TÉCNICO LISBOA

Curriculum Learning for early Alzheimer's Disease diagnosis

Catarina Mendes Faustino Gracias

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisor(s): Prof. Maria Margarida Campos da Silveira

Examination Committee

Chairperson: Prof. João Miguel Raposo Sanches
Supervisor: Prof. Maria Margarida Campos da Silveira
Member of the Committee: Prof. Ana Catarina Fidalgo Barata

October 2021

I declare that this document is an original work of my own authorship and that it fulfils all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Para o meu tio Carlos

Preface

The work presented in this thesis was performed performed at Instituto Superior Técnico, University of Lisbon (Lisbon, Portugal), during the period March-October 2021, under the supervision of Prof. Margarida Silveira.

Acknowledgments

First and foremost, I would like to thank my Mom and my Dad for all their love, patience, dedication and support. You were the ones who made this possible. Also a special thanks to my brother, Miguel, for making me feel at home wherever we are in the world.

A sincere thanks to my Grandmother, for all the advise, optimism and comfort food, now and throughout all my life.

I would like to express my deepest appreciation to Professor Margarida Silveira, not only for trusting me to carry out this project but also for providing me the tools, the motivation and the enthusiasm that made the last 8 months an incredible and fulfilling journey. I could not have had better guidance.

I am also grateful to Dr Durval Campos Costa for delineating and providing the Regions of Interest for Alzheimer's disease.

A special thanks to the Institute for Systems and Robotics (ISR) for providing me the conditions to develop this project and to Professor João Paulo Costeira, for helping me out of the kindness of his heart at all times I needed.

I would also like to thank Professor José Maria Rasquinho, for recognizing and awakening my appreciation and curiosity for science and for encouraging me to discover and follow my own way.

I cannot begin to express my thanks to all my friends. The ones from childhood, who always make the stressfull days feel lighter and turn the calm nights into the best memories. The ones from university, whom I have known for five years but feels like they have been here all the time, I will never forget all the sweat and laughs we shared. The ones that came along they way, like my friend and now housemate, Lucy.

Last but from the bottom of my heart, I would like to give my honest and warm thanks to João, who has been my best friend all these years and now also my love. Thank you for all the times you have read this, all the motivation and inspiration you give me and for making my life so beautiful and joyful.

Resumo

Os estágios iniciais e assintomáticos da doença de Alzheimer, como o défice cognitivo ligeiro, são difíceis de classificar, mesmo por médicos experientes. Por esse motivo, métodos de aprendizadem profunda, como redes neurais de convolução, têm sido implementados com o mesmo propósito, alcançando desempenhos de classificação semelhantes ou até melhores do que os dos próprios médicos. Embora esses métodos tenham a vantagem de que as características das imagens são extraídas automaticamente em vez de manualmente, a sua arquitetura tradicional não permite a incorporação de conhecimento médico. Nesta tese, propomos implementar estratégias de aprendizagem por currículo em redes neuronais de convolução desenhadas para distinguir entre sujeitos saudáveis, com défice cognitivo ligeiro e com doença de Alzheimer. Aprendizagem por currículo é uma estratégia de treino das redes que tenta imitar a maneira como os humanos, neste caso os médicos, aprendem, apresentando primeiro os dados mais fáceis ao modelo e adicionando gradualmente dados mais complexos. Diversas estratégias de aprendizagem por currículo, manuais e automáticas, foram implementadas, incorporando conhecimento médico, para melhorar o desempenho das redes no diagnóstico precoce de Alzheimer. Estas estratégias foram comparadas com modelos usados tradicionalmente e os resultados mostraram claramente que a utilização de aprendizagem por currículo melhora o F1-score (até 3.3%) e a exatidão geral (até 4.5%), particularmente a da DCL (até 11.3%).

Palavras-chave: Doença de Alzheimer, Défice cognitivo ligeiro, Redes neurais de convolução, Aprendizagem por currículo, Conhecimento médico

Abstract

The early and asymptomatic stages of Alzheimer's Disease (AD), such as Mild Cognitive Impairment (MCI) are hard to classify, even by experienced physicians. For this reason, deep learning methods, such as Convolutional neural networks (CNNs), have been implemented for the same purpose, achieving similar or even better classification performance. Although these methods have the advantage that features are automatically extracted rather than handcrafted, their traditional architecture does not allow for the incorporation of medical knowledge. We propose to implement Curriculum Learning (CL) strategies into CNNs designed to diagnose healthy subjects, MCI and AD. CL is a training strategy of the networks that tries to mimic the way humans, and in this cases doctors, learn, by presenting the easier data to the model first and gradually adding more complex data. Several CL strategies, manual and automatic, were implemented, incorporating medical knowledge, to boost the networks performance for early AD diagnosis. They were compared to commonly used baseline deep learning models and the results showed that they clearly improve the F1-score (up to 3.3%) and the overall accuracy (up to 4.5%), particularly that of MCI (up to 11.3%).

Keywords: Alzheimer's disease, Mild cognitive impairment, Convolution neural networks, Curriculum learning, Medical Knowledge

Contents

List of Tables

List of Figures

Nomenclature

Constants

α	Balance factor
δ	Growing factor
θ	Focusing parameter
E	Number of epochs
K	Number of classes
M	Batch size
N	Number of samples

Functions

$\lambda(t)$	Growing function
F	Activation function

Variables

λ	Threshold
b	Bias
s	Sample
t	Epoch
w	Edge's weight
x	Node input
Y	Node value

Arrays

γ	Curriculum
$\hat{p}_y$	Estimated probability distribution

l Loss values

$weight$ Weight values

y Integer class label

List of Acronyms

2D Two dimensional

3D Three dimensional

AD Alzheimer's Disease

AdaDelta Adaptive Delta

AdaGrad Adaptive Gradient

ADAM ADAptive Moment Estimation

ADAS Alzheimer's Disease Assessment Scale

ADNI Alzheimer's Disease Neuroimaging Initiative

AG-CNN Attention-guided CNN

ANN Artificial Neural Network

API Application programming interface

APOE Apolipoprotein E4

APOE Apolipoprotein E

AUC Area under the curve

CAD Computer-assisted diagnosis

CDR Clinical Dementia Rating

CI Cognitive Impairment

CL Curriculum Learning

CNN Convolutional Neural Network

CNNs Convolutional neural networks

CSF Cerebrospinal Fluid

CT Computed Tomography

DASL Deep Active Self-paced Learning

DTI Diffusion Tensor Imaging

EHR Eletronic health records

eMCI early-stage Mild Cognitive Impairment

FC Fully-connected

FDG 18F-Fluorodeoxyglucose

FDG-PET 18F-Fluorodeoxyglucose - Positron Emission Tomography

FL Focal Loss

fMRI functional Magnetic Resonance Imaging

FPR False Positive Rate

IEEE Institute of Electrical and Electronics Engineers

ISBI International Symposium on Biomedical Imaging

lMCI late-stage Mild Cognitive Impairment

LSTM Long short-term memory

MCI Mild Cognitive Impairment

MCIc Mild Cognitive Impairment - converters

MCInc Mild Cognitive Impairment - non converters

MMSE Mini-Mental State Examination

MRI Magnetic Resonance Imaging

NC Normal Control

PET Positron Emission Tomography

ReLU Rectified Linear Unit

ROC Receiver Operating Characteristics

ROIs Regions of Interest

RRNs Recurrent Neural Networks

SDG Stochastic Gradient Descent

SPCL Self-paced Curriculum Learning

SPL Self-paced Learning

SW Sample weights

TeUS Temporal enhanced Ultrasound

TieNet Text-Image embedding Network

US Ultrasound

Chapter 1

Introduction

1.1 Motivation

Alzheimer's Disease (AD) is a progressive and terminal neurodegenerative disorder [1]. It is considered one of the leading causes of death in developed countries, since there is yet no cure available [2]. The clinical research done for AD highly depends on the ability to diagnose AD patients accurately and in an early stage of the disease. An early diagnosis allows to recruit patients for clinical trials, contributing to the ongoing search for treatments and a cure. It also enhances the chances of the available treatments at the time to be relevant for the patients, delaying the disease onset [1, 3].

The early and pre symptomatic stages of AD, such as Mild Cognitive impairment (MCI), are not easily identified by following the traditional diagnostic approaches, where medical doctors collect and analyze the patient data alone [4]. Moreover, the analysis and interpretation of medical datasets is time-consuming and easily influenced by the biases and potential fatigue of human experts [5]. Consequently, AD research can benefit from the use of computer-assisted diagnosis (CAD) systems, which rely on the application of machine learning methods to make faster, earlier and more accurate diagnosis [4]. Currently, Convolutional Neural Networks (CNNs), which allow features being automatically extracted from images rather than handcrafted, have already been successful in AD diagnosis through the classification of medical images [1]. Nevertheless, these recent approaches still have some drawbacks, such as the vulnerability to overfitting problems, which are often related to the small size of available medical datasets, and the fact that they are not usually optimized to incorporate medical knowledge, such as doctors training pattern, i.e., the way or the order by which medical doctors learn/train, or information about cognitive test scores or Regions of Interest (ROIs) for AD diagnosis.

In this paper, as way to overcome these bottlenecks, we propose to develop novel curriculum learning (CL) strategies to more accurately diagnose early AD. The strategies will incorporate medical knowledge, such as the doctors training pattern, scores of the patient's cognitive tests and ROIs for AD diagnosis, into the neural networks. The goal is not is not only to improve their early diagnostic predictions but also to improve the reliability of those diagnosis, as more relevant medical information is used by the models [3].

1.2 Objectives and Original Contributions

This thesis focus on developing curriculum learning strategies for the incorporation of medical domain knowledge into CNNs for AD diagnosis, as a way to improve the model's performance. After developing these strategies, the goal is to use them not only to distinguish between healthy subjects and AD patients, but also to improve the classification of MCI patients, allowing to make a better early AD diagnosis. The classification task will be performed by CNNs using aPositron Emission Tomography (PET) dataset as input. To sum up, the work developed in this project is divided into 2 main goals:

1. Develop curriculum learning strategies, manual and automatic.

2. Use the developed strategies to improve the early AD diagnosis (distinguishing between AD, MCI and healthy controls).

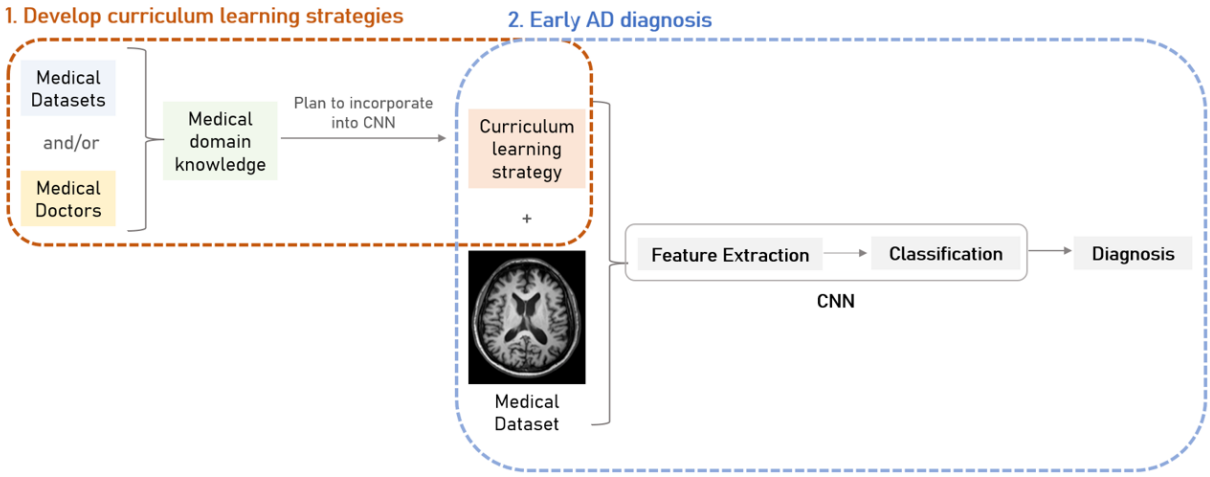

Figure 1.1: Schematic representation of this project's goals: incorporate medical knowledge into networks through curriculum learning techniques; then use the developed strategies to classify AD patients from MCI patients and healthy subjects, as a way to improve early AD diagnosis.

To the extent of our knowledge, this was the first work developing CL strategies, where the learning curriculum was defined using medical knowledge, for improving AD diagnosis. This knowledge was incorporated into the training of neural networks: first the basic concepts of the problem are learnt and only after the more complex aspects are gradually introduced. In order to define data complexity two types of strategies were used: the manual ones, which use information from other medical datasets, such as cognitive test scores, to infer about data complexity; and the automatic ones, that use information about the model's learning performance for the same purpose. All strategies were developed and adjusted to the classification task at hand and consist in novel CL approaches implemented for early AD diagnosis. A paper describing them has been submitted to the 2022 IEEE International Symposium on Biomedical Imaging (ISBI).

After the implementation of the CL strategies and the evaluation of the models' performance on classifying AD, MCI and healthy subjects, we expect to obtain an improvement of such performances when compared to the baseline methods currently used for the same classification task.

1.3 Thesis Outline

This work is composed of 7 chapters. Chapters 2, 3 and 4 are theoretical chapters which cover the background and the state of the art of the topics discussed in this dissertation. Chapter 2 focuses on AD pathophysiology and provides the necessary information about the biomarkers used to diagnose it and evaluate its progress. Chapter 3 provides information about the usage of deep learning models for AD diagnosis. First the relevance of deep learning, more concretely CNNs, in medical image analysis is explained. This is followed by an explanation about how CNNs work, their architecture and training specifications. In the end, several works using CNNs for AD diagnosis are described. Afterwards, details about medical image data are specified: their major pre processing techniques, how to deal with the small size of medical image datasets and how to avoid data leakage. In chapter 4 a summary of the current used techniques to incorporate medical knowledge into deep learning models and their applications is presented. Here the curriculum learning strategies are highlighted and further detailed, since they are the focus of this dissertation. In Chapter 5 the followed methods and original contributions are described and the results are presented and discussed in Chapter 6. In chapter 7 the conclusions of this work are presented as well as future recommendations.

Chapter 2

Alzheimer's disease

AD is a progressive neurodegenerative disorder that slowly destroys memory and the person's ability to reason and function independently [1]. It is considered to be the result of multiple factors rather than a single cause, being the advancing age one of its greatest risk factors. This debilitating disease can be characterized as a combination of cognitive, motor and behavioural deterioration, which eventually becomes overwhelming and devastating both to patients and their families [6].

The disease evolution can differ a lot from patient to patient. However, AD commonly leads to difficulties in communication, learning, recalling new information and performing basic daily activities, such as getting dressed or walking. Patients who suffer from this disease also lose some executive functions (such as planning and judgment) and are usually unaware of their memory or cognitive compromise [7]. There are also some neuropsychiatric symptoms common in AD: apathy and reduced interest in the early clinical stages of AD and depression (50%) and delusions (25%) as the disease progresses [2]. Over time, all cognitive deficits and social dependence increase, the patient's quality of life and motor abilities decrease and, eventually, AD becomes fatal. Patients with mild and moderate AD either progress to advanced-stage disease and die from complications of the decline in brain function, or succumb to comorbid age-related illnesses. These illnesses, such as cardiovascular disease, stroke and cancer, shorten the patients' lifespan [8].

AD is considered one of the leading causes of death in developed countries and the most common type of dementia in the world, accounting for 60% to 70% of cases of progressive cognitive impairment in elderly patients [1][2]. The average duration of illness is 8–10 years [8], its prevalence is higher in women than man and it doubles every 5 years after the age of 60, with higher frequency in those aged 85 years and older [2]. Despite AD being expected to affect 1 out of 85 people in the world by the year 2050 [9], there is still no cure available.

2.1 Pathophysiology and disease evolution

Although AD is the most prevalent mental disease of the world, its pathophysiology, i.e., the mechanisms that cause, result from, or are associated with the disease, are not yet fully understood. Concerning the processes related to AD, the formation of amyloid-β-containing plaques and the deposition of neurofibrillary tangles composed mainly of hyperphosphorylated τ protein, are proposed to be early toxic events in the pathogenesis [7]. The amyloid-β-containing plaques are formed due to the body's failure to clear the amyloid-β peptide from the interstices of the brain, which leads to the accumulation of this protein in and around the neurons. The accumulation of these particles, amyloid-β peptide and hyperphosphorylated τ protein, results in synaptic dysfunction, brain shrinkage and cell death, which is reflected on the slow decline in memory, thinking and reasoning skills [8]. While increased memory loss and confusion are usually the first AD symptoms, the first changes in the brain (such as the presence of abnormal biomarkers, e.g., amyloid-β peptide, τ protein) occur before cognitve decline begins. The brain changes that lead to AD may begin up to 20 years before the symptoms arise [1].

Regarding the evolution of the disease, usually abnormalities are first detected in the brain tissue that involves the frontal and temporal lobes. They slowly progress to other areas of the neocortex according to each patient's rate, reaching wide areas of the cerebral cortex and hippocampus [8]. In the initial stages of AD, patients are classified as having MCI, which is a transitional phase between normal cognitive aging and AD. Patients with MCI usually present objective cognitive impairment since they show signs of memory loss and confusion, but they still have relatively intact functional abilities [10]. Even though MCI is considered a transitional state in AD's progression, only 30-40% of people with MCI develop AD within 5 years (these are called MCIc - MCI converters). There are also MCI patients who never develop AD (these are called MCInc - MCI non-converters) [1]. Moreover, MCI patients can also be divided into two other different categories: early-stage Mild Cognitive Impairment (eMCI) and late-stage Mild Cognitive Impairment (lMCI) [10]. The eMCI subjects represent individuals with milder degrees of cognitive impairment (according to cognitive tests) than the lMCI subjects, and their rate of progression is slower [11]. Figure 2.1 shows a typical chronology of the stages that patients who develop AD go through: they start at an healthy state, i.e., Normal Control (NC), then move to a transitional state, i.e., MCIc, and finally get diagnosed with AD.

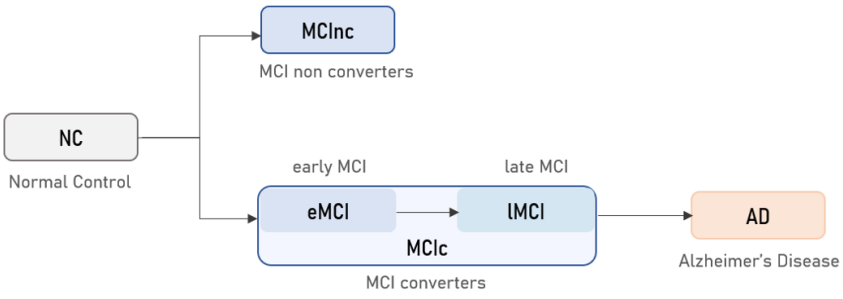

Figure 2.1: Evolution of AD, from NC, i.e., healthy patient, to a transitional state, i.e., MCI to finally being diagnosed with AD.

2.2 Biomarkers

A biomarker is a measurable indicator of some biological state or condition. Unlike symptoms, which are subjective, biomarkers provide an objective way to characterize a disease. Consequently, they can be useful in every step of patient care, by improving the accuracy of diagnosis, monitoring the disease evolution or measuring its severity [7]. Moreover, biomarkers may also be applied to drug development, where they help to assess the effectiveness of the treatment under development [12].

Biomarkers that reflect fundamental features of AD pathophysiology, which allow to differentiate it from other closely related diseases, are presented below and summarized in Table 2.1 [7].

- **Amyloid-β peptide**. A failure to clear this peptide from the interstices of the brain, and consequently its accumulation in brain tissues, is considered an early sign of AD. Detection of the accumulation of amyloid-β can be done through the analysis of the Cerebrospinal Fluid (CSF) (since an higher deposition of the peptide in the brain is reflected less secretion of the peptide to the CSF) or by amyloid PET imaging [13, 14];
- **τ protein**. In AD, chemical changes, i.e., hyper-phosphorylation, cause τ to detach from microtubules and aggregate to other τ molecules, forming toxic tangles inside neurons [15]. The deposition of this particles causes neurodegeneration, which is initially characterized by synaptic dysfunction, followed by the progressive loss of structure or function of neurons [16]. Neurodegeneration can be quantified by 18F-Fluorodeoxyglucose-PET (FDG-PET), which measures synaptic dysfunction and neuronal activity, or by τ protein levels in CSF [14];
- **Brain structure**. In dementia conditions, such as AD, the structure of the brain can be altered. Common alterations are brain atrophy, which is the loss of brain cells, and the destruction of synapses, which allow the neurons to communicate. The loss of brain matter mostly occurs in the medial temporal lobe and can measured by structural Magnetic Resonance Imaging (MRI) [13, 14];
- **Memory and Clinical function**. The progressive damage to brain cells caused by AD causes memory loss, confusion and cognitive decline. The loss of memory and the clinical function can be measured by cognitive tests [14].

Table 2.1: AD's most common biomarkers, the consequences to patients of their changes due to AD and the respective measuring methods.

Biomarker	**Consequence to patient**	**Measuring methods**
Amyloid-β peptide	Amyloid-β-plaques formation	CSF
		Amyloid-PET
τ protein	Neurodegeneration	CSF
		FDG-PET
Brain structure	Brain atrophy	MRI
Memory	Memory loss	Cognitive tests
Clinical function	Cognitive decline	Cognitive tests

Biomarkers may exist before clinical symptoms arise. The amyloid-β peptide, τ protein and brain stucture are biomarkers whose changes cause formation of amyloid-β-containing plaques, neurodegeneration and brain atrophy, respectively. These changes can occur in the early stages of the disease, ergo, can be observed prior to a dementia, i.e., AD, diagnosis. On the other hand, changes in memory and clinical function are the classic indicators of later dementia stages (Figure 2.2) [14].

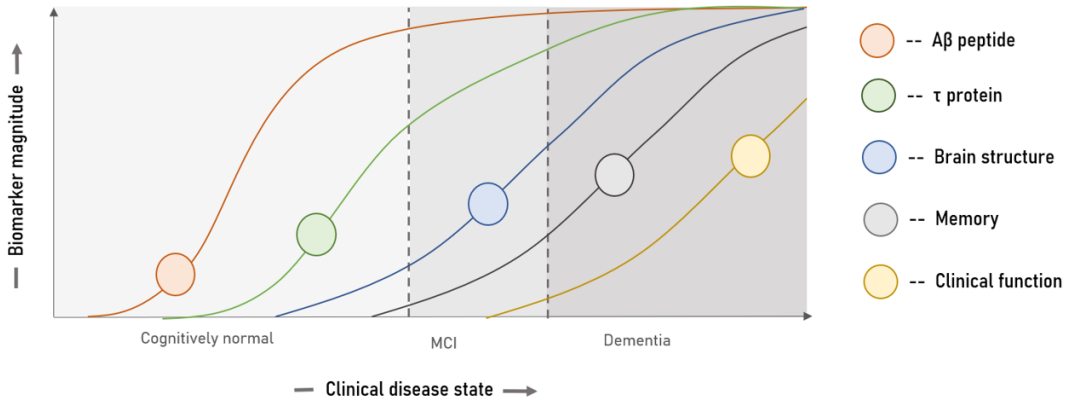

Figure 2.2: Biomarkers magnitude evolution with respect to the clinical disease state [14]

The appropriate monitoring of AD's biomarkers can result in earlier diagnosis and better patient care. By measuring the decline in neurogenesis in the hippocampus, the changes on the ventricles of the brain or by assessing the scores of cognitive tests, an appropriate and timely diagnosis can be done and the evolution of the disease for each individual patient can be followed [17]. Also, the research for specific AD biomarkers will improve the ability to differentiate AD from non-AD dementias and MCI, allowing to better detect and monitor prodromal stages of AD. Therefore, precise, easy to perform and reliable methods to measure these biomarkers are fundamental [7]. Some of these methods are further detailed in sections 2.2.1, 2.2.2 and 2.2.3.

2.2.1 Medical imaging techniques

Medical imaging techniques allow visual representations of the interior of the human body and have an effective role in revealing how the pathology of AD influences the brain. By doing so, they aid the radiologists and physicians to detect, diagnose, or treat diseases earlier and more efficiently [18, 19]. Different imaging modalities usually reflect different temporal and spatial scales information of the brain and because of that they are used for different purposes. For example, for AD, MRI technology is used to detect atrophy of the temporal lobes' medial structures (i.e., hippocampus and entorhinal cortex), while the Diffusion Tensor Imaging (DTI) technique is useful to measure the white matter damage, assessing the disruption in its nerve fibers, by measuring the fiber tract integrity. Moreover, PET technology is considered an appropriate tool for detecting alterations on brain function, since it reflects brain conditions at a molecular and cellular level [19].

2.2.1.1 MRI

Structural MRI has been extensively used to identify brain changes in normal aging, MCI, AD, and other dementias. MRI is useful for AD diagnosis, for measuring treatments that slow progression of neurodegeneration in AD and for ruling out other causes of dementia [20]. The progression of the disease can be seen in MRI scans, from NC to AD, as brain structure changes: ventricle enlargement, increase of medial temporal lobe atrophy and occurrence of white matter lesions (Figure 2.3).

