NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Predictive Modeling of Alzheimer's Disease Using Combined MRI Data

Alícia da Costa Pinho Santos

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Predictive Modeling of Alzheimer's Disease Using Combined MRI Data

by

Alícia da Costa Pinho Santos

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

Supervised by

Mauro Castelli, PhD, NOVA Information Management School

July, 2025

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, July 10th 2025

DEDICATION

To everyone who had the courage (and patience) to read this thesis, thank you. Your bravery will not be forgotten.

ACKNOWLEDGEMENTS

I would like to start by thanking Professor Mauro for accepting this challenge and advise me. His lessons and knowledge truly inspired me to pursue this study and aim high with ambition.

To my parents, siblings and Avó lice thank you so much for all your unwavering support, for your patience during my most stressed-out moments.

To my PRIMA and Meri, thank you for reading my thesis and acting as my personal Grammarly. I am very happy knowing someone actually read it, even if I made you do it!

To my dearest friend Beni Benu, thank you for staying up late with me, drinking coffee, playing Splendor, and procrastinating. You made it all bearable.

To my emigrant friends Kika, Leo, and Teresa, thank you for going abroad just so I have another reason to miss you and constantly plan the next time we'll be together.

To my friend Sérgio that is always annoyingly reminding me how important it is to leave my comfort zone. I appreciate that more than I let on!

And last but definitely not least, to my friend Inês Cal Marques, if it weren't for you, I would've never applied to this master's program or written this thesis. Thank you, Manager!

Thank you all for being there, for being my silent support and for helping me get through the very long journey that was university.

I made it! :)

ABSTRACT

Alzheimer's Disease is a progressive neurodegenerative disorder and the leading cause of dementia worldwide, posing a significant burden on individuals, families, and healthcare systems. Early detection is crucial for enabling timely intervention and personalized care strategies. This thesis investigates the potential of deep learning techniques, particularly Convolutional Neural Networks and Recurrent Neural Networks, to enhance AD diagnosis and progression modeling using MRI data. The study is conducted in two phases: a cross-sectional analysis using the OASIS-1 dataset and a longitudinal analysis using OASIS-2. In the first phase, multiple CNN architectures, including ResNet, DenseNet, and EfficientNet, are evaluated for their ability to classify individuals as demented or non-demented based on 2D slices extracted from structural brain MRI scans. The best-performing CNN model is then employed as a feature extractor in the second phase. Here, extracted features are combined with clinical and demographic data across multiple time points to train RNN models, including LSTM, GRU, BiLSTM, and BiGRU, for disease progression prediction. The integration of spatial and temporal data allows for a comprehensive exploration of how structural brain changes relate to cognitive decline over time. Evaluation metrics such as accuracy, AUC, precision, and recall are used to compare model performance, and findings demonstrate that hybrid CNN-RNN models provide valuable improvements in identifying early-stage Alzheimer's and forecasting progression. This work contributes both methodologically and clinically by proposing an end-to-end deep learning framework that leverages multimodal MRI and clinical data to support more reliable and interpretable tools for early diagnosis and longitudinal monitoring of Alzheimer's Disease.

KEYWORDS

Alzheimer's; Progression; Neuroimaging; Deep learning; Longitudinal-Cross sectional studies;

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AD	Alzheimer’s Disease
ADRC	Alzheimer’s Disease Research Center
BiLSTM	Bidirectional Long Short-Term Memory
BiGRU	Bidirectional Gated Recurrent Units
cMCI	Converter Mild Cognitive Impairment
CNN	Convolutional Neural Network
DBN	Deep Belief Network
DFFNN	Deep Feedforward Neural Network
DL	Deep Learning
EHR	Electronic Health Records
EOAD	Early-Onset Alzheimer's Disease
GAN	Generative Adversarial Network
GCN	Graph Convolutional Network
GRU	Gated Recurrent Units
HC	Health Control
LOAD	Late-Onset Alzheimer's Disease
LSTM	Long Short-Term Memory
MCI	Mild Cognitive Impairment
ML	Machine Learning
MRI	Magnetic Resonance Image
NC	Normal Control
NIA-AA	National Institute of Aging and Alzheimer's Association
OASIS	Open Access Series of Images Studies
PET	Positron Emission Tomography
pMCI	Progressive Mild Cognitive Impairment

RNN	Recurrent Neural Network
ROI	Region of Interest
sMCI	Stable Mild Cognitive Impairment

1. INTRODUCTION

1.1. BACKGROUND AND MOTIVATION

Alzheimer's disease (AD) is a progressive and (currently) incurable neurodegenerative disorder that primarily affects older adults and is the leading cause of dementia worldwide, accounting for 60-70% of all cases. It is characterized by memory loss, cognitive decline and behavioral changes that significantly impact daily functioning. As a result, AD places a considerable burden on individuals, families and healthcare systems. (Mayeux & Stern, 2012)

Despite decades of research and a big progress in understanding the disease, its etiology remains complex and multifactorial, involving a combination of genetic, environmental, and lifestyle factors. This complexity and the fact that with aging populations worldwide, the number of individuals affected is expected to increase exponentially in the upcoming years, early detection becomes crucial.

Currently, no treatments can cure or stop the progression of AD. Available therapies focus on managing symptoms rather than altering the disease course. Therefore, early detection provides an opportunity for more timely clinical interventions and personalized treatment planning, improving patients' lifestyle by delaying the symptoms progression.

Magnetic Resonance Imaging (MRI) has become a standard, a non-invasive tool for assessing structural brain changes associated with AD, such as hippocampal atrophy and cortical thinning. (Zhu et al., 2022) These brain changes often precede clinical symptoms, positioning MRI as a valuable tool for early detection.

The increasing availability of advanced computational methods has opened new possibilities for improving diagnostic support. Machine Learning (ML) and Deep Learning (DL) models, in particular, have shown strong potential in medical imaging. In context of AD, these methods have the ability to uncover patterns in brain structure that may not be visible through traditional clinical assessments. By automating image analysis, they can support faster, more consistent, and potentially more accurate detection of disease-related changes, reducing the need for extensive preprocessing.

Nevertheless, the application of these models in clinical practice remains limited. The potential of ML and DL to contribute to early detection is still underexploited, and further research is needed to refine, validate, and integrate these tools into healthcare systems. Integrating DL models into this process could provide more objective, scalable, and accurate assessments, representing a meaningful step toward earlier and more reliable detection of Alzheimer's disease.

1.2. RESEARCH GAP

There has been a significant increase in the use of medical image, in particular MRI, for identifying structural brain changes related to AD, their ability of identifying characteristic patterns has elevated its importance in the diagnostic process. However, they are typically used only after cognitive symptoms have been identified through clinical assessments, which delays the possibility of early detection, an essential for timely intervention and care planning.

While research using imaging data has expanded considerably in both academic and clinical contexts, its practical application in early detection remains limited. Most existing studies focus on simple classification tasks (e.g., demented vs. non-demented) and are often based on small or homogeneous datasets that do not reflect the diversity of real-world populations.

Additionally, much of the current research relies exclusively on cross-sectional data, missing the potential of longitudinal imaging to capture disease progression over time. The lack of integration between static and temporal data reduces the predictive power and practical applicability of many available models.

Overall, there is a clear need for developing more advanced, flexible, and generalizable models that combine multiple types of information, including clinical scores and repeated scans to improve the early detection and progression prediction of Alzheimer's disease.

1.3. RESEARCH PURPOSE AND OBJECTIVES

The goal of this thesis is to develop predictive models that help reduce the bridge gap in early diagnosis. It evaluates the efficiency of machine learning and deep learning models approaches in predicting Alzheimer's disease by integrating both cross-sectional and longitudinal data from the OASIS-1 and OASIS-2 datasets, respectively. The study aims to improve the accuracy of disease stage classification and progression forecasting, offering more robust tools for early diagnosis and better management strategies.

The methodological approach focuses on advanced deep learning architectures: Convolutional Neural Networks (CNNs) for analyzing cross-sectional MRI data and Recurrent Neural Networks (RNNs) for modeling longitudinal MRI data, with an emphasis on model interpretability and clinical relevance.

Since the main focus is the integration of multiple imaging data types, the model development and evaluation will occur in two main phases:

1. **Initial Classification Phase**: Models ability to classify individuals as demented or non-demented using cross-sectional data, with a particular focus on identifying and interpreting the most influential features driving the predictions.
2. **Progression Prediction Phase**: Evaluate models' ability to classify individuals in transitional states and predict future cognitive decline using longitudinal imaging.

Additionally, relevant clinical and demographic variables, such as age, gender, and education level will be analyzed to explore potential correlations between brain structure changes and individual characteristics.

Overall, the specific objectives of this thesis are:

1. **To explore the predictive potential of cross-sectional MRI data** (OASIS-1) for classifying Alzheimer's disease stages using deep learning models.
2. **To investigate how incorporating longitudinal MRI data (OASIS-2) enhances model performance** assessing its added value in predicting future cognitive decline.
3. **To identify key neuroimaging features and cognitive variables** that contribute most to classification and progression prediction.
4. **To compare different modeling techniques** in terms of accuracy, sensitivity, and interpretability.
5. **To provide insights into the clinical applicability of combining cross-sectional and longitudinal MRI data**, in terms of model performance, generalizability and clinical relevance for early Alzheimer's detection and long-term monitoring.

By addressing these objectives, this thesis aims to contribute both technically and clinically to the field of AI-assisted Alzheimer's research, supporting the development of tools that are not only powerful but also transparent and actionable in real-world healthcare settings.

1.4. RESEARCH CONTRIBUTIONS

By addressing these objectives, this thesis aims to contribute both technically and clinically to the field of AI-assisted Alzheimer's research, supporting the development of tools that are not only powerful but also transparent and actionable in real-world healthcare settings. The combination of cross-sectional and longitudinal data will provide a deeper understanding of the disease progression over time, which is crucial for timely intervention. Additionally, the inclusion of other clinical variables will enhance the model's accuracy and generalizability, ensuring it reflects the complexity of real-world conditions.

This study will focus on overcoming the common challenges associated with deep learning models, ensuring that healthcare practitioners can trust and understand the results, making the tool clinically actionable, providing valuable support for early detection and disease management.

2. LITERATURE REVIEW

This chapter provides a comprehensive review of literature on Alzheimer's disease and the application of deep learning in medical imaging, with particular focus on magnetic resonance imaging and its role in disease classification and progression modeling. The review draws from a broad range of scientific contributions published between 2009 and 2025, combining foundational works with the most recent advances, gathered from databases like PubMed, Google Scholar, arXiv, Elsevier, and IEEE Xplore.

To guide the reader through the theme, this literature review is organized into four main parts. It begins with a general overview of AD, including its clinical characteristics, diagnostic, treatment strategies and current prevalence data. Followed by a section that introduces the foundations of deep learning and its emerging role in healthcare, especially in the domain of medical image analysis.

The core of this review lies in the final two sections. The third one focuses on the use of Convolutional Neural Networks for classifying AD using cross sectional. This means discussing its strengths, typical architecture choices and limitations. Finally, the last section explores temporal modeling approaches, particularly Recurrent Neural Network related, and highlights recent efforts to develop hybrid models, that combine cross-sectional with longitudinal to address existent gaps.

The chapter concludes by identifying the research gap this thesis aims to address and presenting the research question that guides the study:

How can predictive models using cross sectional and longitudinal MRI data improve early detection and classification of Alzheimer's disease?

2.1. INTRODUCTION TO ALZHEIMER'S & EARLY DETECTION

Dementia is a general term for the decline of memory, language, problem solving and other cognitive abilities severe enough to interfere with daily lifestyle. It is not a single disease, but rather a collection of symptoms associated with various neurological disorders and brain injuries, such as Alzheimer. (*Alzheimer's Association 2024 Alzheimer's Disease Facts and Figures*, 2024)

AD is the leading cause of dementia accounting approximately for 60 – 80% of cases worldwide. (Breijyeh & Karaman, 2020; Sorbi & Ferrari, 2021) AD is a progressive neurodegenerative disorder characterized by the progress of cognitive decline, marked by initial manifestations of memory impairment, eventually leading to dependence on daily life activities. (Mayeux & Stern, 2012) Despite extensive research, it remains an incurable condition making early detection and intervention especially critical.

The definition of AD has gradually changed over time, evolving from a single disease entity to a spectrum of related pathologies. Despite this evolution the underlying mechanism of AD remains poorly understood.

Numerous hypotheses have been proposed but none have provided a definite explanation. Historically, AD has been linked to genetic mutations (APP, PSEN1 and PSEN2) strongly associated with early-onset Alzheimer's Disease (EOAD) cases. (Ferrer, 2012) However, beyond genetic conditions, environmental influences are also considered critical, due to the increasing prevalence of Late-Onset Alzheimer's Disease (LOAD) cases.(Mayeux & Stern, 2012)

Several studies have been conducted to investigate disease distribution to better understand their environmental risk, such as gender, age, education and global prevalence. The results demonstrated that socioeconomic factors and life experiences contribute to most of the disparities. Although the most prominent contributors are the different health conditions. (Gustavsson et al., 2023)

Differences in diagnostic criteria and study methodologies constrain the prevalence rates. Therefore, improved and standardized diagnostic methods are crucial in epidemiological research like this one to ensure more accurate and comparable estimates. (Cao et al., 2020; Mayeux & Stern, 2012)

Like the definition of AD itself, diagnostic criteria has evolved over the years. Efforts toward international standardization have led to significant improvements in diagnostic. However, these are still shaped by national health care regulations and strategies. Currently, the most recognized diagnostic frameworks include: The 2011 National Institute on Aging and Alzheimer's Association (NIA-AA) criteria; The 2010 International Working Group revised lexicon; The *Diagnostic and Statistical Manual of Mental and Disorders*. (Atri, 2019) These frameworks integrate cognitive impairments, biomarkers, and specific syndromes, enabling early intervention and more comprehensive diagnostic approach. (Passeri et al., 2022) Still, there remains room for early detection to be better optimized. (Dubois et al., 2021; Eskildsen et al., 2015)

At the moment, the available treatments are primarily palliative, and the effectiveness of emerging therapies remains uncertain. For example, medicine such as *donanemab* and *lecanemab* aim to improve cognitive functions and alleviate symptoms without altering disease progression. (*Alzheimer's Association 2024 Alzheimer's Disease Facts and Figures*, 2024; Breijyeh & Karaman, 2020; Mayeux & Stern, 2012)

The intervention efforts focus on prevention strategies, which fall into three categories:

- Primary prevention – Reducing risk factors in order to delay symptoms manifestation;
- Secondary prevention – Early detection of AD in its initial stages for timely intervention;

- Tertiary prevention – Providing care and pharmacotherapy to stabilize cognition and manage neuropsychiatric symptoms. (Qiu et al., 2009)

Therefore, just as identifying a definitive cause for AD, determining an effective treatment remains equally difficult. This highlights the critical importance of an early diagnostic. For better planning future treatments.

This research focuses on training deep learning models on medical image data to explore potential improvements in the early detection of AD.

2.2. DEEP LEARNING IN HEALTHCARE

Machine learning has become very valuable in medicine applications, particularly in medical imaging tasks. These techniques have significantly advanced many fields in medicine by providing new tools that enhance understanding and offer a possibility for a more rapid and cost-effective diagnostic process. (Chang et al., 2021; Sorbi & Ferrari, 2021)

As Mahesh, B. (2018) mentions, "Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed." However, despite the growing application and potential, traditional machine learning carries risks of misapplication, needing careful implementation to ensure reliability in clinical practice. These limitations are often due to their shallow structures and reliance on many designed features. (Erickson et al., 2017; Wang et al., 2021)

Deep learning has revolutionized Artificial Intelligence by overcoming several challenges that those traditional methods struggled with. It excels at recognizing complex structures in large datasets through backpropagation algorithms, which optimizes internal parameters across multiple layers. (LeCun et al., 2015)

One of the core challenges in healthcare is gaining knowledge and actionable insights from complex, high-dimensional and heterogeneous biomedical data. Modern medical data types, such as electronic health records **(**EHRs), imaging, -omics, sensors, and text are mostly unstructured and poorly annotated, which makes it difficult to interpret without domain expert knowledge.

Deep learning addresses this, by offering end-to-end learning models, excluding the need for manual feature engineering and extracting patterns directly from raw data. Which researchers believe to be a "pass" to enable scalable and unified predictive health systems processing on "millions to billions of patient records and use a single patient representation", improving decision-making. (Miotto et al., 2018)

Several successful implementations highlighted the effectiveness of deep learning. For example, its early use in computer vision was applied on brain MRI scans to predict Alzheimer's disease and its variations. Another example is the approach studied in (Zhang et al., 2019) that won the Parkinson's Progression Marker's Initiative data challenge on subtyping

Parkinson's disease using a temporal deep learning approach (LSTM recurrent neural networks, to identify subtypes of Parkinson's disease based on progression trends). (Miotto et al., 2018)

Despite its promise, deep learning has its own limitations in healthcare. In general Health care data is often very limited, biased, unbalance and with low quality, especially in rare diseases. These uncontrollable factors (for example the quality of the machines used in diagnosis, clinical interpretation, manifested symptoms, financial conditions) contribute to overfitting and poor generalization in deep learning models.

Furthermore, medical datasets are usually heterogenous and noisy and preprocessing this type of data to train robust models remains a major challenge. Additionally, a significant concern is ensuring patients' data privacy when applying AI in computational medicine.

Deep learning is often considered as a "black box" due to its complex architectures and opaque decision-making mechanisms. But in healthcare, interpretability is essential for both doctors and patients to trust and understand decisions. As a result, developing interpretable models is an essential step toward their responsible integration into clinical practice. (Yang et al., 2021)

2.3. DEEP LEARNING IN ALZHEIMER'S DISEASE

Over the years AD has reached several milestones in terms of research and understanding. Advances have been made in identifying potential causes, refining diagnostic criteria and developing treatment strategies. But AD remains an incurable and progressive condition despite medical efforts.

Given its strong correlation with aging, and the fact that global life expectancy is steadily increasing, the number of patients is expected to rise exponentially in the next years. This projects a growth pressure for more advancements.

Imaging studies play a fundamental role in diagnosis, monitoring and assessing neurodegeneration in AD. Among the most widely used imaging techniques for Alzheimer's diagnosis are Positron emission tomography (PET) and MRI, which provide valuable insights into brain structure and function. Currently, neuroimaging data is used to confirm the disease, by excluding other potential conditions. But in recent years the interest in AI-driven approaches in this field has risen substantially, actually, the number of publications on Deep Learning for AD prediction in PET/MRI has grown rapidly. (Zhao et al., 2023)

Due to their ability to dynamically adapt to complex imaging features, extracting patterns for AD classification and analyzing AD progression across different stages while reducing overfitting, deep learning models are becoming central to the development of predictive tools for AD diagnosis. (J. Liu et al., 2018; Maity et al., 2024)

Several architectures have been applied in this type of research, including Deep Feedforward Neural Networks (DFFNN), CNNs, Recurrent Neural Networks (RNNs), Generative Adversarial Networks (GANs) and various graph-based architectures. Each of these approaches brings different strengths to tasks such as classification, segmentation and augmentation. (Khojaste-Sarakhsi et al., 2022)

Table 2.1 below provides an overview of these architectures, their most suitable tasks, key advantages, and example studies.

