings suggest that volumetric features and cognitive assessments remain practical and effective biomarkers for AD prognosis. Furthermore, the developed computational pipeline demonstrates the potential for integrating explainable prognostic models into future clinical decision support systems for risk stratification of MCI patients.

Keywords: Alzheimer's disease, prognosis, MRI, radiomic

Contents

1 Introduction

The global burden of dementia due to Alzheimer's disease (AD) is increasing rapidly. In 2019, approximately 57 million people were living with dementia in the world. Current estimations suggest that by the year 2030, the number of individuals worldwide suffering from various forms of dementia will reach 75 million. AD is recognized as one of the most prevalent subtypes of dementia, accounting for 60–80% of all diagnosed cases. [1], [2] It has also been estimated that the cost of dementia care will rise to 2 trillion US dollars worldwide by 2030 [1]. According to a review and meta-analysis published in 2023, the annual cost of treating mild dementia in the Nordic countries is approximately 20,000–21,000 euros per patient [3]. Increase in life expectancy due to progress in medicine and technology has been accompanied by rising cases of neurocognitive disorders such as AD. [4]

While advanced age represents the primary risk factor for developing AD, the disease's etiology is highly complex and multifactorial. The precise contribution of each of these contributing factors remains unclear, making accurate prediction of disease onset and progression exceptionally difficult. Currently, there is no cure for AD [2], although various treatment options exist [5].

Preventive medication for AD should be started as early as possible, but this is not always the case [4]. A critical stage in the progression of AD is mild cognitive impairment (MCI), which is characterized by measurable cognitive decline that does not yet have a significant impact on daily life [6]. Individuals with MCI

have increased risk of progressing to AD, although not all patients will progress [7]. Accurate prediction of which individuals will progress to AD within a clinically relevant timeframe is essential for treatment planning, because preventive medication is most beneficial for this group [4], [7]–[9].

As the global prevalence and treatment costs of dementia due to AD increase, it is necessary to identify the most cost-effective and reliable prognostic biomarkers capable of predicting the progression from MCI to AD. Non-invasive and widely accessible methods, such as structural magnetic resonance imaging (sMRI) and cognitive assessments hold significant clinical value compared to expensive and invasive positron emission tomography (PET) imaging or cerebrospinal fluid (CSF) analyses. Structural MRI has become a standard tool for evaluating and tracking neurodegenerative changes in AD. [10]–[12] Traditional sMRI approaches often use measurements of brain volume to detect brain atrophy, especially in regions such as the medial temporal lobes. These measurements are well-validated imaging biomarkers of early AD [10], [13], [14], but they may not capture all structural changes related to the pathological picture. Extracting more complex structural metrics, for example, textural features, has the potential to reveal subtle, tissue-level abnormalities of the MCI stage.

More recently, radiomics has been proposed as an approach for cancer research to extract complex quantitative features from medical images and lately this imaging analysis technique has spread to other research fields of medical imaging [15], [16]. Radiomic features are imaging biomarkers that provide detailed information about a specific region of interest (ROI), characterizing image properties such as intensity, texture, and morphological patterns [17], potentially capturing information beyond simple volumetric measures. For diagnostic purposes, MRI radiomics have already shown promising results [18]. However, radiomic analyses involve high-throughput feature extraction, which are sensitive to variations in image acquisition and pro-

cessing steps, and complex feature selection procedures, raising concerns regarding robustness, interpretability and clinical applicability. [17], [19], [20]

Despite growing interest in radiomics, it remains unclear whether these features provide meaningful prognostic improvements over conventional volumetric imaging biomarkers in predicting progression from MCI to AD. Furthermore, existing studies typically extract radiomics from the hippocampus [18], [21] or from the whole brain (gray and white matter) [7], [8], [22]. Consequently, systematic investigations evaluating radiomic patterns across multiple brain regions remain limited. Although the anatomical regions associated with AD progression are relatively well known [10], [13], the specific radiomic characteristics within these regions that are linked to disease conversion remain insufficiently understood. Multi-regional radiomic analysis combined with machine learning (ML) may therefore reveal novel imaging biomarkers that help identify individuals with progressive MCI.

Although cognitive decline and structural brain atrophy are correlated in AD [10], they capture distinct dimensions of the disease by reflecting functional impairment and biological neurodegeneration, respectively. Therefore, evaluating these modalities simultaneously is essential to understand their complementary contributions to AD prognosis.

With the ultimate goal of supporting practical clinical deployment, this thesis evaluates prognostic models using low-cost, non-invasive and clinically accessible MRI-based features together with cognitive assessments. More invasive or expensive modalities, such as PET and CSF biomarkers, are therefore considered outside the scope of this thesis. First, a conventional volumetric model is used to establish baseline predictive performance and identify brain regions associated with conversion from MCI to AD. Subsequently, high-throughput radiomic features extracted from multiple brain regions are investigated for both prognostic modeling and exploratory imaging biomarker discovery. Particular attention is given to evaluating

whether radiomic features derived from specific brain regions may serve as potential imaging biomarkers associated with progression from MCI to AD. The predictive performance of volumetric and radiomic feature sets is systematically compared. In addition, the contribution of cognitive test scores is evaluated to determine the most effective prognostic feature combination.

The modeling approach of this thesis prioritizes interpretability by employing transparent and inherently explainable models rather than complex black-box methods. By focusing on clinically accessible and non-invasive baseline biomarkers, this thesis aims to minimize costs for identification of individuals at high risk of AD conversion. Furthermore, the developed models are evaluated using an independent hold-out test set, providing a more reliable assessment of their generalizability, which is necessary from a clinical perspective. Ultimately, the prognostic models presented in this thesis are designed with clinical translation in mind, serving as foundational algorithms for future clinical decision support systems (CDSS) [23] that facilitate more targeted clinical monitoring, patient stratification, and the development of personalized preventive interventions for AD.

1.1 Research questions

RQ1 Can baseline structural MRI features predict progression from mild cognitive impairment to Alzheimer's disease approximately two years after baseline?

RQ2 How do volumetric and radiomic structural MRI features compare in predicting progression from mild cognitive impairment to Alzheimer's disease, and how does the inclusion of cognitive assessments affect their predictive performance?

RQ3 Which of the evaluated models is the most optimal and practical for integration into a clinical decision support system?

1.2 Thesis content summary

The rest of this thesis is organized as follows. Chapter 2 discusses prognostic biomarkers and provides the theoretical background of sMRI, establishing the foundation for the chosen imaging biomarkers. In addition, the chapter introduces radiomics and previous machine learning studies related to AD prognosis. Chapter 3 describes the materials and methods used in this thesis. ADNI cohort selection process, segmentation, volumetric and radiomic feature extraction, model development, and statistical evaluation procedures will be described in this chapter. Chapter 4 presents the results of the developed prognostic models and compares their performance. Chapter 5 discusses the findings in the context of previous literature, evaluates the strengths and limitations of the proposed methodology, and outlines potential directions for future research. Finally, Chapter 6 summarizes the main conclusions of the thesis.

AI declaration

During the writing process of this thesis, artificial intelligence (AI) was used to refine sentence structures, identify synonyms, and ensure grammatical accuracy. AI was also used to suggest structures for the logical organization of the sections and to find references. Furthermore, it helped with code formatting and debugging. The AI applications used were Gemini and ChatGPT.

2 Background

2.1 Clinical progression from MCI to AD

While advanced age represents the primary risk factor for developing AD, the disease's etiology is highly complex and multifactorial. [24]–[26] It is estimated that the pathological onset of the disease may occur even 20 years before the manifestation of symptoms. Over this long period of time, the gradual accumulation of neurobiological changes ultimately leads, in most cases, to noticeable memory problems, behavioral changes, and speech disorders. [9]

AD is categorized into three different stages: preclinical AD, MCI due to AD, and dementia due to AD. [8], [9] The preclinical stage is asymptomatic phase, during which alterations can be detected in the patient's brain and CSF. However, the presence of preclinical biomarkers does not guarantee progression to MCI or AD. Furthermore, identifying these biological changes requires invasive or specialized diagnostics that are seldom utilized in routine clinical practice for asymptomatic individuals [9], [27]. Therefore, the main focus is on the last two stages.

In contrast to preclinical phase, in MCI due to AD stage, patients typically have recognizable symptoms, although their daily lives remain largely unaffected. MCI is a heterogeneous prodromal stage for many cognitive disorders, making it difficult to define precisely and difficult to predict its progression. [6], [7] For this reason, this thesis refers to this stage as mild cognitive impairment, as not all MCI patients

have AD or develop AD later.

MCI is one of the risk factors for developing the disease [7]. Around 12 - 18% of people over the age of 60 experience symptoms of MCI, and approximately 15% of MCI patients are estimated to progress to dementia, primarily due to AD pathology, within two years. [6], [8] Depending on the individual, patients with MCI may have cognitive impairment in areas such as memory, executive function and language. However, they do not meet the diagnostic criteria for AD or other neurodegenerative disorders. [28]

MCI is typically classified into amnestic and non-amnestic subtypes. Amnestic MCI presents with memory impairment, whereas non-amnestic MCI does not. While the amnestic subtype is widely regarded as the primary early indicator of AD, especially when accompanied by biomarkers like accumulated amyloid-β and tau proteins, both phenotypes carry a risk of progression. [6] Therefore, in order to identify potential progressive MCI (pMCI) patients, this thesis considered MCI as a single group, because regardless of the subtype, it is still possible to develop AD.

The last stage of the disease is dementia due to AD. This stage is characterized by severe cognitive impairment across multiple domains, including memory, executive functions, and language abilities, ultimately resulting in a complete loss of independence and leading to death. [9]

Brain changes in AD are driven by underlying nerve cell degeneration which is fundamentally driven by two key proteinopathies: amyloidopathy and tauopathy [29]. In AD pathology, these proteins malfunction. This dysfunction leads to the accumulation of tau within the neurons which blocks the normal flow of information in the brain. [9], [29] On the other hand, accumulation of beta-amyloid outside of the neurons causes cell death. Both of these negatively affect brain function. [9] Nevertheless, what causes the dysfunction of these proteins and how the accumulation in specific brain regions affects cognitive function is not known. [30]

In many cases, AD is diagnosed after the onset of symptoms, when the medication that prevents cognitive decline is no longer as effective as it would be in the early stage of the disease. Therefore, it is important to identify pMCI subjects as early as possible. [25] Because this population represents a critical window for early intervention and pharmacological treatments [7]–[9], [25], the central focus of this thesis is to identify pMCI subjects from stable MCI (sMCI) subjects.

2.2 Prognostic biomarkers

A biomarker refers to an objectively measurable characteristic that indicates normal biological processes, pathological changes, or responses to treatment [31]. Biomarkers can be obtained from various sources, including biological samples and medical imaging. In the context of neuroimaging, such measures are commonly referred to as imaging biomarkers. [17]

As discussed in Section 2.1, predicting the progression from MCI to AD remains a major clinical challenge due to the heterogeneity of MCI. Prognostic biomarkers aim to address this challenge by estimating the likelihood that an individual will develop AD in the future. [2], [32]

This section reviews prognostic biomarkers for predicting progression from MCI to AD, with a focus on their clinical utility and limitations. The presented approaches are further evaluated in terms of their suitability for this thesis, which focuses on non-invasive methods.

Prognostic and diagnostic methods can be broadly classified into invasive and non-invasive methods [33]. Invasive methods refer to procedures that require obtaining data from within the body, for example, through cerebrospinal fluid collection via lumbar puncture or blood sample. Non-invasive methods, in contrast, refer to procedures that do not require breaking the skin, such as cognitive assessments or MRI. [33], [34]

2.2.1 Invasive biomarkers

The accumulation of tau together with amyloid-β in the brain, particularly in the hippocampus, are a key pathological feature associated with early stages of AD. [30] The level of these proteins can be measured using CSF biomarkers. Reliable indicators of AD are elevated levels of tau protein and decreased levels of amyloid-β in CSF. [7], [35] However, despite their diagnostic accuracy, CSF biomarkers require lumbar puncture, which is an invasive procedure that involves collecting CSF from spinal cord. [34] Although lumbar puncture is generally safe, it may lead to severe complications such as spinal haematoma or bacterial meningitis. [36]

In recent years, blood-based biomarkers have gained increasing attention as a less invasive alternative compared to CSF sample for detecting AD related pathological processes. [24], [37] Blood tests are generally cheaper, faster, and easier to perform in clinical settings. Proteins associated with AD pathology can also be detected in blood samples. In particular, plasma phosphorylated tau (p-tau) has emerged as a promising biomarker for both diagnostic and prognostic purposes. Recent studies have shown that plasma p-tau181 levels are strongly associated with established AD biomarkers, including atrophy, and cognitive decline. In a study based on data from the ADNI, plasma p-tau181 combined with genetic and cognitive measures achieved an AUC of 0.90 for predicting conversion from MCI to Alzheimer's disease within two years. [24]

Regardless of strong predictive performance, biomarkers in the blood have generally much lower concentrations than in CSF, which may limit their reliability and pose challenges for clinical use. [34], [38] Furthermore, the research findings regarding blood-based biomarkers are sometimes contradictory and need additional clinical validation [37], which is why blood-based biomarkers as well as CSF biomarkers are excluded from this thesis.

Although positron emission tomography (PET) is frequently classified as non-

invasive imaging modality [34], it requires the intravenous injection of a radiotracer and involves exposure to ionizing radiation. Consequently, in this thesis, PET is regarded as an invasive technique and therefore, discussed separately from purely non-invasive methods. [39]

PET is a molecular imaging technique that uses radiotracers to visualize biological processes such as the accumulation of amyloid-β plaques and tau tangles in the brain. [32], [34] A variety of tracers have been developed for AD research, enabling the assessment of different aspects of disease pathology. [34] Compared to structural MRI, PET has demonstrated higher accuracy for detecting early pathological changes, allowing identification of AD even during the preclinical stage of the disease. [24], [40]

Two PET techniques have proven particularly effective in predicting the progression from MCI to AD: amyloid PET, which maps protein deposition and fluorodeoxyglucose (FDG) PET, which measures brain glucose metabolism.[32], [41] FDG PET-based models have achieved an AUC of approximately 0.80 for predicting progression from MCI to AD within a 36-month window. [24] However, PET imaging is associated with several practical limitations. It is significantly more expensive and less widely available than MRI, and involves exposure to ionizing radiation. Due to these factors, PET imaging is less suitable for routine clinical use and large-scale studies [34], [40], which is why structural MRI is selected as the primary imaging modality in this thesis.

2.2.2 Non-invasive biomarkers

In contrast to the invasive and high-cost procedures previously discussed, non-invasive prognostic tools provide a more accessible and cost-effective approach to detect early signs of AD. In clinical practice, neuropsychological and cognitive assessments are widely used tools for evaluating a patient's cognitive status. [42], [43]

The Mini-mental state examination (MMSE) is one of the most widely used and accepted cognitive tests to assess cognitive impairment. [33], [42] It is also widely applied as a baseline feature with other cognitive tests and imaging biomarkers in numerous studies focusing on MCI to AD progression [7], [8], [15], [39], [44]. The test consists of a series of tasks designed to assess various domains of cognitive function. [33] The maximum score that can be achieved is 30, with higher scores indicating less cognitive impairment. Despite its widespread use, the MMSE has known limitations, particularly in detecting very early cognitive decline [33]. Studies have reported reduced sensitivity in highly educated individuals and in patients with MCI, where subtle deficits may not be fully captured by the test. [33], [45]

Importantly, the MMSE score alone is insufficient for the clinical diagnosis of MCI or AD [39], [42], and its utility as a standalone prognostic biomarker is similarly limited.

To address the limitations of the MMSE test and obtain more informative baseline measures of cognitive impairment, the Alzheimer's disease assessment scale-cognitive 13 (ADAS-Cog 13) test appears to be a promising option [46]. The ADAS-Cog exists in several versions, which differ in the number and type of tasks included. The original ADAS-Cog 11 consists of 11 items and served as the basis for the development of ADAS-Cog 13, which includes two additional tasks that are designed to increase sensitivity especially to early cognitive changes. [46], [47] Both versions mainly assess dysfunction in memory and language, but ADAS-Cog 13 covers a broader range of cognitive domains. The scores range from 0 to 85, with lower scores indicating less cognitive impairment. [46]

Although there are several cognitive assessments for monitoring cognitive impairment associated with AD [48], this thesis focuses its analysis on these two specific tests: MMSE and ADAS-Cog 13. These assessments were selected due to their common use in clinical practice and their frequent appearance in predictive modeling

literature. [42] Together, MMSE and ADAS-Cog 13 provide important additional information on cognitive impairment, and these two assessments could allow the prognostic model to have more complete representation of baseline cognitive characters alongside structural MRI imaging biomarkers.

2.3 Structural MRI as a prognostic tool

Magnetic resonance imaging (MRI) is a non-invasive imaging technique with high spatial resolution that plays a vital role in the evaluation of neurodegenerative diseases, including AD. [10], [11] MRI offers several important advantages over other imaging techniques. Unlike PET, MRI does not expose patients to ionizing radiation, making it suitable for repeated use in both clinical and research settings. [11] It is widely available and generally more cost-effective than PET imaging. [12] Its ability to provide high-resolution anatomical detail also makes it especially useful for tracking structural brain changes associated with AD. [10]

MRI exploits the magnetic properties of protons, most commonly hydrogen (^{1}H) due to the high water content in the human body. Hydrogen protons have the quantum mechanics property of spin, which is central to the image formation process. When a subject is placed in a strong external magnetic field, hydrogen protons within body tissues align either parallel or antiparallel to the field. A radiofrequency (RF) pulse is then applied at the specific (Larmor) frequency, temporarily disturbing this alignment and causing the protons to absorb energy and become synchronized. After the RF pulse is switched off, the protons gradually return to their equilibrium state through relaxation processes. During this process energy is released and detected by MRI receiver coils as a radiofrequency signal. Differences in relaxation times between tissues generate image contrast. By adjusting imaging parameters, this contrast can be weighted toward different relaxation properties, such as in T1-weighted or T2-weighted images, which enables detailed visualization of anatomical structures. [49]

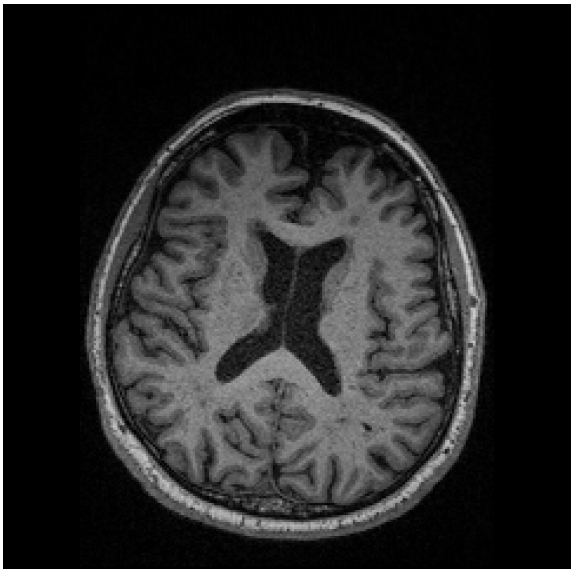

Figure 2.1: Example of a T1-weighted image demonstrating contrast between gray matter, white matter, and CSF.