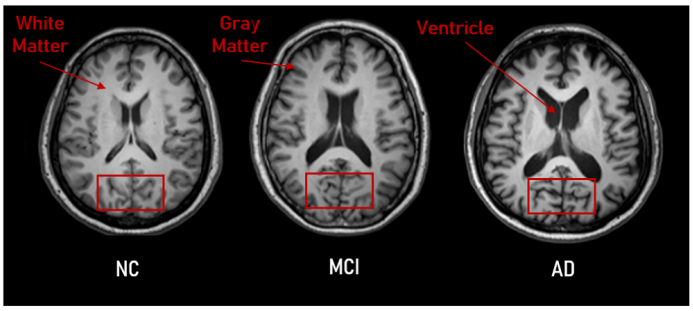

Figure 2.3: MRI scans for patients NC (left), MCI (middle) and AD (right) [21]. The red box identifies the medial temporal lobe, and the white matter, gray matter and ventricles are also identified by the red arrows.

2.2.1.2 PET

PET is an imaging technique that provides information about physiological and biochemical processes of the body.

The 18F-Fluorodeoxyglucose (FDG) is a glucose analogous molecule and the most used PET tracer in the study of AD. In FDG-PET measurements, patients with AD have characteristic reductions in regional brain activity (temporoparietal hypometabolism), which are progressive and correlate with dementia severity (Figure 2.4) [20].

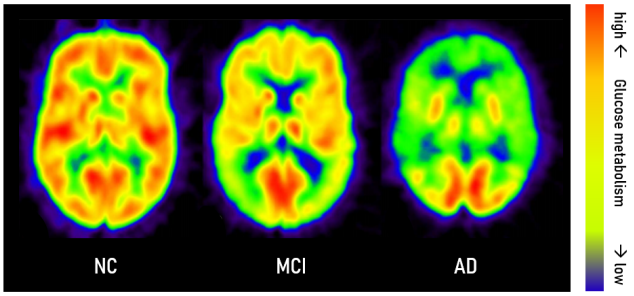

Figure 2.4: PET scans for patients with NC (left), MCI (middle) and AD (right). The colors represent the metabolic rate of glucose: high (red) to low (green). The temporoparietal hypometabolism, i.e, reduction of the metabolic rate of glucose, is evident from the left to the right. [22]

Besides using the glucose metabolism for measuring the loss of brain function, other traces, such as amyloid-β peptide and τ protein homologous, are also commonly used in PET scans as a way to

measure the respective biomarkers accumulation. Unlike FDG-PET, the activity in amyloid-PET or τ-PET scans increases as the patient progressively evolves from NC, to MCI, to AD [22].

2.2.2 Cerebrospinal fluid analysis

The amyloid-β peptide and the τ protein have proven diagnostic accuracy for mild cognitive impairment and dementia due to AD. To study them, a proteomic analysis of CSF must be performed [23]. The CSF fluid analysis is currently most used for improving the distinction between AD and other types of dementia (non-AD), increasing the percentage of appropriately diagnosed patients [24].

Although the amyloid-β peptide and the τ protein, present in the CSF, provide relatively high sensitivity and specificity for early disease detection, they are not suitable for monitoring disease progression [23]. This is sustained by the fact that, despite the fact that these biomarkers represent the earliest detectable changes in the AD course, they have already plateaued by the MCI stage (Figure 2.2), so they can not offer discriminate information about advanced stages of the disease. However, CSF biomarker tests are still very useful, not only for early diagnosis of AD, but also for efficient design of drug intervention clinical trials [25].

2.2.3 Cognitive tests

The measurement of cognition is also a valuable step for distinguishing the early stages of dementia and AD. Some of the most commonly used tests to assess AD-related cognitive decline are: the Alzheimer's Disease Assessment Scale (ADAS), which is frequently used in pharmaceutical trials, the Mini-Mental State Examination (MMSE), which is frequently used by clinicians and researchers interested in cognitive aging, and the Clinical Dementia Rating (CDR), which is commonly used in clinical trials and clinical practice for rating severity, including in early stages of disease [26]. However, the use of different measures across different research centers and studies can make it difficult to compare data across patients or studies.

2.2.3.1 CDR

The CDR is a global rating instrument used to characterize cognitive and functional performance. The CDR score is calculated on the basis of testing six different cognitive and behavioral domains: memory, orientation, judgment and problem solving, community affairs, home and hobbies performance, and personal care. The CDR is based on a scale of 0–5, presented in Figure 2.5 [7], which reflects the degree of Cognitive Impairment (CI).

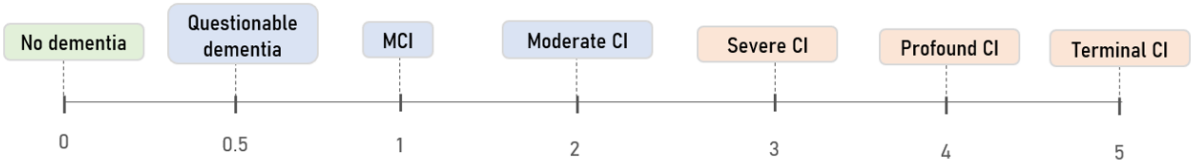

Figure 2.5: CDR is based on a scale of 0–5: no dementia (CDR=0), questionable dementia (CDR=0.5), MCI (CDR=1), moderate CI (CDR=2), severe CI (CDR=3), profound CI (CDR=4) and terminal CI (CDR=5). [7].

2.2.3.2 MMSE

The MMSE is a 30-point test used to measure thinking ability or "cognitive impairment". It is also used to estimate the severity and progression of cognitive impairment and to follow the course of cognitive changes in an individual over time, thus making it an effective way to document an individual's response to treatment. The scores and the corresponding level of dementia, according to Chopra et al. [27], are presented in Figure 2.6.

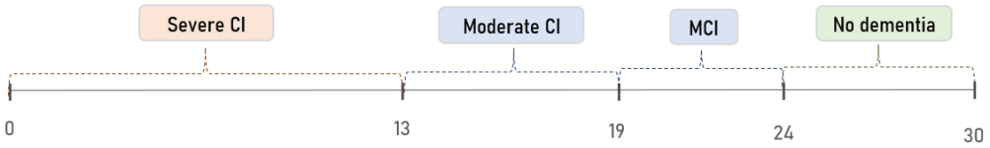

Figure 2.6: MMSE test score indicates the level of dementia: no dementia (MMSE $>$ 24), MCI (19 $<$ MMSE $<$ 23), moderate cognitive impairment (13 $<$ MMSE $<$ 18), and severe cognitive impairment (MMSE $<$ 12) [27].

2.3 Diagnosis

A 'ground truth' diagnosis of AD can only be made by autopsy, which is not clinically helpful [1]. Yet, the use of biomarkers can be very useful when making a diagnosis. An accurate diagnosis of AD, especially at the early stage, is decisive for the treatment of the disease to be relevant. Early diagnosis includes recognition of the pre-demented conditions, before clinical symptoms develop, allowing to identify those who would benefit from therapeutic intervention. This kind of diagnosis plays a significant role in patient care, since patients, by getting the appropriate treatment earlier in the course of the disease, delay the development of symptoms and can maintain their independence longer. Moreover, earlier diagnosis can be extremely helpful to signal patients to clinical trials, which is a crucial step for cure development [18].

At first, the diagnosis of AD was only focused on classifying AD from NC, which is not enough, since numerous times it is already too late for treatment for those patients diagnosed with AD or severe

dementia. Consequently, in recent years, as the importance of early diagnosis gained more relevance, the diagnosis of the disease was not only focused on classifying AD from NC, but also on distinguishing AD and NC from MCI. Furthermore, the ultimate goal is to improve the classification of MCI and to predict if a patient with MCI will develop AD, i.e., classify MCIc from MCInc [1].

Currently, MRI is the most used neuroimaging technique for AD detection [1]. MRI scans provide detailed information about the anatomical structures of the brain, which can help detect and measure brain atrophy patterns in AD. The volumetric and shape analysis of the hippocampus are important for AD diagnosis. However, this is still a challenging task due to hippocampus's irregular shape and blurred boundary in MRI scans. Also, using this region alone may not be sufficient for discriminating MCI from NC subjects. Other regions adjacent to the hippocampus, such as the parahippocampus and amygdala, are also affected in early stages of AD and should be taken in consideration [28]. Moreover, the use of one biomarker alone might not be enough for an early diagnosis of AD, mainly since some of the changes that begin to occur in the brain in the early stages of cognitive decline are detected by different biomarkers than brain structure (figure 2.2). For example, the use of PET images combined with MRI scans might be appropriated for this purpose, since the functional changes detected by PET manifest before the structural changes detected by MRI. Also, the combination of the analysis of multiple biomarkers, such as clinical function, CSF biomarkers and neuroimaging biomarkers, can improve the accuracy in the clinical diagnosis before the development of dementia [8]. Monitoring the decline in neurogenesis in the hippocampus, the changes in brain structure and the scores of cognitive tests together, can be used a diagnosis technique but also as a way to evaluate the progression of the disease [17].

In addition to monitoring biomarkers, there are also many genetic risk factors known for AD that can be helpful when making an early diagnosis, such as age, family history or the presence of the Apolipoprotein E4 (APOE) gene in a person's genome. The healthy allele of this gene encodes the Apolipoprotein E (APOE) enzyme, which is involved in the clearance of the amyloid-β peptide from the brain. The ApoE4 allele encodes a variant of that enzyme (structure or functional changes) that is not effective as the others at promoting the proteolytic break-down of this peptide, promoting its accumulation. Ergo, this allele is believed to be involved in the pathogenesis of AD [8].

Although growing progress has been made in understanding the natural history of AD, particularly the process of evolution of the disease and its risk factors, so far, the causes and mechanisms of AD are not yet fully understood and the cure is still unknown. Nevertheless, the increasing number of trials on drug candidates and the improvements on early diagnostic accuracy show great prospects for progress [8]. The earlier diagnosis will allow patients to have a higher quality of life during the course of the disease, will open doors for the better understanding of the disease mechanisms and, eventually, contribute to the ongoing search for treatments to slow or prevent this devastating disease.

Chapter 3

Deep Learning for Alzheimer's disease diagnosis

3.1 Machine learning in medical diagnosis

3.1.1 Relevance of machine learning in medical diagnosis

With new technologies arising, the amount and diversity of patient data acquired over the years has exponentially increased, leading to complex and heterogeneous health datasets, encompassing imaging data, bio-fluid data, genomics data and behavioural information [4]. The analysis and interpretation, by clinicians, of such datasets can be time consuming and easily influenced by the fatigue of human experts [5]. In response to these challenges, the application of machine learning algorithms to medicine and scientific research has been widely discussed [4].

Neuroimaging was one of the first areas of neurology to benefit from the application of machine learning approaches to improve diagnosis. More specifically, for the case of AD, the use of CAD systems has proven to improve diagnosis accuracy [4]. Medical images, usually 3D images with high resolution, are the most widely used data for AD diagnosis, but also the most complex to analyze, since they contain complex patterns and enormous amount of information [1]. Ergo, to make an accurate diagnosis, medical doctors have to be able to perceive distinctive patterns in such images, allowing them to distinguish between NC, MCI and AD patients. However, to analyse thousands of images and learn to discriminate such patterns is extremely laborious, requiring a lot of practise and time, which most of clinicians do not have, even the most experienced ones [1, 18]. Consequently, CAD systems arose as a way of overcoming the difficulties in the interpretation of medical images. Some of these systems are able to automatically extract informative features that describe the inherent patterns from data and can play a vital role in medical image analysis, since they can assist doctors to faster diagnose and predict the risk of diseases, preventing them in time [18].

3.1.2 Machine Learning

Machine learning is a branch of artificial intelligence which focuses on the use of data and algorithms to imitate the learning process that comes naturally to humans, based on the idea that systems can learn from experience, identify patterns and make decisions without prominent human intervention [29].

In Chollet [30], François Chollet considers that a "machine-learning system is trained rather than explicitly programmed" and describes its learning process as a system which is "presented with many examples relevant to a task, and it finds statistical structure in these examples that eventually allows the system to come up with rules for automating the task" [30].

Machine learning methods are categorized into supervised, unsupervised and reinforcement learning approaches. On the one hand, supervised algorithms require a labelled dataset from which to learn and are subdivided into classification and regression algorithms. Classification algorithms predict the categorical output (for example, the diagnostic category) for each data sample and regression algorithms predict a real-valued variable (for example, the degree of functional impairment) for each data sample [4]. On the other hand, unsupervised learning algorithms use datasets containing features and learn characteristics of the structure of the dataset, such as clustering data samples into groups, or reducing the dimensionality of datasets by generating a simpler representation of highly complex data, without using explicitly-provided labels [31]. In reinforcement learning approaches a reward or punishment is assigned to achieve a desired output.

Supervised machine learning algorithms are currently the most commonly applied to neurodegenerative disease-related data [4]. Moreover, a subfield of machine learning, deep learning, has achieved recent success for medical image classification.

3.1.3 Deep learning

Deep learning is a specific subfield of machine learning. The term deep learning implies the use of deep neural network models, which utilize artificial neural networks (ANN) with more than one hidden layer to carry out the process of machine learning. An ANN tries to mimic the structure and operations of biological neural networks found in the human brain. The typical structure of an ANN, shown in Figure 3.1, consists of layers of interconnected neurons, also called nodes or units. The first layer is the input layer and it is connected to the neurons of the next layer, i.e., a hidden layer, and so on until it reaches the last layer, the output layer. Each connection between two neurons is called an edge and is associated with a numeric number called weight. As exemplified in Figure 3.2, each neuron takes the values from all connected neurons (x) multiplied by the respective edge's weight (w), adds them, and feeds the result into an activation function (F) that may be nonlinear. Furthermore, to adjust the output along with this weighted sum of the inputs, a bias (b) can be added, which allows to shift the activation function by a constant amount. An ANN model learns a function that maps inputs to the desired outputs by adjusting the trainable parameters of the network (weights and biases) to minimize the observed errors between the output of the network and a target output, following an optimization algorithm such as gradient descent [32].

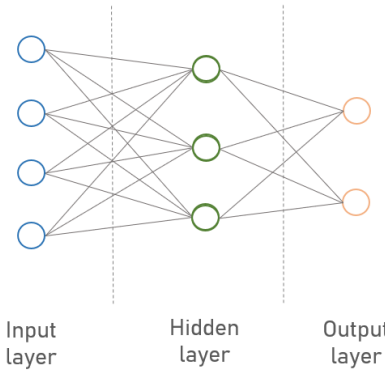

Figure 3.1: Basic architecture of an ANN.

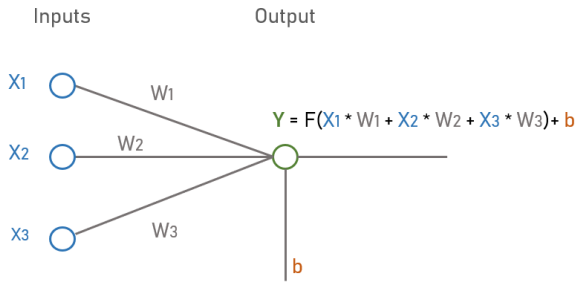

Figure 3.2: Representation of the computation of the value of a node/neuron (Y).

Like all machine learning models, the goal of deep learning is to learn useful representations of the input data that get us closer to the expected output. However, while other approaches of machine learning tend to use only one or two layers of representations of the data, i.e., hidden layers, the process of deep learning relies on learning successive layers of increasingly meaningful representations. Therefore, the “deep” in deep learning stands for this idea of successive hidden layers, in which the depth of the model reflects how many layers are present [30].

Deep neural networks have proven its potential for different classification problems. The use of a particular type of deep neural network, known as a convolutional neural network (CNN), has led to significant performance improvements for image classification [5].

3.1.4 Convolutional Neural Networks

CNNs, like all deep neural networks, try to mimic the biological process of neural networks found in the human brain, only they are inspired by the visual cortex of the brain. Hence, CNNs rapidly became very popular in image-based applications and the most successful deep model for image analysis [1].

These networks have been designed to better utilize spatial information by taking two dimensional (2D)/ three dimensional (3D) images as input and merging the feature extraction and classification tasks. Other major advantage is that they reduce the number of trainable parameters by parameter sharing, in the covolutional layers, forcing a filter used on a single 2D plane to share its weights with all filters used across the same plane [1, 30]. Additionally, the term convolutional in CNNs comes from the fact that these networks have at least one convolutional layer. The convolution layer, where the convolutional operation between the input and the kernel occurs, is responsible for the feature extraction task [30].

Nevertheless, the need of large datasets for the networks to train on can be considered a weakness of these models.

3.1.4.1 Architecture

The architecture of a CNN, shown in Figure 3.3, consists of an input layer, that should receive image data represented by a two/three dimensional matrix, hidden layers and an output layer, which outputs the

predicted label. The hidden layers are made up of several convolutional layers stacked with pooling layers, followed by fully-connected layers. The output layer consists of the last fully-connected layer, with a softmax/sigmoid activation function. The first layers work as feature extractors, extracting discriminative features and the last layers allow task-specific classification using those same features [1].

Convolutional layers Convolutional layer is the first and fundamental hidden layer, which convolves the input image with the learned filters, producing appropriate feature maps, and passes its result to the next layer. It is usually followed by applying a nonlinear activation function such as a Rectified Linear Unit (Rectified Linear Unit (ReLU)) to make all negative value to zero, enabling models to learn more complex representations faster and better.

Pooling layers The pooling layer is used between two convolution layers and is responsible for down-sampling the input feature map, by replacing each non-overlapping block with its maximum or average. This allows to reduce the number of parameters, features and hence, computation costs. After a succession of convolution and pooling layers, the 2D/3D feature maps are flattened into a 1D feature vector that no longer has spatial coordinates.

Fully-connected layers Fully-connected (FC) layers, like the name states, connect all feature elements in the previous layer to the next layer, which is helpful in learning non-linear relationships between the local features. Ergo, they perform like traditional neural networks and contain typically about 90% of the parameters in a CNN.

Softmax layers Softmax or Logistic layer is the last fully-connected layer of CNN, whose name depends on the activation function they use. Logistic is used for binary classification problems and softmax is for multi-classification. The softmax function highlights the largest values in a vector (in a classification problem, those values that represent higher probability of being the right class) while suppressing those that are significantly below the maximum.

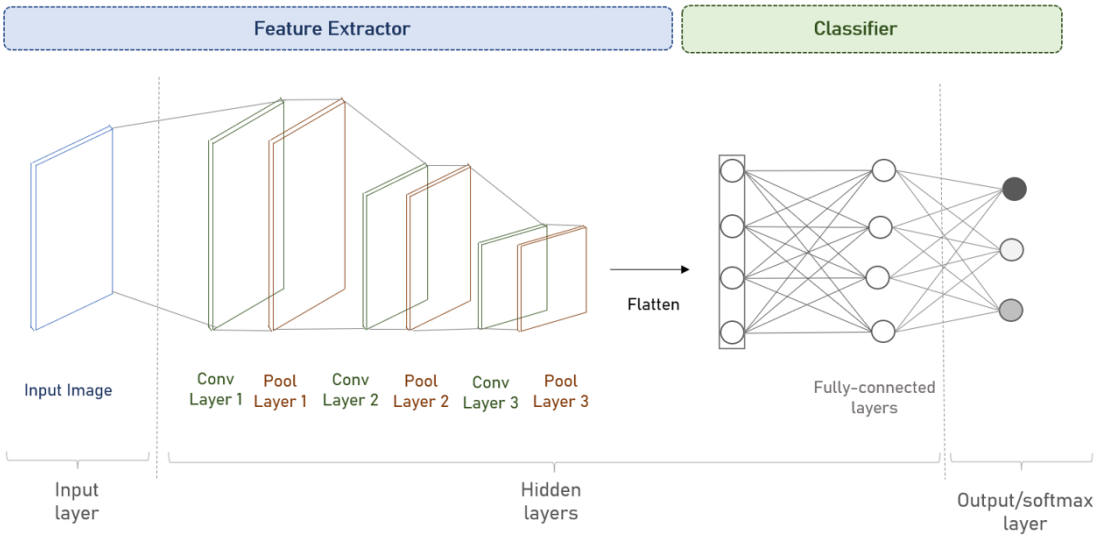

Figure 3.3: Convolutional neural network architecture.

The number of layers of each type, the presence of dropout or/and batch normalization are considered architecture hyperparameters, which can be changed and fine tuned for better classification performance [33].

3.1.4.2 Training

CNNs, just like other supervised machine learning methods, learn their parameters (weights and biases) from a labelled training set. For example, for a image classification task, the training set is composed of images and the respective label, i.e., class, which correspond to the ground truth.

The training procedure corresponds to the tuning of the model's parameters in order to minimize a **loss function**, using an **optimizer** to solve the optimization problem. The loss function measures the difference/error between the network's output and the ground truth data and is the error measure that will be minimized during training. The optimizer is the method that determines how the model will be updated. It takes the loss value, and using the backpropagation algorithm to calculate the gradient of the loss function with respect to the parameters, goes backward (from the last layers to the first layers), and applies the chain rule to compute each parameter's contribution in the loss. Thereby, the model's parameters are adjusted according to the optimizer method, until some stopping criterion is verified [1, 34].

For multi-class classification problems the multi-class cross entropy function is the most commonly used loss function. Regarding the optimizers, some of the most popular are the Stochastic Gradient Descent (Stochastic Gradient Descent (SDG)), the SGD with momentum, the Adaptive Gradient (AdaGrad), the Adaptive Delta (AdaDelta), and the ADAptive Moment Estimation (ADAM) [34]. The ADAM optimizer is the most commonly used in deep learning, since usually allows the network to achieve the smallest training loss in comparison with other optimizers, in the same number of epochs [34].

Furthermore, according to the number of training samples available and their complexity, there are three types of training modes can be defined: ***mini batch***, ***batch*** and ***online/real time***. Mini match is used when the whole training set is too big or too complex to be loaded at once. In this training mode a subset of the training set is loaded at each iteration and the size of each mini batch is called batch size. If the complete training set corresponds to N examples and the batch size is M, then $\frac{N}{M}$ mini batches are formed, i.e., $\frac{N}{M}$ iterations are performed, for each epoch, untill the whole training set is used. The batch mode corresponds to the mini-batch mode when the batch size is set to be the size of the complete training dataset. For each epoch only one iteration with the complete dataset is performed. Moroever, the online/real training mode uses one example in each iteration, matching the number of iterations per epoch to the size of the training set.

Similarly to the architecture hyperparameters, which determine the structure of the network, there are also training hyperparameters that control the behavior of the learning model. Some training hyperparameters that can be adjusted to optimize the learning efficiency of the network are the learning rate, which defines how quickly a network updates its parameters, the weight decay, the number of epochs and the batch size [33]. All these hyperparameters, as well as architecture hyperparameters, are summarized in table 3.1 and further detailed.

Table 3.1: Architecture and training hyperparameters and respective function.

Type	Hyper-parameter	Function/Description
Architecture	Number of layers	*Adjust the model to data complexity: the more complex the data is, the more hidden layers the model should have.*
	Dropout	*Regularization technique to avoid overfitting (increase the validation accuracy) thus increasing the generalizing power.*
	Batch normalization	*Normalize each layer's inputs by using the mean and variance of the values in the current mini-batch.*
	L1/L2 regularization	*Add a penalty term as the model complexity increases, decreasing the importance given to higher terms (avoids very large weights).*
	Activation function	*Introduce nonlinearity to models, which allows deep learning models to learn nonlinear prediction boundaries.*
Training	Parameter initialization	*Define how the parameters, such as weights and biases, are initialized: random initialization or Xavier initialization.*
	Learning rate	*Defines how quickly a network updates its weigths and bias.*
	Momentum	*Specify the amount of old weight change, which is added to the current one, helping to prevent oscillations.*
	Number of epochs	*Number of times the whole training data is shown to the network, responsible for the improvement of the validation accuracy.*
	Batch size	*Number of samples in each batch, which influences the traning and validation accuracy, typical values are 32, 64, 128 and 256.*

3.1.4.3 State of the art

CNNs were first introduced in 1989 by LeCun and colleagues [35]. Although they had immediate success, they have not been widely employed to medical image classification tasks until recently, when wellknown and proven structures have emerged, such as LeNet [36], AlexNet [37], CaffeNet [38], VGGNet [39], GoogLeNet [40], ResNet [41] and DenseNet [42]. Currently, for AD detetion, the main competitor architectures are 3D CNNs and 2D CNNs (with or without recurrent neural networks (RRNs), a type of artificial neural network which recognize data's sequential characteristics) [1]. In this section, related studies using 2D/3D CNNs are discussed, where some prefer to design their own customized architecture, while others use variants of the popular ones. Table 3.2 summarizes information of those studies, as well as their architecture specifications.