Table 2.1 - Deep learning for Alzheimer's disease

Architecture	Type of task	Contributions	Example
CNN	Classification, Segmentation, Detection	Learns spatial hierarchies in images, identifying important features like brain structures and regions affected by Alzheimer's. (Ker et al., 2018; J. Liu et al., 2018; Mienye & Swart, 2024; Schmidhuber, 2015)	(Basaia et al., 2019)
3D CNN	Volumetric Classification	Processes 3D volumetric data capturing relationships between slices in an MRI. Reduces the loss of structure information. (J. Liu et al., 2018; Lundervold & Lundervold, 2019)	(Rahman et al., 2025)
U-Net	Segmentation	Designed for biomedical image segmentation such as brain tissues and atrophy regions. (Ker et al., 2018; Lundervold & Lundervold, 2019; Zunair & Ben Hamza, 2021)	(Hazarika et al., 2022)
RNN / LSTM	Longitudinal Prediction, Progression Modeling, Reconstruction	Good for analyzing time-series data. Models temporal dependencies on the disease evolution. (J. Liu et al., 2018; Lundervold & Lundervold, 2019; Mienye & Swart, 2024)	(Nguyen et al., 2020)
GANs	Image Synthesis, Data Augmentation, Enhancement	Can generate synthetic MRI images to train models when there is limited data. Improves image quality, enhancing realism. (Ker et al., 2018; J. Liu et al., 2018; Lundervold & Lundervold, 2019; Mienye & Swart, 2024)	(Y. Pan et al., 2018)
Autoencoders / VAEs	Feature Compression, Anomaly Detection, Reconstruction	Reduces dimensionality and find useful representations of MRI data. They are helpful for feature extraction and anomaly detection. (Ker et al., 2018; Mienye & Swart, 2024)	(Kumar et al., 2023)

DenseNet / ResNet	Classification, Feature Propagation	Deep connections, designed to improve information flow between layers. Good for deeper CNNs with fewer parameters and robust in small medical datasets. (Ker et al., 2018; Schmidhuber, 2015)	(Esmaeilzadeh et al., 2018)
Hybrid Models	Temporal-Spatial Fusion, Multi-Modal Learning	Combine strengths of multiple architectures (e.g., CNN for spatial + RNN for temporal) and therefore improve robustness and interpretability in AD models. (J. Liu et al., 2018; Lundervold & Lundervold, 2019)	(Lu et al., 2018)

Additionally, emerging complex models like Transformers, Deep Belief Networks (DBNs), and Graph Convolutional Networks (GCNs) are gaining attention in neuroimaging. GCNs, though less common, are becoming useful for modeling population-level patient data. Transformers excel at modeling relationships between imaging and textual data, while DBNs show strong performance in functional MRI through unsupervised hierarchical learning. (Ker et al., 2018; Mienye & Swart, 2024)

However, by the time AD is clinically diagnosed, extensive neuronal loss and neuropathologic changes have already occurred across multiple brain regions. (Mantzavinos & Alexiou, 2017) This reality underscores the importance of early detection methods like neuroimaging. MRI stands out as a key tool since it is a non-invasive, widely accessible, and cost-effective technique that offers a reliable way to detect early structural changes in the brain before major damage occurs.

Deep learning can leverage MRI data to support earlier and more accurate diagnoses, potentially opening the door to timely intervention and better patient outcomes. In particular, CNNs has demonstrated to be a strong performer on extracting spatial patterns from MRI and are the most widely used models in this domain.

2.4. CNN IN ALZHEIMER'S DISEASE

CNN-based approaches are placed in the leader board of the many images understanding challenges. They significantly improved many fields, especially the one of medical image, becoming a powerful and widely adopted technique in research.

Designed to capture spatial hierarchies within images, CNNs automatically learn features across layers, from input to output, without the need for manual extraction. Their ability to handle high-dimensional data and adapt to a variety of tasks through fine-tuning has proven their exceptional effectiveness in recent studies.

The growing success of CNNs in medical image analysis has led to the development of numerous architectures tailored to different tasks. Award-winning models such as LeNet-5, AlexNet, Overfeat, ZFNet, VGGNet, GoogLeNet, ResNet, and Xception have been adapted for medical purposes, including disease classification and lesion detection. (Nigri et al., 2020; Sarvamangala & Kulkarni, 2022) While most CNN-based approaches have been designed 2D images, reflecting the typical structure of datasets, researchers have also studied 3D CNNs to leverage the full volumetric information present in MRI scans. These models capture spatial dependencies across all three dimensions. Making them particularly advantageous for detecting subtle structural changes associated with early-stage neurodegeneration. However, their higher computational cost and larger data requirements often limit their use in favor of more lightweight 2D CNNs.

AD research is no exception. CNNs have become one of the most widely applied deep learning approach for its analysis using MRI scans. Image based diagnosis using CNN has been extensively studied and several methods have been proposed. In fact, the predominant modeling approach is classification – distinguish between cognitively normal individuals, those with mild cognitive impairment (MCI), and those with AD. (Nigri et al., 2020)

Several studies have shown high classification performance using CNNs. For instance, Basaia and coauthors (Basaia et al. 2019) showed that demonstrated that a CNN trained on a single brain structural MRI scan could distinguish AD, converter MCI (cMCI) and stable MCI (sMCI), achieving high levels of accuracy in all the classifications, 99% accuracy for AD vs healthy control (HC) classification and 75% accuracy for cMCI vs. sMCI, both using the ADNI dataset combined with external datasets.

However, training models in static images tend to focus on current disease stages often missing the subtleties that indicate early or preclinical changes, especially patients still classified as MCI.

Studies such as the one conducted by Patil and their colleagues (Patil et al., 2022) have explored the use of early images to study structural changes in the brain, such as hippocampal atrophy and cortical thinning and proven that CNNs have great potential in supporting early diagnosis, particularly when trained on large datasets of individuals with MCI, since they can detect more reliably than traditional manual assessment techniques.

CNNs are known for being very sensitive to subtle changes therefore anatomical changes in brain structure, especially in hippocampus known to be among the first regions affected by Alzheimer's pathology and often missed during manual examinations is often highlighted in CNN-based models due to its to its early structural degradation. (Hu et al., 2014)

In Table 2.2, several research studies are summarized, showing the types of CNN architectures applied, the type of data used, the study designs, and the corresponding results.

Table 2.2 - CNN architectures in early diagnosis

Author	Architecture	Input Data	Task	Performance metrics	Notes
(D. Pan et al., 2021)	3D-CNN with Genetic Algorithm	Structural MRI focusing on Regions of Interest (ROIs)	AD vs. Normal Control (NC); MCI converters vs. NC; MCI converters vs. non-converters.	Accuracy: AD vs. NC – 89%; MCI converters vs. NC – 88%; MCI converters vs. non-converters – 71%	Identified key brain regions (e.g., hippocampus, amygdala) contributing to early AD detection.
(Nawaz et al., 2021)	2D Deep CNN	3D MRI slices	AD vs. MCI vs. NC	Accuracy: 99.89%	Addressed class imbalance; achieved high accuracy in multi-class classification.
(AlSaeed & Omar, 2022)	ResNet50	MRI	AD vs. MCI vs. NC	Accuracy range: 85.7% to 99%	Employed pre-trained ResNet50 for automatic feature extraction, enhancing early diagnosis accuracy.
(Ali et al., 2024)	Custom CNN	MRI	AD vs. HC	Accuracy: 92%; Sensitivity: 90%; Specificity: 94%; AUC: 0.96	Showcased the potential of CNNs in early AD detection, outperforming traditional ML algorithms.

Despite their strengths, CNNs also come with limitations. First the complexity of deep, multilayered CNNs can make it challenging to identify the affected areas of the brain, particularly in elderly patients. Additionally, they are often criticized for the lack of interpretability and the arising issues with data imbalance, where the number of cognitive normal subjects may significantly outnumber those with AD. These limitations can introduce bias into the training process, reduce the model's ability to generalize across different populations or clinical settings, and increase the risk of overfitting, highlighting the need for larger, well-annotated datasets. (Nigri et al., 2020)

To address these challenges, researchers have explored Hybrid models that integrate multiple architectures, for example combine ResNet-50 with Inception V3, which offers several advantages over single CNNs, particularly in addressing some of the limitations inherent to traditional CNNs. Beyond architectural blending, other hybrid approaches involve combining different types of deep learning or even traditional machine learning methods. Among these, one of the most promising strategies for early detection research is the combination of CNNs with RNNs.

CNNs limit the complete detection of AD in the initial stage of the disease and the multi-layered model becomes more complex while identifying the affected areas of the brain in old age people. A hybrid approach can combine the feature extraction power of CNNs with the temporal modeling capabilities of RNNs, especially when working with longitudinal data. This combination not only enhances the model’s ability to extract meaningful features but also improves prediction performance through methods like ensemble averaging or voting mechanisms. It enables the model to learn from multiple aspects of the data, making it particularly effective in early-stage diagnosis. (Podolszańska, 2024)

In this context, a hybrid model that combines CNN with RNN architectures emerges as a highly effective approach early detection. While CNNs are responsible for learning spatial features and identifying pathological conditions in the images, RNNs can be trained to model disease evolution over time, making it possible not only to detect the condition but also and predict its progression.

2.5. TEMPORAL DEEP LEARNING MODELS IN ALZHEIMER'S DISEASE

As previously mentioned, current diagnosis criteria for AD typically begins with manual assessments, which include simple cognitive tests to evaluate the patient’s mental functioning, ending with continued clinical monitoring to track disease progression over time. This longitudinal follow-up is just as critical as the initial assessments, as it provides richer clinical information that can guide treatment decisions and support timely intervention.

For this reason, capturing temporal information is essential in AD research. Understanding its neurodegenerative evolution in each individual, especially during the early or preclinical stages, enriches clinical insight and improves decision-making. Working with longitudinal MRI data, rather than static snapshots, allows researchers and models to observe the disease progression and identify patterns that may be predictive of future decline. Therefore, deep learning models that incorporate temporal data are a natural and necessary advancement for more accurate and personalized diagnosis.

It is known that the direct input of a RNN cannot be an image but it is possible to theoretically flatten the image into a one-dimensional vector. Here is where CNN and RNN work together.

In a hybrid CNN – RNN model, the CNN is responsible for extracting spatial features from each image or slice, which are then used as input for the RNN. For instance, in applications involving longitudinal or volumetric data, such as multiple MRI scans or time series scans, each slide is first processed by the CNN to produce a feature vector. Later, these vectors are fed into RNN in sequence, allowing the model to capture temporal dependencies or inter-slice relationships. This approach enables CNNs to perform Intra-Slice Feature Extraction, while RNNs learns inter-slice or temporal dynamics, effectively modeling the progression of neurodegeneration over time. (M. Liu et al., 2018)

Building upon this hybrid approach is essential to explore the specific RNN architectures that have proven most effective in AD modeling. These architectures are designed to handle sequential data and capture long-term dependencies, which are crucial for modeling how the disease progresses over time. Among the most commonly employed are Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRU), and Bidirectional LSTM (BiLSTM). Each of these has unique properties that make them suitable for different types of temporal analysis in the context of AD. (Cui & Liu, 2019)

Table 2.3 presents a comparative summary of these RNN architectures, adapted from Cui & Liu, 2019, highlighting their complexity, memory handling, and computational demands.

Table 2.3 - RNNs architectures in AD

Feature	Complexity	Gating Mechanism	Long-range Memory	Training Time	Computational Power
RNN	Simple	None	Limited	Fast	Low
LSTM	Complex	3 gates	Excellent	Slow	High
GRU	Moderately complex	2 gates	Good	Faster than LSTM	Moderate

BiLSTM	Very complex	2 gates (x2 directions)	Excellent	Slow (dual processing)	High

These architectures have been widely adopted in AD studies, particularly in combination with CNNs, to process longitudinal MRI data and capture patterns associated with disease progression. Table 2.4 provides an overview of some selected studies that implement CNN-RNN models for AD analysis. It summarizes the architectural choices, types of input data used, and key performance outcomes, offering insight into current trends and model effectiveness in this domain.

Table 2.4 - Hybrid studies with CNN-RNN models

Author	Architectures	Input Data	Task	Performance metrics	Notes
(Cui et al., 2019)	CNN – RNN	MRI	AD vs. HC vs. pMCI vs. sMCI	Accuracy – AD vs. HC: 91.33%; Accuracy – pMCI vs sMCI: 71.71%	Highlighted the effectiveness of combining spatial and temporal features for improved diagnostic performance.
(Khatun et al., 2023)	VGG16 - LSTM	MRI	Classification of AD stages	Accuracy: 98.8%; Sensitivity: 100%; Specificity: 76%.	Superior performance compared to contemporary CNN models, emphasizing the benefit of hybrid architectures.
(Li et al., 2019)	DenseNet - BGRU	MRI	AD vs. MCI vs. NC	AUC – AD vs. NC: 91%; AUC – MCI vs. NC: 75.8%; AUC – pMCI vs. sMCI: 74.6%	Focused on detailed hippocampal analysis, a region critically affected in early AD.

(Jomeiri et al., 2024)	DenseNet - BiLSTM	MRI	AD vs. NC vs. MCI vs. sMCI vs. pMCI	Accuracy – AD vs. NC: 95.28%; Accuracy – NC vs. MCI: 88.19%; Accuracy - sMCI vs. pMCI: 83.51%; Accuracy - MCI vs. AC: 92.14%	Emphasized the importance of longitudinal MRI analysis for accurate diagnosis and understanding disease progression.

While CNNs have proven effective in extracting spatial features from cross-sectional MRI data and RNNs excel at modeling temporal patterns in longitudinal sequences, relatively few studies have explored their combined use to validate and reinforce predictions across both data types. Most existing models focus on one specific task, either classifying patients at a single time point or predicting progression over time without explicitly linking the two of them. This creates a gap in developing holistic models that not only detect AD but also monitor its evolution consistently.

The complementary strengths of CNNs and RNNs, spatial precision and temporal continuity, respectively, offer an opportunity for a more robust and clinically meaningful approach. The goal of integrating these architectures together is to enhance both early-stage detection and progression prediction, ensuring that patterns identified in cross-sectional data are validated over time through longitudinal follow-up. This dual modeling strategy directly addresses the challenges of early detection, model generalization and continuity in patient monitoring, positioning it as a step toward a more reliable and personalized diagnosis of AD.

2.6. Hybrid Deep Learning Approaches for Alzheimer's Disease

The application of deep learning in AD research has demonstrated promising results in tasks such as disease classification, staging and progression prediction. CNNs in particular, have been successful in analyzing cross-sectional MRI data by learning spatial features related to neurodegeneration, while RNNs showed a strong capability in capturing temporal dependencies from longitudinal data. However, current literature is very limited in studies with these two approaches together.

Most existing models focus on classifying patients at a single time point using CNNs or predicting disease evolution using RNNs applied to sequential features. This reflects the current structure of AD diagnostic criteria, where MRI is commonly used to confirm the diagnosis and later to monitor disease progression over time.

This lack of studies, attempting to integrate both types of data to leverage the complementary strengths of these architectures in a single predictive framework creates a gap in two critical objectives: timely and personalized intervention in Alzheimer’s care.

A hybrid approach that uses CNNs to classify patients based on cross -sectional MRI scans and RNNs to validate or refine those predictions through longitudinal follow-up could enhance the robustness and clinical applicability of deep learning models. By combining spatial precision with temporal continuity, models offer an opportunity to better reflect how Alzheimer’s disease manifests and progresses in real patients.

This thesis proposes a predictive modeling strategy that integrates cross-sectional and longitudinal MRI data through a hybrid CNN-RNN architecture. The goal is to explore if disease

stages identified from single time point can be predictive of future progression, and whether those predictions can be validated and refined using longitudinal data. This structure simulates an early diagnostic framework to evaluate the predictive capabilities of deep learning models, potentially offering a data-driven complement to existing diagnostic pathways. The following chapter details the methodology used to develop and evaluate this approach.

3. METHODOLOGY

This chapter presents the methodological approach used to develop predictive models for Alzheimer's disease progression based on MRI data. The study leverages two complementary datasets from the Open Access Series of Imaging Studies (OASIS) project: OASIS-1 for cross-sectional analysis and OASIS-2 for longitudinal analysis.

OASIS is a project developed by a collaboration of researchers and institutions, primarily led by the Washington University Alzheimer's Disease Research center (ADRC), in St. Louis, Missouri. It is a collaborative effort aimed at making neuroimaging data sets of the brain freely available to the scientific community. Both OASIS-1 and OASIS-2 provide not only MRI images but also clinical information in accompanying CSV files, including variables such as sex, age, education level, and brain metrics.

For each dataset, the methodological workflow follows a consistent structure: first, the structure, characteristics, and preprocessing steps of the available data are described. Then, the model architectures, training strategies and evaluation procedures applied to integrate insights from both datasets and predict Alzheimer's disease progression over time are detailed.

The approach involves training predictive models on cross-sectional data and validating their performance using longitudinal data to assess their ability to capture and forecast the trajectory of cognitive decline.

3.1. CROSS-SECTIONAL DATA PREPARATION AND ANALYSIS

3.1.1. MRI Data Acquisition and Organization

OASIS - 1 dataset consists in a collection of 416 subjects, aged between 18 and 96 years, including both men and women. For each of them, 3 to 4 individual T1-weighted MRI scans were obtained during a single imaging session. Among these participants, a hundred of them over the age of 60 were clinically diagnosed with very mild to moderate Alzheimer's disease.

In addition, a reliability dataset is included, containing 20 nondemented subjects, each of whom was imaged again on a visit within 90 days of their initial session.

Each individual's data is organized following the structure below:

1. **3 to 4 images**, collected using the same structural protocol collected in a single session to increase signal-to-noise.
2. **An average image**, obtained by motion-corrected coregistered average of all available data.

3. **A gain-field corrected, atlas-registered image**, mapped to the Talairach and Tournoux 1988 atlas space. (Buckner et al., 2004)

4. **A masked version** of the atlas-registered image in which all non-brain voxels have been assigned an intensity value of 0.

5. **A segmented image**, separating grey/white matter and cerebrospinal fluid (CSF). (Zhang et al., 2001)

All images are in 16-bit big-endian Analyze 7.5 format and are distributed in a zip-compressed archive. The archive is organized by imaging session, with each session labeled by the subject ID using the format OAS1_xxxx_MRy, where 'xxxx' represents a number from 0001 to 9999 and 'y' is an incrementing number that reflects the imaging visit number for the subject. Therefore, OAS1_0037_MR1 refers to the individual 37 first image session.

Each session directory includes 4 subdirectories:

- **TXT file**, a text version of the patient information presented on the XML file that includes additional data.

- **RAW directory**, stores all the original, unprocessed MRI scan images.

- **PROCESSED directory**, includes two subdirectories of processed images:

    - **SUBJ_111**, contains the motion-corrected, averaged image, resampled to 1mm isotropic voxels, while preserving the native acquisition space (MRI images after they were cleaned up and combined into one clear image, keeping the original space but making the resolution better).