AD is characterized by progressive structural brain changes, where neuronal loss and neurodegeneration lead to measurable brain atrophy. Structural neuroimaging provides a useful way to visualize and quantify these alterations. [10], [11]

Among sMRI modalities, T1-weighted imaging is particularly well suited for the assessment of neurodegeneration, as it provides high contrast between gray matter (GM), white matter (WM), and CSF, enabling accurate anatomical delineation as shown in Figure 2.1. [11] These properties make T1-weighted images an effective choice for automated brain segmentation tools [8], [11], which in turn makes them a practical option when aiming to minimize overall processing costs and study brain structures. Segmentation methods are discussed in more detail in the next chapter.

Furthermore, T1-weighted images are widely used in both clinical practice and research on AD, and are integrated into routine diagnostic workflows for AD. [10], [14], [50] As many studies report T1-weighted imaging as a recommended and standard diagnostic tool for AD [11], [14], [39], other sMRI modalities, such as T2-weighted or fluid-attenuated inversion recovery (FLAIR) sequences, are more suited for assessing white matter lesions or vascular-related changes. [11], [13] For these reasons,

T1-weighted MRI is selected as the primary imaging modality in this thesis.

In addition to providing high-resolution anatomical detail, T1-weighted MRI can also support the differentiation between various types of dementia. It is not uncommon for multiple forms of dementia to coexist in a single patient, making diagnosis based on symptoms alone difficult [1]. Structural imaging can aid in distinguishing between these conditions by highlighting region-specific patterns of atrophy associated with different neurodegenerative diseases. [10] In the following sections, these regional patterns and the computational methods used to quantify them are examined in more detail

2.3.1 Regional volumetric analysis

Hippocampal atrophy detected from T1-weighted image is one of the most well-established indicators of AD in the MCI phase [10], [13], [14], but in general, volume loss within structures of the medial temporal lobe (MTL) is considered a key feature of early disease stages. [10] The MTL includes several regions involved in memory processing, such as, the hippocampus, entorhinal cortex, amygdala, and posterior cingulate cortex. Atrophy within these regions has been shown to occur already during the preclinical phase of the disease and is therefore regarded as an important prognostic marker for detecting pMCI subjects. [10], [51]

For instance, using logistic regression and hippocampal volume alone to predict MCI to AD conversion within 36 months has been shown to achieve a predictive accuracy of approximately 60.4% (95% CI 0.52-0.688, AUC$\approx$0.65) [52]. In addition to hippocampal volume, entorhinal cortex (EC) atrophy has been reported as a moderate predictor of conversion from MCI to AD. Using EC volume alone, a logistic regression model achieved an AUC of 0.762 for predicting conversion within three years. The same study also combined hippocampal and EC volumes, which improved predictive performance to an AUC of 0.812. Furthermore, adding demographic

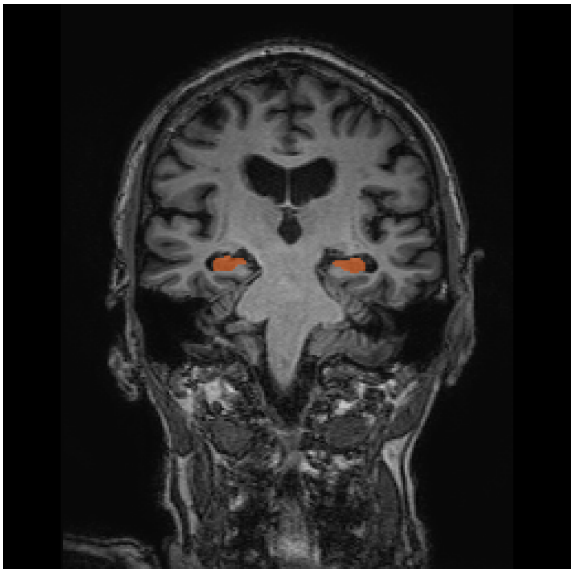

Figure 2.2: Location of the hippocampus in a T1-weighted image.

information such as age and cognitive assessments including MMSE increased the performance to an AUC of 0.854. [53]

Previous studies have also demonstrated significant differences in whole brain atrophy between sMCI and pMCI subjects, which supports the value of volumetric analysis for early prognosis. For instance, research utilizing longitudinal T1-weighted imaging data has shown that pMCI subjects exhibit whole brain atrophy rates approximately 50% higher than sMCI subjects ($p < 0.0005$) after adjusting for baseline volumes, age, and gender. The same study highlights ventricular expansion as another critical prognostic feature indicating that the ventricular expansion rate in pMCI subjects is also approximately 50% greater than in those who remain stable. [54]

To examine regional changes in volume or other computational properties, T1-weighted imaging data must first undergo image segmentation. In the field of medical image analysis, segmentation is usually the first and most important step because it affects the whole outcome of the study. [55], [56]

Segmentation in neuroimaging involves partitioning raw MR images into distinct,

spatially contiguous regions corresponding to tissue types or anatomical structures. This computational process assigns each voxel to a predefined class based on characteristics such as signal intensity and spatial location. A voxel represents a three-dimensional pixel with a unique intensity value and spatial coordinates. By grouping voxels belonging to the same anatomical structure, segmentation produces labeled regions that enable subsequent volumetric analysis. [55] The figure 2.2 shows the segmentation of the right and left hippocampus.

Manual segmentation is one of the image segmentation methods, which is usually the most accurate and consider the "gold standard" method. [57] However, manual delineation is extremely time-consuming and impractical for large datasets such as those used in ML studies. Consequently, modern segmentation methods rely on automated three-dimensional segmentation algorithms. [55]

Automated brain segmentation is commonly performed using established neuroimaging software packages such as FreeSurfer, FSL, and Statistical parametric mapping (SPM) of which FreeSurfer is one of the most used software tools to study volumetric measurements related to AD. [58], [59] These tools implement fully automated pipelines that incorporate important image preprocessing steps, including skull stripping, intensity normalization, and bias field correction, alongside sophisticated segmentation algorithms. Each tool has methodological differences, but all of them aim to produce segmentation based on image registration to an anatomical atlas with prior tissue probabilities to classify each voxel into predefined regions of interest (ROIs). [58], [60]

However, traditional segmentation pipelines such as FreeSurfer's "recon-all" pipeline are computationally very intensive and may require several hours to run a single image. [59], [61] Recently, faster deep learning based segmentation approaches such as SynthSeg have been proposed, which use convolutional neural networks (CNN) to directly predict anatomical labels and enable rapid volumetric analysis while main-

taining accurate segments. [57]

After segmentation, the resulting 3D mask serves as the basis for volumetric calculations. The most straightforward computational approach is voxel counting, where the algorithm sums voxels belonging to a specific anatomical class to calculate the physical volume of the entire ROI. [62], [63]

However, volumetric analysis alone has several limitations. First, volume represents a coarse summary of structural changes within a brain region [64]. A single value describes the entire anatomical structure and may be influenced by individual differences in total head size or sex [65], [66]. Second, volumetric measurements capture macrostructural loss but do not reflect microstructural heterogeneity within ROIs, which may occur even before significant brain atrophy [64]. Consequently, the analysis of structural changes in the brain may require additional information beyond volumetric measures alone.

Supporting this, a previous study [67] showed that texture-based features extracted from the hippocampus improved predictive performance compared to hippocampal volume alone (AUC 0.74 vs 0.67, DeLong test p=0.005) when predicting progression from MCI to AD within two years. This supports the radiomics approach, which enables the extraction of more comprehensive quantitative information from medical images.

2.3.2 Radiomics in neuroimaging

Radiomics is an image analysis process that enables the calculation of a large number of quantitative features from medical images using predefined mathematical formulas. [20], [68] Radiomics was originally developed for cancer research as a high-throughput image analysis tool and has been widely used in studies on gliomas, lung and breast cancer. [15], [16] The term "radiomics" was first introduced by Lambin et al. in 2012 [68] for the purpose of tumor analysis. [69] In recent years, interest

in radiomics research has grown in the field of medical imaging, and this technique has also been applied to predict the progression from MCI to AD. [8], [15], [20]

A primary strength of radiomic features is their ability to detect and quantify microstructural changes that might not be visible in standard two-dimensional images. [16], [64] Radiomic features can be divided into three main groups: morphological features, intensity features, and texture features. [17], [21] These features can then be used to support the prognosis of AD. This section discusses the radiomic pipeline in more detail.

As with volumetric analysis, segmentation is the first step in radiomic pipeline, typically calculated using automated segmentation tools. Segmentation leads to an ROI mask R, where each voxel j is assigned a binary label. Voxels inside the ROI are labeled as 1, whereas voxels outside the ROI are labeled as 0. [17]

$$R_j = \begin{cases} 1, & j \in \mathrm{R} \\ 0, & \text{otherwise} \end{cases}$$

After segmentation, preprocessing is a crucial step in radiomics, as radiomic features are sensitive to different imaging acquisition methods and parameters. Preprocessing aims to improve the reproducibility of the extracted features, allowing radiomics to be used as interpretable and robust biomarkers, while also reducing variability caused by different scanners. Common preprocessing steps include voxel resampling, range re-segmentation and discretization. [20]

Voxel resampling is used to standardize voxel size across images. [19] MRI datasets often contain non-isotropic voxel sizes due to differences in acquisition protocols, which especially affect texture features. Resampling interpolates the image to isotropic voxel spacing, ensuring that spatial relationships between voxels are comparable across MRI data and the image are invariant to rotation. [17], [19], [70] Based on a recent review by Trojani et al. [20], which investigated radiomic preprocessing

parameters across different imaging modalities, most MRI studies resampled voxels to isotropic resolution, with $1\times1\times1$ mm^3 being the most commonly used voxel size.

Range re-segmentation aims to remove specific voxels inside the ROI that fall outside the specified range of gray-level values. This step applies to modalities such as PET, but not to MRI data where intensity values are not standardized and depend on scanner settings and acquisition parameters. [17], [20] Therefore, for MRI data, intensity outlier filtering is typically performed by calculating the mean (μ) and standard deviation (σ) of voxel intensities within the ROI and excluding values that fall outside the interval $\mu \pm 3\sigma$. [17], [20], [70] However, when intensity discretization is applied, additional intensity outlier filtering is generally not recommended for MRI data by the Image Biomarker Standardization Initiative (IBSI), as discretization already reduces intensity variability [17].

Discretization is usually the final preprocessing step and groups image intensities inside the ROI. Discretization converts continuous voxel intensities into a finite number of intervals or bins (Figure 2.3). Two common approaches are fixed bin number and fixed bin width discretization. [17] The fixed bin number method divides the intensity range into a predefined number of bins, while fixed bin width assigns bins based on constant intensity intervals. [17], [70]

According to the review article [20], for MRI data both bin number and bin width strategies have been used either independently or in combination. The use of only bin number or only bin width was nearly equally common.

Other image processing steps before radiomics extraction include harmonization methods. These approaches aim to minimize variability caused by differences in imaging protocols and scanners. Common examples used in MRI radiomics include intensity normalization techniques such as z-score normalization and Nyúl normalization. [19]

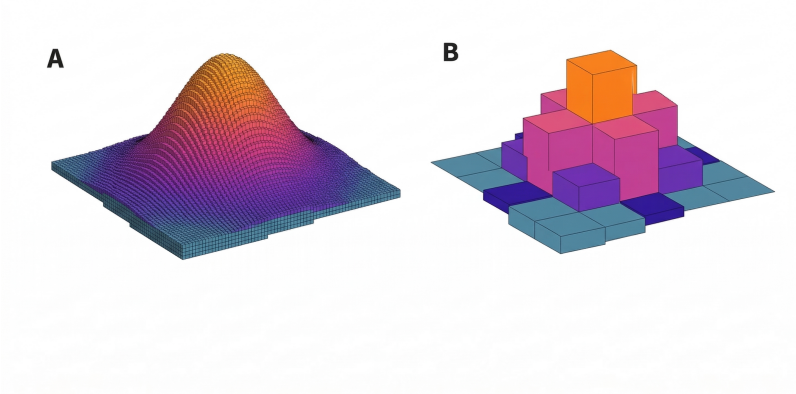

Figure 2.3: Element A. represents raw intensities, while B. shows what happens when intensity discretization is applied [70].

Finally, after preprocessing steps radiomic features can be extracted using the segmented ROI mask. [70] This part transforms the image to mineable high-dimensional data using different formulas. [71]

As mentioned previously, the first type of radiomic features is morphological features. These features describe shape and involve for example volumetric measurements but also many other geometric features of the segmented ROI such as surface area, sphericity, elongation, major, minor axis length and compactness. [17], [21] Morphological features are computed solely from the segmentation mask and are independent of voxel intensity values. Therefore, if range re-segmentation is applied, a separate intensity ROI mask would be required. [17]

The second category of radiomic features is intensity-based features, also referred to as histogram-based features. [21], [70], [71] These features can be divided into four different subgroups: local intensity, intensity-based statistics, intensity histogram, and intensity-volume histogram features. [17] They are primarily calculated from either the unbinned distribution of voxel intensities within the ROI or from a histogram of voxel intensities generated via a discretization method. This category includes features such as mean, median, standard deviation, entropy, and skewness. In the radiomics literature, these features are also commonly described as first-order

statistics [71], which is why this thesis will also use the term first-order statistics to refer to this group of features.

Texture features quantify tissue heterogeneity within the ROI [71]. In addition, this category consists of several different subgroups, of which the most commonly used is the gray-level co-occurrence matrix (GLCM). [16], [71] Other texture feature groups are the gray-level run-length matrix (GLRLM), gray-level size zone matrix (GLSZM), and neighborhood gray-tone difference matrix (NGTDM). [21] Texture features can also be considered second-order statistics. [71]

In addition to different feature types, different image filters can be applied during the extraction process to enhance specific image characteristics [70]. Filters such as Laplacian of Gaussian and wavelet transformations substantially increase the number of extracted radiomic features [72]. Consequently, in prognostic modeling, radiomic features require rigorous validation strategies and feature selection to reduce the risk of overfitting [73].

Overall, radiomics enables a more comprehensive characterization of brain tissue compared to traditional volumetric analysis by capturing not only size-related changes but also intensity distributions and spatial relationships within anatomical regions. This high-dimensional feature representation has the potential to improve the prediction of AD progression by incorporating information that may not be detectable through conventional measures alone. Previous diagnostic studies have already demonstrated promising performance of MRI radiomics in differentiating normal, MCI, and AD subjects [18], while also emphasizing the need for further research before radiomics can be reliably translated into clinical practice. Therefore, this thesis aims to investigate whether radiomics can improve prognostic of AD compared with conventional volumetric measures, while also exploring the potential of region-specific radiomic biomarkers.

2.4 Machine learning for MCI to AD prediction

ML refers to the development of computational algorithms capable of extracting useful patterns from data. Rather than relying on explicit, rule-based programming, these systems identify underlying trends and generate inferences to perform specific analytical tasks. [74] ML has been widely applied to develop prognostic models that distinguish sMCI from pMCI using a wide range of input features and different ML algorithms [4], [8], [41], [44], [75].

To ensure relevance and reliability, studies included in this literature review section were selected based on the following criteria: i) the study formulated the sMCI versus pMCI prediction as a binary classification problem based on the future disease status, ii) utilized sMRI-derived features, iii) reported quantitative performance metrics such as area under the receiver operating characteristic curve (AUC) or classification accuracy when AUC was not available, iv) was published after 2010, and v) evaluated model generalizability using an independent test set. The methodological rationale behind these selection criteria, particularly with respect to validation strategy, performance evaluation, and publication period is discussed further in Section 2.4.3.

While this section is not a systematic review, two comprehensive systematic review articles [4], [41] published in 2021 served as primary sources, analyzing 111 and 116 prognostic ML studies, respectively. From these, six studies met the strict inclusion criteria. In addition, studies published after 2021 were searched separately using the same selection criteria to identify more recent developments not covered by the review articles [4], [41]. Particular attention was given to studies utilizing MRI radiomics, as radiomics-based approaches were not extensively discussed in reviews [4], [41].

2.4.1 Algorithmic approaches

Prognostic studies for AD have used multiple different algorithms as summarized in Tables 2.1 and 2.2 shows. However, most studies reviewed in [41] and [4] applied support vector machines (SVM) to this prediction task. According to [41], SVM was used in 32.6% of the reviewed studies, while logistic regression was the second most commonly used algorithm, appearing in 15.0% of studies.

Table 2.1: Summary of studies using traditional ML methods for predicting progression from MCI to AD. Abbreviations: AB, AdaBoost; APOE4, Apolipoprotein E4; AUC, area under the receiver operating characteristic curve; CDR, Clinical Dementia Rating; CI, confidence interval; GM, gray matter; RF, random forest; SVM, support vector machine.

Author (Year)	**N (sMCI/pMCI)**	**Features**	**ML method**	**AUC (95% CI)**
Lebedev et al. (2014) [76]	35*	cortical thickness, non-cortical volumes, demographics, APOE4	RF	0.83 (0.7-0.965)
Donnelly-Kehoe et al. (2018) [77]	100/100	brain morphometry, demographics, and MMSE	RF SVM AB	0.75 (NA) 0.76 (NA) 0.63 (NA)
Sun et al. (2018) [78]	134/76	GM densities	Lasso SVM	0.68 (NA)
Shu et al. (2021) [7]	203/154	MRI radiomics, ADAS-Cog, CDR, APOE4	SVM	0.79 (NA)
Mieling et al. (2025) [44]	189/189	regional volume and thickness	XGBoost	70% accuracy (NA)
Li et al. (2025) [8]	189/154	MRI radiomics, CDR, ADAS-cog	Decision tree	0.88 (0.80-0.94)

* The study does not define the exact distribution of sMCI and pMCI subjects.