3D CNNs

Since neuroimaging techniques provide 3D images, 3D CNNs became popular for AD detection. However, they are usually complex and associated with a large number of parameters, which combined with small sized datasets might result in overfitting [1]. Multiple AD studies use their own architectures, which can differ much on the number of convolutional layers used, their number of filters and activation function. Basaia et al. [45] used twelve layers and Spasov et al. [46] used seven, both for distinguishing NC from AD and MCI subjects. Bäckström et al. [43] achieved an effective 3D architecture by using five convolutional layers for feature extraction, followed by three fully-connected layers for AD/NC classification. Moreover, Esmaeilzadeh et al. [47] trained a 3D CNN with three convolutional layers on two classes (AD vs. NC) then fine-tuned the model to classify the subjects into three categories, whereas Choi et al. [50] also used a 3D CNN with three convolutional layers, only for discrimination between MCIc

Table 3.2: Summary of several works using CNNs, with MRI or PET images, for AD diagnosis and the respective architecture details.

CNN	Application	Study	Architecture details	
			Design	Combine RNN?
3D CNN	NC *vs* AD	[43]	Customized - 5 conv. layers	☐
		[44]	Based on LeNet	☐
	NC *vs* MCI *vs* AD	[45]	Customized - 12 conv. layers	☐
		[46]	Customized - 7 conv. layers	☐
		[47]	Customized - 3 conv. layers	☐
		[48]	Based on ResNet	☐
		[49]	Based on VGGNet	☐
	MCIc *vs* MCInc	[50]	Customized - 3 conv. layers	☐
2D CNN	NC *vs* AD	[51]	Customized - 3 conv. layers	☐
		[52]	Customized - 3 conv. layers	☐
		[53]	Based on Inception-V3	☒
	NC *vs* MCI *vs* AD	[54]	Customized - 3 conv. layers	☐
		[55]	Customized - 5 conv. layers	☐
		[56]	Customized - 5 conv. layers	☒
	NC *vs* eMCI *vs* lMCI *vs* AD	[57]	Based on GoogleNet	☐

and MCInc. Furthermore, focusing on well-known 3D architectures, Karasawa et al. [48] proposed an effective novel 3DCNN architecture, based on ResNet. Cheng and Liu [44] used a 3D CNN structure inspired by LeNet with four convolutional layers for each image patch. Moreover, Tang et al. [49] built a 3D CNN based on VGGNet, with alleviates gradient vanishing by merging low-level and high-level feature information.

2D CNNs

2D CNNs were the first type of CNNs, which are specifically designed to recognize patterns in two-dimensional images. Most of the studies that used 2D CNNs for 3D images either extract 2D information from the images by splitting volumetric data into image slices (without using RNNs) [51, 52, 54, 55] or they rely on the logic that a 3D image can be treated as a sequence of 2D images (resorting to the use of RNNs) [56, 58]. In the latter, they use RNNs to extract the inter-slice features (similar structures in adjacent slices) while the 2D CNN captures the intra-slice features (similar structures in a single slice) [1].

Regarding the studies that build their customized 2D CNN structure, the ones with three convolutional layers are most common and they have been employed, for example, by Taqi et al. [51], Qiao et al. [52] and Lin et al. [54]. However, five layers have also been used by Awate et al. [55] for classifying subjects as AD or NC. On the other hand, Kazemi and Houghten [57] demonstrated that well known 2D structures, such as AlexNet and GoogLeNet performed well on functional Magnetic Resonance Imaging

(fMRI) images for classifying different stages of AD. Moreover, studies which combine a CNN and an RNN have been increasingly gained relevance. In these studies, the hierarchical 2D CNNs are built to capture the intra-slice features while the gated recurrent unit of RNN is used to extract the inter-slice features for final classification. They have been successfully applied to AD detection by Cheng and Liu [56] and Liu et al. [58].

Some of these methods shown slightly lower accuracies than the ones obtained with 3D CNNs. However, there are some architectures, such as the one proposed by Wegmayr and Haziza [53], a 2D deep model based on Inception-V3, where the model achieved the same accuracy as the 3D-CNN model trained from scratch, only that it trained much quicker because building a 3D CNN requires a larger number of parameters than 2D CNNs.

To conclude, although 2D CNNs are faster to train than 3D CNNs, since the latter have shown better performance results and recently the majority of the available medical images in medical datasets are 3D, they have been more widely employed in research studies for AD.

3.2 Medical image data

3.2.1 Data pre processing techniques

Datasets containing medical image data need to be processed before being used as input for CNNs. There are different methods to obtain images of the human body, such as MRI, fMRI, computerized tomography (CT) or PET, that vary between them. Moreover, even images obtained with the same method can vary a lot, for example, due to differences between patients or to the heterogeneity of hardware and software systems of the medical imaging equipment. To overcome these differences, data needs to be pre processed before being used as input. Actually, the success of a machine learning system using medical images is strongly dependent on effective pre processing techniques [1].

For raw MRI and PET images, which are the most used neuroimaging modalities for AD detection, the most common pre processing techniques performed, as mapped in Figure 3.4, are [1]:

- **Intensity normalization**: corresponds to mapping the intensity of all pixels onto a reference scale, ensuring that similar structures have similar intensities. For this purpose the most used techniques are the N3 algorithm [59], which reduces non-uniformity intensities, and the Gaussian filter, to reduce the noise.

- **Registration**: This technique consists on spatially aligning image scans to a reference anatomical space, i.e., a standard space. It is extremely important when dealing with neuroimaging data, since it allows to compare brain scans of different subjects. Moreover, it is used to co-register multiple modalities.

- **Skull stripping**: Consists in removing the bone of the skull from images.

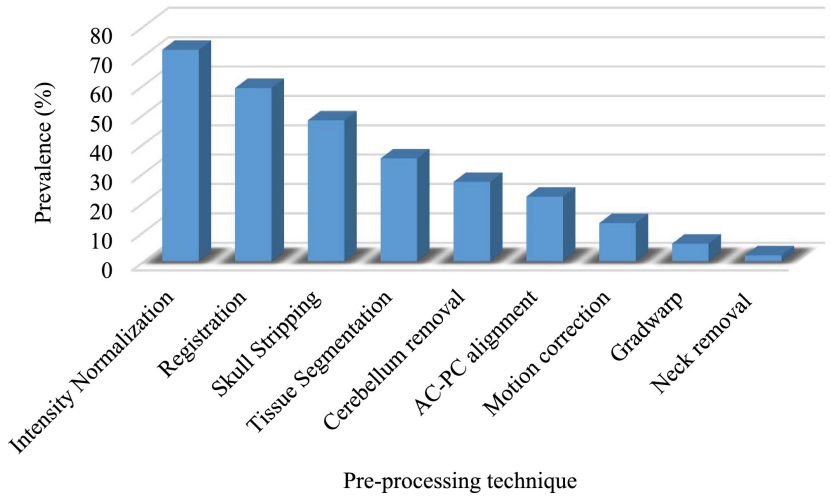

Figure 3.4: The prevalence of each pre processing technique, regarding 114 articles on deep learning for AD detection, according to Ebrahimighahnavieh et al. [1].

3.2.2 Data augmentation

Enormous progress has been made in using deep learning models for image classification, which have been successfully adapted to medical diagnostic tasks. However, the process of data collection for medical domain studies is often associated with high costs and complexity. This is why these studies are usually characterized by limited samples, i.e. small-sized medical datasets [60]. Furthermore, small datasets are often associated with overfitting problems, which occur when the model has too few samples to learn from, making it unable to generalize to new data [30].

One important preprocessing method that has been shown to be effective in training highly discriminative deep learning models and to mitigate the effect of overfitting is data augmentation [61].

Data augmentation is a strategy that consists of "generating more training data from existing training samples, by augmenting the samples via a number of random transformations that yield believable-looking images" [30]. This strategy significantly increases the diversity of data available for training models, without actually collecting new data. Models with data augmentation give better results because augmentation improves the testing accuracy and prevents overfitting problems, as more training data becomes available, i.e., increasing size of the datasets [62]. The most used data augmentation methods are the geometric transformations, such as rotation, horizontal and vertical flip and scaling (zoom in/out), depicted in Figure 3.5. However, some new methods have also been used recently, such as texture transfer and style transfer [63], where there is a merge of the content of one picture with the style of another one, resulting in a completely new picture with characteristics of both images. Additionally, a simple but effective data augmentation technique is the random erasing technique, where a noise-filled rectangle is painted in an image, resulting in changing original pixels values [63].

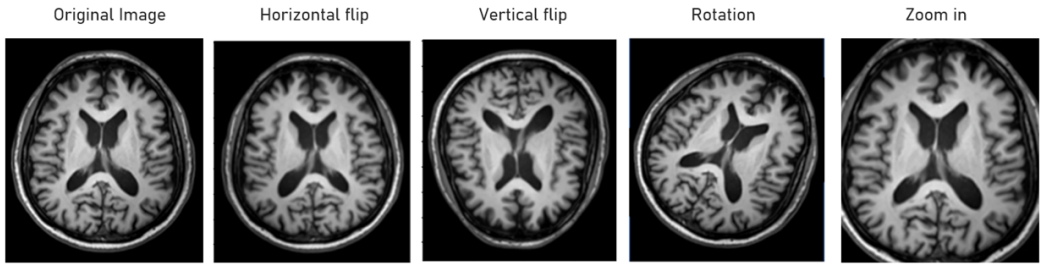

Figure 3.5: Representation of some geometric data augmentation techniques: horizontal flip, vertical flip, rotation and zoom in, from left to right.

3.2.3 Data leakage sources

When training models with medical images, the process of splitting the images into training, validation and test sets needs to be carefully performed and revised so data leakage can be avoided. Data leakage is an undesirable process whereby information is accidentally shared between the training data and the test data, resulting in test evaluation scores that are not representative of real-world unseen data.

For neuroimaging data, the use of test data in any part of the training process leads to bias in performance. For example, the use of images from the same subject in different sets influences the model's performance, since images from the same brain are too similar and are likely to be classified with the same label [33].

The most common causes for data leakage are: wrong data split, late data split and the absence of an independent test set. Wrong data split refers to data from the same subject appearing in several sets (test, training). Late data split occurs when some procedures, such as data augmentation, are performed before dividing the test and training set, which leads to versions of the original image to be treated as different ones and being found in both sets. Furthermore, there is an absence of an independent test set when images from the test set are not only used to test the performance of the model, but are also used in some steps of training or validation.

Chapter 4

Incorporating medical knowledge into CNNs

As mentioned above, the lack of data when training deep learning models can lead to overfiting problems, which are usually solved by regularization or data augmentation techniques. These solutions, even though they effectively improve model's performance, do not introduce any new information [3].

In recent years, introducing information beyond the one available in the datasets at hand has become a promising approach to address the problem of small-sized medical datasets, also improving model's performance and the reliability of the diagnosis. The integration of medical knowledge in deep learning models can span from creating network architectures that mimic how medical doctors are trained, to simulating their diagnostic patterns or paying attention to the regions doctors usually focus on [3].

Additionally, the incorporation of medical knowledge can also work has a way to avoid AD prediction from unrelated cases. This can be achieved by data processing, allowing us to only include data in our training/validation sets that are accurately labelled. In dataset selection, by using domain knowledge, we can exclusively include in the dataset images of AD patients whose symptoms were considered relevant for AD, i.e., patients with very specific AD symptoms, by the clinicians (resorting to information of medical reports, for example). Thereby, when training deep learning models for AD diagnosis, the combination of information derived from neuroimaging data and medical domain knowledge can result in a better combined method. [64, 65].

4.1 Sources of medical knowledge

Medical knowledge can be described as the information about diseases, interpretation of lab tests, etc., which is broadly applicable to decisions about patients. It can be extracted from medical datatsets, medical doctors or medical imaging reports, as detailed in sections 4.1.1, 4.1.2, 4.1.3, respectively. These sections present a detailed description and state-of-the-art review of sources of medical domain knowledge and a few of the more recent and successful methods for the incorporation of that specific knowledge into deep learning models. The methods are summarized in Figure 4.1 and the studies are further described and are also summarized in Table 4.1.

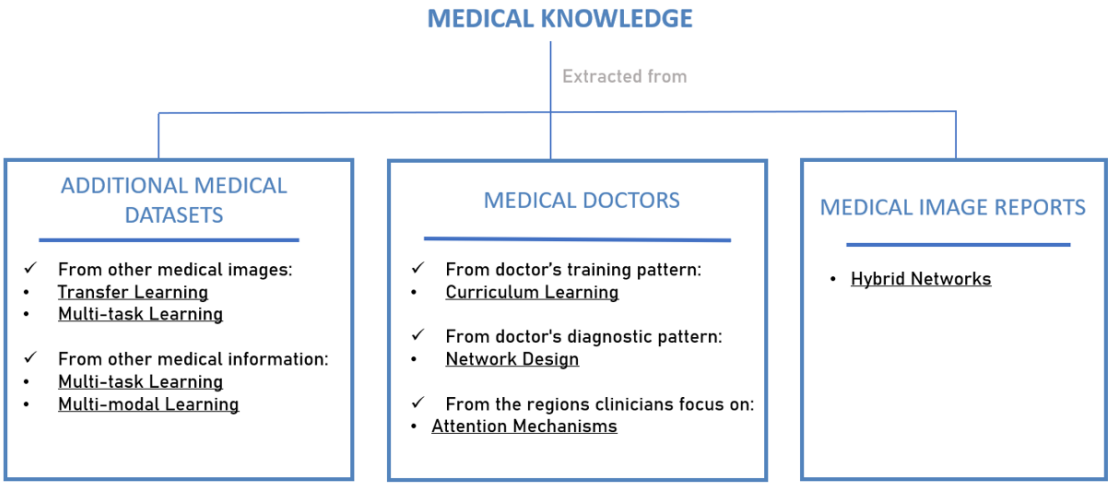

Figure 4.1: Sources of medical knowledge and methods used to incorporate it into CNNs [3].

Table 4.1: Studies incorporating medical knowledge into CNNs, their method of incorporation of such knowledge, the type of image data they used, as well as the area of application and the accuracy (ACC) results ($ACC_{w/}$ when the method is implemented and $ACC_{w/0}$ when it is not).

Method	**Study**	**Application**	**Data used**	**ACC results (%)**	
				$ACC_{w/o}$	$ACC_{w/}$
Transfer learning	[66]	Breast cancer	Mammography, MRI	90	93
	[67]	Prostate cancer	TeUS, B-mode US	72	73
Multi-task learning	[68]	Breast cancer	Mammography	76	78
	[69]	NC *vs* lMCI	fMRI, DTI	85.4	87.80
	[70]	AD *vs* MCI *vs* NC	MRI	40.4	51.2
Multi-modal learning	[71]	AD *vs* NC	MRI, Genetic, EHR	86	88
	[28]	AD *vs* NC	MRI, FDG-PET	87.8	90.15
Curriculum learning	[72]	Breast cancer	MRI, R0I	77	81
	[73]	Thoracic diseases	X-ray, Reports	77.1	80.3
Network design	[74]	Thoracic diseases	X-ray, Radiologist's pattern	84.2	87.1
	[75]	Skin lesions	Dermoscopic images, Pattern	87.4	90.1
Attention mechanism	[76]	Glaucoma Detection	Fundus images, ROI	92.2	95.3
	[77]	AD *vs* NC	MRI, Attention maps	87.8	92.1
	[78]	AD *vs* NC	MRI, Image patches	85.1	97.35
Hybrid networks	[79]	Thoracic diseases	X-ray, Reports	95.7	97.8
	[80]	Bladder cancer	MRI, Reports	84.9	88.6

4.1.1 Additional medical datasets

Medical datasets include a large amount of medical data such as Eletronic health records (EHR), various measurements and medical images. On the one hand, medical image datasets most of the times can be considered compatible, due to the similarities between medical images, allowing us to merge or combine different datasets together. On the other hand, they can also be considered complementary, since images obtained from different medical image modalities can provide complementary information, i.e., structural *vs* functional information. Consequently, in recent years, numerous methods have been developed to incorporate knowledge from different medical datasets into deep learning models, such as Transfer learning, Multi-task learning and Multi-modal learning.

- **Transfer learning**: Transfer learning is a quite popular research topic for image classification in machine learning. It focuses on storing knowledge gained while solving one task, for which we have a large amount of data, and applying it to a different but related one, for which we have a limited amount of data. In deep learning, transfer learning is based on firstly training a network and then copying its first n layers to the first n layers of a target network. As depicted in Figure 4.2, the top layers of the target network are then trained and the first layers can be fine-tuned to the new task or left frozen [81]. This approach has already been broadly used for introducing knowledge from natural images into medical image diagnosis. More recently, the use of images from different medical datasets instead of natural images has also proven to be advantageous [3]. This can be considered preferable since medical images resemble one another in a way that they do not with natural images but also because they can provide complementary information. Transfer learning has already been applied to improve the diagnosis accuracy of breast cancer, by Hadad et al. [66], and prostate cancer, by Azizi et al. [67]), for example. In both studies the models were pre-trained with datasets containing medical images, although from different imaging modalities, of the same regions of the target dataset, such as breast MRI images and mammography images

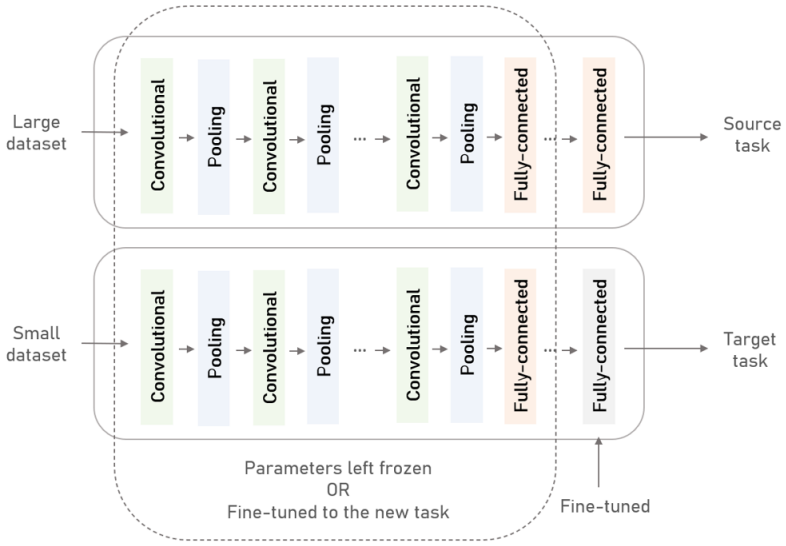

Figure 4.2: Transfer learning scheme.

for breast cancer and Temporal enhanced Ultrasound (TeUS) and B-mode Ultrasound (US) for prostate cancer. Moreover, transfer learning methods have also been proven to be robust for AD detection, by Hon and Khan [82], Maqsood et al. [83] and Ebrahimi-Ghahnavieh et al. [84], where the networks are mainly trained using natural images and then fine-tuned using medical datasets.

- **Multi-task learning** Multi-task learning is also a sub field of machine learning in which multiple learning tasks are solved at the same time, by a shared model, while exploiting commonalities and differences across tasks, as schematized in Figure 4.3. By using different medical datasets with different tasks or different tasks for the same dataset, we can find hidden representations among them and enhance both classification performances [3]. Multi-task learning is a well-planned approach to incorporate medical knowledge of one dataset to another, particularly when training samples from a single modality are limited, which is the case for the majority of the available medical datasets. Recently, Liao et al. [85] proposed a multi-task transfer learning approach for training a deep convolutional neural network for the diagnosis of twelve different types of cancer using multi-task learning. For AD there is also some recent work developed, for example, by Lei et al. [69] and Liu et al. [70] that efficiently used multi-task learning to improve AD diagnosis performance. The former used a multi-task learning model to select discriminative and informative features for fine MCI analysis, allowing to discriminate between different sub-stages of MCI. The latter used MRI data and demographic information of subjects from different medical datasets to buitl a multi-task learning framework for simultaneous brain disease classification and clinical score regression.

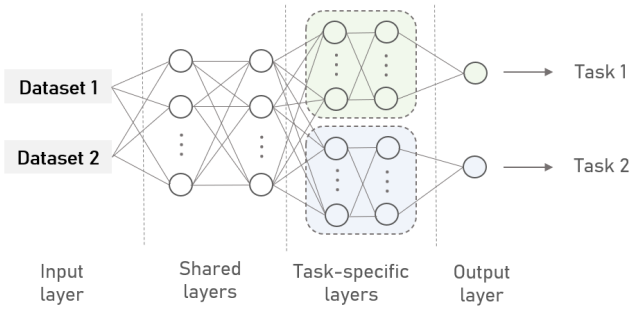

Figure 4.3: Multi-task learning scheme for a dual-task model.

- **Multi-modal learning** The information present in medical datasets is vastly heterogeneous, including data in text, number, image and video format. Even data in the same format, such as images, come from different imaging modalities and hence are associated with very different statistical properties. Multi-modal learning aims to build models that can process and relate information from multiple modalities (Figure 4.4) [86]. Consequently, the deep learning models using this framework are more prone to achieve superior performances due to this ability to extract relationships amongst features from different modalities [71]. Lately, this type of models have been drawing more attention in medical disease classification problems as more and more datasets of different

image modalities became publicly available. For example, Venugopalan et al. [71] combined MRI, PET, biological markers and clinical and cognitive assessments to measure the progression of MCI (early AD). In a similar study, Liu et al. [28] developed a strategy for data fusion to extract complementary information from MRI and PET data modalities. The results show that a performance gain was achieved in both binary classification and multi class classification of AD.

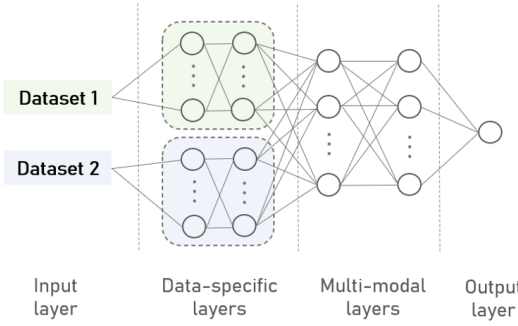

Figure 4.4: Multi-modal learning scheme for a dual-modality model.

4.1.2 Medical doctors

Medical doctors have extensive knowledge about disease patterns, disease mechanisms and disease evolution, not only due to the many years of studying they have gone through but also due to the practical knowledge they acquired practising medicine. However, this knowledge may be difficult to incorporate into CNNs since experienced medical doctors usually combine different types of knowledge in different stages of the diagnosis process, which is difficult to implement in a neural network [3]. The following subsections refer to different types of knowledge from medical doctors, that can be incorporated into deep learning models to improve their diagnostic performance, as well as their incorporation methods.

4.1.2.1 Training pattern

The training process of medical doctors follows the basic learning principle of humans and animals, which starts with learning easier aspects of a task, and then gradually taking more complex examples into consideration. It is possible to incorporate the training pattern of medical doctors in the training process of deep neural networks via curriculum learning [87].

- **Curriculum learning** In curriculum learning, as detailed in in section 4.2, a curriculum is defined detailing the order in which the samples should be used by the networks during train.

4.1.2.2 Diagnostic patterns

Experienced medical doctors and other clinicians follow specific patterns when performing medical tasks, such as preparing patients for examination, taking medical histories, measuring vital signs and reading medical images.

For the evaluation of medical images of different medical specialties, the corresponding clinicians follow their own diagnostic patterns. For instance, radiologists, when reading chest X-ray images, start by examining the whole image. Then, they focus on local lesion areas, analysing their shape, texture, and local characteristics, and afterwards combine the global and local information [88]. Furthermore, dermatologists, when classifying skin lesions from dermoscopic images, follow a different approach: first, they focus on differentiating dermis from epidermis. Then, they go beyond this basic segmentation and find more specific features, such as detecting changes in the epidermis and the presence of immune and nucleated cells, to classify a specific lesion [89]. Incorporating the diagnostic pattern that clinicians follow when reading medical images into the architecture design of deep learning networks has been used as a novel approach to improve their diagnosis performance [3]. One way of doing this is via network design customization.

- **Network design** For a CNN to automatically and efficiently learn the intrinsic image features that are most suitable for the classification purpose, it is crucial that the network design is adjusted to the classification task [90]. As explained before, the choices when customizing the network architecture include deciding the number, the type and how connected the network layers are. Guan et al. [74] proposed a customized network design which consisted of a three-branch CNN to mimic the reading pattern of radiologists when analysing X-ray images. Specifically, the first branch was used for training with global images and the second one, the local branch, for training with discriminative regions from the global image (cropped using the attention heat map automatically generated from the global branch). In the end, the last pooling layers of both the global and local branches are concatenated for fine-tuning of the fusion branch, i.e., the third branch. With a similar purpose, Gonzalez-Diaz [75] presents DermaKNet, a CNN architecture that incorporates subnetworks modeling taks such as lesion-skin segmentation and detection of dermoscopic features. The results showed that the incorporation of these subnetworks not only improved the model's performance, but also improved the interpretability of the diagnosis.