    - **T88_111**, contains the atlas-registered, gain-field corrected images and their brain-masked versions, also resampled to 1mm isotropic voxels (images that are adjusted to match a standard brain template ("atlas") and corrected for lighting differences). This also includes a subdirectory called t4_files that includes the matrices describing the transformation into atlas space.

- **FSL_SEG directory**, includes the grey/white/CSF segmentation image generated from the masked atlas image.

Table 3.1 explains the naming logic for each image presented in the dataset.

Table 3.1 - Images included in the dataset

Name	Description	Dimension	Vox size	Directory
OAS1_xxxx_MRy_mpr-z_anon	Individual scan	256x256x128	1 x 1 x 1.25	RAW
OAS1_xxxx_MRy_mpr_ni_anon_sbj_111	Image averaged across scans	256x256x160	1 x 1 x 1	SUBJ_111
OAS1_xxxx_MRy_mpr_ni_anon_111_t88_gfc	Grain-field corrected atlas registered average	176x208x176	1 x 1 x 1	T88_111
OAS1_xxxx_MRy_mpr_ni_anon_111_t88_masked_gfc	Brain-masked version of atlas registered image	176x208x176	1 x 1 x 1	T88_111
OAS1_xxxx_MRy_mpr_ni_anon_111_t88_masked_gfc_fseg	Brain tissue segmentation	176x208x176	1 x 1 x 1	FSL_SEG

'z' refers to the scan repetition, most sessions include 3 to 4 repetitions and 'i' represents the number of images included in the averaged image.

All this information was gathered from the documentation provided by OASIS (*Open Access Series of Imaging Studies (OASIS)*, n.d.) which summarizes the data description available for researchers.

Besides the nomenclature presented on Table 3.1 there are two additional parameters at the end of each image name: _sag_xxx, _cor_xxx and _tra_xxx this refers to the orientation of each specific image: sag for Sagittal, cor for Coronal and tra for Transverse and the selected slice.

The following images are a representation of what to expect:

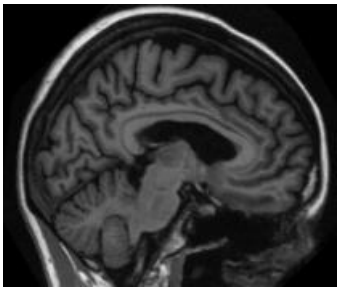

Figure 3.1 - Sagittal Orientation

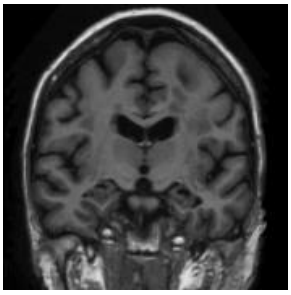

Figure 3.3 - Transverse Orientation

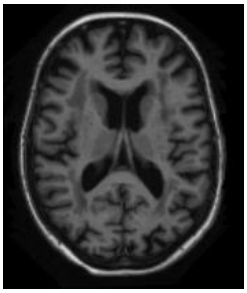

Figure 3.2 - Coronal Orientation

3.1.2. Clinical and Anatomical Data

OASIS-1 provides additional structure datasets in CSV format. Specifically, the dataset named oasis_cross-sectional has descriptive information of each participant, including demographic, clinical and derived anatomic measures:

- **Demographic variables** include gender, handedness, age, years of education and socioeconomic status (SES).
- **Clinical assessments** provided are the Mini-Mental State Examination (MMSE) and Clinical Dementia Rating (CDR).
- **Anatomical measures** derived from the MRI data include the estimated total intracranial volume (eTIV), the atlas scaling factor (ASF) and normalized whole brain volume (nWBV).

This dataset represents a consolidated version of the information provided in the individual text files associated with each participant.

Table 3.2 presents a description of each variable available in the dataset, along with the expected value ranges and formats for each.

Table 3.2 - Variables description

Variable	Description	Values	Notes

ID	Participant ID	OAS1_xxxx_MRy	This dataset combines all folders including the 20 extra sessions
M/F	Gender	M - Male; F - Female	Number
Hand	Handedness	R - Right L - Left	This dataset only includes right-handed participants
Age	Age	18 - 96	.
Educ	Educational Level	1 - less than high school; 2 - high school; 3 - some college; 4 - college; 5 - beyond college.	
SES	Socioeconomic status	1 a 5	
MMSE	Mini-Mental State Examination – Questionnaire to measure cognitive impairment	0 – 30	Higher scores, better cognitive function
CDR	Dementia rating. (All participants with dementia (CDR >0) were diagnosed with probable AD)	0 - nondemented; 0.5 - very mild dementia; 1 - mild dementia; 2 - moderate dementia	(Morris, 1993)
eTIV	Estimated Total Intercranial Volume - Volume inside the skull	~1100 – 2000 Cm^3 (Depends on population)	(Buckner et al., 2004)
nWBV	Normalized Whole Brain Volume - Used to assess brain atrophy	0 – 1	(Fotenos eSt al., 2005)
ASF	Atlas Scaling Factor – Indicates how much the brain needs to be scaled	0.8 – 1.5 (Depends on the brain size)	(Buckner et al., 2004)

Delay	The number of days since the first visit

The other available dataset, oasis_cross-sectional-reliability, includes only information from the 20 participants who completed a second visit in the cross-sectional study. It presents only the anatomical measures recorded and the variable Delay, which represents the number of days since the first visit.

Additionally, a summary table was provided to give an overview of demographic distribution and dementia status.

		Non-Demented				Demented				
Age Group	N	n	mean	male	female	n	mean	male	female	CDR 0.5/1/2
<20	19	19	18.53	10	9	0		0	0	0/0/0
20s	119	119	22.82	51	68	0		0	0	0/0/0
30s	16	16	33.38	11	5	0		0	0	0/0/0
40s	31	31	45.58	10	21	0		0	0	0/0/0
50s	33	33	54.36	11	22	0		0	0	0/0/0
60s	40	25	64.88	7	18	15	66.13	6	9	12/3/0
70	83	35	73.37	10	25	48	74.42	20	28	32/15/1
80s	62	30	84.07	8	22	32	82.88	13	19	22/9/1
≥90	13	8	91.00	1	7	5	92.00	2	3	4/1/0
Total	416	316		119	197	100		41	59	70/28/2

Table 2. Summary of subject demographics and dementia status.

Figure 3.4 - Summary of tabular data

This table provides a brief preview of what to expect when analyzing the tabular data.

3.1.3. Predictive Modeling Strategy and Methodological Choices

The OASIS-1 dataset, which includes both MRI images and structure clinical data, was used to develop the cross-sectional predictive models in this study. The primary objective of this stage is to determine whether it is possible to classify patients as having Alzheimer's disease or not, based on a combination of anatomical imaging and clinical features.

To address this challenge, a modular and scalable modeling pipeline was developed. The pipeline automates dataset construction, preprocessing, model training, evaluation and result comparison. A dataset was created for each image type (RAW, PROCESSED and FSL_SEG) as these directories contain MRI scans that have undergone different medical preprocessing steps. This separation allowed to evaluate how different image versions influence model performance.

3.1.3.1. Exploratory Data Analysis (EDA)

Before model training, an Exploratory Data Analysis (EDA) was conducted to assess the quality, distribution and structure of the clinical and anatomical data from the OASIS-1 dataset, as well as the characteristics of the MRI images. The main goal was to identify potential data quality issues, class imbalances and trends that could influence model design and evaluation. This includes summarizing basic statistics, checking missing values and visualizing feature distributions such as age, MMSE, gender and CDR and some image interpretation like image size and orientation.

Insights from this analysis guided critical preprocessing decisions, including normalization strategies, encoding methods for categorical variables, and feature selection. One key decision emerging from the EDA was the redefinition of the classification target: instead of predicting CDR stages, which are multi-class and often imbalanced, patients were classified into two groups: Alzheimer's vs. non-Alzheimer's. This simplification aimed to enhance classification performance while preserving clinical relevance.

3.1.3.2. Preprocessing

Given that the available data included both MRI images and structured clinical variables, two distinct preprocessing pipelines were implemented: one for image data and one for structured tabular data.

The structured dataset included a variety of demographic, clinical, and anatomical variables, many of which were medically descriptive. Due to the clinical nature of some variables, such as MMSE and CDR scores, missing values were treated cautiously. In cases where medical test values were missing, these entries were excluded rather than imputed, to avoid introducing bias or assumptions into clinically meaningful features. As a result, some reduction in sample size was accepted in favor of data integrity.

The remaining preprocessing steps were deliberately kept simple and transparent. Categorical variables (e.g., gender, handedness, SES) were one-hot encoded or label-encoded where appropriate; Continuous variables (e.g., age, brain volume metrics) were normalized using Standard Scaler; Missing values in non-clinical fields were imputed using simple statistical methods (e.g., KNN imputation); The target variable was binarized for the Alzheimer's vs. non-Alzheimer's classification task, instead of using the original multi-class CDR scale. Additionally, columns that were constant or had minimum variation were dropped.

MRI scans in the OASIS-1 dataset were originally acquired as 3D volumetric brain images in NIfTI format (.nii.gz). However, for public release, the dataset maintainers pre-processed these volumes by extracting representative 2D slices and provided them in (.gif) format. These slices serve as standardized cross-sectional views of the brain and form the basis of the image data used in this study.

To facilitate integration with modern deep learning workflows, all (.gif) images were converted to (.png) format. This conversion was performed for several reasons: (.png) is a lossless format with broader compatibility in Python-based image processing libraries, supports better compression without quality loss, and allows for easier downstream manipulation (e.g., resizing, normalization). The (.png) format also avoids the color palette limitations of (.gif), which could interfere with grayscale medical image fidelity.

Although using 2D images represents a simplification of the original 3D data, it offers important advantages. Working with 2D slices significantly reduces computational load, memory requirements, and training time, making it feasible to run experiments on standard hardware. Furthermore, 2D images are easier to visualize, debug, and interpret, which facilitates model development and evaluation. Importantly, they are also directly compatible with widely available pretrained 2D CNN architectures (e.g., EfficientNet, DenseNet), enabling transfer learning and boosting performance in small datasets.

However, this approach comes with trade-offs. By relying on a single 2D view per subject, spatial context and inter-slice anatomical relationships are lost factors that may be important for detecting subtle brain changes associated with Alzheimer's disease. While 2D CNNs are effective and accessible, they may not capture the full complexity of brain structure that 3D models could potentially leverage.

Despite these limitations, the use of pre-extracted 2D images provided a practical and robust foundation for this study's cross-sectional modeling phase, striking a balance between accessibility, clinical relevance, and methodological consistency.

Beyond resizing and normalization, minimal preprocessing was applied to preserve critical diagnostic features. Excessive preprocessing was avoided to reduce the risk of discarding subtle but important anatomical variations. Image normalization strategies varied by model: pretrained CNNs applied standard ImageNet mean and standard deviation normalization, while the manual CNN used min–max normalization specific to the dataset.

Although some spatial information is inevitably lost in the 3D-to-2D conversion, this approach enabled the use of accessible and efficient 2D CNN pipelines without compromising interpretability or model performance for this stage of the study.

In both structured and image preprocessing, the overarching goals were to maintain clinical relevance, avoid data leakage, and ensure full compatibility with the downstream modelling strategies.

3.1.3.3. Data Splitting Strategies

To ensure robust model evaluation and prevent data leakage, the choice of data splitting strategy was made with careful consideration. Going through the various possibilities, some of the approaches, such as manual splitting or standard K-Folds, were discarded early in the

process. These methods do not account for the distribution of the target variable (in the case of standard K-Folds) or they risk introducing bias due to arbitrary partitioning (as in manual splits). Given the presence of repeated scans and an unequal class distribution, such methods could lead to data leakage or poor generalization.

Finally, two cross-validation strategies were considered for this study: **Stratified K-Folds** and **Group K-Folds**. These methods were chosen after evaluating the characteristics of the OASIS-1 dataset, particularly the potential for class imbalance and the presence of repeated measures (multiple images from the same subject).

- **Stratified K-Fold Cross-Validation** ensures that each fold maintains the same proportion of classes as the overall dataset. This is particularly beneficial in scenarios with class imbalance, helping the model learn equally from both classes during training and ensuring fair performance evaluation. In this study, the class split of approximately 135 Alzheimer's cases to 100 non-Alzheimer's cases, while not extreme, was considered significant enough to warrant stratification.

- **Group K-Fold Cross-Validation** was used to address the issue of repeated measures. In OASIS-1, some subjects have multiple associated images (due to repeated scans or preprocessing variations). Group K-Folds ensures that all data points belonging to the same subject are kept within a single fold—either for training or validation—but never both. This prevents the model from learning subject-specific features that could result in overfitting.

These two strategies are complementary: stratification mitigates class imbalance, while grouping prevents information leakage from repeated samples. However, combining them, Stratified Group K-Folds, poses implementation challenges, as most standard libraries do not support simultaneous stratification and grouping and the complexity of manually combining it could lead to other implantation errors and conclusions. Therefore, while this combination was considered, it was not implemented in the initial pipeline.

Ultimately, models were evaluated using both Stratified K-Folds and Group K-Folds separately to understand their individual impact on model performance and robustness.

3.1.3.4. Modelling Approaches

Convolutional Neural Networks were chosen as the core modelling architecture due to their proven effectiveness in processing visual data, especially in medical imaging tasks. CNNs are designed to capture spatial hierarchies and local features (e.g., edges, textures, and shapes), which are essential for detecting subtle structural brain changes associated with Alzheimer's disease. This architecture eliminates the need for manual feature engineering and enables the model to learn complex patterns directly from the raw image data.

Importantly, the use of CNNs supports the longitudinal goal of this thesis. The CNN developed in this cross-sectional phase is intended to serve as a reusable feature extractor for the longitudinal analysis using the OASIS-2 dataset. The learned image representations will be used as input features for sequence models, which are better suited to modelling temporal progression. This forward-compatible design ensures methodological consistency and strengthens the integration between the cross-sectional and longitudinal components.

To select the most effective modelling approach for classifying MRI images, both pretrained CNNs and a manually designed CNN were considered.

Pretrained models, such as EfficientNet, DenseNet and RegNet were chosen due to their proved performance on large-scale image classification tasks. These models leverage transfer learning using weights previously trained on ImageNet datasets, allowing the model to improve generalization, significantly reduce training time and mitigate overfitting, which is particularly valuable given the limited size of OASIS-1 dataset.

- **ResNet-18** is known for its use of *residual connections*, which help mitigate the vanishing gradient problem in deep networks. ResNet-18 is relatively lightweight and was selected to test whether a simpler architecture could achieve good performance without high computational cost.
- **DenseNet-121** introduces dense connections between layers, promoting feature reuse and improving gradient flow. It is particularly effective when working with small datasets and often performs well with fewer parameters than comparable models.
- **EfficientNet-B0** balances depth, width, and resolution to achieve optimal performance with fewer resources. B0 is the smallest version in the family and was chosen for its efficiency and strong performance in medical imaging benchmarks.
- **VGG-16** remains widely used due to its simple and interpretable architecture. Its inclusion allows for comparisons against a well-established baseline in deep learning literature and helps evaluate whether more modern models truly outperform classic architectures.

These models were selected *instead of alternatives* (e.g., Inception, MobileNet, or deeper ResNets) to provide a balance between performance, interpretability, and computational efficiency.

The manual CNN was designed and trained from scratch, serving both as a baseline and as a flexible architecture tailored to the characteristics of the data. Like the pretrained models, it processes MRI images and structured data jointly. It includes parallel branches: one branch for processing images through convolutional layers, and another for structured data through dense layers. These branches are then merged before the final classification layers. This

design allows full control over the architecture and facilitates experimentation with model complexity and regularization.

Across all model types, consistent training procedures were applied. Binary cross-entropy was used as the loss function, Adam optimizer was selected for training and early stopping was implemented based on validation loss to avoid overfitting. Finally, for pretrained models, ImageNet normalization statistics were used, while the manual CNN employed min–max normalization based on the dataset.

This dual-model approach enables a fair and direct comparison between pretrained and manually built CNNs in a multimodal classification task. It also provides insights into the trade-offs between using off-the-shelf architectures and custom-designed networks in the context of medical diagnosis and eventual longitudinal modelling.

3.1.3.5. Evaluation Strategy

Model performance was assessed using a consistent evaluation framework across all configurations, including both pretrained and manually designed models. The primary classification metrics were:

- **Accuracy**: the overall proportion of correctly classified instances;
- **Precision**: the proportion of positive predictions that were correct;
- **Recall (Sensitivity)**: the proportion of actual positive cases correctly identified;
- **F1-Score**: the harmonic mean of precision and recall, providing a balanced measure for imbalanced datasets.

These metrics were computed for each fold during cross-validation and averaged to obtain final performance estimates. To further support interpretability, confusion matrices were generated to visualize model behaviour on Alzheimer's vs. non-Alzheimer's classifications.

All models were trained using early stopping based on validation loss to prevent overfitting. Only the best-performing model weights were retained for evaluation within each fold. The evaluation process was identical across all model architectures, data splits, and image types to enable a fair and systematic comparison.

To identify the best-performing model, the average values of the main classification metrics were used as the basis for comparison. Additionally, training curves, including training loss, validation loss, and accuracy were recorded at each epoch, allowing for detailed monitoring of model convergence and overfitting.

In summary, this study employed a structured and scalable modelling pipeline to evaluate the effectiveness of both pretrained and manually designed Convolutional Neural Networks in classifying Alzheimer's disease based on multimodal data. Multiple CNN architectures,

designed to process both image and structured clinical inputs, were tested across two distinct cross-validation strategies to assess the influence of data partitioning on model performance and robustness.

3.2. OASIS-2 LONGITUDINAL MRI DATA IN NONDEMENTED AND DEMENTED OLDER ADULTS

3.2.1. Dataset construction

The second part of this study utilizes OASIS-2 dataset, that consists in a longitudinal collection of 150 subjects aged 60 to 96. Each subject was scanned on two or more visits, separated by at least one year for a total of 373 imaging sessions. For each subject, 3 or 4 individual T1-weighted MRI scans were obtained in single scan sessions.

Out of the 150 participants, 72 remained nondemented throughout the study, 64 were characterized as demented at the time of their initial visits and remained so for subsequent scans, including 51 individuals with mild to moderate Alzheimer's disease. Other 14 subjects were characterized as nondemented at the time of their initial visit and were subsequently characterized as demented at a later visit.

As with OASIS-1, the dataset consists of:

- Compressed folders of MRI scans per session.
- A CSV file with tabular clinical and demographic data.

In addition to previously used variables (e.g., age, gender, MMSE, CDR), OASIS-2 includes two new longitudinal-specific columns:

Table 3.3 - Additional variables description

Variables	Description	Values	Notes
Visit	Number of the scan session	1 – 5	
Group	Classification of the patient transition	Demented; Nondemented and Converted	If a patient is demented it is because it stayed demented from the first to the last visit.

However, unlike OASIS-1, which provided 2D preprocessed images, the OASIS-2 dataset contains raw 3D volumetric MRI data in (.hdr) and (.img) format. Since the models in this study

were trained on 2D slices, a preprocessing step was required to convert the 3D data into suitable 2D inputs.

This conversion was performed using NiBabel, a specialized Python library designed for handling medical imaging data formats commonly used in neuroimaging, such as NIfTI (.nii.gz) and Analyze (.hdr/.img).