Despite the variety of algorithms tested in the literature, findings in [41] reported that the choice of algorithm alone did not have a statistically significant impact on predictive performance across studies. Although non-linear models generally achieved slightly higher performance, imaging features appeared to provide

greater benefit for linear and generalized linear models. These findings suggest that predictive performance may depend more strongly on feature quality and study design than algorithmic complexity. Consequently, simpler and more interpretable linear models may still provide competitive performance, especially since model transparency is highly desirable in both neuroimaging research and CDSS [23], [79].

While the review findings in [41] suggested that algorithm choice has a limited impact on performance, the second review [4] noted that deep learning (DL) algorithms consistently achieve the highest predictive performance. DL is a subcategory of ML based on multi-layered neural networks that automatically extract features from raw data [74], [80]. Table 2.2 summarizes recent prognostic DL approaches, illustrating their strong predictive capabilities. Convolutional neural networks (CNN) are the most commonly used DL algorithms in recent literature [24]. However, despite outperforming traditional ML methods, DL models are inherently limited by their "black-box" nature. This lack of interpretability poses a significant challenge in clinical prognostic tasks, where understanding the neurobiological and physiological rationale behind a prediction is crucial [79].

Table 2.2: Summary of studies using DL methods for predicting progression from MCI to AD. Abbreviations: CNN, convolutional neural networks.

Author (Year)	**N (sMCI/pMCI)**	**Features**	**DL method**	**AUC (95% CI)**
Spasov et al. (2019) [81]	228/181	brain volumes, demographics, neuropsychological, APOE4	CNN	0.93 (NA)
Zhang et al. (2021) [82]	251/162	3D brain volume	CNN	0.87 (NA)

Directly comparing all of these studies summarized in Table 2.1 and 2.2 is challenging, as methodological differences, such as preprocessing techniques, validation strategies, and dataset characteristics, are highly variable across the literature. Al-

though DL models are less interpretable, these complex classification models are becoming increasingly common. Studies that achieve the highest predictive accuracy generally use a combination of multimodal and multidimensional data alongside complex models. When comparing the two tables, it is evident that none of the traditional approaches in Table 2.1 achieved a higher AUC than the leading study [81] presented in Table 2.2. Furthermore, across both traditional ML and DL methodologies, models consistently benefit from the inclusion of complementary clinical information, such as demographic variables, cognitive assessments, and genetic data. [4]

2.4.2 Prognostic features

The review in [41] categorized the utilized input features into several broad groups, including imaging features, cognitive features, and demographic and genetic features, although each category also contained multiple subcategories. Among these, imaging biomarkers were one of the most frequently utilized feature classes. Structural T1-weighted MRI data were used in 69% of the reviewed studies [41], again highlighting the central role of sMRI in prognostic modeling of AD. Similarly, [4] reported that whole brain volumes represented the most commonly used imaging feature category, appearing in 70 reviewed studies.

As shown in Table 2.1, many studies favored volumetric measurements and cortical thickness using FreeSurfer processing pipelines. For example, studies in [44], [76], [77] used brain morphometry features together with demographic variables as input features. Interestingly, a more recent study [44] achieved an accuracy of 70% using an XGBoost model utilizing similar regional volume and thickness features, highlighting the enduring relevance of these established imaging biomarkers. In contrast to regional volumetric and thickness measures, also voxelwise gray matter(GM) density methods were employed, achieving an AUC of 0.68 [78].

Regardless of the large number of studies utilizing MRI-based features, relatively few investigations have focused specifically on MRI radiomics [8]. Nevertheless, radiomics-based ML approaches have been increasingly applied for the classification of cognitively normal (CN), MCI, and AD subjects using structural T1-weighted MRI [21], [83], [84]. These studies aim to evaluate whether quantitative radiomic features can capture subtle tissue-level alterations associated with different stages of neurodegeneration. For example, a radiomics-based classification model presented in [84] achieved cross-validated AUC of 0.97 ($\pm$ 0.0175) when differentiating MCI from AD subjects. Furthermore, a review by [18] highlighted the diagnostic promise of MRI radiomics, noting that more research related to radiomics is needed.

More recently, a few prognostic studies using MRI radiomic features have been published [7], [8], [22]. For example, a prognostic radiomics-based model was developed in [8], using baseline T1-weighted MRI from 343 MCI subjects. Extracting 756 initial features from whole-brain gray and white matter, they utilized dimensionality reduction to isolate 11 key features to build a radiomic signature using logistic regression. The radiomic signature alone achieved a moderate AUC of 0.78. Highly predictive variables included gray matter volume counts, maximum 3D diameters, and texture features such as cluster shade. By combining this signature with neuropsychological scores using a decision tree, the model achieved an accuracy of 0.88 in the validation set. However, in the discussion section, the same value reported as accuracy is referred to as AUC, which causes uncertainty in the interpretation of the reported performance.

Studies [8] and [7] serve as the primary radiomic reference studies for this thesis, as they have developed precisely the kind of multimodal prognostic models that will be developed in this thesis. Both studies found that when cognitive tests were combined with radiomic features, the AUCs improved significantly compared to using the radiomic signature alone. However, uncertainties in performance met-

rics reported by [8], combined with a broader lack of direct comparisons between high-dimensional radiomics and conventional volumetry, highlight a critical gap in the current literature. To address this, this thesis conducts a rigorously validated, transparent comparative evaluation of these exact modeling strategies, alongside an exploratory analysis of MRI radiomics. This analysis further determines the most robust and computationally efficient sMRI features for future integration into a CDSS.

2.4.3 Methodological challenges in prognostic modeling

Although ML methods have shown promising results in predicting progression from MCI to AD, substantial methodological challenges remain. As summarized in Tables 2.1 and 2.2, reported predictive performances vary widely across studies. These deviations are largely driven by differences in dataset characteristics, modeling choices, and validation procedures, which complicate comparisons between studies and raise uncertainty regarding which features are most informative for modeling AD prognosis. [41] Furthermore, the lack of reporting on confidence intervals in most studies presented in this Section exacerbates these discrepancies and obscures the true reliability of these results.

In predictive modeling using neuroimaging data, a major challenge is the high dimensionality of the feature space relative to the typically limited sample sizes available [79]. Structural MRI datasets may contain hundreds or thousands of imaging variables, particularly when radiomic features are extracted. In radiomics studies, the number of extracted features may exceed the number of subjects by several orders of magnitude [7], [8], making robust feature selection and validation particularly important. Such high-dimensional feature spaces are often characterized by a low signal-to-noise ratio in addition to strong correlations between features. Under these conditions, the risk of overfitting increases, meaning the model memorizes the

training data, including noise, and fails to generalize to unseen data [79], [85].

Furthermore, prognostic datasets frequently suffer from class imbalance, where usually the number of sMCI subjects typically outweighs pMCI subjects within a given timeframe. If not addressed through resampling techniques or appropriate evaluation metrics, models may become biased toward the majority class, leading to high overall accuracy despite poor sensitivity for identifying subjects who progress to AD. [86] Older studies (published before 2010) also frequently used relatively small sample sizes, which further limited reliability. For example, the study [52] published in 2008, discussed in Section 2.3.1, had only 53/73 pMCI/sMCI subjects, but also [76] in Table 2.1 had a very small sample size.

To mitigate the risks of high dimensionality and overfitting, feature selection plays a central role in prognostic modeling using radiomics. Feature selection aims to reduce dimensionality by identifying the most informative variables while removing redundant or noisy features.[41] Commonly applied methods include filter approaches based on statistical tests or correlation analysis, wrapper methods such as recursive feature elimination, and embedded methods integrated directly into the learning algorithm, such as LASSO regularization [21], [80].

However, improper feature selection procedures are frequently found in prognostic studies [41]. The review article noted that some studies performed automated feature selection on the whole dataset before doing the train-test split. This introduces data leakage and results in overly optimistic performance estimates. Therefore, feature selection should always be conducted exclusively within the training data during each validation iteration. [41]

Validation strategy represents another critical methodological issue [41]. Reliable evaluation of predictive performance requires strict separation between training and testing data [41]. Cross-validation (CV) methods are commonly used within the training data to estimate model performance and optimize model parameters [85].

Among CV methods, 10-fold CV and leave-one-out cross-validation (LOOCV) are the most commonly used approaches [4]. In k-fold CV, the dataset is divided into k subsets, where the model is iteratively trained on k-1 folds and evaluated on the remaining fold [87]. The review [4] discussed that while CV methods are primarily used to estimate model performance across multiple train-test splits, nested CV is considered best practice in ML for model optimization and hyperparameter tuning to further reduce the risk of overly optimistic performance estimates [85].

Crucially, a strict inclusion criterion for the studies summarized in this section (Tables 2.1 and 2.2) was the use of an independent hold-out test set for final model evaluation. Reliable evaluation of prognostic ML models requires assessment on previously unseen data in order to estimate how well the model generalizes beyond the training cohort. As noted in [41], evaluating a model using the same dataset for both training and testing provides only training performance and does not reliably reflect real-world generalizability. Furthermore, the same study reported that approximately one quarter of the reviewed studies had misused the test set during model development or evaluation. By prioritizing studies evaluated on previously unseen data, the performance metrics presented here provide a more realistic estimate of generalizability and are therefore more comparable to this thesis.

Further complicating cross-study comparisons, the follow-up periods used across these studies to determine disease progression also vary. In Tables 2.1 and 2.2 the follow-up time ranges from 18 months [78], [81] to 24-36 months [7], [8], [76], [82], while in some cases the exact follow-up time is not explicitly defined [44], [77]. This also complicates direct performance comparisons. In general, longer follow-up periods are associated with increased prognostic uncertainty, as baseline biomarkers become less predictive over extended time intervals. [88]

Ultimately, a critical bottleneck in deploying these models is clinical feasibility. For these prognostic techniques to effectively assist clinicians in everyday practice,

there must be a practical balance between the most advanced, high-performing algorithms and the data that are actually accessible in standard clinical workflows. Relying on highly invasive or expensive multimodal data limits real-world applicability. Furthermore, while standard regional volumetry and complex deep learning models have been extensively studied, there remains a significant research gap regarding the prognostic utility of quantitative MRI radiomics. Consequently, this thesis focuses on bridging this gap by maximizing predictive performance using essential, easily obtainable data, specifically MRI radiomics combined with standard demographic and cognitive assessments. Furthermore, the models are evaluated using a previously unseen independent test set to provide a more realistic estimate of model generalizability and better approximate real-world clinical deployment.

3 Materials and methods

3.1 Alzheimer's Disease Neuroimaging Initiative

Data used in this thesis were downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). ADNI is a large longitudinal multicenter study established in 2004 with the aim of validating biomarkers for clinical trials in AD and supporting the development of disease-modifying treatments. ADNI participants include cognitively normal (CN) individuals, MCI and AD subjects. [24]

The ADNI study collects comprehensive multimodal data, including MRI, PET, biological markers, clinical evaluations, and neuropsychological assessments to characterize MCI and early AD progression. By the end of 2020, more than 5000 studies using ADNI data had been published. [24] This thesis reflects the data available in 2025. All available ADNI phases were downloaded, including ADNI1, ADNI2, ADNIGO, ADNI3, and ADNI4. Nevertheless, no participants were included from the ADNI4 phase because the data required for follow-up was not available. Ultimately, the participants were distributed across the phases as follows: ADNI1 256, ADNI2 233, ADNIGO 64, and ADNI3 57.

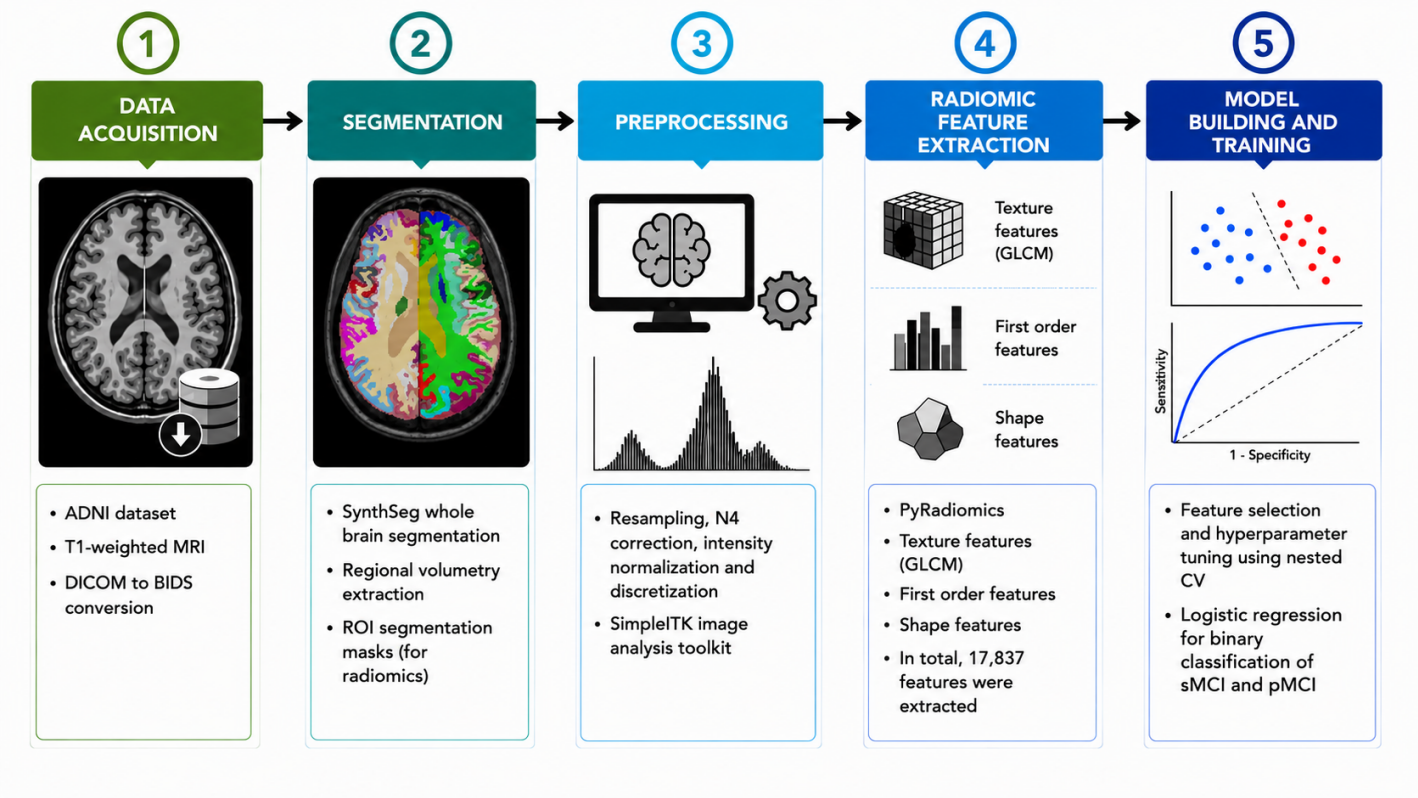

Figure 3.1: Overview of the complete analysis pipeline, illustrating the progression from raw T1-weighted MRI acquisition through preprocessing, feature extraction, and model training to final sMCI versus pMCI classification.

3.1.1 Participants

The entire dataset was initially divided into 3T and 1.5T subsets based on magnetic field strength. The aim was to utilize as much MRI data as possible, which is why participants from all ADNI phases were initially included. The ADNI1 phase contains the majority of the 1.5T images, which were included for participants who did not have 3.0T images available at baseline. In the ADNI1 phase, a participant could belong to a group in which no 3T images were acquired at all. A more detailed description of the imaging protocols related to the ADNI1 phase can be found here: ADNI MRI technical procedures manual.

To determine the first time point, the VISCODE2 (VISCODE2=bl) variable provided by ADNI was used, which is consistent regardless of the phase. At baseline,

the 3T subset consisted of 2657 subjects. Individuals whose SynthSeg segmentation failed and those without baseline diagnostic information were excluded. This left 2,501 subjects, whose diagnoses were distributed as follows: 1,136 CN, 1,010 MCI, and 336 AD. Many subjects had multiple images taken during a single scanning session. To eliminate duplicates, images with a note of repetition in the series description were removed and otherwise, the first image was included.

In the 1.5T subset, a total of 823 participants had 1.5T images at baseline, which were processed in the same way as the 3T subset. After accounting for missing segmentation data and removing participants with missing diagnosis information, 819 participants remained. From this subset, 450 participants were identified who had no 3T images available and this group was merged into the 3T subset.

All of these steps were also performed to the second time point data, which corresponded to the 2-year follow-up visit (VISCODE2=m24). The m24 visit was selected as the primary follow-up time point because it provided the largest number of participants who had progressed from MCI to AD while maintaining a sufficiently large overall sample size.

For this thesis, only participants diagnosed with MCI at baseline were considered. Participants whose follow-up dementia diagnosis was attributed to a non-Alzheimer's etiology (DXDDUE) were excluded to ensure that conversion events reflected progression to Alzheimer's disease. It is also noted on the ADNI study website that the screening procedures already aim to exclude subjects with non-Alzheimer's etiologies (adni.loni.usc.edu/data-samples). 36 subjects converted back to the CN group during the follow-up period and were excluded.