4.1.2.3 Regions clinicians focus on

As previously stated, clinicians follow their specific diagnostic patterns when reading medical images. Furthermore, their vision always focuses on selective parts of the medical images.The information about the regions that clinicians focus on provides valuable information about which brain areas are particularly related to the diagnosis of a disease [77, 78]. Therefore, the integration, in deep learning models, of which regions medical doctors focus on when reading medical images, has proven to achieve higher performance compared to the traditional CNN models [3]. One of the methods by which this can be done is the attention mechanism.

- **Attention mechanism** The attention mechanism was first introduced by Bahdanau et al. [91], in 2015, in natural language processing (NLP), but it was quickly adapted for other different applications, such as computer vision and speech processing. As neural networks try to mimic the way the human brain analyzes and processes information in a simplified manner, attention mechanism

also attempt to implement, in deep neural networks, the human action of selectively focusing on a few relevant things, while ignoring others. For the interpretation of medical images, the knowledge of the regions which greater attention should be given to, is represented by attention maps. Attention maps are usually learnt by the network, but they can also be derived from the areas doctors focus on when reading images. The information to design these areas can be collected from medical annotations or by eye tracking. Attention maps generally are represented as a grid of numbers that indicate what locations of a image are important for a given task. For example, Li et al. [76], in 2019, purposed an attention-guided CNN (AG-CNN) for glaucoma detection. In this approach, the attention maps of ophthalmologists were collected through eye-tracking and implemented in the structure of the network. The experiment results showed that the proposed AG-CNN approach significantly improves the performance of CNN-based glaucoma detection. For AD diagnosis, Jin et al. [77] proposed a novel attention mechanism approach that not only improved the classification performance, but also worked as a biomarker explorer, capturing significant brain regions for AD classification. More recently, in 2021, Zhang et al. [78] also proposed an attention mechanism for AD classification that achieved a performance among the top ranks and improved in discriminating MCI subjects (MCIc and MCnc).

4.1.3 Medical imaging reports

Medical imaging reports, also referred as diagnostic imaging reports, reflect the knowledge of experts, since they contain the interpretation of medical image data and clinical findings, provided by a specialist, such as a radiologist [3, 92]. Incorporating this knowledge into CNNs designed for disease diagnosis has been considered as a rising approach [93]. As medical reports are generally handled by RNNs [94] and medical images by CNNs, the incorporation of information from medical reports can be done via hybrid networks, containing both CNNs and RNNs. This differentiates from typical Multi-modal Learning architectures since they do not imply the use of RNNs.

- **Hybrid networks** Hybrid deep neural networks were built to support mixed inputs. Their architecture is an aggregation of multiple networks, with reflects in good flexibility and wide applicability [95]. For the diagnosis of brain diseases, Vatian et al. [93] used a novel framework for fusing medical images and the corresponding reports, which yield better results than the models without fusion of textual conclusions of radiologists. Wang et al. [79] proposed the Text-Image embedding network (TieNet), consisting of a CNN-RNN architecture, to classify the common thorax disease in chest X-rays. TieNet, by using both image features and text embeddings extracted from associated reports, improved classification results (with a 6% increase on average of the Area under the curve (AUC) compared with the baseline CNN purely based on medical images. Moreover, Zhang et al. [80] introduced a novel neural network, named TandemNet. It consists of a dual-attention model that allows interactions between images and semantic information. The presented results show that incorporating information from diagnostic reports significantly improves the cancer diagnostic performance over the baseline method.

4.2 Curriculum learning

One way of incorporating medical knowledge into deep learning models is to train them by mimicking the training pattern of medical doctors. During medical school, medical students train by learning concepts and tasks with increasing difficulty [87]. Usually they start with the easiest tasks, such as identifying lesions in medical images, and they gradually move to more challenging ones, such as determining if a lesion is malignant or benign [73].

In curriculum learning strategies, a curriculum is designed, which defines the order in which the data are presented to the model, instead of being randomly presented. It has been an active research topic for computer vision and, more recently, it has been used to improve the training of deep networks, specifically CNNs trained for image recognition [96]. Usually, the curriculum is predefined (manual strategies). However, since defining a good curriculum manually is not an easy task, some strategies rely on learning the curriculum from the data, simultaneously with network training (automatic strategies).

- **Manual strategies.** Frameworks in which the curriculum is defined *a priori*, i.e, before training the model. The data is fed into the model, according to the previously defined curriculum. They were further subdivided into two different groups:
  - **Complexity focused** Strategies focused on progressively training the network with more difficult or complex tasks and/or samples.
  - **ROI focused** Strategies focused on gradually training with more information of each sample/image, i.e., wider or different ROIs.
- **Automatic strategies.** These strategies are the ones where the whole dataset is fed into the model at once and the curriculum is automatically generated by the network itself, depending how it follows through the training data, so as to maximise the learning efficiency [97]. They were also subdivided into two different groups:
  - **Self-paced learning** In these strategies, the student, i.e., the network, is able to control the amount of information it consumes and the duration of time they need to learn the information properly. They do not allow for the incorporation of extra medical knowledge.
  - **Self-paced curriculum learning** Strategies in which the student, i.e., the neural network, takes into account both prior knowledge and its learning progress during training, to automatically choose sub-tasks or samples from a given set for it to train on.

Curriculum learning has recently shown to improve the performance of CNNs for several medical image classification tasks [72, 73, 98]. They have been used for classification of breast cancer malignancy, thoracic lesions, bone fracture, histological images and medical image segmentation. Despite the recent success of curriculum learning strategies for medical image classification, they have still not been applied to networks for AD diagnosis.

The following sections revise the use of these strategies for medical image classification problems, which are also summarized in Table 4.2.

Table 4.2: Most recent implementations of curriculum learning strategies for disease classification. SENS: Sensitivity, TPR: True Positive Rate.

Curriculum strategy		Classification of	Study (year)	Data	Data purpose	Results (%)	
						w/o CL	w/ CL (%)
Manual	Complexity focused	Thoracic disease	[73] (2018)	Chest X-ray	Disease classification	$AUC = 77.08$	$AUC = 80.27$
				Radiology reports	Define the curriculum		
		Femur fracture	[99] (2019)	X-ray	Fracture classification	$F_1 - score = 81.71$	$F_1 - score = 86.57$
				Experts annotations	Define the curriculum		
		Histological image	[100] (2020)	Histological images	Polyp classification	$AUC = 83.7$	$AUC = 88.2$
				Experts annotations	Define the curriculum		
	ROI focused	Breast cancer	[72] (2019)	MRI	Cancer classification	$AUC = 50$	$AUC = 89$
				Lesion patches	Pre train the model		
			[101] (2017)	Mammography	Cancer classification	$AUC = 65$	$AUC = 92$
				Lesion masks	Pre train the model		
		Thoracic disease	[102] (2019)	Chest C-ray	Disease classification	$ACC = 96.1$	$ACC = 97.2$
				Patch images	Pre train the model		
			[103] (2017)	Chest CT	Lung nodule detection	$SENS = 79.4$	$SENS = 88.3$
				Nodule patches	Pre train the model		
Automatic	SPL	Breast cancer	[104] (2020)	Histological images	Cancer classification	$ACC = 91.42$	$ACC = 92.87$
				Loss values	Define the curriculum		
			[98] (2019)	MRI	Cancer classification	$AUC = 86$	$AUC = 90$
				Loss values	Define the curriculum		
	SPCL	Thoracic disease	[105] (2018)	CT images	Nodule segmentation	$TPR = 74$	$TPR = 95$
				Loss values + Annotations	Define the curriculum		
		Histological image	[106] (2020)	Fundus images	Glaucoma detection	$AUC = 99.08$	$AUC = 99.45$
				Evidence maps + Features	Define the curriculum		

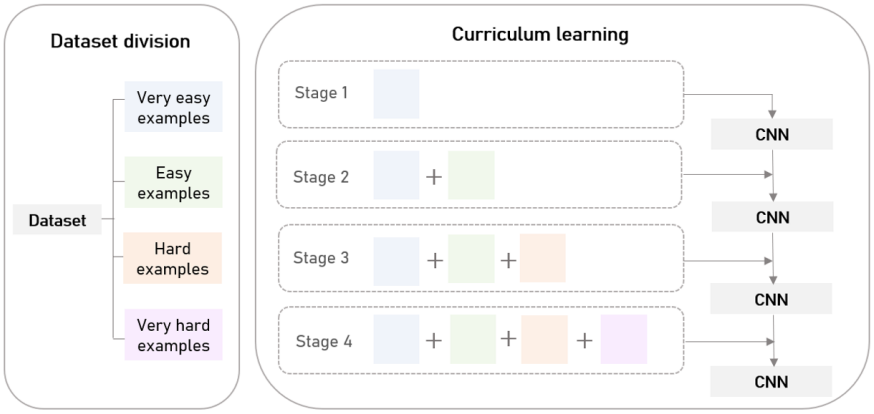

Figure 4.5: Complexity focused curriculum learning strategy: fedding a CNN with progressively more complex tasks or samples.

4.2.1 Manual curriculum strategies

4.2.1.1 Complexity focused

As depicted in Figure 4.5, complexity focused strategies consist on feeding the CNN with samples, i.e., medical images, ordered by difficulty. The easier tasks/samples are first used for training and then only more complex ones are added. Furthermore, it is common to keep previously introduced samples in the pool of training samples rather than replacing them with new ones [107].

An easy task can be considered training the network using only the easiest disease categories. An easy disease category is a state of a disease which its classification is considered straightforward, such as classifying a patient as healthy or with IV stage of cancer, whereas intermediate disease stages (such as stage I or II of cancer) are considered difficult disease categories and are only added later in the training process. Ergo, distinguish healthy subjects from stage IV cancer patients can considered an easy task and distinguishing subjects with different cancer stages is considered an hard task. Furthermore, an easy sample is a sample that clearly belongs to the disease category which is associated. For example,a mammography being classified as stage IV cancer by multiple different annotators, i.e., medical doctors, is an easy sample. Whereas a hard sample is when the class label attributed to that sample is not unanimous among all annotators [100].

In short, in complexity focused strategies, the network can be trained with gradually more complex tasks, with gradually more complex samples or even a combination of both, i.e., first training the network with the easier samples of the easier tasks and gradually adding more complex samples and tasks.

For identifying and distinguishing various chest abnormalities, an attention-guided curriculum learning strategy was used by Tang et al. [73]. In this framework, the disease severity level (such as mild, moderate and severe) was used to separate the data samples, of 14 thoracic disease categories (such as, among others, pneumonia and pneumothorax), by order of increasing difficulty. Then, for each disease category, the CNN matures and converges gradually by seeing samples from “easy” to “hard”. This training pattern mimics the way in which medical students learn how to read radiographs. Compared to the benchmark results, the presented framework achieved higher AUC scores for all disease categories

except hernia.

Similar to Tang et al. [73], Jiménez-Sánchez et al. [99] used a curriculum learning approach to improve the classification of proximal femur fractures. The X-ray samples were divided into different severity categories according to medical decision trees and annotations of multiple experts. The results achieved show that this medical knowledge-based curriculum learning performs better in terms of accuracy (up to 15%), achieving the accuracy of experienced trauma surgeons.

For histopathology image classification, Wei et al. [100] proposed a curriculum learning approach which achieved an AUC of 88.2%, an improvement of 4.5%, compared to baseline frameworks. Since medical image datasets are labeled by multiple annotators, the annotator agreement is used to map the difficulty of a given example: the greater the discrepancy of the labels of a given example between different annotators, the greater the difficulty of that training sample and, therefore, the later it should be added to train.

4.2.1.2 ROI focused

The strategies focused on progressively adding more complex regions/sections of the images were called ROI focused strategies. They consist on fine-tuning of the model on complete images after training on lesion-specified patches of those images, such as ROIs (Figure 4.6). The fine-tuning of the network using the whole image is an important step so that no information present on the image is disregarded. For example, classification of malignancy for breast cancer is usually tackled by first localizing individual lesions and then classifying them with respect to malignancy. However, if we would only focus on those individual lesion areas, and do not take all the image into account, other regions that are not labeled as lesions but contain global medically relevant information could often be disregarded. It is important to notice that, in these strategies, the complete image and the ROI correspond to the same image, i.e., same size, only the ROI has zero-valued pixels outside the region of interest.

Haarburger et al. [72] and Lotter et al. [101] used multi scale lesion-specific curriculum learning strategies applied to mammogram datasets. The deep learning models were trained in two stages. The first consists of a simpler task: training a classifier to estimate the probability of the presence of a lesion in a given image patch. The second stage involves aggregation across patches, consisting in a more

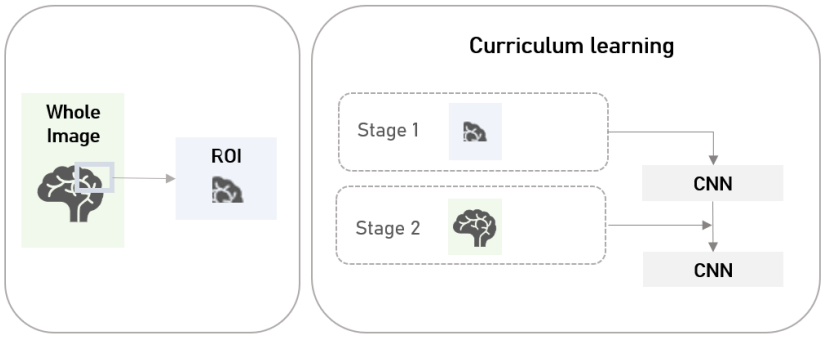

Figure 4.6: Curriculum learning strategy focused on feeding the CNNs with more complex sections of the images.

difficult task: image-level training. In both cases the use of curriculum learning enabled successful training of the network, improving the AUC from 0.50% to 0.89% and from 0.50% to 0.92%, respectively. Park et al. [102] also used a similar curriculum learning strategy, to improve the classification performance when training a Restnet-50 to classify various lesions in chest X-rays. In this approach, thoracic abnormalities were first identified, by learning of patch images around abnormal lesions, and then the resnet-50 was fine-tuned using the entire images. Furthermore, Jesson et al. [103] proposed a curriculum adaptive sampling approach, for lung nodule detection, that topped the LUNA16 nodule detection benchmark. In this framework, patches of growing sizes were continuously fed as input to the network, until the algorithm learnt how to distinguish lung nodules from their initial surroundings.

4.2.2 Automatic curriculum learning strategies

4.2.2.1 Self-paced learning

Self-paced learning (SPL) is an automatic curriculum lerning strategy where data are sorted while training, based on sample training loss, as depicted in Figure 4.7 [108]. A threshold, λ, is defined and the a self-paced function selects samples with loss below λ, which are considered easy, to be used as training samples in the next epoch. During training the threshold is updated, according to a growing factor, δ, from including only the lower loss samples in training, to including all samples in the final epochs. This strategy does not take prior medical knowledge into account [109].

Asare et al. [104] proposed a semisupervised learning framework that uses self-training with self-paced learning in the classification of breast cancer histopathological images. The approach consists of, among others, a selection algorithm for picking out training samples for retraining the model. When retraining the model, first relatively high confidence samples are chosen (“easy” samples), then gradually “hard” samples are added to the training data. This strategy prevents retraining the model with noisy samples, i.e., higher loss samples, in the beginning, avoiding mistake reinforcement. The effectiveness of the proposed method was also demonstrated by the authors. For the same purpose, Maicas et al. [98] proposed a novel training approach, inspired by how radiologists are trained, for breast screening

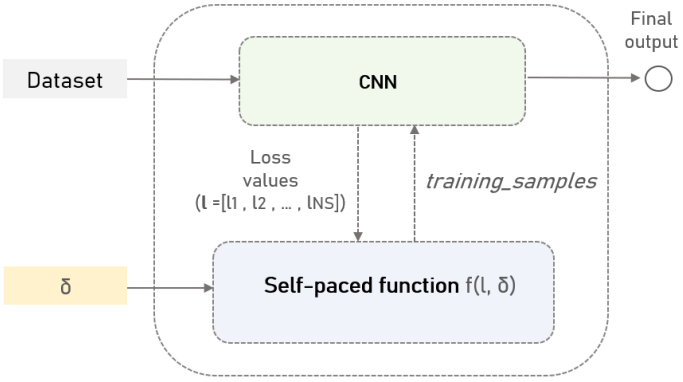

Figure 4.7: Self-paced learning scheme where the self-paced function is represented in blue, which takes as input the training losses of the samples, l and the growing factor, δ, and returns the training curriculum for the next train.

classification, using dynamic contrast enhanced MRI images. The network, instead of randomly selecting tasks to train on, it samples tasks that can achieve a higher improvement on their performance. The classification performance achieved by this approach was shown better, when compared to state of art baseline approaches: DenseNet, multiple instance learning and multi-task learning.

Furthermore, to deal with very large images such as 3D CT scans, Berger et al. [110] proposed an adaptive sampling algorithm, named *isample*, that not only improved the performance of multi-organ CT segmentation, but also increased the speed of the training process. The adaptive sampling algorithm uses a-posterior error maps, generated throughout training, to focus sampling on difficult regions. The networks is encouraged to train the tasks for which it shows more difficulty, resulting in improved learning. This sampling approach improved the accuracy of segmentation for aorta, lung and abdomen segmentation, when compared to training networks without *isample*.

4.2.2.2 Self-paced curriculum learning

Self-paced curriculum learning (SPCL) is a semi-automatic curriculum learning approach which results from the merge of manual curriculum learning strategies with SPL. On the one hand, in manual curriculum learning, the curriculum is predetermined by prior knowledge, and remains fixed thereafter. This type of method is called an “instructor-driven” approach, since it heavily relies on the quality of prior knowledge while ignoring feedback the learner’s feedback. On the other hand, in SPL, the curriculum is dynamically determined to adjust to the learning pace of the learner. However, SPL is unable to deal with prior knowledge and is therefore considered a “student-driven” approach.

In SPCL strategies, the model takes as input a predetermined curriculum, where the prior knowledge is encoded. During training, just like in SPL, the SPCL algorithm takes the model’s feedback, such as the training loss of the samples, and uses this information to iteratively update the training curriculum, in each epoch. It is considered an “instructor-student-collaborative” learning mode, as opposed to “instructor-driven” or “student-driven” [111].

Jiang et al. [111] has empirically shown the advantage of SPCL on two tasks: matrix factorization and multimedia event detection. In the clinical field, Wang et al. [105] and Rongchang and Li [106] have also proven the efficiency of SPCL strategies. Wang et al. [105] developed a novel Deep Active Self-paced Learning (DASL) strategy, for pulmonary nodule segmentation, with the purpose of dealing with unannotated samples. This strategy is based on a combination of active learning and self-paced learning frameworks. The latter consists in gradually incorporating easy-to-hard samples into training. In this study, to classify unannotated samples as easy or hard samples, the algorithm considers both prior knowledge, about those samples, and the learning progress made during training. The success of this approach was proven by the results achieved: DASL strategy performs much better than the model trained without DASL using the same amount of annotated samples. Moreover, Rongchang and Li [106] used an adaptive dual-curriculum learning framework for glaucoma diagnosis to overcome the training bias that comes from the normal-abnormal class imbalance and from the presence of rare but significant images. The dual-curriculum reflects the predetermined curriculum, which contains information about the medical complexity of data samples, and the update of that curriculum, through the contribution of

evidence maps generated by the model itself. This framework improves the convergence speed of the training process and obtains the better performance compared to benchmark procedures.

4.2.3 Comparison between manual and automatic strategies

On the one hand, manual curriculum learning strategies are the most used for incorporating curriculum learning in the classification of medical images. They have proven to efficiently improve the accuracy of such classifications. However, they still present a major disadvantage when compared to the automatic ones: the need to design or choose the appropriate curriculum for the model to train with. Sometimes, even for teachers, it is hard to choose the best way to present information for their students to learn. The same problem arises when building the curriculum for curriculum learning approaches. Even though they mainly focus on simply dividing tasks or examples into 'easy' and 'hard' or dividing images into important sections or less important sections, gathering and incorporating the necessary medical knowledge to make that division is not always as straightforward as it seems. In addition, the predefined curriculum stays fixed during the training process, which can be seen as lack of flexibility, since these strategies ignore, to some extent, the feedback of the network that is being trained [109]. Regarding the ROI focused strategies, they can provide an added advantage when compared to other methods, such as attention mechanisms. They allow the network to focus on regions that are most important for the classification task at hand, similarly to attention mechanisms, but they do it without disregarding global image information that can also be medically relevant (by fine-tuning the network with the complete images).

On the other hand, the use of automatic curriculum learning strategies in deep learning has promising applications. Although they have become a cornerstone of recent successes in deep reinforcement learning [112], they have also been recently used for classification tasks [97, 113, 114], as shown above, and yield similar or better results than the benchmark approaches. Their major disadvantage comes from the fact that they can be laborious to implement.

Regarding SPL, the fact that these strategies do not incorporate external knowledge into the neural networks, lowers their robustness, with contrast to manual strategies and SPCL. Nevertheless, as stated before, they are considered advantageous since they are more dynamic (consider model feedback) when compared to manual strategies and they avoid the step of carefully defining curriculum, which usually is the most important and time consuming step of manual CL approaches and SPCL [109].

Regarding SPCL, it represents a general learning paradigm that combines the advantages from both manual curriculum learning and SPL. Not only it inherits and further generalizes the theory of SPL, but also it complements it, by introducing a flexible way to incorporate prior knowledge into the networks [111].

Chapter 5

Methodology

In this thesis multiple manual and automatic CL strategies will be implemented in CNNs to improve the classification of AD, particularly that of MCI. Most of the strategies will incorporate medical knowledge, such as cognitive test scores and ROIs, in the process of building the curriculum. In order to implement them, first the data had to be collected and processed: image data that corresponds to the object of classification of the CNNs and data regarding cognitive test scores and ROIs, which will allow to define te curriculum, as detailed in section 5.1. Furthermore, details about how the CL was implemented and the baseline methods used are further presented in sections 5.3 and 5.4, respectively.

5.1 Data selection and processing

All data used in the presented experiments were extracted from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. ADNI is a global research study that actively supports the investigation and development of treatments that slow or stop the progression of AD. Progressively, as more data became available, the extent to which the ADNI data was used evolved, and is currently divided into four phases, according to the overall objectives of each phase: ADNI1, ADNI GO, ADNI2 and ADNI3.

ADNI1 primary goal was to develop biomarkers as outcome measures for clinical trials, ADNI GO focus on earlier stages of disease, ADNI2 intends to develop biomarkers as predictors of cognitive decline and ADNI3 studies the use of functional imaging techniques in clinical trials. The data used in this study was the data available for ADNI1 phase, which only includes data from NC, MCI and AD participants. It comprises:

- **Clinical data**: information about each subject, including recruitment, demographics, physical examinations, and cognitive assessment data (cognitive test scores);
- **Imaging data**: MRI and FDG-PET images;
- **Genetic data**: Genotyping and DNA sequencing of all participants;
- **Biospecimen data**: includes blood, urine, and cerebrospinal fluid (CSF) values.

Since the available dataset did not allow to separate MCI subjects into MCIc and MCInc or eMCI and lMCI, the classification task in this dissertation was focused on distinguishing between 3 classes only: NC, MCI and AD. The dataset contained the information described above (clinical and demographic information) for each brain scan, as well as information about the time it was acquired.

5.1.1 Imaging data

The neuroimaging data selected corresponds to FDG-PET brain scans. All scans were normalized, averaged and co-registered by ADNI researchers. The volumes were further normalized in the range of [0,1] and cropped from 60x128x128 to 40x98x98, in order to remove most of the area surrounding the brain, which does not include relevant information for the classification task.

The selected dataset comprises 1393 FDG-PET images from 406 different subjects. Each image is labeled as NC, MCI or AD and some subjects were followed over a 24 month period, contributing with brain scans from month 0, month 6, month 12 and month 24, while others left the study earlier. This information is further detailed in Table 5.1. Moreover, it was verified that all images from the same subject have the same label, i.e., the cognition level (normal, mild dementia or dementia).

Table 5.1: Demographic and clinical profile of the groups studied $(mean \pm standard\,deviation)$.

FDG-PET	NC	MCI	AD	Total
Sex (% M)	63.8	66.2	59.9	64.2
Age	76.9 ± 4.8	76 ± 7.3	76.5 ± 7.1	76.3 ± 7
MMSE	29.1 ± 1.1	26.6 ± 3.2	21.6 ± 4.4	26.1 ± 4.1
CDR	0.02 ± 0.2	0.5 ± 0.2	0.95 ± 0.5	0.47 ± 0.43
Month 0	102	207	95	404
Month 6	94	188	86	368
Month 12	85	177	74	336
Month 24	84	142	59	285
Total images	365	714	314	1393
Subjects	104	207	95	406

5.1.2 Regions of interest (ROIs)

For images, a region of interest is a subset of pixels that are considered to be informative for a specific purpose. For the FDG-PET dataset, ten ROIs for AD were delineated and provided by an experienced physician, Professor Dr. Durval Campos Costa. They were rearranged into 8 different ones: symetrical ROIs with respect to the vertical axis of the coronal section of the brain were merged into one (since AD is not related to a specific brain hemisphere) and also all ROIs were merged into one major ROI.