Each MRI scan in the dataset is a volumetric image composed of three spatial dimensions, typically representing the brain as a 3D grid of voxels (volumetric pixels). These volumes capture detailed anatomical information across different planes: sagittal (left to right), coronal (front to back), and axial (top to bottom).

Using NiBabel, the 3D structure of each MRI was loaded and accessed as a matrix of intensity values. From each volume, one representative 2D slice was extracted from the midpoint of each of the three anatomical planes. These central slices were chosen to ensure consistency across patients and to focus on regions of the brain known to show structural changes associated with Alzheimer's Disease.

Once extracted, the 2D slices were processed into a standardized image format. The intensity values were normalized to ensure consistent brightness and contrast, and all slices were resized to a fixed dimension suitable for deep learning input. Finally, each image was saved as a grayscale PNG file, labeled with its corresponding subject ID and orientation.

This slice-based approach allowed the study to retain important diagnostic features from the original 3D scans while leveraging the simplicity and computational efficiency of 2D CNN models. The selection of sagittal, coronal, and axial slices ensured that each model had access to complementary views of brain anatomy, including structures such as the hippocampus, lateral ventricles, and frontal lobes which are known to be affected in dementia.

3.2.2. Feature Extraction

RNN models cannot directly accept raw image data as input because they require fixed-length vectors or sequences rather than high-dimensional image matrices. Therefore, to enable sequential modeling, the raw image information must first be encoded into a numerical format compatible with these models. This is where the first part of the study becomes crucial.

To address this, the CNN developed in the initial phase, designed to learn spatial brain patterns relevant for classification tasks, is employed as a feature extractor. It processes each MRI slice and transforms it into a fixed-length feature vector that summarizes the most salient information in a compact form.

Feature extraction can be understood as the process by which the CNN compresses complex image data into a lower-dimensional representation, preserving meaningful patterns learned

during training. These feature vectors act as numerical summaries of the original images, encoding spatial information in a way suitable for sequential modeling.

For the longitudinal dataset, which includes multiple imaging sessions per subject, this approach is especially valuable. Since RNNs expect sequences of vectors rather than raw images, the pretrained CNN from the cross-sectional phase converts each MRI slice into one feature vector. These vectors, combined with structured clinical data if available, form a more suitable input for the RNN models to capture temporal progression patterns across multiple visits.

In summary, CNN-based feature extraction bridges the gap between raw MRI images and sequential modeling frameworks, enabling the effective use of longitudinal imaging data for disease progression analysis.

3.2.3. EDA and Preprocessing

Since the data used in the longitudinal part comes from the same repository as OASIS-1, there were no significant differences between the available variables. The preprocessing steps implemented were the same as in the first part to ensure the models were trained consistently. After preprocessing, a basic exploratory data analysis (EDA) was performed to better understand the variables that were excluded in the first part as well as the new variables introduced in the longitudinal dataset.

3.2.4. Model Approach

To address the temporal nature of the longitudinal dataset, models specifically designed to handle sequential data were required. Recurrent Neural Networks were chosen as the core modeling architecture due to their effectiveness in capturing temporal dependencies and learning from patterns that evolve over time.

Unlike feedforward networks, RNNs maintain a form of memory across time steps, allowing them to model how a subject's condition changes over multiple visits. This makes them particularly suitable for longitudinal medical data, where understanding progression is essential.

To ensure a robust modeling process, the study focused on well-established RNN variants that are commonly used in sequence prediction tasks. Specifically, four architectures were considered:

- **LSTM** networks are capable of learning long-term dependencies by mitigating the vanishing gradient problem common in vanilla RNNs. Their gated structure allows them to selectively retain or discard information over time, making them ideal for medical progression modeling.

- **GRU** are a streamlined alternative to LSTMs, using fewer gates and parameters while achieving similar performance. They are computationally efficient and well-suited for smaller datasets or faster training cycles.

- **BiLSTM** model extends the LSTM by processing sequences in both forward and backward directions. In clinical settings, where future and past context can both inform a current state, bidirectional models can improve performance by incorporating more context.

- **BiGRU** is similar to BiLSTM, this model enhances GRU by adding bidirectional processing. It provides the benefits of GRU's efficiency while capturing information from both temporal directions.

These models were selected instead of alternatives like Transformers or 1D CNNs because the dataset is relatively small and structured around clear time steps. While Transformers are state-of-the-art in many domains, they typically require large datasets and more computational resources. In contrast, LSTM and GRU variants are well-suited for limited, structured, and temporally aligned medical datasets such as OASIS-2.

Furthermore, bidirectional versions of these models were included to evaluate whether having access to both past and future visits (when applicable in training) would improve performance in capturing complex temporal patterns associated with Alzheimer's progression.

3.2.5. Evaluation strategy

The evaluation strategy in the longitudinal phase closely mirrors that of the cross-sectional phase, ensuring consistency in how model performance is assessed across both stages of the study. However, due to differences in data structure and modeling goals, some adjustments were necessary.

Unlike the cross-sectional setup, where K-fold cross-validation was used, the longitudinal modeling employed a single train-test split at the subject level. This approach allowed the computation of Receiver Operating Characteristic (ROC) curves and the corresponding Area Under the Curve (AUC), which could not be reliably calculated during cross-validation.

- **ROC Curve**: Plots the true positive rate against the false positive rate across different classification thresholds.

- **AUC**: Provides a scalar measure of a model's ability to distinguish between progression and non-progression; higher values indicate better discriminative power.

The inclusion of ROC and AUC adds a probabilistic layer to the evaluation process, complementing traditional threshold-based metrics such as accuracy, precision, recall, and F1 score.

Due to the direct and non-iterative nature of the data split, confusion matrices were not used in this phase.

In conclusion, the evaluation framework adopted for the longitudinal phase was designed to align with the methodological consistency established in the cross-sectional phase, while also incorporating adjustments that leverage the strengths of sequence-based modeling. With both structured data and CNN-derived image features available, and a clearly defined set of metrics, the methodology is now fully prepared for implementation. The next chapter presents the empirical results derived from this pipeline and analyzes model behavior in the context of Alzheimer's disease progression.

4. EMPIRICAL STUDY

In this chapter, all the conducted experiments are presented and critically analyzed. Like the previous chapter, the content is organized according to the two main phases of this project. The first phase focuses on identifying a convolutional neural network (CNN) architecture capable of effectively classifying MRI scans into Alzheimer's and non-Alzheimer's cases. The selected model will later be used as a feature extractor in the longitudinal analysis phase and combined with structured clinical variables across multiple time points to model disease progression and evaluate temporal patterns associated with cognitive decline.

4.1. CROSS SECTIONAL ANALYSIS

To determine the best performing CNN, multiple approaches were explored. The primary goal was to design a flexible pipeline using various libraries and components that would allow for the testing of different combinations of preprocessing steps, splitting methods, and model architectures. From exploratory data analysis (EDA) to model evaluation, many configurations were implemented. However, several models and setups were discarded early due to poor initial performance.

The central research question guiding this phase was: **"How can CNN-based methods be used to classify MRI scans into the four stages of Alzheimer's disease with high accuracy and generalizability?"**

The next sections provide a detailed explanation of the empirical exploration of this pipeline, covering exploratory data analysis, dataset construction, validation strategies, model training, and the challenges encountered throughout the process.

The experimental pipeline was divided into two main scripts:

- **Dataset Construction Script** - Responsible for extracting image data and building structured datasets.
- **Model Evaluation Script** - Contains the implementation of data splitting strategies and CNN training and evaluation routines.

4.1.1. Dataset Construction

The dataset construction pipeline is composed of four main classes:

- **DataSetBuilder -** This class manages access to the MRI data inside compressed folders and builds structured datasets from the available image types: **RAW**, **PROCESSED**, and **FSL_SEG**. Each dataset includes relevant metadata (e.g., patient ID, image path, image orientation) and is saved in a format suitable for further analysis and model input. Additionally, the structure dataset with the demographic and clinical data of the patient is also loaded.

- **EDA** – Handles exploratory data analysis, including statistical summaries, distribution visualizations, class balance evaluation, and extraction of preliminary insights into the dataset composition and quality.
- **StructurePreprocessor** – Responsible for dataset cleaning procedures, such as removing duplicates, handling missing values, and ensuring consistency across metadata fields. This step ensures the structural integrity of the dataset before image-level preprocessing.
- **ImagePreprocessor** – Manages image preprocessing operations, including normalization, resizing, and optional transformations. These steps ensure that the MRI scans are standardized before being passed into CNN architectures.

These classes complement each other and the final output consists of datasets ready for training and evaluation.

The process begins with dataset construction (as previously described), resulting in three datasets, one for each image type with the following fields: Patient_ID, Orientation, Image Dimensions, Voxel Size, Session and Image Path.

The datasets are composed as follows:

- **RAW**: 1688 images from 436 subjects
- **PROCESSED**: 2180 images from 436 subjects
- **FSL_SEG**: 436 images from 436 subjects

It is important to clarify that although the subject count is listed as 436, each subject can have multiple sessions (e.g., MR1 and MR2) and these are treated independently in the dataset for classification purposes. This decision is justified by the goal of training the model to classify MRI scans individually, rather than modeling subject-specific longitudinal changes at this stage.

4.1.1.1. Preliminary Exploratory Data Analysis (Pre-EDA)

Before implementing preprocessing strategies, a preliminary EDA was performed to assess data completeness and identify potential issues such as missing values.

All patients had corresponding MRI images, however, the structured dataset (containing demographic and clinical data) included missing values across several fields:

- **Education**: 201 missing entries
- **Socioeconomic Status (SES)**: 220 missing entries
- **MMSE** (Mini-Mental State Examination): 201 missing entries
- **CDR** (Clinical Dementia Rating): 201 missing entries

- **Delay** (Time between sessions): 416 missing entries — only available for MR2

Among the available clinical variables, MMSE (Mini-Mental State Examination) and CDR (Clinical Dementia Rating) are particularly important, as they are widely used to assess cognitive impairment and track the progression of Alzheimer's disease. In this study, CDR was chosen as the ground truth label for classification tasks.

An analysis of missing values revealed that the 201 missing entries in the CDR field corresponded exclusively to patients with no clinical diagnosis (labeled 0), which was confirmed by cross-referencing the total number of diagnosed subjects in the OASIS repository. Although MMSE is a valuable indicator of cognitive status, it is inherently more subjective and exhibited a high number of missing values. Given the risk of introducing bias by imputing such subjective clinical data, no imputation was performed for MMSE.

To ensure label consistency and data integrity, only subjects with valid (non-missing) CDR values were retained. While this filtering step reduced the overall dataset size, it significantly improved the reliability of the labels used for model training and evaluation. After this reduction, the remaining missing values were limited to the SES (Socioeconomic Status) and Delay columns, both of which were handled appropriately during the preprocessing stage.

4.1.1.2. Exploratory Data Analysis and Cleaning process

From this point onwards, all evaluations were conducted exclusively on subjects for whom labels were available. As a result, a new population size was established for each dataset:

- **RAW**: 913 images from 235 subjects
- **PROCESSED**: 1175 images from 235 subjects
- **FSL_SEG**: 235 images from 235 subjects

STRUCTURED DATA ANALYSIS AND PREPROCESSING

To better understand the characteristics of the structured data, a comprehensive distribution analysis was conducted. This included visual evaluations of demographic and clinical variables using histograms and boxplots.

The boxplots did not reveal any unusual dispersion or significant presence of outliers, which contributed to higher confidence in the consistency of these variables. In contrast, the histograms provided more valuable insights. Several variables exhibited a well-distributed spread, indicating good data coverage, even if some values appeared less frequently (Figure 0.1 on the Appendix A).

However, the distribution of the target variable, CDR, raised concerns. The CDR scores range from 0 (no dementia) to 2 (moderate dementia), with non-zero values indicating the presence of Alzheimer’s disease. Unfortunately, the dataset was highly imbalanced, with most of the

subjects having a CDR of 0. This imbalance posed a risk of producing biased and overfitted models.

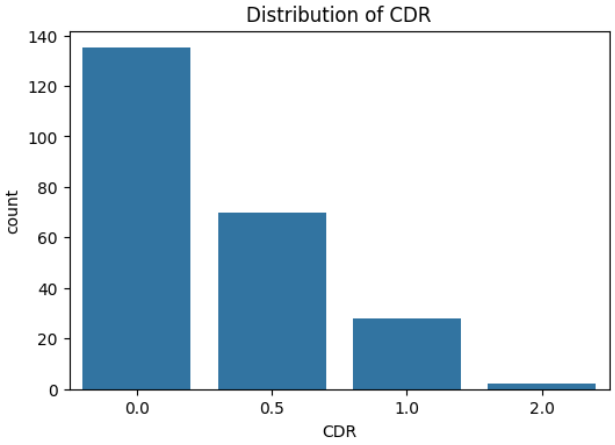

Figure 4.1 - Target variable distribution

Given this imbalance, a change in modeling strategy was warranted. Instead of attempting a multi-class classification (0, 0.5, 1, 2), the problem was reframed as a binary classification task: distinguishing between Alzheimer's (CDR > 0) and non-Alzheimer's (CDR = 0). This simplification helped address the imbalance and improve model robustness.

The MMSE (Mini-Mental State Examination) variable also presented distribution issues. Over 100 individuals scored a perfect 30, which aligns with the high number of subjects with CDR = 0. However, for subjects with lower scores, the distribution was highly skewed, with few values exceeding 40 instances. Despite its clinical importance, the skewness and sparsity in MMSE values suggest it may not be ideal for inclusion as a predictive input variable.

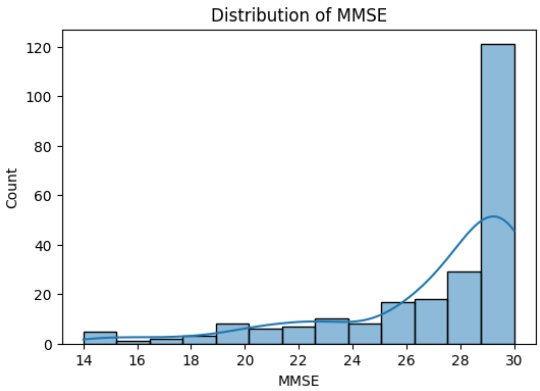

Figure 4.2 - MMSE distribution (Cross sectional)

A correlation matrix was also computed to assess relationships among variables. One notable finding was the strong negative correlation (-0.99) between ASF and eTIV.

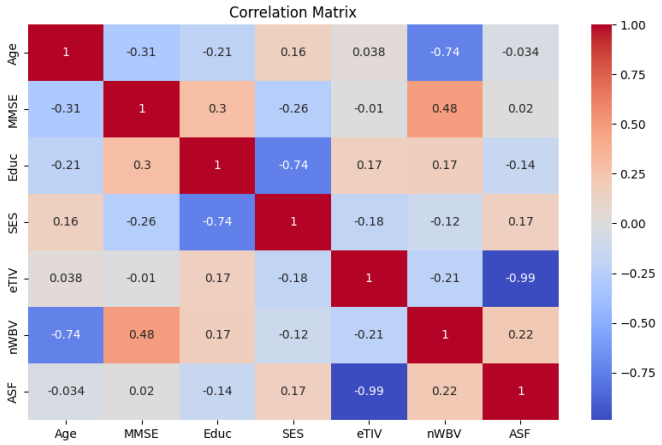

Figure 4.3 - Correlation matrix (Cross sectional)

This high collinearity indicates that using both variables in the same model could introduce multicollinearity, potentially distorting coefficient interpretation and model reliability. Between the two, eTIV is generally more interpretable, as it reflects a biological measure related to total intracranial volume and is commonly used in neurodegeneration research. In contrast, ASF is a scaling factor used mainly for preprocessing and lacks clear biological meaning. Therefore, it was decided to include eTIV and exclude ASF from the model inputs.

Additional preprocessing steps were also performed to improve data quality and modeling performance:

- The columns "Delay" and "Hand" were dropped because of excessive missing values and low variability, respectively.
- The "M/F" (gender) variable was encoded as binary.
- All continuous variables were normalized using StandardScaler, which standardizes features by removing the mean and scaling to unit variance, ensuring comparability across features and aiding convergence in gradient-based models.
- Remaining missing values were imputed using KNN imputation, a method that considers the values of neighboring instances to fill in missing data. This approach was chosen due to its ability to preserve local data structure and correlations, which are often critical in medical datasets.

These steps ensured a clean, standardized dataset, optimized for training reliable machine learning models.

IMAGE PREPROCESSING

In parallel with the structured data preparation, a dedicated preprocessing pipeline was applied to the MRI images to ensure consistency and suitability for deep learning models.

All images were first extracted from (.gif) formats archives and converted to 2D slices in (.png) format. During this process, the image path metadata was updated to reflect the location of the preprocessed images rather than the original archive.

Before applying any preprocessing strategies, a visual inspection was conducted on a representative sample of images from each dataset (RAW, PROCESSED, and FSL_SEG) alongside an evaluation of their structural metadata.