Although the m24 visit code was used to identify the follow-up assessment, the actual interval between baseline MRI acquisition and follow-up diagnosis varied across participants. Therefore, this interval was calculated directly from examination dates and restricted to a minimum of 2 years and a maximum of 3 years. Seven

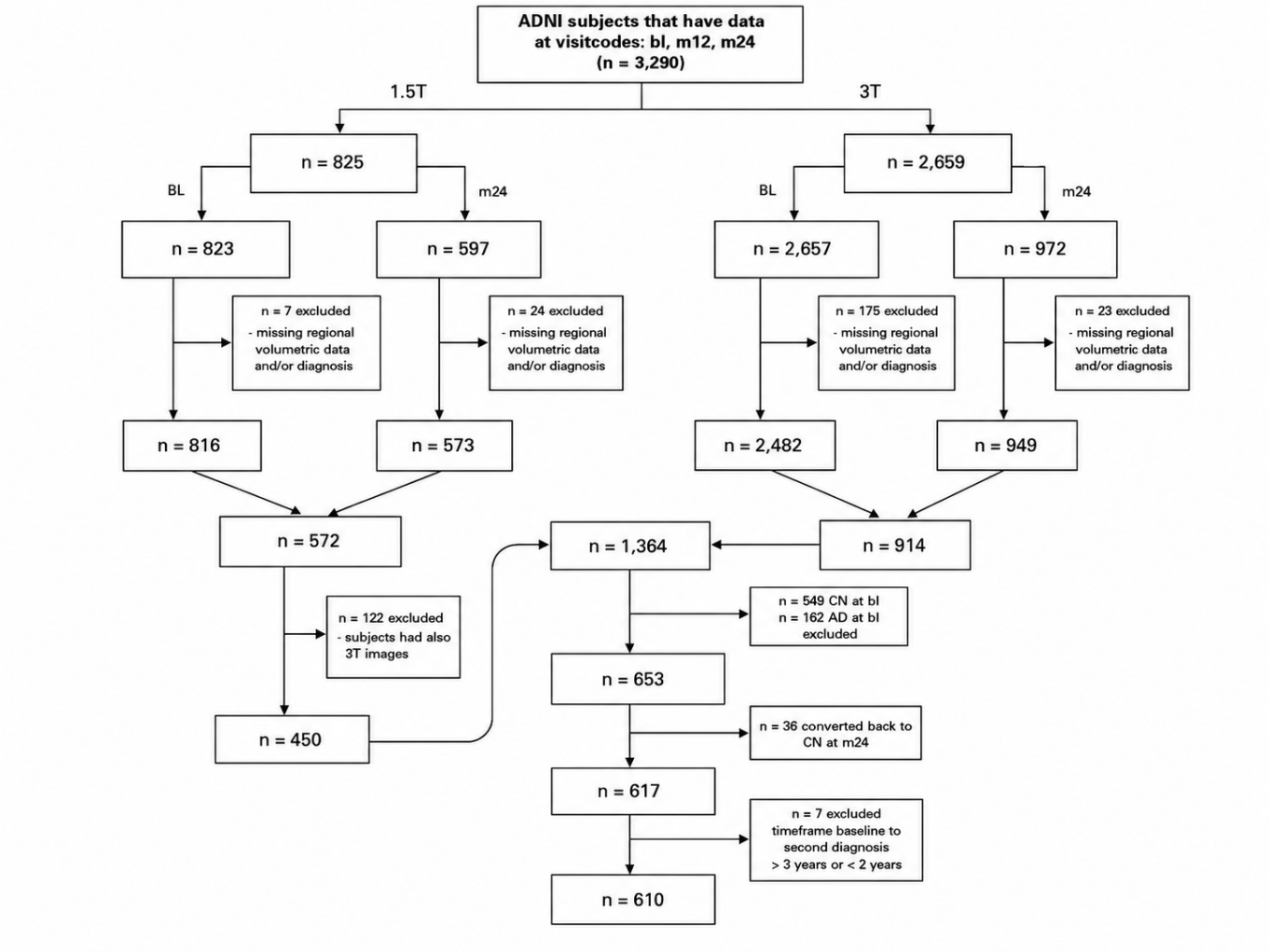

Figure 3.2: Flowchart of participant selection and exclusion procedures used to derive the final cohort from the ADNI dataset. Abbreviations: BL, baseline; T, Tesla.

subjects fell outside this interval and were therefore excluded from the analysis. The mean interval between the baseline MRI scan and the follow-up diagnosis was 2.1 years (SD 0.14 years). Participants diagnosed with AD at follow-up were classified as progressive MCI (pMCI), whereas those who remained diagnosed with MCI were classified as stable MCI (sMCI).

MMSE and ADAS-Cog 13 assessments were required to occur within 60 days before or after the baseline MRI scan. A previous study [44] defined the acceptable interval as less than six months. However, a stricter criterion was applied in

this thesis to ensure closer temporal alignment between cognitive assessments and MRI acquisition. Assessments obtained outside the 60-day window were treated as missing values, resulting in 17 missing MMSE scores and 28 missing ADAS-Cog 13 scores. Missing cognitive scores were treated as missing and imputed using multivariate imputation by chained equations (MICE). The imputation model included age, years of education, sex, MMSE score, and ADAS-Cog 13 score. Imputed MMSE values were rounded to the nearest integer and constrained to the valid range of 0–30, whereas imputed ADAS-Cog 13 values were constrained to the range of 0–85.

The final study cohort included 610 participants (Figure 3.2), including 444 sMCI and 166 pMCI cases (Table 3.1).

Table 3.1: Baseline characteristics of the MCI population. Group differences between sMCI and pMCI were assessed using the Mann–Whitney U test for numerical variables (age, education, MMSE, and ADAS-Cog 13) and the chi-squared test for the categorical variable (sex). To account for multiple comparisons, a Bonferroni-corrected significance threshold of $p<0.01$ was applied.

	All	**sMCI**	**pMCI**	**p-value**
ADNI, N	610	444	166	
Age [y]	72.87 (7.46)	72.48 (7.56)	73.91 (7.11)	0.0238
Education [y]	16.00 (2.74)	16.05 (2.73)	15.86 (2.76)	0.4109
MMSE	27.62 (1.82)	28.0 (1.71)	26.62 (1.73)	<0.0001
ADAS-Cog 13	16.56 (6.60)	14.72 (5.86)	21.49 (5.89)	<0.0001
Female [%]	40.98%	40.99%	40.96%	1.00

3.2 MRI data download and NIfTI conversion

All data were downloaded to the Puhti supercomputer provided by CSC Finland, where all image processing was performed up to the point of running the ML models. Once the imaging data was downloaded, the first step was to convert the DICOM (Digital Imaging and Communications in Medicine) data into NIfTI (Neuroimaging

Informatics Technology Initiative) format. DICOM is the standard output format for clinical MRI scanners, designed to comprehensively store individual 2D image slices into one DICOM file alongside extensive patient and acquisition metadata. However, due to its complexity, it is standard practice in neuroimaging related studies to convert these files into the NIfTI format. NIfTI combines multiple 2D image slices into a single 3D image file making it more efficient for computational analysis. [89]

To manage this large-scale conversion, the workflow was automated using Bash scripts executed within the Puhti supercomputing environment. The pipeline deployed the dcm2bids Apptainer container, which utilized dcm2niix version 1.0, which allowed the acquisition time to be appended to the image filenames, enabling them to be reliably matched with other MRI metadata later based on the acquisition time. For this thesis, all T1-weighted images from the entire ADNI study (approximately 22,000 images in total) were initially converted.

3.3 MRI data acquisition

The final analysis included both 1.5T and 3.0T T1-weighted images. The acquisition sequences predominantly corresponded to magnetization prepared rapid acquisition gradient echo (MPRAGE), but equivalent spoiled gradient echo with an inversion recovery preparation sequences (SPGR/FSPGR) were also included. Across the different phases of the study, brain scans were performed on scanners from multiple manufacturers. In this thesis, the final analysis included following scanner types Philips (105), Siemens (320), and GE Healthcare (185). Because ADNI is a multisite study, the final analysis included data from 62 different sites. Consequently, specific acquisition parameters varied across the cohort, such as repetition time (TR) ranging from 6.5ms to 3s, and echo time (TE) ranging from 2.8ms to 4.4ms. The slice thickness for the acquired T1-weighted sequences was predominantly 1.2 mm,

although a small number of scans had a thickness of 1.0 mm.

3.4 Structural image analysis

3.4.1 Segmentation

Segmentation was performed directly on the raw T1-weighted images using SynthSeg. SynthSeg [57] is a state-of-the-art segmentation tool that utilizes a CNN trained on a highly randomized synthetic dataset. This training strategy makes the model contrast-independent and highly robust to extreme variations in scanner types, field strengths, and voxel resolutions. Due to this inherent robustness, the raw images could be segmented directly without the need for image preprocessing steps. For this thesis, SynthSeg version 2.0 was utilized, with all computations performed on the Puhti supercomputer. In addition to generating the segmentation masks, SynthSeg automatically calculates regional volumetric measurements when -vol flag is added and these were extracted simultaneously during this step. Volumetric measurements were calculated from 101 distinct ROIs, comprising 48 ROIs from each hemisphere and unpaired structures, such as brain stem, CSF, the third and fourth ventricles.

3.4.2 Radiomics feature extraction

All image processing steps and radiomic feature extraction were performed using a custom-built Python (v3.10) pipeline, utilizing SimpleITK (v2.5.3) [90] for image manipulation and PyRadiomics [91] (pyradiomics.readthedocs.io/en/latest/, v3.0.1) for feature extraction. Pyradiomics is the most commonly utilized software package for radiomic research and is highly IBSI-compliant, with the majority of its extracted features aligning directly with the IBSI guidelines [80]. Similarly to the SynthSeg segmentation process, this pipeline was executed on the Puhti supercomputer to ensure computational efficiency across the large dataset.

Preprocessing steps

As discussed in section 2.3.2, preprocessing is one of the most important steps in radiomic analysis. Unlike the deep learning segmentation method, a precise preprocessing pipeline was required prior feature extraction to harmonize the structural T1-weighted imaging data, which originated from different scanners and acquisition protocols. To maximize data consistency, preprocessing steps were executed in the following order adhering to established best practices in current neuroimaging literature. [8], [17], [20], [84]

First, the raw T1-weighted images were resampled to match the spatial geometry (spacing, origin, and direction) of their corresponding SynthSeg segmentation masks. Because SynthSeg generates its output masks at an isotropic voxel size of $1\times1\times1$ mm^3, this step resampled the voxel sizes of the raw images to be uniform across the entire dataset. B-spline interpolation (tricubic) was used to preserve texture details and to prevent blockness that could be occur with nearest neighbour interpolation algorithm. This guaranteed voxel-wise alignment between the anatomical images and the ROI masks.

Intensity non-uniformities (bias fields) are a common source of variability in sMRI [19]. To correct low-frequency intensity non-uniformities caused by magnetic field inhomogeneities, the images were processed using the `N4BiasFieldCorrectionImageFilter` class provided by SimpleITK. The correction was performed on full-resolution images. A multi-resolution convergence strategy was employed, with a maximum of 100 iterations at each of four resolution levels ([100, 100, 100, 100]) and a convergence threshold of 0.001. A binary mask derived from SynthSeg segmentation was used to restrict bias field estimation to brain tissue, thereby reducing the influence of background noise.

Since T1-weighted MRI intensities are arbitrary, intensity normalization is an important step to harmonize scanner effects, as discussed in the section 2.3.2. z-

score normalization was applied to standardize the intensity distribution across all images. The mean (μ) and standard deviation (σ) were calculated exclusively from voxels within the brain mask to capture statistics of the brain tissue, ignoring the background and non-brain tissue. The image intensities were transformed using the following formula:

$$x' = \frac{x - \mu}{\sigma} \times 100$$

where x represents the intensity values in the raw image and x' is the normalized image. A scaling factor of 100 was applied to expand the dynamic range, ensuring numerical stability.

Feature extraction

Radiomic features were extracted from the preprocessed images using the PyRadiomics open-source library. To capture multi-scale texture information, features were calculated from both the original images and images filtered with a Laplacian of Gaussian (LoG) filter (spatial scaling factors $\sigma = 1.0, 2.0, 3.0$ mm) in the same way as in the [8].

Features were extracted from all available regions defined by the SynthSeg segmentation mask. To optimize computational performance and memory usage, the preCrop setting was enabled. This parameter crops the input image to the bounding box of the respective ROI prior to any feature extraction or filter application. While this cropping reduces memory consumption generally, it is fundamentally critical for optimizing the computationally heavy LoG filter convolutions. To prevent edge artifacts and ensure sufficient neighborhood context for the LoG filter [91], a padDistance of 10 voxels was applied around each bounding box.

Discretization of image intensities which is a crucial step for texture analysis was performed using a fixed bin width approach (binWidth = 5). This value was selected

based on the Z-scored intensity range of the normalized images (approximately -300 to +300). A bin width of 5 yields a bin count of 30 to 90 bins across ROIs. According to the PyRadiomics documentation, the optimal number of bins is 30 to 130, as this should ensure good reproducibility and performance [91].

For each ROI, the following PyRadiomics feature classes were extracted: shape features, first-order statistics, and gray level co-occurrence matrix (GLCM) features. Shape features were computed only from the original, unfiltered images, whereas the other two feature classes were extracted from both original and LoG-filtered images. The final dataset consisted of a total of 17,837 radiomic features per subject.

Prior to any model building, several anatomical regions were excluded from the radiomic dataset based on physiological relevance, methodological consistency, and statistical robustness. The following regions were removed from further analysis: CSF, brain stem, cerebellum, ventral diencephalon (DC), frontal pole, temporal pole and corpus callosum.

CSF was excluded because, as a fluid-filled region lacking cellular structure, texture-based radiomic features were not considered biologically meaningful. The brain stem and cerebellum were excluded to focus the analysis on cortical and subcortical regions more directly associated with AD-related neurodegeneration. The frontal pole, temporal pole, as well as ventral DC were excluded due to their relatively small size, which increases susceptibility to segmentation inaccuracies and unstable radiomic measurements. The corpus callosum was excluded to maintain consistency between the radiomic and volumetric feature sets. After these exclusions, the remaining radiomic dataset had 15,653 features.

3.5 Train-test split

To prevent data leakage and provide unbiased performance evaluation, the dataset was split into training (70%) and independent test (30%) sets using the `train_test-`

`_split()` function from scikit-learn [87]. Stratification was performed based on the outcome variable, defined as conversion status at follow-up, to preserve the proportions of sMCI and pMCI subjects between the training and test sets. A fixed random state was used to ensure reproducibility of the data split. The independent test set was not used during feature selection, model training, or hyperparameter optimization. As shown in Table 3.2, all demographic and clinical variables were comparable between the training and test sets. This demonstrates that the train-test split was executed successfully and that the test set accurately represents the training cohort.

Table 3.2: Baseline demographic and clinical characteristics of the training and test sets. Numerical variables are presented as mean (standard deviation) and compared using the Mann–Whitney U test. Categorical variables are presented as percentages and evaluated using the Chi-square test of independence. To account for multiple comparisons across the variables, a Bonferroni-corrected significance threshold of $p < 0.0125$ was applied.

Variables	**Training set**	**Test set**	**p-value**
N	427	183	–
Age(years)	72.68(7.37)	73.30(7.66)	0.380
MMSE	27.64(1.83)	27.59(1.79)	0.773
ADAS-Cog 13	16.56(6.76)	16.57(6.20)	0.735
Female[%]	40.28%	42.62%	0.653

In addition, Table 3.3 shows that the class proportions remain highly consistent between the training and test sets. Notably, while demographic variables such as age and sex show no significant difference after Bonferroni correction, the baseline cognitive assessments (MMSE and ADAS-Cog 13) exhibit highly significant variance between the sMCI and pMCI groups across both data splits.

Table 3.3: Baseline characteristics stratified by sMCI and pMCI groups within the training and test sets. Statistical tests and significance thresholds follow the exact methodology detailed in Table 3.2.

Variables	**Training set**			**Test set**		
	sMCI	**pMCI**	**p-value**	**sMCI**	**pMCI**	**p-value**
N	311	116	–	133	50	–
Age(years)	72.29(7.55)	73.74(6.81)	0.045	72.93(7.57)	74.32(7.82)	0.259
MMSE	28.05(1.69)	26.55(1.77)	<0.001	27.90(1.75)	26.78(1.66)	<0.001
ADAS-Cog 13	14.74(6.07)	21.41(6.11)	<0.001	14.65(5.35)	21.67(5.40)	<0.001
Female[%]	39.87%	41.38%	0.864	43.61%	40.00%	0.785

3.6 Prognostic models

To address the research questions, seven predictive models based on different combinations of imaging, demographic, and clinical features were constructed to classify sMCI versus pMCI subjects.

Two volumetric models (Vol, Vol+Cog) were constructed using structural brain volume measurements derived from SynthSeg. Although SynthSeg produced volumetric estimates for 101 brain regions, dimensionality was reduced by averaging bilateral structures and normalizing all volumes by total intracranial volume (TIV) to account for head-size variation. Throughout the rest of this thesis, the term 'volume' refers to TIV-corrected mean regional volumes unless otherwise specified. This feature engineering resulted in 52 regional volumetric features. To ensure consistency between volumetric and radiomics models, regions excluded from the radiomics analysis were also removed from the volumetric feature set, resulting in a final set of 45 volumetric regions.

Four radiomics models were constructed. Due to the high dimensionality of the radiomic feature space, these models required more stringent feature selection procedures, which are described in Section 3.7.2.

The first two radiomics models (Rad, Rad+Cog) used the radiomics feature

set extracted from a broad set of anatomically segmented bilateral brain regions described in Section 3.4.2. These features were not averaged across hemispheres or normalized by TIV, in contrast to the volumetric feature set.

The second two radiomic models (ROI-Rad, ROI-Rad+Cog) used a volumetry-informed ROI strategy. In this approach, a subset of brain regions was selected based on their association with MCI-to-AD conversion in the volumetric training data. Radiomic features were then filtered by selecting all features corresponding to the identified anatomical regions based on feature naming conventions. Specifically, all radiomic variables whose labels contained any of the selected region names were retained, forming a reduced feature space for the ROI-informed radiomics models.

This approach enabled comparison between a full radiomic feature space and a radiomic model focused on regions showing the strongest volumetric association with AD progression.

In addition to imaging-based models, a clinical model (Cog) was constructed using MMSE, ADAS-Cog 13, age, and sex. This model was included to evaluate the predictive value of clinical and demographic variables independently of imaging features and to assess whether cognitive measures improve prediction beyond imaging-derived features.

Table 3.4: Overview of the predictive models evaluated in this thesis. All models had demographics (age and sex) as input variables.