Table 5.2 summarizes the information about the ROIs provided (available ROIs) and the ones used in the project (selected ROIs).

Table 5.2: Available ROIs provided by Professor Dr. Durval Campos Costa, their name, percentage of brain area they occupy and the ROIs selected for this project.

Selected ROIs	Available ROIs	Name	Brain area (%)
1+2	1	Left lateral temporal	4.51
	2	Right lateral temporal	
3+4	3	Left mesial temporal	0.94
	4	Right measial temporal	
5	5	Inferior frontal gyrus/Orbitofrontal	0.84
6	6	Inferior anterior cingulate	0.71
7+8	7	Left dorsolateral parietal	2.66
	8	Right dorsolateral parietal	
9	9	Superior anterior cingulate	1.33
10	10	Posterior cingulate and precuneus	1.28
ROI_{ALLl}	All Rois	———————-	12.29

5.1.3 Cognitive test data

The scores of the cognitive tests were retrieved from the ADNI dataset. Out of the available tests, the CDR and the MMSE were the ones used in this project, which were previously explained in section 2.2.3.

The clinical profile of the groups studied was evaluated by the bar plots presented in Figures 5.1 and 5.2. They inform about the correlation between the label of the images (NC, MCI or AD) and the corresponding score of the cognitive tests. For example, according to the CDR test, a NC subject should present a CDR of zero [7]. However, when analysing Figure 5.1, it is verified that some of the brain scans that were considered as NC, i.e., no dementia, do not correspond to zero CDR values, but to score values of 0.5 instead (which are associated with questionable dementia in the CDR scale). In the scope of this project the former are the images considered as “easy” (where the label and the respective cognitive test score are in agreement). The former consist in “hard samples”, since the image label and the cognitive test score do not match.

Moreover, it was verified that, for a few number of images, the scores of the cognitive tests were not registered in the dataset, and are therefore presented as unknown.

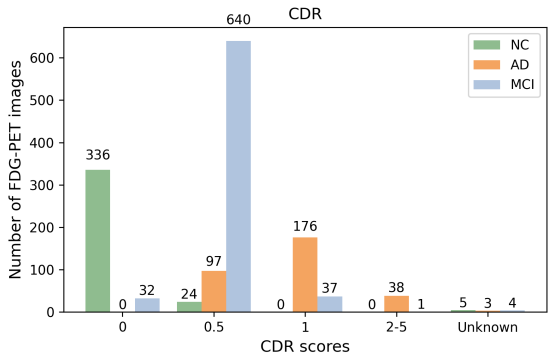

Figure 5.1: FDG-PET images grouped by the scores of the CDR test.

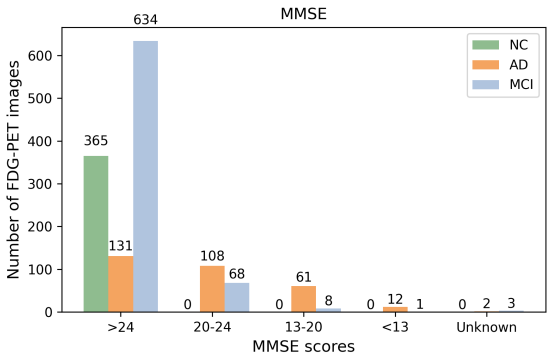

Figure 5.2: FDG-PET images grouped by the scores of the MMSE test.

5.2 Building and evaluating the deep learning model

5.2.1 Model Architecture

In order to chose the model architecture best suited for the selected dataset, different architectures were tested and compared: 3D-CNN, 2D-CNN with Long short-term memory (LSTM) and a 3D-ResNet. Although the Resnet architecture yielded better results, in terms of accuracy, the architecture chosen to be applied was the 3D-CNN, since its computational costs were significantly lower. The Resnet architecture has a total number of 46 225 539 trainable parameters, while the 3D-CNN has only 637 203, which makes it approximately 75 times less complex, and therefore faster at making the predictions. It was important to chose a lower complexity architecture since the use of curriculum learning strategies can bring additional training costs.

The 3D-CNN architecture selected, summarized in Table 5.3, corresponds to a basic CNN architecture which has already proven to be efficient by Pereira [115], for the classification of FDG-PET and MRI brain scans as AD and NC. Its architecture consists of three convolutional blocks where the 3D convolutional layer is composed of 8, 16 and 32 filters, respectively, with ReLU activation function. Each convolutional layer is followed by a 3D max-pooling layer and a batch normalization layer. The output of the last convolution blocks is then flattened and fed into a fully connected classifier network with 64 units and a softmax layer in the end, allowing the classification into 3 classes: NC, MCI and AD.

Table 5.3: Architecture of the 3D CNN model.

Layer Type	Parameters	Filters/Units
Convolutional	ReLU	3x3x3x8
Max Pooling	2x2x2, Stride-2	-
Batch Norm.	-	-
Convolutional	ReLU	3x3x3x8
Max Pooling	2x2x2, Stride-2	-
Batch Norm.	-	-
Convolutional	ReLU	3x3x3x8
Max Pooling	2x2x2, Stride-2	-
Batch Norm.	-	-
Flatten	-	-
Dense	-	64
Batch Norm.	-	-
Dense	-	64
Batch Norm.	-	-
Dense	Softmax	3

5.2.2 Training and testing the model

5.2.2.1 Training, validation and test sets

To train and evaluate the models a 5-fold cross-validation was performed. The percentage of each label in each fold remained representative of the original dataset, as it is displayed in Table 5.4.

In order to avoid data leakage, all subjects, instead of all images, were considered for the splitting step, to guarantee that brain scans from the same subject were not present in different folds. Although subjects are anonymous, they have a unique ID code. The 406 subjects were divided into five different folds and five different models were trained. As detailed in Figure 5.3, each model used one of those folds for testing (20% of the dataset) and the remaining four for training (80% of the dataset). For each train, the subjects in the training set were further divided into subjects for training the model (80% out of the subjects of the original training test) and subjects for the validation of the model (20% out of the subjects of the original training test). In the end, to convert the subjects into images, all images from the same subject were added to the corresponding set, originating the final training set (with 64% of the images), the final validation set (with 16% of the images) and the final test set (with 20% of the images).

Table 5.4: Information regarding each of the 5 folds generated to perform five-fold cross validation.

Data	NC (%)	MCI (%)	AD (%)
Original dataset	19.86	50.52	29.62
Fold 1	25.37	52.94	21.69
Fold 2	19.20	55.44	25.36
Fold 3	31.62	47.77	20.61
Fold 4	36.39	48.88	18.73
Fold 5	17.71	55.71	26.58

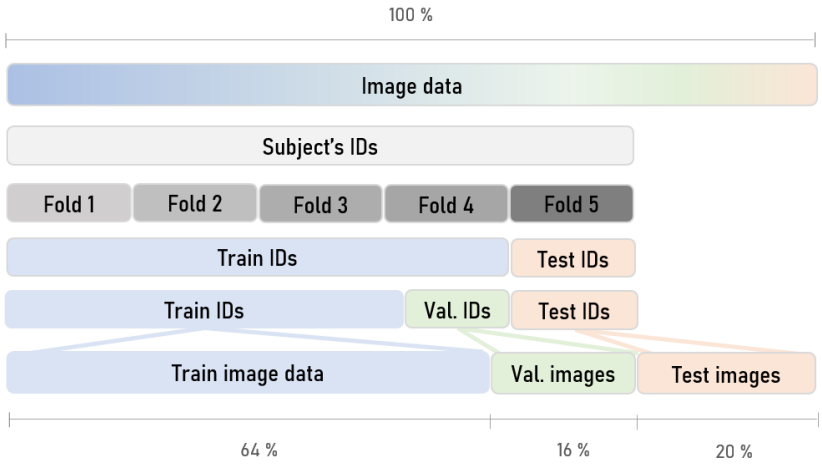

Figure 5.3: Representation of the data division into five folds and further division into training, validation and test sets for the model that used the fifth fold for testing.

5.2.2.2 Hyperparameters

The selected architecture hyperparameters are described in section 5.2.1.

Regarding the training hyperparameters, the networks were trained using the Adam optimizer with an initial learning rate of 0.001, β_1 = 0.9 and β_2 = 0.999. During training, the metric evaluated was the accuracy and the categorical cross-entropy was chosen as the loss function. Data generators were used for both training and testing, using the mini batch training mode with a batch size of 16, for a total number of 100 epochs, using an early stop criterion monitoring the validation loss with a patience of 50 epochs. Moreover, all weights were inicialized using the Glorot (also called Xavier) uniform initializer.

5.2.2.3 Class imbalance

In the FDG-PET dataset, described in Table 5.1, there is a clear class imbalance problem. The distribution of samples across the known classes is not symmetrical, there are approximately double of MCI labeled images (majority class) than the NC and AD labeled ones (minority classes). Moreover, during training in the CL strategies there is also a class imbalance always present, since the curriculum is built based on sample complexity and not based on the number of samples of each class. To solve this imbalanced classification problem, a weighted training strategy was applied during training,in which, for each round of train, the weight of each class was inversely proportional to the class frequency in the train set. Thereby, over-represented classes in the training set are penalized and under-represented ones are favoured.

5.2.3 Evaluating the model

For this multi class classification problem, multiple evaluation metrics were used. Accuracy, F1-score, AUC and model run time were used to evaluate the strategies implemented and to compare the respective results. Additionally, the p-value was also obtained, to assess the results' statistical relevance.

Accuracy: Accuracy is the closeness of the measurement results to a known true value. In this multi-class classification problem, accuracy corresponds to the percentage of predicted labels that match the true label.

$$Accuracy\,(\%) = \frac{Number\ of correctly\ classified\ images}{Number\ of\ images} * 100 \tag{5.1}$$

F1-score: F1-score rates how successful a classifier is and is a suitable measure for classification problems on imbalanced datasets. The F1-score (Equation 5.4) can be interpreted as a weighted average of the precision (Equation 5.2) and recall (Equation 5.3). In multi-class cases, the F1-score is the average of the F1-score of each class, weighted by support (the number of true instances for each label), which accounts for class imbalance.

$$Precision\,(class\,a = a) = \frac{TP(class\,a = a)}{TP(class\,a = a) + FP(class\,a = a)} \tag{5.2}$$

$$Recall\,(class\,a = a) = \frac{TP(class\,a = a)}{TP(classa = a) + FN(class\,a = a)} \tag{5.3}$$

$$F1\text{-}score\,(class\,a = a) = \frac{2 * Precision(class\,a = a) * Recall(class\,a = a)}{Precision(class\,a = a) + Recall(class\,a = a)} \tag{5.4}$$

AUC: The AUC measures the ability of a classifier to distinguish between classes, representing the degree of separability. AUC corresponds to the area under the Receiver Operating Characteristics (ROC) curve, which is a probability curve that plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold values. It is designed for binary classification problems, but it can be extended to multiclass classification problems. In our evaluation, the ”One vs All” technique was used: the weighted AUC score for each class against all other classes was computed, and afterwards all AUC scores were further averaged.

Model run time: The time it takes for a model to finish a complete train and test.

p-value: The p-value reflects the statistical significance of a measurement. A measurement is statistically significant if the p-value of the statistical test is small enough to reject its null hypothesis. In this project, the null hypothesis of the test is that the results of the CL strategies and those of the baseline methods come from the same distribution. The statistical test used was Wilcoxon test, which is an analogous of T-test for muticlass problems and the most common, and implemented, threshold value for p-value is 0.05 (p-value $<$ 0.05).

5.3 Incorporating curriculum learning

Curriculum learning strategies require the inputs to be gradually introduced into the neural networks. For this to happen, the curriculum must defined before the training of the model (in the case of manual strategies) or iteratively defined before each epoch as the model is trained (in the case of automatic strategies). Section 5.3.1 explains how CL was introduced into the CNNs (for manual strategies) and section 5.3.2 describes the proposed curriculum learning approaches and their implementation details.

5.3.1 How to use curriculum learning in deep learning models

The strategies used to incorporate curriculum learning into deep learning models rely on the idea that, instead of using a complete dataset at once as input, the data must be sequentially fed into the model. However, there are multiple different ways of implementing this idea of sequence.

The models were trained by gradually adding new data to the training set, yet maintaining the prior data, as a way to "avoid forgetting" the information learnt in the first place. In this project, each time new data are added to the training set, it accounts for a different round of training of the model. If the curriculum consists on having n different training sets, the model goes through n rounds of training. Moreover, after each round of training, the last fully connected layer of the model (the one that contains the information about the predicted label) is replaced by a randomly initialized one, while all weights an bias are maintained.

The model was trained using curriculum learning as schematized in Figure 5.4: each time new data are added to the training set, the last FC layer is replaced by a randomly internalized one, while other weights and biases are only updated. The architecture of the network and each of its layers is maintained throughout the complete training, where the data is sequentially added, by order of complexity, creating a continuously growing training dataset.

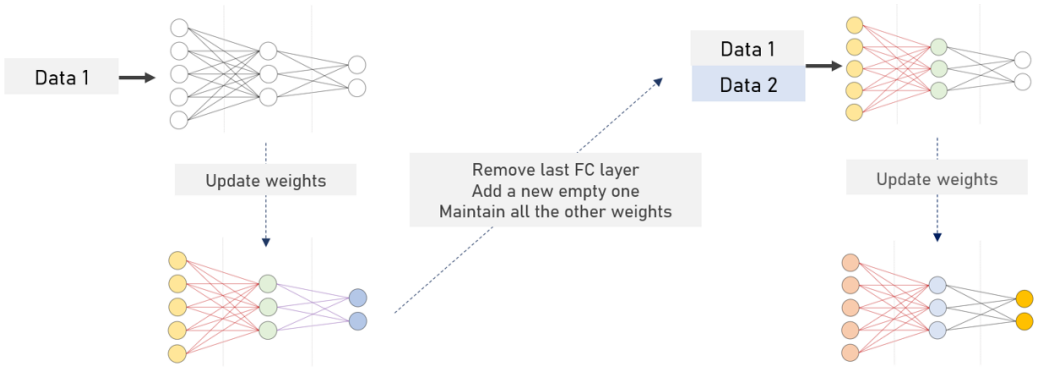

Figure 5.4: Representation of the method used for implementing curriculum learning: retraining the model with a growing dataset. The change of colors of the nodes and edges represents the values of the weights and biases being updated. White represent randomly initialized values and equal colors represent values being maintained.

5.3.2 Curriculum learning strategies

To improve early AD diagnosis from medical images, curriculum learning strategies were applied to CNNs. Different strategies were implemented, nine manual, three automatic, eight use medical knowledge to build the curriculum (such as cognitive test scores and ROI) and four do not. As schematized in Figure 5.5, the manual strategies are further subdivided into complexity focused strategies, ROI focused strategies, mixed strategies and replicate automatic strategies, while the automatic ones are subdivided into self-paced learning and self-paced curriculum learning. All these strategies differ either on how the curriculum is built or on the information they use to build it.

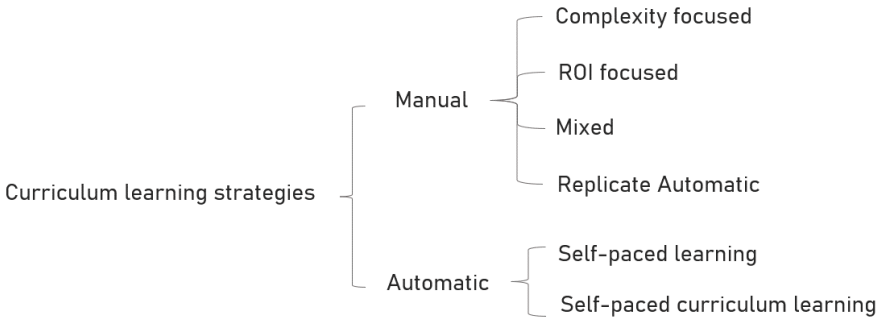

Figure 5.5: Different curriculum learning strategies performed.

5.3.2.1 Complexity focused

In complexity focused strategies the model is first trained with easier data (or tasks) and gradually more complex data (or tasks) are introduced. To achieve this, first the notions of easy sample, hard sample, easy task and hard task needed to be defined in the context of the problem. The MMSE and CDR scores were used for this purpose.

- **Easy sample**: an image was considered an easy sample if its label (NC, MCI and AD) and its corresponding MMSE or CDR score were in agreement. For example, according to the CDR scale (See figure 2.5), a score of zero is associated with no dementia, i.e, NC. Therefore, all images labeled as NC with a CDR score of zero are considered easy samples.
- **Hard sample**: Sample from the dataset which its label and its correspondent score of the cognitive test (CDR or MMSE) do not match. For example, all images labeled as NC with a CDR score different from zero are considered hard samples. Additionally, those images for which the CDR and MMSE value were not available in the dataset (unknown values) are also considered hard.
- **Easy task**: A task is considered easy when the model is asked to distinguish between two distinct/discriminative classes, such as classifying NC from AD.
- **Hard task**: A task is considered hard when the model is asked to distinguish between more than two classes or between two similar classes, which are not so discriminative, such as classifying AD and MCI or NC and MCI.

Three complexity focused strategies were implemented, one focused only on task complexity (Task strategy) and two focused on both sample and task complexity (one based on MMSE and the other on CDR).

Task strategy: In this manual approach, the samples are fed into the network ordered by task complexity. It follows the transfer learning proposal of Grassi et al. [116], yet it is adapted to a curriculum learning strategy consisting in two rounds of training: in the first the model is trained with only AD and NC samples (samples from only two classes and easier to distinguish between them), and only in the second round the MCI samples are added (samples from three classes and harder to distinguish between them).

MMSE based strategy: The training samples were divided, based on the present definition of easy and hard samples and easy and hard tasks, in order to build the curriculum. MMSE test scores were used to manually build the curriculum schematized in Figure 5.6, based on increasing complexity of the samples and increasing complexity of the tasks.

CDR based strategy: CDR test scores were used to manually build the curriculum schematized in Figure 5.7, also based on increasing complexity of the samples and increasing complexity of the tasks.

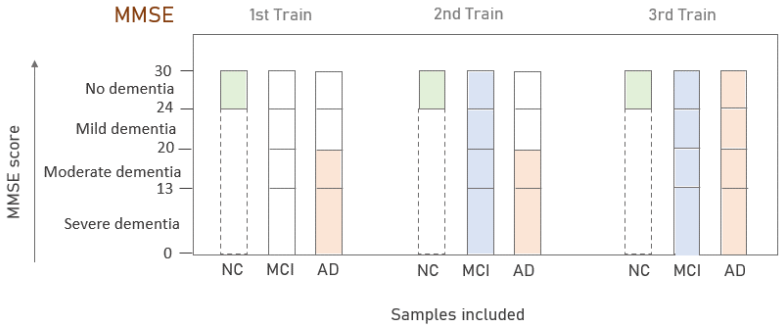

Figure 5.6: Manually defined curriculum based on MMSE scores. The NC, MCI and AD samples included in each round of training are represented in green, blue and orange, respectively, and their MMSE scores are represented in the vertical axis.

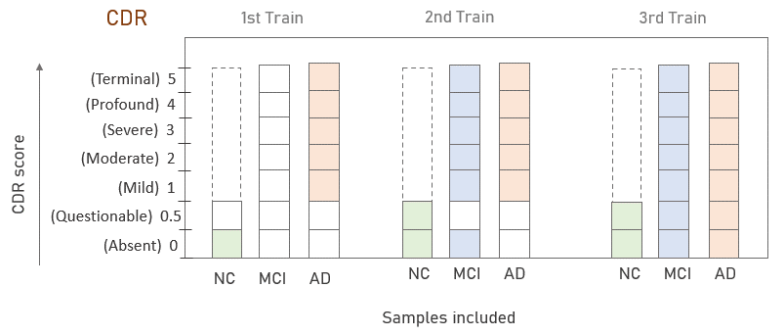

Figure 5.7: Manually defined curriculum based on CDR scores. The NC, MCI and AD samples included in each round of training are represented in green, blue and orange, respectively, and their CDR scores are represented in the vertical axis.

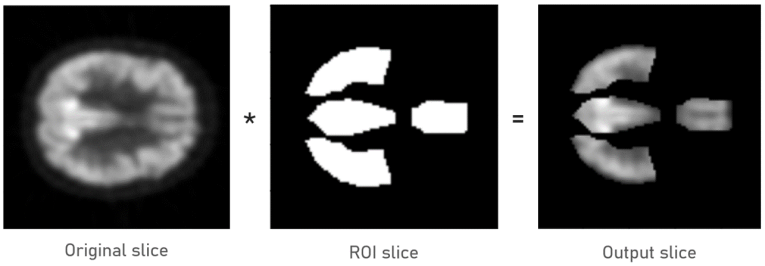

Figure 5.8: FDG-PET image slice filtered by the ROI mask.

In these last two strategies the data is fed into the model based on a predetermined curriculum, which is built based on sample and task complexity. Just like medical students learn, the model will start by learning easy concepts, training first using only an easy task (classify NC from AD). Moreover, to guarantee that the discriminative features of the AD and NC are well learnt, without noisy information, only the easy samples of that task are used in the first round of training. Then, in the second round, a more difficult task is performed, which consists in classifying AD from MCI and NC. In this round, MCI samples are added to the training data, which now comprises mostly easy samples from three classes. In the last round, all hard samples are added to the training data, i.e., all samples are used for training, meaning that an hard task is performed using all easy and hard samples of that task.

5.3.2.2 ROI focused

ROI focused strategies, revised in section 4.2.1.2, focus on progressively adding to the training set more complex regions of the images. The model was first trained with the images of dataset multiplied, through pixel wise multiplication, by a ROI mask (1 inside the ROI and 0 outside), and then it was retrained (fine-tuned) using the complete images. Figure 5.8 exemplifies how the original image and the ROI mask produce the output image, which is then fed into the model in the first round of training.

5.3.2.3 Mixed

Curriculum learning mixed strategies result from merging complexity focused strategies and ROI focused strategies. It follows the principle that the model should be presented first with only the most discriminative regions of the easiest examples (easiest samples multiplied by the ROI mask), and only afterwards include all brain regions, i.e. the complete image, of both easy and hard samples. Thereby, three strategies were defined:

Mix 1 strategy (based on CDR and ROI): follows the strategy described in Figure 5.7, only that the first two trains use the samples multiplied by the ROI mask, while the last train and test are performed using the complete images (without multiplying them by the ROI mask).

Mix 2 strategy (based on MMSE and ROI): similarly to the strategy described in Figure 5.6, only the first two trains use the images multiplied by the ROI mask and the last train and test are performed using the complete images.

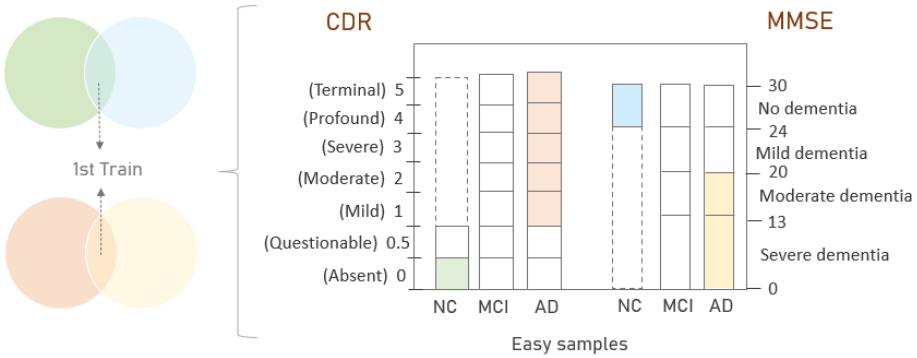

Figure 5.9: Representation of the samples used in the first training stage (on the left) of the mixed strategy based on CDR, MMSE and ROI. They correspond to NC samples considered easy by both CDR (in green) and MMSE (in blue) and only AD considered easy by both CDR (in orange) and MMSE (in yellow).

Mix 3 strategy (based on CDR, MMSE and ROI): the model goes though two training stages: first it is fed with samples that are considered easy according to both MMSE and CDR scores, as schematized in Figure 5.9, multiplied by the ROI mask. In the second train, the model trains with all (both easy and hard) complete brain scans.

Multiple experiences were conducted to find out the optimal way to apply the mixed strategies to the deep learning model. One of them was, for example, to verify if it was better to train the model always using the image data multiplied by the ROI mask (in all training stages and test) or to use the image data multiplied by the ROI mask in the first trains and the complete images for the last trains and test. The latter has proven to be better and was the one applied to the models in this project.

5.3.2.4 Self-paced learning

In the SPL algorithm, as explained in section 4.2.2.1, the training of the network is embedded in the algorithm itself, allowing it to iteratively control the learning curriculum. In this strategy data are sorted while training based on sample training loss [108]. A threshold, λ, is defined and the samples with loss below (above) λ are considered easy (hard). During training the threshold is updated, according to a growing factor, δ, from including only the lower loss samples, to including all samples in the final epochs. This strategy does not take prior medical knowledge into account.