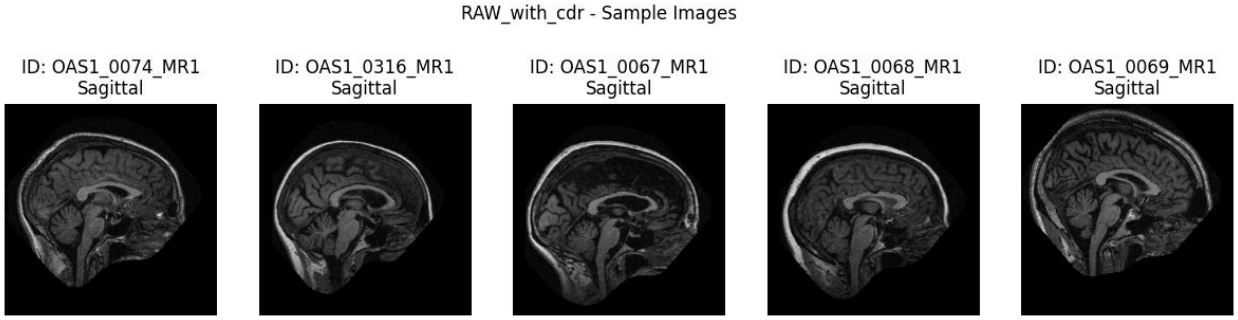

Figure 4.4 - Images presented on the RAW dataset

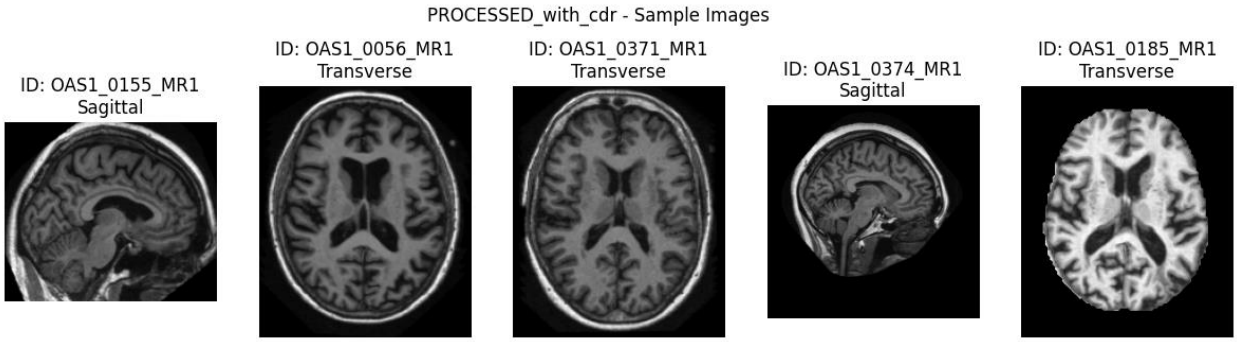

Figure 4.5 - Images presented on the PROCESSED dataset

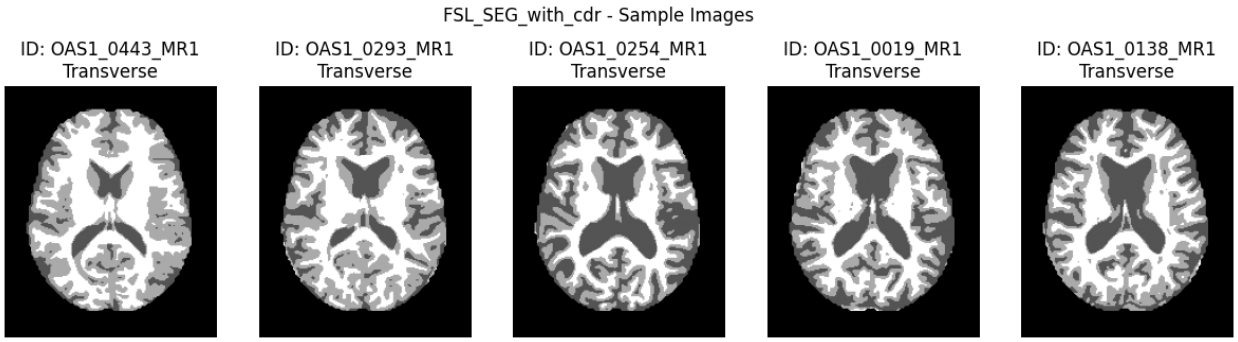

Figure 4.6 - Images presented on the FSL_SEG dataset

The key insight from this inspection was that the datasets contained three distinct types of MRI scans, each with different characteristics:

- RAW: sagittal slices with a fixed resolution of 256×256

- FSL_SEG: transverse slices with a resolution of 176×208
- PROCESSED: scans with multiple orientations and variable resolutions, making them the most heterogeneous group

This variability required models to generalize effectively across different scan types. Accordingly, a preprocessing strategy was developed to standardize inputs while preserving clinically relevant information.

Image resizing was a critical step in this process. Several target dimensions were tested, for example, 128×128, 224×224, and 256×256, to evaluate trade-offs between image clarity and computational efficiency.

For the PROCESSED dataset, 224×224 was selected as the standard, offering a balanced compromise between resolution and model performance. RAW and FSL_SEG images retained their original dimensions during modeling, as they were already consistent and sufficiently optimized.

To enhance model robustness and prevent overfitting, data augmentation techniques were selectively applied during training. These included slight rotations, horizontal flips, and contrast adjustments. However, augmentation was used cautiously, as excessive transformations risk distorting clinically meaningful features.

In addition, non-informative metadata columns such as Vox_size, Dimensions, and Session were removed after resizing and format conversion. The Orientation variable, which provides spatial context about the scan direction, was preserved and encoded using One-Hot Encoding to ensure compatibility with learning algorithms.

Finally, all preprocessed images were linked to the structured data using subject and session identifiers. This enabled a seamless fusion of image and tabular features, laying the groundwork for multi-modal modeling strategies.

4.1.2. Model Evaluation

The modeling pipeline was structured into three main components: PretrainedModel, ManualCNN, and Split. These core classes were responsible for defining and training the models, as well as handling dataset splitting strategies.

- **PretrainedModel** - Encapsulates a set of widely-used convolutional neural networks (CNNs) pretrained on ImageNet, including **ResNet18**, **DenseNet121**, **EfficientNet-B0**, and **VGG16**. These models serve as strong baselines for transfer learning, especially when working with limited medical imaging data.
- **ManualCNN** - A custom-designed convolutional neural network built from scratch. This model was included to evaluate the performance of a lighter, domain-specific architecture compared to heavyweight pretrained models.

- **Split** - Implements different dataset partitioning strategies to evaluate generalization under varied data constraints.

In addition to these main components, several supporting classes were implemented:

- **Evaluator** - Handles the computation of key evaluation metrics (accuracy, precision, recall, F1-score, AUC), confusion matrix plotting, and metric logging.
- **EarlyStopping** - Monitors validation performance and halts training if no improvement is observed after a defined number of epochs, helping to prevent overfitting.
- **GetData** - Manages data loading, preprocessing transformations, and batch generation.
- **Utils** - A utility class providing helper functions used throughout the pipeline (e.g., setting seeds, logging results, and saving models).

The main idea of this pipeline was to evaluate all the possible combinations and try to understand which one performs better for all the three data sets.

4.1.2.1. Configuration Setting

Several configuration options were evaluated during the development of the modeling pipeline, particularly regarding data splitting strategies and model selection. Given the multimodal nature of the datasets, each combining structured clinical information with image data, is necessary to give special attention when determining how to partition the data in a way that preserved both label distribution and subject integrity.

SPLITTING METHODS

All three datasets include structured and image data. The RAW and PROCESSED datasets contain multiple MRI slices per patient, meaning that traditional data partitioning methods risk introducing information leakage if images from the same subject appear in both training and validation sets. This was a key consideration in the selection of appropriate splitting strategies.

Initially, four methods were considered: manual train/test splitting, standard K-Fold cross-validation, Stratified K-Fold, and Group K-Fold.

Manual splitting was quickly excluded due to its lack of reproducibility and potential to introduce arbitrary biases. For relatively small datasets, random partitions may fail to represent the true distribution of labels or subjects, undermining the validity of model evaluation.

Standard K-Fold cross-validation was also discarded. While it provides robust cross-validation in general settings, it does not preserve class distribution across folds. Given the moderate class imbalance in the dataset, this could lead to skewed performance metrics.

Stratified K-Fold and Group K-Fold were the ones that remain in the final pipeline after empirical evaluation and theoretical analysis.

Stratified K-Fold maintains the same proportion of classes in each fold as in the overall dataset, it supports fair evaluation and reduces the risk of biased learning caused by underrepresented classes in training or validation sets.

Group K-Fold can handle repeated measures. It ensures that all data from a single individual remains within one fold, avoids evaluating the model on data from patients it has already learned from during training.

Together, Stratified and Group K-Fold offered complementary benefits: one addressed class imbalance, while the other mitigated subject-level data leakage. Although a combined Stratified Group K-Fold would theoretically offer the best of both worlds, it was not implemented due to practical limitations in existing libraries and the added complexity it would introduce to the pipeline.

MODEL ARCHITECTURES

The primary architecture employed was a flexible deep neural network capable of integrating image features from pre-trained convolutional neural networks (CNNs) with structured clinical features such as demographic and cognitive scores. To explore different design trade-offs, four pre-trained models and one manually implemented CNN were tested.

The pre-trained models chosen were ResNet18, DenseNet121, VGG16, and EfficientNet-B0, selected for their strong performance in computer vision tasks and architectural diversity. Each CNN was used without its classification head, retaining only the feature extraction layers. The extracted image features were then concatenated with embeddings generated from structured clinical data, which were processed through a small multi-layer perceptron (MLP). The resulting combined vector was passed to a fully connected classification head. This design allows the model to jointly learn from both visual and structured inputs, improving performance in multimodal settings.

In addition, a simpler CNN was manually implemented as a baseline. This architecture consisted of two convolutional layers with ReLU activations and max pooling, followed by a dense layer for the structured input. A final classification head processed the concatenated image and structured features. While it does not leverage transfer learning from large-scale datasets, this model provides a lightweight and interpretable alternative for comparison.

To support these architectures, a preprocessing pipeline was established to adapt the input images to the expected formats of each model. This included resizing and data augmentation. Since medical imaging can contain subtle but clinically relevant features, augmentations were carefully chosen and optionally adjustable. For training, transformations such as horizontal flipping, rotation, and color jittering were applied. Normalization using ImageNet statistics

was performed only for the pre-trained models to match the input distributions expected by their feature extractors.

Beyond architecture and preprocessing, several techniques were incorporated to address key modeling challenges. Class imbalance was mitigated through two strategies:

- **Class-weighted loss** – Using CrossEntropyLoss with class weights calculated based on label frequencies.
- **SMOTE oversampling** – Applied only to the structured features. Synthetic samples were generated and paired with the nearest neighbor's image to expand underrepresented classes.

Although alternative strategies for handling class imbalance, such as focal loss, synthetic image augmentation (e.g., MixUp, CutMix), or GAN-based oversampling, have demonstrated effectiveness in computer vision tasks, they were deliberately excluded from this study.

The limited size of the dataset was a primary concern and aggressive augmentation techniques pose the risk of introducing visual artifacts or distorting clinically relevant anatomical features, which is particularly problematic in neuroimaging applications. Given the subtle and localized nature of structural differences associated with dementia, preserving the fidelity of MRI images was prioritized.

Instead, a more conservative and interpretable strategy was adopted. Class-weighted loss was used during training to penalize misclassification of underrepresented classes, while SMOTE was applied solely to the structured data. This decision was based on the understanding that structured features are less susceptible to semantic distortion when synthetically generated. To maintain multimodal consistency, each synthetic structured sample created via SMOTE was paired with the image of its nearest neighbor in the original dataset. Although this approach does not involve direct oversampling of image data, it ensures balanced class representation while preserving the integrity of the visual modality. Overall, this design reflects a deliberate trade-off between methodological rigor and the practical constraints of medical data, with an emphasis on reproducibility and clinical interpretability.

To prevent overfitting and unnecessary computation, early stopping was implemented. The model monitored validation loss, and training was halted if no significant improvement was observed for a predefined number of epochs. After several configurations, a patience of 2 epochs was found to be optimal, especially given the limited size of the dataset.

Overall, this modular and extensible architecture integrated both image-based and structured data through PyTorch components. The combination of diverse CNN backbones, tailored preprocessing, balancing strategies, and regularization helped build robust and reproducible

classification models, well-suited for the challenges of neuroimaging-based dementia assessment.

Finally, the three datasets were evaluated across 10 experimental runs, covering all combinations of cross-validation strategies and model architectures. Each model was tested using consistent evaluation metrics: accuracy, precision, recall, and F1-score.

Since each model was trained and validated across multiple folds, the final performance scores were reported as averages across all folds, offering a more stable and generalizable estimate of model effectiveness.

Results were visualized using:

- Line plots showing training and validation metrics across epochs for the best-performing fold.
- Bar charts comparing mean performance scores of each model.
- Confusion matrices to examine the distribution of predicted vs. actual labels.

These visualizations support a comprehensive comparison of model architectures, not only in terms of raw performance but also with respect to their stability and generalization across different cross-validation configurations.

4.2. LONGITUDINAL ANALYSIS

As in the cross-sectional phase, multiple modeling approaches were explored to identify the best-performing RNN architecture for longitudinal prediction. However, while the initial phase focused heavily on preprocessing steps, model selection, and fold-based validation, the emphasis in this phase shifted toward model configurations and dataset arrangements. This was possible because the preprocessing pipeline and feature extraction steps had already been defined and reused from the cross-sectional methodology.

The guiding research objective for this phase was: **"How effectively can RNN-based models leverage longitudinal imaging and clinical data to identify disease progression over time?"** which frames the exploration of how temporal dependencies and sequential modeling can support the analysis of neurodegenerative change.

The empirical phase was organized around a modular and reusable pipeline that enabled flexible experimentation with different architectures and data formats. This pipeline was implemented using a combination of Python libraries (such as PyTorch, NumPy, and Pandas), and was divided into two core components:

- **Dataset Construction Script**: Responsible for loading CNN-extracted image features and clinical data and evaluating them.

- **Model Evaluation Script**: Implements RNN training, validation, and testing routines, including performance tracking and metrics computation.

The following sections present a detailed walkthrough of the experimental procedures, including exploratory data analysis (EDA), dataset formulation, validation strategies, and the modeling process—highlighting both technical design decisions and challenges encountered throughout.

4.2.1. Dataset Construction

This pipeline was composed of four main classes:

- **Dataset** - Manages access to the MRI data stored in compressed folders and constructs structured datasets from the available image types (in this case, only RAW data). It extracts relevant metadata and organizes it into a format compatible with downstream analysis. Additionally, it loads the tabular dataset containing demographic and clinical information. This class mirrors the structure and functionality used in the cross-sectional phase.
- **Preprocessor** - Applies the same data cleaning and normalization procedures developed during the cross-sectional phase, ensuring consistency in data handling across both study components.
- **FeatureExtractor** - Utilizes the pretrained CNN model (selected in the cross-sectional study) to convert 2D MRI slices into fixed-length feature vectors that will serve as compact representations of each image, enabling compatibility with RNN models.
- **EDA**: Performs statistical analysis and visualization of the longitudinal dataset. This includes summary statistics, distribution plots, class balance checks, and an initial review of data quality and completeness.

4.2.1.1. Image extraction

As previously mentioned, all MRI images in the longitudinal dataset were provided in their original 3D format, stored inside compressed ZIP archives. Since this study was designed to work with 2D image data, a transformation from 3D NIfTI files to 2D slices was necessary during the data extraction process.

To maintain consistency across the dataset and ensure that the most informative regions of the brain were captured, three slices were extracted from each scan, one for each anatomical orientation: sagittal, coronal, and axial. This decision was based on the clinical relevance of these orientations, as different parts of the brain that are affected by Alzheimer's disease (such as the hippocampus, frontal lobe, and brain symmetry) are best visualized from different angles.

With each orientation, it was necessary to select the most representative slice from the 3D volume. Although the dataset was not extremely large, performing manual slice selection for each subject would have been time-consuming and potentially inconsistent, especially given

the limited clinical background of the researcher. As a result, an automated and standardized method was preferred.

Choosing the middle slice for each orientation was considered a reasonable and effective approach. These mid-volume slices are typically centered on key anatomical regions of the brain and are less likely to include peripheral noise or partial structures. In the context of Alzheimer's disease, where brain regions such as the hippocampus, ventricles, and frontal lobes are of diagnostic importance, these central slices often provide a good balance of informative content across patients.

Additionally, several patients had more than one scan per session. To avoid over-representing certain patients and to maintain a uniform dataset structure, only the first available scan per session was used. This approach also ensured that the number of images per session remained consistent across subjects.

The resulting dataset followed the same structure as the one used in the cross-sectional phase, with the only difference being that the input images were derived from 3D longitudinal scans.

4.2.1.2. Feature extraction

After preprocessing implemented identically to the cross-sectional phase, all the same variables were retained in the input dataset, ensuring consistency in the types and formats of features used across both study phases.

Each row of the dataset corresponds to one image, a 2D slice extracted from the original 3D MRI volume and all input images were cleaned, resized, and normalized following the exact procedures used during CNN training in the first part of the study.

The previously trained CNN was used as a feature extractor, meaning it processed each image and output a fixed-length feature vector that summarizes the most relevant spatial characteristics learned during training. These vectors represent high-level, compressed representations of the original images, making them suitable for sequential modeling.

In this study, the CNN was based on a ResNet-18 architecture, producing a 512-dimensional feature vector for each image. Due to the large number of images and the high dimensionality of the output, these feature vectors were initially stored using Python's Pickle format (.pkl), which efficiently serializes Python objects for storage. Later, the vectors were combined with their corresponding target labels and saved in a (.npz) (NumPy zip) format. This choice ensures compact storage and fast loading during model training, helping to avoid memory bottlenecks and improve computational efficiency. This feature extraction pipeline allows raw imaging data to be translated into a numerical format compatible with RNN models while maintaining consistency with the preprocessing and structure established during the cross-sectional phase.

4.2.1.3. Exploratory Data Analysis

The exploratory data analysis (EDA) conducted in this phase was intentionally limited. Since the same structured variables from the cross-sectional phase were used and had previously shown minimal influence on the classification task. Only a brief analysis was performed to reassess their potential usefulness in the longitudinal setting.

The goal of this EDA was to evaluate whether any of the previously excluded tabular variables might now contribute meaningful information when paired with sequential data. Distribution plots and a correlation matrix were generated. However, most variables displayed high inter-correlation and redundancy, while others (such as MMSE) appeared heavily skewed. No additional patterns or strong associations emerged that could directly enhance the classification of progression in the context of this study.

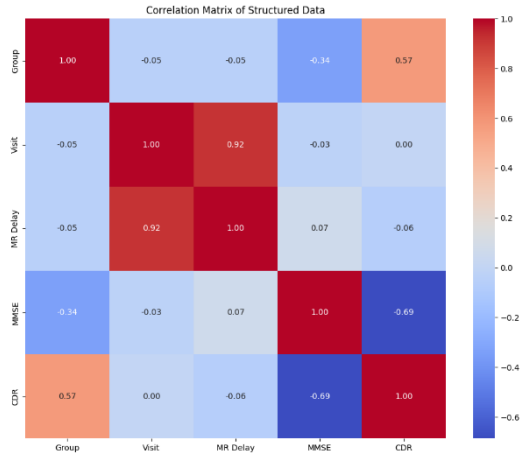

Figure 4.7 - Correlation matrix (Longitudinal)

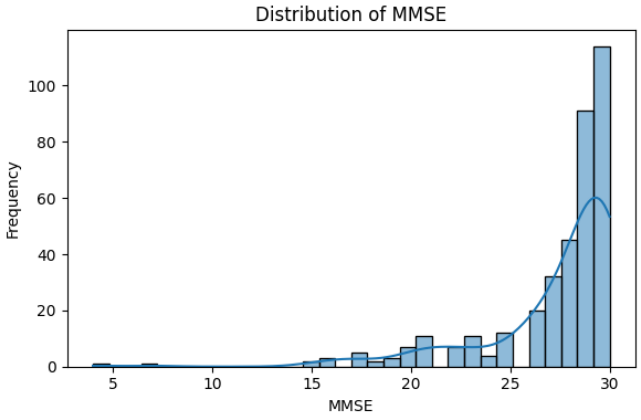

Figure 4.8 - MMSE distribution (Longitudinal)

Given these findings, the structured data was not used directly for modeling in this phase. Instead, the focus shifted to defining a meaningful target variable for dementia progression.

To capture cognitive decline over time, two key clinical variables from the dataset were considered: CDR and Group. While the *Group* variable flags whether a subject converted from non-demented to demented and if stays or not nondemented, it does not account for changes within already demented individuals. The CDR score, on the other hand, provides a graded measure of cognitive impairment and is more sensitive to intra-dementia progression.

This led to the creation of a new binary progression label, defined as follows:

- 0 (No progression) – Subjects initially labeled as *non-demented* (CDR = 0) who remained non-demented across all visits.
- 1 (Progression) – Subjects who were:

- Initially labeled as *demented* (CDR > 0), regardless of whether their CDR increased.
- Initially *non-demented* but later converted (CDR increased from 0 to a higher value).