Model	**Input features**
Cog	MMSE, ADAS-Cog 13
Vol	TIV-corrected regional volumetric features
Vol+Cog	Vol features + MMSE, ADAS-Cog 13
Rad	Full radiomics feature set
Rad+Cog	Rad features + MMSE, ADAS-Cog 13
ROI-Rad	Radiomics features from volumetry-informed ROIs
ROI-Rad+Cog	ROI-Rad features + MMSE, ADAS-Cog 13

All seven predictive models used logistic regression (LR) with elastic net regularization as the final classification algorithm implemented in Python using the scikit-

learn ML package [87]. Random forest was evaluated as an alternative non-linear classifier during model development. As it achieved performance comparable to or slightly lower than LR, the latter was selected due to its simplicity and interpretability. This model choice strongly aligns with current literature, which recommends LR-based frameworks to ensure clinical trust and usability within a CDSS [23].

3.7 Model training and validation pipeline

3.7.1 Preprocessing

To ensure robust data transformations to the LR model and prevent data leakage, all preprocessing steps were embedded within a scikit-learn `Pipeline` utilizing a `ColumnTransformer`. Numerical variables, such as age, cognitive scores, volume features, and radiomic features were standardized to a mean of zero and unit variance using a `StandardScaler`. The categorical variable, sex, was transformed using a `OneHotEncoder`.

3.7.2 Feature selection

For the low-dimensional volume models and clinical model, feature selection was performed inherently during model training using the embedded properties of the elastic net penalty. Elastic net regularization combines both L_1 (Lasso) and L_2 (Ridge) penalties, making it highly advantageous for both volumetric and radiomic data, where features frequently exhibit strong multicollinearity [92]. The L_1 penalty enforces sparsity by shrinking the coefficients of uninformative features to exactly zero, effectively performing embedded feature selection. Simultaneously, the L_2 penalty stabilizes the model by grouping highly correlated features together rather than arbitrarily discarding them, as a pure Lasso penalty might do. [93]

Conversely, due to the extreme high dimensionality of the radiomic feature space,

relying solely on the embedded elastic net penalty was computationally impractical and prone to severe overfitting. While there is no universal consensus on the optimal feature selection strategy for radiomics, recent literature suggests that methods such as Lasso, ANOVA, and Minimum Redundancy Maximum Relevance (mRMR) should be considered first to achieve high predictive performance while effectively reducing model complexity [94]. Therefore, ANOVA and mRMR were integrated into a multi-step feature selection procedure to reduce dimensionality before classification.

The radiomic feature selection procedure consisted of three sequential stages. The first step was performed prior to the CV modeling pipeline, whereas the last two steps were executed exclusively on the training data during each CV iteration to prevent data leakage. First, to rapidly reduce dimensionality, an initial unsupervised variance filter was applied, discarding any features with a variance below 0.01. Because this step evaluates only the internal variance of the features and does not utilize the target outcome, it safely reduces the computational burden without introducing data leakage. The remaining features were then fed into the automated two-step feature selection pipeline. An analysis of variance (ANOVA) F-value filter (SelectKBest) was utilized to isolate the top 300 features most highly associated with the target outcome. Next, the mRMR algorithm implemented in the Feature-engine package [95] was performed. The mutual information difference (MID) criterion was used to identify an optimal subset of features, with the maximum number of selected features restricted to 12. The maximum feature count of 12 was chosen to maintain a conservative events-per-variable (EPV) ratio and to reduce overfitting risk [96], [97].

3.7.3 Cross-validation and hyperparameter tuning

To optimize hyperparameters while simultaneously providing a robust, unbiased estimation of model performance, a nested CV strategy was utilized for all models. As discussed in Section 2.4.3, nested CV reduces the risk of overly optimistic performance estimates by separating hyperparameter optimization from performance estimation.

The nested CV architecture consisted of an outer loop to evaluate generalizability and an inner loop dedicated strictly to hyperparameter tuning, both utilizing Stratified K-Fold CV to maintain class balance. For the volumetric and clinical models, the outer loop employed 10 folds, and the inner hyperparameter optimization was conducted using Bayesian Optimization `BayesSearchCV()` from the scikit-optimize (skopt) library for 30 iterations, optimizing directly for the AUC. In contrast, to maintain computational feasibility given the high-dimensional feature space, the radiomics models utilized a 5-fold outer loop and 20 optimization iterations. Across all models, the inner tuning loop was fixed at 5 folds.

The outer CV loop was used exclusively for performance estimation, whereas hyperparameter optimization was conducted within the corresponding inner folds. After completion of the nested CV procedure, Bayesian optimization was repeated on the full training set to determine the final hyperparameter configuration. The final model was then trained using the complete training set and evaluated on the independent hold-out test set.

3.8 Statistical analysis

The final predictive models were evaluated on the independent hold-out test set to assess their generalization performance on previously unseen data. Model performance was quantitatively assessed using a comprehensive set of classification metrics,

specifically the AUC as the primary performance metric, overall accuracy, balanced accuracy, sensitivity, and specificity. Binary classifications were determined by calculating an optimal decision threshold on the training set using Youden's J statistic ($J = \text{Sensitivity} + \text{Specificity} - 1$). This approach maximizes the true positive rate while minimizing the false positive rate [98]. The inclusion of balanced accuracy was particularly critical for this analysis, as it provides a robust evaluation metric that accounts for the inherent class imbalance between sMCI and pMCI subjects within the cohort [44].

To quantify uncertainty in model performance, 95% confidence intervals (CIs) were estimated using bootstrap resampling with 1,000 iterations. The lower and upper confidence bounds corresponded to the 2.5th and 97.5th percentiles of the bootstrap distributions, respectively. Finally, to directly address the core research questions of this thesis, pairwise comparisons of model AUCs were performed using DeLong's test to evaluate whether observed differences in discrimination performance were statistically significant.

4 Results

4.1 Structural MRI models

All imaging-based models achieved AUC values above 0.70 on the independent test set, indicating that baseline sMRI contained prognostic information regarding future conversion from MCI to AD.

The Vol model demonstrated moderate discriminative performance for predicting conversion from MCI to AD. In the training set, the model achieved a mean AUC of 0.76 ± 0.073. Evaluation on the independent test set resulted in an AUC of 0.73 (95% CI: 0.65–0.81), (Table 4.1).

Table 4.1: Predictive performance of the sMRI models on the independent test set. AUC values are presented with 95% CI obtained by bootstrap resampling.

Metrics	**Vol**	**Rad**	**ROI-Rad**
AUC (95% CI)	0.73 (0.65-0.81)	0.75 (0.67-0.83)	0.79 (0.70-0.86)
Balanced accuracy	0.69	0.71	0.73
Accuracy	0.69	0.73	0.70
Sensitivity	0.68	0.68	0.78
Specificity	0.69	0.74	0.68

18 volumetric regions were retained in the final LR model as predictive features. Figure 4.1 presents the standardized coefficients of the selected volumetric predictors. Positive coefficients indicate increased probability of progression, whereas negative coefficients indicate association with sMCI status. These 18 regions (Figure

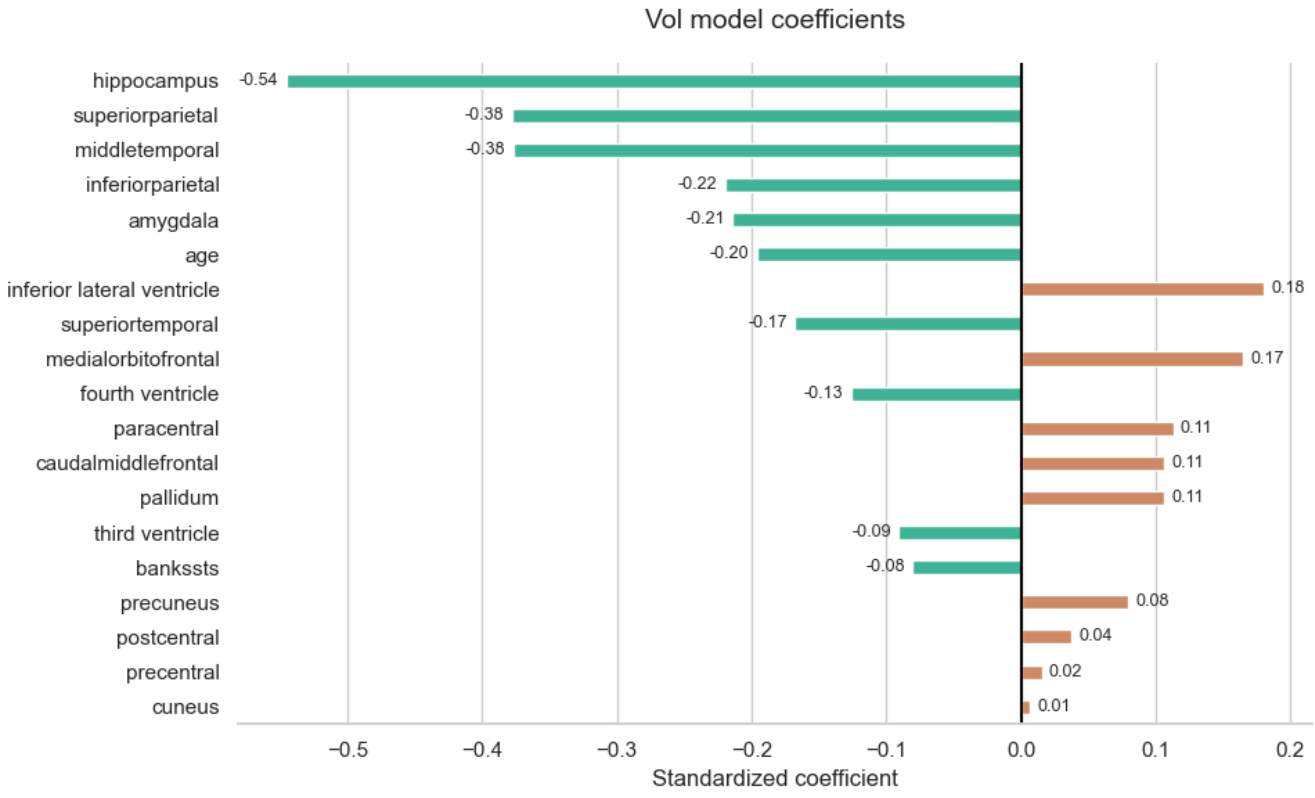

Figure 4.1: Standardized coefficients of the volumetric model.

4.1) were subsequently used to define the ROI-Rad feature set.

Although the Rad and ROI-Rad models demonstrated comparable performance during nested CV (mean AUC 0.78 $\pm$0.050 and 0.77 $\pm$0.038, respectively), ROI-Rad generalized better to the independent test set, achieving an AUC of 0.79 compared with AUC of 0.75 for Rad (Figure 4.2). Although the difference reached nominal significance according to DeLong's test ($p = 0.04$), it did not remain significant after Bonferroni correction ($\alpha < 0.0167$) (Table 4.2).

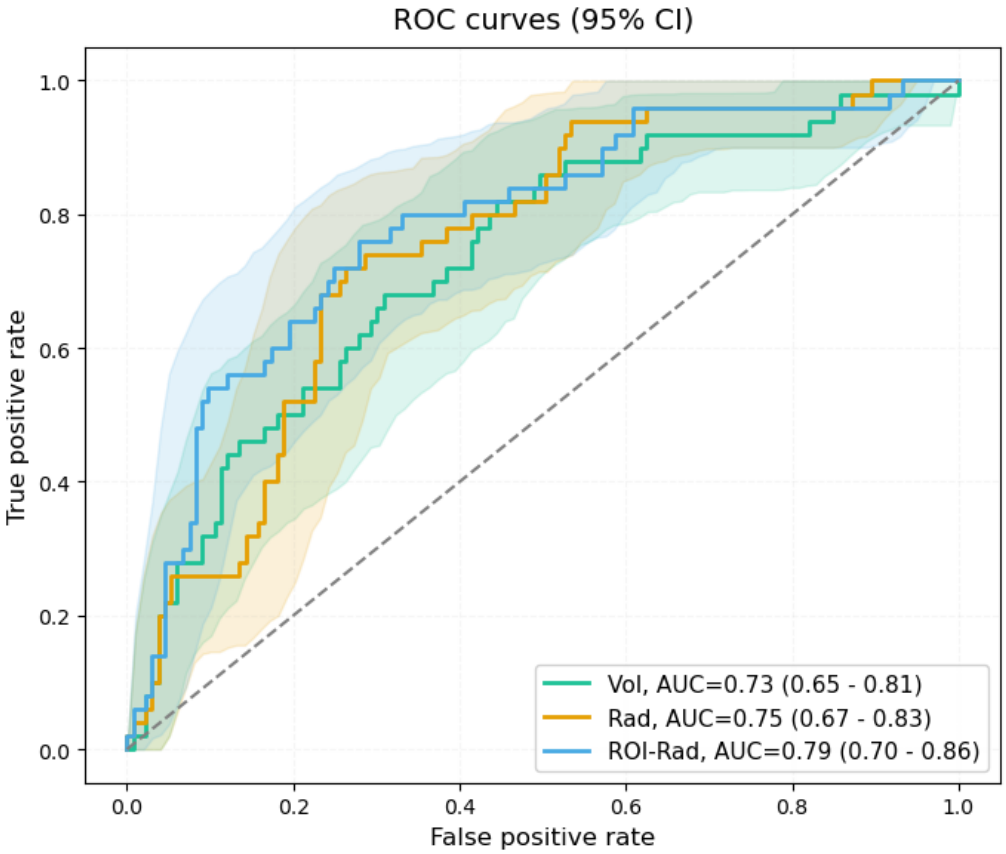

Figure 4.2: Receiver operating characteristic (ROC) curves for the Vol, Rad, and ROI-Rad models evaluated on the independent test set. Shaded areas indicate 95% CI.

4.2 Comparative evaluation of volumetric and radiomic models

When comparing the Rad model with the entire radiomic feature set as input to the Vol model, the radiomics approach yielded a marginal increase in performance (ΔAUC = 0.02) in the independent test set. However, this difference was not statistically significant (p = 0.515), indicating that the full radiomics model did not provide a significant prognostic advantage over conventional volumetric features.

Consequently, comparison of the optimized ROI-Rad model with the baseline Vol model demonstrated an increase in discrimination performance(ΔAUC = 0.06). Although this difference did not reach statistical significance according to DeLong's

Table 4.2: Pairwise comparative evaluation of the sMRI prognostic models using DeLong's test to address RQ2. Differences in AUC (ΔAUC) and corresponding p-values are reported for the independent test set. Because three pairwise model comparisons were performed, statistical significance was assessed using a Bonferroni-corrected threshold of $p < 0.0167$.

Models	**ΔAUC**	**p-value**
Vol vs Rad	0.02	0.515
Vol vs ROI-Rad	0.06	0.087
Rad vs ROI-Rad	0.04	0.043

test ($p = 0.087$), the ROI-Rad model consistently achieved higher performance across multiple evaluation metrics.

4.3 Effect of cognitive tests

As previously established in the baseline cohort characteristics (Tables 3.1 and 3.3), MMSE and ADAS-Cog 13 scores exhibited significant differences between the stable and progressing MCI groups, suggesting that these scores contain prognostic potential. Consistent with this observation, the integration of these cognitive metrics with the sMRI features generally improved predictive performance. As shown in Table 4.3, all multimodal models achieved an AUC of 0.80 or higher, with the ROI-Rad+Cog model achieving the highest overall predictive performance in this thesis (AUC = 0.82).

Furthermore, the multimodal models demonstrated consistent performance between nested CV and independent test set evaluation. Vol+Cog and Rad+Cog achieved mean CV AUCs of 0.82 ± 0.051 and 0.82 ± 0.045, respectively, while ROI-Rad+Cog achieved a mean CV AUC of 0.81 ± 0.049. The corresponding test set AUCs remained similar, suggesting good generalization performance and limited evidence of substantial overfitting (Figure 4.3).

In addition, the Cog model, without any imaging features, achieved an AUC of

Table 4.3: Predictive performance of the multimodal models on the independent test set. AUC values are presented with 95% CI obtained by bootstrap resampling.

Metrics	**Vol+Cog**	**Rad+Cog**	**ROI-Rad+Cog**
AUC (95% CI)	0.80 (0.73-0.87)	0.80 (0.74-0.87)	0.82 (0.75-0.89)
Balanced accuracy	0.70	0.72	0.74
Accuracy	0.67	0.73	0.74
Sensitivity	0.76	0.70	0.74
Specificity	0.64	0.74	0.74

0.81 (95% CI 0.74-0.87) on the independent test set. Among all models, this clinical model achieved the highest sensitivity of 0.80. However, its specificity (0.65) was comparable to that of the Vol+Cog model (0.64), representing the lowest specificity values among the evaluated models.

To address RQ2, pairwise DeLong's tests were conducted to assess whether the addition of cognitive measures resulted in statistically significant improvements in AUC on the independent test set (Table 4.4). A significant improvement was observed for the volumetric model (Vol vs. Vol+Cog, $p = 0.005$). In contrast, although the full radiomics model showed a higher AUC after the addition of cognitive measures (Rad vs. Rad+Cog, $p = 0.030$), the result did not remain statistically significant after Bonferroni correction (corrected significance threshold: $p < 0.0125$). These findings suggest that cognitive assessments provide complementary prognostic information beyond volumetric MRI features, whereas their added value for radiomics-based models was less evident in the present dataset.