The SPL implementation was based on a simple SPL PyTorch implementation provided by Wenig [117], however it was further adapted to Keras. The parameters, such as the threshold and growing factor were defined according to Ghasedi et al. [118], but were also adapted to the FDG-PET dataset.

The algorithm, described in Algorithm 1, works as follows: at each epoch, the network receives as input the list of examples that it will use for training, i.e, $training\_samples$, and outputs that list, updated, that should be used as input by the model in the next epoch. The $training\_samples$ list is updated based on the loss of each sample, which are calculated every time the model is trained. In the first epoch the model trains with the complete dataset, to initialize the losses of all samples. Then, a threshold is defined in such a way that only 2% of the samples are included in the next train (corresponding

Algorithm 1 Self-paced learning algorithm

procedure SPL ▷ The SPL algorithm automatically defines the curriculum
  $N = number\,of\,samples$
  $E = number\,of\,epochs$
  $training\_samples = [s_1, s_2, ..., s_N]$
  $\lambda = threshold$
  $\delta = growing\,factor$
  for t in $[0, E]$ **do**: ▷ The model is trained E times
    Train the model using $training\_samples$
    $losses = [l_{s_1}, l_{s_2}, ..., l_{s_N}]$ ▷ Save the loss for each training sample
    if $t = 0$ **then**:
      Initial λ is defined ▷ Define initial threshold on the first epoch
    end if
    $updated\_samples = [\,]$
    foreach $x \in [0, ..., N]$ **do**: ▷ Build the $updated\_samples$ array
      if $l_{s_x} <= \lambda$ **then**:
        $updated\_samples = updated\_samples + [s_x]$
      end if
    end for
    if $length(updated\_samples) <= batch\_size$ **then**:
      Add low loss samples to $updated\_samples$ ▷ Avoid too few training samples
    end if
    $training\_samples = updated\_samples$ ▷ Update $training\_samples$ for next train
    $\lambda = \lambda * \delta$ ▷ Update threshold
  end for
end procedure

to the samples with lower losses). Afterwards, in the following epochs, the losses are updated and the threshold increases according to the growing factor, such that more samples are included in the $updated\_samples$ list. However, if this list contains less samples than the number of the batch size, the next samples with lowest loss are added to the $training\_samples$ list, to guarantee that the model always has enough samples to train with. The length of the $training\_samples$ list keeps increasing, from 2% of its full capacity, until it reaches 100%, around $\frac{3}{4}$ of the total number of epochs.

The growing factor was defined in such a way that around $\frac{3}{4}$ of the total number of epochs all samples are included by the algorithm in the training of the model. The value implemented in this strategy, which respects the former constraint, is $\delta = 1.5$.

5.3.2.5 Self-paced curriculum learning

As detailed in section 4.2.2.2, SPCL takes into account both knowledge prior to training and the model's learning progress during training. The SPCL algorithm implemented, described in Algorithm 2, was inspired in the implementation provided by Jiang et al. [111] for a multimedia event detection, yet adapted to the current classification problem.

Similarly to SPL, SPCL works as follows: at each epoch, the network receives as input the list of samples that it will use for training, i.e, $training\_samples$ and outputs a list of training examples, i.e, $updated\_samples$, that should be used by the model itself for training in the next epoch. However, SPCL complements the SPL algorithm, by updating the list of training examples using not only the loss of each

Algorithm 2 Self-paced curriculum learning algorithm

procedure SPL ▷ The SPL algorithm automatically defines the curriculum
  $N = number\,of\,samples$
  $E = number\,of\,epochs$
  $training\_samples = [s_1, s_2, ..., s_N]$
  $\gamma = [\gamma_{s_1}, \gamma_{s_2}, ..., \gamma s_{s_N}]$ ▷ Predetermined curriculum
  $\lambda(t)$ ▷ Growing function
  for t in $[0, E]$ **do**: ▷ The model is trained N_e times
    Train the model using $training\_samples$
    $losses = [ls_1, ls_2, ..., ls_n]$ ▷ Save the normalized loss for each training sample
    $\gamma = \gamma \odot losses$ ▷ Update curriculum
    $threshold = \lambda(t)$ ▷ Update threshold
    $updated\_samples = [\,]$
    foreach $x \in [0, ..., N]$ **do**: ▷ Build the $updated\_samples$ array
      if $\gamma_{s_x} <= threshold$ **then**:
        $updated\_samples = updated\_samples + [s_x]$
      end if
    end for
    if $length(updated\_samples) <= batch\_size$ **then**:
      Add low γs samples to $updated\_samples$ ▷ Avoid too few training samples
    end if
    $training\_samples = updated\_samples$ ▷ Update $training\_samples$ for next train
  end for
end procedure

sample but also a predefined curriculum. In the first epoch the model trains with the complete dataset, to initialize the losses of all samples. The predetermined curriculum, $\gamma = [\gamma s_1, \gamma s_2, ..., \gamma s_N]$, is built as vector with values in the range of [0,1], where a lower values means that the corresponding samples are easier and should be learnt in earlier epochs. This vector is updated during training through element wise multiplication ($\odot$) with the $losses$ vector. The updated γ vector contains information regarding not only prior medical knowledge, but also information about the model's feedback during training: lower values in γ represent samples that were both easy for the model to learn (low loss) and considered easy in the predefined curriculum. Based on this updated vector, a threshold is defined in such a way that only 2% of the samples are included in the second train (corresponding to the samples with lower γs values). Those training samples are then added to the $updated\_samples$. In each of the following epochs the losses and the curriculum are updated, respectively, and the threshold increases according to the growing function, $\lambda(t)$, such that more samples are included in the $updated\_samples$ list. Like in SPL, if these list contains less samples than the number of the batch size, the next samples with lowest γs values are added to the $training\_samples$. The length of the $training\_samples$ list keeps increasing, from 2% of its full capacity, until it reaches 100%, around $\frac{3}{4}$ of the total number of epochs.

Implementation details

Two SPCL strategies were implemented, SPCL 1 and SPCL 2, differing only in the information used to build the predetermined curriculum. The predetermined curriculum, γ, and the growing function, $\lambda(t)$, were built and adjusted for SPCL 1 and SPCL 2.

- The predetermined curriculum, γ, must be an array with values in [0,1], where each instance of

the array, γ_{si}, represents the order that each training sample, s_i, should be added to train.

SPCL 1: In order to built γ, the first step was to define the order in which samples would be fed into the model. For this purpose, the training samples were separated into three different groups, following the MMSE-based division of Figure 5.6, since it yielded better results than the CDR.

- **Group A**: contains only easy AD and NC samples (easy samples of an easy classification task). These consist on the samples used in the first round of training of the strategy described in Figure 5.6.
- **Group B**: contains only MCI samples. These consist on the samples added in the second round of training of the strategy described in Figure 5.6.
- **Group C**: contains only hard AD samples, which correspond to the samples added in the third round of training of the strategy described in Figure 5.6.

It was assumed that all samples in the same group should have the same weight and the samples in group A, the easier ones, should correspond to the lower values, since they are the ones that should be learnt first in the training process. Taking this into account, and after testing multiple possible values for the weights, each entry of the predefined curriculum vector, γ_{s_i}, was defined according to Equation 5.5, where $i \in [0, N]$.

$$\gamma_{s_i} = \begin{cases} 0.33, & s_i \in A \\ 0.66, & s_i \in B \\ 0.99, & s_i \in C \end{cases} \tag{5.5}$$

It is important to notice that the training losses obtained using categorical cross entropy are normalized between 0 and 1 so that the effect of the predetermined curriculum (which is also in [0,1]) is noticed when they are multiplied.

SPCL 2: In this implementation the predetermined curriculum follows the curriculum of the task strategy, where the model first trains only with NC and AD samples and afterwards MCI samples are added. γ_{s_i}, was defined according to equation 5.6, where $i \in [0, N]$.

$$\gamma_{s_i} = \begin{cases} 0.33, & s_i \in NC \cup AD \\ 0.99, & s_i \in MCI \end{cases} \tag{5.6}$$

- The growing function ,$\lambda(t)$, as explained before, dictates how the threshold grows. Similarly to the work of Ghasedi et al. [118], in the first epoch ($t = 0$), the threshold, $\lambda = \lambda(0)$, is defined so that training starts with only 2% of samples. Afterwards, $\lambda(t)$ exponentially increases until all samples are included in training, around $\frac{3}{4}$ of the maximum number of epochs, in epoch t=75.

The growing functions allows the threshold to grow in such a way that all samples are progressively included in train, as exemplified in Figure 5.10. In the first epoch the minimum threshold is defined. Then, as the threshold grows, more samples are gradually included in training (those below it) until the model is training with all samples available.

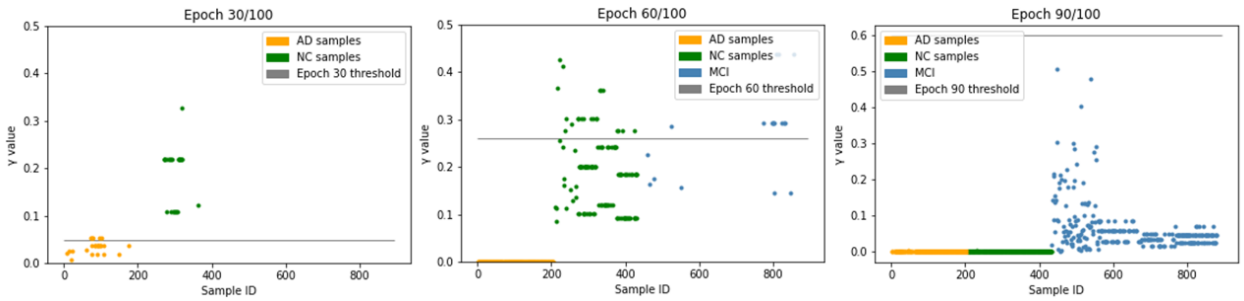

Figure 5.10: Representation of the samples (s_i) used for training by the SPCL 1 model in epoch 30, 60 and 90, their γ_{s_i} value and the threshold (grey line), which determines the samples that should be used for training in the next epoch (samples below it).

5.3.2.6 Replicate automatic strategy

The last strategy implemented was to manually build a curriculum that mimics the automatically defined curriculum of the automatic strategies. Thereby, a curriculum with three rounds of training was built, exemplified in Figure 5.11, where in the first round, the model is fed only with AD labeled samples, adding the NC labeled samples in the second round and finishing in the third round by feeding the complete dataset into the model (AD, NC and MCI labeled samples). This curriculum is equivalent to the one automatically generated in SPL. Figure 5.12 shows the samples used for training in the first epochs of the SPL strategy (a), as well as the samples used in the middle epochs (b)) and in the last epochs (c)).

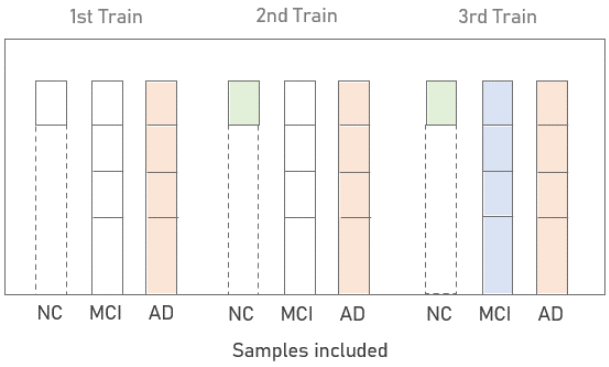

Figure 5.11: Manually defined curriculum, based on the automatically generated curriculum presented in Figure 5.12, where samples are gradually added to the next train, first (1st), second (2nd) and third (3rd), respectively. The NC, MCI and AD samples included in each round are represented in green, blue and orange, respectively.

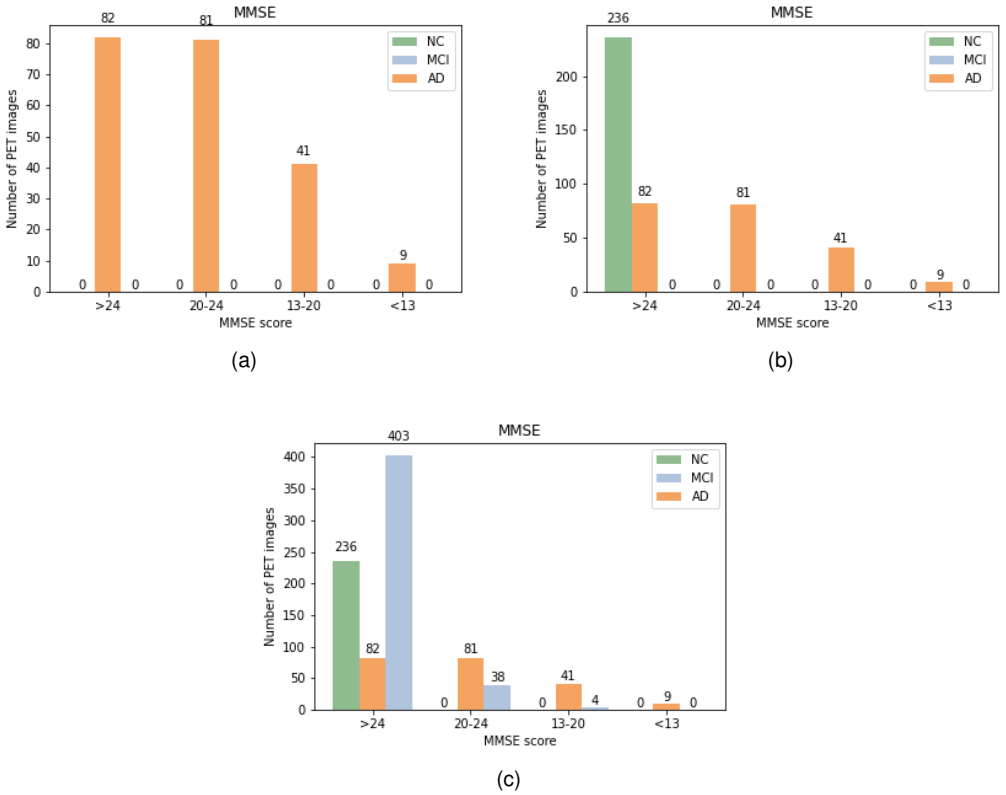

Figure 5.12: Print of the automatically generated curriculum by the SPL algorithm for a training dataset comprising 894 samples (236 NC, 445 MCI and 213 AD): (a) **first epochs**: the model trained with only AD samples, (b) **middle epochs**: all the AD and NC samples were used and (c) **final epochs**: all samples available were used for training.

5.4 Baseline methods

In order to proper evaluate the curriculum learning results, they must be compared to the results of baseline strategies. Three baseline methods were applied: the Simple model, the Focal loss model and the Sample weights model. Although none of the baseline methods use curriculum learning, the Focal loss model takes into account the model's feedback and Sample weights model takes into account medical knowledge prior to training, such as MMSE scores.

5.4.1 Simple model

This strategy corresponds to train a CNN, with the same architecture as the CL strategies, by presenting the entire dataset was to the network at every training epoch.

5.4.2 Focal loss

Another baseline strategy used was training the model in a similar way as the simple model, but using the Focal loss (FL) [119] as loss function instead of the sparse categorical loss. FL is particularly useful in cases where there is a class imbalance, which was the case of the selected dataset. In this

strategy, the FL function implemented is an α-balanced variant of FL and is expressed as follows [119] (Equation5.7):

$$FL(y, \hat{p}_y) = -\alpha(1 - \hat{p}_y)^{\theta} * log(\hat{p}_y) \tag{5.7}$$

$y = [0, ..., K-1]$ is an integer class label (K denotes the number of classes), $\hat{p}_y = [\hat{p}_0, ..., \hat{p}_{K-1}]$ is a vector representing an estimated probability distribution over the K classes and α represents the balance factor. θ is a focusing parameter which smoothly adjusts the rate at which easy examples are down weighted. Easily missclassified samples are considered hard samples and are associated with higher loss values. FL tries to handle the class imbalance problem by assigning more weight to hard (easy) samples by increasing (decreasing) the value of θ.

In our implementation the values $\alpha = 0.25$ and $\theta = 2$ were used. These were the values used by Zhao et al. [120], which used a FL function to predict AD's progression, from NC to MCI and to AD.

5.4.3 Sample weights

In the Sample weights (SW) strategy the model was trained in the same way as the simple model, only rather than using class weights, sample weights were implemented. Each sample was associated with a specific weight during training. This weight specifies how much influence each sample in a batch should have, in the computation of the total loss. It is important to notice that this strategy can not be considered a CL strategy since, despite different samples having different weights, they are all randomly presented to the model, not in a specific order.

In order to fairly compare this strategy to the curriculum learning ones, the weights of the samples were defined following the same logic of the predetermined curriculum. Here, easier samples are associated with higher weights in the beginning so that they are given more relevance. Then their weight decreases as we evolve through the epochs. Contrarily, harder samples are associated with lower weights in the first epochs, which increase as the model progresses through the epochs. The samples were divided into the same 3 groups (A, B and C) as the ones described in section 5.3.2.5 and the weight value, $weight_{s_i}$, of each sample, s_i, is defined according to table 5.5, where $i \in [1, N]$ and N corresponds to the total number of training samples.

Table 5.5: Value of $weight_{s_i}$ with respect to s_i and the epoch number (t).

$\boldsymbol{weight_{s_i}}$		s_i		
		$s_i \in A$	$s_i \in B$	$s_i \in C$
Epoch number (t)	t $<$ 30	1.33	1	0.77
	30 $<$ t $<$ 60	1	1.33	0.77
	t $>$ 60	0.77	1	1.33

Chapter 6

Results and Discussion

6.1 Computational specifications

The experiments were performed on a single NVIDIA GeForce GTX 1070 GPU with 8GB of memory, in a machine with an Intel Core i7-6800K @ 3.40GHz CPU. Additionally, all experiments were carried out in Python 3.6 and all deep learning implementations were based on deep learning libraries Tensorflow and Keras, which is a high-level application programming interface (API) of Tensorflow.

6.2 Baseline methods results

The results obtained for the baseline methods are displayed in Table 6.1. The Simple model presents the poorest overall accuracy, F1-score and MCI accuracy. Out of the baseline models, it can be considered the least suitable for the selected dataset and for early AD diagnosis. This can be due to the fact that the Simple model does not take into account the model's learning feedback nor medical knowledge prior to training, like the Focal loss and Sample weights models do, respectively. On the one hand, the Focal loss model shows a slightly improvement of 1.3% in overall accuracy, when compared to the Simple one. On the other hand, the Sample weights model improved the overall accuracy, from 82.7% to 85.3%, and the MCI accuracy by and 22.5%, when compared to the Simple model. These results show that taking the model's feedback into account (Focal loss) and incorporating medical knowledge into the models (Sample weights) is advantageous for improving the overall and MCI accuracy, being that the later had a higher contribution for such improvements. However, in the Sample weights strategy,

Table 6.1: Results of baseline models: overall and class specific accuracy, F1-score (F1), area under the curve (AUC) and training time as $(mean \pm standard\,deviation)$.

Model	Accuracy (%)				F1 (%)	AUC (%)	Time (min)
	Overall	NC	MCI	AD			
Simple	82.7 ± 0.8	$\mathbf{94.6 \pm 2.5}$	71.1 ± 1.1	$\mathbf{95.6 \pm 1.8}$	83.0 ± 0.8	96.2 ± 0.7	52
Focal Loss	84.0 ± 0.5	86.2 ± 6.3	76.4 ± 5.1	92.9 ± 4.3	83.7 ± 0.3	$\mathbf{97.2 \pm 0.1}$	55
Sample weights	$\mathbf{85.3 \pm 0.6}$	76.7 ± 1	$\mathbf{93.6 \pm 1.5}$	78.2 ± 2.8	$\mathbf{85.3 \pm 0.6}$	92.8 ± 4.1	55

the improvement of MCI accuracy is achieved at the cost of AD and NC accuracy, and AUC value, which suffered a 3.4% decrease.

6.3 Curriculum learning results

6.3.1 Manual strategies

For the ROI focused strategy, multiple analysis were performed in order to assess which ROIs were best for the classification task at hand. The most relevant were:

- **Training models using ROIs**: Two types of training strategies were performed, one consisted on training a model once using all images multiplied by a ROI mask and the other on training a model using a ROI focused curriculum learning strategy. In the former, 10 separated models were trained, each one of them using as input the entire training dataset multiplied by one of the ROIs so that the model learns how to classify AD, NC and MCI using only the pixels inside each ROI. In the latter, 10 separated models were also trained, but using curriculum learning instead. In the first round of training of each model, only the pixels inside the ROI were used and in the second round the complete images were considered. The results obtained are presented in Table 6.2, which highlight ROI 5, 7+8, 9 and 10 with highest accuracy for discriminating AD, MCI and NC.
- **Pixel average analysis**: In order to understand how discriminative each ROI is, a pixel average analysis was performed. This analysis consisted in averaging all pixel values of each label, for all images, for each ROI, after normalization. After that, multiple plots were produced in order to explore how different is the pixel average of the different labels, in each ROI. When analysing Figure 6.1, we can verify that ROI 5, 9 and 10 are the most discriminative, showing almost no overlap of the pixel average in the three classes. Pixel average of ROI 6, ALL ROI and ROI 3+4 mostly overlap in two classes, AD and NC for the fist two and NC and MCI for the last one. For the other ROIs, the pixel averages overlap in all three classes.
- **Literature research**: A thorough literature research was performed in order to understand the state-of-the-art about the most relevant brain regions to discriminate AD patients from healthy controls. According to Rondina et al. [121], medially located posterior cortical regions (such the medial parietal cortex, encompassing the precuneus and posterior cingulate gyrus) are the most discriminating regions in PET images. Similarly, Yokoi et al. [122] concluded that the precuneus/posterior cingulate cortex play an important role in developing dementia in AD, by studying these regions significance to discriminate AD from NC in PET and fMRI. Moreover, Hiscox et al. [123] performed a voxel-based morphometry of both cortical and subcortical grey matter, which revealed volume reductions due to AD in the hippocampus, middle, superior temporal gyri and precuneus.

Gathering the results from the trained models using ROIs and the pixel average analysis, it was verified that the most discriminating ROIs according to these match most discriminative regions for AD found in bibliography, as shown in table 6.2.

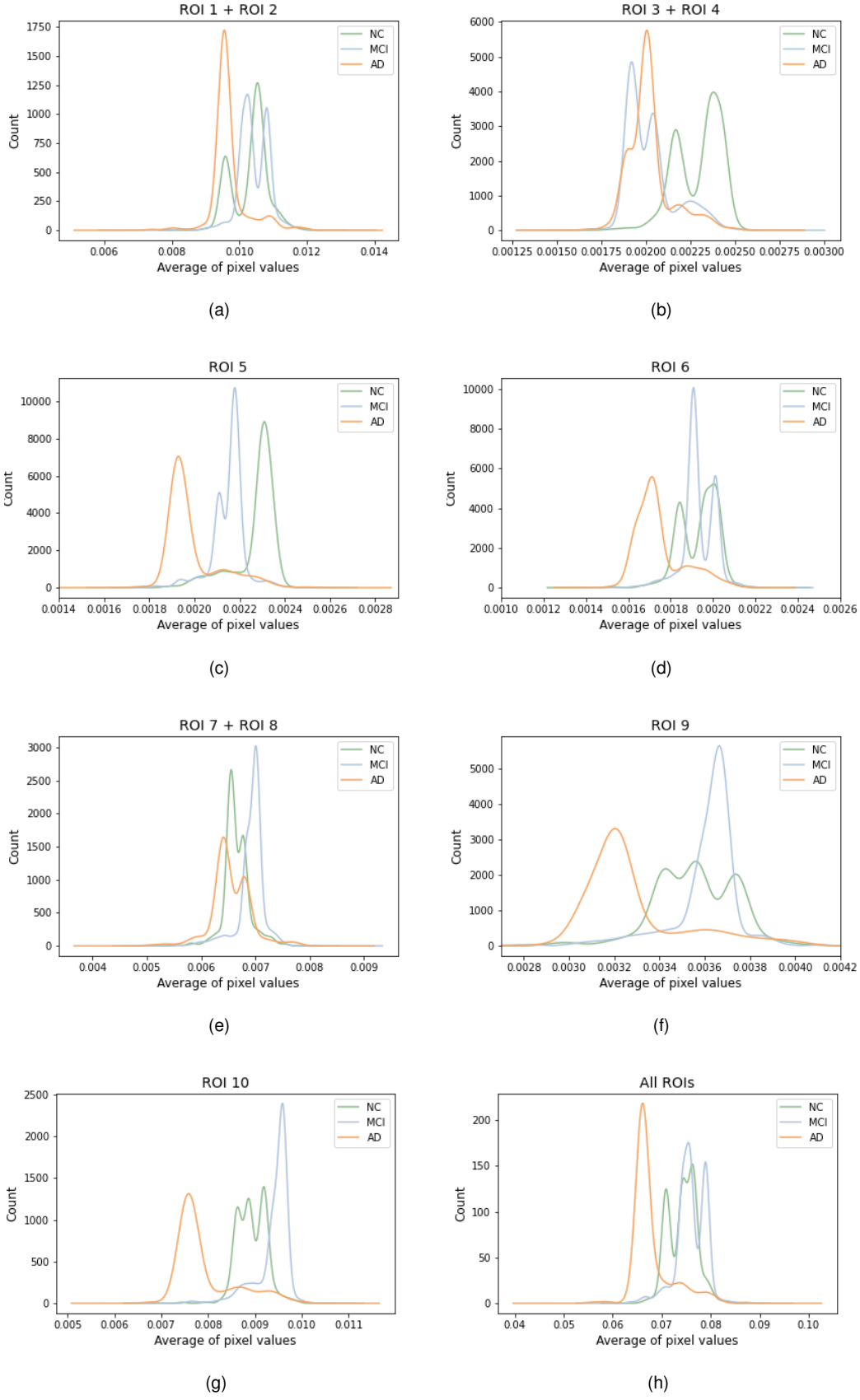

Figure 6.1: Histogram of the average of the pixel values inside: (a) ROI 1+2, (b) ROI 3+4, (c) ROI 5, (d) ROI 6, (e) ROI 7+8, (f) ROI 9, (g) ROI 10 and (h) ALL ROI, for all images labeled as NC (in green), as MCI (in blue) and as AD (in orange).