This definition captures both conversion to dementia and continuation of an already progressive condition, aligning more closely with the longitudinal nature of the data and the study's goal of identifying early or ongoing cognitive decline.

4.2.2. Model Evaluation

In contrast to the more complex pipeline used in the cross-sectional phase, the longitudinal modeling pipeline was intentionally streamlined. It relied primarily on two classes:

- **RNN** – Responsible for grouping input data chronologically by subject and handling the implementation of different recurrent neural network architectures.
- **Evaluator** – Computes key evaluation metrics (accuracy, precision, recall, F1-score, and AUC), generates performance plots, and logs results for comparison.

The primary goal of this phase was to explore how different modeling strategies performed across the three orientation-specific datasets. While the scope of combinations tested was more limited than in the cross-sectional phase, the design remained systematic. Specifically, two different *chronological grouping strategies* were used to structure the input sequences, and four different RNN-based architectures were evaluated. This resulted in a total of eight modeling combinations (2 grouping methods × 4 models).

These experiments aimed to identify which configurations, both in terms of data sequencing and model type, yielded the most accurate predictions of dementia progression over time.

4.2.2.1. Configuration Setting

All configurations in this phase were designed to enhance model performance. Since the input to the model consisted exclusively of vectors extracted by a CNN, the focus shifted entirely to how these vectors were grouped and sequenced to reflect the longitudinal structure of the data.

GROUPING STRATERGY

Before applying any RNN-based modeling, it was essential to group the feature vectors corresponding to the same MRI session. As mentioned earlier, each 3D scan was processed to extract three 2D slices (one per orientation), each producing an individual feature vector. If these vectors were treated independently, it could result in data leakage, where the model unintentionally learns from information seen during training. Therefore, vectors from the same session were first aggregated to form a single representation per MRI visit.

The next level of grouping focused on how to organize the sessions chronologically for each subject, and two primary approaches were tested:

- **Full Subject Grouping**: All sessions of a subject were grouped into a single chronological sequence. These were ordered by visit number and padded when necessary, as not all subjects had the same number of visits. This strategy provides a holistic view of patient progression over time and enables the model to capture long-term patterns between the earliest and latest available visits.
- **Sliding Session Pairs**: In this configuration, pairs of consecutive sessions were grouped (e.g., Visit 1 → Visit 2, Visit 2 → Visit 3). This approach breaks down progression into smaller temporal windows, allowing the model to learn finer-grained transitions. It also increases the number of training samples, which is especially beneficial for smaller datasets like this one (approximately 150 subjects).

Each grouping method serves a distinct purpose. The full-subject grouping captures the broader progression trajectory, useful for understanding long-term changes. The sliding session pairs focus on short-term differences between adjacent visits, potentially enhancing sensitivity to early signs of progression and improving generalization due to the increased training data.

MODEL ARCHITECTURES

In this phase, the modeling focused exclusively on pretrained RNN architectures designed to capture temporal dependencies in longitudinal data. Four pretrained RNN variants were selected for evaluation: LSTM, GRU, BiLSTM and BiGRU.

The rationale for including both unidirectional (LSTM, GRU) and bidirectional (BiLSTM, BiGRU) models was to explore whether modeling the sequence in both temporal directions improves the detection of progression patterns, which can be especially relevant when progression signals are subtle or complex.

The implementation was straightforward, relying mostly on standard hyperparameters such as number of hidden layers, hidden layer size, dropout rate (to reduce overfitting), optimizer choice (e.g., Adam) and learning rate.

To prevent overfitting and reduce unnecessary computation, early stopping was employed. Training was monitored using the F1 score on the validation set, halting if no improvement was observed for 5 consecutive epochs, a patience value found to balance adequate training time against overfitting risk given the dataset size.

Each model was evaluated across 15 experimental runs, covering all combinations of the two grouping strategies and the four RNN architectures. Evaluation metrics included accuracy, precision, recall and F1-score.

In addition, Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) scores were computed, offering a probabilistic perspective on classification performance beyond threshold-dependent metrics.

Results were visualized with:

- Line plots illustrating training and validation metrics over epochs to assess convergence and stability.
- ROC/AUC charts for comprehensive model comparison.

These visualizations and metrics collectively facilitated an in-depth comparison of model architectures, revealing not only their raw predictive performance but also their robustness and generalization capabilities across different data grouping configurations.

Overall, this phase successfully implemented and evaluated four pretrained RNN architectures on longitudinal MRI data, carefully considering different sequence grouping strategies and model configurations. The combined use of standard training practices, early stopping, and multiple evaluation metrics provided a thorough understanding of each model's predictive capabilities and stability.

These efforts laid a solid foundation for the next chapter, where the detailed results and comparisons of these models will be presented and analyzed to identify the best approach for progression classification.

5. RESULTS AND DISCUSSION

In this chapter, we present the outcomes of the modeling process, along with the rationale behind key strategic decisions. As in previous chapters, the results are divided into two complementary parts: one for the cross-sectional approach and one for the longitudinal approach.

5.1. CROSS SECTIONAL

As previously described, the main objective of the cross-sectional stage is to train a convolutional neural network (CNN) capable of extracting meaningful features from MRI scans. These learned features will then be reused during the longitudinal modeling phase for temporal analysis.

After performing exploratory data analysis (EDA) and preprocessing, three datasets were prepared for model training:

- **RAW**: 913 images from 235 subjects
- **PROCESSED**: 1175 images from 235 subjects
- **FSL_SEG**: 235 images from 235 subjects

Each dataset shared the same structure: a target variable (y), a column containing the image path, and a set of structured (tabular) features (x). The intention was to evaluate how well the CNN model could handle each image modality, ensuring versatility and generalizability for the next stage of the study.

The initial approach was to run the full training pipeline using all three datasets, to compare their performance and draw conclusions about their suitability for feature extraction. However, it quickly became apparent that FSL_SEG and RAW should be excluded from further evaluation, for different reasons:

FSL_SEG

This dataset contains only one image per subject, making it the smallest among the three. While training was stable across several model configurations, the validation accuracy consistently hovered around 0.65, indicating limited generalization capabilities and potential overfitting. The small size and lack of diversity made it unsuitable for robust CNN training.

Additionally, the images in FSL_SEG had undergone significant processing (including skull stripping, segmentation, and spatial normalization). These transformations, while useful for specific types of neuroimaging analysis, make the dataset less compatible with the goal of building a model that can generalize MRI scans in the longitudinal phase. Therefore, this dataset was excluded from the final training pipeline.

RAW

The RAW dataset showed strong results in terms of model performance. With approximately 3 to 4 images per subject, it provided a decent number of training examples. However, all images were captured in the same orientation (sagittal). While this uniformity made training easier since the model faced less variation, it limited the model's exposure to the variety of spatial patterns present in other orientations. This is particularly problematic because sagittal views tend to capture fewer Alzheimer's-related anatomical features than coronal or axial slices.

Despite high training and validation accuracy, the model trained on RAW data lacked the diversity needed for optimal generalization. Relying solely on sagittal images increases the risk of bias and misclassification when the model is later exposed to more complex or varied inputs.

Given the shortcomings of the other datasets, all final results presented in this chapter are based on the PROCESSED dataset. This dataset offers several key advantages, it includes five images per subject, providing a richer representation of brain anatomy, contains images in three orientations, for sagittal and transverse, each subject has two images, increasing intra-orientation variability.

This dataset addresses both the limitations of FSL_SEG (insufficient size and over-processing) and RAW (lack of orientation diversity). It is therefore the most suitable dataset for training a CNN intended for downstream feature extraction in a longitudinal setting.

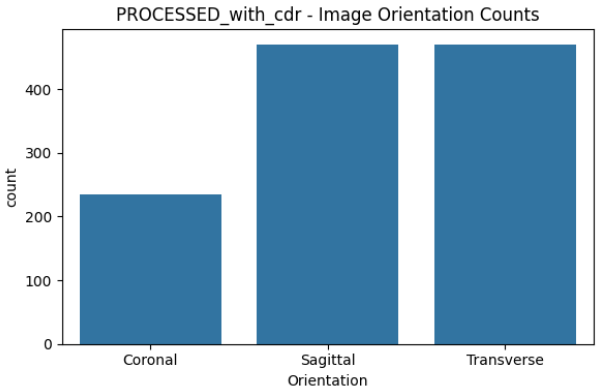

Figure 5.1 - PROCESSED orientation distribution

5.1.1. Processed Final Configurations

After excluding the FSL_SEG and RAW datasets, all training and evaluation efforts focused exclusively on the PROCESSED dataset. Several model configurations were tested with the aim of identifying the most effective setup for extracting relevant features from the available data. Since this dataset includes multiple images per subject and a mix of orientations, the training strategies were designed specifically to account for these characteristics.

Initial experiments evaluated standard image resolutions commonly used in CNNs, such as 128×128 and 256×256. However, the original images in the PROCESSED dataset vary in size

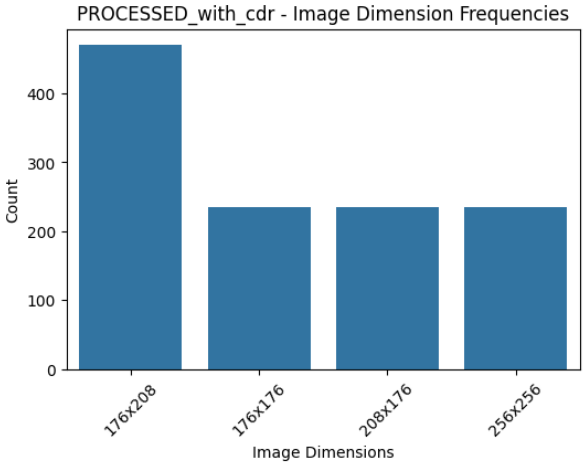

Figure 5.2 - PROCESSED dimensions distribution

which made the use of extreme resizing less desirable. Downscaling too aggressively risked discarding important anatomical information, while upscaling could introduce artifacts and unnecessary computational burden.

To strike a balance, an intermediate resolution of 224×224 was selected as the final image size. This resolution is a commonly accepted standard in pretrained CNN architectures, ensuring compatibility and preserving sufficient image detail for effective feature extraction.

Two main data splitting strategies were evaluated during experimentation: Stratified K-Folds Cross-Validation and Group K-Folds Cross-Validation.

While stratified K-Folds yielded slightly better validation performance (likely due to more evenly distributed labels), it was ultimately excluded due to a critical risk of data leakage. Since the PROCESSED dataset includes multiple images per subject, along with associated tabular features, using stratified K-Folds could result in the same subject appearing in both training and validation sets, albeit with different images. This overlap could allow the model to "memorize" subject-specific patterns rather than generalize, thereby inflating performance metrics and undermining the model's ability to perform in real-world conditions.

To prevent this, Group K-Folds was selected as the final splitting method. Although it may reduce variability between folds and increase the difficulty of training, it provides a more realistic and unbiased estimate of model performance by strictly separating subjects across training and validation data.

With this, all final model evaluations and experiments were conducted with the following settings:

- Image resolution: 224×224

- Data splitting method: Group K-Folds Cross-Validation
- Number of folds: 5
- Tabular data and image input combined through a custom CNN + MLP architecture

This configuration ensures that the model learns from a diverse and balanced representation of the PROCESSED dataset, while avoiding biases and overfitting due to patient duplication across splits.

5.1.2. Processed Results

To identify the best model for feature extraction, a set of convolutional neural networks were evaluated under consistent conditions using the PROCESSED dataset. The evaluation followed a structured cross-validation pipeline (5-fold Group K-Fold), with results compared across multiple metrics: validation loss, accuracy, precision, recall, and F1 score.

VGG16 was excluded early from this comparative analysis. Despite being a popular architecture, its performance was consistently poor across all datasets, including extremely low validation scores and a tendency to overfit. Additionally, the architecture's depth and outdated design resulted in longer training times without meaningful improvement, making it unsuitable for the current task. For these reasons, it was removed alongside the RAW and FSL_SEG datasets.

5.1.2.1. ResNet-18

ResNet-18 is a lightweight residual network known for its ability to mitigate the vanishing gradient problem through shortcut connections, making it efficient for training even with limited data.

Table 5.1 - ResNet performance results

Fold	Val Loss	Accuracy	Precision	Recall	F1 Score
0	0.6622	0.7191	0.7761	0.7191	0.7189
1	0.7193	0.6511	0.6511	0.6511	0.6511
2	0.6204	0.6979	0.7124	0.6979	0.7006
3	0.7921	0.6128	0.6336	0.6128	0.5810
4	0.6773	0.7319	0.7500	0.7319	0.6965

ResNet-18 demonstrated relatively consistent performance, with F1 scores ranging between 0.65 and 0.72. However, Fold 3 appeared as an outlier, possibly due to more complex patient cases or imbalanced class representation in that group. This highlights a limitation of Group K-Fold splitting, especially with a limited number of subjects: certain folds may inherently contain more challenging samples.

On average, ResNet-18 achieved an accuracy of 68.3% (95% CI: 62.96%–73.56%) and an F1-score of 66.9% (95% CI: 60.94%–72.99%), indicating reasonably stable performance under varying training-validation splits. Confidence intervals were computed using the standard error of the mean across folds and reflect the robustness of the model's results.

The model showed a tendency to prioritize precision over recall, meaning it was slightly better at correctly identifying demented patients while occasionally missing some. This is relevant in medical screening tasks where false positives might be preferable to false negatives. Training typically converged quickly, with early stopping occurring between 3 to 6 epochs, suggesting efficient learning.

5.1.2.2. DenseNet-121

DenseNet-121 is a deeper architecture that connects each layer to every other layer in a feed-forward fashion, allowing it to reuse features and reduce the number of parameters while maintaining high representational power.

Table 5.2 - DenseNet performance results

Fold	Val Loss	Accuracy	Precision	Recall	F1 Score
0	0.5851	0.6936	0.6887	0.6936	0.6878
1	0.6811	0.6255	0.6293	0.6255	0.6207
2	0.6427	0.6340	0.6286	0.6340	0.6302
3	0.6369	0.6894	0.6286	0.6894	0.6885
4	0.5707	0.6894	0.7099	0.6979	0.7018

DenseNet-121 showed a more balanced performance across folds and across metrics. Its F1 scores are stable and its precision and recall values are consistently close, indicating that it does not favor one class over the other. A very desirable behavior in clinical prediction settings.

The average performance across folds was strong and relatively stable, with an accuracy of 66.8% (95% CI: 62.64%–70.99%) and an F1-score of 66.6% (95% CI: 62.04%–71.11%). The relatively narrow confidence intervals reinforce the observed consistency and imply that the model is robust to variation in the training-validation split.

Compared to ResNet-18, DenseNet-121 generally required one or two more epochs to converge before triggering early stopping. This may reflect its greater depth and capacity to generalize, especially in a limited data regime. The results support DenseNet's ability to extract and propagate richer features through dense connections while remaining computationally efficient.

5.1.2.3. EfficientNet_B0

EfficientNet is a newer family of CNNs that scale width, depth, and resolution in a balanced way. EfficientNet-B0, the base model, is known for achieving strong performance while being highly efficient computationally.

Table 5.3 - EfficientNet performance results

Fold	Val Loss	Accuracy	Precision	Recall	F1 Score
0	0.6263	0.6894	0.7106	0.6894	0.6922
1	0.6577	0.6851	0.7210	0.6851	0.6740
2	0.6742	0.6383	0.7203	0.6383	0.6319
3	0.6472	0.6766	0.6766	0.6766	0.6742
4	0.6369	0.6809	0.7026	0.6809	0.6864

EfficientNet-B0 maintained consistent accuracy and precision across folds. The precision was slightly higher than recall in all folds, indicating a cautious prediction strategy (more conservative about labeling a patient as demented) which may be suitable for clinical screening scenarios where overdiagnosis is less critical than missing early-stage cases.

Statistical reporting revealed an average accuracy of 67.4% (95% CI: 65.12%–69.70%) and an F1-score of 67.2% (95% CI: 64.42%–69.92%). The confidence intervals were moderately tight, suggesting a stable and reliable performance across the five validation folds.

Although the validation loss remained slightly higher than DenseNet's, the model still achieved competitive F1 scores. Training often stopped earlier (e.g., epoch 4), suggesting fast convergence but possibly also early saturation in learning.

All of this supports the model’s reputation for efficiency and effective feature extraction, even when applied to relatively small medical imaging datasets.

5.1.2.4. Best Model

To provide a clearer picture, the average performance across all five folds for each model is summarized below:

Table 5.4 - Models performance comparison

Model	Val Loss	Accuracy	Precision	Recall	F1 Score
Resnet18	0.6943	0.6826	0.7046	0.6826	0.6696
Densenet121	0.6681	0.6681	0.6691	0.6681	0.6658
Efficientnet_b0	0.6487	0.6741	0.7062	0.6741	0.6717

While all models performed within a close range, EfficientNet-B0 achieved the highest average F1 score and lowest validation loss, indicating strong overall balance and generalization. However, these average metrics alone do not capture variability across folds.

To quantify the stability of each model, 95% confidence intervals were computed for the core evaluation metrics. EfficientNet-B0 achieved an F1-score of 67.2% (95% CI: 64.4%–69.9%), followed closely by ResNet-18 at 66.9% (95% CI: 60.9%–73.0%), and DenseNet-121 at 66.6% (95% CI: 62.0%–71.1%). These overlapping intervals suggest that, statistically, there is no significant performance difference between the models. Thus, additional clinical and practical factors were considered in selecting the final model.

A more nuanced perspective is provided by the confusion matrices of each model, which reveal specific trade-offs between sensitivity and specificity:

Table 5.5 - CNN Models performance comparison

Model	True Positives	False Positives	True Negatives	False Negatives
Resnet18	95	45	69	**26**
Densenet121	92	33	70	**40**
Efficientnet_b0	94	31	65	**45**

- ResNet-18 had the lowest number of false negatives (26), indicating stronger sensitivity to identifying demented patients.
- DenseNet-121 produced 40 false negatives.
- EfficientNet-B0, despite its strong average scores, had the highest number of false negatives (45).

In medical screening tasks, minimizing false negatives is especially critical, as missed diagnoses can lead to delayed care or mismanagement. In contrast, false positives are typically less harmful and can be corrected in follow-up assessments.

For this reason, ResNet-18 was selected as the final model for feature extraction in the longitudinal phase of the study. Its efficient training, competitive metrics, and clinically favorable decision patterns make it the most appropriate choice.

A graphic with performance metric curves for each model's best-performing fold is provided in Appendix A.

Building on this foundation, the next phase of the study focuses on the longitudinal modeling of disease progression. Leveraging the features extracted using ResNet-18, combined with structured clinical data, the longitudinal analysis aims to capture temporal patterns and trajectories that underpin cognitive decline and dementia progression over time. This approach will enable more precise characterization and prediction of patient outcomes, ultimately contributing to improved clinical decision-making and personalized care.

5.2. LONGITUDINAL

The initial input data consisted of several rows, each representing a combination of three vectors corresponding to orientation slices per patient per session. Initially, grouping two sessions per patient was considered a better approach. However, this method proved to be suboptimal. Models trained using session-based grouping performed very poorly, with some models reaching an accuracy of 0, clearly indicating serious issues with generalization or potential data leakage.