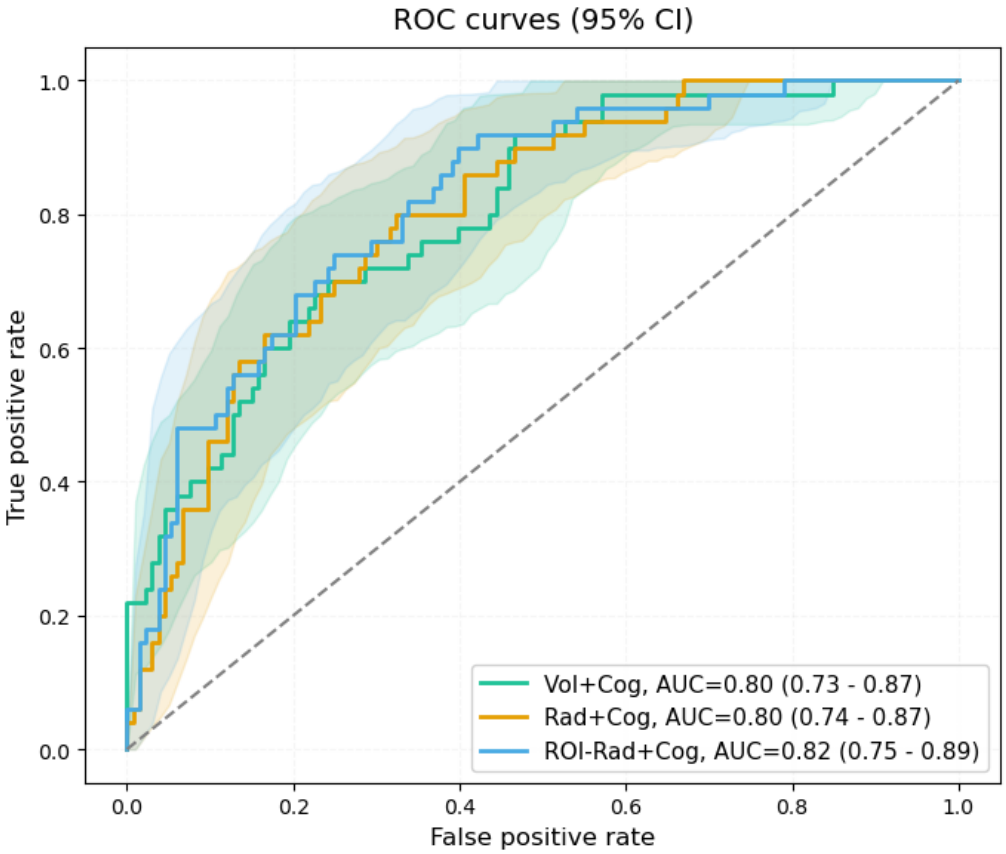

Figure 4.3: ROC curves for the Vol+Cog, Rad+Cog, and ROI-Rad+Cog models evaluated on the independent test set. Shaded areas indicate 95% CI.

Table 4.4: Pairwise comparison of imaging-based prognostic models using DeLong's test. Differences in AUC (ΔAUC) and corresponding p-values are reported for the independent test set. Because four pairwise model comparisons were performed, statistical significance was assessed using a Bonferroni- corrected threshold of $p < 0.0125$.

Models	**ΔAUC**	**p-value**
Vol vs Vol+Cog	0.07	0.005
Rad vs Rad+Cog	0.05	0.030
ROI-Rad vs ROI-Rad+Cog	0.03	0.181
Cog vs ROI-Rad+Cog	0.01	0.597

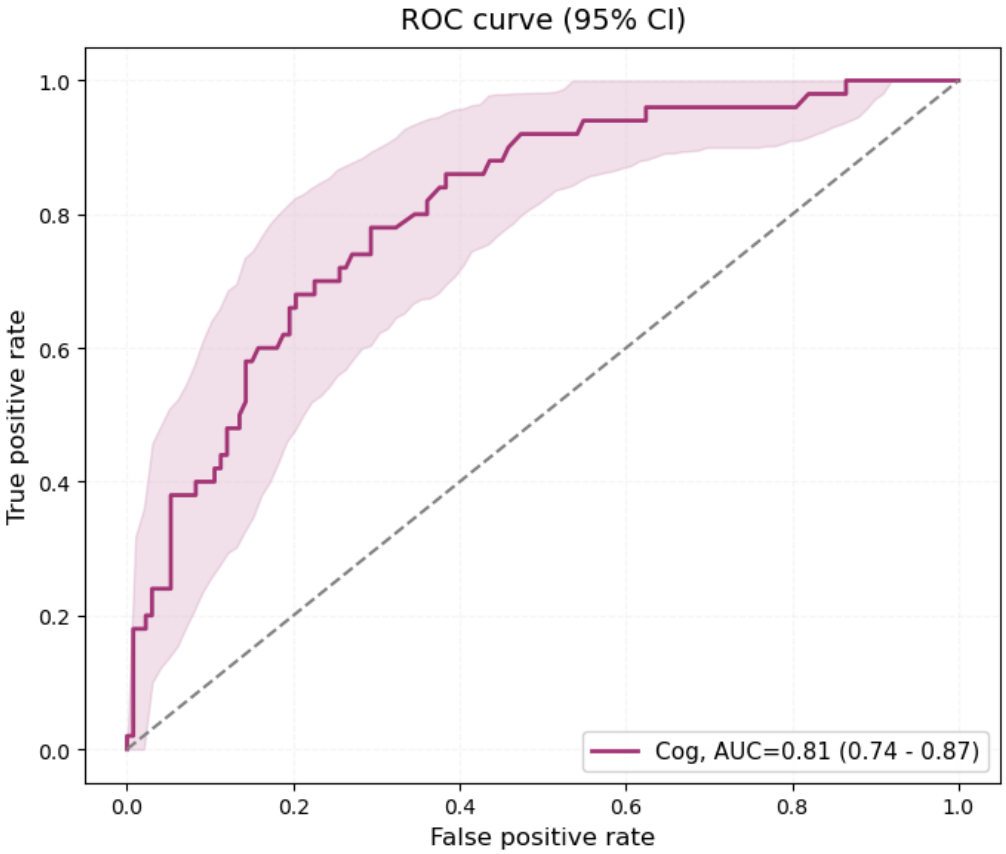

Figure 4.4: ROC curve for the Cog model evaluated on the independent test set. Shaded areas indicate 95% CI.

Moreover, this same trend was observed when cognitive scores were added to the highly optimized ROI-Rad model, the performance increase did not result in a statistically significant improvement ($p = 0.181$). This also suggests that the performance gain associated with cognitive information was smaller for the radiomics models.

Finally, to benchmark the multimodal approach against the clinical model, the highest performing ROI-Rad+Cog model was compared directly against the Cog model. While the multimodal model achieved higher absolute predictive performance, the difference was not statistically significant ($p = 0.597$).

4.4 Feature analysis

As shown in Figure 4.1, the largest coefficients in the Vol model were associated with the hippocampus, superior parietal cortex, middle temporal gyrus, inferior parietal cortex, and amygdala. These regions were also retained in the final Vol+Cog model, indicating that these regions remained important predictors of AD conversion even when cognitive test variables were included.

In the high-dimensional models, substantial overlap in selected predictors was observed between the Rad and ROI-Rad models. Most notably, four distinct radiomic features were consistently selected across all radiomics models, completely independent of regional constraints or the integration of clinical cognitive scores: the 10th percentile of the right amygdala (LoG 3.0mm), the least axis length of the left hippocampus, the GLCM contrast of the right middle temporal gyrus (LoG 1.0mm), and the 90th percentile of the left amygdala (LoG 1.0mm). In both Rad and ROI-Rad models the 10th percentile of the right amygdala and the least axis length of the left hippocampus exhibited the highest absolute coefficients, underscoring their primary prognostic importance.

Although volume-related radiomics features were available for all anatomical regions, only two regions and their mesh volume or voxel volume were also among predictors in the final radiomics models. A volume-based feature from the left middle temporal gyrus was retained in all radiomics models. Interestingly, while the mesh volume of the left entorhinal cortex was selected in both Rad models, this region was not retained in the volumetric models. In addition, in the Rad+Cog model, the first-order skewness of the left entorhinal cortex achieved a larger absolute coefficient than the MMSE score. However, across all other multimodal combinations of sMRI and cognitive test models, the ADAS-Cog 13 and MMSE scores received the largest absolute coefficients, being the most influential predictors when included in the models.

Detailed standardized coefficient plots for the final models are provided in Appendix A.

5 Discussion

The primary objective of this thesis was to evaluate and compare the prognostic value of conventional volumetric features, high-dimensional radiomic features, and established cognitive assessments in predicting the conversion from MCI to AD approximately two years after baseline. In all independent test set evaluations, predictive models demonstrated good discriminative capabilities between sMCI and pMCI groups, confirming that baseline sMRI and cognitive assessments contain critical prognostic information. All multimodal models achieved an AUC of 0.80 or higher. Among the evaluated models, ROI-Rad+Cog achieved the highest predictive performance.

When compared with previous studies, the baseline volumetric model developed in this thesis achieved performance that was broadly comparable to more complex XGBoost model in [44], which also included regional thickness measures. The relatively simple LR model, based on regional volume measurements calculated using the fast SynthSeg tool, was able to capture a substantial proportion of the prognostic signal available. In addition, the sMRI models demonstrated robust predictive validity, relying on anatomical features that strongly align with established pathways of AD progression [11], [13]. The biological plausibility of the Vol model is further supported by the retained predictors (Figure 4.1). Larger hippocampal volume was associated with the sMCI status, consistent with extensive evidence demonstrating hippocampal atrophy as one of the earliest and most robust imaging biomarkers of

AD progression [13]. Other highly weighted regions included the amygdala, middle temporal cortex, and parietal regions, all of which have been repeatedly linked to AD pathology and disease progression [11], [13], [99].

Regarding the comparison between volumetric and radiomics-based approaches, the results provide only partial support for the hypothesis that radiomic features improve prognostic performance. The radiomics model utilizing the full feature space (Rad) achieved only a modest improvement over the volumetric baseline, and this difference was not statistically significant. Although ROI-Rad achieved the highest AUC among the imaging-only models, the improvement over the volumetric model did not remain statistically significant after correction for multiple comparisons. These findings highlight both the potential and limitations of radiomics approaches. Although ROI-Rad achieved the highest imaging-only AUC, the lack of statistically significant improvement relative to the volumetric model suggests that any additional prognostic information provided by radiomics is modest. Importantly, these modest performance gains achieved by radiomics must be balanced against the increased methodological complexity required for feature extraction, dimensionality reduction, and model optimization [17], [20]. In contrast, volumetric measures are easier to compute, easier to interpret, and more readily transferable to clinical practice. Therefore, the incremental predictive benefit of radiomics should be considered alongside its additional computational and methodological burden.

An important observation was that the ROI-Rad approach was not based on prior biological assumptions alone. Instead, the regions were identified through the volumetric training data and subsequently used to constrain the ROI feature set. The resulting performance improvement suggests that in this thesis volumetric measures appear useful for identifying anatomically relevant regions, whereas radiomic features provide additional characterization of tissue properties within those regions.

The feature analysis further supports the hypothesis that radiomic features

may provide additional prognostic information beyond conventional volumetric measurements. In radiomic models, the most influential predictors repeatedly originated from structures classically associated with AD pathology, including the hippocampus, amygdala, EC and middle temporal regions. The previously mentioned study [44], which examined volume and thickness measurements, found that EC thickness, in particular, is associated with the progression of MCI to AD. An intriguing finding was the appearance of EC radiomic features in the final Rad and Rad+Cog models despite the absence of entorhinal volume measures in the final volumetric models. The fact that the skewness of the left EC captured from the original image had an even higher coefficient than the MMSE score coefficient in the Rad+Cog model is a noteworthy observation, because the EC together with the hippocampus is widely recognized as one of the first regions affected by AD pathology, due to tau protein related accumulation occurring in this region during the earliest stages before widespread cortical involvement [13], [99]. The right EC mesh volume was also retained in the final Rad+Cog model. One possible interpretation of this is that radiomic features capture subtle tissue-level alterations within the entorhinal cortices that are not captured in the TIV-corrected and averaged EC volumetric measures. However, this hypothesis cannot be verified within the present thesis and requires further validation.

In addition, when the performance between the Rad and ROI-Rad models is compared, a fundamental challenge related to high-dimensional ML known as the curse of dimensionality can be observed [16]. The initial radiomic feature space comprised more than 15,000 variables for the training set of 427. This $p >> n$ paradigm [80], increases the likelihood that the algorithm will identify false correlations that exist only in the training data. While nested CV and feature selection methods were employed to mitigate this, the full Rad model remained susceptible to overfitting due to the overwhelming ratio of noise to potential signal, which may

have contributed to slight overfitting of the Rad model, resulting in a small drop in performance when evaluated on the independent test set.

However, the ROI-Rad approach succeeded in overcoming this limitation. From an ML perspective, integrating this domain knowledge effectively acts as a form of regularization. By artificially restricting the hypothesis space to regions exhibiting macro-structural atrophy, the pipeline effectively suppressed high-dimensional noise. This intervention appears to have improved the bias-variance tradeoff [80]. It introduced a biologically sound inductive bias that substantially reduced the model's variance. Consequently, the ROI-Rad model demonstrated superior generalization capabilities on the independent hold-out test set compared to the Rad model. This suggests that when using high-dimensional neuroimaging data, relying solely on automated feature selection is often insufficient, and that hybrid pipelines leveraging domain-specific structural priors are essential for building robust, generalizable prognostic models.

Compared with the two previous radiomics studies [7], [8], this thesis extracted radiomic features from a substantially broader range of brain regions. Both studies focused on WM and GM for radiomic feature extraction, while study [7] also included CSF. Despite these differences in feature extraction regions, some radiomic features were consistently identified across studies. Notably, cluster shade was included in the radiomic signatures reported in both previous studies and was also retained in all final PyRadiomics models developed in this thesis, suggesting that this texture feature may capture disease-related image heterogeneity that is robust across different extraction approaches.

Direct comparison of model performance between studies is challenging because radiomic features were extracted from different anatomical regions, using different software packages, and incorporated into different ML pipelines. For example, study [8] generated a radiomics signature using PyRadiomics features that was sub-

sequently used to calculate a radiomics score for a tree-based classification model. Although study [8] reported a higher AUC than the ROI-Rad+Cog model developed in this thesis, the ROI-Rad model achieved a higher AUC than the radiomics signatures reported in studies [7] and [8] (AUC = 0.69 and AUC = 0.78, respectively). These findings may indicate that extracting radiomic features from specific smaller anatomical regions can provide useful discriminatory information and with further optimization of the algorithms, radiomic features extracted and larger training set, this approach could evolve into an exceptionally effective predictive tool. Nevertheless, methodological differences between the studies limit the ability to draw direct conclusions regarding which approach is superior.

Another notable finding was the strong performance of the cognitive assessments. Consistent with previous studies [8], [77], the inclusion of MMSE and ADAS-Cog 13 generally improved predictive performance, particularly for the volumetric model. However, the performance gain was smaller for the radiomic models and did not remain statistically significant after correction for multiple comparisons.

The cognitive-only model also demonstrated remarkably strong performance, achieving predictive accuracy comparable to that of the best multimodal model. No statistically significant difference was observed between the Cog and ROI-Rad+Cog models. The strong performance of the Cog model may reflect the fact that neuropsychological assessments directly capture the functional consequences of neurodegeneration. While imaging biomarkers quantify structural changes, cognitive tests measure the clinical expression of these changes. Consequently, cognitive assessments may remain highly effective predictors over relatively short follow-up intervals such as the approximately two-year period examined in this thesis.

A major methodological strength of this thesis lies in its efficient and balanced computational pipeline, which integrated DL for image segmentation with inherently explainable ML for clinical prognostic classification. Traditional sMRI segmentation

pipelines, such as FreeSurfer, are computationally very heavy as discussed in Section 2.3.1, in contrast to the fast SynthSeg segmentation. To further mitigate computational bottlenecks, the entire image processing and radiomic extraction pipeline was executed on the Puhti supercomputer (CSC Finland). While deploying high-throughput custom Bash and Python pipelines in a high-performance computing environment required a more complex technical implementation, it significantly accelerated processing times for the large ADNI cohort.

However, several limitations regarding the thesis cohort and data pipeline must be acknowledged. First, the sample size remains relatively modest for a high-dimensional radiomics analysis, and the inherent class imbalance limits statistical power, even though the pMCI cohort was marginally larger than in comparable benchmark studies [7], [8]. Furthermore, while an independent test set was used, all participants originated from the ADNI cohort. Therefore, external validation in independent clinical populations is required before a broader clinical deployment.

Second, critical technical limitations relate to the radiomics extraction pipeline. The MRI data were acquired across multiple sites using diverse scanners and protocols. While this reflects real-world clinical heterogeneity, radiomic features are very sensitive to acquisition parameters and scanner-specific effect [16], [20]. Although careful preprocessing was performed to reduce these effects, some scanner- and site-related noise inevitably remains. For example, the ComBat method could have been used to further reduce the scanner effect [19]. Additionally, there were some minor methodological challenges between the segmentation phase and the feature extraction phase. For instance, radiomic features were calculated from separately resampled images rather than the exact intermediate resampled images generated by SynthSeg during segmentation. Furthermore, because explicit skull-stripping was omitted prior to extraction, and image preprocessing was performed post-segmentation, subtle variations in the segmentation masks and resulting radiomic

features cannot be ruled out. Finally, as radiomic features are highly software-dependent, direct comparisons with studies utilizing tools other than PyRadiomics should be interpreted with caution.

While the predictive modeling intentionally emphasized explainable ML, this approach serves as both a strength and a limitation. Rather than employing complex "black-box" algorithms for the final models, the use of LR with elastic net penalty provided transparent, standardized feature coefficients. In the context of clinical decision support system, this interpretability is a critical requirement [23]. Nevertheless, this may not represent the optimal modeling strategy. More sophisticated ML approaches may potentially capture more complex patterns and possibly achieve higher accuracy. For example, decision tree [8], [80] or DL architecture [100] may have achieved different levels of predictive performance. Therefore, the present results should not be interpreted as evidence that LR represents the absolute upper bound of predictive accuracy for radiomic-based models AD prognosis.

The findings of this thesis open several doors for future studies. Future engineering efforts should focus on integrating the most optimal Vol+Cog model into a fully automated CDSS capable of retrieving MRI data and relevant non-invasive clinical information directly from hospital information systems. This pipeline could provide clinicians with an interpretable prognostic risk estimate while minimizing additional workflow burden.

Second, confirming the predictive value of the EC radiomics requires more comprehensive biological validation. Since the EC is classically associated with the earliest accumulations of tau protein, future multimodal studies should combine sMRI radiomic signatures with molecular imaging, such as Tau-PET. Establishing a direct correlation between radiomic texture alterations and molecular pathology would fundamentally validate radiomics as a proxy for early microstructural neurodegeneration.

Finally, this thesis focused exclusively on baseline prognostic biomarkers. Future computational studies should explore longitudinal ML models to analyze how radiomic features change over time. By incorporating the trajectory of morphological and textural changes across multiple time points, the clinical understanding of progressive neurodegeneration and predictive accuracy could be further enhanced.