Table 6.2: Summary of ROI information. The last two columns highlight the most discriminative ROIs for the classification of NC, MCI and AD, from the perspective of the pixel average method and literature research, respectively.

ROI name	Brain Region	Accuracy simple train (%)	Accuracy CL train (%)	Pixel average	Literature research
1+2	Left and right lateral temporal	80.0	82.8		
3+4	Left and right mesial temporal	83.0	84.2		
5	Inferior frontal gyrus/Orbitofrontal	84.4	84.9	✓	✓
6	Inferior anterior cingulate	74.9	84.6		
7+8	Left dorsolateral parietal	84.6	84.8		
9	Right dorsolateral parietal	83.9	85.1	✓	✓
10	Posterior cingulate and precuneus	83.8	84.5	✓	✓
All Rois	—	83.0	85.7		
5+9+10	—	**86.5**	**86.7**		

Taking into account the results of these three analysis, a new ROI was built: ROI 5+9+10. It results from the merge of ROI 5, 9 and 10. The models were trained once more, this time using ROI 5+9+10, and the performance improved when compared to the models where only one ROI was used at the time (Table 6.2). The results presented ahead for the ROI focused strategies used ROI 5+9+10 and All ROIs as the ROI mask. The latter was considered as a comparison strategy, since ROI 5+9+10 only included about 3.45% of the pixels of each brain scan.

Furthermore, in the Mixed strategies, the ROI mask used was the one corresponding to all ROIs.

The results of all manual curriculum learning strategies implemented are presented in Table 6.3. The Replicate strategy, which uses a replica of the curriculum of automatic strategies, is the one that presents highest F1-score and MCI accuracy. However, the ROI strategy using ROI 5+9+10 has the ROI mask is the one with highest overall accuracy, directly followed by the Replicate strategy and the Task strategy, which was the fastest manual method.

On the one hand, comparing the strategies that incorporate the scores of the cognitive tests in the process of building the curriculum, the MMSE has proven to be the best regarding both overall and MCI accuracy. On the other hand, comparing the strategies that incorporate ROI information, the use of

Table 6.3: Results of manual curriculum learning strategies: overall and class specific accuracy, F1-score (F1), area under the curve (AUC) and training time as $(mean \pm standard\,deviation)$.

Model	Accuracy (%)				F1 (%)	AUC (%)	Time (min)
	Overall	NC	MCI	AD			
Task	86.6 ± 0.8	81.2 ± 4.2	89.7 ± 3.2	84.7 ± 6.9	86.6 ± 0.8	96.8 ± 1.0	$\mathbf{97}$
MMSE	86.3 ± 1.2	82.7 ± 6.4	91.7 ± 5.1	77.8 ± 3.8	86.7 ± 1.3	97.0 ± 0.5	174
CDR	86.1 ± 1.4	82.7 ± 5.9	90.0 ± 5.3	81.9 ± 4.8	86.1 ± 1.4	$\mathbf{97.1 \pm 0.1}$	172
ROI ALL	85.7 ± 1.1	77.4 ± 4.3	89.9 ± 5.9	81.1 ± 7.2	85.8 ± 1.1	96.2 ± 1.2	103
ROI 5+9+10	$\mathbf{86.7 \pm 1.8}$	84.0 ± 7.7	87.4 ± 7.8	84.4 ± 7.0	86.8 ± 1.8	95.3 ± 2.7	99
Mix 1	85.6 ± 1.0	83.7 ± 8.6	88.0 ± 6.2	83.9 ± 3.5	85.7 ± 0.9	97.0 ± 0.7	149
Mix 2	86.1 ± 1.2	80.5 ± 6.9	88.4 ± 3.5	$\mathbf{86.6 \pm 4.8}$	86.0 ± 1.3	96.9 ± 0.9	147
Mix 3	85.7 ± 0.8	$\mathbf{87.8 \pm 5.3}$	85.4 ± 3.8	83.2 ± 1.7	85.9 ± 0.6	96.7 ± 0.5	89
Replicate	86.6 ± 1.4	79.3 ± 3.0	$\mathbf{92.8 \pm 4.5}$	81.0 ± 3.4	$\mathbf{87.0 \pm 1.4}$	96.3 ± 0.7	163

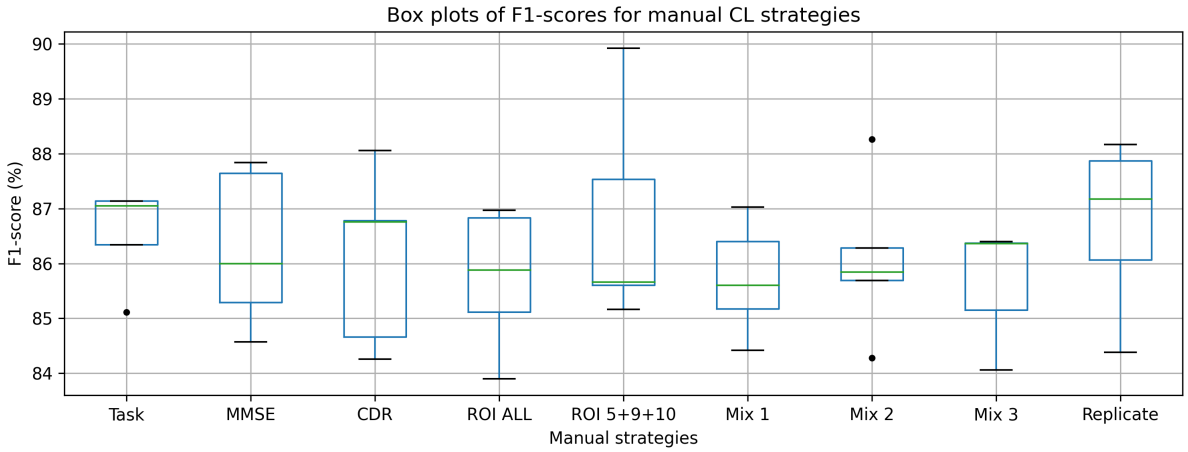

Figure 6.2: Box plots of the F1-score results for the manual curriculum learning strategies, where the maximum, minimum and median (in green) values are indicated.

ROI 5+9+10 has shown to be advantageous, improving the overall accuracy by 1%, when compared to All ROI. Regarding the mixed strategies, we can see that there is no advantage in combining both information of cognitive tests and ROI to build the curriculum, since the results were the poorest out of all manual curriculum learning strategies. Regarding the AUC, we can observe that the strategies MMSE, CDR, and Mix 2 are the best ones at distinguishing between the three classes, obtaining the highest AUC values.

Figure 6.2 allows to visually compare the results of the 5-fold cross validation method, and further corroborate the interpretation of the results of Table 6.3. Each box plot corresponds to the F1-scores of the 5 models trained for each manual strategy. The Task and Replicate strategies are still the ones with the best results, i.e., higher median. Nevertheless, this figure also gives us information about the dispersion of data, showing that, despite the fact that median of the Replicate strategy is slightly higher, the results of the Task strategy have lower dispersion, and therefore are more robust. It can also be verified that the ROI 5+9+10 strategy achieves the maximum F1-score, around 90%. However, it also presents the higher dispersion of the results.

To sum up, both medical knowledge prior to training (such as information about task complexity, cognitive test scores and ROIs) and the model's feedback (as the one used in the Replicate strategy) are suitable for building a training curriculum. The MMSE and CDR strategies, which use both task complexity information and the cognitive test scores to built the curriculum, have shown that using different types of medical knowledge can be advantageous. Nevertheless, mixing information of cognitive tests and ROI, such as the proceeding in the mixed strategies, did not show any improvement.

Table 6.4: Results of automatic curriculum learning strategies: overall and class specific accuracy, F1-score (F1), area under the curve (AUC) and training time as $(mean \pm standard\,deviation)$.

Model	Accuracy (%)				F1 (%)	AUC (%)	Time (min)
	Overall	NC	MCI	AD			
SPL	85.9 ± 0.8	86.1 ± 2.7	85.2 ± 1.3	88 ± 3.8	86.1 ± 0.9	96.6 ± 0.9	$\mathbf{43}$
SPCL 1	$\mathbf{87.2 \pm 1.2}$	85.6 ± 2.9	$\mathbf{88.4 \pm 3.5}$	86.6 ± 2.8	$\mathbf{87.3 \pm 1.4}$	$\mathbf{97 \pm 0.7}$	48
SPCL 2	86.4 ± 1.7	$\mathbf{87.9 \pm 4.3}$	83.9 ± 6.6	$\mathbf{88.9 \pm 3.7}$	86.5 ± 1.7	95.9 ± 1.5	48

6.3.2 Automatic strategies

Table 6.4 presents the results obtained for the automatic strategies. Analysing it, we can observe that SPL is the fastest approach and SPCL yields the best results. SPCL strategies, in comparison with SPL, require the extra work of building the curriculum. Nevertheless, they complement SPL and prove that incorporating medical knowledge prior to training into the models brings an added advantage. SPCL 1 improves the overall accuracy of SPL by 1.3%, achieving the best overall performance.

Comparing the two SPCL strategies, the first one shows higher overall and MCI accuracy, F1-score and AUC. This shows that combining information about the MMSE scores and task complexity (SPCL 1 proceeding) for building the training curriculum is more advantageous than using only information about task complexity (SPCL 2 proceeding).

Furthermore, Figure 6.3 allows to evaluate the robustness of the results for the automatic strategies. Regarding the median, the results corroborate those of Table 6.4: SPCL 1 is the best strategy, followed by SPCL 2 and SPL. Both SPCL strategies present a similar dispersion of data and similar maximum value. SPL presents the lower data dispersion. However, this only happens due to the fact that two of the results are considered outliers.

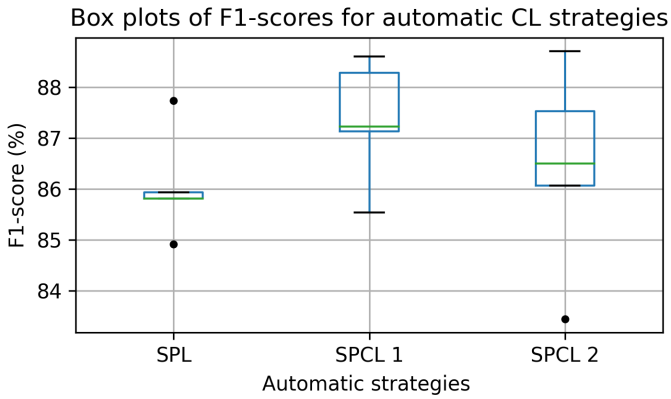

Figure 6.3: Box plots of the F1-score results for the automatic curriculum learning strategies, where the maximum, minimum and median (in green) values are indicated.

6.4 Comparison between strategies

6.4.1 Classification results

The results per class, for all methods implemented, are presented in Figure 6.4. The overall accuracy and F1-scores for all methods are also displayed in Figures 6.5 and 6.6, respectively.

Comparing the curriculum learning strategies with the baseline methods, we can observe that, even though the curriculum learning strategies decrease the accuracy of the classification of AD and NC individually, they improve the MCI classification, maintaining an higher F1-score and overall accuracy (Figures 6.5 and 6.6). Nevertheless, this is only verified when we compare the curriculum strategies with the Simple and Focal loss models. The Sample weights strategy, which incorporates medical knowledge, is the one to achieve the best MCI accuracy, however, it is also the one with the worst AD and NC accuracy, by far.

The incorporation of the model's feedback and medical knowledge into the models has shown to be effective to improve overall and MCI accuracy. This is true for baseline methods (Focal loss and Sample weights) as well for all curriculum learning strategies applied. However, the later show less discrepancy between MCI and AD/NC accuracies, achieving similar MCI accuracy to the Sample weights model, but always better accuracy for AD and NC. Thereby, curriculum learning strategies can be considered superior to all baseline methods, even those which incorporate extra information.

Comparing the manual and automatic strategies, by analysing Figure 6.4, it can be verified that the manual ones, despite achieving higher MCI accuracy, they also have higher uncertainty and higher discrepancy between MCI and AD/NC accuracies, making the automatic ones the most robust. Regarding Figures 6.5 and 6.6, the SPCL 1 strategy (automatic) has yielded the best results in terms of overall accuracy and F1-score, followed by five strategies, four of them manual, Task, ROI 5+19+10, Replicate and MMSE, and one automatic, SPCL 2. Concerning the model running time, although baseline methods are faster than manual strategies, the automatic ones are even faster and yield better results.

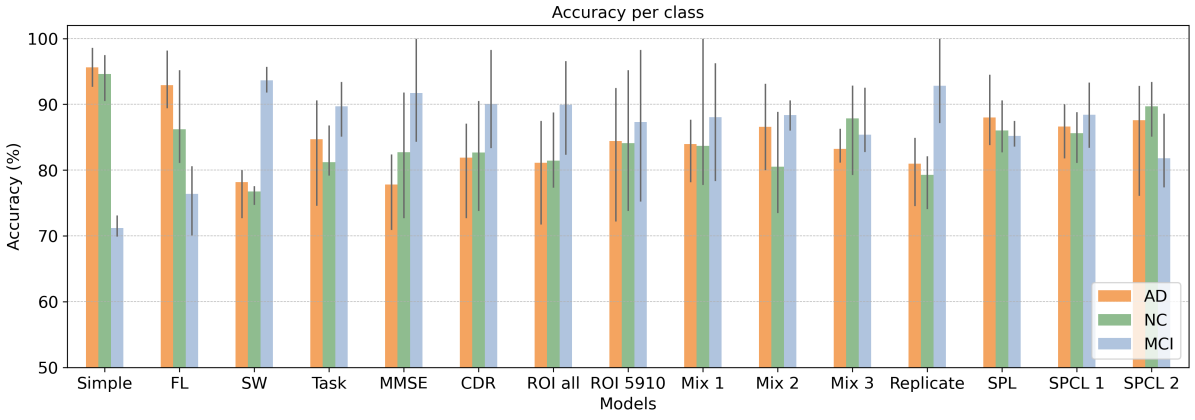

Figure 6.4: Bar plots representing the accuracy per class (AD, NC and MCI), for all the implemented models with error lines indicating the variability of data (minimum and maximum value). FL: Focal loss; SW: Sample weights.

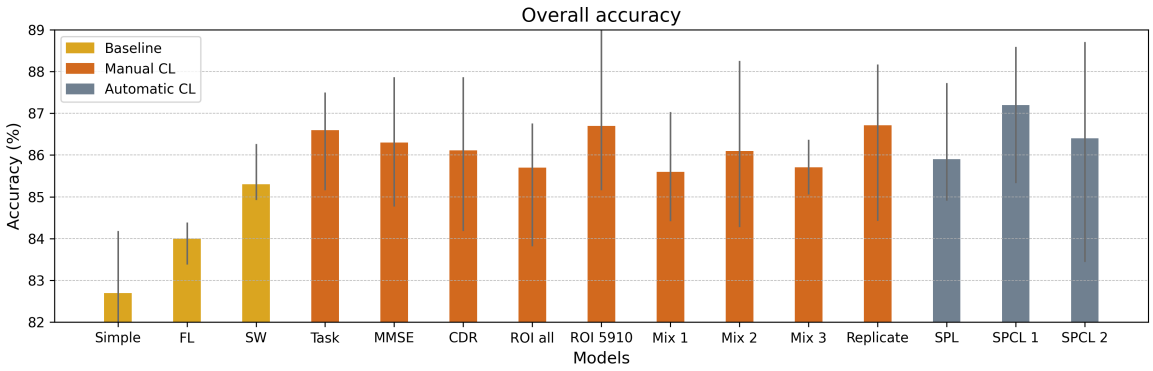

Figure 6.5: Overall accuracy results for all strategies implemented with error lines indicating the variability of data (minimum and maximum value). FL: Focal loss; SW: Sample weights.

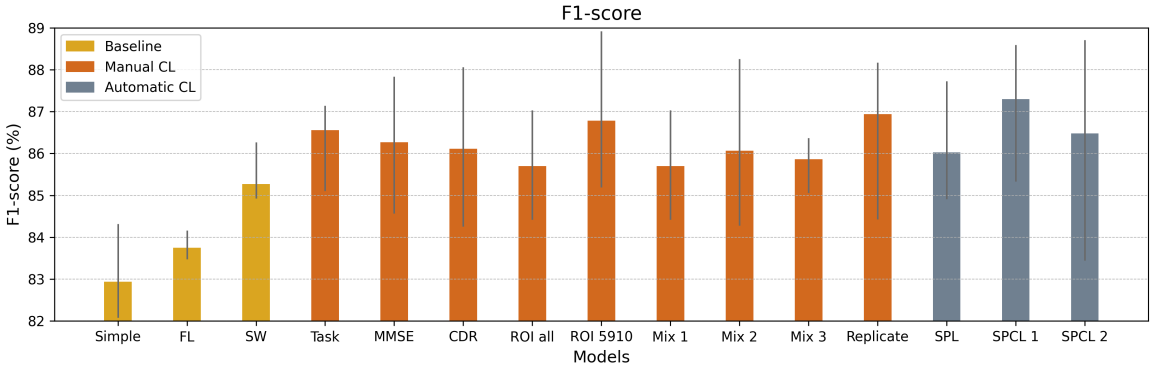

Figure 6.6: F1-score results for all strategies implemented with error lines indicating the variability of data (minimum and maximum value). FL: Focal loss; SW: Sample weights.

Regarding the results of the replicate strategy, it has shown to be one of the best implemented strategies. These results were rather surprising since in the first round of training the model is not learning how to classify different labeled images, it trains only with AD images. We expected this to have a negative impact in the results, such as a higher AD accuracy and a decrease in the ability of the model to distinguishing between classes, since in the first round of training this model learns to classify every image as AD. However, this was not verified. Moreover, the results of the Replicate strategy were significantly better than the ones obtained for SPL (Figures 6.5 and 6.6), which uses an equivalent curriculum, only in a different way. This could be explained by the fact that in the SPL strategy the model has less epochs to train with all available samples. Therefore, the model might not have enough epochs to adjust to the complete dataset and learn the best weights and biases, which is reflected in lower performance.

The results obtained show that all the proposed curriculum learning strategies improve both overall and MCI classification (early AD), and F1-score performances. Additionally, the incorporation of medical knowledge into the process of building the curriculum has also proven to be advantageous for early AD diagnosis, since all strategies that incorporate it yield better MCI accuracy results than SPL, Focal loss and Simple model, which do not take it into account (Figure 6.4).

6.4.2 Statistical relevance

To assess the statistical significance of the results obtained, the Wilcoxon signed-rank test was used. This is a non-parametric statistical paired test, which allows to compare matched samples from two different populations, without assuming a Gaussian distribution.

Table 6.5 presents the p-values of the statistical tests performed between the results of the curriculum learning strategies and each of the baselines methods. The p-values were lower than the threshold (0.05) for four out of the six best curriculum strategies, such as Task, MMSE, Replicate and SPCL 1. This indicates that the difference between their results and those of the baseline methods are statistically relevant. Regarding the ROI strategies, despite ROI 5+9+10 being one of the strategies with best results, the p-values are too big for the null hypothesis (the two populations/results have the same distribution with the same median) to be rejected. One explanation to justify the fact that the difference between the results of this model and those of the Simple model are not statistically relevant is that they do not use different information. In the ROI strategy, no new information is added in comparison to the Simple model. For both methods, all the images are all presented at once. The difference relies on the fact that, in the ROI strategy, a smaller portion of that images (correspondent to the ROI) is presented in the first training epoch, and only in the second train all complete images are fed into the model. Moreover, regarding the SPCL 1 results, they show statistical relevance when compared to the results of the Simple model and Focal loss, but not when compared to the Sample weights model. This can be explained by the fact that these two models use equivalent curricula.

In general, it was verified that the difference between the results of curriculum learning strategies and baseline methods are statistically relevant, which contributes for the robustness of these strategies.

Additionally, Table 6.6 presents the p-values between the results of comparable curriculum learning strategies, whether by the fact that they use equivalent or similar curricula or whether to compare the effect of incorporating medical knowledge *vs* not incorporating it. In the first row of the table, in order to further compare the results of automatic strategies, the p-values between those that do not incorporate

Table 6.5: P-value between the predictions of the baseline methods and curriculum learning strategies. The p-values below the threshold (0.05) are highlighted in gray.

p-value	Baseline methods		
	Simple model	Focal loss	Sample weights
Task	$\mathbf{0.006}$	$\mathbf{0.011}$	$\mathbf{0.020}$
MMSE	$\mathbf{0.007}$	$\mathbf{0.003}$	$\mathbf{1.56e{-}10}$
CDR	0.098	0.073	$\mathbf{0.012}$
All ROI	0.061	0.062	0.065
ROI 5+9+10	0.061	0.067	0.065
MMSE+ROI	$\mathbf{0.024}$	$\mathbf{0.025}$	0.064
CDR+ROI	$\mathbf{0.046}$	$\mathbf{0.037}$	$\mathbf{0.027}$
MMSE+CDR+ROI	$\mathbf{0.026}$	$\mathbf{0.005}$	$\mathbf{0.003}$
Replicate	$\mathbf{0.004}$	$\mathbf{0.019}$	$\mathbf{0.014}$
SPL	0.052	0.065	0.061
SPCL 1	$\mathbf{0.038}$	$\mathbf{0.028}$	0.075
SPCL 2	0.056	0.052	0.05

Table 6.6: P-value between the predictions of curriculum learning strategies. The p-values below the threshold (0.05) are highlighted in gray.

p-value	SPCL 1	SPCL 2	Replicate	Task	MMSE	CDR
SPL	0.087	0.12	**0.009**	—	—	—
Task	—	—	—	—	**5.10e−11**	**3.04e−4**
SPCL 1	—	—	—	—	**2.41e−6**	—
SPCL 2	—	—	—	**7.33e−5**	—	—

medical knowledge (SPL) and those that do (SPCL), were obtained. It was verified that there is no statistically relevance between these results, despite the fact that their accuracy and F1-scores differ up to 1.3% and 1.2%, respectively. In the second row, the strategies compared were Task, which only incorporates the knowledge of task complexity, and MMSE and CDR, which incorporate both knowledge of task complexity and the cognitive test results. The low p-values reflect that there is a significant difference in the results when the information of MMSE and CDR scores are combined with task complexity to build the curriculum. Moreover, to fairly compare automatic strategies with manual ones, we obtained the p-values between the results of these two types of strategies using the same curriculum: SPL *vs* Replicate, SPCL 1 *vs* MMSE and SPCL 2 *vs* Task. The p-values are bellow the threshold in all these three cases, which allows us to conclude that the differences in performance between the results of manual and automatic strategies are statistically relevant.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

This thesis was, as far as we know, the first work investigating the use of curriculum learning for early AD diagnosis from neuroimaging. Twelve different CL strategies, nine manual and three automatic, incorporating different kinds of medical knowledge were implemented. The knowledge could be in the form of task complexity (Task and SPCL 2), ROI information (ROI all and ROI 5+9+10) or a combination of different types of information, such as mixing cognitive test scores with task complexity (MMSE, CDR, SPCL 1) or with ROI information (Mix 1, Mix 2 and Mix 3). Moreover, one automatic strategy (SPL) and one manual (Replicate) were also defined, for comparison purposes, since none of them incorporated any kind of medical knowledge.

The results show that all the proposed curriculum learning strategies improve both overall and MCI classification (early AD) performances, highlighting their superiority regarding the baseline methods. Out of the manual strategies, the ROI focused strategy, the Task focused strategy and the Replicate were the ones to yield the best overall results and the MMSE strategy obtained the best MCI accuracy. The automatic strategies have shown to be the best ones, in terms of performance, robustness and time. In fact, SPCL 1 has obtained the best overall accuracy and F1-score. For all CL strategies, the incorporation of medical knowledge (in the form of ROI information, task complexity information and cognitive test scores scores) into the curriculum learning strategies has proven to be advantageous, further improving the overall accuracy, F1-score and MCI accuracy.

The results obtained in this thesis show that the order in which data is fed into the CNNs, for early AD diagnosis, is meaningful. They have also demonstrated the added advantage of using medical domain knowledge for building the curricula. That said, curriculum learning strategies incorporating medical knowledge allow for better early AD diagnosis, which can contribute to the ongoing search for treatments to prevent or delay the onset of this devastating disease.

7.2 Future Work

The investigation performed in this thesis can be of great importance to the clinical research developed towards finding therapeutics and a cure for AD. Even though the results obtained were distinctly positive, there is still a lot of room for improvements. For example, other types of external information, such as medical imaging reports or evidence maps obtained during training, could be used for developing different curricula for the curriculum learning strategies. Moreover, other strategies could be developed for determining the curriculum. For instance, the creation of two different deep learning models, one for classifying the images and another for defining the curriculum. The latter could be a regression neural network that receives as input the images, the encoded medical knowledge and the classification model's feedback and returns a value for each image, which represents the order in which they should be added to the training set.

Regarding improving the early AD diagnosis, to make a more accurate prediction, these strategies could be applied to a dataset that allows for MCIc *vs* MCInc distinction, allowing to distinguish early AD from other unrelated dementia cases.

Additionally, since the proposed approaches were successfully implemented for the current classification problem, they could also be applied to other type of input images, different from PET, such as MRI or others, or yet adapted to the diagnosis of other neurogenerative disorders, like Parkinson's or Huntington's disease.

Bibliography

[1] M. A. Ebrahimighahnavieh, S. Luo, and R. Chiong. Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review. *Computer methods and programs in biomedicine*, 187:105242, 2020.

[2] C. L. Masters, R. Bateman, K. Blennow, C. C. Rowe, R. A. Sperling, and J. L. Cummings. Alzheimer's disease. *Nature reviews disease primers*, 1:15056, 2015.

[3] X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, and S. Yu. A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis*, page 101985, 2021.

[4] M. A. Myszczynska, P. N. Ojamies, A. M. Lacoste, D. Neil, A. Saffari, R. Mead, G. M. Hautbergue, J. D. Holbrook, and L. Ferraiuolo. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nature Reviews Neurology*, 16(8):440–456, 2020.

[5] M. I. Razzak, S. Naz, and A. Zaib. Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps*, pages 323–350, 2018.

[6] A. Association. 2008 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 4(2):110–133, 2008.

[7] T. K. Khan. Biomarkers in Alzheimer's disease. *Academic Press*, 2017.

[8] C. L. Masters, R. Bateman, K. Blennow, C. C. Rowe, R. A. Sperling, and J. L. Cummings. Alzheimer's disease. *Nature reviews disease primers*, 1:15056, 2015.

[9] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arrighi. Forecasting the global burden of Alzheimer's disease. *Alzheimer's & dementia*, 3(3):186–191, 2007.

[10] E. C. Edmonds, C. R. McDonald, A. Marshall, K. R. Thomas, J. Eppig, A. J. Weigand, L. Delano-Wood, D. R. Galasko, D. P. Salmon, M. W. Bondi, et al. Early versus late MCI: Improved MCI staging using a neuropsychological approach. *Alzheimer's & Dementia*, 15(5):699–708, 2019.

[11] P. S. Aisen, R. C. Petersen, M. C. Donohue, A. Gamst, R. Raman, R. G. Thomas, S. Walter, J. Q. Trojanowski, L. M. Shaw, L. A. Beckett, et al. Clinical core of the Alzheimer's Disease Neuroimaging Initiative: progress and plans. *Alzheimer's & Dementia*, 6(3):239–246, 2010.

[12] L. J. Thal, K. Kantarci, E. M. Reiman, W. E. Klunk, M. W. Weiner, H. Zetterberg, D. Galasko, D. Praticò, S. Griffin, D. Schenk, et al. The role of biomarkers in clinical trials for Alzheimer disease. *Alzheimer disease and associated disorders*, 20(1):6, 2006.

[13] K. Blennow and H. Zetterberg. Biomarkers for Alzheimer's disease: current status and prospects for the future. *Journal of internal medicine*, 284(6):643–663, 2018.

[14] ADNI Alzheimer's disease neuroimaging initiative. Accessed: 2021-05-21.

[15] J. Avila, J. J. Lucas, M. Perez, and F. Hernandez. Role of tau protein in both physiological and pathological conditions. *Physiological reviews*, 84(2):361–384, 2004.

[16] L. Crews and E. Masliah. Molecular mechanisms of neurodegeneration in Alzheimer's disease. *Human molecular genetics*, 19(R1):R12–R20, 2010.

[17] J. Albright, A. D. N. Initiative, et al. Forecasting the progression of Alzheimer's disease using neural networks and a novel preprocessing algorithm. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 5:483–491, 2019.

[18] L. Zhang, M. Wang, M. Liu, and D. Zhang. A survey on deep learning for neuroimaging-based brain disorder analysis. *Frontiers in neuroscience*, 14, 2020.

[19] F. E.-Z. A. El-Gamal, M. M. Elmogy, M. Ghazal, A. Atwan, M. F. Casanova, G. N. Barnes, A. S. El-Baz, and H. Hajjdiab. Medical imaging diagnosis of early Alzheimer's disease. *Frontiers in bioscience (Landmark edition)*, 23:671–725, 2018.

[20] M. Weiner and Z. Khachaturian. The use of MRI and PET for clinical diagnosis of dementia and investigation of cognitive impairment: a consensus report. *Alzheimer's Assoc Chicago, IL*, 1:1–15, 2005.

[21] J. E. Park, J. Yun, S. J. Kim, W. H. Shim, J. S. Oh, M. Oh, J. H. Roh, S. W. Seo, S. J. Oh, and J. S. Kim. Intra-individual correlations between quantitative THK-5351 PET and MRI-derived cortical volume in Alzheimer's disease differ according to disease severity and amyloid positivity. *PloS one*, 14(12), 2019.

[22] L. Rice and S. Bisdas. The diagnostic value of FDG and amyloid PET in Alzheimer's disease — a systematic review. *European journal of radiology*, 94:16–24, 2017.

[23] P. E. Khoonsari, A. Häggmark, M. Lönnberg, M. Mikus, L. Kilander, L. Lannfelt, J. Bergquist, M. Ingelsson, P. Nilsson, K. Kultima, et al. Analysis of the cerebrospinal fluid proteome in Alzheimer's disease. *PloS one*, 11(3), 2016.

[24] O. Hansson, S. Lehmann, M. Otto, H. Zetterberg, and P. Lewczuk. Advantages and disadvantages of the use of the csf amyloid β (aβ) 42/40 ratio in the diagnosis of Alzheimer's disease. *Alzheimer's research & therapy*, 11(1):1–15, 2019.

[25] J.-H. Kang, M. Korecka, J. B. Toledo, J. Q. Trojanowski, and L. M. Shaw. Clinical utility and analytical challenges in measurement of cerebrospinal fluid amyloid-β1–42 and τ proteins as Alzheimer disease biomarkers. *Clinical chemistry*, 59(6):903–916, 2013.

[26] S. Balsis, J. F. Benge, D. A. Lowe, L. Geraci, and R. S. Doody. How do scores on the ADAS-Cog, MMSE, and CDR-SOB correspond? *The Clinical Neuropsychologist*, 29(7):1002–1009, 2015.

[27] A. Chopra, T. A. Cavalieri, and D. J. Libon. Dementia screening tools for the primary care physician. *Clinical Geriatrics*, 15(1):38, 2007.

[28] S. Liu, S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, D. Feng, M. J. Fulham, et al. Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. *IEEE Transactions on Biomedical Engineering*, 62(4):1132–1140, 2014.

[29] J. Alzubi, A. Nayyar, and A. Kumar. Machine learning from theory to algorithms: an overview. *Journal of physics: conference series*, 1142(1):012012, 2018.

[30] F. Chollet. *Deep learning with Python*. Simon and Schuster, 2017.

[31] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[32] S. Ruder. An overview of gradient descent optimization algorithms. *arXiv:1609.04747*, 2016.

[33] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, O. Colliot, et al. Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical image analysis*, 63:101694, 2020.

[34] S. Khan, H. Rahmani, S. A. A. Shah, and M. Bennamoun. A guide to convolutional neural networks for computer vision. *Synthesis Lectures on Computer Vision*, 8(1):1–207, 2018.

[35] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[36] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[38] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. pages 675–678, 2014.

[39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[40] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[41] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[42] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[43] K. Bäckström, M. Nazari, I. Y.-H. Gu, and A. S. Jakola. An efficient 3D deep convolutional network for Alzheimer's disease diagnosis using mr images. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 149–153, 2018.

[44] D. Cheng and M. Liu. Classification of Alzheimer's disease by cascaded convolutional neural networks using PET images. *International Workshop on Machine Learning in Medical Imaging*, pages 106–113, 2017.

[45] S. Basaia, F. Agosta, L. Wagner, E. Canu, G. Magnani, R. Santangelo, M. Filippi, A. D. N. Initiative, et al. Automated classification of Alzheimer's disease and mild cognitive impairment using a single mri and deep neural networks. *NeuroImage: Clinical*, 21, 2019.

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

[47] S. Esmaeilzadeh, D. I. Belivanis, K. M. Pohl, and E. Adeli. End-to-end Alzheimer's disease diagnosis and biomarker identification. *International Workshop on Machine Learning in Medical Imaging*, pages 337–345, 2018.

[48] H. Karasawa, C.-L. Liu, and H. Ohwada. Deep 3D convolutional neural network architectures for Alzheimer's disease diagnosis. *Asian conference on intelligent information and database systems*, pages 287–296, 2018.

[49] H. Tang, E. Yao, G. Tan, and X. Guo. A fast and accurate 3D fine-tuning convolutional neural network for Alzheimer's disease diagnosis. *International CCF Conference on Artificial Intelligence*, pages 115–126, 2018.

[50] H. Choi, K. H. Jin, A. D. N. Initiative, et al. Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. *Behavioural brain research*, 344:103–109, 2018.

[51] A. M. Taqi, A. Awad, F. Al-Azzo, and M. Milanova. The impact of multi-optimizers and data augmentation on Tensorflow convolutional neural network performance. *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 140–145, 2018.

[52] J. Qiao, Y. Lv, C. Cao, Z. Wang, and A. Li. Multivariate deep learning classification of Alzheimer's disease based on hierarchical partner matching independent component analysis. *Frontiers in aging neuroscience*, 10:417, 2018.

[53] V. Wegmayr and D. Haziza. Alzheimer classification with MR images: Exploration of CNN performance factors. 2018.

[54] W. Lin, T. Tong, Q. Gao, D. Guo, X. Du, Y. Yang, G. Guo, M. Xiao, M. Du, X. Qu, et al. Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment. *Frontiers in neuroscience*, 12:777, 2018.

[55] G. Awate, S. Bangare, G. Pradeepini, and S. Patil. Detection of Alzheimers disease from mri using convolutional neural network with tensorflow. *arXiv preprint arXiv:1806.10170*, 2018.

[56] D. Cheng and M. Liu. Combining convolutional and recurrent neural networks for Alzheimer's disease diagnosis using PET images. *2017 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 1–5, 2017.

[57] Y. Kazemi and S. Houghten. A deep learning pipeline to classify different stages of Alzheimer's disease from fmri data. *2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–8, 2018.

[58] M. Liu, D. Cheng, W. Yan, A. D. N. Initiative, et al. Classification of Alzheimer's disease by combination of convolutional and recurrent neural networks using FDG-PET images. *Frontiers in neuroinformatics*, 12:35, 2018.

[59] J. G. Sled, A. P. Zijdenbos, and A. C. Evans. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE transactions on medical imaging*, 17(1):87–97, 1998.

[60] T. Shaikhina and N. A. Khovanova. Handling limited datasets with neural networks in medical applications: A small-data approach. *Artificial intelligence in medicine*, 75:51–63, 2017.

[61] Z. Hussain, F. Gimenez, D. Yi, and D. Rubin. Differential data augmentation techniques for medical imaging classification tasks. *AMIA Annual Symposium Proceedings*, page 979, 2017.

[62] S. Afzal, M. Maqsood, F. Nazir, U. Khan, F. Aadil, K. M. Awan, I. Mehmood, and O.-Y. Song. A data augmentation-based framework to handle class imbalance problem for Alzheimer's stage detection. *IEEE Access*, 7:115528–115539, 2019.

[63] A. Mikołajczyk and M. Grochowski. Data augmentation for improving deep learning in image classification problem. *2018 international interdisciplinary PhD workshop (IIPhDW)*, pages 117–122, 2018.

[64] B. Ljubic, S. Roychoudhury, X. H. Cao, M. Pavlovski, S. Obradovic, R. Nair, L. Glass, and Z. Obradovic. Influence of medical domain knowledge on deep learning for Alzheimer's disease prediction. *Computer methods and programs in biomedicine*, 197:105765, 2020.

[65] S. Qiu, G. H. Chang, M. Panagia, D. M. Gopal, R. Au, and V. B. Kolachalama. Fusion of deep learning models of mri scans, mini–mental state examination, and logical memory test enhances diagnosis of mild cognitive impairment. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:737–749, 2018.

[66] O. Hadad, R. Bakalo, R. Ben-Ari, S. Hashoul, and G. Amit. Classification of breast lesions using cross-modal deep learning. *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 109–112, 2017.

[67] S. Azizi, P. Mousavi, P. Yan, A. Tahmasebi, J. T. Kwak, S. Xu, B. Turkbey, P. Choyke, P. Pinto, B. Wood, et al. Transfer learning from RF to B-mode temporal enhanced ultrasound features for prostate cancer detection. *International journal of computer assisted radiology and surgery*, 12 (7):1111–1121, 2017.

[68] R. K. Samala, H.-P. Chan, L. M. Hadjiiski, M. A. Helvie, K. H. Cha, and C. D. Richter. Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms. *Physics in Medicine & Biology*, 62(23):8894, 2017.

[69] B. Lei, N. Cheng, A. F. Frangi, E.-L. Tan, J. Cao, P. Yang, A. Elazab, J. Du, Y. Xu, and T. Wang. Self-calibrated brain network estimation and joint non-convex multi-task learning for identification of early Alzheimer's disease. *Medical image analysis*, 61:101652, 2020.

[70] M. Liu, J. Zhang, E. Adeli, and D. Shen. Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis. *IEEE Transactions on Biomedical Engineering*, 66(5):1195–1206, 2018.

[71] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Scientific reports*, 11(1):1–13, 2021.

[72] C. Haarburger, M. Baumgartner, D. Truhn, M. Broeckmann, H. Schneider, S. Schrading, C. Kuhl, and D. Merhof. Multi scale curriculum CNN for context-aware breast MRI malignancy classification. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, page 495–503, 2019.

[73] Y. Tang, X. Wang, A. P. Harrison, L. Lu, J. Xiao, and R. M. Summers. Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. *International Workshop on Machine Learning in Medical Imaging*, pages 249–258, 2018.

[74] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. *arXiv preprint arXiv:1801.09927*, 2018.

[75] I. Gonzalez-Diaz. Dermaknet: Incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis. *IEEE journal of biomedical and health informatics*, 23 (2):547–559, 2018.

[76] L. Li, M. Xu, X. Wang, L. Jiang, and H. Liu. Attention based glaucoma detection: A large-scale database and CNN model. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10571–10580, 2019.

[77] D. Jin, J. Xu, K. Zhao, F. Hu, Z. Yang, B. Liu, T. Jiang, and Y. Liu. Attention-based 3D convolutional network for Alzheimer's disease diagnosis and biomarkers exploration. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1047–1051, 2019.

[78] J. Zhang, B. Zheng, A. Gao, X. Feng, D. Liang, and X. Long. A 3d densely connected convolution neural network with connection-wise attention mechanism for Alzheimer's disease classification. *Magnetic resonance imaging*, 78:119–126, 2021.

[79] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058, 2018.

[80] Z. Zhang, P. Chen, M. Sapkota, and L. Yang. Tandemnet: Distilling knowledge from medical images using diagnostic reports as optional semantic references. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 320–328, 2017.

[81] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? *arXiv preprint arXiv:1411.1792*, 2014.

[82] M. Hon and N. M. Khan. Towards Alzheimer's disease classification through transfer learning. *2017 IEEE International conference on bioinformatics and biomedicine (BIBM)*, pages 1166–1169, 2017.

[83] M. Maqsood, F. Nazir, U. Khan, F. Aadil, H. Jamal, I. Mehmood, and O.-y. Song. Transfer learning assisted classification and detection of Alzheimer's disease stages using 3D MRI scans. *Sensors*, 19(11):2645, 2019.

[84] A. Ebrahimi-Ghahnavieh, S. Luo, and R. Chiong. Transfer learning for Alzheimer's disease detection on MRI images. *2019 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, pages 133–138, 2019.

[85] Q. Liao, Y. Ding, Z. L. Jiang, X. Wang, C. Zhang, and Q. Zhang. Multi-task deep convolutional neural network for cancer diagnosis. *Neurocomputing*, 348:66–73, 2019.

[86] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[87] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[88] C. Qin, D. Yao, Y. Shi, and Z. Song. Computer-aided detection in chest radiography based on artificial intelligence: a survey. *Biomedical engineering online*, 17(1):1–23, 2018.

[89] S. Chan, V. Reddy, B. Myers, Q. Thibodeaux, N. Brownstone, and W. Liao. Machine learning in dermatology: current applications, opportunities, and limitations. *Dermatology and therapy*, 10(3): 365–386, 2020.

[90] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen. Medical image classification with convolutional neural network. *2014 13th international conference on control automation robotics & vision (ICARCV)*, pages 844–848, 2014.

[91] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[92] T. Gong, C. L. Tan, T. Y. Leong, C. K. Lee, B. C. Pang, C. T. Lim, Q. Tian, S. Tang, and Z. Zhang. Text mining in radiology reports. *2008 Eighth IEEE International Conference on Data Mining*, pages 815–820, 2008.

[93] A. Vatian, N. Gusarova, N. Dobrenko, A. Klochkov, N. Nigmatullin, A. Lobantsev, and A. Shalyto. Fusing of medical images and reports in diagnostics of brain diseases. *Proceedings of the 2019 the International Conference on Pattern Recognition and Artificial Intelligence*, pages 102–108, 2019.

[94] I. Banerjee, Y. Ling, M. C. Chen, S. A. Hasan, C. P. Langlotz, N. Moradzadeh, B. Chapman, T. Amrhein, D. Mong, D. L. Rubin, et al. Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. *Artificial intelligence in medicine*, 97:79–88, 2019.

[95] Z. Yuan, Y. Jiang, J. Li, and H. Huang. Hybrid-DNNs: Hybrid deep neural networks for mixed inputs. *arXiv preprint arXiv:2005.08419*, 2020.

[96] G. Hacohen and D. Weinshall. On the power of curriculum learning in training deep networks. *arXiv*, 2019.

[97] A. Graves, M. G. Bellemare, J. Menick, R. Munos, and K. Kavukcuoglu. Automated curriculum learning for neural networks. *International conference on machine learning*, pages 1311–1320, 2017.

[98] G. Maicas, A. P. Bradley, J. C. Nascimento, I. Reid, and G. Carneiro. Training medical image analysis systems like radiologists. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019.

[99] A. Jiménez-Sánchez, D. Mateus, S. Kirchhoff, C. Kirchhoff, P. Biberthaler, N. Navab, M. A. G. Ballester, and G. Piella. Medical-based deep curriculum learning for improved fracture classification. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 694–702, 2019.

[100] J. Wei, A. Suriawinata, B. Ren, X. Liu, M. Lisovsky, L. Vaickus, C. Brown, M. Baker, M. Nasir-Moin, N. Tomita, L. Torresani, J. Wei, and S. Hassanpour. Learn like a pathologist: Curriculum learning by annotator agreement for histopathology image classification. *arXiv*, 2020.

[101] W. Lotter, G. Sorensen, and D. Cox. A multi-scale CNN and curriculum learning strategy for mammogram classification. *arXiv*, 2017.

[102] B. Park, Y. Cho, G. Lee, S. M. Lee, Y.-H. Cho, E. S. Lee, K. H. Lee, J. B. Seo, and N. Kim. A curriculum learning strategy to enhance the accuracy of classification of various lesions in chest-PA X-ray screening for pulmonary abnormalities. *Nature*, 2019. doi: 10.1038/s41598-019-51832-3.

[103] A. Jesson, N. Guizard, S. H. Ghalehjegh, D. Goblot, F. Soudan, and N. Chapados. CASED: curriculum adaptive sampling for extreme data imbalance. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 639–646, 2017.

[104] S. K. Asare, F. You, and O. T. Nartey. A semisupervised learning scheme with self-paced learning for classifying breast cancer histopathological images. *Computational Intelligence and Neuroscience*, 2020, 2020.

[105] W. Wang, Y. Lu, B. Wu, T. Chen, D. Z. Chen, and J. Wu. Deep active self-paced learning for accurate pulmonary nodule segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 723–731, 2018.

[106] Z. Rongchang and S. Li. EGDCL: An adaptive curriculum learning framework for unbiased glaucoma diagnosis. *Proceedings of the Computer Vision—ECCV*, 07 2020.

[107] I. Oksuz, B. Ruijsink, E. Puyol-Antón, J. R. Clough, G. Cruz, A. Bustin, C. Prieto, R. Botnar, D. Rueckert, J. A. Schnabel, et al. Automatic CNN-based detection of cardiac MR motion artefacts using k-space data augmentation and curriculum learning. *Medical image analysis*, 55:136–147, 2019.

[108] M. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. *Advances in neural information processing systems*, 23:1189–1197, 2010.

[109] X. Wang, Y. Chen, and W. Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[110] L. Berger, H. Eoin, M. J. Cardoso, and S. Ourselin. An adaptive sampling scheme to efficiently train fully convolutional networks for semantic segmentation. *Annual Conference on Medical Image Understanding and Analysis*, pages 277–286, 2018.

[111] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann. Self-paced curriculum learning. *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[112] R. Portelas, C. Colas, L. Weng, K. Hofmann, and P.-Y. Oudeyer. Automatic curriculum learning for deep RL: A short survey. *arXiv preprint arXiv:2003.04664*, 2020.

[113] N. Milano and S. Nolfi. Automated curriculum learning for embodied agents a neuroevolutionary approach. *Scientific Reports*, 11(1):1–14, 2021.

[114] T. Matiisen, A. Oliver, T. Cohen, and J. Schulman. Teacher-student curriculum learning. *IEEE transactions on neural networks and learning systems*, 31(9):3732–3740, 2019.

[115] P. M. T. Pereira. *Transfer Learning methods for Alzheimer's Disease Diagnosis*. Masters dissertation in electrical and computer engineering, Instituto Superior Técnico, Universidade Técnica de Lisboa, 2021.

[116] M. Grassi, D. A. Loewenstein, D. Caldirola, K. Schruers, R. Duara, and G. Perna. A clinically-translatable machine learning algorithm for the prediction of Alzheimer's disease conversion: further evidence of its accuracy via a transfer learning approach. *International psychogeriatrics*, 31 (7):937–945, 2019.

[117] P. Wenig. Self-paced learning for machine learning. `https://towardsdatascience.com/self-paced-learning-for-machine-learning-f1c489316c61`, 15 Feb. 2020.

[118] K. Ghasedi, X. Wang, C. Deng, and H. Huang. Balanced self-paced learning for generative adversarial clustering network. pages 4391–4400, 2019.

[119] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[120] Y. Zhao, B. Ma, P. Jiang, D. Zeng, X. Wang, and S. Li. Prediction of Alzheimer's disease progression with multi-information generative adversarial network. *IEEE Journal of Biomedical and Health Informatics*, 25(3):711–719, 2020.

[121] J. M. Rondina, L. K. Ferreira, F. L. de Souza Duran, R. Kubo, C. R. Ono, C. C. Leite, J. Smid, R. Nitrini, C. A. Buchpiguel, and G. F. Busatto. Selecting the most relevant brain regions to discriminate Alzheimer's disease patients from healthy controls using multiple kernel learning: A comparison across functional and structural imaging modalities and atlases. *NeuroImage: Clinical*, 17:628–641, 2018.

[122] T. Yokoi, H. Watanabe, H. Yamaguchi, E. Bagarinao, M. Masuda, K. Imai, A. Ogura, R. Ohdake, K. Kawabata, K. Hara, et al. Involvement of the precuneus/posterior cingulate cortex is significant for the development of Alzheimer's disease: a PET (THK5351, PiB) and resting fMRI study. *Frontiers in aging neuroscience*, 10:304, 2018.

[123] L. V. Hiscox, C. L. Johnson, M. D. McGarry, H. Marshall, C. W. Ritchie, E. J. Van Beek, N. Roberts, and J. M. Starr. Mechanical property alterations across the cerebral cortex due to Alzheimer's disease. *Brain communications*, 2(1):fcz049, 2020.