Although this grouping allowed the model access to more apparent patterns, there was too much similarity between sessions and too little distinguishable variation. As a result, the model often failed to generalize, possibly confusing repeated patterns for previously seen data. Surprisingly, instead of leveraging these patterns, the models consistently failed, sometimes predicting completely at random, as indicated by an accuracy of zero. This suggests that the model was unable to detect meaningful differences between sessions, making it ineffective for any two-session grouping strategy.

Consequently, all subsequent experiments were conducted using grouping by subject. While this approach significantly reduced the dataset size, it led to a noticeable improvement in model performance. This trade-off between dataset size and generalization capacity underscores the importance of avoiding session-based overlap, which can introduce data leakage and mislead model evaluation.

Table 5.6 - RNN Models Performance comparison

Model	Val Loss	Accuracy	Precision	Recall	F1 Score
LSTM	0.6762	0.5667	0.5789	0.6875	0.6286
GRU	0.7039	0.6000	0.6111	0.6875	0.6471
BiLSTM	0.7005	0.5667	0.5789	0.6875	0.6286
BiGRU	0.7336	0.5333	0.5500	0.6875	0.6111

This table shows that there was no significant performance difference among the tested models. The GRU model achieved the highest accuracy (60.00%), precision (0.6111), and F1 score (0.6471), indicating that it handled false positives better and achieved the most balanced results in those specific metrics. However, its AUC score was the lowest (0.50), suggesting poor ability to distinguish between classes across different thresholds.

The LSTM and BiLSTM models produced nearly identical outcomes across all core metrics. Interestingly, BiLSTM achieved the highest AUC (0.61), suggesting slightly better discriminative ability overall — despite its otherwise average performance. This indicates that no single model was consistently superior across all evaluation metrics. The use of bidirectionality in both BiLSTM and BiGRU did not yield clear performance advantages for this specific task or dataset.

In contrast, the BiGRU model showed the weakest overall performance, with the lowest accuracy (53.33%), F1 score (0.6111), and the highest validation loss (0.7336), possibly indicating overfitting or difficulty in extracting meaningful features from the limited data.

Interestingly, all models achieved the same recall (0.6875), indicating equal ability to detect positive cases. However, differences in precision and F1 scores reveal that some models, particularly GRU, were more prone to false positives, while others, like BiLSTM, were slightly better at class discrimination as reflected by AUC.

As previously noted, in medical contexts, key evaluation metrics include recall, ROC, and AUC. High recall is crucial for minimizing missed diagnoses (false negatives), while ROC curves allow for threshold-independent performance assessment by evaluating the trade-off between sensitivity and specificity. AUC condenses this performance into a single value and reflects a model's overall capacity to discriminate between classes.

Although recall values were relatively high, the ROC curves and corresponding AUC scores point to weak class discrimination. AUC values ranged from 0.50 to 0.61, with curves clustering close to the diagonal line, indicating behavior close to random guessing. This suggests that, despite identifying positives reasonably well, the models lacked consistent decision boundaries across thresholds.

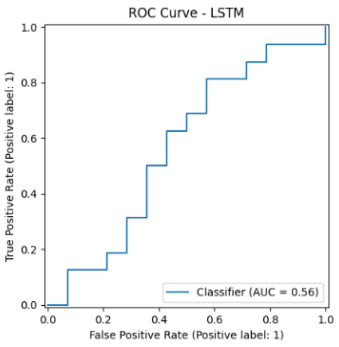

Figure 5.3 - LSTM ROC curve

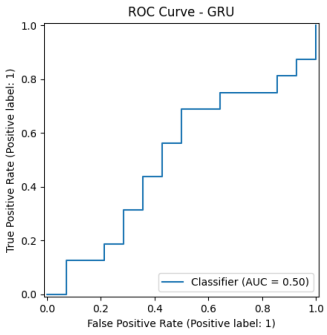

Figure 5.4 - GRU ROC curve

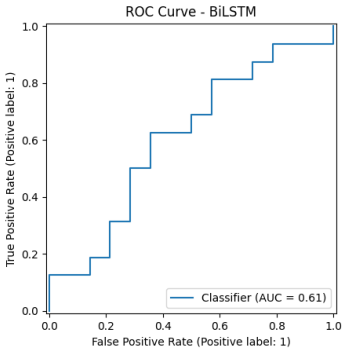

Figure 5.5 - BiLSTM ROC curve

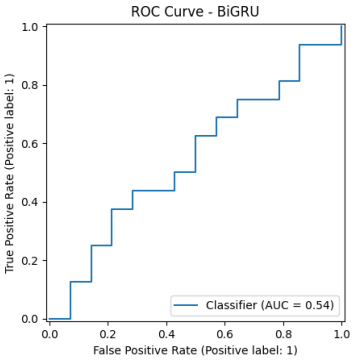

Figure 5.6 - BiGRU ROC curve

These observations are reinforced by the validation loss values, all above 0.65. This indicates low model confidence and poor generalization to unseen data, likely a result of overfitting due to the small and relatively homogeneous dataset.

The primary objective of this study was to assess model performance using minimally processed data, simulating real-world medical conditions where full preprocessing may not always be feasible or appropriate. In clinical practice, imaging data is typically preprocessed through spatial and intensity normalization, denoising, artifact correction, contrast enhancement, resampling, and especially segmentation, which isolates anatomical structures or lesions. These steps help reduce variability and emphasize the most diagnostically relevant regions.

However, aggressive or unguided preprocessing can risk removing subtle but clinically meaningful features. For that reason, such preprocessing is typically performed or supervised by medical professionals, not general data analysts. In this study, preprocessing was deliberately minimized to preserve raw clinical signals and avoid the risk of discarding relevant information. While this choice reflects a cautious and realistic clinical approach, it also came with a trade-off: the dataset, composed of only 150 subjects and lacking enhancement, offered insufficient variation for the models to learn robustly.

As a result, none of the tested architectures, regardless of complexity, were able to generalize well or demonstrate consistently strong performance. The dataset's limited size and homogeneity were the dominant constraints on model performance.

In this longitudinal analysis, no statistical significance testing was performed. This decision was made due to the weak overall performance of the models, the low AUC scores, and the limited size and diversity of the dataset. Given these constraints, the inclusion of formal statistical validation was considered inappropriate, as it would not meaningfully contribute to the interpretation of results. Nonetheless, future work with larger and more heterogeneous samples should incorporate statistical validation techniques to assess the robustness and reliability of model performance.

In summary, although no model demonstrated clear superiority, the GRU architecture showed slightly better performance in conventional metrics (accuracy, precision, F1), while BiLSTM achieved the best AUC. More complex architectures, such as bidirectional networks, did not yield meaningful improvements in this setting. These findings highlight the crucial role of dataset size and diversity in model generalization, as well as the importance of clinically guided preprocessing when developing machine learning models for medical imaging data.

6. CONCLUSIONS AND FUTURE WORKS

The primary objective of this study was to investigate the potential of deep learning models for the early prediction of Alzheimer's Disease using MRI data. This was approached through a two-stage framework: Classification of Alzheimer's vs. non-Alzheimer's cases from cross-sectional MRI data using convolutional neural networks and Modeling disease progression vs. non-progression using longitudinal data with recurrent neural networks, such as GRU, LSTM, and their bidirectional variants.

This design was intended to ultimately enable early-stage prediction from a single MRI input by combining diagnostic classification with progression modeling.

While the full end-to-end pipeline could not be fully realized due to dataset limitations, the study offered important insights into the capabilities and constraints of deep learning architectures in real-world medical imaging scenarios.

Several challenges were encountered throughout the research. Limited computational resources and deep learning expertise initially delayed experimentation, but these hurdles were gradually addressed through alternative training strategies and model tuning. However, the most critical limitation was the size and homogeneity of the dataset, which included only 150 subjects. This lack of variation significantly restricted the models' ability to generalize, resulting in low AUC scores and high validation losses, both indicative of overfitting and poor class separability.

Despite these limitations, some findings emerged. The GRU model demonstrated slightly better performance on conventional classification metrics such as accuracy and F1 score, while BiLSTM achieved the highest AUC, indicating better performance across decision thresholds. However, no model showed clear superiority across all evaluation metrics and the differences in performance were not statistically significant.

Additionally, the study highlights several critical considerations for medical AI research. Deep learning models require not only more data, but more diverse and clinically representative data. While this work deliberately minimized preprocessing to retain raw clinical signals, some clinically guided preprocessing, such as segmentation or artifact correction, could enhance learning without removing diagnostically relevant features.

Combining cross-sectional and longitudinal data, as attempted here, is a valuable direction, as it captures both static and dynamic aspects of disease progression, which may be essential for early diagnosis. But several future work directions are recommended.

- **Expand the dataset:** Increasing both the number and diversity of subjects is paramount. Multi-center datasets such as those from ADNI can provide the scale and variability necessary for meaningful model generalization.
- **Integrate clinical preprocessing:** Future studies should explore the selective use of preprocessing techniques (e.g., segmentation, normalization, denoising) performed under clinical supervision to enhance model learning while preserving critical diagnostic information.

- **Refine the early prediction pipeline:** With better data and preprocessing, the two-stage approach, classification followed by progression modeling, can be restructured into a single, end-to-end pipeline capable of predicting AD at its earliest stages.
- **Strengthen interdisciplinary collaboration:** Close cooperation with neurologists and radiologists is essential to ensure models align with clinical understanding and offer real-world utility. Their input can help shape both feature selection and the interpretation of model outputs.
- **Enhance interpretability through visual explanation tools:** incorporate techniques like Grad-CAM to highlight important regions in medical images, use saliency maps and SHAP to explain model decisions on clinical data and integrate attention mechanisms in temporal models to identify key time points for disease progression.

Although this study faced important limitations, it contributes to the growing body of work exploring AI in early Alzheimer's detection. It demonstrates both the potential and pitfalls of using deep learning in low-preprocessing, small-dataset conditions, realistic challenges in many clinical settings. By addressing these issues through data expansion, methodological refinement, and clinical collaboration, future research can move closer to delivering robust, trustworthy tools for early AD prediction.

BIBLIOGRAPHICAL REFERENCES

Ali, M. U., Kim, K. S., Khalid, M., Farrash, M., Zafar, A., & Lee, S. W. (2024). Enhancing Alzheimer's disease diagnosis and staging: A multistage CNN framework using MRI. *Frontiers in Psychiatry*, *15*. https://doi.org/10.3389/fpsyt.2024.1395563

AlSaeed, D., & Omar, S. F. (2022). Brain MRI Analysis for Alzheimer's Disease Diagnosis Using CNN-Based Feature Extraction and Machine Learning. *Sensors*, *22*(8), Article 8. https://doi.org/10.3390/s22082911

Alzheimer's Association 2024 Alzheimer's Disease Facts and Figures. (2024).

Atri, A. (2019). The Alzheimer's Disease Clinical Spectrum: Diagnosis and Management. *Medical Clinics of North America*, *103*(2), 263–293. https://doi.org/10.1016/j.mcna.2018.10.009

Basaia, S., Agosta, F., Wagner, L., Canu, E., Magnani, G., Santangelo, R., & Filippi, M. (2019). Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage: Clinical*, *21*, 101645. https://doi.org/10.1016/j.nicl.2018.101645

Breijyeh, Z., & Karaman, R. (2020). Comprehensive Review on Alzheimer's Disease: Causes and Treatment. *Molecules*, *25*(24), Article 24. https://doi.org/10.3390/molecules25245789

Buckner, R. L., Head, D., Parker, J., Fotenos, A. F., Marcus, D., Morris, J. C., & Snyder, A. Z. (2004). A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: Reliability and validation against manual measurement of total intracranial volume. *NeuroImage*, *23*(2), 724–738. https://doi.org/10.1016/j.neuroimage.2004.06.018

Cao, Q., Tan, C.-C., Xu, W., Hu, H., Cao, X.-P., Dong, Q., Tan, L., Yu, J.-T., & Zhu, L.-Q. (2020). The Prevalence of Dementia: A Systematic Review and Meta-Analysis. *Journal of Alzheimer's Disease*, *73*(3), 1157–1166. https://doi.org/10.3233/JAD-191092

Chang, C.-H., Lin, C.-H., & Lane, H.-Y. (2021). Machine Learning and Novel Biomarkers for the Diagnosis of Alzheimer's Disease. *International Journal of Molecular Sciences*, *22*(5), 2761. https://doi.org/10.3390/ijms22052761

Cui, R., & Liu, M. (2019). RNN-based longitudinal analysis for diagnosis of Alzheimer's disease. *Computerized Medical Imaging and Graphics*, *73*, 1–10. https://doi.org/10.1016/j.compmedimag.2019.01.005

Cui, R., Liu, M., & Alzheimer's Disease Neuroimaging Initiative. (2019). RNN-based longitudinal analysis for diagnosis of Alzheimer's disease. *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society*, *73*, 1–10. https://doi.org/10.1016/j.compmedimag.2019.01.005

Dubois, B., Villain, N., Frisoni, G. B., Rabinovici, G. D., Sabbagh, M., Cappa, S., Bejanin, A., Bombois, S., Epelbaum, S., Teichmann, M., Habert, M.-O., Nordberg, A., Blennow, K., Galasko, D., Stern, Y., Rowe, C. C., Salloway, S., Schneider, L. S., Cummings, J. L., & Feldman, H. H. (2021). Clinical diagnosis of Alzheimer's disease: Recommendations of the International Working Group. *The Lancet. Neurology*, *20*(6), 484–496. https://doi.org/10.1016/S1474-4422(21)00066-1

Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). Machine Learning for Medical Imaging. *Radiographics*, *37*(2), 505–515. https://doi.org/10.1148/rg.2017160130

Eskildsen, S. F., Coupé, P., Fonov, V. S., Pruessner, J. C., & Collins, D. L. (2015). Structural imaging biomarkers of Alzheimer's disease: Predicting disease progression.

Neurobiology of Aging, *36*, S23–S31. https://doi.org/10.1016/j.neurobiolaging.2014.04.034

Esmaeilzadeh, S., Belivanis, D. I., Pohl, K. M., & Adeli, E. (2018). End-To-End Alzheimer's Disease Diagnosis and Biomarker Identification. *Machine Learning in Medical Imaging. MLMI (Workshop)*, *11046*, 337–345. https://doi.org/10.1007/978-3-030-00919-9_39

Ferrer, I. (2012). Defining Alzheimer as a common age-related neurodegenerative process not inevitably leading to dementia. *Progress in Neurobiology*, *97*(1), 38–51. https://doi.org/10.1016/j.pneurobio.2012.03.005

Fotenos, A. F., Snyder, A. Z., Girton, L. E., Morris, J. C., & Buckner, R. L. (2005). Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD. *Neurology*, *64*(6), 1032–1039. https://doi.org/10.1212/01.WNL.0000154530.72969.11

Gustavsson, A., Norton, N., Fast, T., Frölich, L., Georges, J., Holzapfel, D., Kirabali, T., Krolak-Salmon, P., Rossini, P. M., Ferretti, M. T., Lanman, L., Chadha, A. S., & van der Flier, W. M. (2023). Global estimates on the number of persons across the Alzheimer's disease continuum. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *19*(2), 658–670. https://doi.org/10.1002/alz.12694

Hazarika, R. A., Maji, A. K., Syiem, R., Sur, S. N., & Kandar, D. (2022). Hippocampus Segmentation Using U-Net Convolutional Network from Brain Magnetic Resonance Imaging (MRI). *Journal of Digital Imaging*, *35*(4), 893–909. https://doi.org/10.1007/s10278-022-00613-y

Hu, Z., Wu, L., Jia, J., & Han, Y. (2014). Advances in longitudinal studies of amnestic mild cognitive impairment and Alzheimer's disease based on multi-modal MRI techniques. *Neuroscience Bulletin*, *30*(2), 198–206. https://doi.org/10.1007/s12264-013-1407-y

Jomeiri, A., Navin, A. H., & Shamsi, M. (2024). Longitudinal MRI analysis using a hybrid DenseNet-BiLSTM method for Alzheimer's disease prediction. *Behavioural Brain Research*, *463*, 114900. https://doi.org/10.1016/j.bbr.2024.114900

Ker, J., Wang, L., Rao, J., & Lim, T. (2018). Deep Learning Applications in Medical Image Analysis. *IEEE Access*, *6*, 9375–9389. IEEE Access. https://doi.org/10.1109/ACCESS.2017.2788044

Khatun, M., Islam, M. M., Rifat, H. R., Shahid, M. S. B., Talukder, M. A., & Uddin, M. A. (2023). Hybridized Convolutional Neural Networks and Long Short-Term Memory for Improved Alzheimer's Disease Diagnosis from MRI Scans. *2023 26th International Conference on Computer and Information Technology (ICCIT)*, 1–6. https://doi.org/10.1109/ICCIT60459.2023.10441274

Khojaste-Sarakhsi, M., Haghighi, S. S., Ghomi, S. M. T. F., & Marchiori, E. (2022). Deep learning for Alzheimer's disease diagnosis: A survey. *Artificial Intelligence in Medicine*, *130*, 102332. https://doi.org/10.1016/j.artmed.2022.102332

Kumar, S., Payne, P. R. O., & Sotiras, A. (2023). Normative Modeling using Multimodal Variational Autoencoders to Identify Abnormal Brain Volume Deviations in Alzheimer's Disease. *Proceedings of SPIE--the International Society for Optical Engineering*, *12465*, 1246503. https://doi.org/10.1117/12.2654369

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, F., Liu, M., & Alzheimer's Disease Neuroimaging Initiative. (2019). A hybrid Convolutional and Recurrent Neural Network for Hippocampus Analysis in Alzheimer's Disease. *Journal of Neuroscience Methods*, *323*, 108–118. https://doi.org/10.1016/j.jneumeth.2019.05.006

Liu, J., Pan, Y., Li, M., Chen, Z., Tang, L., Lu, C., & Wang, J. (2018). Applications of deep learning to MRI images: A survey. *Big Data Mining and Analytics*, *1*(1), 1–18. Big Data Mining and Analytics. https://doi.org/10.26599/BDMA.2018.9020001

Liu, M., Cheng, D., Yan, W., & Alzheimer's Disease Neuroimaging Initiative. (2018). Classification of Alzheimer's Disease by Combination of Convolutional and Recurrent Neural Networks Using FDG-PET Images. *Frontiers in Neuroinformatics*, *12*. https://doi.org/10.3389/fninf.2018.00035

Lu, D., Popuri, K., Ding, G. W., Balachandar, R., & Beg, M. F. (2018). Multimodal and Multiscale Deep Neural Networks for the Early Diagnosis of Alzheimer's Disease using structural MR and FDG-PET images. *Scientific Reports*, *8*, 5697. https://doi.org/10.1038/s41598-018-22871-z

Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift Für Medizinische Physik*, *29*(2), 102–127. https://doi.org/10.1016/j.zemedi.2018.11.002

Maity, R., Raja Sankari, V. M., Snekhalatha, U., Velu, S., Alahmadi, T. J., Alhababi, Z. A., & Alkahtani, H. K. (2024). Early detection of Alzheimer's disease in structural and functional MRI. *Frontiers in Medicine*, *11*, 1520878. https://doi.org/10.3389/fmed.2024.1520878