6 Conclusions

The results of this thesis demonstrate that baseline sMRI features can be used to predict progression from MCI to AD approximately two years after baseline, confirming their value as prognostic imaging biomarkers. Both sMRI and cognitive assessments provide valuable prognostic information. Although high-dimensional radiomic features achieved competitive predictive performance compared to conventional volumetric features, they did not yield statistically significant improvements. When evaluating multimodal models, adding cognitive assessments to the volumetric model (Vol) significantly improved the performance of the model, whereas no significant improvements were observed in the radiomic models. The strongest predictive performance of the combined region of interest radiomics and cognitive tests model (ROI-Rad+Cog) underscores the potential of radiomics when features are extracted from relevant brain regions. However, the model with volumetric features and cognitive assessments provided the most favorable balance between predictive performance, interpretability, and implementation complexity. Therefore, the present findings support this multimodal model as the most practical non-invasive approach for the future development of CDSS.

References

[1] World Health Organization, *Global Action Plan on the Public Health Response to Dementia 2017–2025*. World Health Organization, 2017.

[2] M. J. Garcia, R. Leadley, J. Ross, *et al.*, "Prognostic and predictive factors in early alzheimer's disease: A systematic review", *Journal of Alzheimer's Disease Reports*, vol. 8, no. 1, pp. 203–240, 2024.

[3] L. Jönsson, A. Tate, O. Frisell, and A. Wimo, "The costs of dementia in europe: An updated review and meta-analysis", *Pharmacoeconomics*, vol. 41, no. 1, pp. 59–75, 2023.

[4] S. Grueso and R. Viejo-Sobera, "Machine learning methods for predicting progression from mild cognitive impairment to alzheimer's disease dementia: A systematic review", *Alzheimer's research & therapy*, vol. 13, no. 1, p. 162, 2021.

[5] N. C. Fox, C. Belder, C. Ballard, *et al.*, "Treatment for alzheimer's disease", *The Lancet*, vol. 406, no. 10510, pp. 1408–1423, 2025.

[6] R. C. Petersen, "Mild cognitive impairment", *CONTINUUM: lifelong Learning in Neurology*, vol. 22, no. 2, pp. 404–418, 2016.

[7] Z.-Y. Shu, D.-W. Mao, Y.-y. Xu, Y. Shao, P.-P. Pang, and X.-Y. Gong, "Prediction of the progression from mild cognitive impairment to alzheimer's

disease using a radiomics-integrated model", *Therapeutic Advances in Neurological Disorders*, vol. 14, 2021.

[8] Y. Li, P. Yi, M. Jin, Y. Li, and W. Chen, "A radiomics model predicts progression from mild cognitive impairment to alzheimer's disease using structural mri", *Scientific Reports*, vol. 15, no. 1, p. 35 679, 2025.

[9] Alzheimer's Association, "2019 alzheimer's disease facts and figures", *Alzheimer's & dementia*, vol. 15, no. 3, pp. 321–387, 2019.

[10] G. B. Frisoni, N. C. Fox, C. R. Jack Jr, P. Scheltens, and P. M. Thompson, "The clinical use of structural mri in alzheimer disease", *Nature reviews neurology*, vol. 6, no. 2, pp. 67–77, 2010.

[11] P. Huang and M. Zhang, "Magnetic resonance imaging studies of neurodegenerative disease: From methods to translational research", *Neuroscience Bulletin*, vol. 39, no. 1, pp. 99–112, 2023.

[12] C. A. Raji and T. L. Benzinger, "The value of neuroimaging in dementia diagnosis", *CONTINUUM: Lifelong Learning in Neurology*, vol. 28, no. 3, pp. 800–821, 2022.

[13] L. Pini, M. Pievani, M. Bocchetta, *et al.*, "Brain atrophy in alzheimer's disease and aging", *Ageing Research Reviews*, vol. 30, pp. 25–48, 2016, Brain Imaging and Aging.

[14] X. Zheng, J. Cawood, C. Hayre, S. Wang, and A. D. N. I. Group, "Computer assisted diagnosis of alzheimer's disease using statistical likelihood-ratio test", *Plos one*, vol. 18, no. 2, 2023.

[15] H. Zhou, J. Jiang, J. Lu, *et al.*, "Dual-model radiomic biomarkers predict development of mild cognitive impairment progression to alzheimer's disease", *Frontiers in neuroscience*, vol. 12, p. 1045, 2019.

[16] N. Beig, K. Bera, and P. Tiwari, "Introduction to radiomics and radiogenomics in neuro-oncology: Implications and challenges", *Neuro-Oncology Advances*, vol. 2, no. Supplement_4, pp. iv3–iv14, 2020.

[17] A. Zwanenburg, M. Vallières, M. A. Abdalah, *et al.*, "The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping", *Radiology*, vol. 295, no. 2, pp. 328–338, 2020.

[18] R. Shahidi, M. Baradaran, A. Asgarzadeh, *et al.*, "Diagnostic performance of mri radiomics for classification of alzheimer's disease, mild cognitive impairment, and normal subjects: A systematic review and meta-analysis", *Aging clinical and experimental research*, vol. 35, no. 11, pp. 2333–2348, 2023.

[19] Y. Li, S. Ammari, C. Balleyguier, N. Lassau, and E. Chouzenoux, "Impact of preprocessing and harmonization methods on the removal of scanner effects in brain mri radiomic features", *Cancers*, vol. 13, no. 12, 2021.

[20] V. Trojani, M. C. Bassi, L. Verzellesi, and M. Bertolini, "Impact of preprocessing parameters in medical imaging-based radiomic studies: A systematic review", *Cancers*, vol. 16, no. 15, p. 2668, 2024.

[21] D.-H. Shih, Y.-H. Wu, T.-W. Wu, Y.-K. Wang, and M.-H. Shih, "Classifying dementia severity using mri radiomics analysis of the hippocampus and machine learning", *IEEe Access*, vol. 12, pp. 160 030–160 051, 2024.

[22] S. Aghajanian, F. Mohammadifard, I. Mohammadi, *et al.*, "Longitudinal structural mri-based deep learning and radiomics features for predicting alzheimer's disease progression", *Alzheimer's Research & Therapy*, vol. 17, no. 1, p. 182, 2025.

[23] Q. Xu, W. Xie, B. Liao, *et al.*, "Interpretability of clinical decision support systems based on artificial intelligence from technological and medical per-

spective: A systematic review", *Journal of healthcare engineering*, vol. 2023, no. 1, 2023.

[24] D. P. Veitch, M. W. Weiner, M. Miller, *et al.*, "The alzheimer's disease neuroimaging initiative in the era of alzheimer's disease treatment: A review of adni studies from 2021 to 2022", *Alzheimer's & Dementia*, vol. 20, no. 1, pp. 652–694, 2024.

[25] S. Grueso and R. Viejo-Sobera, "Machine learning methods for predicting progression from mild cognitive impairment to alzheimer's disease dementia: A systematic review.", *Alzheimer's research and therapy*, vol. 13,1, 2021.

[26] J. S. Benoit, W. Chan, L. Piller, and R. Doody, "Longitudinal sensitivity of alzheimer's disease severity staging", *American Journal of Alzheimer's Disease & Other Dementias*, vol. 35, 2020.

[27] B. Dubois, H. Hampel, H. H. Feldman, *et al.*, "Preclinical alzheimer's disease: Definition, natural history, and diagnostic criteria", *Alzheimer's & Dementia*, vol. 12, no. 3, pp. 292–323, 2016.

[28] K. Palmer, L. Fratiglioni, and B. Winblad, "What is mild cognitive impairment? variations in definitions and evolution of nondemented persons with cognitive impairment", *Acta Neurologica Scandinavica*, vol. 107, pp. 14–20, 2003.

[29] J. A. Gutiérrez-Vargas, J. F. Castro-Álvarez, J. F. Zapata-Berruecos, K. Abdul-Rahim, and A. Arteaga-Noriega, "Neurodegeneration and convergent factors contributing to the deterioration of the cytoskeleton in alzheimer's disease, cerebral ischemia and multiple sclerosis", *Biomedical Reports*, vol. 16, no. 4, p. 27, 2022.

[30] C. Reitz, L. Honig, J. P. Vonsattel, M. X. Tang, and R. Mayeux, "Memory performance is related to amyloid and tau pathology in the hippocampus",

Journal of Neurology, Neurosurgery & Psychiatry, vol. 80, no. 7, pp. 715–721, 2009.

[31] J. P. O'connor, E. O. Aboagye, J. E. Adams, *et al.*, "Imaging biomarker roadmap for cancer studies", *Nature reviews Clinical oncology*, vol. 14, no. 3, pp. 169–186, 2017.

[32] S. Teipel, A. Drzezga, M. J. Grothe, *et al.*, "Multimodal imaging in alzheimer's disease: Validity and usefulness for early detection", *The Lancet Neurology*, vol. 14, no. 10, pp. 1037–1053, 2015.

[33] J. M. Fernández Montenegro, B. Villarini, A. Angelopoulou, E. Kapetanios, J. Garcia-Rodriguez, and V. Argyriou, "A survey of alzheimer's disease early diagnosis methods for cognitive assessment", *Sensors*, vol. 20, no. 24, p. 7292, 2020.

[34] I.-H. Oh, W.-R. Shin, J. Ahn, *et al.*, "The present and future of minimally invasive methods for alzheimer's disease diagnosis", *Toxicology and Environmental Health Sciences*, vol. 14, no. 4, pp. 309–318, 2022.

[35] K. Blennow and H. Zetterberg, "Biomarkers for alzheimer's disease: Current status and prospects for the future", *Journal of internal medicine*, vol. 284, no. 6, pp. 643–663, 2018.

[36] J. M. Costerus, M. C. Brouwer, and D. van de Beek, "Technological advances and changing indications for lumbar puncture in neurological disorders", *The Lancet Neurology*, vol. 17, no. 3, pp. 268–278, 2018.

[37] G. Garcia-Escobar, R. M. Manero, A. Fernández-Lebrero, *et al.*, "Blood biomarkers of alzheimer's disease and cognition: A literature review", *Biomolecules*, vol. 14, no. 1, p. 93, 2024.

[38] H. Hampel, S. E. O'Bryant, J. L. Molinuevo, *et al.*, "Blood-based biomarkers for alzheimer disease: Mapping the road to the clinic", *Nature Reviews Neurology*, vol. 14, no. 11, pp. 639–652, 2018.

[39] A. Marcisz, A. D. N. Initiative, and J. Polanska, "Can t1-weighted magnetic resonance imaging significantly improve mini-mental state examination-based distinguishing between mild cognitive impairment and early-stage alzheimer's disease?", *Journal of Alzheimer's Disease*, vol. 92, no. 3, pp. 941–957, 2023.

[40] C. Suh, W. Shim, S. Kim, *et al.*, "Development and validation of a deep learning–based automatic brain segmentation and classification algorithm for alzheimer disease using 3d t1-weighted volumetric images", *American Journal of Neuroradiology*, vol. 41, no. 12, pp. 2227–2234, 2020.

[41] M. Ansart, S. Epelbaum, G. Bassignana, *et al.*, "Predicting the progression of mild cognitive impairment using machine learning: A systematic, quantitative and critical review", *Medical Image Analysis*, vol. 67, p. 101 848, 2021.

[42] S. Balsis, J. F. Benge, D. A. Lowe, L. Geraci, and R. S. Doody, "How do scores on the adas-cog, mmse, and cdr-sob correspond?", *The Clinical Neuropsychologist*, vol. 29, no. 7, pp. 1002–1009, 2015.

[43] T. Tong, P. Thokala, B. McMillan, R. Ghosh, and J. Brazier, "Cost effectiveness of using cognitive screening tests for detecting dementia and mild cognitive impairment in primary care", *International journal of geriatric psychiatry*, vol. 32, no. 12, pp. 1392–1400, 2017.

[44] M. Mieling, M. Yousuf, and N. Bunzeck, "Predicting the progression of mci and alzheimer's disease on structural brain integrity and other features with machine learning", *GeroScience*, pp. 1–25, 2025.

[45] A. J. Mitchell, "The mini-mental state examination (mmse): Update on its diagnostic accuracy and clinical utility for cognitive disorders", in *Cognitive screening instruments: A practical approach*, Springer, 2017, pp. 37–48.

[46] J. K. Kueper, M. Speechley, and M. Montero-Odasso, "The alzheimer's disease assessment scale–cognitive subscale (adas-cog): Modifications and responsiveness in pre-dementia populations. a narrative review", *Journal of Alzheimer's Disease*, vol. 63, no. 2, pp. 423–444, 2018.

[47] J. Podhorna, T. Krahnke, M. Shear, J. E Harrison, and A. D. N. Initiative, "Alzheimer's disease assessment scale–cognitive subscale variants in mild cognitive impairment and mild alzheimer's disease: Change over time and the effect of enrichment strategies", *Alzheimer's research & therapy*, vol. 8, no. 1, p. 8, 2016.

[48] K. K. Tsoi, J. Y. Chan, H. W. Hirai, S. Y. Wong, and T. C. Kwok, "Cognitive tests to detect dementia: A systematic review and meta-analysis", *JAMA internal medicine*, vol. 175, no. 9, pp. 1450–1458, 2015.

[49] D. B. Plewes and W. Kucharczyk, "Physics of mri: A primer", *Journal of magnetic resonance imaging*, vol. 35, no. 5, pp. 1038–1054, 2012.

[50] H. Yang, C. Dong, Y. Cai, *et al.*, "Advances in the use of structural and diffusion magnetic resonance imaging for characterizing scd and mci due to alzheimer's disease", *Frontiers in Neuroscience*, vol. 19, p. 1 596 459, 2025.

[51] S. M. Nestor, R. Rupsingh, M. Borrie, *et al.*, "Ventricular enlargement as a possible measure of alzheimer's disease progression validated using the alzheimer's disease neuroimaging initiative database", *Brain*, vol. 131, no. 9, pp. 2443–2454, 2008.

[52] A. Fleisher, S. Sun, C. Taylor, *et al.*, "Volumetric mri vs clinical predictors of alzheimer disease in mild cognitive impairment", *Neurology*, vol. 70, no. 3, pp. 191–199, 2008.

[53] D. Devanand, G. Pradhaban, X. Liu, *et al.*, "Hippocampal and entorhinal atrophy in mild cognitive impairment: Prediction of alzheimer disease", *Neurology*, vol. 68, no. 11, pp. 828–836, 2007.

[54] M. C. Evans, J. Barnes, C. Nielsen, *et al.*, "Volume changes in alzheimer's disease and mild cognitive impairment: Cognitive associations", *European radiology*, vol. 20, no. 3, pp. 674–682, 2010.

[55] I. Despotović, B. Goossens, and W. Philips, "Mri segmentation of the human brain: Challenges, methods, and applications", *Computational and mathematical methods in medicine*, vol. 2015, no. 1, 2015.

[56] M. A. Balafar, A. R. Ramli, M. I. Saripan, and S. Mashohor, "Review of brain mri image segmentation methods", *Artificial Intelligence Review*, vol. 33, no. 3, pp. 261–274, 2010.

[57] B. Billot, D. N. Greve, O. Puonti, *et al.*, "Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining", *Medical image analysis*, vol. 86, 2023.

[58] F. Klauschen, A. Goldman, V. Barra, A. Meyer-Lindenberg, and A. Lundervold, "Evaluation of automated brain mr image segmentation and volumetry methods", Wiley Online Library, Tech. Rep., 2009.

[59] E. Khadhraoui, T. Nickl-Jockschat, H. Henkes, D. Behme, and S. J. Müller, "Automated brain segmentation and volumetry in dementia diagnostics: A narrative review with emphasis on freesurfer", *Frontiers in Aging Neuroscience*, vol. 16, 2024.

[60] K. M. Pohl, J. Fisher, W. E. L. Grimson, R. Kikinis, and W. M. Wells, "A bayesian model for joint segmentation and registration", *NeuroImage*, vol. 31, no. 1, pp. 228–239, 2006.

[61] T. N. Akudjedu, L. Nabulsi, M. Makelyte, *et al.*, "A comparative study of segmentation techniques for the quantification of brain subcortical volume", *Brain imaging and behavior*, vol. 12, no. 6, pp. 1678–1695, 2018.

[62] A. Ramani, J. H. Jensen, and J. A. Helpern, "Quantitative mr imaging in alzheimer disease", *Radiology*, vol. 241, no. 1, pp. 26–44, 2006.

[63] P. Koussis, P. Toulas, D. Glotsos, E. Lamprou, D. Kehagias, and E. Lavdas, "Reliability of automated brain volumetric analysis: A test by comparing neuroquant and volbrain software", *Brain and behavior*, vol. 13, no. 12, 2023.

[64] K. Zhao, Y. Ding, Y. Han, *et al.*, "Independent and reproducible hippocampal radiomic biomarkers for multisite alzheimer's disease: Diagnosis, longitudinal progress and biological basis", *Science Bulletin*, vol. 65, no. 13, pp. 1103–1113, 2020.

[65] D. H. Mathalon, E. V. Sullivan, J. M. Rawles, and A. Pfefferbaum, "Correction for head size in brain-imaging measurements", *Psychiatry Research: Neuroimaging*, vol. 50, no. 2, pp. 121–139, 1993.

[66] C. M. Leonard, S. Towler, S. Welcome, *et al.*, "Size matters: Cerebral volume influences sex differences in neuroanatomy", *Cerebral cortex*, vol. 18, no. 12, pp. 2920–2931, 2008.

[67] L. Sørensen, C. Igel, N. Liv Hansen, *et al.*, "Early detection of alzheimer's disease using m ri hippocampal texture", *Human brain mapping*, vol. 37, no. 3, pp. 1148–1161, 2016.

[68] P. Lambin, E. Rios-Velazquez, R. Leijenaar, *et al.*, "Radiomics: Extracting more information from medical images using advanced feature analysis", *European journal of cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[69] Z. Liu, S. Wang, D. Dong, *et al.*, "The applications of radiomics in precision diagnosis and treatment of oncology: Opportunities and challenges", *Theranostics*, vol. 9, no. 5, 2019.

[70] J. E. Van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler, "Radiomics in medical imaging—"how-to" guide and critical reflection", *Insights into imaging*, vol. 11, no. 1, p. 91, 2020.

[71] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images are more than pictures, they are data", *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.

[72] R. T. Larue, J. E. van Timmeren, E. E. de Jong, *et al.*, "Influence of gray level discretization on radiomic feature stability for different ct scanners, tube currents and slice thicknesses: A comprehensive phantom study", *Acta oncologica*, vol. 56, no. 11, pp. 1544–1553, 2017.

[73] R. T. Larue, G. Defraene, D. De Ruysscher, P. Lambin, and W. Van Elmpt, "Quantitative radiomics studies for tissue characterization: A review of technology and methodological procedures", *The British journal of radiology*, vol. 90, no. 1070, 2017.

[74] S. Cui, H.-H. Tseng, J. Pakela, R. K. Ten Haken, and I. El Naqa, "Introduction to machine and deep learning for medical physicists", *Medical physics*, vol. 47, no. 5, 2020.

[75] N. Goel, S. I. Thomopoulos, T. Chattopadhyay, and P. M. Thompson, "Predictive modeling of alzheimer's disease prognosis using anatomical & diffusion mri", in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2023, pp. 1–5.

[76] A. Lebedev, E. Westman, G. Van Westen, *et al.*, "Random forest ensembles for detection and prediction of alzheimer's disease with a good between-cohort robustness", *NeuroImage: Clinical*, vol. 6, pp. 115–125, 2014.

[77] P. A. Donnelly-Kehoe, G. O. Pascariello, J. C. Gómez, A. D. N. Initiative, *et al.*, "Looking for alzheimer's disease morphometric signatures using machine learning techniques", *Journal of neuroscience methods*, vol. 302, pp. 24–34, 2018.

[78] Z. Sun, Y. Qiao, B. P. Lelieveldt, M. Staring, A. D. N. Initiative, *et al.*, "Integrating spatial-anatomical regularization and structure sparsity into svm: Improving interpretation of alzheimer's disease classification", *NeuroImage*, vol. 178, pp. 445–460, 2018.

[79] L. Kohoutová, J. Heo, S. Cha, *et al.*, "Toward a unified framework for interpreting machine-learning models in neuroimaging", *Nature protocols*, vol. 15, no. 4, pp. 1399–1435, 2020.

[80] V. Riberdy, A. Guida, J. Rioux, and K. Brewer, "Radiomics in preclinical imaging research: Methods, challenges and opportunities", *npj Imaging*, vol. 3, no. 1, p. 45, 2025.

[81] S. Spasov, L. Passamonti, A. Duggento, P. Lio, N. Toschi, A. D. N. Initiative, *et al.*, "A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to alzheimer's disease", *Neuroimage*, vol. 189, pp. 276–287, 2019.

[82] J. Zhang, B. Zheng, A. Gao, X. Feng, D. Liang, and X. Long, "A 3d densely connected convolution neural network with connection-wise attention mechanism for alzheimer's disease classification", *Magnetic Resonance Imaging*, vol. 78, pp. 119–126, 2021.

[83] A. Lin, Y. Chen, Y. Chen, *et al.*, "Mri radiomics combined with machine learning for diagnosing mild cognitive impairment: A focus on the cerebellar gray and white matter", *Frontiers in Aging Neuroscience*, vol. 16, 2024.

[84] E. Y. Cheung, A. C. Chau, F. H. Tang, and A. D. N. Initiative, "Radiomics-based artificial intelligence differentiation of neurodegenerative diseases with reference to the volumetry", *Life*, vol. 12, no. 4, p. 514, 2022.

[85] G. Varoquaux, P. R. Raamana, D. A. Engemann, A. Hoyos-Idrobo, Y. Schwartz, and B. Thirion, "Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines", *NeuroImage*, vol. 145, pp. 166–179, 2017.

[86] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: Review of a decade of research", *Artificial intelligence review*, vol. 57, no. 10, p. 273, 2024.

[87] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[88] C. Cabral, P. M. Morgado, D. C. Costa, M. Silveira, A. s Disease Neuroimaging Initiative, *et al.*, "Predicting conversion from mci to ad with fdg-pet brain images at different prodromal stages", *Computers in biology and medicine*, vol. 58, pp. 101–109, 2015.

[89] X. Li, P. S. Morgan, J. Ashburner, J. Smith, and C. Rorden, "The first step for neuroimaging data analysis: Dicom to nifti conversion", *Journal of neuroscience methods*, vol. 264, pp. 47–56, 2016.

[90] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek, "The design of simpleitk", *Frontiers in neuroinformatics*, vol. 7, p. 45, 2013.

[91] J. J. Van Griethuysen, A. Fedorov, C. Parmar, *et al.*, "Computational radiomics system to decode the radiographic phenotype", *Cancer research*, vol. 77, no. 21, 2017.

[92] W. A. Noortman, D. Vriens, J. Bussink, *et al.*, "Multicollinearity and redundancy of the pet radiomic feature set", *European radiology*, vol. 35, no. 11, pp. 6905–6916, 2025.

[93] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net", *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.

[94] W. Zhang, Y. Guo, and Q. Jin, "Radiomics and its feature selection: A review", *Symmetry*, vol. 15, no. 10, p. 1834, 2023.

[95] S. Galli, "Feature-engine: A python package for feature engineering for machine learning", *Journal of Open Source Software*, vol. 6, no. 65, p. 3642, 2021.

[96] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein, "A simulation study of the number of events per variable in logistic regression analysis", *Journal of clinical epidemiology*, vol. 49, no. 12, pp. 1373–1379, 1996.

[97] A. Chalkidou, M. J. O'Doherty, and P. K. Marsden, "False discovery rates in pet and ct studies with texture features: A systematic review", *PloS one*, vol. 10, no. 5, 2015.

[98] I. Fernando, "A binary probability decision tree with youden's j statistic: A simpler machine learning algorithm for medical diagnosis", in *2025 19th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI-Winter)*, IEEE, 2025, pp. 245–250.

[99] C. R. Jack, D. S. Knopman, W. J. Jagust, *et al.*, "Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade", *The Lancet Neurology*, vol. 9, no. 1, pp. 119–128, 2010.

[100] C. Scapicchio, M. Gabelloni, A. Barucci, D. Cioni, L. Saba, and E. Neri, "A deep look into radiomics", *La radiologia medica*, vol. 126, no. 10, pp. 1296–1311, 2021.

Appendix A Model coefficients

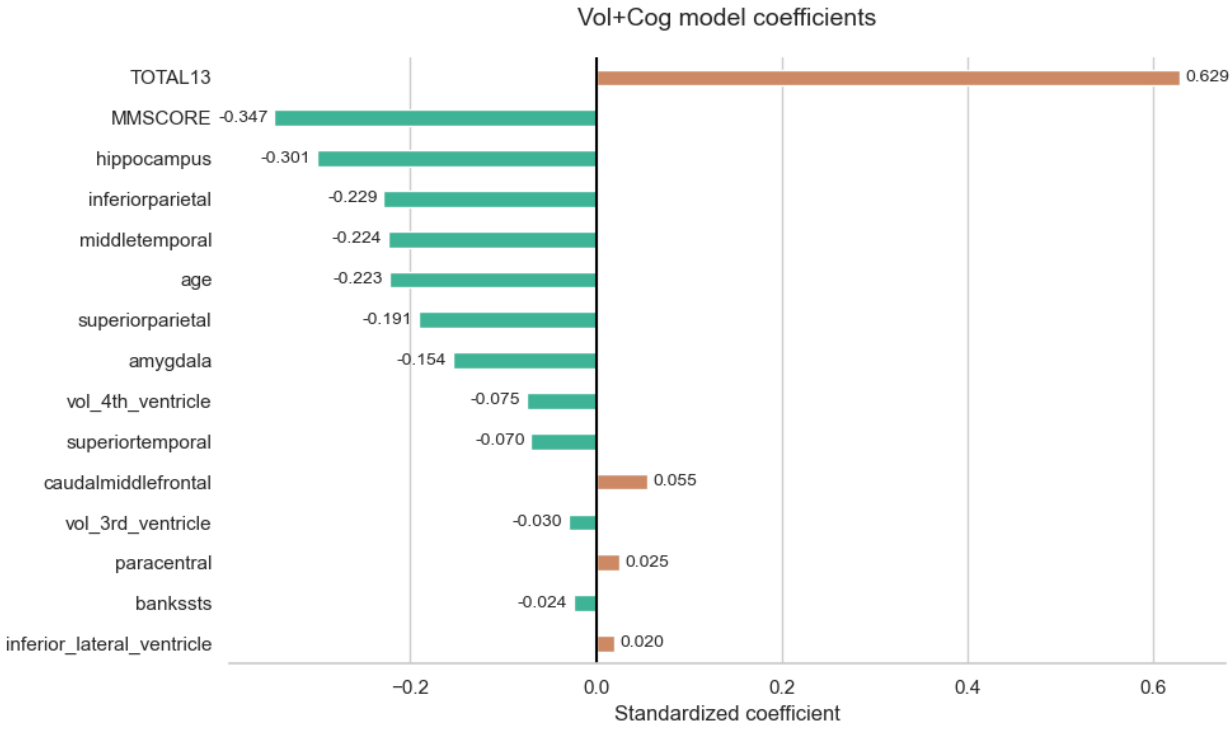

Figure A.1: Standardized coefficients of the Vol+Cog model.

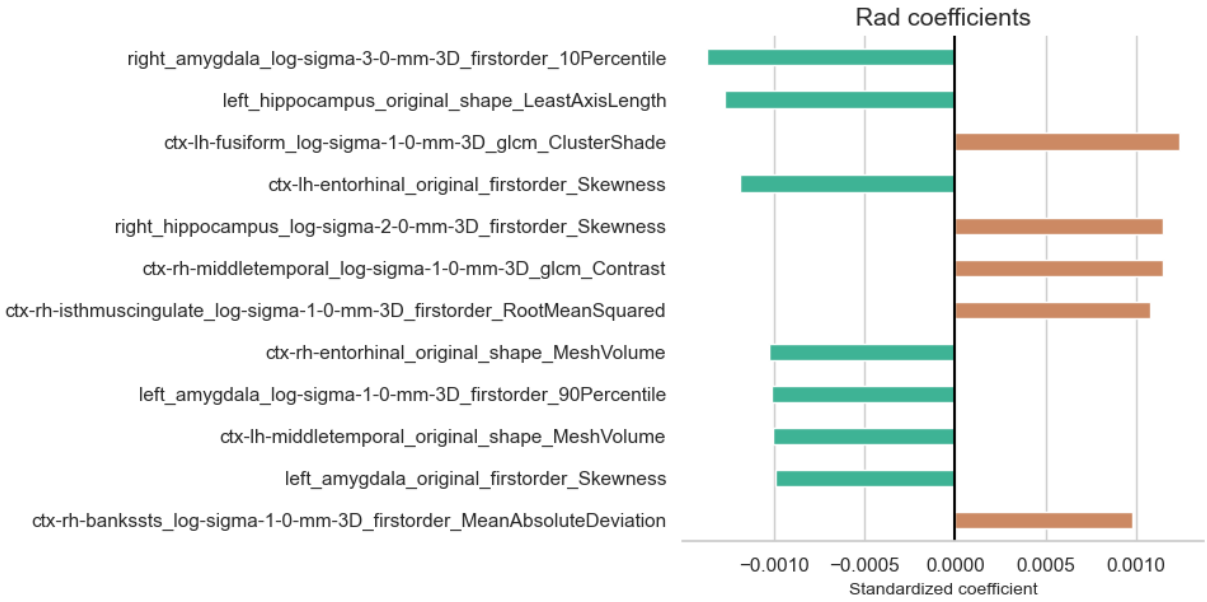

Figure A.2: Standardized coefficients of the Rad model.

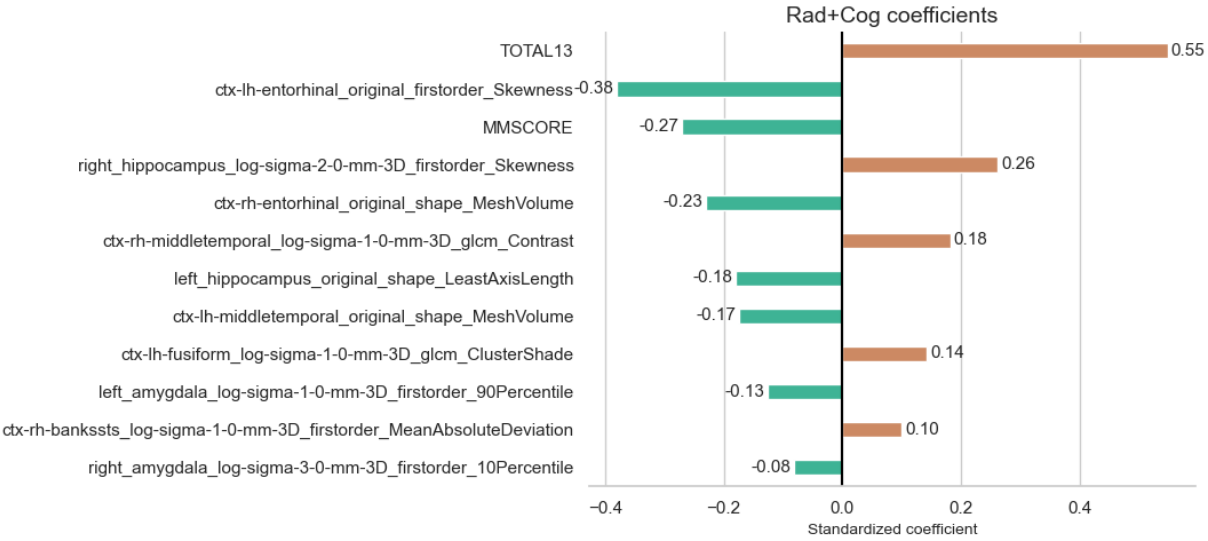

Figure A.3: Standardized coefficients of the Rad+Cog model.

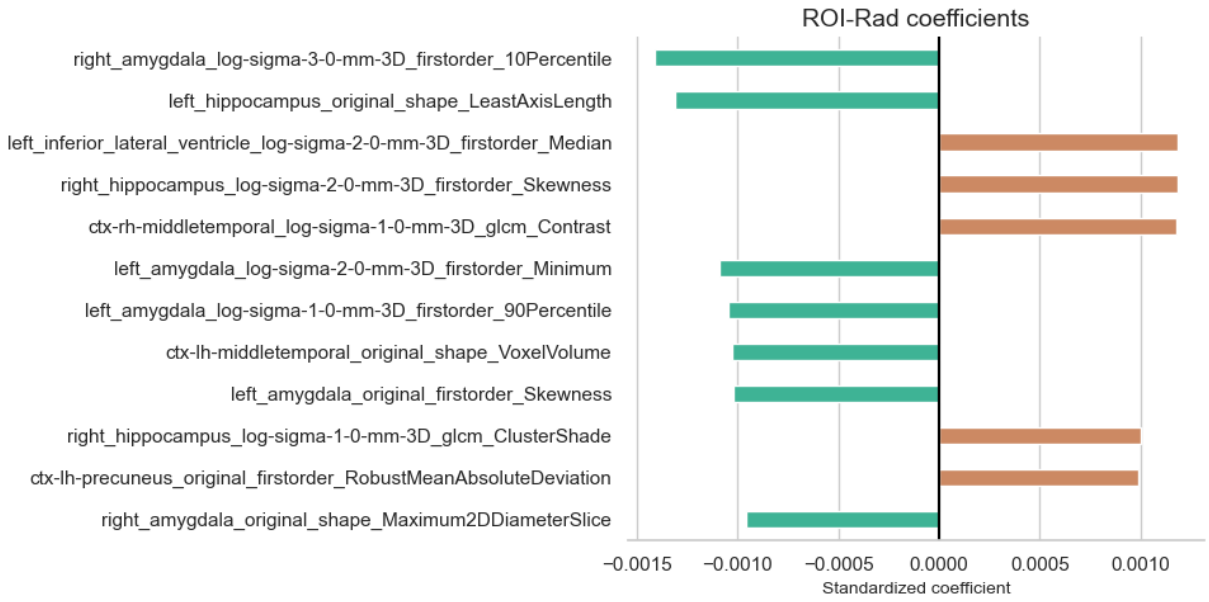

Figure A.4: Standardized coefficients of the ROI-Rad model.

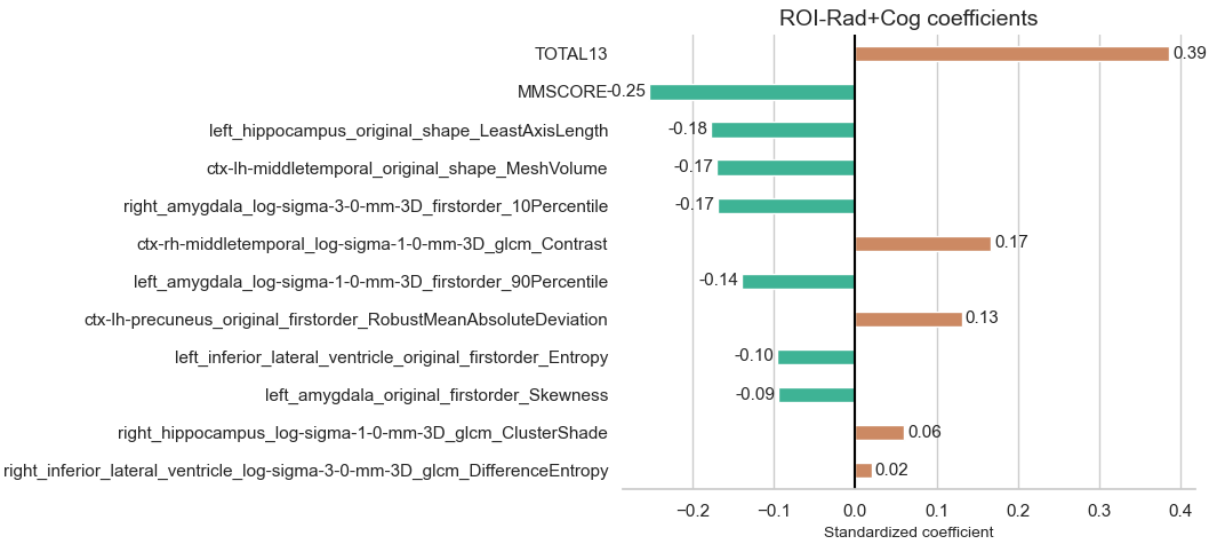

Figure A.5: Standardized coefficients of the ROI-Rad+Cog model.