Mantzavinos, V., & Alexiou, A. (2017). Biomarkers for Alzheimer's Disease Diagnosis. *Current Alzheimer Research*, *14*(11), 1149–1154. https://doi.org/10.2174/1567205014666170203125942

Mayeux, R., & Stern, Y. (2012). Epidemiology of Alzheimer Disease. *Cold Spring Harbor Perspectives in Medicine*, *2*(8), a006239. https://doi.org/10.1101/cshperspect.a006239

Mienye, I. D., & Swart, T. G. (2024). A Comprehensive Review of Deep Learning: Architectures, Recent Advances, and Applications. *Information*, *15*(12), Article 12. https://doi.org/10.3390/info15120755

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, *19*(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

Morris, J. C. (1993). The Clinical Dementia Rating (CDR). *Neurology*, *43*(11), 2412-2412-a. https://doi.org/10.1212/WNL.43.11.2412-a

Nawaz, A., Anwar, S. M., Liaqat, R., Iqbal, J., Bagci, U., & Majid, M. (2021). *Deep Convolutional Neural Network based Classification of Alzheimer's Disease using MRI data* (No. arXiv:2101.02876). arXiv. https://doi.org/10.48550/arXiv.2101.02876

Nguyen, M., He, T., An, L., Alexander, D. C., Feng, J., & Yeo, B. T. T. (2020). Predicting Alzheimer's disease progression using deep recurrent neural networks. *NeuroImage*, *222*, 117203. https://doi.org/10.1016/j.neuroimage.2020.117203

Nigri, E., Ziviani, N., Cappabianco, F., Antunes, A., & Veloso, A. (2020). *Explainable Deep CNNs for MRI-Based Diagnosis of Alzheimer's Disease* (No. arXiv:2004.12204). arXiv. https://doi.org/10.48550/arXiv.2004.12204

Pan, D., Zeng, A., Zou, C., Rong, H., & Song, X. (2021). Early detection of Alzheimer's disease using 3D convolutional neural networks. *Alzheimer's & Dementia*, *17*(S4), e053169. https://doi.org/10.1002/alz.053169

Pan, Y., Liu, M., Lian, C., Zhou, T., Xia, Y., & Shen, D. (2018). Synthesizing Missing PET from MRI with Cycle-consistent Generative Adversarial Networks for Alzheimer's Disease Diagnosis. *Medical Image Computing and Computer-Assisted Intervention : MICCAI ...*

International Conference on Medical Image Computing and Computer-Assisted Intervention, *11072*, 455–463. https://doi.org/10.1007/978-3-030-00931-1_52

Passeri, E., Elkhoury, K., Morsink, M., Broersen, K., Linder, M., Tamayol, A., Malaplate, C., Yen, F. T., & Arab-Tehrany, E. (2022). Alzheimer's Disease: Treatment Strategies and Their Limitations. *International Journal of Molecular Sciences*, *23*(22), 13954. https://doi.org/10.3390/ijms232213954

Patil, V., Madgi, M., & Kiran, A. (2022). Early prediction of Alzheimer's disease using convolutional neural network: A review. *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery*, *58*(1), 130. https://doi.org/10.1186/s41983-022-00571-w

Podolszańska, J. (2024). Development and Optimization of Deep Learning Systems for MRI Analysis in Alzheimer's Disease Monitoring. *Journal of Telecommunications and Information Technology*, 56–61. https://doi.org/10.26636/jtit.2024.4.1815

Qiu, C., Kivipelto, M., & von Strauss, E. (2009). Epidemiology of Alzheimer's disease: Occurrence, determinants, and strategies toward intervention. *Dialogues in Clinical Neuroscience*, *11*(2), 111–128. https://doi.org/10.31887/DCNS.2009.11.2/cqiu

Rahman, A. U., Ali, S., Saqia, B., Halim, Z., Al-Khasawneh, M. A., AlHammadi, D. A., Khan, M. Z., Ullah, I., & Alharbi, M. (2025). Alzheimer's disease prediction using 3D-CNNs: Intelligent processing of neuroimaging data. *SLAS Technology*, *32*. https://doi.org/10.1016/j.slast.2025.100265

Sarvamangala, D. R., & Kulkarni, R. V. (2022). Convolutional neural networks in medical image understanding: A survey. *Evolutionary Intelligence*, *15*(1), 1–22. https://doi.org/10.1007/s12065-020-00540-3

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Sorbi, S., & Ferrari, C. (2021, July). *The complexity of Alzheimer's disease: An evolving puzzle*. https://doi.org/10.1152/physrev.00015.2020

Wang, P., Fan, E., & Wang, P. (2021). Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognition Letters*, *141*, 61–67. https://doi.org/10.1016/j.patrec.2020.07.042

Yang, S., Zhu, F., Ling, X., Liu, Q., & Zhao, P. (2021). Intelligent Health Care: Applications of Deep Learning in Computational Medicine. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/fgene.2021.607471

Zhang, X., Chou, J., Liang, J., Xiao, C., Zhao, Y., Sarva, H., Henchcliffe, C., & Wang, F. (2019). Data-Driven Subtyping of Parkinson's Disease Using Longitudinal Clinical Records: A Cohort Study. *Scientific Reports*, *9*, 797. https://doi.org/10.1038/s41598-018-37545-z

Zhao, Y., Guo, Q., Zhang, Y., Zheng, J., Yang, Y., Du, X., Feng, H., & Zhang, S. (2023). Application of Deep Learning for Prediction of Alzheimer's Disease in PET/MR Imaging. *Bioengineering*, *10*(10), 1120. https://doi.org/10.3390/bioengineering10101120

Zhu, L.-Y., Shi, L., Luo, Y., Leung, J., & Kwok, T. (2022). Brain MRI Biomarkers to Predict Cognitive Decline in Older People with Alzheimer's Disease. *Journal of Alzheimer's Disease*, *88*(2), 763–769. https://doi.org/10.3233/JAD-215189

Zunair, H., & Ben Hamza, A. (2021). Sharp U-Net: Depthwise convolutional network for biomedical image segmentation. *Computers in Biology and Medicine*, *136*, 104699. https://doi.org/10.1016/j.compbiomed.2021.104699

APPENDIX A

The following images are additional graphics that were printed during the two phases with exploratory data analysis and model training.

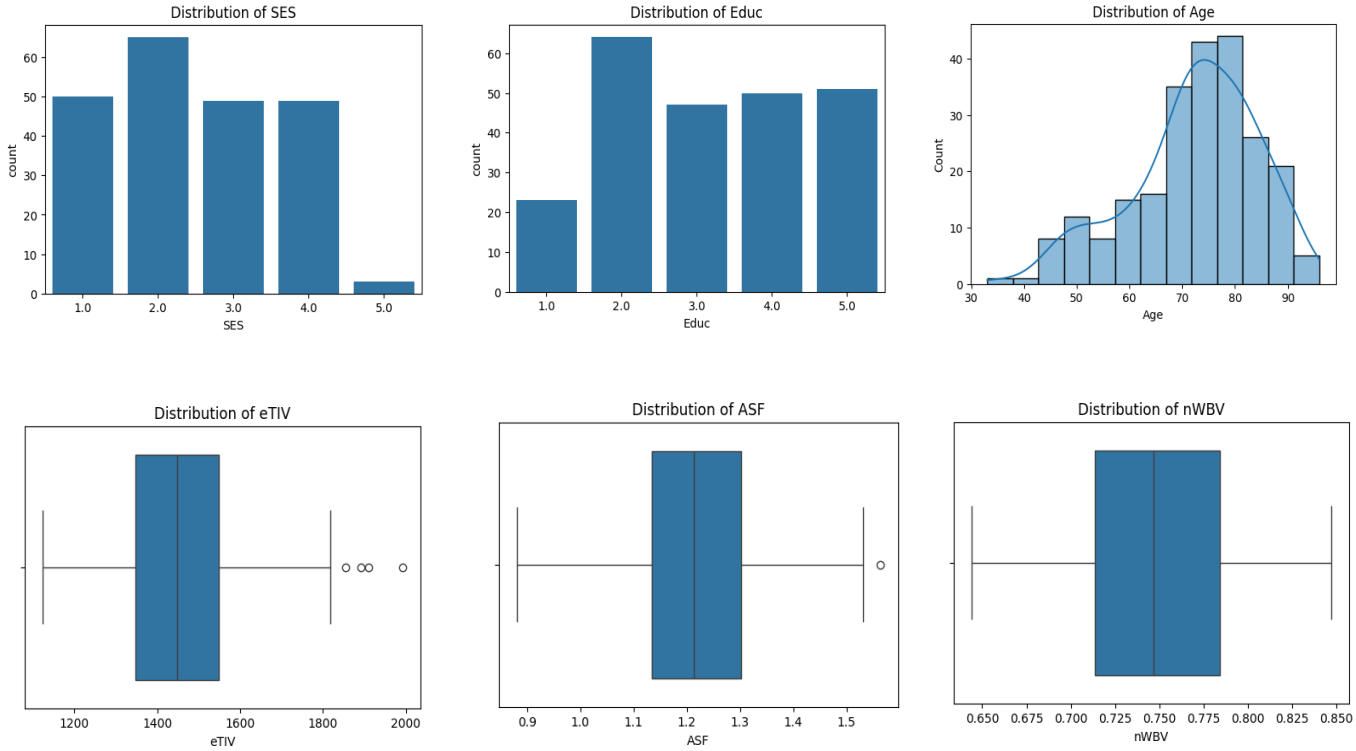

Figure 0.1 - Additional Cross Sectional EDA

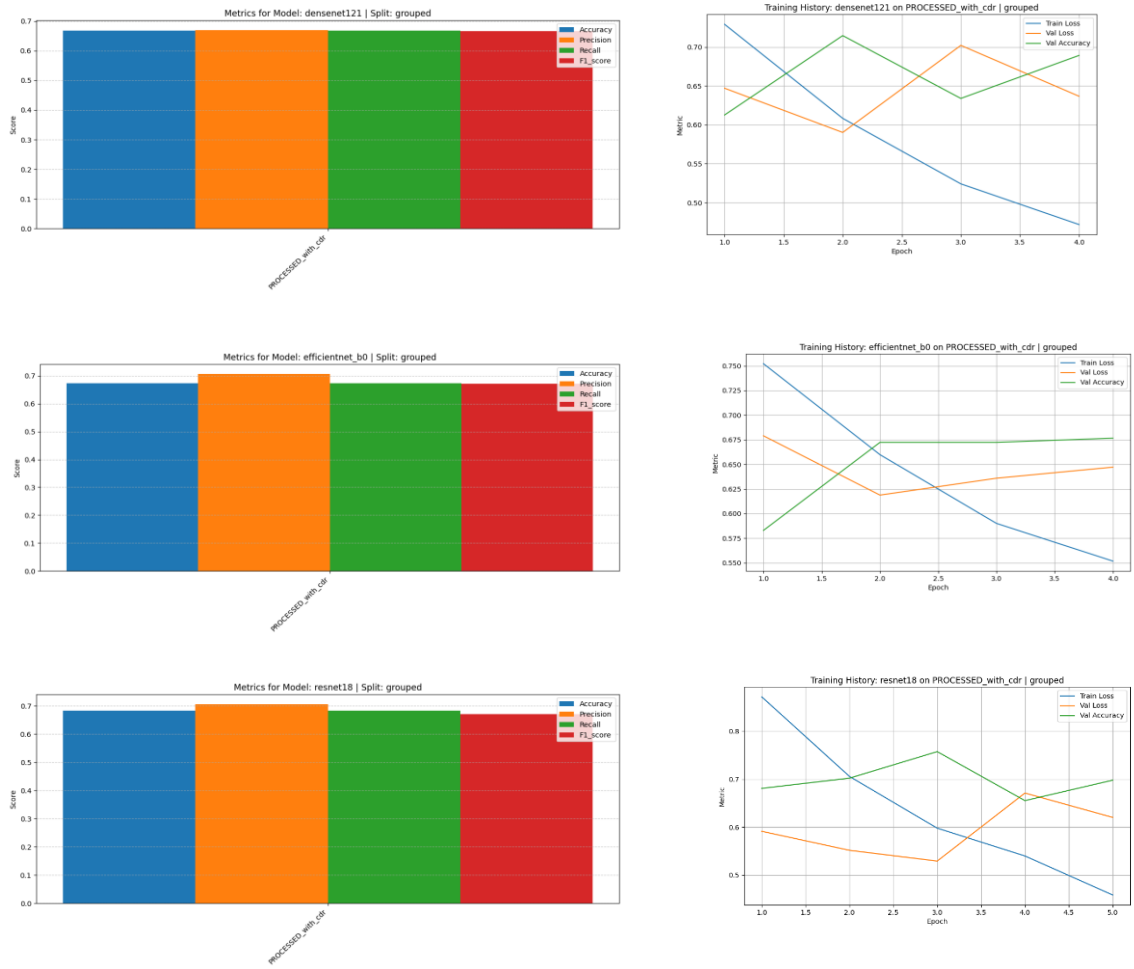

Figure 0.2 - Cross Sectional models visualization

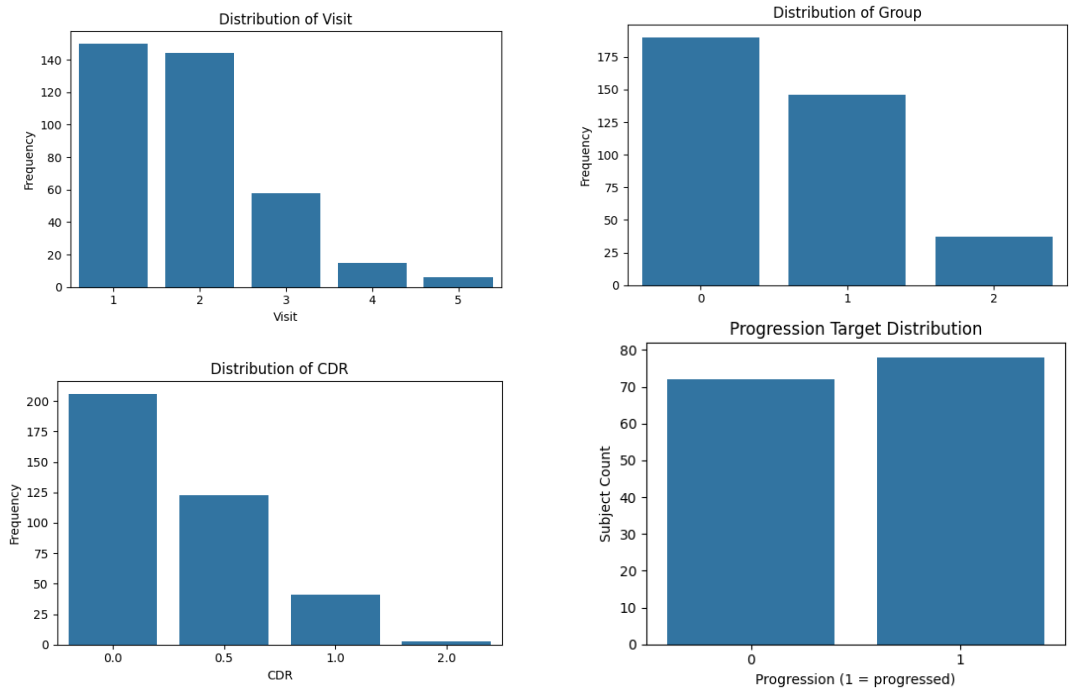

Figure 0.3 - Additional Longitudinal EDA

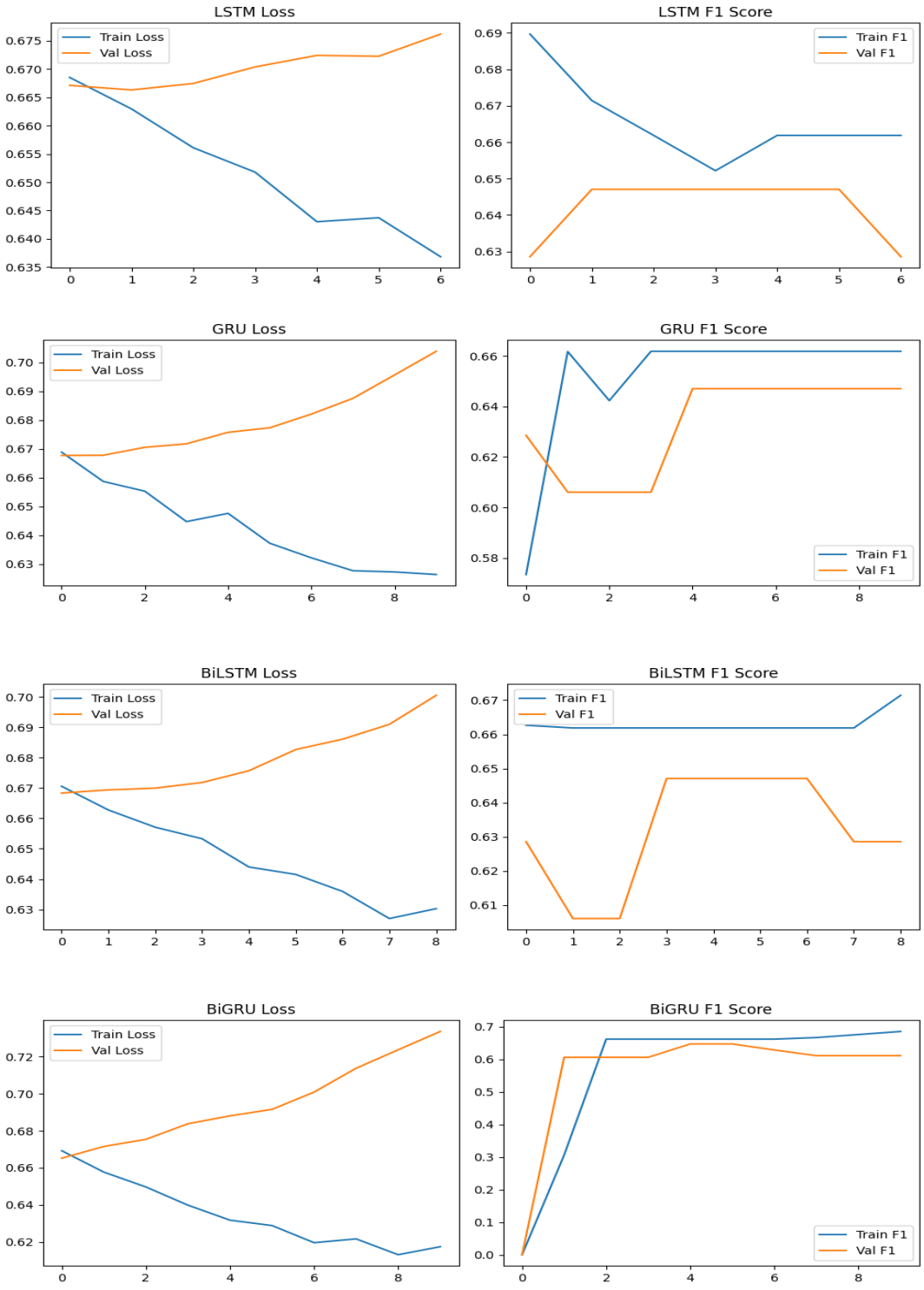

Figure 0.4 - Longitudinal models visualization

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa