Diagnosis of Neurodegenerative Diseases using Deep Learning

Ekin Yagis

A thesis presented for the degree of

Doctor of Philosophy (Ph.D)

School of Computer Science & Electronic Engineering

University of Essex

Colchester, United Kingdom

Date of submission April 2022

To my parents Gülay and Serdar

for their endless love and support

Abstract

Automated disease classification systems can assist radiologists by reducing workload while initiating therapy to slow disease progression and improve patients' quality of life. With significant advances in machine learning (ML) and medical scanning over the last decade, medical image analysis has experienced a paradigm change. Deep learning (DL) employing magnetic resonance imaging (MRI) has become a prominent method for computer-assisted systems because of its ability to extract high-level features via local connection, weight sharing, and spatial invariance. Nonetheless, there are several important research challenges when advancing toward clinical application, and these problems inspire the contributions presented throughout this thesis.

This research develops a framework for the classification of neurodegenerative diseases using DL techniques and MRI. The presented thesis involves three evolution stages. The first stage is the development of a robust and reproducible 2D classification system with high generalisation performance for Alzheimer's disease (AD), mild cognitive impairment (MCI), and Parkinson's disease (PD) using deep convolutional neural networks (CNN).

The next phase of the first stage extends this framework and demonstrates its use on different datasets while quantifying the effect of a highly observed phenomenon called data leakage in the literature. Key contributions of the thesis presented in this stage are a thorough analysis of the literature, a discussion on the potential flaws of the selected studies, and the development of an open-source evaluation system for neurodegenerative disease classification using structural MRI. The second stage aims to overcome the problems stem from investigating 3D data with 2D models. With this goal, a 3D CNN-based diagnostic framework is developed for classifying AD and PD patients from healthy controls using T_1-weighted brain MRI data. The last stage includes two phases with a focus on AD and MCI diagnosis. The first phase proposes a new autoencoder-based deep neural network structure by integrating supervised prediction and unsupervised representation. The second phase introduces the final contribution of the thesis which is a novel ensemble approach that may also be used to predict diseases other than neurodegenerative ones (e.g., tuberculosis (TB)) using a modality apart from MRI.

Publications

The following publications and manuscripts were a result of work conducted during this doctoral study:

- Yagis, E., García Seco de Herrera, A., Abolghasemi V., Andritsch J., Pinpo N., Chaichulee S., Citi, L. (2021). Mild Cognitive Impairment Diagnosis from Brain MRI using an Ensemble Deep Learning Method in Physics in Medicine Biology - IOPscience 2022. (Under preparation).

- Yagis, E., García Seco de Herrera, A., Abolghasemi V., Andritsch J., Pinpo N., Chaichulee S. (2021). Ensemble Deep Learning Architectures for Automated Diagnosis of Pulmonary Tuberculosis using Chest X-ray in IEEE Journal of Biomedical and Health Informatics. (Submitted).

- Yagis, E., Atnafu, S.W., García Seco de Herrera, A. et al. Effect of data leakage in brain MRI classification using 2D convolutional neural networks. Sci Rep 11, 22544 (2021). https://doi.org/10.1038/s41598-021-01681-w

- Yagis, E., García Seco de Herrera, A., & Citi, L. (2021). Convolutional Autoencoder based Deep Learning Approach for Alzheimer's Disease Diagnosis using Brain MRI in IEEE 34rd International Symposium on Computer-Based Medical Systems (CBMS).

- Yagis, E., Citi, L., Diciotti, S., Merzi, C., Workalemahu Atnafu, S., & Garcia Seco De Herrera, A. (2020). 3D Convolutional Neural Networks for Diagnosis of Alzheimer's Disease via structural MRI in IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Rochester, MN, USA, 2020 pp. 65-70. doi: 10.1109/CBMS49503.2020.00020

- Yagis, E., García Seco de Herrera, A., & Citi, L. (2019, November). Generalization Performance of Deep Learning Models in Neurodegenerative Disease Classification. In 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).

Acknowledgements

I would like to express my deep gratitude to my supervisors, Dr. Alba Garcia Seco de Herrera and Prof. Luca Citi, for their trust, guidance, and encouragement. I am grateful for their unique supervision throughout my doctoral chapter. You have set an example of excellence as a researcher, mentor, and instructor.

Next, I would like to thank my thesis committee members, Dr. Maria Kyropoulou and Dr. Borzoo Rassouli, and my viva examiners, Dr. Haider Raza and Dr. Juan Domingo Gispert López, for all of their interest in my work; your discussion, ideas, and feedback have been absolutely invaluable.

My sincere thanks also go to our collaborators, Prof. Stefano Diciotti, Dr. Chiara Marzi, Selamawet Workalemahu Atnafu, Dr. Vahid Abolghasemi, Dr. Jarutas Andritsch, and finally Dr. Sitthichok Chaichulee, for sharing their insight, constructive feedback, and valuable experience and time.

I'm also grateful to my dearest friends in London and Istanbul for their encouragement and company along the way.

I would especially like to thank my amazing family for the endless love, support, and constant inspiration they have given me from the beginning of my life. They are the greatest luck that I have in this world. And to my best friend, my playmate and my lifelong partner Arın, nothing would have been the same without you.

Finally, longing for a new journey, I hope I never lose my curiosity, and always have the courage to be myself.

Contents

List of Figures

List of Tables

List of Acronyms

AD	Alzheimer's disease.
ADNI	Alzheimer's Disease Neuroimaging Initiative.
AIBL	Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing.
CAE	Convolutional autoencoder.
CDR	Clinical dementia rating.
CN	Cognitively normal.
CNN	Convolutional neural network.
CSF	Cerebrospinal fluid.
CV	Cross-validation.
DL	Deep learning.
DT	Decision tree.
fMRI	Functional magnetic resonance imaging.
GM	Gray matter.
GPU	graphical processing unit.

ILSVRC ImageNet Large Scale Visual Recognition Challenge.

LR Logistic regression.

MAPT Microtubule-associated protein tau.

MCI Mild cognitive impairment.

ML machine learning.

MMSE Mini-mental state examination.

MRI Magnetic resonance imaging.

MSE mean squared error.

OASIS Open Access Series of Imaging Studies.

PCA principle component analysis.

PET Positron emission tomography.

pMCI Progressive mild cognitive impairment.

ReLU rectified linear unit.

resNet Residual neural network.

RF Random forest.

ROI Region of interest.

SD Standard deviation.

SGD Stochastic gradient descent.

sMCI Stable mild cognitive impairment.

SNR Signal-to-noise ratio.

SVM Support vector machine.

T1wMRI T1-weighted magnetic resonance imaging.

WM White matter.

Chapter 1

Introduction

"I live my life in widening circles that reach out across the world."

- Rainer Maria Rilke, *Rilke's Book of Hours*

A brief introduction to this doctoral dissertation is provided in this chapter. It begins by presenting the motivation for this research. Then, a description of the outline of this dissertation is given. Finally, the major scientific accomplishments of this chapter are stated.

1.1 Motivations

Neurodegenerative disease is a term that refers to a heterogeneous group of disorders characterised by the progressive and irreversible degeneration of cells in the nervous system. Neurodegenerative diseases could cause problems related to movement (called

ataxias) or mental functioning (called dementias). While some of the physical or cognitive signs associated with neurodegenerative disorders can be eased with therapy, there is currently no proven cure for common neurodegenerative diseases, including Alzheimer's (AD), Parkinson's (PD), and Huntington's disease.

One of the critical challenges in neurodegenerative disease research is that the accurate and early diagnosis of disorders is difficult. Because the course of the diseases typically begins several years before symptoms such as dementia and ataxia occur, it is of considerable significance to predict whether diseases will develop in a given subject as early as possible. For instance, AD patients usually experience diagnostic symptoms at later stages after irreversible neural damage occurs. Therefore, early detection of AD is crucial to starting treatments to decelerate the progress of the disease and maximize patients' quality of life.

Early detection and biomarker identification could result in delays in disease progression, identifying treatable symptoms, initiating more effective pharmacological therapies, and early psychosocial care organizations.

As the disease progresses, the structure of the brain undergoes some changes, such as the shrinkage of the cerebral cortex and hippocampus and the expansion of ventricles [8, 9]. Through numerous medical imaging techniques such as MRI, PET, and computed tomography (CT), some of these changes can be detected earlier. Notably, a T_1-weighted MRI scan of the brain reveals high-resolution structural information of the brain and can

be used to identify atrophic changes in the temporal lobes [10].

With the rapid advances in machine learning (ML) and scanning, early detection of neurodegenerative diseases may be possible via computer-assisted systems using neuroimaging data. Among all these, deep learning (DL) utilising MRI has become a prominent tool due to its capability to extract high-level features through local connectivity, weight sharing, and spatial invariance.

This Ph.D. thesis is funded by the School of Computer Science and Electronic Engineering (CSEE) doctoral scholarship. Furthermore, I'd also like to acknowledge that the work presented in Chapter 7 was partially supported by the University of Essex GCRF QR Engagement Fund provided by Research England (Grant number G026).

1.2 Thesis overview

This thesis is composed of eight chapters. The contents of each chapter are illustrated as follows.

- **Chapter 1** provides an introduction of the thesis, summarises the motivation behind the work conducted, and presents the research significance. It also introduces the organisation of the thesis and lists the scientific contributions.
- **Chapter 2** introduces the background information related to this thesis. This covers: i) pathogenesis and epidemiology of the most common neurodegenerative diseases:

AD and PD together with mild cognitive impairement (MCI), ii) neuroimaging techniques, iii) data pre-processing procedures iv) machine learning (ML) approaches (i.e., conventional techniques and deep learning (DL) methods), and v) datasets used in the thesis.

- **Chapter 3** details the design of AD and PD classification experiments using 2D convolutional neural networks (CNN). It includes image pre-processing methods as well as 2D models and transfer learning strategies. The impact of erroneous cross-validation on model performance is also presented.

- **Chapter 4** extends the previous chapter by quantifying the extent of the overestimation of classification accuracy in the case of incorrect slice-level cross-validation. Finally, it presents the true performance of 2D CNN models trained with subject-level and slice-level CV data split for the classification of AD and PD patients and aims to clarify data leakage problems in the literature.

- **Chapter 5** demonstrates our 3D CNN-based diagnostic framework for classifying AD patients from healthy controls using T_1-weighted brain magnetic resonance imaging (MRI) data.

- **Chapter 6** introduces a new autoencoder-based deep neural network structure by integrating supervised prediction and unsupervised representation for AD diagnosis.

- **Chapter 7** presents a novel ensemble DL method for automated diagnosis by

combining a multi-scale CNN and a convolutional autoencoder. In this chapter, the model has been tested on another modality than MRI and used to predict pulmonary tuberculosis as well as AD and MCI.

- **Chapter 8** completes this dissertation by reviewing the primary goals and outlining the contributions. Finally, it recalls the key results and presents future research directions.

1.3 Scientific contributions of the thesis

The main contributions of this thesis are the following:

- The development and release of a framework for 2D CNN based classification of AD/MCI and PD using T_1-weighted brain MRI data [11, 12] (see Ch.3 and Ch.4).
- The conduction of an exhaustive literature survey and the review the potential flaws in various studies in the literature [12] (see Ch.4).
- A quantitative assessment of the effect of data leakage caused by the adoption of incorrect slice-level cross-validation, rather than subject-level, using three 2D CNN models for the classification of patients with AD and PD [12] (see Ch.3 and Ch.4).
- The implementation of volumetric CNN-based approach for the diagnosis of AD and the visualisation of the spatial attention of CNN's decision [13] (see Ch.5).

- The implementation of a novel ensemble method which has a potential to be used clinically on not only neurodegenerative diseases but also pulmonary tuberculosis (see Ch.7).

Chapter 2

Background

This chapter presents some background for the research presented in this thesis. This includes the pathogenesis and epidemiology of the most common neurodegenerative diseases, neuroimaging techniques and data pre-processing procedures, ML approaches, and datasets used in the thesis. The subsets created from the given datasets for experimentations in the thesis are not covered in this chapter but rather presented in Chapters 3, 4, 5, and 6. Furthermore, each chapter will have a full literature review based on the approach of the chapter.

2.1 Neurodegenerative diseases

Neurodegenerative disorders are defined by the progressive loss of structure or function of neurons [14]. Amyotrophic lateral sclerosis, multiple sclerosis, PD, AD, Huntington's

disease, multiple system atrophy, and prion disorders are examples of neurodegenerative diseases. These disorders are deemed incurable since there is no known technique to reverse the progressive degeneration of neurons [15]. The three diseases discussed in this thesis, namely AD, MCI and PD are briefly introduced in this section.

2.1.1 Alzheimer's disease

AD is a neurodegenerative illness characterised by extracellular plaques containing β-amyloid (βA) and intracellular neurofibrillary tangles containing tau. In healthy neurons, tau protein normally stabilises the microtubules [16]. However, abnormal changes in brain chemistry cause tau protein molecules to detach from microtubules and form neurofibrillary tangles destroying the brain cells' ability to communicate with other cells [17]. The most frequent symptom of AD is trouble with short-term memory, although impairment in expressive speech, visuospatial processing, and executive (mental agility) skills also occurs [18]. Patients usually experience diagnostic symptoms at later stages after irreversible neural damage occurs.

According to several recent research, AD can begin 20 years or more before symptoms occur and the condition is clinically diagnosed [2, 19–22]. Only after a certain stage, patients may experience diagnostic symptoms such as deterioration in memory and decline in cognitive abilities when the irreversible neurological damage already occurs. Therefore, early and accurate diagnosis of AD is crucial and may be possible via computer-assisted

analytical techniques. Receiving an early diagnosis of AD allows individuals to take advantage of numerous therapies, plan their future, and improve their life quality.

2.1.1.1 Epidemiology

The global frequency of all-cause dementia is estimated to rise from 57.4 million in 2019 to 152.8 million by 2050 [23]. According to studies employing MRI and positron emission tomography (PET) to assess the burden of AD, MCI with AD pathology accounts for 50% of all instances of MCI, while dementia attributable to AD accounts for 60%–90% of all dementia cases [24, 25]. It is predicted that by 2050, half (51%) of all people 65 and older will be facing AD [26].

The major risk factor for both dementia and AD is age [27]. When it comes to genetic risk factors, scientists have discovered evidence of a relationship between AD and genes on four chromosomes, designated 1, 14, 19, and 21. The APOE gene, which is found on chromosome 19, has been linked to late-onset AD, which is the most frequent form of the disease in adults over the age of 65 [28]. Carrying a specific type of APOE gene called the APOE ϵ 4 allele raises the risk of dementia by 3–4 times in heterozygotes (prevalence 25%) and 12–15 times in homozygotes (prevalence 2%) compared to carrying APOE ϵ 3 [29,30]. On the other hand, in the case of early-onset AD, almost all instances of dominantly inherited AD are caused by mutations in the APP (amyloid precursor protein), PSEN1 (presenilin 1), and PSEN2 (presenilin 2) genes. When people with these gene abnormalities exhibit symptoms, they are

almost always under 65 years old [18].

2.1.1.2 Pathogenesis

The pathophysiology of AD can be interpreted as positive ('overt') lesions visible under a microscope, such as tau-containing neurofibrillary tangles, β-amyloid (βA) containing plaques, activated glia, or expanded endosomes. Alternatively, AD might be considered as a negative ('covert') phenomenon, namely the loss of synaptic homeostasis, neurons, or neural network integrity [18].

The (βA) peptide is produced by the metabolism of amyloid precursor protein (APP), a 695–770 amino acid type I transmembrane glycoprotein [31]. An extracellular protease known as (α)-secretase cleaves APP near the membrane [32, 33]. The (βA) peptide's deposition in the brain is thought to be the first phase in the AD process, as summarised in Figure 2.1 [31, 34] [35, 36]. An accumulation occurs 15–20 years before clinical symptoms appear, owing to a peptide elimination problem in the brain (see Figure 2.2) [37].

Earlier research implicated (βA) fibrils as the neurotoxic factor causing cellular death, memory loss, and other AD symptoms [38, 39]. However, additional research over the last two decades has revealed that oligomeric or prefibrillar forms of the (βA) peptide are the most toxic to neuronal cells [39–41].

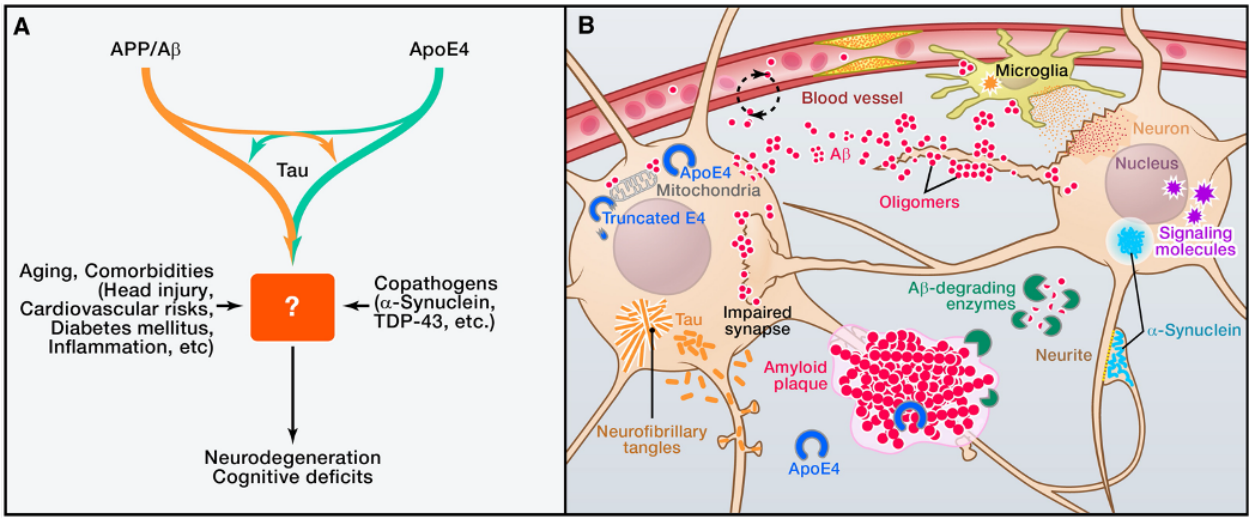


Figure 2.1: Pathogenesis of Alzheimer's Disease. (A) Alzheimer's disease (AD) is most likely caused by copathogenic interactions between many variables, including APP/(βA), APOE 4, tau, a-synuclein, TDP-43, ageing, and other comorbidities. It is unclear how they interact to harm neuronal function and survival. (B) (βA) oligomers disrupt synaptic functioning and associated signalling pathways, altering neuronal activity and causing glial cells to produce neurotoxic mediators. Neuronal processes are displaced and distorted by fibrillar amyloid plaques. APOE 4 is a lipid transport protein that reduces (βA) clearance while promoting its deposition. Tau, which is typically found in axons, becomes mislocalised and forms aggregates termed neurofibrillary tangles in the soma and dendrites of neurons (NFTs). Self-assembly of (α)-synuclein into pathogenic oligomers and bigger aggregates is also possible (Lewy bodies). Reprinted from [1].

2.1.2 Mild Cognitive Impairment

MCI is a condition characterised as having memory concerns above what was anticipated for their age and who showed a minor memory impairment but did not significantly interfere with everyday activities [42]. It is the earliest symptomatic stage of cognitive impairment in which individuals retain the capacity to conduct most everyday tasks independently. As a result,

it differs from dementia, in which cognitive losses are more severe and extensive and affect daily life. Yet, moderate cognitive impairment with memory complaints and impairments (amnestic MCI) has been demonstrated to have a significant probability of progression to dementia, specifically Alzheimer's type dementia [42]. MCI is frequently indicated by a worldwide grade of 0.5 on the Clinical Dementia Rating (CDR) scale [43].

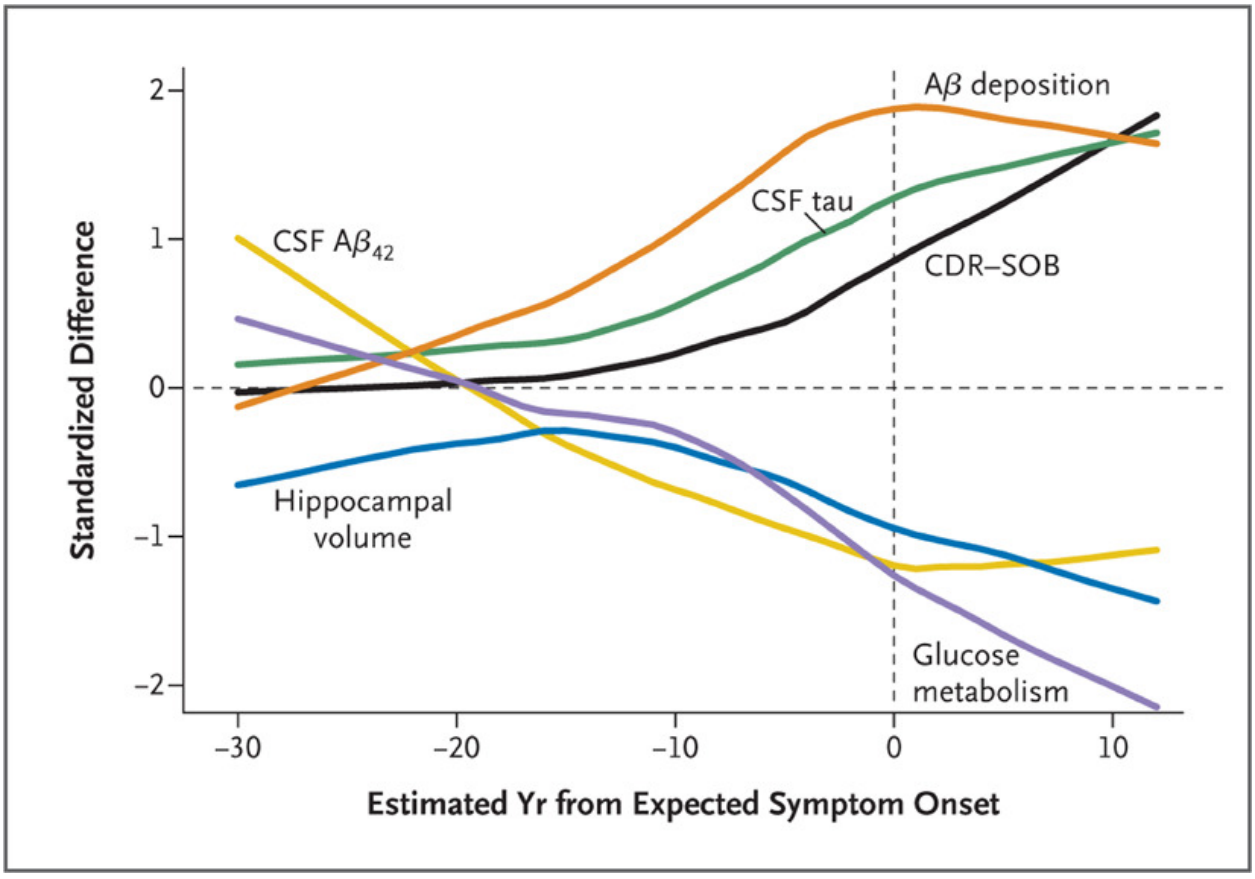


Figure 2.2: The amyloid cascade model proposed by Bateman et al. [2]. As time passes, Alzheimer's disease (AD) biomarkers grow increasingly aberrant, with amyloid buildup leading to greater tau pathology and neurodegeneration. Mild dementia (Clinical Dementia Rating (CDR) 1) appeared an average of 3.3 years before the predicted onset of symptoms. Reprinted from [2].

2.1.2.1 Epidemiology

The yearly conversion rate from MCI to AD has been found to be between 10% and 15% [44]. After six years of follow-up, roughly 80% of MCI patients will have changed to AD (MCI converters [MCI-C]). In contrast, other MCI patients will remain stable or convert back to normal (MCI non-converters [MCI-NC]) [44, 45].

There are no cures for those who already have AD, and current therapies can only slow the progression of the condition [46]. As a result, early diagnosis of MCI and prognosis prediction models are highly needed.

2.1.2.2 Pathogenesis

Though MCI was previously regarded as a transitional stage between normal ageing and AD dementia, AD pathology is only one of the numerous mechanisms that might contribute to MCI [47]. Cerebrovascular disease, psychiatric disease (particularly depression), and various non-AD neurodegenerative pathologies, such as frontotemporal lobar degeneration, Lewy body disease, limbic-predominant age-related TDP-43 encephalopathy (LATE), hippocampal sclerosis, primary age-related tauopathy (PART), and others, are other causes of MCI [48–52].

Even though the underlying pathogenesis of the condition is still mostly unknown, previous research suggested that weighted gene co-expression network analysis (WGCNA) might be used to investigate the relations between genes and clinical aspects of

neurodegenerative disorders [53]. Six genes have been identified as being involved in the pathological alterations associated with MCI and AD [54].

2.1.3 Parkinson's disease

PD is a neurological disorder caused by the progressive death of dopamine-producing cells in the brain [55, 56]. It is the second most common neurodegenerative disorder after AD, affecting around 2 to 3% of the population over the age of 65 [57]. An estimated 7 to 10 million people worldwide have been affected by PD and related disorders in 2018 [58]. The neuropathological hallmarks include neuronal loss in the substantia nigra, resulting in striatal dopamine insufficiency, and intracellular inclusions containing an aggregation of alpha-synuclein [59].

2.1.3.1 Epidemiology

The condition usually appears between the ages of 65 and 70. In population-based cohorts, onset before the age of 40 is seen in less than 5% of cases [60]. Men are slightly more likely than women to develop the condition [61, 62]. The reasons for the male preponderance in PD are unknown, although some theories include estrogen's protective effect in women, differences in gender-specific exposure to environmental risk factors, and genetic susceptibility genes on the sex chromosomes [62]. The disease's prevalence is estimated to be between 100 and 200 persons per 100,000, with an annual incidence of 15 people per

100,000. The Global Burden of Disease (GBD) chapter revealed that more than 10 million people worldwide were affected by PD in 2017 [63]. Unfortunately, the number of people with PD is expected to increase substantially by 2030 [64].

2.1.3.2 Pathogenesis

The presence of Lewy bodies, intracellular inclusions of aggregated α-synuclein, neuroinflammation, and degeneration of dopaminergic neurons in the substantia nigra are the pathological markers of PD [59, 65]. Although the cause and pathogenesis of selective dopamine neuron loss and α-synuclein accumulation remain unknown, growing lines of evidence from environmental risk factors and early-onset genetics point to a convergence between energy metabolism and protein disposal in the development of PD [65]. These findings imply that mitochondrial and ubiquitin-proteasome system dysfunction can play a crucial role in the etiology of PD [66].

2.2 Neuroimaging technique: Structural MRI

Structural MRI and functional MRI are two types of MRI. The former, structural MRI, is a common imaging technique in both research and clinical practice [67]. Functional imaging can be seen as a technology that gives dynamic physiological information, whereas structural imaging provides static anatomical information. BOLD (blood oxygen level dependent) method, perfusion (whether by endogenous or exogenous contrast), blood flow,

and cerebrospinal fluid (CSF) pulsation data are therefore included in fMRI [68].

The contrast between tissue compartments, mainly grey and white matter, is used in MRI assessments of brain anatomy [69]. MRI signal differs between tissue types in general because grey matter has more cell bodies (e.g., neurons and glial cells) than white matter, which is predominantly composed of long-distance nerve fibers (myelinated axons) and supporting glial cells. The so-called T1 relaxation period of hydrogen atoms in tissue is crucial for contrast in structural MRI [70]. T1 time is influenced by a variety of biomolecules. Recent research indicates that lipid content is especially important for T1 time in the brain and the ensuing contrast between grey and white matter [71, 72].

2.3 Medical image pre-processing

Pre-processing raw images is the initial step in most data-driven investigations. It is necessary for a quantitative analysis to be successful as it improves the effectiveness of the subsequent segmentation, feature extraction, or classification procedures.

MRI scans may contain a variety of artifacts. Pre-processing these images includes removing artifacts, modifying image resolution, and addressing contrast discrepancies caused by differing capture devices and settings [73].

Bias field correction, intensity rescaling, standardisation, skull stripping, and registration are generally included in the image preparation procedure [74]. Various causes of artifacts may be present depending on the data modalities and the scientific topic of interest; therefore,

different corrective strategies should be used [75, 76]. The most prevalent pre-processing approaches for T_1-weighted images, which are also employed in the thesis, will be presented in this part. These techniques will be discussed in further depth in later chapters based on the unique demands of each experimental method.

2.3.1 Bias field correction

A bias field signal is a low-frequency, very smooth signal that corrupts MRIs, particularly those produced by outdated MRI machines [77, 78]. Before submitting damaged MRIs to algorithms, a pre-processing step that involves bias field signal correction is required [79]. The non-parametric non-uniformity intensity normalisation (N3) technique, available in the Freesurfer software package, and the N4 algorithm provided in ITK 9 are two prominent ways for correcting these intensity inhomogeneities [80–82].

2.3.2 Intensity rescaling and standardisation

MRI scans are classified into two types: contrast and non-contrast [83]. That causes MR images to often have various intensity ranges and intensity distribution, potentially affecting image pre-processing stages [84]. Unfortunately, one of the fundamental drawbacks of MRI methods has been that intensities do not have a consistent meaning, even within the same protocol, for the same body area, for images taken on the same scanner, and for the same subject [84]. Scaling of the minimum to maximum intensity

range of the given image to a fixed standard range could be used to deal with different intensity ranges on a simple level [85]. On the other hand, histogram equalisation is mainly used for intensity standardisation [86].

2.3.3 Skull stripping

Image analysis algorithms may encounter difficulties when dealing with non-brain tissues [87]. Skull stripping is the procedure of separating brain tissue (cortex and cerebellum) from its surroundings (skull and nonbrain area) [88]. Since the CSF space and skull are dark in T_1-weighted images, the edges between the brain and the skull are well-marked; however, even strong edges may be unsettled due to finite resolution during MRI acquisition or the presence of other anatomical partial structures within the brain [89].

2.3.4 Image registration

Image registration is the process of spatially aligning two images using a set of geometric modifications so that voxels at corresponding locations contain comparable information across different scans/subjects regardless of their different anatomy or acquisition modality.

There are two kinds of registration algorithms based on transformation models: linear registration and non-linear registration [90]. Linear registration is commonly used and often consists of a six-parametric rigid transformation (rotation and translation on the x, y, and z coordinate axes) or a 12-parametric affine transformation (rotation, translation, scaling, and

shearing on x, y, and z coordinate axes) [77]. Non-linear registration has a higher degree of elasticity and can mimic local deformation than linear registration [91].

Image registration could align images of the same subject, images of different subjects, or be performed between an atlas and an individual patient [92]. The Talairach stereotactic space, which was created from a single post-mortem female brain, is the most often used atlas for image alignment [93]. The alignment is most usually performed using an affine transformation, but additional degrees of freedom are occasionally employed [94]. Another common brain spaces or templates that are used as spatial normalisation targets are the Montreal Neurological Institute (MNI) templates [95]. MNI templates are characterised by differences in origin, orientation, and larger dimensions and do not refer to the same brain structures as Talairach coordinates [96].

The Statistical Parametric Mapping (SPM) software package and the Advanced Normalisation Tools (ANTs) provide solutions for both linear and non-linear registration [97, 98].

2.4 Machine learning

ML is an area of artificial intelligence (AI) and computer science that focuses on utilising data and algorithms to replicate how humans learn [99]. The primary goal is for computers to learn autonomously without human involvement and then adapt their activities accordingly based on previous examples [100]. It is a multidisciplinary field allowing computers to speak

with humans, drive autonomously, track down suspects and detect cancer [101–104].

ML has three main learning paradigms: supervised and unsupervised learning and reinforcement learning [105]. Classification and regression are two major prediction problems in supervised learning [106]. In classification tasks, data points correspond to a limited number of categories [107]. In contrast, regression tasks offer continuous numerical outputs within a range, such as a measurement or product price. The primary aim of a regression problem is to come up with a mapping function based on the input and output variables [108].

Unsupervised learning uses ML algorithms to analyse and cluster data in datasets that are neither classified nor labeled [105]. Some examples of unsupervised learning algorithms include k-means clustering, hierarchical clustering, principle component analysis, and autoencoders [109–112].

In reinforcement learning, on the other hand, algorithms learn to react to their surroundings on their own via a trial-and-error method [113]. The most renowned reinforcement learning algorithms are designed to compete with human specialists in games such as chess or Go [114].

Despite the lack of a clear distinction between the two stages, older algorithms, such as support vector machines (SVM) and decision trees (DT), are commonly referred to as conventional ML approaches [115]. Contrarily, DL automatically extracts high-level features from the raw data due to its stacked structure, which is in a hierarchy of increasing complexity

and abstraction [116].

The conventional methodologies, as well as the DL methods, are briefly introduced in this section.

2.4.1 Conventional machine learning

Conventional ML systems are mostly based on hand-crafting features from the dataset and using those features for predicting outcomes. They need domain expertise and human interaction, making them unsuitable for many complex tasks [116]. The most common algorithms used in the field are SVMs, DTs, and logistic regression (LR).

SVMs are one of the most popular supervised ML methods, especially in neuroimaging analysis, due to their flexibility and simplicity when applied to a wide range of problems [117]. The fundamental goal of this approach is to use various forms of kernel functions to project nonlinear separable samples onto another higher dimensional space [118].

SVM was first introduced in late 1990s for binary classification [119]. Given the training set $T = (x_i, y_i)$ of $y_i \in [-1, +1]$, the purpose of the binary classification is to identify the hyperplane which divides the space into two half-spaces for two different classes of inputs. For a linear classifier, a hyperplane is a linear function of x, $f(x) = wx + b$, such that $y_i(f(x)) = y_i(wx + b) > 0$ where w is weight vector and b is bias.

As a result, the separating hyperplane may be represented as follows:

$$f(x) = wx + b = 0 \tag{2.1}$$

Thus, the optimal hyperplane is the one maximising the margin between the two classes. After making a decision boundary that separates between the two classes as wide as possible, SVM automatically assigns new samples to any of the two-class labels depending on their location to the line [120].

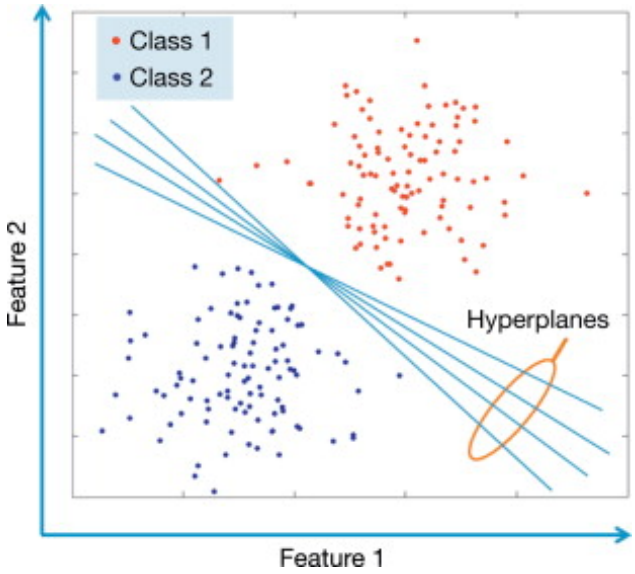


Figure 2.3: Illustration of a support vector machine classification results indicating that a large number of hyperplanes may give an equally good separation between the two classes. Reprinted from [3].

A DT is another supervised learning method where observations draw mapping to conclusions about its target value. The leaves in the tree structures indicate class labels,

non-leaf nodes are features, and branches are feature conjunctions that lead to classifications [121]. One of the most significant benefits of DTs is that they need minimum data preparation and are simple to explain to non-technical persons [122]. They do, however, have a tendency to overfit the training data without proper pruning or constraining tree development, resulting in poor generalisation.

Despite its name, LR is another widely used model in classification tasks [123]. Given that the classes in supervised classification tasks are discrete, the purpose of the methods is again to determine the decision boundaries between the classes. It assumes that $y|x$ is the Bernoulli distribution. The formula of LR can be found in equation 2.2.

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{2.2}$$

Because of the oversimplified assumption of linear decision boundaries, logistic regression is frequently one of the first techniques to be implemented to solve classification problems [124]. Furthermore, logistic regression is thought to be less prone to overfitting due to the linear, noncomplex decision boundaries [123].

2.4.2 Deep learning

DL refers to a class of ML algorithms that is able to learn from data like the other ML algorithms. However, unlike the conventional ML models, DL no longer requires a hand-crafted feature extraction by automating the process. DL's hierarchical design for feature

learning is another distinguishing trait [125]. A DL algorithm tries to mimic how the human brain learns by taking an obscure job, such as differentiating a pattern and breaking it down into many tiers of simpler tasks [126]. During the early phases of development, the model receives a large amount of data, which necessitates substantial processing time and power to decide the output. However, as training progresses, neural connections get stronger and adapt to accommodate continual learning [127]. As per Andrew Ng, "The analogy to DL is that the rocket engine is the DL models, and the fuel is the huge amounts of data we can feed to these algorithms [128]."

DL has become a popular class of ML algorithms in computer vision and has been successfully employed in various tasks, including multimedia analysis (image, video and audio analysis), natural language processing, and robotics [129]. The most often used deep networks, CNNs and Recurrent Neural Networks (RNN), will be discussed in the following sections: 2.4.2.1 and 2.4.2.2, with a focus on CNN as it is widely used throughout the thesis. Moreover, a theoretical background regarding another type of feed forward neural network which is mainly used in Chapter 6 and Chapter 7 will also be given in 2.4.2.3. To find more details regarding the concept of DL including mathematical and conceptual background, techniques used in industry, and research perspectives, please see [130].

2.4.2.1 Convolutional neural networks

CNN is a type of feed-forward neural network that is mainly applied to the pattern and image recognition [131].

An input and output layer and other hidden layers are the main building blocks of a CNN. Convolutional layers, pooling layers, activation functions, and fully connected (FC) layers are common hidden layers [132]. The convolution layer is the main component that extracts and creates a number of feature maps from the original input image or the preceding layer's output. Starting at the top left corner of an image, the convolution filter advances horizontally over each row of pixels. The activation layer comes immediately after the convolution layer. It is either a sigmoid, hyperbolic tangent or rectified linear units (ReLu) layer that introduces nonlinearity to the convolution layer to avoid overfitting and inflating gradient effects. The pooling layer decreases the dimensionality of the feature maps by combining the outputs of a cluster of neurons from the previous layer into a single neuron in the following layer [133]. Pooling may be classified into three types: maximum, average, and sum pooling [134]. Finally, the image goes through dense or FC layers, which provide network output. The link between the features retrieved by preceding convolutional and pooling layers and the target is learned by FC layers.

The difference between the predicted and real labels is measured using the loss function. In classification tasks, cross-entropy loss, which measures the distance between the output distribution and the true distribution, is extensively utilised [135]. Mean squared error

(MSE) loss and hinge loss are other common loss functions that are widely considered in the literature [136].

The weights of a neural network model are traditionally set to a small random number, while the biases are set to zero. A parameter optimisation algorithm determines the weight update with the goal of minimising the error function [137]. By updating the weights in the opposite direction of the gradient of the performance function with respect to the weights, the error is minimised. Because each weight is changed separately, using a partial derivative is essential in this procedure. In addition, a scalar parameter called 'learning rate' is added to govern the weight change step size. Another forward pass occurs once the weights have been changed. When either a pre-determined number of iterations or a minimal error rate is met, the learning process comes to an end. Gradient Descent, Gradient Descent with Momentum, Scaled Conjugate Gradient, and BFGS Quasi-Newton are some of the optimisation techniques that have been applied for training CNNs [138, 139].

Although LeCun [140] invented the first CNN architecture, LeNet, in 1988, low processing and memory capabilities rendered the technique impossible to deploy until around 2010. The model, which counts as a standard template of CNN, consists of seven layers: two convolutional layers associated with pooling layers, followed by three FC layers. In 2012, Krizhevesky et al. [133] proposed a deeper and wider CNN model called AlexNet and made a monumental impact in the research community by winning the most difficult

ImageNet challenge for visual object recognition called the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). VGGNet [141], which comprises 16 convolutional layers, was the runner-up in ILSVRC 2014. It was intriguing because the architecture demonstrated that employing numerous little 3x3 filters is more efficient than using a few larger filters. The notion of skip connections is suggested by He et al. [5] for deep CNN training, and ResNet gained prominence in 2015. Following that, most succeeding networks, such as Inception-ResNet, Wide ResNet, ResNeXt, and others, employed this approach [142–144]. Recently, a new family of convolutional networks such as EfficientNetV2 [145], and hybrid models that combine convolution and self-attention like CoAtNet [146] attracted major attention due to their success on the well-established ImageNet dataset.

2.4.2.2 Recurrent neural network

A recurrent neural network (RNN) is a type of artificial neural network that works with time-series data or sequential data. It is characterised by its "memory," which allows it to impact the current input and output by using information from previous inputs. While typical deep neural networks presume that inputs and outputs are independent of one another, the output of recurrent neural networks relies on the sequence's prior parts [147].

RNNs are commonly employed to address ordinal or temporal problems such as natural language processing [148], handwriting recognition [149], audio recognition [150], and image

captioning [151].

2.4.2.3 Autoencoders

Autoencoder is an unsupervised artificial neural network that consists of two parts: an encoder and a decoder. While the encoder tries to learn efficient representations of the input in a reduced dimension, the decoder part of the network reconstructs the input as close to the original as possible using latent representation coming from the encoding part. In other words, an autoencoder aims to learn an approximation to the identity function by minimising the reconstruction error between input and output. In this work, the MSE is used as reconstruction error between the input image x and the reconstructed image at the output $\hat{x}_i = g(f(x_i))$:

$$\mathcal{L} = \frac{1}{N}\sum_{i}(x_i - g(f(x_i)))^2 \tag{2.3}$$

Autoencoders are mainly used for data dimensionality reduction and image denoising as well as learning latent representations that can be used to generate novel data samples.

The number of convolutional layers, filter size of convolutional layers, and convolutional kernel size are the three main hyperparameters in the CAE.

2.5 Datasets

The datasets utilized in this dissertation are briefly presented in this section. In later chapters, the subsets formed from the datasets for the specific purpose of the experiments are described in depth.

2.5.1 Datasets for AD

The Alzheimer's Disease Neuroimaging Initiative (ADNI), the Open Access Series of Imaging Studies (OASIS), and the Australian Imaging, Biomarkers and Lifestyle (AIBL) are three publicly available datasets that have been primarily utilized in the research of AD. The first two datasets utilized in the thesis are briefly described in the following sections.

2.5.1.1 ADNI

ADNI[1] is a research initiative that brings together researchers to collect, validate, and utilize several types of data such as clinical, genetic, MRI, PET, and biospecimen to validate biomarkers for AD [6]. ADNI was formed in 2004 and launched three different phases so far, namely ADNI 1, ADNI GO/2, and now ADNI 3. In addition to the first phase, ADNI 2 contains information from 150 elderly controls, 100 early mild cognitive impairment (EMCI) subjects, 150 late mild cognitive impairment (LMCI) subjects, and 150 mild AD patients.

[1]The details regarding acquisition protocols can be found at `http://adni.loni.usc.edu/methods/documents/mri-protocols/`.

2.5.1.2 OASIS

OASIS[2] is a project that is intended to promote future discoveries in AD by providing neuroimaging datasets freely to the scientific community. The project released data in three different phases: OASIS 1-Cross-sectional, OASIS 2-Longitudinal, and OASIS-3-Longitudinal. OASIS 1 includes overall 416 subjects (316 HC and 100 AD) aged 18 to 96.

2.5.2 Dataset for PD

The main publicly available dataset for PD research, namely the Parkinson's Progression Markers Initiative (PPMI), is presented in this subsection.

2.5.2.1 PPMI

The PPMI is a publicly available long-term observational research project that collects clinical, imaging, genetic, and biochemical information and helps researchers identify biomarkers of PD progression. The imaging dataset consists of a set of three-dimensional brain slices of 452 PD patients (292 males and 160 females) and 204 HC (134 males and 70 females). The average age of the patients is 61, where the minimum age is 30, and the maximum age is 89.

[2]More details about the OASIS-1 data can be found at `https://www.oasis-brains.org/files/oasiscross-sectionalfacts.pdf`.

Chapter 3

2D CNN for the automated diagnosis of neurodegenerative diseases using structural MRI

Over the past decade, ML gained considerable attention from the scientific community and has progressed rapidly as a result [152]. Given its ability to detect subtle and complicated patterns, DL has been utilised widely in neuroimaging studies for medical data analysis and automated diagnostics with varying degrees of success [153]. In this chapter, two state-of-the-art CNN models has been implemented classification of two most common neurodegenerative diseases, namely AD and PD, using MRI. The impact of the data division strategy on the model performance is demonstrated by comparing the results derived from two different split

approaches. The performance of the CNN models are first evaluated by dividing the dataset at the subject level in which all of the MRI slices of a patient are put into either training or test set. It is then observed that pooling together all slices prior to applying cross-validation, as erroneously done in a number of previous studies, leads to inflated accuracies by as much as 26% for the classification of the diseases. This chapter is based on [11].

3.1 Introduction

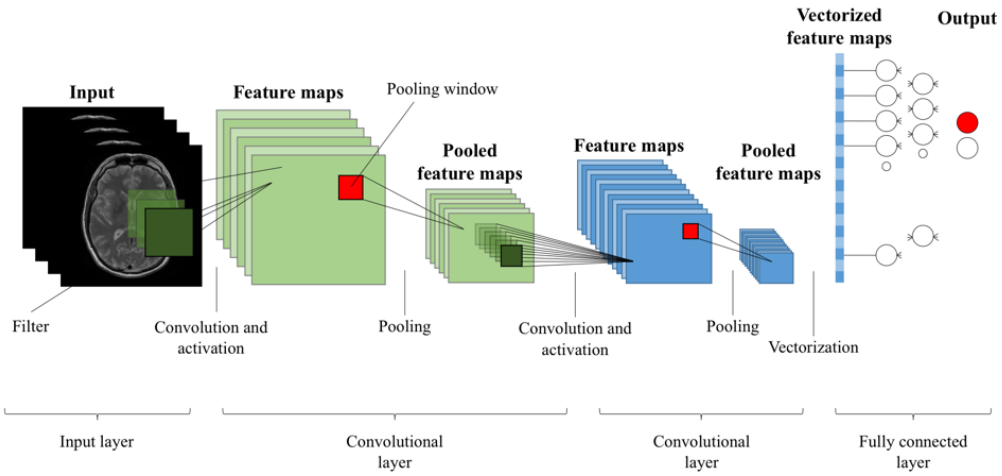


Figure 3.1: The architecture of a convolutional neural network (CNN) model used in medical image classification. (Modified from the Figure in [4])

DL models have attracted a great deal of research interest in medical imaging due to their advantages and successes in various fields such as image and speech recognition, automation, security, computer-aided diagnosis (CAD), just to name a few [154]. In particular, medical image analysis using DL opened a new door into CAD. In recent years, CNNs have been used

to detect and classify a range of diseases from cancer to neurological disorders [155–158].

The CNN models used in these studies are mostly utilised on well-known big datasets such as ImageNet [159] and MNIST [160]. A sample CNN architecture used in medical image classification can be seen in Figure 3.1. Model training and testing are generally done by splitting the dataset into three subsets: training, validation, and test. Training and validation are used to learn parameters and decide whether training is complete, whereas test data are used to evaluate model performance on new previously unseen data. However, CNN models may not perform well when presented with the new data as well as previously believed [161]. A recent work in computer vision has indicated that the true generalisation performance of even classic CIFAR-10 photograph classification CNNs to new data are questionable and lower than previous results [162]. In domains such as disease detection, that kind of mismatch can cause serious problems as the researchers could design models which perform well on the specific test set but are incapable of generalising, and fail when new data are presented [163].

It has been long known that having an appropriate data division is crucial to achieve a generalisation performance [164, 165]. There are various statistical sampling techniques such as simple random sampling [166], deterministic methods [167], DUPLEX [168], and stratified sampling [169] which may be used in different types of data to decrease the variance of the model performance.

To measure the model's ability to adapt properly to new, previously unseen data, the

ideal test set should be the reflection of the data that could be encountered elsewhere. Most image classification algorithms split the data into training, validation, and test sets at random, assuming that all images are independent of one another. This assumption, however, may fail in medical imaging when patches or slices are derived from a 3D image at various time periods, resulting in data from the same individual appearing in numerous sets.

Thus, in medical image classification, the accuracy on a test set which is randomly sampled from the data may not reflect the model's performance on new, previously unseen data and may create a major bias which can be explained as data leakage [170, 171]. The concept of data leakage and its detailed quantitative assessment will be further explained in Chapter 4.

In this chapter, the generalisation performance of the networks on the classification of the two most common neurological disorders is assessed. The contributions of this chapter are as follows:

- A solid framework for PD and AD classification using CNNs and MR images is proposed;
- AD is further classified into its prodromal stage, mild cognitive impairment (MCI)
- Two state of the art CNN models together with a smart data selection algorithm are presented and tested on two public datasets: PPMI and OASIS;
- The impact of the data division strategy on the model performances is demonstrated

by comparing the results based on two different split approach, one of which affected by data leakage.

This chapter is organised as follows: In Section 3.2, there is an overview on the related work. Section 3.3 describes the steps of the proposed methodology in detail. Classification results are presented in Section 3.5 and discussed in Section 3.6. Finally, Section 3.7 concludes the chapter with some remarks and indicates possible future directions.

3.2 Related work

Disease	Study	No. of subjects	No. of MRIs	Data division method	Accuracy (%)
PD	Sivaranjini et al., 2019 [172]	182	7646 slices (2D)	4:1 train/test split by MRI slices	88.9
PD	Esmaeilzadeh et al. 2018 [173]	452	452 volumes (3D)	8.5:1:0.5 train/development/test split by augmented patches from MRI	100
AD	Jain et al.,2019 [174]	150	3000 slices (2D)	8:2 train/test split, by augmented MRI slices	95
AD	Hon and Khan, 2017 [175]	200	6400 slices (2D)	4:1 train/test split by MRI slices, 5-fold cross-validation	92.3
AD	Farooq et al., 2017 [176]	355	38024 slices (2D)	3:1 train/test split by MRI slices	98.8
AD	Sarraf and Tofighi, 2016 [177]	n/a	90300 slices (2D)	3:1:1 train/validation/test split, 5-fold cross validation	96.85
AD	Wu et al., 2018 [178]	457	21936 slices (2D)	2:1 train/test split, 5-fold cross validation	97.58
AD	Payan and Montana, 2015 [179]	n/a	100 volumes (3D)	8:1:1 train/validation/test split, by patches from MRI	89.47

Table 3.1: Summary of the studies with potential of data leakage. Studies perform PD and AD classification using 2D or 3D convolutional neural networks (CNNs) with structural magnetic resonance imaging.

In recent years, several neuroimaging studies have utilised ML algorithms for detection and diagnosis of PD [180–182]. Various modalities such as MRI, PET, fMRI, and single photon emission computed tomography (SPECT) are used within these research to diagnose PD [183, 184]. In 2018, Esmaeilzadeh *et al.* [173] used 3D CNN for simultaneous classification and regression of PD diagnosis based on MRI and personal information (i.e., age and gender). They achieved 100% accuracy on both test and validation sets. In that

chapter, they reached to the conclusion that Superior Parietal part on the right hemisphere of the brain is very critical in the diagnosis of PD. Lei *et al.* [185] performed a multi-class classification of three different clinical statuses: PD, SWEDD, and healthy conditions (HC) via SVM. They concluded that the classification performance with multi-modality features (GCD) combined with CSF biomarkers and clinical scores (DSSM) is always better than those without additional features. Recently, Sivaranjini *et al.* [172] utilised AlexNet to diagnose PD. The image dataset with 80% of the input data are used for training, and the remaining 20% is used for testing. Through TL, they achieved an accuracy of 88.9% on the classification of MRI slices. However, they did not test their model with subjects that were not included in the training data.

For the diagnosis of AD, on the other hand, Sarraf *et al.* [177] used a CNN model using functional magnetic resonance imaging (fMRI) and MRI. The data was divided into three parts: training (60%), validation (20%), and test (20%). They achieved 99.9% accuracy for functional MRI data and 98.84% for MRI data, respectively. However, data division was not done at the subject-level leading data from the same subject to be in both the training and test sets.

In [179], Payan and Montana designed a classification system that combines sparse autoencoders and CNNs. They divided ADNI dataset into training set (1,731 samples - 76.5%), validation set (306 samples -13.5%) and test set (228 samples - 10%) and achieved 95.39% classification accuracy with both 2D CNNs and 3D CNNs. Again, they did not

perform subject level division. Lastly, Hon *et al.* [175] utilised two state-of-the-art architectures, namely VGG16 and Inception V4 to classify AD. They used 5-fold cross-validation to obtain the results, with an 80% - 20% split between training and test. By using a pre-trained model for transfer learning (TL), they reported 92.3% accuracy with VGG16 model and 96.25% with Inception model.

The phenomenon known as data leakage, is indeed a serious problem in the literature. Still, many chapters published in the area are suffering from biased results most probably caused by limited experience with medical data. A recent work by Wen *et al.* [186] is also illustrated the presence of data leakage across various studies which use ML in AD classification. They performed a rigorous literature search on AD and grouped the studies into three categorize: (a) studies without data leakage; (b) studies with potential data leakage and (c) studies with clear data leakage. They observed data leakage in 42% of surveyed chapters.

3.3 Methods

In this section, the data splitting, the pre-processing steps and finally, the model architectures are briefly described.

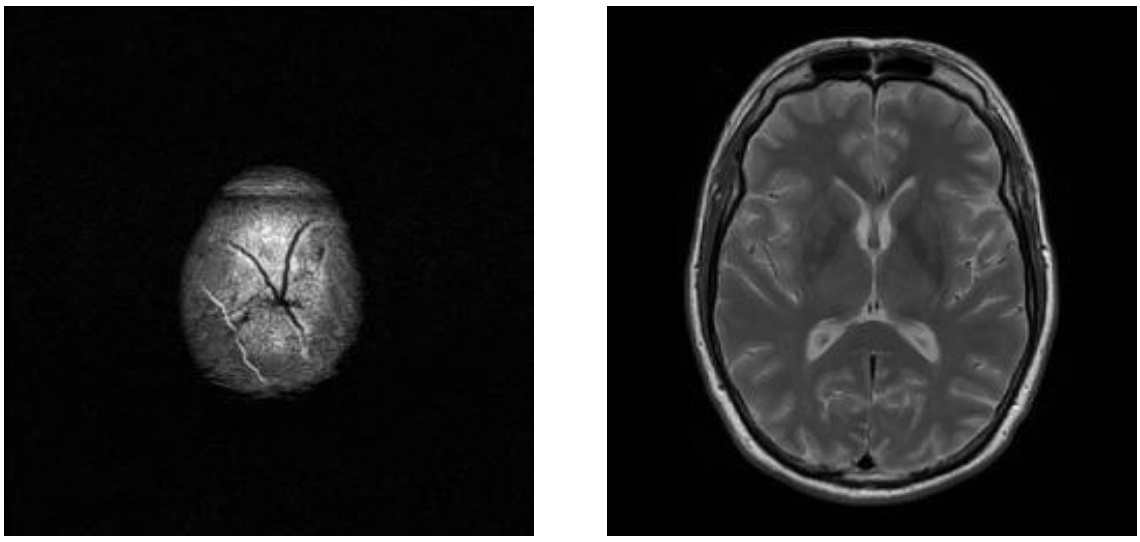
3.3.1 Data Splitting

Throughout the work, it has been realised that a common misconception occurs in many different studies which use ML algorithms in 3D medical imaging. Performance of the models was often determined by dividing the pooled slices into training and test sets [172, 175–177, 187] (see Table 3.1). Thus, training and test sets included the different brain slices of the same subjects. Unfortunately, in that case, the high accuracy may stem from high intra-subject correlation. To test the hypothesis, two different data splitting approaches are employed. First, the data is divided by subject, in which all of the MRI slices of a subject are placed either in the training or in the test set. Then, in the second part, all the slices are pooled together and the overall set is split randomly, meaning that the different slices of the same patient could appear both in the training and test sets. The preliminary exploration on the issue has been given in this chapter with a focus on the models created to classify AD+MCI vs. HC, MCI vs. HC and PD vs. HC. The extent of data leakage in the literature and its quantification will be thoroughly discussed in the Chapter 4.

3.3.2 Image Pre-processing

The input of the 2D CNNs that has been utilised in the proposed approach is the set of 2D slices extracted from the MRI volume. Typically, each MRI volume contains many slices that correspond to a different cross-section of the brain.

To increase the performance of classification, it has been decided to pick the most



(a) Non-informative slice in terms of the amount of the grey matter visible

(b) Informative slice in terms of the amount of the grey matter visible

Figure 3.2: Example of two magnetic resonance imaging (MRI) slices of a Parkinson Disease (PD) subject from the Parkinson's Progression Markers Initiative (PPMI) dataset.

informative slices to train the network. It is known that a significant matter intensity loss with changes in the striatum region is observed in PD when compared with HC [188]. By calculating the image entropy for each slice, it has been aimed to select the slices that can illustrate such a degenerated structure [175].

Two sets of MRI slices that belong to a PD patient are shown in Figure 3.2. The slice on the left of the figure is not very informative in terms of the amount of grey matter it reveals when compared to the slice on the right.

Entropy is a measure of histogram dispersion which illustrates the variation in a slice [189]. In the case of an image which has been perfectly histogram equalised, all 256 such states

are equally occupied, and the entropy of the image is maximum [190]. On the other hand, if all of the pixels of an image have the same value, the entropy is zero. Therefore, if the entropy of the image is reduced, its information is reduced as well. Thus, to obtain the most informative slices for network training, an entropy threshold has been determined (4.5, based on the empirical analysis).

For a slice, the entropy can be calculated as follows:

$$H = -\sum_{i=1}^{M} p_i \log p_i$$

where M is the number of grey levels (256 for 8-bit images) and p_i is the probability of a pixel having grey level intensity.

After eliminating the slices which fail to carry the necessary information, normalisation was performed on the remaining MRI slices to obtain an unvaried contrast and intensity range. For this reason, each MRI slice in the data set was normalised to the range $(0, 1)$. To be compatible with the pre-trained models of VGG16 and Resnet50, the slices were resized to be 224×224. The models are presented thoroughly in the subsection 3.3.3.

The AD slices were subjected to the same pre-processing structure.

3.3.3 CNN Models

VGG and ResNets, two extensively utilized architectures in disease detection frameworks, are employed.

3.3.3.1 VGG16

VGG16 is a 16-layer network built by Oxfords Visual Geometry Group (VGG) and presented in their chapter entitled "Very Deep Convolutional Networks for Large-Scale Image Recognition" [191]. It won the ImageNet competition in ILSVRC-2014 with the accuracy of 92.7%. It replaces large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) in the Alexnet with multiple 3×3 kernel-sized filters.

The input to the first layer is a fixed-size 224×224 RGB image. The image is then passed through a stack of convolutional layers as well as max pooling layers. Finally, convolutional layers are followed by three Fully-Connected (FC) layers and the soft-max layer for 1000-way ILSVRC classification. The architecture of VGG16 is shown in the Figure 3.3.

3.3.3.2 Resnet50

Residual neural network (ResNet) ranked first in the ILSVRC 2015 classification competition with top-5 error rate of 3.57%. He *et al.* [5] ease the training process of deep neural networks while making their model deeper than those used previously. They reformulate the layers as learning residual functions with reference to the layer inputs,

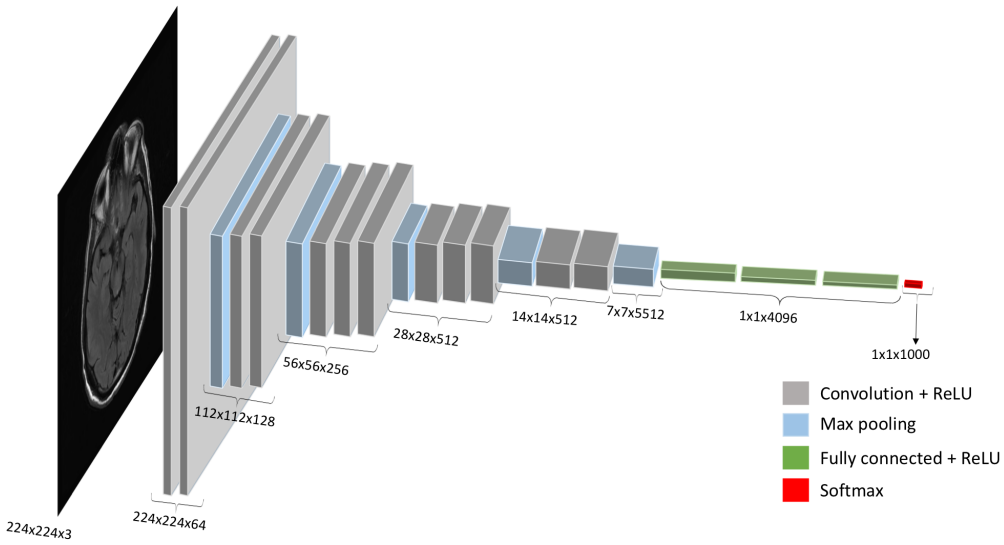


Figure 3.3: The architecture of the VGG16 model adopted for magnetic resonance imaging (MRI) data.

rather than learning unreferenced functions. Residual neural networks solve the problem known as vanishing gradient. When the network is too deep, the gradients of the loss function approaches zero, making the network hard to train. As a result, the weights are not updated, and thus learning cannot be achieved. With ResNets, the gradients can flow directly through the skip connections backward from latter layers to initial filters. The building block of a sample residual neural network structure is shown below in the Figure 3.4.

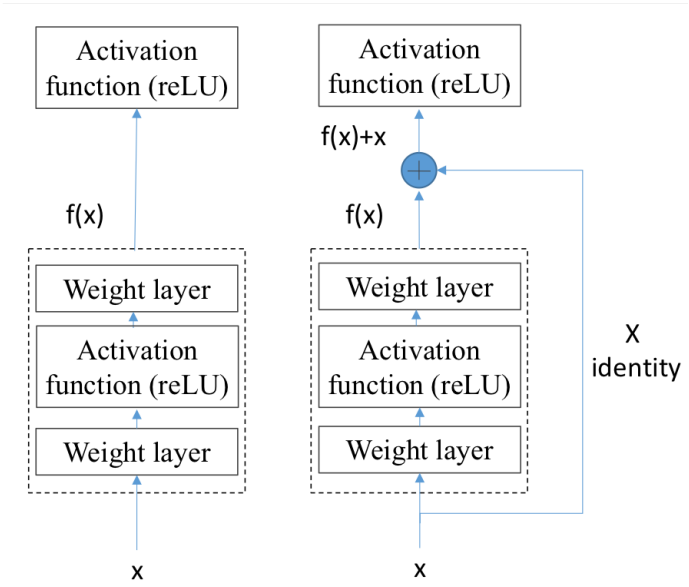


Figure 3.4: A building block of a regular learning (left) and a residual learning (right) (from He, 2016 [5]).

3.4 Evaluation framework

In this section, the datasets used in the experiments are presented together with the training protocols of the models.

3.4.1 Datasets

In this chapter two datasets were used, namely Parkinson's Progression Markers Initiative (PPMI) database [192] for PD and OASIS [7] for AD.

Dataset	Diagnosis	No. of patients	Sex	Age	No. of MR slices)
PPMI	PD	204	101 M, 103 F	30-89	3015
	HC	204	134 M, 70 F	30-89	3015

Table 3.2: Demographic information of PPMI dataset.

3.4.1.1 PPMI

The axial T2 weighted MRI slices used to classify PD in this chapter are from the PMMI database (Table 3.2). The reason behind using T2 weighted MRI for PD is that T2 weighted sequences are better at detecting changes in tissue properties [193]. As a result, the data has the potential to monitor the structural changes of the brain caused by PD, such as the reduced volume of caudate and putamen [188].

The PPMI subset used in this chapter consists of 408 subjects, with 204 HC and 204 PD subjects. It has 6569 MRI slices derived from HC and 4467 slices from PD subjects. 7030 slices in total were randomly picked for the slice-based PD subset. Of these, 3515 slices were PD, and the remaining 3515 were HC. In the case of random division, 80 % of these slices were used in training, while the remainder were assigned to the test set to prove the effect of data leakage. For the subject-based case, the data was divided by patient meaning that the MRI slices of 164 patients from each class were placed in the training set and the slices of 40 AD patients and 40 HC were assigned to the test set.

Dataset	Diagnosis	No. of patients	Sex	Age	No. of MR slices)
OASIS	AD	100	65 M, 35 F	18-96	3200
	HC	100	38 M, 62 F	18-96	3200

Table 3.3: Demographic information of OASIS-1 dataset.

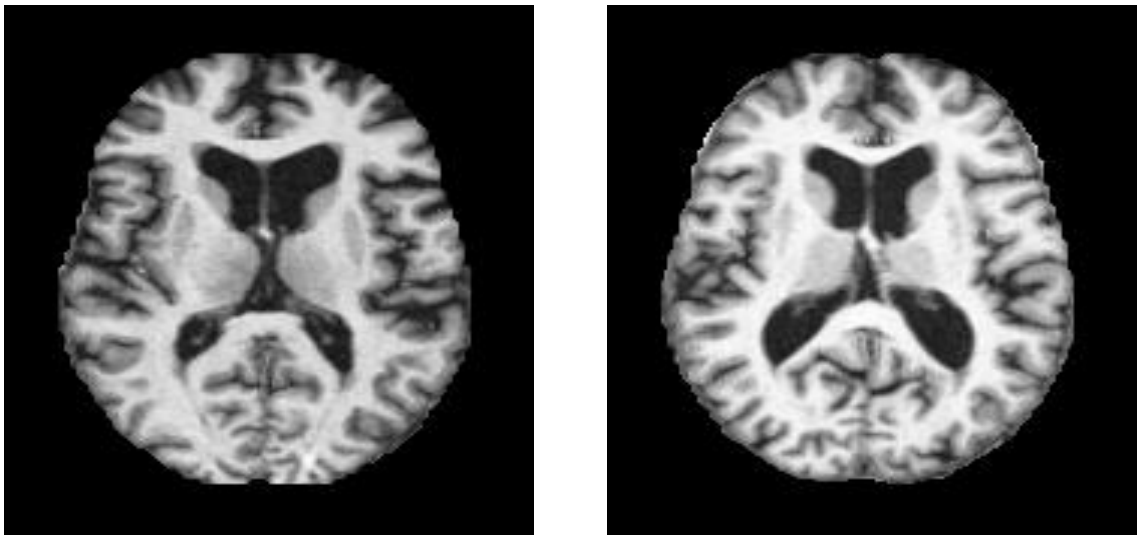
3.4.1.2 OASIS

For classification of AD, cross-sectional, structural MRI data from the OASIS database was used (Table 3.2). For the random split tests, the exact data set which were used in Hon *et al.*'s work [175] was employed in order to replicate their approach while avoiding bias.[1] The subset they have used in their work consists of cross-sectional T_1-weighted MRI scans. In their experiments, they randomly picked 200 subjects, 100 of whom were chosen from the AD group, while the other 100 from the HC group. The sample MRI slices from OASIS data can be seen in Figure 3.5.

For the subject based case, a similar subset from the OASIS database was created by picking 200 subjects, half of whom were AD patients, while the other half was HC. MRI slices of 80 subjects from each class are used to train the model, while the other subjects took part in testing process. MRI scans from OASIS database are in hdr/img file format. For pre-processing, the scannings were first converted into NIfTI format, then into 2D (jpg) format.

The decision criteria of AD in this work was that a variable called CDR with 0 suggested

[1]The subset Hon *et al.* created from the OASIS data are accessible at `https://github.com/marciahon29/Ryerson{_}MRP`



(a) A sample magnetic resonance imaging slice of a Alzheimer's disease patient

(b) A sample magnetic resonance imaging slice of a health control

Figure 3.5: Example of two Magnetic resonance imaging (MRI) slices of an Alzheimer's Disease (AD) subject and healthy control (HC) from OASIS database.

HC and any value greater than 0 implied AD. For that reason, from the clinical perspective, the AD dataset included MCI patients as well since MCI is staged clinically at the 0.5 level on the CDR scale.

OASIS-1 dataset includes two different data: Raw and processed. Processed images are the brain-masked version of atlas registered image that are used in both types of experiments.

3.4.2 Model training protocols and transfer learning

Acquiring large sets of labeled data in medical imaging is a hard task as it is mostly sealed due to privacy and institutional policies, or expensive to label. To avoid the common problem of overfitting which generally stems from small data set and deep networks, transfer learning (TL) is employed to train a model efficiently on a smaller data set.

The idea behind TL is that many deep neural networks trained on images exhibit a common behavior: the first layers extract generic features and perform general operations such as edge detection or colour blob detection [194]. Such low level features might be applicable to many datasets and tasks. Thus, when a network is pre-trained on an extremely large dataset, such as ImageNet, comprising 1.4 million images with 1000 classes, knowledge extracted from there can be applied to the given task of interest. Even for cross-domain application, such as networks trained on natural images used with medical images, TL has been proved to be robust [195].

For transfer learning, the fine-tuning approach was followed, where the last three layers

of the pre-trained model are modified. The weights of the other layers of the model were frozen during fine-tuning to prevent overfitting. For VGG16, 50 epochs were used with a batch size of 40. The stochastic gradient descent and Adagrad optimisation algorithms were used to minimize cross-entropy type of error. For Resnet50, 100 epochs with batch size of 32 were used. The optimisation model was stochastic gradient descent. The loss function was categorical cross-entropy. In this work, the optimal configuration of several hyperparameters including number of epochs and batch size was determined using the traditional trial-and-error technique proposed by Ortiz-Rodriguez et al. [196].

Data selection method and pre-processing part mentioned in Section 3.3 are implemented in MATLAB [197]. Then, DL methods are executed using Keras [198] with a TensorFlow [199] backend. Architectures as well as the pre-trained weights were available to download in open source repositories of the models.

3.5 Experimental results

The main aim was to differentiate AD+MCI and PD patients from HC by analysing MRI data derived from two different databases via the CNN models and to show the importance of data division method on the generalisation performance of the models. Table 3.4 illustrates the accuracy results of the two models across two separate datasets using subject-level data splitting and random splitting after pooling all slices.

As it can be seen from the Table 3.4, both VGG16 and Resnet models can classify

	PD (Data 1: PPMI)		AD (Data 2: OASIS)	
	RD	SbD	RD	SbD
VGG16	82.8%	61.2%	90.47%	64.3%
Resnet50	88.6%	67.3%	92.5%	67.1%

Table 3.4: Tested architectures and their corresponding average accuracy on two dataset (PPMI and OASIS) using two data divisions (RD-Random Division, SbD-Subject-based Division).

PD from HC with more than 82% accuracy when data was randomly split (biased split). However, on subject based split (unbiased split), a large drop in accuracy (17% to 25%) was observed for classification of the disease. Again, for AD classification, the same pattern was detected. When data was divided at subject level, classification accuracy of VGG16 model is 64.3% whereas Resnet50 model achieves 67.1%. Alarmingly, pooling then splitting at slice level can inflate the classification accuracy by 26.1 percent points compared to the subject level split. A further experimentation was also performed to classify the prodromal stage of AD, known as MCI from HC. MCI diagnosis itself is indeed a very challenging problem because there are no significant changes in the brain structures of MCI patients compared to HCs. In the AD+MCI dataset used in this chapter, there are 21 subjects whose CDR score is 0.5. When VGG16 model was used to differentiate those from HCs, 62% classification accuracy is achieved using subject based split whereas with Resnet50 framework, 63.4% is attained.

3.6 Discussion

Comparison of classification performances across studies is an arduous task as each chapter has various pre-processing stages, validation approach or hyperparameter selection. In studies which create subsets from publicly available datasets, the selection of the subset is often a random process, which makes it impossible to replicate the work accurately [175]. Moreover, some of the studies do not provide sufficient implementation details, especially about the validation procedures adopted, with the risk that the reported performances are affected by significant bias. Dividing the data at the slice-level in medical image classification is a significant problem which is currently widespread in the field. The results show that this may artificially inflate the accuracy of classifiers by as much as 26 percentage points.

To evaluate prospective clinical feasibility of automated diagnosis, unbiased and accurate assessment of the model performances is crucial. Despite the previous works' impressive accuracy, there are still some serious issues that must be resolved, as well as room for improvement in medical image classification and automated diagnosis.

3.7 Conclusion

In this chapter, a transfer learning-based method is utilised to detect two most common neurological diseases from structural MR images. Two state-of-the-art architectures, namely

VGG16 and Resnet, are employed to classify PD subject from HC and AD+MCI subjects from HC. A second experiment was carried out to differentiate the prodromal stage of AD, known as MCI from HC. The proposed models are tested on MRI slices from the PMMI and OASIS brain imaging datasets, where MRI slices of more than 300 patients are used to train the models. The results of two data split approaches are compared across separate data sets, and it is shown that there is a large overestimation in accuracy when slices from all subjects are pooled together prior to validation.

The large discrepancy of accuracies between two types of data division suggests that the test accuracy from the random division approach is not a valid measure of performance on new subjects. Subject level tests are required to show the accurate performance of the classification model.

While it is certain that most researchers are well aware of the issue and would never split data from the same subject into test and training data, it is noticed that this remains a severe issue in the literature. With the recent advances in ML and AI, more and more people are becoming interested in applying these techniques to biomedical imaging and there is a real and growing risk that many of them will not be familiar with the possible issues and the good practices.

Optimising the hyperparameters of the models and expanding the datasets via collaborations may be crucial to achieving better results. With these efforts, it is aimed to solve the problem behind the low accuracy of subject-level tests, achieve better patient

group classification, and ease the diagnosis of neurodegenerative disorders in the near future.

In the next chapter, Chapter 4, modified versions of the models will be investigated together with the effect of deep fine-tuning to further quantify the effect of data leakage in the literature.

Chapter 4

Effect of data leakage in brain MRI classification using 2D CNNs

In recent years, 2D CNNs have been extensively used for predicting diagnosis in neurological diseases from MRI data due to their potential to discern subtle and intricate patterns. Despite the high performances reported in numerous studies, developing CNN models with good generalisation abilities is still a challenging task due to possible data leakage introduced during cross-validation (CV). The effect of data leakage caused by 3D MRI data splitting based on a 2D slice-level, rather than a subject-level, is quantitatively investigated in this chapter, employing three 2D CNN models for the classification of patients with AD and PD. The experiments showed that slice-level CV erroneously boosted the average slice level accuracy on the test set by 30% on OASIS, 29% on ADNI, 48% on PPMI and 55% on a local

de-novo PD Versilia dataset. Further tests on a randomly labeled OASIS-derived dataset produced about 96% of (erroneous) accuracy with 2D slice-level split and an outcome of 50%, as expected from a randomised experiment, for a subject-level data split. Overall, the extent of the effect of overfitting due to an erroneous slice-based CV data is severe, especially for small datasets. The adoption of subject-based CV in 2D CNNs studies is strongly recommended. This chapter is the result of a collaborative effort with the University of Bologna. It was published as a journal paper in Nature Scientific Reports, with Selamawet Workalemahu Atnafu sharing first authorship (Yagis et al., 2021) [12].

4.1 Introduction

Deep CNNs hierarchically learn high level and complex features from input data, hence eliminating the need for handcrafting features, as in the case of conventional ML schemes [130]. The application of these methods in neuroimaging is rapidly growing (see [200, 201] for reviews). Several studies employed DL methods for image improvement and transformation [202–207]. Other studies performed lesion detection and segmentation [208–210] and image-based diagnosis using different CNNs architectures [211, 212]. DL has also been applied to more complex tasks, including identifying patterns of disease subtypes, determining risk factors, and predicting disease progression (see, e.g. [201, 213] for reviews). Early works applied stacked auto encoders [214, 215], and deep belief networks [216] to differentiate neurological patients

from healthy subjects using data collected from different neuroimaging modalities, including MRI, PET, resting-state-functional MRI (rsfMRI) and the combination of these modalities [152]. Some authors reported very high accuracies in classifying patients with neurological diseases, such as AD and PD. For a binary classification of AD vs. HCs, Hon and Khan [175] reported accuracy up to 96.25% using a transfer learning strategy. Sarraf et al. [177] classified subjects as AD or HCs with a subject-level accuracy of 100% by adopting LeNet-5 and GoogleNet network architectures. In other studies, CNNs have been used for performing multi-class discrimination of subjects. Recently, Wu and colleagues [178] adopted a pre-trained CaffeNet and achieved accuracies of 98.71%, 72.04%, and 92.35% for a three-way classification between HCs, stable MCI and progressive MCI patients, respectively. In another work by Islam and Zhang [217], an ensemble system of three homogeneous CNNs has been proposed and an average multi-class classification accuracy of 93.18% was found on an OASIS dataset. For the classification of PD, Esmaeilzadeh et al. [173] distingueshed PD patients from HCs based on MRI and demographic information (i.e., age and gender). With the proposed 3D model, they achieved 100% accuracy on the test set. In another chapter by Sivaranjini and Sujatha [172], a pre-trained 2D CNN AlexNet architecture was used to classify PD patients vs. HCs, resulting in an accuracy of 88.9%.

Although very good performances have been shown by using DL for classification of neurological disorders, there are still many challenges that need to be addressed, including

complexity and difficulty in interpreting the results due to highly nonlinear computations, non-reproducibility of the results and data/information and, especially, data overfitting (see Davatzikos et al. and Vieira et al. [152, 213] for reviews).

A poor generalisation ability on real-world data and overly optimistic results may be due to data leakage – a process caused by incorporating information of test data into the learning process [218]. A more subtle version of this problem is when the test data are disjoint from the training data but come from a distribution that is more similar to that of the training set than one would expect from new data [219, 220]. In 3D medical imaging such as MRI or CT, dividing the overall data randomly causing slices or patches from the same patient to be in both training and test sets and leads to a biased assessment.

In this chapter, the issue of data leakage in one of the most common class of DL models, i.e., 2D CNNs, caused by incorrect dataset split of 3D MRI data is addressed. Specifically, the effect of data leakage in different datasets of T_1-weighted brain MRI of HCs and patients with neurological disorders is quantified using a nested CV strategy. In particular, three 2D CNNs are adopted for the classification of 1) AD patients using two public and international datasets, namely OASIS and ADNI and, 2) de-novo PD patients using a public and a private dataset, namely Parkinson's Progression Markers Initiative (PPMI) and Versilia, respectively. The main focus of this work was on both large and small datasets in order to evaluate a possible increase of performance overestimation when a smaller dataset was used, as it is often the case in clinical practice.

4.2 Related work

While concluding that data leakage leads to overfitting will surprise few practitioners, the extent to which this is happening in neuroimaging applications and the quantitative effect on performance, especially in small datasets, is mostly unknown. While the work that consists of this chapter was published, an independent work by Wen et al. [186] was discovered that corroborated several of the results on the problem of data leaking. They successfully suggested a framework for reproducible assessment of AD classification methods. However, the architectures have not been trained and tested on smaller datasets typical of clinical practice and they mainly employed hold-out model validation strategies, rather than cross-validation (CV) – that gives a better indication of how well a model performs on unseen data [221, 222]. Moreover, the authors focused on illustrating the effect of data leakage on the classification of AD patients only.

Unfortunately, the problem of data leakage incurred by incorrect data split is not only limited within the area of AD classification but can also be seen in various other neurological disorders. It is more common to observe the data leakage in 2D architectures, yet some forms of data leakage, such as late split, could be present in 3D CNN studies as well. Moreover, although deep complex classifiers are more prone to overfitting, also conventional ML algorithms may be affected by data leakage. A summary of these works with clear and potential data leakage is given in Tables 4.1 and 4.2, respectively. Other works with insufficient information to assess data leakage are reported in Table 4.3.

Disorder	References	Groups (number of subjects)	Machine learning model	Data split method	Type of data leakage	Accuracy (%)
AD/MCI	Farooq et al.	AD-MCI-HC (36)	2D CNN	4:1 train/test slice-level split	Wrong split	96.00
	Ramzan et al.	AD-HC (200)	2D CNN (VGG16)	4:1 train/test slice-level split	Wrong split	96.25
	Jain et al.	AD-MCI-HC (150)	2D CNN (VGG16)	4:1 train/test slice-level split	Late and wrong split	95.00
	Khagi et al.	AD-HC (56)	2D CNN (AlexNet, GoogLeNet,ResNet50, new CNN)	6:2:2 train/validation/test slice-level split	Wrong split	98.00
	Sarraf et al.	AD-HC (43)	2D CNN (LeNet-5)	3:1:1 train/validation/test slice-level split	Wrong split	96.85
	Wang et al.	MCI-HC (629)	2D CNN	Data augmentation+10:3:3 train/validation/test split by MRI slices	Wrong split and augmentation before split	90.60
	Puranik et al.	AD/EMCI-HC (75)	2D CNN	17:3 train/test split by MRI slices	Wrong split	98.40
	Basheera et al.	AD-MCI-HC (1820)	2D CNN	4:1 train/test split by MRI slices	Wrong split	90.47
	Nawaz et al.	AD-MCI-HC (1726)	2D CNN	6:2:2 slice level split	Wrong split	99.89

Table 4.1: Summary of the previous studies performing classification of neurological disorders using MRI and with clear data leakage (see also Supplementary Table S1 for a detailed description). AD Alzheimer's disease, HC healthy controls, MCI mild cognitive impairment.

Disorder	References	Groups (number of subjects)	Machine learning model	Data split method	Type of data leakage	Accuracy (%)
AD/MCI	Farooq et al.	AD-MCI-LMCI-HC (355)	2D CNN (GoogLeNet and modifed ResNet)	3:1 train/test (potential) slice level split	Wrong split	98.80
	Ramzan et al.	HC-SMC- EMCIMCI-LMCI-AD (138)	2D CNN (ResNet-18)	7:2:1 train/validation/test (potential) slice-level split	Wrong split	100
	Raza et al	AD-HC (432)	2D CNN (AlexNet)	4:1 train/test (potential) slice level split	Wrong split	98.74
	Pathak et al.	AD-HC (266)	2D CNN	3:1 (potential) slice level split	Wrong split	91.75
ASD	Libero et al.	ASD-TD (37)	Decision tree	unclear	Entire data set used for feature selection	91.90
	Zhou et al.	ASD-TD/HC (280)	Random tree classifer	4:1 train/test split	Entire data set used for feature selection	100
PD	Sivaranjini et al.	PD-HC (182)	2D CNN	4:1 train/test split by MRI slices	Wrong split	88.90
TBI	Lui et al.	TBI-HC (47)	Multilayer perceptron	tenfold CV	Entire data set used for feature selection	86.00
Brain tumor	Hasan et al.	Tumor-HC (600)	MGLCM+2D CNN+SVM	tenfold CV	Wrong split and entire data set used for feature selection	99.30

Table 4.2: Summary of the previous studies performing classifcation of neurological disorders using MRI and suspected to have potential data leakage (see also Supplementary Table S2 for a detailed description). AD Alzheimer's disease, ASD Autism spectrum disorder, EMCI early mild cognitive impairment, HC healthy controls, LMCI late mild cognitive impairment, MCI Mild cognitive impairment, MGLCM modifed grey level co-occurrence matrix, PD Parkinson's disease, SMC subjective memory concerns, TBI traumatic brain injury, TD typically developing.

Disorder	References	Groups (number of subjects)	Machine learning model	Data split method	Accuracy (%)
AD/MCI	Al-Khuzaie et al.	AD-HC (240)	2D CNN	(Potential) slice-level split	99.30
	Wu et al.	AD-HC (457)	2D CNN	Data augmentation+2:1 train/test split by MRI slices	97.58

Table 4.3: Summary of the previous studies performing classification of neurological disorders using MRI and that provide insufficient information to assess data leakage (see also Supplementary Table S3 for a detailed description). AD Alzheimer's disease, HC healthy controls, MCI mild cognitive impairment.

4.3 Methods

4.3.1 Overview

A schematic diagram of the overall T_1-weighted MRI data processing pipeline adopted for all AD and de-novo PD datasets (see 4.4.1) is shown in Figure 4.1. Briefly, after a pre-processing stage which includes registration to a standard space, skull-stripping and slice selection based on entropy calculation 4.3.2, three different 2D CNN architectures have been trained and tested 4.3.3 to quantitatively assess the effect of data leakage on performance. The model's training and validation are conducted in a nested CV loop using two different data split strategies 4.4.2: a) subject-level split, avoiding any form of data leakage and b) slice-level split, in which different slices of the same subject are contained both in the training and the test folds (data leakage will occur).

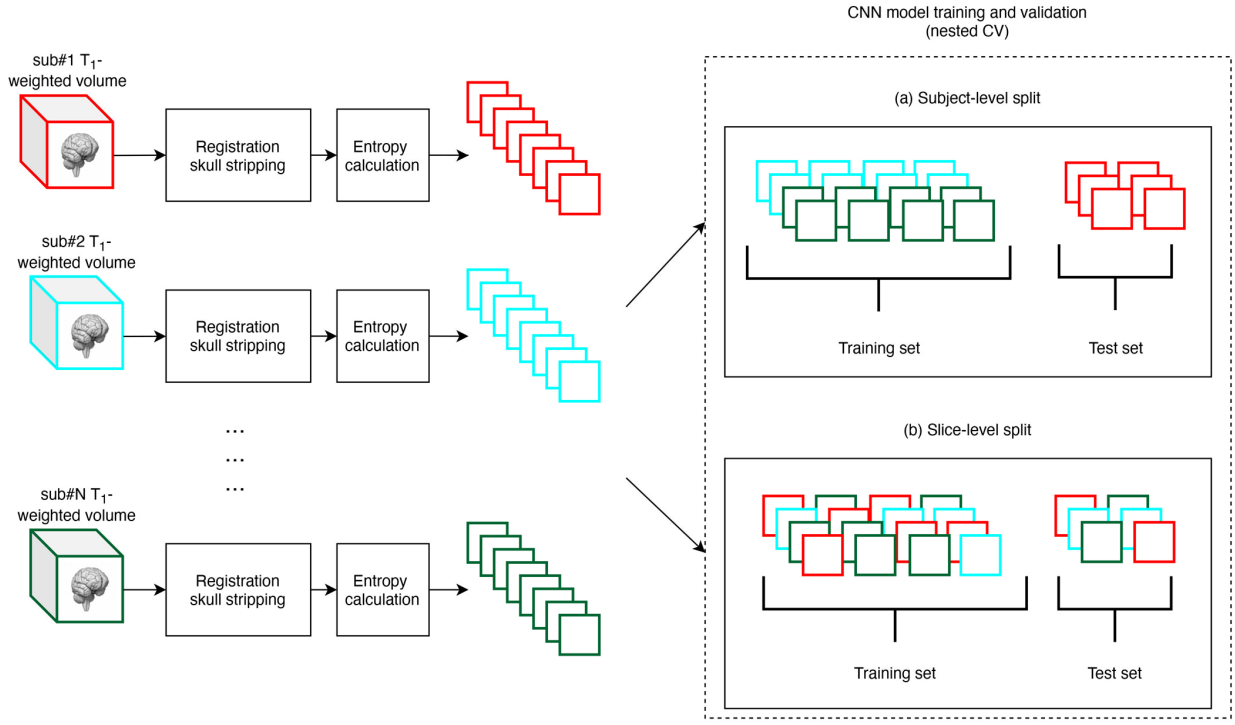


Figure 4.1: Schematic diagram of the overall T_1-weighted MRI data processing and validation scheme. First, a pre-processing stage included co-registration to a standard space, skull-stripping and slices selection based on entropy calculation. Then, CNNs model's training and validation have been performed on each dataset in a nested CV loop using two different data split strategies: a) subject-level split, in which all the slices of a subject have been placed either in the training or in the test set, avoiding any form of data leakage; b) slice-level split, in which all the slices have been pooled together prior to CV, then split randomly into training and test set.

4.3.2 T_1-weighted MRI data pre-processing

All T_1-weighted MRI data went through two pre-processing steps (see Figure 4.1). In the first stage, co-registration to a standard template space and skull stripping were applied to geometrically re-align all the images and remove non-brain regions. In the second stage, the collection of a subset of axial images using an entropy-based slice selection approach has been carried out.

4.3.2.1 Co-registration to a standard template space and skull stripping

For the OASIS datasets, publicly available pre-processed data was used (gain-field corrected, brain masked, and co-registration) [223]. Briefly, the brain masks from OASIS were obtained using an atlas-registration-based method, and their quality was controlled by human experts [7] and each volume has been co-registered to the Talairach and Tournoux atlas. Each pre-processed T_1-weighted volume had data matrix size of 176 $\times$ 208 $\times$ 176 and a voxel size of 1 mm $\times$ 1 mm $\times$ 1 mm [223]. For all other datasets, each individual T_1-weighted volume was co-registered to the MNI152 standard template space (at 1 mm voxel size – available in the FSL version 6.0.3 package) by using the SyN algorithm included in ANTs package (version 2.1.0) with default parameters [224]. Then, the brain mask of the standard template space has been applied to each co-registered volume. Each pre-processed T_1-weighted volume had data matrix size of 182 $\times$ 218 $\times$ 182 and voxel size of 1 mm $\times$ 1 mm $\times$ 1 mm. Figure 4.2 illustrates sample pre-processed T_1-weighted slices

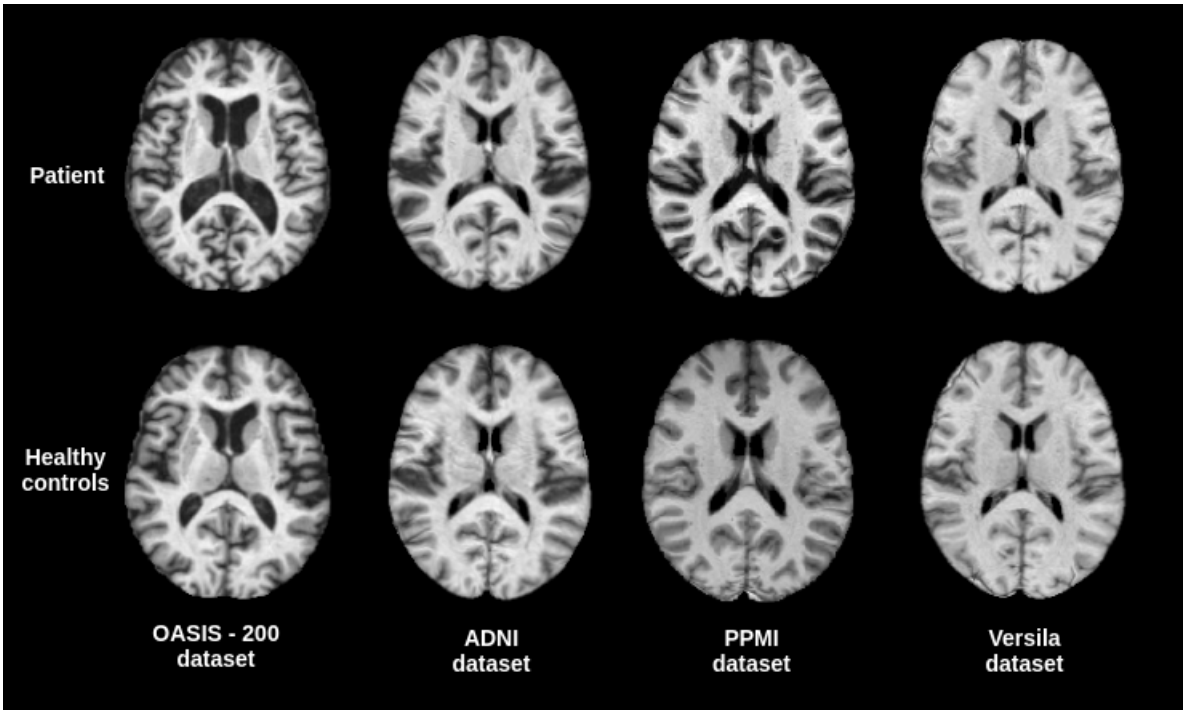


Figure 4.2: Sample pre-processed T_1-weighted axial images from OASIS-200, ADNI, PPMI and Versilia datasets.

from OASIS-200, ADNI, PPMI and Versilia datasets.

4.3.2.2 Entropy-based slice selection

Each T_1-weighted slice generally conveys a different amount of information. Given that developing a 2D CNN model was a main interest, a preliminary slice selection was employed based on the amount of information, retaining, for each T_1-weighted volume, only the eight axial slices that showed the highest entropy [175]. Specifically, for a slice with k greyscale levels and with each grey level having a probability of occurrence p_k (estimated as its relative frequency in the image), the Shannon entropy E_s was computed as:

$$E_s = \sum_k p_k \log(p_k) \tag{4.1}$$

To be consistent with the input sizes of the proposed 2D CNN models, all slices were resized to 224 × 224 pixels by fitting a cubic spline between the 4-by-4 neighborhood pixels [225]. Voxel-wise feature standardisation has also been applied to make training the CNNs easier and to achieve faster convergence, i.e., for each voxel, an average value of all greyscale values within the brain mask has been subtracted and scaled by the standard deviation (within the brain mask) [226].

4.3.3 Model architectures

Since the number of subjects in each dataset may not be sufficient to train with high accuracy a 2D CNN model from scratch, a ML technique called transfer learning that allows employing pre-trained models has been used, i.e., model parameters previously developed for one task (source domain) to be transferred to the target domain for weight initialisation and feature extraction. In particular, CNN layers hierarchically extract features starting from the general low-level features to those specific to the target class and, using transfer learning, the general low-level features can be shared across tasks.

This chapter extends the previous chapter by quantifying the extent of the overestimation of classification accuracy in the case of incorrect slice-level cross-validation. Notably, in this chapter, modifications of previously used models are employed. A

pre-trained VGG16 [191] and ResNet-18 [5] models were used, as detailed in the following sections. The transfer learning approach and VGG16 architectures used in this chapter are similar to those employed in [175] as their results triggered proposed investigation of data leakage.

4.3.3.1 VGG16-based models

VGG16 consists of five convolutional blocks, with alternating convolutional and pooling layers, and three fully connected layers [227].

In transfer learning, the most common approach is copying the first n layers of the pre-trained network to the first n layers of a target network, and then randomly initialising the remaining layers to be trained on the target task. Depending on the size of the target dataset and the number of parameters in the first n layers, these copied features can be left unchanged (i.e., frozen) or fine-tuned during the training of the network on a new dataset. It is well accepted that if the target dataset is relatively small, fine-tuning may cause overfitting, whereas if the target dataset is large, then the base features can be fine-tuned to improve the performance of the model without overfitting. To investigate the effect of fine-tuning, two different variants of VGG16 architecture, namely VGG16-v1 and VGG16-v2 (see Figure 4.3) have been tested. The former model has been used as a feature extractor where the weights for all of the network are frozen except that of the final fully connected layer. The three topmost layers have been replaced by randomly initialised fully

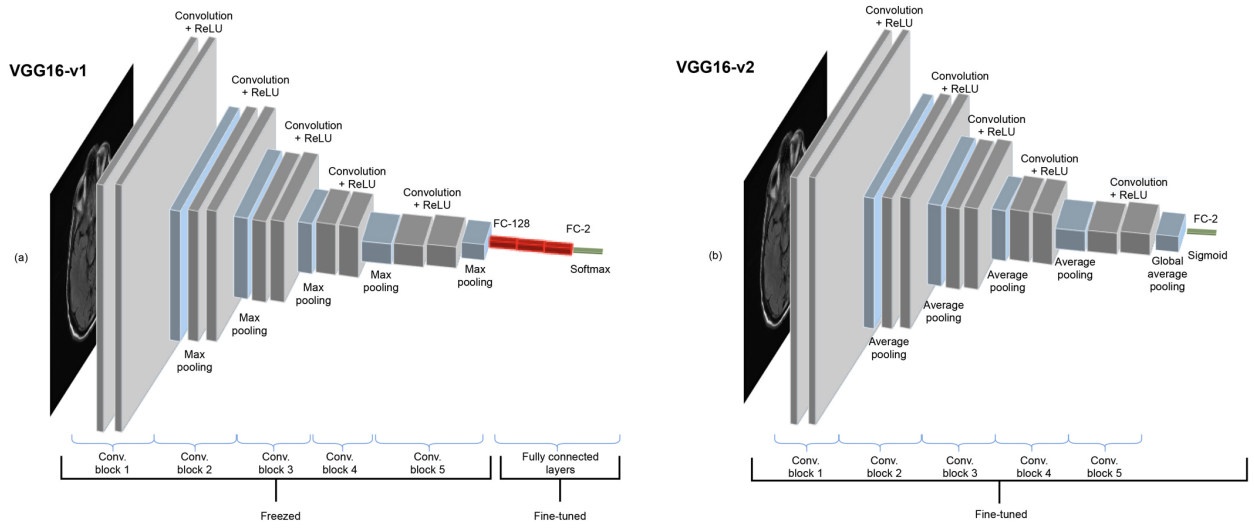


Figure 4.3: The two different networks based on the VGG16 architecture are shown. Each coloured block of layers illustrates a series of convolutions. (a) The first model, named as VGG16-v1 consists of five convolutional blocks followed by three fully connected layers. Only the last three fully connected layers are fine-tuned. (b) On the other hand, the second model, VGG16-v2, has 5 convolutional blocks followed by a global average pooling layer and all the layers are fine-tuned.

connected layers with rectified linear unit (ReLU) activation. The weights are initialised according to the Xavier initialisation heuristic [228] to prevent the gradients from vanishing or exploding. The VGG16-v2 model has been utilised as a weight initialiser where the weights are derived from the pre-trained network and fine-tuned during training. The fully connected layers have been replaced by a randomly initialised global average pooling (GAP) layer suggested by Lin and colleagues [229] to reduce the number of parameters and, rather than freezing the CNN layers, all layers were fine-tuned .

4.3.3.2 ResNet-18 based model

It has been long believed that deeper networks have the ability to learn more complex non-linear relationships than shallower networks with the same number of neurons, and thus network depth is of great importance on model performance [230]. However, many studies revealed that deeper networks often converge at a higher training and test error rate when compared to their shallower counterparts [231]. Therefore, stacking more layers to the plain networks may eventually degrade the model's performance while complicating the optimisation process. To overcome this issue, He et al. [231] introduced deep residual neural networks and achieved top-5 test accuracies with their models on the popular ImageNet test set. The model was proposed as an attempt to solve the vanishing gradients and the degradation problems using residual blocks. With these residual blocks, the feature of any deeper unit can be computed as the sum of the activation of a shallower unit and the residual function. This architecture causes the gradient to be directly propagated to shallower units making ResNets easier to train. There are different versions of ResNet architecture with various numbers of layers. Unlike the previous chapter, in this work, ResNet-18 architecture, which is an 18-layer residual DL network consisting of five stages, each with a convolution and identity block [231] was used. In the model, one fully connected layer with sigmoid activation has been added at the end of the network – a common practice in binary classification tasks as it takes a real-valued input and squashes the output to range between 0 and 1. Since the network is relatively smaller and has a lower number of parameters than VGG16, the

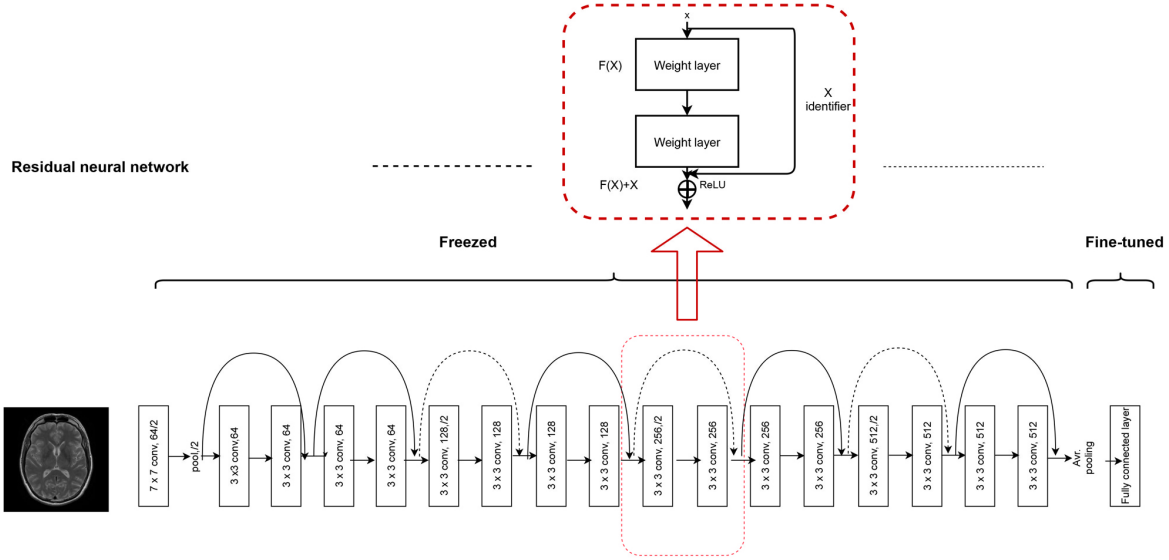


Figure 4.4: A modified ResNet-18 architecture with an average pooling layer at the end is shown. The upper box represents a residual learning block with an identity shortcut. Each layer is denoted as (filter size, channels); layers labeled as "freezed" indicates that the weights are not updated during backpropagation, whereas when they are labeled as "fine-tuned" they are updated. The identity shortcuts can be directly used when the input and output are of the same dimensions (solid line shortcuts) and when the dimensions increase (dotted line shortcuts). ReLU = rectified linear unit.

weights and biases of all the transferred layers are fine-tuned while the newly added fully connected layer has been trained starting from randomly initialised weights. The architecture of proposed ResNet-18 model can be seen in Figure 4.4.

4.4 Evaluation framework

This section gives a comprehensive summary of the datasets as well as a list of evaluation approaches.

4.4.1 Datasets

In this chapter, the scans collected by three public and international datasets of T_1-weighted images of patients with AD (the OASIS dataset [7] and the ADNI dataset [6] and de-novo PD (the PPMI dataset) were adopted. An additional private de-novo PD dataset, namely the Versilia dataset, has also been used. A summary of demographics of the datasets used in this chapter is shown in Table 4.4. In the following sections, a detailed description of all datasets will be reported.

4.4.1.1 OASIS-200, OASIS-34 and OASIS-random datasets

The T_1-weighted images of 100 AD patients [(59 women and 41 men, age 76.70 ± 7.10 years, mean ± standard deviation (SD)] and 100 HCs (73 women and 27 men, age 75.50 ± 9.10 years, mean ± SD) were used from the OASIS-1 chapter – a cross-sectional cohort of the OASIS brain MRI dataset [7], freely available at `https://www.oasis-brains.org/`. In particular, the same scans that were previously selected by other authors (Hon & Khan, 2017 [175]) and the previous chapter were employed, and this dataset was called OASIS-200. The subject identification numbers (IDs) and demographics of these subjects were specified

Datasets	Patients	Healthy controls
OASIS-200		
Number of subjects	100	100
Age (range, years)	62–96	59–94
Age (mean±SD, years)	76.70±7.10	75.50±9.10
Gender (women/men)	59/41	73/27
ADNI		
Number of subjects	100	100
Age (range, years)	56–89	58–95
Age (mean±SD, years)	74.28±7.96	75.04±7.11
Gender (women/men)	44/56	52/48
PPMI		
Number of subjects	100	100
Age (range, years)	34–82	31–83
Age (mean±SD, years)	61.71±9.99	61.91±11.52
Gender (women/men)	40/60	36/64
Versilia		
Number of subjects	17	17
Age (range, years)	48–78	54–77
Age (mean±SD, years)	64±7.21	64.00±7.00
Gender (women/men)	4/13	5/12

Table 4.4: Demographic features of subjects belonging to OASIS-200, ADNI, PPMI, and Versilia datasets. The same information for the OASIS-34 datasets has been reported in Supplementary Table S5. AD Alzheimer's disease, ADNI Alzheimer's Disease Neuroimaging Initiative, OASIS open access series of imaging studies, PD Parkinson's disease, PPMI Parkinson's Progression Markers Initiative, SD standard deviation.

in Table S1 (Supporting Information). No significant difference in age (p = 0.15 at t-test) was found between the two groups, while a significant (borderline) difference in gender was observed (p = 0.04 at χ^2-test). In OASIS-1, AD diagnosis, as well as the severity of the disease, were evaluated based on the global CDR score derived from individual CDR scores for the domains memory, orientation, judgment and problem solving, function in community affairs, home and hobbies, and personal case [232, 233]. Subjects with a global CDR score of 0 have been labeled as HCs, while scores 0.5 (very mild), 1 (mild), 2 (moderate) and 3 (severe) have been all labeled as AD. As mentioned in the previous chapter, the AD dataset contained MCI patients as well because MCI is clinically staged at the 0.5 level on the CDR scale.

All T_1-weighted images have been acquired on a 1.5 T MR scanner (Vision, Siemens, Erlangen, Germany), using a Magnetisation Prepared Rapid Gradient Echo (MPRAGE) sequence in a sagittal plane [repetition time (TR) = 9.7 ms, echo time (TE) = 4.0 ms, flip angle = 10°, inversion time (TI) = 20 ms, delay time (TD) = 200 ms, voxel size = 1 mm × 1 mm × 1.25 mm, matrix size = 256 × 256, number of slices = 128] [7]. Since the problem of overfitting during the training of ML models is strongly dependent on the size of the dataset, an additional subset of the OASIS-200 dataset, which has the same size as the Versilia dataset (see section 2.2.4), called OASIS-34, was created by randomly selecting 34 subjects (17 AD patients and 17 HCs). Given that the performance of a model trained on a small sample dataset could depend on the selected samples, ten instances of OASIS-34

dataset were created by randomly sampling from OASIS-200 dataset ten times independently. The subject IDs included in each instance are found in Table S2 (Supporting Information). No significant differences in age (p>0.05 at t-test) and gender (p > 0.05 at χ^2-test) were observed between the two groups in all OASIS-34 instances except in one case in which a gender difference was found (p = 0.01 at χ^2-test) [see Table S3 (Supporting Information) for details]. The proposed models have been then trained independently on the ten different OASIS-34 datasets, and the average results have been computed. Finally, a new dataset called OASIS-random was created to further quantify the amount of the overestimation of classification accuracy in the case of a slice-level split. For each subject in the OASIS-200 dataset, a fake random label of either AD patient or HC was issued. In this case, as the label and the MRI data are statistically independent, any accuracy significantly above the 50% chance level can be ascribed to overfitting.

4.4.1.2 ADNI dataset

In addition to the OASIS-200 dataset, another dataset for AD was explored in the chapter. The T_1-weighted MRI data of 100 AD+MCI patients (44 women and 56 men, age 74.28 ± 7.96 years, mean ± SD) and 100 HCs (52 women and 48 men, age 75.04 ± 7.11 years, mean ± SD) was considered. No significant difference in age (p = 0.24 at t-test) and gender (p = 0.26 at χ^2-test) was found between the two groups. AD+MCI patients have been randomly chosen from the ADNI 2 dataset (available at `http://adni.loni.usc.edu/`) –

a cohort of ADNI that extends the work of ADNI 1 and ADNI-GO studies (Petersen et al., 2010). Led by the Principal Investigator Michael W. Weiner, MD, ADNI was launched in 2003 with the aim of investigating if biological markers (such as MRI and PET) can be combined to define the progression of MCI and early AD. MPRAGE T_1-weighted MRI scans acquired by 3 T scanners [6 Siemens (Erlangen, Germany) MRI scanners and 6 Philips (Amsterdam, Netherlands) scanners] in a sagittal plane (voxel size = 1 mm $\times$ 1 mm $\times$ 1.2 mm) have been utilized. The image size of the T_1-weighted data acquired from the Siemens and Philips scanners were 176 $\times$ 240 $\times$ 256 and 170 $\times$ 256 $\times$ 256, respectively. Since ADNI 2 is a longitudinal dataset, more than one scan was available for each subject. The first scan of each participant has been chosen to produce a cross-sectional dataset. Table S4 (Supporting Information) provides subject IDs and the acquisition date of the specific scan used in the chapter. The MRI acquisition protocol for each MRI scanner can be found at `http://adni.loni.usc.edu/methods/documents/mri-protocols/`. In ADNI 2 dataset, subjects have been categorised as AD patients or HCs based on whether subjects have complaints about their memory and by considering a combination of neuropsychological clinical scores [6].

4.4.1.3 PPMI dataset

100 de-novo PD subjects (40 women and 60 men, age 61.71 $\pm$ 9.99, mean $\pm$ SD) and 100 HCs (36 women and 64 men, age 61.91 $\pm$ 11.52, mean $\pm$ SD) have been selected from

the publicly available PPMI dataset . No significant difference in age (p = 0.44 at t-test) and gender (p = 0.56 at χ^2-test) was found between the two groups. The criterion used to recruit de-novo PD patients, and HCs were defined by Marek et al. [234]. Briefly, PD patients were selected within two years of diagnosis with a Hoehn and Yahr score <3 [235], at least two of resting tremor, either bradykinesia or rigidity (must have either resting tremor or asymmetric bradykinesia) or a single asymmetric resting tremor or asymmetric bradykinesia and dopamine transporter (DAT) or vesicular monoamine transporter type 2 (VMAT-2) imaging showing a dopaminergic deficit. HCs were free from any clinically significant neurological disorder [234]. The T_1-weighted scans were collected at baseline using MR scanners manufactured by Siemens (11 scanners at 3 T and 5 scanners at 1.5 T), Philips Medical Systems (10 scanners at 3 T and 11 scanners at 1.5 T), GE Medical Systems (11 scanners at 3 T and 24 scanners at 1.5 T) and another anonymous one (5 scanners at 1.5 T). It has been found that three subjects had missing MRI protocols. The details of the MRI protocols of all scanners can be found in Table S5 (Supporting Information).

4.4.1.4 Versilia dataset

Seventeen (4 women and 13 men, age 64 ± 7.21 years, mean ± SD) patients with de-novo parkinsonian syndrome consecutively referred to a Neurology Unit for the evaluation of PD over a 24-month interval (from June 2012 to June 2014) were recruited in this dataset. More details about clinical evaluation can be found in [236]. Seventeen HCs (5 women and 12

men, age 64 ± 7 years, mean ± SD) with no history of neurological diseases and normal neurological examination were recruited as controls. No significant difference in age (p = 0.95 at t-test) and gender (p = 0.70 at $\tilde{\chi}^2$-test) was found between the two groups. All subjects underwent high-resolution 3D T_1-weighted imaging on a 1.5 T MR scanner system (Magnetom Avanto, software version Syngo MR B17, Siemens, Erlangen-Germany) equipped with a 12-element matrix radiofrequency head coil and SQ-engine gradients. The SQ-engine gradients had a maximum strength of 45 mT/m and a slew rate of 200 T/m/s. T_1-weighted MR images were acquired with an axial high resolution 3D MPRAGE sequence with TR = 1900 ms, TE = 3.44 ms, TI = 1100 ms, flip angle = 15o, slice thickness = 0.86 mm, field of view (FOV) = 220 mm×220 mm, matrix size = 256×256, number of excitations (NEX) = 2, number of slices = 176.

4.4.2 Models training and validation

Each 2D CNN model has been trained and validated using a nested CV strategy – a validation scheme that allows to examine the unbiased generalisation performance of the trained models along with performing, at the same time, hyperparameters optimisation [237]. It involves nesting two k-fold CV loops where the inner loop is used for optimising model hyperparameters, and the outer loop gives the unbiased estimate of the performance of the best model. It is especially suitable when the amount of data available is insufficient to allow separate validation and test sets [237]. A schematic diagram of the

procedure is illustrated in Figure 4.5. It starts by dividing the dataset into k folds and one-fold is kept as a test set (outer CV), while the other k-1 folds are split into inner folds (inner CV). The model hyperparameters are chosen from the hyperparameter space through a grid search based on the average performance of the model over the inner folds. In particular, the learning rate was varied in the set 10-5, 3 $\times$ 10-5, 10-4, 3 $\times$ 10-4, 10-3 and the learning rate decay was varied in 0, 0.1, 0.3, 0.5. The chosen model is then fitted with all the outer fold training data and tested on the unseen test fold, resulting in an unbiased estimation of the model's prediction error. Specifically, a 10-fold CV has been chosen because it offers a favorable bias-variance tradeoff [238, 239]. In all experiments, batch size was equal to 128 epoch number was 50. Due to its ability to adaptively updating individual learning rates for each parameter, an Adam optimizer was used [240]. Each selected slice of the 3D T_1-weighted volume has been classified independently and the final model's performance was stated using the mean slice-level accuracy, separately, on the training set and test set folds of the outer CV. CNNs model's training and validation has been conducted on each dataset in a nested CV loop using two different data split strategies: a) subject-level split, in which all the slices of a subject have been placed either in the training set or in the test set, avoiding any form of data leakage; b) slice-level split, in which all the slices have been pooled together prior to CV, then split randomly into training and test set. In this case, for each slice of the test set, a set of highly correlated slices coming from the MR volume of the same subject ended up in the training set, giving

rise to data leakage, as shown pictographically in Figure 4.1. CNN models were carried out using a custom-made software in Python language (version 3.6.8) using the following modules: CUDA v.9.0.176, TensorFlow-gpu v.1.12.0, Keras v.2.2.4 [241], Scikit-learn v.0.20.2 [242], Nibabel v.2.3.3 and OpenCV v.3.3.0. All the code, model architectures and model weights will be made publicly available at time of acceptance of this manuscript. The training and validation of CNN models were performed on a workstation equipped with a 12 GB G5X frame buffer NVIDIA TITAN X (Pascal) GPU with 64 GB RAM, 8 CPUs, 3584 CUDA cores and 11.4 Gbps processing speed. The average computational time for CNN training on a dataset of 34 and 200 subjects were 5.68 hours (VGG16-v1), 5.63 hours (VGG16-v2), 2.94 hours (ResNet-18) and 33.93 hours (VGG16-v1), 33.82 hours (VGG16-v2), 14.12 hours (ResNet-18), respectively. The total computational time for this chapter was thus about 17 days.

4.5 Experimental results

The detailed performances of the three CNN models on all datasets are reported in Table 4.5. For AD classification, accuracies on the test set, using subject-level CV, were below 71% for large datasets (OASIS-200 and ADNI), whereas they were below 59% for smaller datasets (OASIS-34). As regards de novo PD classification, they were around 50% for both large (PPMI) and small (Versilia) datasets. Conversely, in all datasets, slice-level CV erroneously produced very high classification accuracies on the test set (higher than 94% and 92% on large

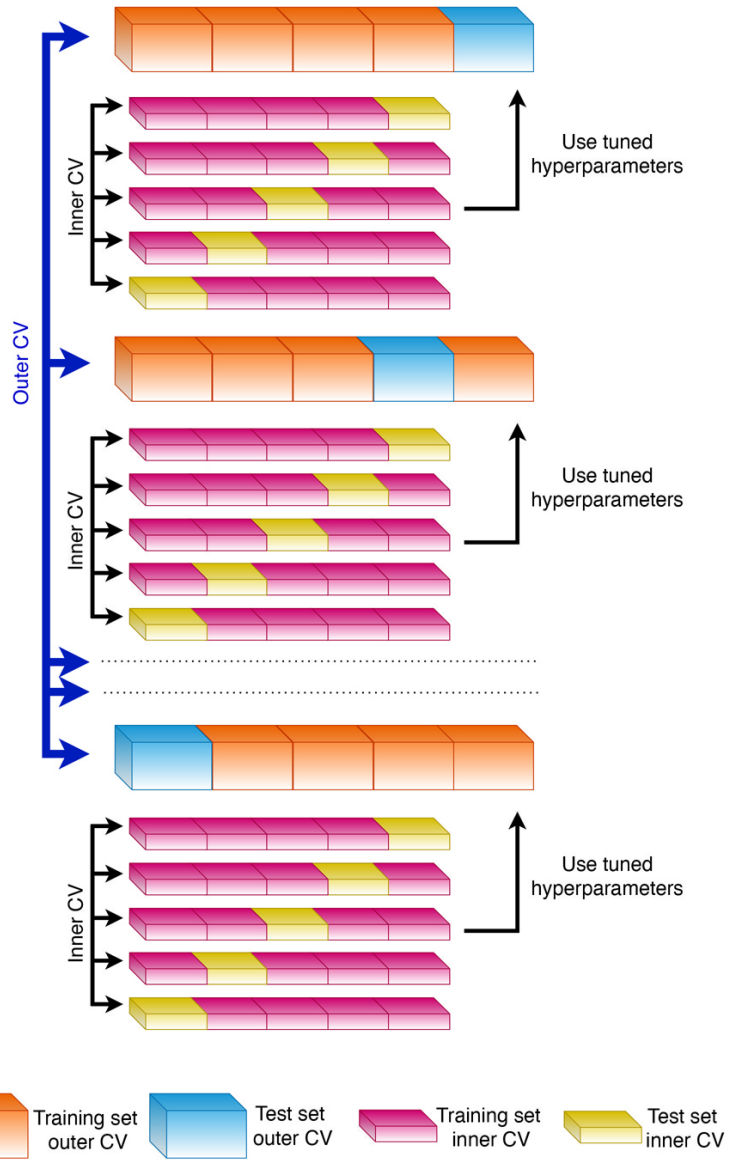


Figure 4.5: A scheme of nested CV is represented: the inner CV loop is used to optimize hyperparameters, whereas the outer loop estimates the selected models' performance.

and small datasets, respectively), leading to deceptive, over-optimistic results (Table 4.5). The worst-case stemmed from the randomly labeled OASIS dataset, which resulted in a model with unacceptably high performances (accuracy on the test set more than 93%) using

Dataset	Network architecture	Training set accuracy (%)		Test set accuracy (%)		
		Subject-level split	Slice-level split	Subject-level split	Slice-level split	Difference
OASIS-200	VGG16-v1	95.93	99.85	66.0	94.18	28.18
	VGG16-v2	95.13	100	66.13	96.99	30.86
	ResNet-18	100	100	68.87	98.96	30.1
OASIS-34	VGG16-v1	88.94	100	54.35	99.19	44.84
	VGG16-v2	96.94	100	54.34	99.33	44.99
	ResNet-18	100	100	57.49	98.96	41.47
OASIS-random	VGG16-v1	63.38	100	53.37	95.93	42.56
	VGG16-v2	69.17	100	49.25	94.81	45.56
	ResNet-18	84.49	99.09	50.8	93.74	42.94
ADNI	VGG16-v1	91.09	100	70.12	95.31	25.19
	VGG16-v2	80.49	100	66.49	95.24	28.75
	ResNet-18	100	100	68.68	96.87	30.19
PPMI	VGG16-v1	76.8	100	48.24	93.99	45.75
	VGG16-v2	73.19	100	46.93	94.37	47.44
	ResNet-18	100	100	48.06	96.12	44.06
Versilia	VGG16-v1	99.72	100	53.86	95.97	42.11
	VGG16-v2	76.89	100	42.97	97.8	54.83
	ResNet-18	99.90	95.13	51.36	92.63	41.27

Table 4.5: Mean slice-level accuracy on the training and test set of the outer CV over fivefold nested CV has been reported for three 2D CNN models (see "Materials and methods" section), all datasets, and two data split methods (slice-level and subject-level). Te difference between accuracy using slice-level and subject-level split in the test set has also been reported.

slice-level CV, whereas classifcation results obtained using a subject-level CV were about 50%, in accordance with the expected outcomes for a balanced dataset with completely random labels.

An additional experiment, similar to the one described in the previous chapter, was carried out to differentiate the prodromal stage of AD, known as MCI from HC. There are 21 participants with a CDR score of 0.5 in the OASIS 200 dataset used in this chapter. When the VGG16-v1 model is employed to distinguish those from HCs, subject-based split achieves 59% classification accuracy, VGG16-v2 achieves 62% whereas the Resnet18 framework achieves 64.4%. On the other hand, there are 24 participants with a CDR score of 0.5 in the ADNI dataset. The highest classification accuracy was achieved with the Resnet18 (65.4%, 61%

for VGG16-v1 and 58% for VGG18-v2).

It should be highlighted that the diagnosis of MCI was not the primary goal for this chapter, and that these findings are based on limited datasets that should be confirmed by larger investigations. Because there are no significant variations in the brain architecture of MCI patients compared to HCs, MCI diagnosis is a complex problem that should be studied as a separate research question. In Chapter 7, MCI diagnosis will be investigated in detail.

4.6 Discussion

The extent of the overestimation of the model's classification performance caused by the use of an inaccurate slice-level CV is quantified in this chapter, which is unfortunately common in the neuroimaging literature (see Table 4.1). More specifically, the performance of three 2D CNN models (two VGG variants and one ResNet-18) trained with subject-level and slice-level CV data splits has been demonstrated for the classification of AD and PD patients from HCs using T_1-weighted brain MRI data. The results revealed that pooling slices of MRI volumes for all subjects and then dividing randomly into training and test set leads to significantly inflated accuracies (in some cases from barely above chance level to about 99%). In particular, slice-level CV erroneously increased the average slice level accuracy on the test set by 40-55% on smaller datasets (OASIS-34 and Versilia) and 25-45% on larger datasets (OASIS-200, ADNI, PPMI). Furthermore, the t-test was used to determine whether the difference is statistically significant. In this test, a p-value of less than 0.05

neared zero, suggested that there is enough variation in the sample to account for probable mean differences.

An additional experiment has also been conducted in which all the labels of the subjects were fully randomised (OASIS-random dataset). Even under such circumstances, using the slice-level split, an erroneous 95% classification accuracy has been achieved on the test set with all models, whereas 50% accuracy has been found using a subject-level data split, as expected from a randomised experiment. This large (and erroneous) increase in performance could be due to the high intra-subject correlation among T_1-weighted slices, resulting in a similar information content present in slices of the same subject [243]. In AD classification, three previous studies [175–177], using similar deep networks (VGG16, ResNet-18 and LeNet-5, respectively), reported higher classification accuracies (92.3%, 98.0% and 96.8%, respectively) than ours. However, there is strong indication that these performances are massively overestimated due to a slice-level split. In particular, in one of this works [175], the presence of data leakage was further corroborated by the source code accompanying the chapter and confirmed by the data. In fact, when the same dataset of Hon and Khan [175] (OASIS-200 dataset) was used, the proposed VGG16 models achieved only 66% classification accuracy with subject-level split, whereas they boosted to about 97% with a slice-level split. Similar findings were presented by Wen et al. [186] who used an ADNI dataset with 330 HCs and 336 AD patients. Indeed, using baseline data, they reported a 79% of balanced accuracy in the validation set with subject-level split which

increased up to 100% with slice-level split. One of the main issues in the classification of neurological disorders using DL is data scarcity [215]. Not only because labeling is expensive, but also because privacy reasons and institutional policies make acquiring large sets of labeled imaging data even more challenging [244]. To show the impact of data size on model performance, 10 small subsets from OASIS dataset (OASIS-34 datasets) were created. As expected, when the data was reduced, lower classification accuracies were obtained with all the networks using the subject-level data split method. However, when the slice-level method was used, the models achieved erroneous better results on OASIS-34 than on OASIS-200 dataset. Similarly, models trained on the Versilia dataset (34 subjects) produced inflated results with the slice-level split. Overall, these results point out that the data leakage is extremely relevant especially when small datasets are used – a situation which may be unfortunately common in clinical practice. In this chapter, the effect of fine-tuning on model performance was also evaluated. To assess whether models could perform better when fine-tuning is started from the earlier layers, two different VGG variants, namely VGG16-v1 and VGG16-v2 were created. The former was used as fixed feature extractor – i.e. the pre-trained weights were frozen, thus also reducing the computational load. The latter was used as a weight initialiser – this feature may help to improve the performance depending on the size of the dataset and parameters. In the conducted experiments, the difference between the classification accuracies of the two VGG variants was very low. Indeed, Kandel and Castelli [245], showed that fine-tuning the top

layers could be sufficient for shallow networks such as VGG to achieve good results, whereas fine-tuning the entire network can produce better results for deep networks like InceptionV3. It is well-known that data leakage leads to inflating performance. Nevertheless, the degree of overestimation quantified through the experiments was surprising. Unfortunately, in the literature, the precise application of CV is frequently not well-documented and the source code is not available. This situation leaves the neuroimaging community unable to trust the (sometimes) promising results published. Regardless of the network architecture, the number of subjects, and the level of complexity of the classification problem, all experiments that applied slice-level CV yielded very high classification accuracies on the test set as a result of incorporating different slices of the same subject in both the training and test sets. This data leakage also yields to the concept of dataset shift in which the testing (unseen) data experience a phenomenon that leads to a change in the distribution of a single feature, a combination of features, or the class boundaries [246, 247]. There are various potential explanations for dataset shift, but leakage leads to two of the most important ones: sample selection bias and non-stationary settings. The difference in distribution in the first case is due to the fact that the training examples were collected using a biased approach, and so do not reliably represent the operational environment where the classifier is to be deployed. The second reason occurs when there is a temporal or spatial difference between the training and test environments [247].

Considering classifications on 2D MR images, in order to prevent data leakage and to get trustable results, it is crucial that the CV split to be done based on the subject-level. This assures the training and validation sets to be completely independent and confirms that no information is leaking from the test set into the training set during the development of the model. With recent advances in ML, more and more people are becoming interested in applying these techniques to biomedical imaging, and there is a real and growing risk that many of them will not be familiar with the possible issues and good practices. The need for documenting how the CV is built, the architecture utilized, and how the various hyperparameter choices/tunings are chosen, as well as presenting their values when available, is also stressed. Besides, it would be also necessary to make the source codes available to the neuroimaging community so that the results will be reproducible [248]. All the source code can be found in a Github repository at https://github.com/Imaging-AI-for-Health-virtual-lab/Slice-Level-Data-Leakage, and a Docker image can be downloaded at from here. The supplementary materials for all chapters are also presented in the repository including the subject IDs and associated demographics for all datasets.

4.7 Conclusion

In conclusion, training a 2D CNN model for analysing 3D brain image data should be performed using a subject-level CV to prevent data-leakage. The adoption of slice-based CV results in very optimistic model performances, especially for small datasets, as the extent

of the overestimation due to data leakage is severe. The main limitation of this chapter is that it does not assess all forms of data leakage, such as late split and hyperparameters optimisation in the test set, which are both likely to occur also in 3D CNN studies. Late split occurs when the data augmentation step is performed before separating the test set from the training data. In that case, the augmented data generated from the same original image can be seen in both training and test data, leading to inflated performance [186]. Still, using the same test set for optimising the training hyperparameters as well as evaluating the model performance is an additional form of data leakage [237]. Finally, data leakage also occurs when feature selection is performed based on the whole dataset prior to carrying out cross-validation [237,249]. An evidence of all these data leakage issues in the recent literature has been stated in Table 4.1.

Chapter 5

3D CNN for the classification of neurodegenerative diseases using structural MRI

This chapter describes an investigation of the classification accuracy based on three publicly available data sets, namely, ADNI, OASIS, and PPMI by building a 3D VGG variant convolutional network (CNN). 3D models have been used to avoid information loss and to further learn more abstract level spatial representation. A pre-processing stage has also been employed to enhance the effectiveness and classification performance of the model. The proposed model achieved 76.5% classification accuracy on ADNI, 71% on OASIS dataset and 66% on PPMI dataset with 5-fold cross-validation (CV). These results

are comparable to other studies using various convolutional models. However, the subject-based divided dataset has only one MRI of a single patient to prevent possible data leakage, whereas some other studies have different screenings of the same patients "over a time period" in their datasets. This chapter is based on [13].

5.1 Introduction

Throughout the last decade, multiple studies have been focusing on the automatic diagnosis of neurodegenerative diseases using different methods [250–252]. Among those, DL has come to the fore as one of the most promising tools to address neurodegenerative disease diagnosis and prognosis. In DL models, discriminative features may be extracted automatically from the raw data resulting in end-to-end learning design.

In this chapter, an end-to-end AD+MCI/HC and PD/HC classifier, which takes T_1 weighted MRI as input will be presented. A 3D VGG (a deep neural network model implemented by Oxford Visual Geometry Group (VGG)) variant CNN was implemented to overcome the limitations regarding the feature extraction from brain MRI and preserve spatial relations. Figure 5.1 provides an illustration of the network architecture.

The chapter is organised as follows: after this introduction, a brief of related work is given in Section 5.2. Section 5.3 provides the details of the proposed model, including the datasets and classification algorithm of CNN. Experimental results are presented in Section 5.5. Finally, Section 5.7 concludes the chapter with some final remarks.

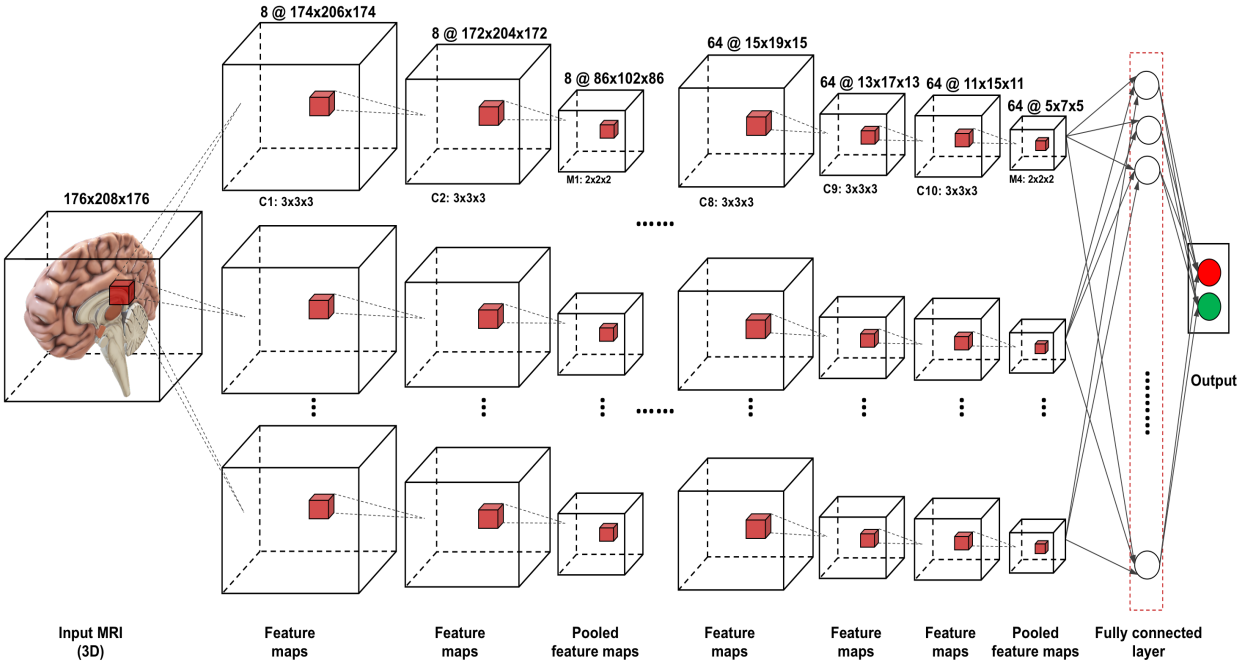


Figure 5.1: Overview of the 3D convolutional neural network (CNN) architecture. 3D boxes show input and feature maps.

5.2 Related Work

Various studies used a set of 2D slices extracted from the MRI volume as input to the 2D CNN architectures [11, 175, 176, 217, 253–255]. Farooq et al. [176] used a 2D CNN model for 4-way classification of Alzheimer's into AD, MCI, LMCI and HC using structural MR images. Sarraf et al. [177] utilised CNN and the famous architecture LeNet-5 to classify functional MRI data of AD's patients from HCs. In [175], Hon et al. used VGG16 and Inception V4 to classify AD using transfer learning. Finally, in 2019, Jain et al. [174] presented the CNN model for the 3-way AD classification. However, in most of these studies, it is not clear if

data division was done at the subject-level, calling into question the validity of the results due to potential data leakage [11, 186, 256]. Another possible problem in the 2D approach is the loss of information from 3D MRI when sliced and analysed by 2D convolutional filters.

Some studies addressed 3D networks to solve the issue of insufficient information in the 2D slice-level approach [257, 258]. Even though these models are computationally more expensive, they have a higher capability to extract discriminative features from three-dimensional MRI data. Korolev et al. [259] used 3D residual neural network architecture together with several regularisation techniques for AD classification. In 2018, Hosseini-Asl et al. [260], utilised a pre-trained 3D-Adaptive CNN classifier with used scans from the CADDementia dataset for the classification of AD vs. HC. However, the details regarding CV methodology and classification decisions are not presented in this chapter. Wang et al. [261] proposed an ensemble of 3D densely connected convolutional networks (3D-DenseNets) for three-class AD, MCI, and HC diagnosis. In their model, MRI scans of the same patients that are over three years apart are employed as different samples, incorporating information of test data into the learning process. Rieke et al. [262] trained a 3D CNN for AD classification accuracy. At the end of their visualisation efforts, they showed that the model focuses on the medial temporal lobe. Yang et al. [263] also provided visual explanations regarding the AD from Deep 3D CNNs. They utilised 3D-VGGNet together with 3D-ResNet. Finally, in 2019, Oh et al. [258] develop a volumetric CNN-based approach for the AD classification task.

On the PD side, Chakraborty et al. [264] recently developed a 3D CNN architecture for learning the intricate patterns in the MRI scans for the detection of PD. In 2021, Dhinagar et al. [265] proposed another 3D CNN model for classification of AD and PD, with similar findings.

It should be noted that the classification performances of these studies are hard to compare as they have trained and tested the models with different sets of participants. The studies also differ in terms of the pre-processing stages, hyperparameter selection, cross-validation (CV) procedure, and evaluation metrics.

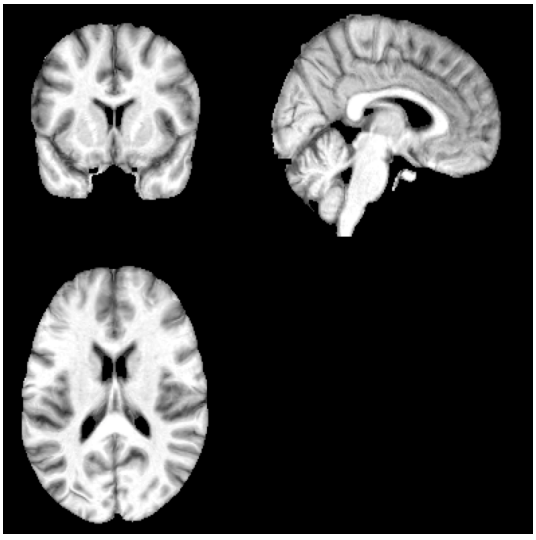
5.3 Methods

In this section, the main components of the proposed framework are presented. The pre-processing steps of T1 -weighted MRI data and the model details are briefly explained.

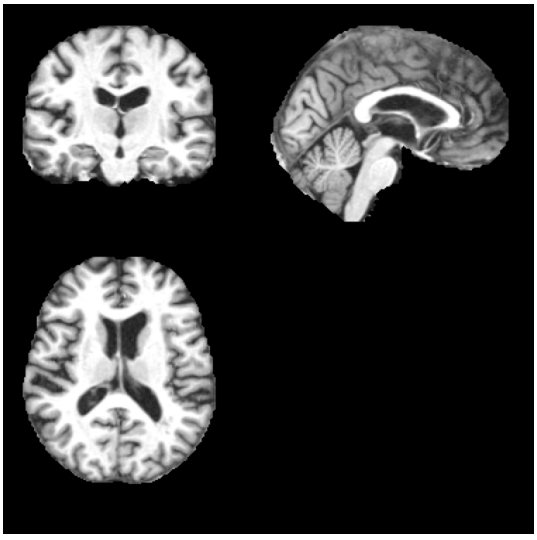
5.3.1 Data pre-processing

Even though CNN models do not require any pre-processing beforehand, an accurate image pre-processing stage could be key to increase the effectiveness of learning and help to achieve a good classification performance, particularly in the domain of MRI [266, 267]. All the data has been transformed into a standardised structure by performing co-registration with a standard template and skull stripping.

For ADNI, each T_1-weighted image has been co-registered with the SyN method using



(a) A sample T_1-weighted MRI slices of a Alzheimer's disease (AD) patient from ADNI dataset after pre-processing - in coronal, sagittal, and axial view (left, right and bottom respectively).



(b) A sample T_1-weighted MRI slices of a Alzheimer's disease (AD) patient from OASIS dataset after pre-processing - in coronal, sagittal, and axial view (left, right and bottom respectively)

Figure 5.2: Example of six magnetic resonance imaging (MRI) slices of two Alzheimer's Disease (AD) subjects from ADNI and OASIS databases [6, 7].

standard T_1-weighted template MNI152 at 1 mm [224]. After co-registration, the brain mask of the standard space has been applied to each volume to remove extracranial tissues. The final size of the ADNI T_1-weighted MRI volumes is 182 x 218 x 182 with 1mm x 1mm x 1mm voxel size.

When it comes to the OASIS dataset, gain-field corrected data was used. An additional brain masking and re-sampling operations are performed. The final dimension of the 3D volume is 176 x 208 x 176 with 1mm x 1mm x 1mm voxel size [223]. The sample MRI slices

from ADNI and OASIS datasets after the pre-processing stage can be seen in Figure 5.2.

The same pre-processing approach as described in Chapter 4.3.2 was used for the PPMI dataset.

5.3.2 CNN Models: 3D CNNs

A 3D CNN model inspired by VGG-16 architecture was created. The model differs from prior networks in the literature such as Voxnet [268] and its variants [269] with major modifications in terms of number of layers, filter sizes, pooling methods and input shape. As a side note, the work of Zunair et al. provided additional motivation to investigate the Tuberculosis predition problem (See Chapter 7) and test the models using different modalities [269].

The model has four convolutional blocks, among which the first two contain two convolutional layers each, and the latter two have three convolutional layers followed by a pooling layer with filter size 2x2x2. The overview of the 3D CNN architecture is shown in Figure 5.3. A convolutional and a pooling layer has several feature maps, and in most cases, the number of feature maps increases as layers grow. The calculation of the j^{th} feature map is given by:

$$y^j = f(W_j * x + b_j) \tag{5.1}$$

where y^j be the 3D array of the j^{th} feature map in a hidden layer, x be the 3D array of the input, b^j be the scalar bias and W^j be the 3D filter with a size of $w \times h \times d$. f corresponds to an activation function, and $*$ stands for the convolution operation. The convolution

operation $[W_j * x](m, p, q)$, is represented as follows:

$$\sum_{u=0}^{w-1} \sum_{v=0}^{h-1} \sum_{k=0}^{d-1} W_j(w-u, h-v, d-w) x(m+u, p+v, q+k) \tag{5.2}$$

After the convolutional blocks, a dropout layer with a probability of 0.5 is applied to avoid overfitting. It is followed by three fully connected layers with 128, 64, and 2 neurons, respectively. The last fully-connected layer with softmax activation provides the output label. The model has been trained with categorical cross-entropy loss and the Adam optimizer with a learning rate of 0.0001 and a batch size of 2 for 200 epochs. Binary cross-entropy loss is computed as:

$$L(y, p) = -(y \log p + (1-y) \log (1-p)) \tag{5.3}$$

where y is the actual label and p is the predicted label.

Training and validation of the proposed models were performed on a NVidia RTX2080 GPU.

5.4 Evaluation framework

This section provides a full overview of the dataset as well as a variety of assessment techniques. The data collection and dataset utilised in the experiments are discussed in further depth in 5.4.1, whilst model training and validation procedures are described in the

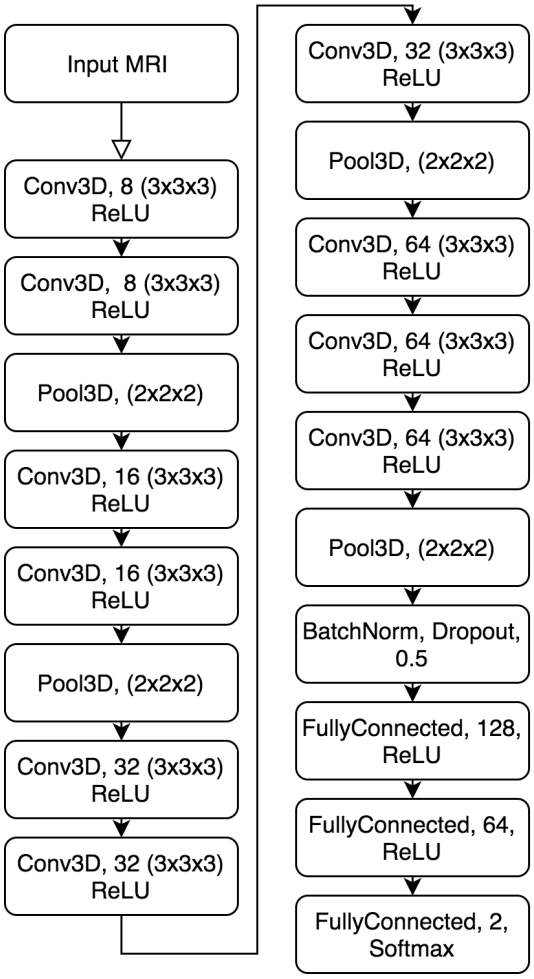


Figure 5.3: The architecture of the convolutional neural network (CNN) model used in the AD classification tasks.

subsection 5.4.2.

5.4.1 Datasets

In this chapter, two primary publicly available datasets on AD and related dementia are used: ADNI dataset [6] and OASIS [7] dataset together with a publicly available PD dataset. These

Dataset	Diagnosis	No. of patients	Sex	Age (mean ± SD)
ADNI	AD	100	44 F, 56 M	74.3 ± 7.9
	HC	100	52 F, 48 M	75.0 ± 7.1
OASIS	AD	100	59 F, 41 M	76.7 ± 7.1
	HC	100	73 F, 27 M	75.5 ± 9.1

Table 5.1: Demographic information of subjects from ADNI and OASIS datasets

datasets are described in detail below. The characteristics of the subjects included in this chapter are given in Table 5.1.

It should be emphasised that, as in Chapters 3 and 4, the decision criteria for AD in this work were that a variable named CDR with a value of 0 suggested HC and any value more than 0 inferred AD. As a result, from a clinical standpoint, the AD dataset contained MCI patients as well, because MCI is clinically staged at the 0.5 level on the CDR scale.

5.4.1.1 ADNI

In this chapter, a subset of ADNI 2 dataset with 200 structural T_1-weighted MRI scans was used. From ADNI 2 dataset, 200 subjects has been randomly picked, 100 of whom were chosen from the AD group (44 women and 56 men, age 74.28 ± 7.96 years, mean ± SD), while the other 100 from the HC group (52 women and 48 men, age 75.04 ± 7.11 years, mean ± SD). In order to make a comparative analysis between 2D models and 3D models, the same subjects (same number of patients and subject IDs) as in earlier studies (See Chapter 4) have been utilised. Only the first scan of each patient has been added to the dataset. Patients with a CDR score of 0 are labeled as HC subjects, whereas the ones whose CDR

rating higher than 0 are considered as AD subjects. MPRAGE T_1-weighted MR images have been acquired using 3 T scanners, and consisted of 176 × 240 × 256 (Siemens) and 170 × 256 × 256 (Philips) voxels with a size of approximately 1 mm × 1 mm × 1.2 mm.

5.4.1.2 OASIS

For the experiments, T_1-weighted MRI scans of 100 healthy subjects (73 women and 27 men, age 75.5 ± 9.1 years, mean ± SD) and 100 AD patients (59 women and 41 men, age 76.7 ± 7.1 years, mean ± SD) have been selected to create a subset of OASIS-1 dataset. Again, the CDR score was 0 for the HC subjects, 0.5 (very mild), 1 (mild), 2 (moderate), and 3 (severe) were for the AD subjects. MPRAGE T_1-weighted MR images have been acquired using a 1.5 T Siemens scanner. They are in the size of 256 × 256 × 128 with voxel size 1 mm × 1 mm × 1.25 mm.

5.4.1.3 PPMI

From the publicly accessible PPMI dataset 100 de-novo PD subjects (40 women and 60 men, age 61.71 ± 9.99, mean ± SD) and 100 HCs (36 women and 64 men, age 61.91 ± 11.52, mean ± SD) have been selected. Marek et al. [234] developed the criterion utilised to enrol de-novo PD patients and HCs. The T_1-weighted scans were collected at baseline using MR scanners manufactured by Siemens (11 scanners at 3 T and 5 scanners at 1.5 T), Philips Medical Systems (10 scanners at 3 T and 11 scanners at 1.5 T), GE Medical Systems (11

scanners at 3 T and 24 scanners at 1.5 T) and another anonymous one (5 scanners at 1.5 T). It has been found that three subjects had missing MRI protocols.

5.4.2 Model training and validation

The model has been evaluated using five-fold CV. The average accuracy is obtained by repeating 5 times the full 5-fold cross-validation starting from five different splits of the data into folds. The architecture was built using Keras with TensorFlow backend [199, 241].

5.5 Experimental results

The model was tested on three different test sets, each of which contains 40 subjects. Using 5-fold CV, the model achieves $(76.5 \pm 0.09)\%$ (mean $\pm$ standard deviation) on ADNI dataset, $(71.0 \pm 0.03)\%$ (mean, standard deviation) classification accuracy on the OASIS dataset, and $(66.5 \pm 0.08)\%$ (mean $\pm$ standard deviation) on PPMI dataset. The results are comparable to other studies that use different convolutional models for AD+MCI vs. HC and PD vs. HC classification.

Dataset	Accuracy	Sensitivity	Specificity	Precision	F1-score
ADNI-200	0.76	0.73	0.80	0.78	0.75
OASIS-200	0.71	0.70	0.72	0.71	0.70
PPMI-200	0.66	0.66	0.67	0.66	0.66

Table 5.2: The model's performance on different dataset.

A further MCI vs. HC classification experiment was also carried out to differentiate the

prodromal stage of AD. In the OASIS AD+MCI dataset presented in this chapter, there are 21 persons with a CDR score of 0.5. When the 3D CNN model is used to identify those from HCs, it obtains a classification accuracy of 62%. The ADNI dataset, on the other hand, has 24 persons with a CDR score of 0.5. To identify MCI from HC, the 3D CNN model attained a classification accuracy of 64%. As expected, in this experiments the model performance degraded due to a small training dataset. As noted in earlier chapters, MCI diagnosis is a difficult subject that requires further exploration because there are no significant variations in the brain architecture of MCI patients compared to HCs. The diagnosis of AD and its prodromal stage will be the main focus for the next two chapters of this thesis.

5.6 Discussion

All the datasets were divided by subjects, and only one screening of a patient was included in the dataset in order to prevent possible data leakage. For instance, Rieke et al. [262] reported (78 ± 0.04)% classification accuracy with a similar architecture using ADNI 1 datasets, which contains MRI scans of the subjects up to three-time points (screening, 12 and 24 months; sometimes multiple scans per visit). Following such procedure may cause the scans of the same subject to be in both testing and training set, which could affect the model performance. Moreover, PD diagnosis is known to be more difficult than AD classification since conclusive diagnosis of PD may necessitate the use of other imaging modalities, such as DAT-PET, to complement structural characteristics from T_1-weighted brain MRI.

5.7 Conclusion

In this chapter, a deep 3D CNN model has been presented for the diagnosis of AD+MCI and PD patients using structural brain MRI. The model performance was demonstrated on two primary AD datasets, namely ADNI and OASIS and one PD dataset called PPMI. Without any feature extraction stage, the model managed to achieve $(76.4 \pm 0.09)\%$ (mean, standard deviation) and $(71.0 \pm 0.03)\%$ accuracy for classification of AD subjects from HC on ADNI and OASIS datasets respectively. In future work, it is desired to expand this chapter and archive better classification accuracy through optimising the network shape and hyperparameters. Moreover, explainable AI techniques would be used to investigate which brain regions or patterns are most important to the model and shed light on the rationale behind the model's predictions.

Chapter 6

Autoencoder based deep neural network architectures for automated diagnosis

Rapid and accurate diagnosis of AD is critical for patient treatment, especially in the early stages of the disease. While computer-assisted diagnosis based on neuroimaging holds vast potential for helping clinicians detect disease sooner, there are still some technical hurdles to overcome. This chapter presents an end-to-end disease detection approach using convolutional autoencoders by integrating supervised prediction and unsupervised representation. The 2D neural network is based upon a pre-trained 2D convolutional autoencoder to capture latent representations in structural brain MRI scans. Experiments

on the OASIS brain MRI dataset revealed that the model outperforms a number of traditional classifiers in terms of accuracy using a single slice. This chapter is based on [270].

6.1 Introduction

In neurodegenerative diseases research, the clinical understanding of neuroimaging scans can be complex, as brain modifications can be challenging to discern from those due to healthy ageing. Especially in the early stages of an illness, detecting disease-related changes from MRI scans could be extremely problematic. Thus, in the last few years, there has been a research interest in modeling the deviation of brain structure due to neurodegeneration [271].

Among those, DL-based approaches quickly stand out as they automatically discover discriminative features in the training data collection even when the raw data is used as input [154]. Here, in medical image analysis, one of the biggest challenges is the high dimensionality of the input [272]. For instance, even though there are only several hundred MRIs in the OASIS, each image has more than six million dimensions ($176 \times 176 \times 208$) [7].

DL models are subject to noise and redundant information encoded in high-dimensional data, which can lead to unstable and erroneous predictions [273]. Training a supervised DL model with high dimensionality and low-quality image data might result in overfitting and/or unstable behaviour, particularly when training data is scarce or uneven [274]. For this reason, just the most essential information from the data needs to be collected [275].

Autoencoders are an unsupervised dimensionality reduction technique that has been demonstrated to filter out noise and redundant information while providing robust and stable feature representations [276].

In this work, to parse neuroanatomical alterations in AD, an end-to-end DL approach has been proposed based on deep convolutional autoencoders (CAE) using MRI [277, 278]. An autoencoder is an artificial neural network built to recreate its input [279]. Deep CAEs consist of two parts. The first component, which is the encoding function of the model, learns how to compress the original input in a latent representation. The second part, known as the decoder, learns to recreate the input data as near as possible to the original using the latent representation [280]. In this work, a 26-layer deep CAE model has been used to retrieve a lower-dimensional representation of the data, which contains all the important information needed to describe the original data point. Then, those latent space representations extracted from brain MRI data are used to differentiate subjects with AD and MCI from HC. It has also been searched for cerebral atrophy patterns to discover the early changes in the brain characterising AD. To this end, the intermediate activations were visualised across different convolutional layers to understand why the model makes certain decisions. A 26-layer deep CAE model has been chosen based on the work of Teganya and Romero [281] as they have observed that the best performance in this case is achieved around number of layers equal to 26 layers.

The rest of the chapter is organised as follows: The following section 6.2 briefly describes

the related work in the literature. In section 6.3, the detailed methodology of the proposed framework is presented. Section 6.4 explains evaluation methods and section 6.5 presents the obtained results and following by discussion. Finally Section 6.6 discusses the results and Section 6.7 concludes this work.

6.2 Related works

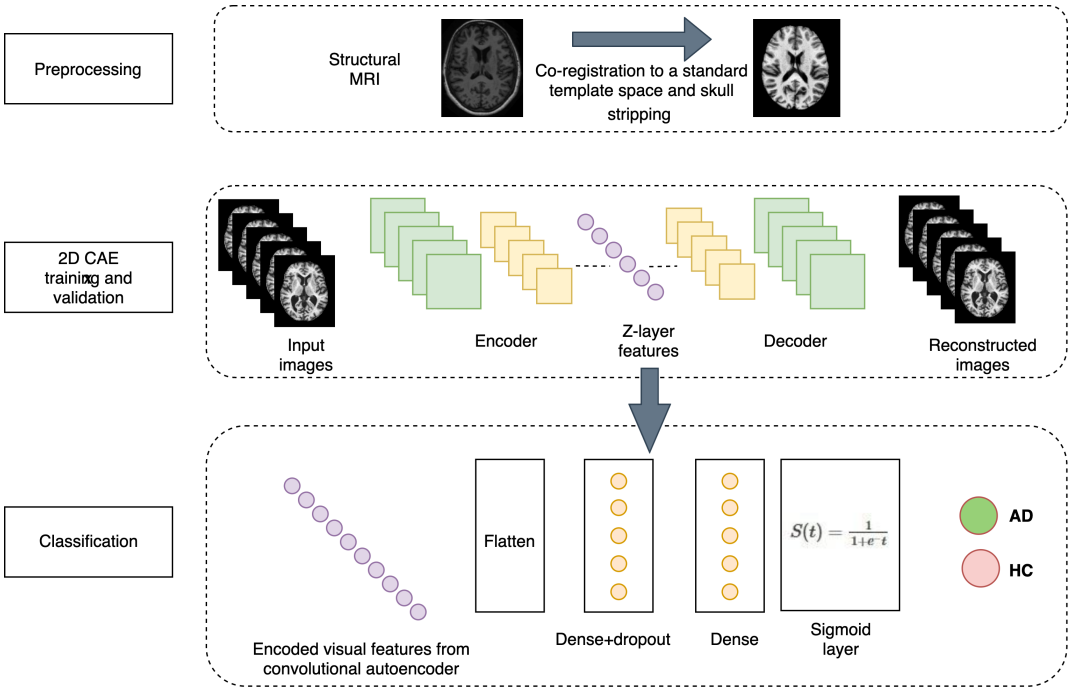


Figure 6.1: Overview of the proposed autoencoder based deep neural network architectures for automated diagnosis.

A set of 2D slices extracted from the MRI volume was used as input to 2D CNN architectures in various studies for the purpose of AD

diagnosis [11, 175, 176, 217, 253–255, 282]. Among those, not many studies explored the possibility of integrating CAE into their framework to learn an efficient representation of data. Martinez et al. [283] proposed a deep CAE architecture to extract data-driven features and stated that in the case of neuropsychological assessment variables such as the Mini-Mental State Exam (MMSE) or the AD Assessment Scale (ADAS11) ratings, imaging-derived markers could forecast clinical variables with correlations above 0.6 [284, 285]. In 2020, Oh et al. [258] used volumetric CAE-based unsupervised learning for the AD vs. HC classification task, then applied supervised transfer learning to solve the progressive mild cognitive impairment (pMCI) vs. stable mild cognitive impairment (sMCI) classification task. Basu et al [286] proposed a model which consists of a 3D convolutional variational autoencoder and a Multi-Layer Perceptron (MLP) to predict the likelihood of the next disease label. Lastly, in 2021 Ferri et al. [287] presented an ANN with stacked autoencoders to differentiate AD and HC using resting-state electroencephalogram (rsEEG), MRI, and rsEEG + MRI features.

6.3 Methods

An end-to-end AD diagnostic framework that extracts latent representations for each class from a brain MRI with a 2D-CAE and then performs classification with a stacked CNN is proposed. The methodology is structured by two main components: 2D CAE training/validation for latent space representation and disease classification using latent

representation, as shown in Figure 6.1. In section 6.3.1 the data selection procedure together with the pre-processing steps are described. CAE architecture and training strategy is illustrated in section 6.3.2, followed by proposed classification approach in section 6.3.3. Finally, visualisation of activations is described in section 6.3.4.

6.3.1 MRI pre-processing

The publicly available pre-processed version of OASIS data (gain-field corrected, brain masked, and co-registration) [223] has been used in the experiments. In that version, an atlas-registration-based method was used to create the OASIS brain masks [7]. The Talairach and Tournoux atlases were also used for co-registration of each volume. The data matrix size of each pre-processed T_1-weighted volume was $176 \times 208 \times 176$, and the voxel size was $1mm \times 1mm \times 1mm$ (see Han et al. [223]). From these volumes, the middle axial slice (the 106th) has been selected as input for the models. In the work of Mendoza-Leon et al. [288], it has been shown that this axial location corresponds to the anatomical slice, which has a higher degree of disease-associated information due to its high individual content-based image retrieval performance. When the disease label was used as the criteria of interest, the performance results are evaluated by mean average precision values for axial plane [289]. This finding was interpreted as an indication of a higher degree of disease-related knowledge, making them good candidates for a single-slice classification method. However, it should be noted that the index of the selected slice is heavily

dependent on the dataset/atlas used. Mendoza-Leon et al. [288] have been used the same dataset; therefore, the same procedure has been followed while selecting the single slice candidate. Before feeding the network, the MRI slices in the dataset are normalised in the range $[0, 1]$ to receive an unvaried contrast and intensity range.

6.3.2 Convolutional autoencoder

In the proposed model, the encoder has three convolution blocks, where each block has a convolution layer (a kernel size of 3×3) followed by a batch normalisation layer. After the first and second convolution blocks, a max-pooling layer (a kernel size of 2×2) is used to downsample the output features of the convolutions. In the decoder, there are two convolution blocks with convolutional layers (a kernel size of 3×3) with ReLU activations and batch normalisation layers. Here, upsampling layers (a kernel size of 2×2) are used after the first and second convolution blocks. Moreover, batch normalisation is used to standardise the layer's input for each mini-batch and stabilise the learning process. The details of the network can be seen in Figure 6.2.

6.3.3 Classification model

In the classification part, the exact same encoder architecture that was used in the convolutional autoencoder model has been employed. After the last convolutional layer of the encoder, there is a 'flatten' layer in which the two-dimensional matrix of features is

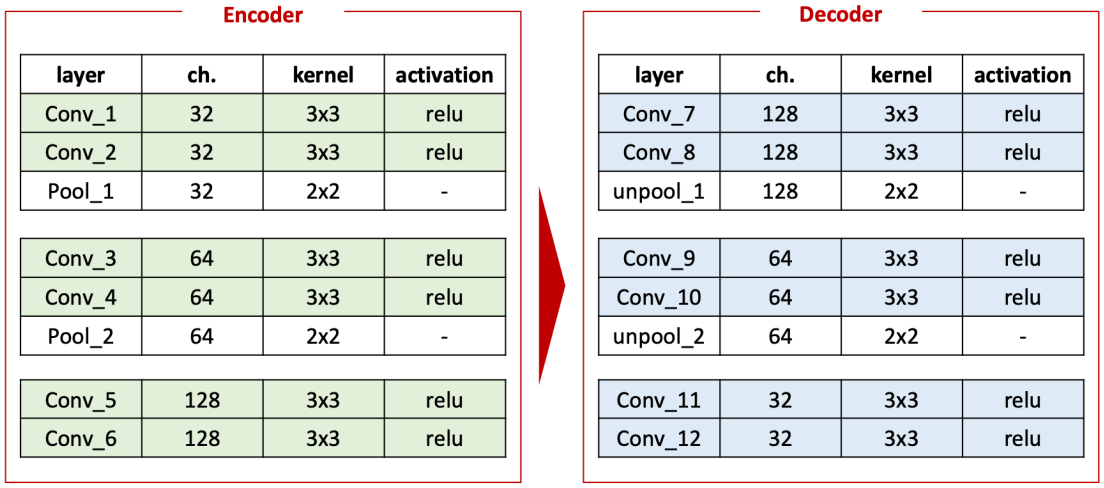


Figure 6.2: Detailed architecture of the proposed convolutional autoencoder.

flattened into a vector that can be fed into dense layers. Flatten layer is followed by two dense layers with 256 and 128 nodes, respectively. A dropout of 0.2 was added to the first dense layer together with ReLU activation. In the output layer, the sigmoid activation function has been used. In the training process, the premier step was freezing the first 15 layers coming from the pre-trained autoencoder and train only the dense layers. Then, all the layers were fine-tuned in the second stage. After multiple trials and errors, the optimal hyperparameters were determined. The model has been trained for 400 epochs each time with a batch size of 32 using Adam optimizer with a 0.001 learning rate. Binary cross entropy has been used as a loss function. As overfitting was a big concern due to the proposed relatively complex model with a small dataset, dropout regularisation has been implemented to prevent the network from overfitting.

6.3.4 Visualisation

In the literature, several methods for comprehending and visualising convolutional networks have been created mostly to interpret the learned features in a neural network [141,290,291]. In the context of proposed research, the activations or, in other words, feature maps of the network during the forward pass, have been used. Feature maps are created by applying filters to the input image or the feature map output of the previous layers. The internal representations for the input for each of the layers in the model are shown by visualisation.

The effect of applying the filters in the first convolutional layer, as seen in Figure 6.3, is a variety of representations of the axial brain image with various features illuminated. Some draw attention to shapes, while some concentrate on the background or the foreground. The feature maps closest to the model's input catch a lot of fine detail in the picture, while the feature maps reveal less and less detail as we go further into the model.

6.4 Evaluation framework

This section reports a detailed description of the dataset together with the range of evaluation methods. In the subsection 6.4.1, the data collection and dataset used in the experiments are described in more detail, whereas in the subsection 6.4.2 model training and validation strategies are presented.

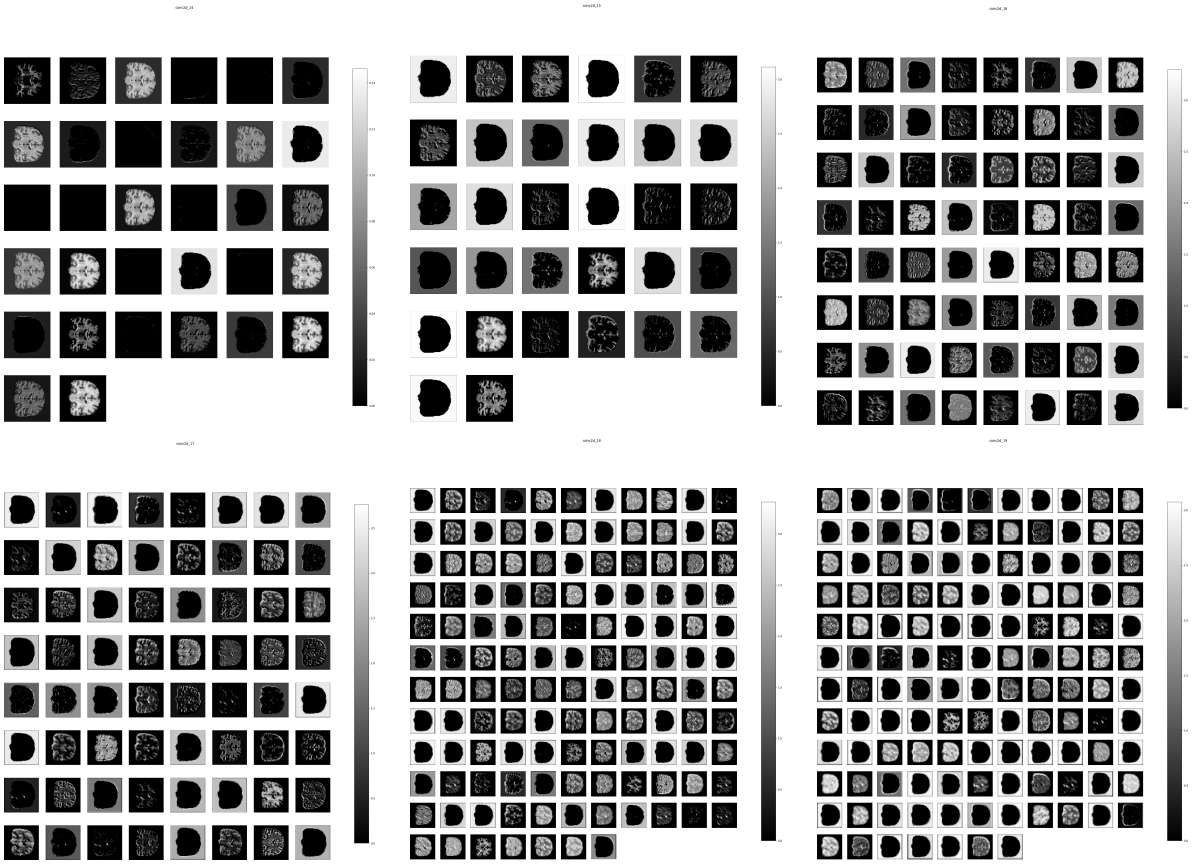


Figure 6.3: Visualisation results of selected convolutional layer feature maps. First row, from top to bottom: first, second and third convolutional layers. Second row, from top to bottom: fourth, fifth and sixth convolutional layers.

6.4.1 OASIS dataset description

In this chapter, the publicly available OASIS dataset[1], has been used [7]. The T_1-weighted images of 100 AD patients and 100 HCs have been selected from the OASIS-1 chapter – a cross-sectional cohort of the OASIS brain MRI dataset [7]. In the dataset, there was no substantial difference in age ($p = 0.15$ at t-test), but there was a significant (borderline) difference in gender ($p = 0.04$ at χ^2-test) between the two classes. The clinical characteristics of the subjects included in this chapter are summarised in Table 6.1. T_1-weighted images were acquired on a 1.5 T MR scanner (Vision, Siemens, Erlangen, Germany) in the sagittal plane using a Magnetisation Prepared Rapid Gradient Echo (MPRAGE) series [7].

The global CDR score derived from individual CDR ratings is used in OASIS-1 to assess the diagnosis of AD as well as the seriousness of the disorder. On the CDR scale, MCI is staged at the 0.5 mark. In the scope of the conducted experiments, HCs had the CDR scores 0, while scores of 0.5 (very mild), 1 (mild), 2 (moderate), and 3 (severe) were all labeled as AD. For that reason, from the clinical perspective, the AD dataset included MCI patients as well since MCI is staged clinically at the 0.5 level on the CDR scale, making classification task more challenging compared to AD vs. HC. In the AD+MCI dataset used in this chapter, there are 21 subjects whose CDR score is 0.5.

[1] `https://www.oasis-brains.org/`

Dataset	Diagnosis	No. of patients	Sex	Age (mean ± SD)
OASIS	AD	100	59 F, 41 M	76.7 ± 7.1
	HC	100	73 F, 27 M	75.5 ± 9.1

Table 6.1: Demographic features of subjects belonging to OASIS dataset.

6.4.2 Model training and validation

The MSE (see Equation 7.3) has been used as an evaluation measure to show how well the AE is capable of reconstructing unseen images. Moreover, in the scope of this experiment, the peak signal-to-noise ratio (PSNR) has been calculated as a quality measurement between the original and a reconstructed image. A measure of image quality is required when comparing reconstructed outcomes. MSE and PSNR ratio are two widely used metrics. One drawback of MSE is that it is highly dependent on image intensity scaling. By scaling the MSE according to the image range, PSNR prevents this problem, and it is calculated as:

$$\mathrm{PSNR} = 10 \log_{10} \frac{{\mathrm{Max_I}}^2}{\mathrm{MSE}} \tag{6.1}$$

where Max_I is the maximum pixel value.

To measure the prediction performance of the model, accuracy and F1 score have been used as evaluation metrics.

For AD+MCI classification task, the given model is fit to the training data for 400 epochs with a batch size of 32. Out of 200 subjects, 140 (70 AD, 70 HC) were picked for training the autoencoder. 20% of the training data (28 subjects) were used as validation to control

model generalisation, and to interrupt training when generalisation stops improving. Thus, train data shape is (112, 176, 176, 1) whereas validation data shape is (28, 176, 176, 1). The remaining 60 subjects have been chosen to use in the subsequent experiments for testing the full model with unseen patients while avoiding data leakage (see Chapter 3 and 4. A first-order gradient-based optimisation algorithm called Adam has been utilised with adaptive learning rates (alpha = 0.0001, beta1 = 0.9, beta2 = 0.999).

The average accuracy is obtained by repeating the full experimentation five times. The architecture was built using Keras (v2.3.1) with TensorFlow backend (v2.0.0) [199,241]. The training and validation of CNN models were performed on a workstation equipped with an NVidia RTX2080 GPU. The average computational time for model training was 3.2 hours.

6.5 Experimental results

In this section, the detailed performance of the end-to-end stacked autoencoder model is reported together with the disease prediction performance.

6.5.0.1 Reconstruction capability

The image quality of the restored image increases as the PSNR grows. In the experiments, 33.34 dB PSNR of reconstructed validation images has been achieved. The visualisation of sample test images and reconstructed test images can be seen in Figure 6.4.

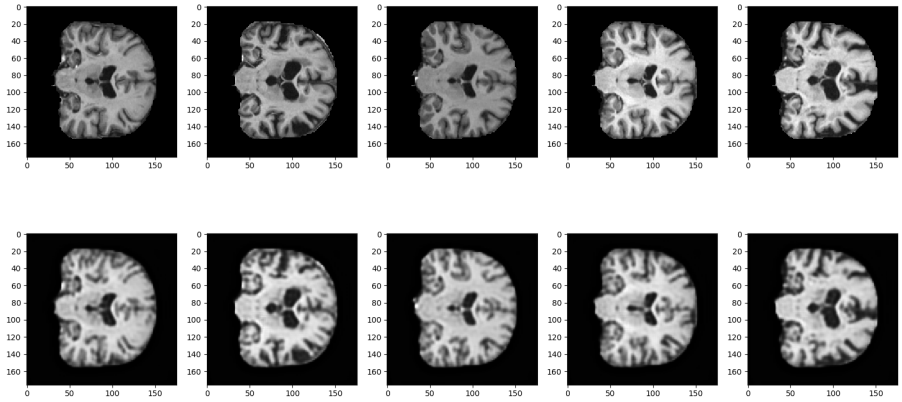


Figure 6.4: Sample test images (above) and reconstruction of test images (below) using the autoencoder based reconstruction approach.

6.5.0.2 Performance of the classification

The mean percentage ± standard deviation of accuracy, over 400 iterations are reported in Table 6.2. The autoencoder-based classification model achieves 77 ± 0.02 with an F1 scores of 0.74 and 0.79 for the AD+MCI vs. HC diagnosis task. When the autoencoder-based model is used to detect the prodromal stage of AD, 64.3% classification accuracy is achieved for MCI vs. HC task.

Class	Accuracy	Sensitivity	Specificity	Precision	F1-score
0 (HC)	0.77	0.67	0.87	0.84	0.74
1 (AD+MCI)	0.77	0.67	0.87	0.72	0.79

Table 6.2: Classification performance on the test set. The accuracy, sensitivity, specificity, and F1 scores for each class are listed.

6.6 Discussion

A single 2D slice of MRI volume has been used for each subject in the framework, which provided many functional benefits. First, it reduced the computational time and resources drastically as the processing complexity, and memory bandwidth demands of 2D CNN models are smaller than 3D CNN models [292]. Second, by using 2D MRI slices, clinical researchers could take advantage of the most recent CNN architectures, which are often implemented in 2D due to the availability of large 2D image datasets such as ImageNet, CIFAR, and so on [293].

Given the scale of the networks and the small amount of data used, overfitting was a major concern. In order to prevent the network from overfitting, the dropout method is adopted. Moreover, during the experiments, it has been realised that after epoch 400, the error on training data kept decreasing, whereas validation loss started to increase to a considerable value. Thus, early stopping is employed to avoid overfitting during training.

MCI patients are considered to be at a higher risk of dementia, primarily of Alzheimer type [294]. MCI is clinically challenging to distinguish from cognitively stable HC. Thus, in most AD diagnosis frameworks, they are not added to the dataset, which explains the higher accuracy reported in previous works [283].

Compared to other methods where multiple MRI slices are used with well-known deep 2D CNN architectures such as VGG16 and ResNet50, the proposed method outperforms in terms of diagnosis accuracy [11]. The experimental results show that when latent representations

are learned in a way that promotes sparsity, classification accuracy improves [295].

6.7 Conclusion

An early and precise AD diagnosis is needed, and an automated diagnostic tool helped by neuroimaging data analysis would offer a more detailed and effective solution, as well as potentially increase diagnostic precision. In this chapter, a CAE-based DL method has been presented for classifying AD+MCI vs. HC subjects using single 2D brain MRI slices. Experimental results on the test set demonstrated the effectiveness of the proposed method in the classification of AD+MCI vs. HC. As far as classification accuracy is concerned, the proposed approach outperforms the conventional approach where deep CNNs use whole MRI slices as input instead of the latent representations [11]. By integrating supervised prediction and unsupervised representation together, the model achieves 77% classification accuracy using only one MRI slice for each subject.

Chapter 7

Ensemble deep learning methods for automated diagnosis

In recent years, DL models have demonstrated dramatically higher performance on a variety of medical imaging tasks, such as cancer classification [296–298], organ and tumour segmentation [299, 300], or AD and PD detection [13, 214] including TB detection and diagnostics. However, the overall performance of a DL classification system is strongly reliant on the quantity of training data, and data may be sparse for many medical image classification applications including TB diagnostics. Transfer learning, data augmentation, and architectural modifications such as dropout have been used widely in the field to deal with data limitations [301]. Furthermore, DL models also often have high variance and can become stuck in local loss minima during training. Ensemble approaches that aggregate

the output of numerous DL models have been proven to have more generalizability than a single model, according to empirical studies. Ensemble methods may improve the robustness of the predictions and the model while lowering variance and bias [302].

In this work, a novel end-to-end ensemble DL architecture is proposed for automated diagnosis together with two independent models. The model classifies patients by combining supervised prediction with unsupervised representation. These parts are interconnected and trainable from end-to-end, providing a direct link between raw data and the clinical outcome of interest. This approach can potentially be converted into a valuable extra tool to assist physicians. As a result, our objective was to create a robust and generalisable automated diagnostic approach in the semi-supervised learning framework. Part of this chapter is the result of a collaborative effort with the Prince of Songkla University and itt has been submitted to IEEE Journal of Biomedical and Health Informatics.

7.1 Introduction

The proposed approach was first evaluated on ADNI-200 and OASIS-200 datasets. On top of these, an additional larger dataset was created from ADNI database to perform three additional binary classifications: AD vs. HC, MCI vs. AD and MCI vs. HC as OASIS-200 and ADNI-200 contained only 21, 24 MCI patients, respectively.

In the second half of this chapter, another research question is addressed: Could the presented approach be applied to conduct classification on other diseases? With this aim,

the semi-supervised ensemble method has also been tested on tuberculosis detection. Tuberculosis (TB) is an airborne infectious disease caused by the bacillus Mycobacterium tuberculosis (Mtb) [303]. Mtb is predominantly a pulmonary pathogen, but it can infect practically any part of the body [304]. Medical science has assisted in effectively treating TB infections since the middle of the nineteenth century [305]. Unfortunately, TB continues to have a devastating impact on people's lives. According to the World Health Organization (WHO), an estimated 10 million people fell ill with TB in 2019, of whom 1.4 million died of it [306]. In 2019, WHO reported that Thailand was still ranked within the 30 high TB burden countries with an incidence rate of 150 cases per 100,000 populations [307]. Not only general populations who infected with TB but also in the healthcare workers section who work in patient care units at the hospital suffered from the impact of TB [307, 308]. Songklanagarind Hospital[1], the most prominent university hospital in Southern Thailand, is still facing challenges in controlling and managing TB cases with over 96,000 chest X-ray (CXR) images taken in 2019. In Thailand, the reported TB diagnosis was only 59% of the expected number of cases.

As for the diagnostic tests of TB, the findings of a sputum smear take a few days, while the results of culture take a few weeks [309]. These standard tests not only cause delays in isolating infected patients but also have limited sensitivity [310]. Accordingly, medical chest imaging is a vital tool for the early detection of active TB disease. Even though computed

[1]Songklanagarind Hospital: `https://hospital.psu.ac.th`

tomography (CT) has recently attracted more attention, conventional radiography remains the standard of practice, especially in low and medium resource areas [311] such as Thailand as it is more accessible. Following WHO recommendations, Thailand screens TB nationwide using CXR [312]. Approximately $80,000 - 90,000$ CXR images are obtained annually and collected in the Songklanagarind Hospital Information System (HIS). The volumes have exceeded the capability to be interpreted by the radiologists, where the general practitioners have to interpret the CXR instead, which may lead to possible misdiagnosis.

While early diagnosis and treatment can accelerate progress in the management of TB and minimise its complications, the lack of medical expertise, particularly in low-resource settings, can lead to misdiagnosis and poor detection rates. Missed or delayed diagnosis of TB can have devastating consequences for patients and the community by slowing treatment, prolonging the duration of infectivity, increasing disease transmission, and driving up medical expenses and mortality rates [313]. Thus, a robust Artificial Intelligence (AI) tool based on automated diagnosis to analyse CXR is an alternative, cost-effective, and rapid method to combat this disease. In addition, it could reduce radiologists' workload and increase the accuracy of CXR interpretation. In this chapter, the proposed methods has been tested on two publicly available datasets provided by the National Library of Medicine - National Institutes of Health (NLM/NIH) of the United States of America (USA): the Montgomery County dataset and the Shenzhen No.3 People's Hospital dataset [314]. Moreover, a private clinical dataset was used to evaluate the proposed methodology. This dataset was generated

at the Songklanagarind Hospital in Thailand. A detailed description of each is reported in the "Evaluation Framework" section.

The rest of this chapter is structured as follows: Section 7.2 goes over the previous works in the literature including both AD/MCI and TB, Section 7.3 presents details of the materials and methods including the datasets used for both AD/MCI and TB classification while Section 7.5 part 7.5.1 illustrates the results of AD/MCI experiments and part 7.5.2 presents TB experiments. Finally, Section 7.6 discusses the results in detail and Section 7.7 concludes this work.

7.2 Related works

One of the vital tools that have been used to diagnose diseases is medical imaging. Especially information obtained from medical imaging plays an essential role in patient care procedure and further interventional procedures guidance. Therefore with this increase in workload on both radiologist and physicians, automated, computer-based analysis of medical images emerged as early as 1967 [315]. Ensemble learning, in addition to CNNs, has been found to be useful in medical imaging analysis. The typical 3D form of medical imaging data, as well as its frequent limited availability, might pose a hurdle when training classifiers [316]. By mixing numerous trained models, ensemble learning can be used to circumvent these restrictions [317].

7.2.1 Deep learning methods for AD/MCI classification

In the detailed review of Logan et al. [317], it has been illustrated that in addition to CNNs, the following model architectures have been proposed for AD classification using ensemble approaches: hierarchical ensemble learning with deep neural net [255], learning-using-privileged-information (LUPI) algorithms [318], sparse regression models [319], and instance transfer learning [320].

In 2020, An et al. [321] introduced a deep ensemble learning framework with the goal of leveraging deep learning algorithms to integrate multisource data and tap the 'knowledge of experts.' At the voting layer, two sparse autoencoders are trained for feature learning in order to reduce attribute correlation and, eventually, diversify the basic classifiers. As a meta classifier, the neural network is utilized. Venugopalan et al. [322], on the other hand, utilized stacked denoising auto-encoders to extract features from clinical and genetic data, and employed 3D-convolutional neural networks (CNNs) for imaging data in their framework. They identified the hippocampus, amygdala brain regions, and the Rey Auditory Verbal Learning Test (RAVLT) as the main distinguishing traits of AD, which are congruent with existing AD literature.

7.2.2 Deep learning methods for TB classification

In 2017, Liu *et al.* [323] presented a method using CNN for TB detection in a large imbalanced and less-category dataset to classify TB cases. The data was trained using shuffle sampling

with cross-validation with 85.68% accuracy on TB classification.

Yadav *et al.* [324], proposed a TB screening system with the FastAI tool that provides a quick modification and mixed and matched low-level components to crated a new approach [325] for CXR image. Their learning model was trained on the NIH Chest X-ray dataset of 14 common thorax diseases including TB. The learning model initially learned at low image resolution before increasing the resolution while training using the coarse-to-fine knowledge transfer technique. The authors employed the ResNet-50 model [5]. Lakhani and Sundaram [326] presented deep convolutional neural networks (DCNNs) and obtained their best result using an ensemble of the AlexNet and GoogLeNet DCNNs. In 2018, Li *et al.* [327] proposed a CNN model using feature extraction of Conv and the unsupervised features of AutoEncoder as AE-CNN block to detect abnormalities in the classification of TB using the whole region of interest (ROI) images. Norval *et al.* [328] investigated TB detection from CXR images using a hybrid approach. The proposed method combined the original statistical computer-aided detection and CNN, which included image pre-processing and segmentation techniques. This hybrid approach helped to improve the contrast of the images.

In 2020, Rahman *et al.* [329] used transfer learning technique on 9 different CNN models (ResNet18, ResNet50, ResNet101, ChexNet, InceptionV3, Vgg19, DenseNet201, SqueezeNet, and MobileNet) for TB and non-TB normal cases classification from CXR images on a NLM/NIH Shenzhen dataset, Belarus dataset, NIAID TB dataset and RSNA

dataset [314]. As a result, the improvement of classification accuracy notably showed in the image segmentation dataset with the best model achieved a accuracy of 96.47%, a precision of 96.62%, a sensitivity of 96.47%, a F1-score of 96.47%, and a specificity of 96.51%.

In 2021, Rahman *et al.* [330] investigated the method of detecting TB from CXR images on NLM/NIH Shenzhen dataset. Their model used three pre-trained neural networks, ResNet101, VGG19, and DenseNet201, along with extreme gradient boosting. Their results showed that the DenseNet201 with extreme gradient boosting had an accuracy of $99.92 \pm 0.14\%$.

Additionally, ensemble learning is another method to retrieve a better predictive result by combining the prediction from different classification models into a new robust classifier model [331]. Ayaz *et al.* [332], proposed a novel TB detection technique that combines hand-crafted features with CNN models through Ensemble Learning. Similarly, Lakhani *et al.* [326] proposed the detection of TB by using two CNN models to classify pulmonary TB and normal CXR images with an ensemble technique added to the classification model to improve the efficiency of the classifier. They achieved the area under the receiver operating characteristic (AUROC) of 0.99 and 0.97 on the Shenzhen and Montgomery datasets, respectively which are similar to the results obtained in the experiments. A modality-specific ensemble DL model proposed by Rajaraman and Antani [333] has enhanced the generalisation performance using pre-trained customised CNN model and modality-specific features.

7.3 Methods

This section presents the pre-processing steps applied to raw CXR images followed by the proposed NN models. Two single DL models and a novel ensemble method are presented for TB diagnosis. The first model is based on a convolutional autoencoder architecture. The second model is a multi-scale residual neural network with deep layer aggregation. An ensemble DL model based on the fusion of the first two models is also proposed (see Section 7.3.4). Finally, the metrics employed to evaluate the performance of the proposed models are described in detail.

7.3.1 Data pre-processing

7.3.1.1 Data pre-processing for MRI

The complete pre-processing processes for ADNI-200 and OASIS-200 datasets have previously been described in Chapter 5.3.1. In this chapter, a new dataset from ADNI, named ADNI-516, was generated on top of the previously utilized dataset to examine the model's MCI classification performance on a larger cohort.

All images in ADNI-516 were affinely aligned to the MNI152 template space and intensity-normalised [334]. AD/MCI patients have been randomly chosen, along with the HC, from the ADNI 2 dataset (available at `http://adni.loni.usc.edu/)` [6]).

7.3.1.2 Data pre-processing for chest X-ray

The primary goal of data pre-processing is to improve image quality so that objects of interest, such as nodules and fibrotic scars, are more visible. Such an image quality enhancement has a significant impact on the performance of the subsequent processing steps [335, 336].

Before starting the pre-processing stage, all the digital imaging and communications in medicine (DICOM) based scans was converted into PNG files. Histogram processing was then applied as an important part of pre-processing step. Histogram equalisation is a typical approach for improving contrast in the anatomic region of interest in an input image. By spreading out the most common intensity levels, it improves contrast in low-contrast areas [337].

Let n_i be the number of occurrences of grey level i. The probability of an occurrence of a pixel of level i in the image is:

$$p_x(i) = p(x = i) = \frac{n_i}{n}, \quad 0 \leq i < L \tag{7.1}$$

where L is the number of possible intensity values, often 256 whereas n is the total number of pixels in the image, and $p_x(i)$ is the image's histogram for pixel value i, normalised to $[0, 1]$. Let p represent a normalised histogram of f with a bin for each intensity level. The

histogram equalised image g will be given as:

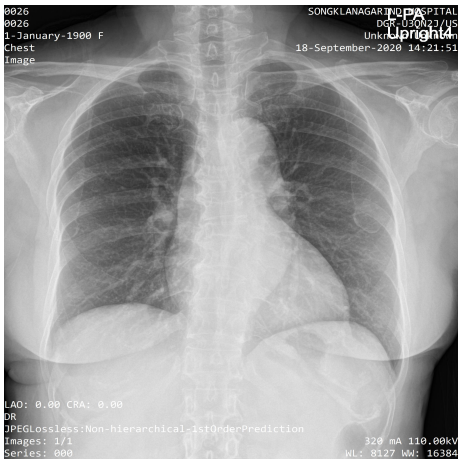
$$g_{i,j} = \lfloor \{(L-1) \sum_{n=0}^{f_{i,j}} p_n \rfloor \tag{7.2}$$

where $\lfloor \cdot \rfloor$ returns the closest integer. The images were scaled to a $[0, 1]$ pixel value range and histogram equalisation has been performed.

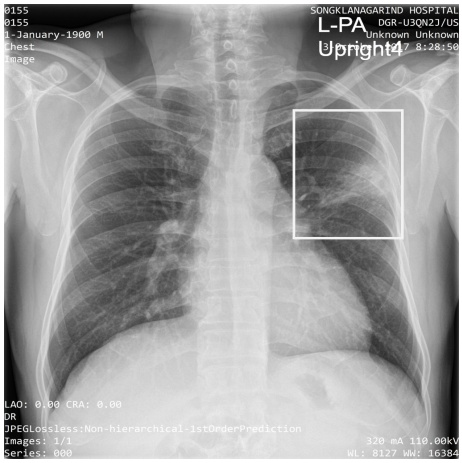
Additionally, automated central cropping was performed to pick lung fields and eliminate the presence of embedded markings on the CXR. Lung field cropping provides two advantages. While it lowers the amount of information lost due to down scaling, it also normalises geometric images. The images were finally downsampled to 512×512 pixels due to GPU memory constraints. For both NLM collections, first black margins have been cleaned as the images from these collections include large black margins around the borders, then histogram equalisation has been performed. After this stage, central cropping has been employed again to pick the lung field and to get rid of the embedded markings on the CXR. Figure 7.1 shows the original CXR images and the same CXR images after the pre-processing step.

7.3.2 Convolutional Autoencoder based DL (CAE-NN)

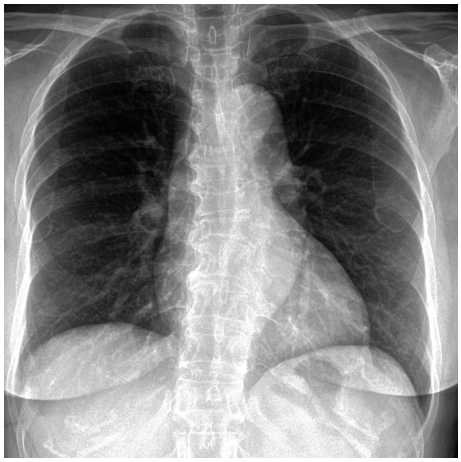
An autoencoder (AE) is a form of neural network that does not require data to be labeled, making it an unsupervised learning technique. In a nutshell, an autoencoder is trained for



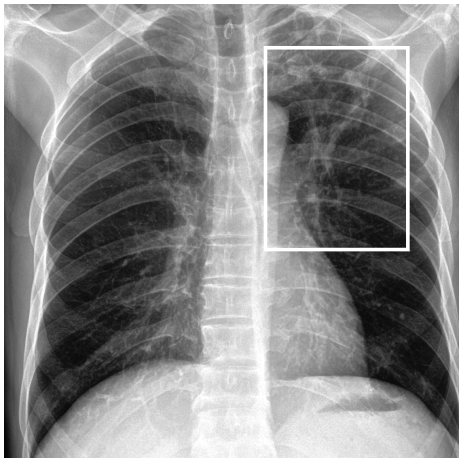
(a) Original CXR with normal case



(b) Original CXR with active pulmonary tuberculosis. The white rectangle denotes reticulonodular infiltration.



(c) Pre-processed CXR with normal findings



(d) Pre-processed CXR with active pulmonary tuberculosis. The white rectangle denotes reticulonodular infiltration.

Figure 7.1: Example of the original and the pre-processing CXR of a healthy patient and a patient with active pulmonary tuberculosis

replicating the input to the output. There is a hidden layer that provides a 'compressed' code that lies in a space called latent space for representing the input. Briefly, an autoencoder is composed of two major components: an encoder that converts the input into code and a decoder that converts the code into a reconstruction of the input. The dimensionality of AEs' latent spaces is smaller than that of the original input, implying that its code cannot store a complete duplicate of the input data and requiring the model to learn how to represent the same data with fewer dimensions.

7.3.2.0.1 Latent representation extraction The four primary hyperparameters in the convolutional autoencoder (CAE) are the number of convolutional layers, the number of filters, the convolutional kernel size, and the number of strides [270]. Figure 7.2 details the architecture of the proposed convolutional autoencoder. The encoder in the proposedmodel consists of six convolution blocks, each with a convolution layer (with a kernel size of 3×3 and the strides of 2 to half the size of features), an instance normalisation layer, and a parametric rectified linear unit (PReLU). PReLU, is an activation function that generalises the classic rectified unit by adding a slope for negative values [231]. This arrangement allows the embedding of a large CXR image into the latent vector. Inversely, the decoder in the proposed model consists of six transposed convolution blocks, each with a transposed convolution layer (with a kernel size of 3×3 and the strides of 2 to double the size of features), an instance normalisation layer, and a PReLU activation. The decoder reconstructs the CXR image from the latent vector. Each convolution layer is followed by a PReLU activation,

allowing the encoding and decoding functions to be non-linear.

It should be noted that no label information is utilised throughout the training phase, therefore this is a completely unsupervised approach.

Encoder

Layer	Channels	Kernel	Strides	Output
INPUT	1	–	–	512x512
CONV1	32	3x3	2	256x256
CONV2	32	3x3	2	128x128
CONV3	64	3x3	2	64x64
CONV4	64	3x3	2	32x32
CONV5	128	3x3	2	16x16
CONV6	128	3x3	2	8x8

Decoder

Layer	Channels	Kernel	Strides	Input
CONVT1	128	3x3	2	8x8
CONVT2	128	3x3	2	16x16
CONVT3	64	3x3	2	32x32
CONVT4	64	3x3	2	64x64
CONVT5	32	3x3	2	128x128
CONVT6	32	3x3	2	256x256
OUTPUT	1	–	–	512x512

Each convolution (CONV) or transposed convolution (CONVT) layer is followed by an instance normalisation layer and a parameterised rectified linear unit (PReLU) layer.

Figure 7.2: Detailed architecture of the proposed convolutional autoencoder including the number of channels and kernel size in each layer.

7.3.2.0.2 Classification The classifier is based on the use of latent features extracted using the deep convolutional autoencoder network. After the encoder's last convolutional layer, the two-dimensional matrix of features was fed to two 3×3 convolutional layers with 256 and 128 output nodes, respectively. Each of the convolutional layers was followed by an instance normalisation and a PReLU activation. The resulting features were then flattened into one dimensional vector then fed to a dense layer with one output.

The sigmoid activation function was used in the output layer. The training process begins by freezing the weight layers of the pre-trained autoencoder and training just the classifier layers. Then, in the second stage, all of the layers are fine-tuned. The overview of the

proposed method is shown in Figure 7.3.

Convolutional layers enable CAEs to filter out noise and produce strong and stable feature representations while simultaneously lowering the input dimension size. It qualifies them for dealing with high-dimensional noisy images.

One advantage of CAEs over standard dense autoencoders for image processing is that there is usually a significant loss of information while stacking and slicing the data. Instead of stacking the data like in traditional autoencoders, the convolutional layers of CAEs may efficiently maintain the spatial information of the input image data and extract information.

7.3.3 Multi-scale convolutional neural network (MS-CNN)

A multi-scale residual neural network [338] with deep layer aggregation [339] is proposed. A residual network is typically composed of residual blocks stacked sequentially. Each residual block consists of a stack of convolutional layers with non-linearity and a shortcut connection. In this work, the original ResNet network [340] is extended with multi-scale backbone modules and hierarchical layer connections.

ResNet residual block with grouped convolution [340] and multi-scale feature representation [338] was employed in this chapter (see Fig. 7.3). This work follows the implementation of Gao *et al.* [338]. The proposed residual block is made of (1) a convolutional layer with a kernel size of 1, (2) a hierarchical multi-scale module of three convolutional layers with a kernel size of 3 and a cardinality of 8, and (3) a convolutional

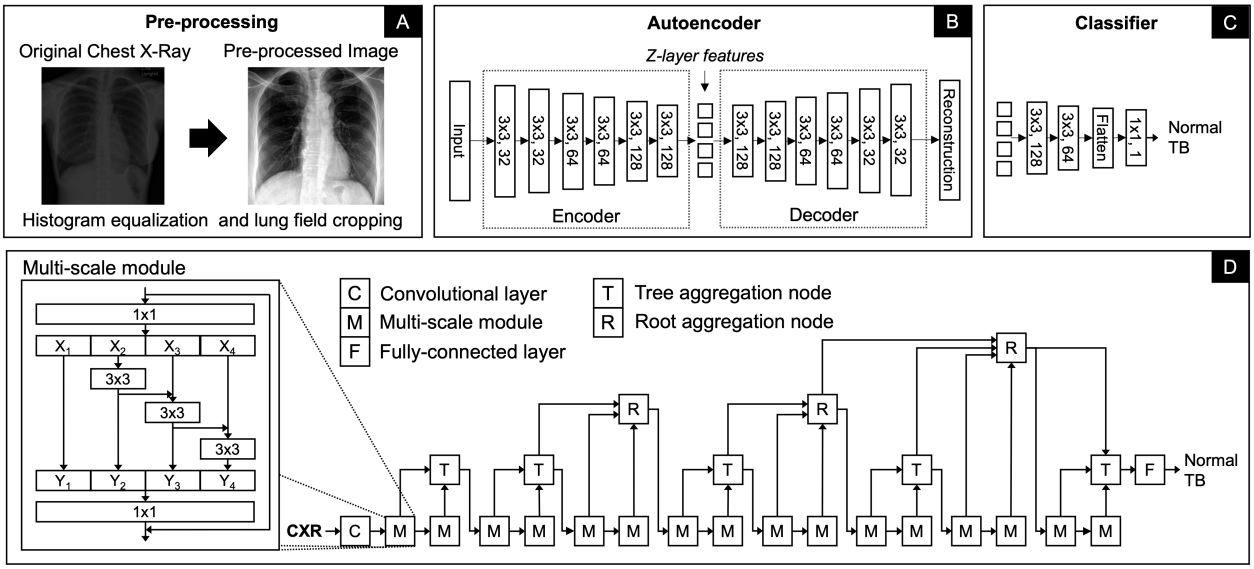


Figure 7.3: Overview of the architectures used for feature extraction and classification for automated pulmonary tuberculosis detection. (A) Pre-processing: chest X-rays have been improved using histogram equalisation and lung field cropping; (B) autoencoder: autoencoder has been trained and features selected from the bottleneck layer; (C) classifier: classification is performed; (D) multiscale convolutional neural network: end-to-end classification has been performed - also features selected from the last convolutional layer for ensemble learning.

layer with a kernel size of 1. In the multi-scale module, the input is split into 4 groups with equal size. Each subset is processed through each 3×3 convolutional layer, except the first group. The output from the first convolutional layer is also added to the input of the second convolutional layer. Similarly, the output from the second convolutional layer is added to the input of the third convolutional layer. Each pass to the convolutional layer enlarges a receptive field. All resulting feature maps are then concatenated together. Batch normalisation was applied after each convolutional layer in the multi-scale module. This

creates the combinatorial explosion effect such that the output of the residual block has feature maps with different receptive field sizes [338].

Multiple multi-scale residual blocks are combined in a hierarchical tree structure through the deep layer aggregation scheme similar to the work by Yu *et al.* [339] (see Fig 7.3). The hierarchical aggregation has iterative connections joining neighbouring residual blocks into a tree, and hierarchical connections joining multiple trees, helping the propagation of feature maps and gradients across the network. With multi-scale residual blocks, the hierarchical tree structure can promote even stronger multi-scale feature representation to the network.

The proposed model is composed of a convolutional layer with a kernel size of 7 with a stride of 2 followed by sixteen multi-scale residual blocks arranged in the deep layer aggregation structure. Five tree aggregation nodes and three root aggregation nodes were used, for a total of eight iterative aggregation nodes. Each aggregation node was composed of a convolutional layer with a kernel size of 1, batch normalisation, and a PReLU activation. Prior to each tree aggregation node, a convolutional layer with a kernel size of 7 with a stride of 2 was applied to half the size of feature maps. The multi-scale network has one output and a sigmoid activation function.

7.3.4 Ensemble learning

Ensemble learning is a type of learning strategy in which many 'base' models are combined to execute tasks such as supervised and unsupervised learning rather than a single model.

Increasing the variety among the base classifiers is one of the key reasons for the success of ensemble techniques, as noted in [341]. To create a diverse classifier, two different DL models have been used. The two-dimensional feature maps of the last convolutional layer of the encoder in the CAE-NN model and the feature maps of the last convolutional layer of the MS-CNN model were concatenated . Both sets of the feature maps have the same size of feature maps at a downsampling factor of 2^6 of the input size. The combined feature maps were then fed to the same classifier as in the CAE-NN model for the classification of normal and TB images.

7.4 Evaluation framework

A description of the datasets and metrics used to evaluate the proposed systems is given in this section.

7.4.1 Datasets

7.4.1.1 AD/MCI datasets

7.4.1.1.1 ADNI-516 On top of the previously used datasets for AD+MCI/HC (see the details regarding to OASIS-200 and ADNI-200 datasets in 4.4.1.1 and in 4.4.1.2), another dataset was created from ADNI to perform three additional binary classifications: AD vs. HC, MCI vs. AD and MCI vs. HC. The dataset consists of 172 AD, 172 HC and

172 MCI patients. All images were affinely aligned to the MNI152 template space and intensity-normalised [334]. AD+MCI patients have been randomly chosen from the ADNI 2 dataset (available at `http://adni.loni.usc.edu/)` – a cohort of ADNI that extends the work of ADNI 1 and ADNI-GO studies [6]). MPRAGE T_1-weighted MRI scans acquired by 3 T scanners [6 Siemens (Erlangen, Germany) MRI scanners and 6 Philips (Amsterdam, Netherlands) scanners] in a sagittal plane (voxel size = 1 mm $\times$ 1 mm $\times$ 1.2 mm) have been utilized. The image size of the T_1-weighted data acquired from the Siemens and Philips scanners were 176 $\times$ 240 $\times$ 256 and 170 $\times$ 256 $\times$ 256, respectively. Since ADNI 2 is a longitudinal dataset, more than one scan was available for each subject. The first scan of each participant has been chosen to produce a cross-sectional dataset. The MRI acquisition protocol for each MRI scanner can be found at `http://adni.loni.usc.edu/methods/documents/mri-protocols/`. In ADNI 2 dataset, subjects have been categorised as AD patients or HCs based on whether subjects have complaints about their memory and by considering a combination of neuropsychological clinical scores [6].

7.4.1.2 TB datasets

To evaluate the efficiency of the proposed methodology, two public and international datasets provided by the National Library of Medicine - National Institutes of Health (NLM/NIH)[2] of the United States of America (USA),namely the Montgomery County

[2] `https://www.nlm.nih.gov/`

dataset and Shenzhen No.3 People's Hospital dataset [314] were used. In addition, a private dataset provided by the Prince of Songkla University, which was built based on the cases of the Songklanagarind Hospital was employed. Table 7.1 presents the demographic and diagnosis data of the participants of each dataset.

Dataset	**Number of cases**		**Gender (%)**		**Age**
	TB	**HC**	**Male**	**Female**	**(years)**
Montgomery County, United States	58	80	46	54	40.1 ± 18.7
Shenzhen No.3 People's Hospital, China	336	326	69	31	35.6 ± 14.7
Songklanagarind Hospital, Thailand	268	274	50	50	51.2 ± 18.1

Table 7.1: Patient demographics and diagnosis information of the chapter population.

7.4.1.2.1 NLM Collection – Montgomery County X-ray Dataset (MC) The CXR images in this dataset were obtained from the TB control programme of Montgomery County's Department of Health and Human Services in Montgomery County, Maryland, USA [314]. This collection comprises 138 posterior-anterior X-rays, 80 of which are normal and 58 of which are abnormal with TB symptoms. The CXRs were taken with a Eureka stationary X-ray machine (CR) and are supplied as 12-bit grey level images in Portable Network Graphics (PNG) format.

7.4.1.2.2 NLM Collection – Shenzhen Hospital X-ray Dataset (SZ) This set of CXR was collected at Shenzhen No.3 Hospital in Shenzhen, Guangdong Province,

China [314]. It consists of 326 normal CXRs and 336 abnormal CXRs exhibiting different TB symptoms. The CXRs are publicly available in PNG format [314][3].

7.4.1.2.3 Songklanagarind Hospital Dataset (SK) The Songklanagarind Hospital dataset was collected by the Department of Radiology, Faculty of Medicine, Prince of Songkla University, in Thailand. The CXRs were collected from patients who had CXR with corresponding chest CT taken between November 2015 and December 2020. Patients with underlying lung diseases or with HIV positive serology were excluded.

CXRs without the posteroanterior and anteroposterior positions, CXRs with more than one image taken in the same position, and CXRs with poor image quality were also excluded. Each CXR (normal and with active pulmonary TB) was read by experienced thoracic radiologists. The active pulmonary TB cases were confirmed by sputum culture for Mtb. Then, the CXRs were exported in an uncompressed Digital Imaging and Communications in Medicine (DICOM) format from the institutional Picture Archiving and Communication System (PACS). The images were then converted to Portable Network Graphics (PNG) format. As a result, the dataset consisted of 542 CXRs, of which 274 images were CXRs with normal findings (sample of image see Figure 7.1a) and 268 images were CXRs with active pulmonary TB (sample of image see Figure 7.1b).

[3] `https://lhncbc.nlm.nih.gov/LHC-publications/pubs/TuberculosisChestXrayImageDataSets.html`

7.4.2 Model training and validation

AEs are made of an encoder and a decoder that may be trained concurrently to minimise a loss function between an input and the reconstruction of the input. The MSE and the binary cross-entropy are two often used loss functions for training autoencoders (BCE). To measure how well the AE can recreate unseen images the MSE is applied to the test images. MSE is used as reconstruction error between the input image x and the reconstructed image at the output $\hat{x}_i = g(f(x_i))$:

$$\mathcal{L} = \frac{1}{N}\sum_i (x_i - g(f(x_i)))^2 \tag{7.3}$$

Furthermore, the PSNR was used as a quality measurement between the original and reconstructed images in the scope of this investigation. When comparing reconstructed outcomes, a metric of picture quality is necessary. PSNR is calculated as:

$$\text{PSNR} = 10 \log_{10} \frac{{\text{Max}_\text{I}}^2}{\text{MSE}} \tag{7.4}$$

where Max_I is the maximum pixel value.

The proposed models predict whether a CXR belongs to the normal class or the pulmonary TB class. In order to evaluate the classification performance of the proposed methods with the three datasets (see Section 7.4.1) four measures are given: accuracy, sensitivity, specificity, and F1-score.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{7.5}$$

Sensitivity measures the ability of the model to correctly classify a chest radiograph as pulmonary tuberculosis:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{7.6}$$

Specificity measures the ability of the model to correctly classify a chest radiograph as normal findings:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{7.7}$$

Sensitivity and specificity are inversely proportional, i.e. as sensitivity increases, specificity decreases and the other way around. F_1-Score measuring the ability of the model to correctly classify a chest radiograph with pulmonary tuberculosis by taking into account all misclassified samples is defined as:

$$\text{F}_1\text{-Score} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})} \tag{7.8}$$

The 5-fold cross-validation approach is employed to build a robust model. Cross-validation is a useful method for dealing with the overfitting problem. In this approach, the dataset is divided into five mutually exclusive subsets, with one of them serving as the test set each time and the model being trained five times. It utilises all of

the data for training/testing and then takes the average of the cross-validation results, making the assessment findings more stable.

For data augmentation, a random affine transformation is employed with a rotation range between $-\frac{\pi}{8}$ and $-\frac{\pi}{8}$, a scale range between 0.80 and 1.20, and a translation range between -64 and 64 pixels as well as random elastic transformation with a grid spacing of 64 pixels and a magnitude between 0 and 2. Data augmentation is employed on the fly at the training time. To eliminate data leakage, the data is separated initially, and then augmentation is conducted exclusively on the training set for each cycle [11].

All networks were implemented using the Project MONAI framework [342] version 0.5 on the Nvidia GeForce RTX 2080. A cross entropy loss is used and the network is trained using the NovoGra method [343].

7.5 Experimental Results

This section details the results produced for both AD, MCI classification and TB classification using the techniques and evaluation framework presented in Sections 7.3 and 7.4, respectively.

7.5.1 Experiments part I: Ensemble method for AD/MCI diagnosis

The performance of the ensemble method on TB classification has motivated the investigation of the use of a similar approach in MCI/AD/HC diagnosis. For this purpose, first, the ensemble method has been tested on ADNI-200 and OASIS-200 datasets in order to have a comparative analysis between proposed technique and the previous results.

Dataset	Model	Accuracy	F1 - score
ADNI-200	Ensemble method	82 ± 0.02%	0.78
OASIS-200	Ensemble method	80 ± 0.15%	0.83

Table 7.2: Classification performance on the test set. The mean percentage ± standard deviation of accuracy and F1 score are listed.

On top of the previously used datasets for AD+MCI/HC, another dataset was created from ADNI to perform three additional binary classifications: AD vs. HC, MCI vs. AD and MCI vs. HC. The dataset consists of 172 AD, 172 HC and 172 MCI patients. All images were affinely aligned to the MNI152 template space and intensity-normalised [334]. Multiscale CNN and convolutional autoencoder based ensemble classifier achieved 82%, 70%, and 65% (on AD vs. HC, MCI vs. AD, and MCI vs. HC, respectively).

7.5.2 Experiments part II: Ensemble method for TB diagnosis

As the PSNR increases, the image quality of the recovered image improves. In the trials, an average PSNR of 30.86 dB, 31.71 dB, and 35.51 dB was obtained for rebuilt validation

Metrics	Dataset								
	Songklanagarind Hospital (SK)			Montgomery County (MC)			Shenzhen No.3 People's Hospital (SZ)		
	MS-CNN	CAE-NN	Ensemble	MS-CNN	CAE-NN	Ensemble	MS-CNN	CAE-NN	Ensemble
AUROC	0.95 ± 0.02	0.93 ± 0.03	**0.96 ± 0.03**	0.76 ± 0.10	0.65 ± 0.14	**0.77 ± 0.09**	0.94 ± 0.02	0.90 ± 0.03	**0.98 ± 0.01**
Accuracy	**0.92 ± 0.03**	0.88 ± 0.04	0.92 ± 0.05	0.75 ± 0.03	0.71 ± 0.10	**0.77 ± 0.09**	0.89 ± 0.03	0.85 ± 0.03	**0.95 ± 0.04**
Sensitivity	0.86 ± 0.03	**0.91 ± 0.0**8	0.89 ± 0.07	**0.80 ± 0.15**	0.69 ± 0.17	0.73 ± 0.23	0.92 ± 0.05	0.87 ± 0.04	**0.95 ± 0.04**
Specificity	**0.96 ± 0.03**	0.85 ± 0.10	0.95 ± 0.05	0.69 ± 0.25	0.74 ± 0.16	**0.83 ± 0.14**	0.86 ± 0.02	0.83 ± 0.07	**0.95 ± 0.06**
PPV	**0.96 ± 0.03**	0.87 ± 0.07	0.95 ± 0.05	0.81 ± 0.11	0.79 ± 0.08	**0.88 ± 0.09**	0.86 ± 0.02	0.84 ± 0.06	**0.95 ± 0.06**
NPV	0.88 ± 0.03	**0.91 ± 0.06**	0.90 ± 0.06	**0.76 ± 0.14**	0.64 ± 0.12	0.73 ± 0.17	0.92 ± 0.05	0.87 ± 0.03	**0.95 ± 0.04**

Table 7.3: Classification performance of the proposed models on all datasets without data augmentation. MS-CNN stands for Multi-scale CNN whereas CAE-NN is short for Convolutional Autoencoder based classifier. The performance metrics that are used in the experiments are Area under the receiver operating characteristics (AUROC), accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

images for the Montgomery County dataset, the Shenzhen Hospital dataset, and the Songklanagarind Hospital dataset, respectively.

Extensive experiments have been conducted to verify the effectiveness of the proposed models to diagnose pulmonary TB using CXR (see Section 7.2). In this work, three datasets are used as presented in Section 7.4.1. The CXR images from the three datasets were all pre-processed (see Section 7.3.1) and used for DL model training and validation.

Tables 7.3 and 7.4 show the classification performance of the proposed models for all datasets without and with data augmentation, respectively. Figure 7.5 shows the comparison of AUROC rates among three models (without and with data augmentation) with all datasets (SK, MC, and SZ). Data augmentation resulted in a slight increase in performance on all datasets. MS-CNN achieved higher AUROCs than CAE-NN for all datasets. The ensemble of both methods resulted in a further increase in performance. CAE-NN had larger standard

Metrics	Dataset								
	Songklanagarind Hospital (SK)			Montgomery County (MC)			Shenzhen No.3 People's Hospital (SZ)		
	MS-CNN	CAE-NN	Ensemble	MS-CNN	CAE-NN	Ensemble	MS-CNN	CAE-NN	Ensemble
AUROC	**0.97 ± 0.02**	0.94 ± 0.03	**0.97 ± 0.02**	0.74 ± 0.13	0.70 ± 0.13	**0.77 ± 0.06**	0.98 ± 0.01	0.95 ± 0.02	**0.99 ± 0.01**
Accuracy	**0.93 ± 0.03**	0.88 ± 0.04	**0.93 ± 0.02**	**0.77 ± 0.08**	0.72 ± 0.10	**0.77 ± 0.04**	0.94 ± 0.02	0.91 ± 0.02	**0.96 ± 0.02**
Sensitivity	0.91 ± 0.08	0.89 ± 0.06	**0.94 ± 0.06**	**0.86 ± 0.11**	0.80 ± 0.15	0.81 ± 0.12	0.95 ± 0.03	0.94 ± 0.03	**0.99 ± 0.01**
Specificity	**0.96 ± 0.03**	0.88 ± 0.08	0.92 ± 0.02	0.63 ± 0.15	0.62 ± 0.17	**0.73 ± 0.12**	**0.94 ± 0.05**	0.88 ± 0.07	**0.94 ± 0.04**
PPV	**0.96 ± 0.03**	0.89 ± 0.06	0.93 ± 0.02	0.77 ± 0.07	0.75 ± 0.08	**0.80 ± 0.06**	**0.94 ± 0.05**	0.89 ± 0.06	**0.94 ± 0.03**
NPV	0.92 ± 0.07	0.89 ± 0.05	**0.94 ± 0.06**	**0.79 ± 0.16**	0.72 ± 0.13	0.78 ± 0.09	0.95 ± 0.03	0.94 ± 0.03	**0.99 ± 0.01**

Table 7.4: Classification performance of the proposed models on all datasets with data augmentation. MS-CNN stands for Multi-scale Convolutional Neural Network whereas CAE-NN is short for Convolutional Autoencoder based classifier. The performance metrics that are used in the experiments are : Area under the receiver operating characteristics (AUROC), accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

deviations than MS-CNN. The ensemble model had an AUROC of 0.97, 0.77, and 0.99 for the SK, MC, and SZ datasets, respectively. The performance of the models for the MC dataset was lower compared to the other two datasets in nearly all metrics.

7.6 Discussion

As it can be seen from the results presented in Section 7.5.1, the ensemble method outperforms all the previously used frameworks in Chapters 4, 5, and 6 for AD+MCI/HC when tested on ADNI-200 and OASIS-200. For the binary classification of AD vs. HC, MCI vs. AD and MCI vs. HC, multiscale CNN and convolutional autoencoder based ensemble classifier achieved state-of-the-art results.

When it comes to TB classification, the performance of the suggested ensemble

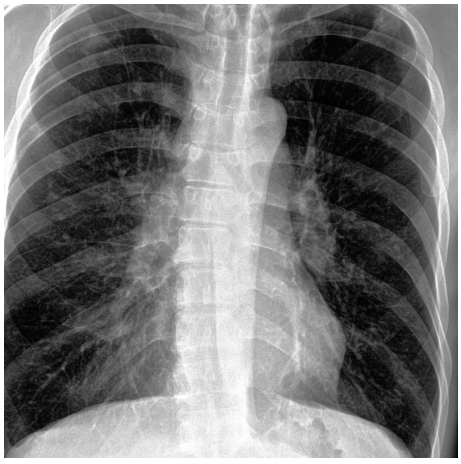
technique is compared with the two alternative DL models in Tables 7.3 and 7.4. Three separate datasets were used to validate the performance of the proposed models, containing participants from three different nations (Thailand, USA, and China) and ethnics, in various formats and of varying quality (DICOM and PNG). The models are trained and tested using the same datasets, one of which is from Thailand, a high TB burden area. Tables 7.3 and 7.4 show that, in this two-class scenario, all the networks perform very well in identifying TB and normal images across the three datasets. Therefore, the model can be applied to different ethnic groups.

The SK dataset differs slightly from the MC and SZ datasets, as the SK dataset contains only active TB cases, while the MC and SZ datasets contain both active and inactive TB. It is not surprising that the MC dataset achieved lower performance compared to the other datasets, as it contains a smaller number of cases. Since CXRs vary widely in terms of patient anatomy and abnormalities, the performance of the model could increase if more cases are included in the dataset.

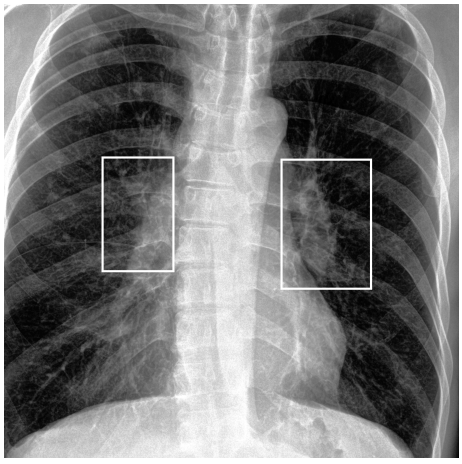
The use of the institutional dataset (SK Dataset) allows us to extend the analysis to full-text radiologist reports. On the SK dataset with the single DL model (either MS-CNN or CAE-NN), most false positive predictions were CXRs with some abnormalities but without active TB, e.g., prominent heart size, dilated aorta, suboptimal inspiration, degenerative spine, and slightly oblique position. In addition, most false-negative predictions were CXRs with minimal abnormalities, indeterminate TB activity, or abnormalities in the upper long

zones. Figure 7.4 showed a sample of false-negative result cases from a single DL model. Figure 7.4a is the original CXR image in which both MS-CNN and CAE-NN models could not identify this image as the active TB. Figure 7.4b is the same CXR image of Figure 7.4a that has been further investigated by checking with the patient's sputum smear test result and an expert radiologist. The sputum smear test showed that the patient has an active TB and the expert radiologist identified the active TB position on the CXR image. As a consequence of this examination, an interesting issue was discovered: the pattern of a lesion in the picture was not typical, as only necrotic mediastinal nodes were seen, with no infiltration. In addition, the lesion was covered by part of the rib bone as illustrated in Figure 7.4b. Nevertheless, when using the ensemble model, the image was correctly classified as a positive result in this case.

On the MC dataset, as observed in its detailed findings, most false-negative predictions were CXRs with abnormalities in the upper long zones and inactive TB scars. On the SZ dataset, half of the false-negative predictions were CXRs with abnormalities in the upper long zones, and one-third of the false-negative predictions were CXRs with the reactivation of old TB lesions. Most of the false positives and false negatives were from cases that were not prominently represented in the datasets. Further investigation on bone suppression algorithms may help to improve the classification for the detection of TB in the upper lung regions or oblique positions. It is worth noting that the SK dataset contains only CXRs with active TB, while the MC and SZ datasets contain CXRs with both active TB and scars from



(a) The miss-classified CXR image (false negative)



(b) The necrotic mediastinal nodes identified by an expert radiologist denoting with the white rectangles

Figure 7.4: Example of miss-classified result as the false negative CXR image from the single DL model (a patient with active pulmonary tuberculosis (positive) but classified as negative).

previous TB infection.

The ensemble model improves the accuracy of the CAE-NN and MS-CNN models. The false-negative rates of the ensemble model with data augmentation were 6%, 19%, and 1% for the SK, MC, and SZ datasets, respectively. The reported false-negative rates from the interpretation of expert radiologists were from 7% to 9% [344, 345]. However, the interpretation of the general physician and other specialists had a false negative rate ranging from 15% to 30% [346].

The ensemble DL model has been shown to be a successful solution since it is applicable to varied modalities and disease structures. Therefore, the model might be a helpful triage tool,

especially for the annual health checkup and preoperative screening in developing countries where the availability of expert radiologists is limited. These tools stand a chance to be a critical part of the diagnostic process when the numbers of radiologists are limited and low-cost solutions are required.

7.7 Conclusions

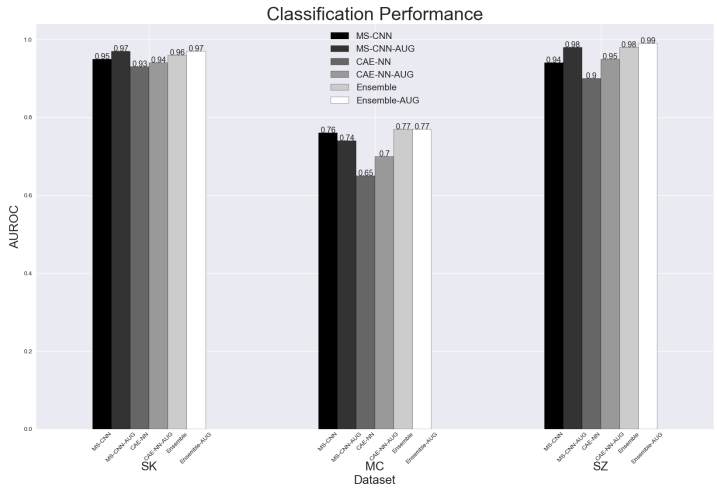


Figure 7.5: Bar chart showing the performance of different classification models on three datasets

This work presents a novel ensemble method together with two individual DL-based classification approaches for automatic diagnosis. The performance of the DL models was evaluated on two different diseases using overall five publicly available datasets as well as one private dataset.

The experimental findings clearly show that the proposed ensemble learning method outperforms other tested models for AD and MCI diagnosis. The proposed ensemble method achieves 82%, 70%, and 65% AUROC rates for AD vs. HC, MCI vs. AD, and MCI vs. HC, respectively.

For pulmonary TB prediction and classification, the two suggested DL-driven techniques yield precise and effective results. With pre-processing, the AUROC rates of MS-CNN for TB detection were 97%, 74%, and 98% on Songklanagarind, Montgomery, and Shenzhen datasets, respectively. Convolutional autoencoder-based NN classifier achieved 94%, 70%, and 95% (on Songklanagarind, Montgomery, and Shenzhen datasets, respectively). The ensemble learning technique, on the other hand, outperforms other evaluated models and achieves the state-of-the-art for automated pulmonary TB detection using CXRs. On the Songklanagarind, Montgomery, and Shenzhen datasets, the suggested ensemble technique obtains 97%, 77%, and 99% AUROC rates, respectively.

The models' performance suggests that they have the potential to be a very useful and rapid diagnostic tool in the future, perhaps saving a substantial number of people who die each year as a consequence of delayed or inadequate diagnosis.

Chapter 8

Conclusions

Deep learning in medical image processing is rapidly advancing, with the potential to revolutionise everyday clinical practise. Early illness detection and tracking lead to better patient care and monitoring of disease-modifying therapy, therefore it's crucial. This thesis proposes various DL-based methods for neurodegenerative disease diagnosis and demonstrates their usage on a variety of public and private datasets.

In Chapter 3, the diagnosis of two common neurological diseases from structural MR images is performed using a transfer learning-based technique. Two cutting-edge architectures, namely VGG16 and Resnet, are used to distinguish PD and AD+MCI patients from HC.

In Chapter 4, the use of erroneous slice-level CV, which is sadly widespread in neuroimaging literature has been revealed and the extent of overestimation in classification

performances is assessed. It is concluded that slice-based CV leads to highly optimistic model performances, especially for small datasets. It is commonly known that data leakage results in increased model performance. Nonetheless, the degree of overestimation determined by the experimentation was unexpected.

In Chapter 5, a deep 3D CNN model has been proposed for the diagnosis of AD+MCI and PD patients using structural brain MRI. The model's performance was validated on two key AD datasets, ADNI and OASIS, as well as one PD dataset called PPMI. 3D models have been utilised to prevent information loss and to learn more abstract level spatial representation.

In Chapter 6, a CAE-based DL method has been presented for classifying AD+MCI vs. HC subjects using single 2D brain MRI slices.

Finally in Chapter 7, a novel end-to-end semi-supervised ensemble DL architecture for automated diagnosis is introduced, along with two DL-driven approaches. The model detects patients by combining supervised prediction and unsupervised representation. These components are interconnected and trainable from the beginning to the end, providing a direct link between raw data and intended clinical output. The proposed ensemble method has been tested on various diseases including a non-neurodegenerative one and successfully reached state-of-the-art.

Key contributions of this dissertation are i) the development and release of several 2D and 3D CNN based frameworks for mainly AD and PD diagnosis using T_1-weighted brain MRI data, ii) the conduction of an exhaustive literature survey and the review of the flaws

in selected studies, iii) the quantitative assessment of data leakage caused by the adoption of incorrect slice-level CV, rather than subject-level CV, using three 2D CNN models for the classification of patients with AD and PD and iv) the implementation of a novel ensemble method which has a potential to be used clinically not only on neurodegenerative diseases but also for other diseases such as pulmonary tuberculosis.

The lack of openness and generalisability in the literature remains a significant obstacle to the clinical translation of DL approaches. Another explanation for the low translation rate of DL to clinical practice is that many studies, including this one, use only binary classification to discriminate between two categories of diseases. The majority of the published research demonstrated competitive performance in distinguishing AD patients from HCs. However, because the patient is already demented, the therapeutic use of this model may be restricted. More exciting challenges, such as the finding of early biomarkers, remain unanswered. The requirement of sizeable training data in DL models is one of the main challenges in deep learning-based medical image analysis. However, medical image datasets are typically small since patient privacy prohibits building large-scale datasets, and ground-truth requires experienced radiologists to laboriously annotate results. When large volumes of medical imaging data become available, poor performance tasks like sMCI vs pMCI may be improved. Furthermore, medical images, such as MRIs, are often rather large. Spatial 3D data and 3D models also require a great deal of GPU memory. Deep learning models have high capacity and complexity, which may result in poor generalisation

ability on outlier data. Some of these challenges have been addressed in this thesis by using semi-supervised structures, data augmentation, and transfer learning. Furthermore, by releasing open-source frameworks, it is envisaged that a baseline performance would be provided and future researchers would be able to increase transparency and generalisability.

The models' performance shows that once confirmed and validated through larger investigations, they have the potential to be a highly helpful and quick diagnostic tool in the future, potentially saving significant number of individuals who die each year as a result of the delayed or insufficient diagnosis.

Bibliography

[1] Y. Huang and L. Mucke, "Alzheimer's mechanisms and therapeutic strategies," *Cell*, vol. 148, no. 6, pp. 1204–1222, 2012.

[2] R. J. Bateman, C. Xiong, T. L. Benzinger, A. M. Fagan, A. Goate, N. C. Fox, D. S. Marcus, N. J. Cairns, X. Xie, T. M. Blazey, *et al.*, "Clinical and biomarker changes in dominantly inherited Alzheimer's disease," *N Engl J Med*, vol. 367, pp. 795–804, 2012.

[3] J. Dukart, "Basic concepts of image classification algorithms applied to study neurodegenerative diseases," in *Brain mapping: An encyclopedic reference*, pp. 641–646, Elsevier, 2015.

[4] R. C. Gonzalez and R. E. Woods, "Image processing," *Digital image processing*, vol. 2, p. 1, 2007.

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

[6] R. C. Petersen, P. Aisen, L. A. Beckett, M. Donohue, A. Gamst, D. J. Harvey, C. Jack, W. Jagust, L. Shaw, A. Toga, *et al.*, "Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization," *Neurology*, vol. 74, no. 3, pp. 201–209, 2010.

[7] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *Journal of cognitive neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[8] S. Lehericy, M. Baulac, J. Chiras, L. Pierot, N. Martin, B. Pillon, B. Deweer, B. Dubois, and C. Marsault, "Amygdalohippocampal mr volume measurements in the early stages of Alzheimer's disease.," *American Journal of Neuroradiology*, vol. 15, no. 5, pp. 929–937, 1994.

[9] M. Bobinski, M. De Leon, J. Wegiel, S. Desanti, A. Convit, L. Saint Louis, H. Rusinek, and H. Wisniewski, "The histological validation of post mortem magnetic resonance imaging-determined hippocampal volume in Alzheimer's disease," *Neuroscience*, vol. 95, no. 3, pp. 721–725, 1999.

[10] J. Mortimer, K. Gosche, K. Riley, W. Markesbery, and D. Snowdon, "Delayed recall, hippocampal volume and Alzheimer's neuropathology: findings from the nun study," *Neurology*, vol. 62, no. 3, pp. 428–432, 2004.

[11] E. Yagis, A. G. S. De Herrera, and L. Citi, "Generalization performance of deep learning models in neurodegenerative disease classification," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1692–1698, IEEE, 2019.

[12] E. Yagis, S. W. Atnafu, A. García Seco de Herrera, C. Marzi, R. Scheda, M. Giannelli, C. Tessa, L. Citi, and S. Diciotti, "Effect of data leakage in brain mri classification using 2d convolutional neural networks," *Scientific reports*, vol. 11, no. 1, pp. 1–13, 2021.

[13] E. Yagis, L. Citi, S. Diciotti, C. Marzi, S. W. Atnafu, and A. G. S. De Herrera, "3d convolutional neural networks for diagnosis of Alzheimer's disease via structural mri," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 65–70, IEEE, 2020.

[14] S. Przedborski, M. Vila, V. Jackson-Lewis, *et al.*, "Series introduction: Neurodegeneration: What is it and where are we?," *The Journal of clinical investigation*, vol. 111, no. 1, pp. 3–10, 2003.

[15] K. J. Barnham, C. L. Masters, and A. I. Bush, "Neurodegenerative diseases and oxidative stress," *Nature reviews Drug discovery*, vol. 3, no. 3, pp. 205–214, 2004.

[16] M. D. Weingarten, A. H. Lockwood, S.-Y. Hwo, and M. W. Kirschner, "A protein factor essential for microtubule assembly," *Proceedings of the National Academy of Sciences*, vol. 72, no. 5, pp. 1858–1862, 1975.

[17] I. Grundke-Iqbal, K. Iqbal, Y.-C. Tung, M. Quinlan, H. M. Wisniewski, and L. I. Binder, "Abnormal phosphorylation of the microtubule-associated protein tau in alzheimer's cytoskeletal pathology," *Proceedings of the National Academy of Sciences*, vol. 83, no. 13, pp. 4913–4917, 1986.

[18] D. S. Knopman, H. Amieva, R. C. Petersen, G. Chételat, D. M. Holtzman, B. T. Hyman, R. A. Nixon, and D. T. Jones, "Alzheimer's disease," *Nature Reviews Disease Primers*, vol. 7, no. 1, pp. 1–21, 2021.

[19] V. L. Villemagne, S. Burnham, P. Bourgeat, B. Brown, K. A. Ellis, O. Salvado, C. Szoeke, S. L. Macaulay, R. Martins, P. Maruff, *et al.*, "Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study," *The Lancet Neurology*, vol. 12, no. 4, pp. 357–367, 2013.

[20] E. M. Reiman, Y. T. Quiroz, A. S. Fleisher, K. Chen, C. Velez-Pardo, M. Jimenez-Del-Rio, A. M. Fagan, A. R. Shah, S. Alvarez, A. Arbelaez, *et al.*, "Brain imaging and fluid biomarker analysis in young adults at genetic risk for autosomal dominant Alzheimer's disease in the presenilin 1 e280a kindred: a case-control study," *The Lancet Neurology*, vol. 11, no. 12, pp. 1048–1056, 2012.

[21] C. R. Jack Jr, V. J. Lowe, S. D. Weigand, H. J. Wiste, M. L. Senjem, D. S. Knopman, M. M. Shiung, J. L. Gunter, B. F. Boeve, B. J. Kemp, *et al.*, "Serial pib and mri in normal, mild cognitive impairment and Alzheimer's disease: implications for sequence of pathological events in Alzheimer's disease," *Brain*, vol. 132, no. 5, pp. 1355–1365, 2009.

[22] H. Braak, D. R. Thal, E. Ghebremedhin, and K. Del Tredici, "Stages of the pathologic process in Alzheimer's disease: age categories from 1 to 100 years," *Journal of Neuropathology & Experimental Neurology*, vol. 70, no. 11, pp. 960–969, 2011.

[23] G. . D. F. Collaborators *et al.*, "Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019," *The Lancet Public Health*, 2022.

[24] R. C. Petersen, P. Aisen, B. F. Boeve, Y. E. Geda, R. J. Ivnik, D. S. Knopman, M. Mielke, V. S. Pankratz, R. Roberts, W. A. Rocca, *et al.*, "Mild cognitive impairment due to Alzheimer's disease in the community," *Annals of neurology*, vol. 74, no. 2, pp. 199–208, 2013.

[25] E. K. Degenhardt, M. M. Witte, M. G. Case, P. Yu, D. B. Henley, H. M. Hochstetler, D. N. D'Souza, and P. T. Trzepacz, "Florbetapir f18 pet amyloid neuroimaging and characteristics in patients with mild and moderate Alzheimer's dementia," *Psychosomatics*, vol. 57, no. 2, pp. 208–216, 2016.

[26] A. Association *et al.*, "2018 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 14, no. 3, pp. 367–429, 2018.

[27] W. M. van der Flier and P. Scheltens, "Epidemiology and risk factors of dementia," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 76, no. suppl 5, pp. v2–v7, 2005.

[28] J. Poirier, P. Bertrand, S. Kogan, S. Gauthier, J. Davignon, and D. Bouthillier, "Apolipoprotein e polymorphism and Alzheimer's disease," *The Lancet*, vol. 342, no. 8873, pp. 697–699, 1993.

[29] A. Gharbi-Meliani, A. Dugravot, S. Sabia, M. Regy, A. Fayosse, A. Schnitzler, M. Kivimäki, A. Singh-Manoux, and J. Dumurgier, "The association of apoe $\varepsilon 4$ with cognitive function over the adult life course and incidence of dementia: 20 years follow-up of the whitehall ii study," *Alzheimer's research & therapy*, vol. 13, no. 1, pp. 1–11, 2021.

[30] S. J. van der Lee, F. J. Wolters, M. K. Ikram, A. Hofman, M. A. Ikram, N. Amin, and C. M. van Duijn, "The effect of apoe and other common genetic variants on the onset of Alzheimer's disease and dementia: a community-based cohort study," *The Lancet Neurology*, vol. 17, no. 5, pp. 434–444, 2018.

[31] R. J. O'brien and P. C. Wong, "Amyloid precursor protein processing and Alzheimer's disease," *Annual review of neuroscience*, vol. 34, pp. 185–204, 2011.

[32] S. Lammich, E. Kojro, R. Postina, S. Gilbert, R. Pfeiffer, M. Jasionowski, C. Haass, and F. Fahrenholz, "Constitutive and regulated α-secretase cleavage of Alzheimer's amyloid precursor protein by a disintegrin metalloprotease," *Proceedings of the national academy of sciences*, vol. 96, no. 7, pp. 3922–3927, 1999.

[33] C. Duyckaerts, B. Delatour, and M.-C. Potier, "Classification and basic pathology of Alzheimer's disease," *Acta neuropathologica*, vol. 118, no. 1, pp. 5–36, 2009.

[34] V. W. Chow, M. P. Mattson, P. C. Wong, and M. Gleichmann, "An overview of app processing enzymes and products," *Neuromolecular medicine*, vol. 12, no. 1, pp. 1–12, 2010.

[35] M. P. Murphy and H. LeVine III, "Alzheimer's disease and the amyloid-β peptide," *Journal of Alzheimer's disease*, vol. 19, no. 1, pp. 311–323, 2010.

[36] C. Haass and D. J. Selkoe, "Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid β-peptide," *Nature reviews Molecular cell biology*, vol. 8, no. 2, pp. 101–112, 2007.

[37] H. Hampel, J. Hardy, K. Blennow, C. Chen, G. Perry, S. H. Kim, V. L. Villemagne, P. Aisen, M. Vendruscolo, T. Iwatsubo, *et al.*, "The amyloid-β pathway in Alzheimer's disease," *Molecular psychiatry*, pp. 1–23, 2021.

[38] A. E. Roher, M. O. Chaney, Y.-M. Kuo, S. D. Webster, W. B. Stine, L. J. Haverkamp, A. S. Woods, R. J. Cotter, J. M. Tuohy, G. A. Krafft, *et al.*, "Morphology and toxicity of aβ-(1-42) dimer derived from neuritic and vascular amyloid deposits of Alzheimer's disease," *Journal of Biological Chemistry*, vol. 271, no. 34, pp. 20631–20635, 1996.

[39] B. Seilheimer, B. Bohrmann, L. Bondolfi, F. Müller, D. Stüber, and H. Döbeli, "The toxicity of the Alzheimer's β-amyloid peptide correlates with a distinct fiber morphology," *Journal of structural biology*, vol. 119, no. 1, pp. 59–71, 1997.

[40] G.-f. Chen, T.-h. Xu, Y. Yan, Y.-r. Zhou, Y. Jiang, K. Melcher, and H. E. Xu, "Amyloid beta: structure, biology and structure-based therapeutic development," *Acta Pharmacologica Sinica*, vol. 38, no. 9, pp. 1205–1235, 2017.

[41] M. Verma, A. Vats, and V. Taneja, "Toxic species in amyloid disorders: Oligomers or mature fibrils," *Annals of Indian Academy of Neurology*, vol. 18, no. 2, p. 138, 2015.

[42] S. Gauthier, B. Reisberg, M. Zaudig, R. C. Petersen, K. Ritchie, K. Broich, S. Belleville, H. Brodaty, D. Bennett, H. Chertkow, *et al.*, "Mild cognitive impairment," *The lancet*, vol. 367, no. 9518, pp. 1262–1270, 2006.

[43] C. P. Hughes, L. Berg, W. Danziger, L. A. Coben, and R. L. Martin, "A new clinical scale for the staging of dementia," *The British journal of psychiatry*, vol. 140, no. 6, pp. 566–572, 1982.

[44] R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, E. G. Tangalos, and E. Kokmen, "Mild cognitive impairment: clinical characterization and outcome," *Archives of neurology*, vol. 56, no. 3, pp. 303–308, 1999.

[45] D. Shigemizu, S. Akiyama, S. Higaki, T. Sugimoto, T. Sakurai, K. A. Boroevich, A. Sharma, T. Tsunoda, T. Ochiya, S. Niida, *et al.*, "Prognosis prediction model for conversion from mild cognitive impairment to alzheimer's disease created by integrative analysis of multi-omics data," *Alzheimer's research & therapy*, vol. 12, no. 1, pp. 1–12, 2020.

[46] D. Siedlecki-Wullich, J. Català-Solsona, C. Fábregas, I. Hernández, J. Clarimon, A. Lleó, M. Boada, C. A. Saura, J. Rodríguez-Álvarez, and A. J. Miñano-Molina, "Altered micrornas related to synaptic function as potential plasma biomarkers for alzheimer's disease," *Alzheimer's research & therapy*, vol. 11, no. 1, pp. 1–11, 2019.

[47] R. C. Petersen, "Mild cognitive impairment," *CONTINUUM: Lifelong Learning in Neurology*, vol. 22, no. 2 Dementia, p. 404, 2016.

[48] G. A. Jicha, J. E. Parisi, D. W. Dickson, K. Johnson, R. Cha, R. J. Ivnik, E. G. Tangalos, B. F. Boeve, D. S. Knopman, H. Braak, *et al.*, "Neuropathologic outcome of mild cognitive impairment following progression to clinical dementia," *Archives of neurology*, vol. 63, no. 5, pp. 674–681, 2006.

[49] I. Carrière, A. Fourrier-Reglat, J.-F. Dartigues, O. Rouaud, F. Pasquier, K. Ritchie, and M.-L. Ancelin, "Drugs with anticholinergic properties, cognitive decline, and dementia in an elderly general population: the 3-city study," *Archives of internal medicine*, vol. 169, no. 14, pp. 1317–1324, 2009.

[50] J. L. Robinson, S. Porta, F. G. Garrett, P. Zhang, S. X. Xie, E. Suh, V. M. Van Deerlin, E. L. Abner, G. A. Jicha, J. M. Barber, *et al.*, "Limbic-predominant age-related tdp-43 encephalopathy differs from frontotemporal lobar degeneration," *Brain*, vol. 143, no. 9, pp. 2844–2857, 2020.

[51] E. L. Abner, R. J. Kryscio, F. A. Schmitt, D. W. Fardo, D. C. Moga, E. T. Ighodaro, G. A. Jicha, L. Yu, H. H. Dodge, C. Xiong, *et al.*, "Outcomes after diagnosis of mild cognitive impairment in a large autopsy series," *Annals of neurology*, vol. 81, no. 4, pp. 549–559, 2017.

[52] L. E. McCollum, S. R. Das, L. Xie, R. de Flores, J. Wang, S. X. Xie, L. E. Wisse, P. A. Yushkevich, D. A. Wolk, A. D. N. Initiative, *et al.*, "Oh brother, where art tau? amyloid, neurodegeneration, and cognitive decline without elevated tau," *NeuroImage: Clinical*, vol. 31, p. 102717, 2021.

[53] S. Mukherjee, C. Klaus, M. Pricop-Jeckstadt, J. A. Miller, and F. L. Struebing, "A microglial signature directing human aging and neurodegeneration-related gene networks," *Frontiers in neuroscience*, vol. 13, p. 2, 2019.

[54] R. Tang and H. Liu, "Identification of temporal characteristic networks of peripheral blood changes in Alzheimer's disease based on weighted gene co-expression network analysis," *Frontiers in aging neuroscience*, vol. 11, p. 83, 2019.

[55] R. Postuma and J. Montplaisir, "Predicting Parkinson's disease–why, when, and how?," *Parkinsonism & related disorders*, vol. 15, pp. S105–S109, 2009.

[56] R. Armañanzas, C. Bielza, K. R. Chaudhuri, P. Martinez-Martin, and P. Larrañaga, "Unveiling relevant non-motor Parkinson's disease severity symptoms using a machine learning approach," *Artificial intelligence in medicine*, vol. 58, no. 3, pp. 195–202, 2013.

[57] S. Halbgebauer, M. Nagl, H. Klafki, U. Haußmann, P. Steinacker, P. Oeckl, J. Kassubek, E. Pinkhardt, A. C. Ludolph, H. Soininen, *et al.*, "Modified serpina1 as risk marker for parkinson's disease dementia: Analysis of baseline data," *Scientific reports*, vol. 6, p. 26145, 2016.

[58] U. Parkinson's, "The incidence and prevalence of parkinson's in the UK," *London, UK*, 2018.

[59] A. Kouli, K. M. Torsney, and W.-L. Kuan, "Parkinson's disease: etiology, neuropathology, and pathogenesis," *Exon Publications*, pp. 3–26, 2018.

[60] A. Reeve, E. Simcox, and D. Turnbull, "Ageing and Parkinson's disease: why is advancing age the biggest risk factor?," *Ageing research reviews*, vol. 14, pp. 19–30, 2014.

[61] T. F. Sheet, "National institute of neurological disorders and stroke, national institutes of health," *Updated Feb*, vol. 18, 2011.

[62] K. Wirdefeldt, H.-O. Adami, P. Cole, D. Trichopoulos, and J. Mandel, "Epidemiology and etiology of parkinson's disease: a review of the evidence," *European journal of epidemiology*, vol. 26, no. 1, p. 1, 2011.

[63] S. L. James, D. Abate, K. H. Abate, S. M. Abay, C. Abbafati, N. Abbasi, H. Abbastabar, F. Abd-Allah, J. Abdela, A. Abdelalim, *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017," *The Lancet*, vol. 392, no. 10159, pp. 1789–1858, 2018.

[64] M. Wanneveich, F. Moisan, H. Jacqmin-Gadda, A. Elbaz, and P. Joly, "Projections of prevalence, lifetime risk, and life expectancy of Parkinson's disease (2010-2030) in france," *Movement Disorders*, vol. 33, no. 9, pp. 1449–1455, 2018.

[65] J. L. Eriksen, Z. Wszolek, and L. Petrucelli, "Molecular pathogenesis of parkinson disease," *Archives of neurology*, vol. 62, no. 3, pp. 353–357, 2005.

[66] A. N. M. Copas, S. F. McComish, J. M. Fletcher, and M. A. Caldwell, "The pathogenesis of Parkinson's disease: A complex interplay between astrocytes, microglia, and t lymphocytes?," *Frontiers in Neurology*, vol. 12, 2021.

[67] S. J. Teipel, M. Grothe, S. Lista, N. Toschi, F. G. Garaci, and H. Hampel, "Relevance of magnetic resonance imaging for early detection and diagnosis of Alzheimer's disease," *Medical Clinics*, vol. 97, no. 3, pp. 399–424, 2013.

[68] M. Symms, H. Jäger, K. Schmierer, and T. Yousry, "A review of structural magnetic resonance neuroimaging," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 75, no. 9, pp. 1235–1244, 2004.

[69] G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson, "The clinical use of structural mri in Alzheimer's disease," *Nature Reviews Neurology*, vol. 6, no. 2, pp. 67–77, 2010.

[70] M. T. Vlaardingerbroek and J. A. Boer, *Magnetic resonance imaging: theory and practice*. Springer Science & Business Media, 2013.

[71] J. S. O'Brien and E. L. Sampson, "Lipid composition of the normal human brain: gray matter, white matter, and myelin," *Journal of lipid research*, vol. 6, no. 4, pp. 537–544, 1965.

[72] C. Leuze, M. Aswendt, E. Ferenczi, C. W. Liu, B. Hsueh, M. Goubran, Q. Tian, G. Steinberg, M. M. Zeineh, K. Deisseroth, *et al.*, "The separate effects of lipids and proteins on brain mri contrast revealed through tissue clearing," *Neuroimage*, vol. 156, pp. 412–422, 2017.

[73] J. V. Manjón, "Mri preprocessing," in *Imaging Biomarkers*, pp. 53–63, Springer, 2017.

[74] A. Carré, G. Klausner, M. Edjlali, M. Lerousseau, J. Briend-Diop, R. Sun, S. Ammari, S. Reuzé, E. Alvarez Andres, T. Estienne, *et al.*, "Standardization of brain mr images across machines and protocols: bridging the gap for mri-based radiomics," *Scientific reports*, vol. 10, no. 1, pp. 1–15, 2020.

[75] W. C. Peh and J. H. Chan, "Artifacts in musculoskeletal magnetic resonance imaging: identification and correction," *Skeletal radiology*, vol. 30, no. 4, pp. 179–191, 2001.

[76] E. Pusey, R. B. Lufkin, R. Brown, M. A. Solomon, D. D. Stark, R. Tarr, and W. Hanafee, "Magnetic resonance imaging artifacts: mechanism and clinical significance.," *Radiographics*, vol. 6, no. 5, pp. 891–911, 1986.

[77] M. B. T. Noor, N. Z. Zenia, M. S. Kaiser, S. A. Mamun, and M. Mahmud, "Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detection of Alzheimer's disease, parkinson's disease and schizophrenia," *Brain informatics*, vol. 7, no. 1, pp. 1–21, 2020.

[78] S. K. Jindal, S. Banerjee, R. Patra, and A. Paul, "Deep learning-based brain malignant neoplasm classification using mri image segmentation assisted by bias field correction and histogram equalization," in *Brain Tumor MRI Image Segmentation Using Deep Learning Techniques*, pp. 135–161, Elsevier, 2022.

[79] J. Juntu, J. Sijbers, D. V. Dyck, and J. Gielen, "Bias field correction for mri images," in *Computer recognition systems*, pp. 543–551, Springer, 2005.

[80] B. Fischl, "Freesurfer," *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.

[81] 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*, vol. 17, no. 1, pp. 87–97, 1998.

[82] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4itk: improved n3 bias correction," *IEEE transactions on medical imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.

[83] Y.-D. Xiao, R. Paudel, J. Liu, C. Ma, Z.-S. Zhang, and S.-K. Zhou, "Mri contrast agents: Classification and application," *International journal of molecular medicine*, vol. 38, no. 5, pp. 1319–1326, 2016.

[84] L. G. Nyúl and J. K. Udupa, "On standardizing the mr image intensity scale," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 42, no. 6, pp. 1072–1081, 1999.

[85] P. Juszczak, D. Tax, and R. P. Duin, "Feature scaling in support vector data description," in *Proc. asci*, pp. 95–102, Citeseer, 2002.

[86] A. Madabhushi and J. K. Udupa, "Interplay between intensity standardization and inhomogeneity correction in mr image processing," *IEEE Transactions on Medical Imaging*, vol. 24, no. 5, pp. 561–576, 2005.

[87] J. G. Park and C. Lee, "Skull stripping based on region growing for magnetic resonance brain images," *NeuroImage*, vol. 47, no. 4, pp. 1394–1407, 2009.

[88] F. Ségonne, A. M. Dale, E. Busa, M. Glessner, D. Salat, H. K. Hahn, and B. Fischl, "A hybrid approach to the skull stripping problem in mri," *Neuroimage*, vol. 22, no. 3, pp. 1060–1075, 2004.

[89] J. Swiebocka-Wiek, "Skull stripping for mri images using morphological operators," in *IFIP International Conference on Computer Information Systems and Industrial Management*, pp. 172–182, Springer, 2016.

[90] H. Lester and S. R. Arridge, "A survey of hierarchical non-linear medical image registration," *Pattern recognition*, vol. 32, no. 1, pp. 129–149, 1999.

[91] X. Zhang, Y. Feng, W. Chen, X. Li, A. V. Faria, Q. Feng, and S. Mori, "Linear registration of brain mri using knowledge-based multiple intermediator libraries," *Frontiers in neuroscience*, vol. 13, p. 909, 2019.

[92] K. K. Brock, S. Mutic, T. R. McNutt, H. Li, and M. L. Kessler, "Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the aapm radiation therapy committee task group no. 132," *Medical physics*, vol. 44, no. 7, pp. e43–e76, 2017.

[93] J. Talairach, "Co-planar stereotaxic atlas of the human brain-3-dimensional proportional system," *An approach to cerebral imaging*, 1988.

[94] D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, "Medical image registration," *Physics in medicine & biology*, vol. 46, no. 3, p. R1, 2001.

[95] A. C. Evans, D. L. Collins, S. Mills, E. D. Brown, R. L. Kelly, and T. M. Peters, "3d statistical neuroanatomical models from 305 mri volumes," in *1993 IEEE conference record nuclear science symposium and medical imaging conference*, pp. 1813–1817, IEEE, 1993.

[96] A. R. Laird, J. L. Robinson, K. M. McMillan, D. Tordesillas-Gutiérrez, S. T. Moran, S. M. Gonzales, K. L. Ray, C. Franklin, D. C. Glahn, P. T. Fox, *et al.*, "Comparison of the disparity between talairach and mni coordinates in functional neuroimaging data: validation of the lancaster transform," *Neuroimage*, vol. 51, no. 2, pp. 677–683, 2010.

[97] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.

[98] B. B. Avants, N. Tustison, G. Song, *et al.*, "Advanced normalization tools (ants)," *Insight j*, vol. 2, no. 365, pp. 1–35, 2009.

[99] J. G. Carbonell, R. S. Michalski, and T. M. Mitchell, "An overview of machine learning," *Machine learning*, pp. 3–23, 1983.

[100] A. Kaplan and M. Haenlein, "Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence," *Business Horizons*, vol. 62, no. 1, pp. 15–25, 2019.

[101] M. B. Hoy, "Alexa, siri, cortana, and more: an introduction to voice assistants," *Medical reference services quarterly*, vol. 37, no. 1, pp. 81–88, 2018.

[102] A. E. Sallab, M. Abdou, E. Perot, and S. Yogamani, "Deep reinforcement learning framework for autonomous driving," *Electronic Imaging*, vol. 2017, no. 19, pp. 70–76, 2017.

[103] P. Singhal, P. K. Srivastava, A. K. Tiwari, and R. K. Shukla, "A survey: Approaches to facial detection and recognition with machine learning techniques," in *Proceedings of Second Doctoral Symposium on Computational Intelligence*, pp. 103–125, Springer, 2022.

[104] M. Elgamal, "Automatic skin cancer images classification," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 4, no. 3, pp. 287–294, 2013.

[105] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, pp. 381–386, 2020.

[106] T. Hastie, R. Tibshirani, and J. Friedman, "Overview of supervised learning," in *The elements of statistical learning*, pp. 9–41, Springer, 2009.

[107] J. G. Bazan, H. S. Nguyen, S. H. Nguyen, P. Synak, and J. Wróblewski, "Rough set algorithms in classification problem," in *Rough set methods and applications*, pp. 49–88, Springer, 2000.

[108] M. A. Hardy, *Regression with dummy variables*, vol. 93. Sage, 1993.

[109] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[110] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[111] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[112] W. Wang, Y. Huang, Y. Wang, and L. Wang, "Generalized autoencoder: A neural network framework for dimensionality reduction," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 490–497, 2014.

[113] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.

[114] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, *et al.*, "A general reinforcement learning algorithm that masters chess, shogi, and go through self-play," *Science*, vol. 362, no. 6419, pp. 1140–1144, 2018.

[115] V. Grollemund, P.-F. Pradat, G. Querin, F. Delbot, G. Le Chat, J.-F. Pradat-Peyre, and P. Bede, "Machine learning in amyotrophic lateral sclerosis: achievements, pitfalls, and future directions," *Frontiers in neuroscience*, vol. 13, p. 135, 2019.

[116] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of big data*, vol. 2, no. 1, pp. 1–21, 2015.

[117] S. B. Kotsiantis, I. Zaharakis, P. Pintelas, *et al.*, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, no. 1, pp. 3–24, 2007.

[118] W. S. Noble, "What is a support vector machine?," *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[119] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[120] A. Ben-Hur, C. S. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch, "Support vector machines and kernels for computational biology," *PLoS computational biology*, vol. 4, no. 10, p. e1000173, 2008.

[121] L. Tan, "Code comment analysis for improving software quality," in *The Art and Science of Analyzing Software Data*, pp. 493–517, Elsevier, 2015.

[122] V. Kotu and B. Deshpande, *Data science: concepts and practice*. Morgan Kaufmann, 2018.

[123] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," *Journal of biomedical informatics*, vol. 35, no. 5-6, pp. 352–359, 2002.

[124] S. Ray, "A quick review of machine learning algorithms," in *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pp. 35–39, IEEE, 2019.

[125] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 1395–1403, 2015.

[126] M. A. Nielsen, *Neural networks and deep learning*, vol. 25. Determination press San Francisco, CA, USA, 2015.

[127] Y. Bengio, "Deep learning of representations: Looking forward," in *International conference on statistical language and speech processing*, pp. 1–37, Springer, 2013.

[128] C. Garling, "Andrew ng: Why 'deep learning'is a mandate for humans, not just machines," *Wired, May*, 2015.

[129] W. G. Hatcher and W. Yu, "A survey of deep learning: Platforms, applications and emerging research trends," *IEEE Access*, vol. 6, pp. 24411–24432, 2018.

[130] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[131] M. V. Valueva, N. Nagornov, P. A. Lyakhov, G. V. Valuev, and N. I. Chervyakov, "Application of the residue number system to reduce hardware costs of the convolutional neural network implementation," *Mathematics and Computers in Simulation*, vol. 177, pp. 232–243, 2020.

[132] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

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

[134] D. Scherer, A. Müller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," in *International conference on artificial neural networks*, pp. 92–101, Springer, 2010.

[135] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research*, vol. 134, no. 1, pp. 19–67, 2005.

[136] K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," *arXiv preprint arXiv:1702.05659*, 2017.

[137] H. H. Aghdam and E. J. Heravi, "Guide to convolutional neural networks," *New York, NY: Springer*, vol. 10, no. 978-973, p. 51, 2017.

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

[139] D. F. Shanno, "Conditioning of quasi-newton methods for function minimization," *Mathematics of computation*, vol. 24, no. 111, pp. 647–656, 1970.

[140] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[141] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[142] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[143] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[144] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

[145] M. Tan and Q. V. Le, "Efficientnetv2: Smaller models and faster training," *arXiv preprint arXiv:2104.00298*, 2021.

[146] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," *arXiv preprint arXiv:2106.04803*, 2021.

[147] A. Graves, "Supervised sequence labelling," in *Supervised sequence labelling with recurrent neural networks*, pp. 5–13, Springer, 2012.

[148] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[149] V. Carbune, P. Gonnet, T. Deselaers, H. A. Rowley, A. Daryin, M. Calvo, L.-L. Wang, D. Keysers, S. Feuz, and P. Gervais, "Fast multi-language lstm-based online handwriting recognition," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 23, no. 2, pp. 89–102, 2020.

[150] W. Feng, N. Guan, Y. Li, X. Zhang, and Z. Luo, "Audio visual speech recognition with multimodal recurrent neural networks," in *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 681–688, IEEE, 2017.

[151] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, 2014.

[152] S. Vieira, W. H. Pinaya, and A. Mechelli, "Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications," *Neuroscience & Biobehavioral Reviews*, vol. 74, pp. 58–75, 2017.

[153] H. Yu, L. T. Yang, Q. Zhang, D. Armstrong, and M. J. Deen, "Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives," *Neurocomputing*, vol. 444, pp. 92–110, 2021.

[154] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.

[155] J. Bernal, K. Kushibar, D. S. Asfaw, S. Valverde, A. Oliver, R. Martí, and X. Lladó, "Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review," *Artificial intelligence in medicine*, 2018.

[156] D. A. Ragab, M. Sharkas, S. Marshall, and J. Ren, "Breast cancer detection using deep convolutional neural networks and support vector machines," *PeerJ*, vol. 7, p. e6201, 2019.

[157] H. Chougrad, H. Zouaki, and O. Alheyane, "Deep convolutional neural networks for breast cancer screening," *Computer methods and programs in biomedicine*, vol. 157, pp. 19–30, 2018.

[158] M. Kirienko, M. Sollini, G. Silvestri, S. Mognetti, E. Voulaz, L. Antunovic, A. Rossi, L. Antiga, and A. Chiti, "Convolutional neural networks promising in lung cancer t-parameter assessment on baseline fdg-pet/ct," *Contrast Media & Molecular Imaging*, vol. 2018, 2018.

[159] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[160] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010.

[161] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLoS medicine*, vol. 15, no. 11, p. e1002683, 2018.

[162] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do cifar-10 classifiers generalize to cifar-10?," *arXiv preprint arXiv:1806.00451*, 2018.

[163] A. Blum and M. Hardt, "The ladder: A reliable leaderboard for machine learning competitions," *arXiv preprint arXiv:1502.04585*, 2015.

[164] Y. Xu and R. Goodacre, "On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning," *Journal of Analysis and Testing*, vol. 2, no. 3, pp. 249–262, 2018.

[165] J. Larsen and C. Goutte, "On optimal data split for generalization estimation and model selection," in *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No. 98TH8468)*, pp. 225–234, IEEE, 1999.

[166] S. Lohr, "Sampling: Design and analysis.," 1999.

[167] G. J. Bowden, H. R. Maier, and G. C. Dandy, "Optimal division of data for neural network models in water resources applications," *Water Resources Research*, vol. 38, no. 2, pp. 2–1, 2002.

[168] R. D. Snee, "Validation of regression models: methods and examples," *Technometrics*, vol. 19, no. 4, pp. 415–428, 1977.

[169] J. E. Trost, "Statistically nonrepresentative stratified sampling: A sampling technique for qualitative studies," *Qualitative sociology*, vol. 9, no. 1, pp. 54–57, 1986.

[170] N. Kriegeskorte, W. K. Simmons, P. S. Bellgowan, and C. I. Baker, "Circular analysis in systems neuroscience: the dangers of double dipping," *Nature neuroscience*, vol. 12, no. 5, p. 535, 2009.

[171] S. Rathore, M. Habes, M. A. Iftikhar, A. Shacklett, and C. Davatzikos, "A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages," *NeuroImage*, vol. 155, pp. 530–548, 2017.

[172] S. Sivaranjini and C. Sujatha, "Deep learning based diagnosis of parkinson's disease using convolutional neural network," *Multimedia Tools and Applications*, pp. 1–13, 2019.

[173] S. Esmaeilzadeh, Y. Yang, and E. Adeli, "End-to-end parkinson disease diagnosis using brain mr-images by 3d-cnn," *arXiv preprint arXiv:1806.05233*, 2018.

[174] R. Jain, N. Jain, A. Aggarwal, and D. J. Hemanth, "Convolutional neural network based Alzheimer's disease classification from magnetic resonance brain images," *Cognitive Systems Research*, vol. 57, pp. 147–159, 2019.

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

[176] A. Farooq, S. Anwar, M. Awais, and S. Rehman, "A deep cnn based multi-class classification of Alzheimer's disease using mri," in *2017 IEEE International Conference on Imaging systems and techniques (IST)*, pp. 1–6, IEEE, 2017.

[177] S. Sarraf and G. Tofighi, "Classification of Alzheimer's disease using fmri data and deep learning convolutional neural networks," *arXiv preprint arXiv:1603.08631*, 2016.

[178] C. Wu, S. Guo, Y. Hong, B. Xiao, Y. Wu, Q. Zhang, A. D. N. Initiative, *et al.*, "Discrimination and conversion prediction of mild cognitive impairment using convolutional neural networks," *Quantitative Imaging in Medicine and Surgery*, vol. 8, no. 10, p. 992, 2018.

[179] A. Payan and G. Montana, "Predicting Alzheimer's disease: a neuroimaging study with 3d convolutional neural networks," *arXiv preprint arXiv:1502.02506*, 2015.

[180] M. D. Abràmoff, P. T. Lavin, M. Birch, N. Shah, and J. C. Folk, "Pivotal trial of an autonomous ai-based diagnostic system for detection of diabetic retinopathy in primary care offices," *Npj Digital Medicine*, vol. 1, no. 1, p. 39, 2018.

[181] G. An, K. Omodaka, K. Hashimoto, S. Tsuda, Y. Shiga, N. Takada, T. Kikawa, H. Yokota, M. Akiba, and T. Nakazawa, "Glaucoma diagnosis with machine learning based on optical coherence tomography and color fundus images," *Journal of healthcare engineering*, vol. 2019, 2019.

[182] T. A. Shaikh and R. Ali, "Applying machine learning algorithms for early diagnosis and prediction of breast cancer risk," in *Proceedings of 2nd International Conference on Communication, Computing and Networking*, pp. 589–598, Springer, 2019.

[183] G. Garraux, C. Phillips, J. Schrouff, A. Kreisler, C. Lemaire, C. Degueldre, C. Delcour, R. Hustinx, A. Luxen, A. Destée, *et al.*, "Multiclass classification of fdg pet scans for the distinction between Parkinson's disease and atypical parkinsonian syndromes," *NeuroImage: Clinical*, vol. 2, pp. 883–893, 2013.

[184] M. Tahmasian, L. M. Bettray, T. van Eimeren, A. Drzezga, L. Timmermann, C. R. Eickhoff, S. B. Eickhoff, and C. Eggers, "A systematic review on the applications of

resting-state fmri in Parkinson's disease: does dopamine replacement therapy play a role?," *Cortex*, vol. 73, pp. 80–105, 2015.

[185] H. Lei, Y. Zhao, Y. Wen, Q. Luo, Y. Cai, G. Liu, and B. Lei, "Sparse feature learning for multi-class parkinson's disease classification," *Technology and Health Care*, vol. 26, no. S1, pp. 193–203, 2018.

[186] 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*, vol. 63, p. 101694, 2020.

[187] H. Mohsen, E.-S. A. El-Dahshan, E.-S. M. El-Horbaty, and A.-B. M. Salem, "Classification using deep learning neural networks for brain tumors," *Future Computing and Informatics Journal*, vol. 3, no. 1, pp. 68–71, 2018.

[188] U. Saeed, J. Compagnone, R. I. Aviv, A. P. Strafella, S. E. Black, A. E. Lang, and M. Masellis, "Imaging biomarkers in parkinson's disease and parkinsonian syndromes: current and emerging concepts," *Translational neurodegeneration*, vol. 6, no. 1, p. 8, 2017.

[189] J. P. Pluim, J. A. Maintz, and M. A. Viergever, "Interpolation artefacts in mutual information-based image registration," *Computer vision and image understanding*, vol. 77, no. 2, pp. 211–232, 2000.

[190] M. Ahmad, M. Z. Alam, Z. Umayya, S. Khan, and F. Ahmad, "An image encryption approach using particle swarm optimization and chaotic map," *International Journal of Information Technology*, vol. 10, no. 3, pp. 247–255, 2018.

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

[192] K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury, *et al.*, "The parkinson progression marker initiative (PPMI)," *Progress in neurobiology*, vol. 95, no. 4, pp. 629–635, 2011.

[193] B. Heim, F. Krismer, R. De Marzi, and K. Seppi, "Magnetic resonance imaging for the diagnosis of parkinson's disease," *Journal of neural transmission*, vol. 124, no. 8, pp. 915–964, 2017.

[194] K. Nogueira, O. A. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognition*, vol. 61, pp. 539–556, 2017.

[195] M. Gao, U. Bagci, L. Lu, A. Wu, M. Buty, H.-C. Shin, H. Roth, G. Z. Papadakis, A. Depeursinge, R. M. Summers, *et al.*, "Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 6, no. 1, pp. 1–6, 2018.

[196] J. M. Ortíz Rodríguez, M. d. R. Martínez Blanco, J. M. Cervantes Miramontes, H. R. Vega Carrillo, *et al.*, *Robust design of artificial neural networks methodology in neutron spectrometry*. IntechOpen, 2013.

[197] MATLAB, *version 9.5.0.944444 (R2018b)*. Natick, Massachusetts: The MathWorks Inc., 2018.

[198] F. Chollet, "keras." `https://github.com/fchollet/keras`, 2015.

[199] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265–283, 2016.

[200] H. Greenspan, B. Van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[201] G. Zaharchuk, E. Gong, M. Wintermark, D. Rubin, and C. Langlotz, "Deep learning in neuroradiology," *American Journal of Neuroradiology*, vol. 39, no. 10, pp. 1776–1784, 2018.

[202] K. Bahrami, F. Shi, X. Zong, H. W. Shin, H. An, and D. Shen, "Reconstruction of 7t-like images from 3t mri," *IEEE transactions on medical imaging*, vol. 35, no. 9, pp. 2085–2097, 2016.

[203] X. Han, "Mr-based synthetic ct generation using a deep convolutional neural network method," *Medical physics*, vol. 44, no. 4, pp. 1408–1419, 2017.

[204] R. Li, W. Zhang, H.-I. Suk, L. Wang, J. Li, D. Shen, and S. Ji, "Deep learning based imaging data completion for improved brain disease diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 305–312, Springer, 2014.

[205] F. Liu, H. Jang, R. Kijowski, T. Bradshaw, and A. B. McMillan, "Deep learning mr imaging–based attenuation correction for pet/mr imaging," *Radiology*, vol. 286, no. 2, pp. 676–684, 2018.

[206] R. Vemulapalli, H. Van Nguyen, and S. K. Zhou, "Deep networks and mutual information maximization for cross-modal medical image synthesis," in *Deep Learning for Medical Image Analysis*, pp. 381–403, Elsevier, 2017.

[207] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen, "Image reconstruction by domain-transform manifold learning," *Nature*, vol. 555, no. 7697, pp. 487–492, 2018.

[208] P. D. Chang, "Fully convolutional deep residual neural networks for brain tumor segmentation," in *International workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries*, pp. 108–118, Springer, 2016.

[209] Q. Dou, H. Chen, L. Yu, L. Zhao, J. Qin, D. Wang, V. C. Mok, L. Shi, and P.-A. Heng, "Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1182–1195, 2016.

[210] O. Maier, C. Schröder, N. D. Forkert, T. Martinetz, and H. Handels, "Classifiers for ischemic stroke lesion segmentation: a comparison study," *PloS one*, vol. 10, no. 12, p. e0145118, 2015.

[211] 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*, vol. 62, no. 4, pp. 1132–1140, 2014.

[212] S. M. Plis, D. R. Hjelm, R. Salakhutdinov, E. A. Allen, H. J. Bockholt, J. D. Long, H. J. Johnson, J. S. Paulsen, J. A. Turner, and V. D. Calhoun, "Deep learning for neuroimaging: a validation study," *Frontiers in neuroscience*, vol. 8, p. 229, 2014.

[213] C. Davatzikos, "Machine learning in neuroimaging: Progress and challenges," *Neuroimage*, vol. 197, p. 652, 2019.

[214] S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, and D. Feng, "Early diagnosis of Alzheimer's disease with deep learning," in *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*, pp. 1015–1018, IEEE, 2014.

[215] H.-I. Suk and D. Shen, "Deep learning-based feature representation for ad/mci classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 583–590, Springer, 2013.

[216] D. Kuang, X. Guo, X. An, Y. Zhao, and L. He, "Discrimination of adhd based on fmri data with deep belief network," in *International Conference on Intelligent Computing*, pp. 225–232, Springer, 2014.

[217] J. Islam and Y. Zhang, "Brain mri analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks," *Brain informatics*, vol. 5, no. 2, p. 2, 2018.

[218] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 6, no. 4, pp. 1–21, 2012.

[219] A. Torralba, A. A. Efros, *et al.*, "Unbiased look at dataset bias.," in *CVPR*, vol. 1, p. 7, Citeseer, 2011.

[220] A. Ashraf, S. Khan, N. Bhagwat, M. Chakravarty, and B. Taati, "Learning to unlearn: Building immunity to dataset bias in medical imaging studies," *arXiv preprint arXiv:1812.01716*, 2018.

[221] A. Blum, A. Kalai, and J. Langford, "Beating the hold-out: Bounds for k-fold and progressive cross-validation," in *Proceedings of the twelfth annual conference on Computational learning theory*, pp. 203–208, 1999.

[222] S. Yadav and S. Shukla, "Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification," in *2016 IEEE 6th International conference on advanced computing (IACC)*, pp. 78–83, IEEE, 2016.

[223] X. Han, R. Kwitt, S. Aylward, S. Bakas, B. Menze, A. Asturias, P. Vespa, J. Van Horn, and M. Niethammer, "Brain extraction from normal and pathological images: a joint pca/image-reconstruction approach," *NeuroImage*, vol. 176, pp. 431–445, 2018.

[224] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ants similarity metric performance in brain image registration," *Neuroimage*, vol. 54, no. 3, pp. 2033–2044, 2011.

[225] G. Bradski and A. Kaehler, "Learning opencv, ed. m. loukides," 2011.

[226] C. Laurent, G. Pereyra, P. Brakel, Y. Zhang, and Y. Bengio, "Batch normalized recurrent neural networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2657–2661, IEEE, 2016.

[227] X. Wang, J. Liu, T. Qiu, C. Mu, C. Chen, and P. Zhou, "A real-time collision prediction mechanism with deep learning for intelligent transportation system," *IEEE transactions on vehicular technology*, vol. 69, no. 9, pp. 9497–9508, 2020.

[228] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, JMLR Workshop and Conference Proceedings, 2010.

[229] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

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

[231] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," 2015.

[232] J. C. Morris, "Current vision and scoring rules the clinical dementia rating (cdr)," *Neurology*, vol. 43, pp. 2412–2414, 1993.

[233] J. C. Morris, M. Storandt, J. P. Miller, D. W. McKeel, J. L. Price, E. H. Rubin, and L. Berg, "Mild cognitive impairment represents early-stage Alzheimer's disease," *Archives of neurology*, vol. 58, no. 3, pp. 397–405, 2001.

[234] K. Marek, S. Chowdhury, A. Siderowf, S. Lasch, C. S. Coffey, C. Caspell-Garcia, T. Simuni, D. Jennings, C. M. Tanner, J. Q. Trojanowski, *et al.*, "The Parkinson's progression markers initiative (PPMI)–establishing a pd biomarker cohort," *Annals of clinical and translational neurology*, vol. 5, no. 12, pp. 1460–1477, 2018.

[235] M. M. Hoehn, M. D. Yahr, *et al.*, "Parkinsonism: onset, progression, and mortality," *Neurology*, vol. 50, no. 2, pp. 318–318, 1998.

[236] C. Tessa, N. Toschi, S. Orsolini, G. Valenza, C. Lucetti, R. Barbieri, and S. Diciotti, "Central modulation of parasympathetic outflow is impaired in de novo Parkinson's disease patients," *PloS one*, vol. 14, no. 1, p. e0210324, 2019.

[237] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC bioinformatics*, vol. 7, no. 1, pp. 1–8, 2006.

[238] T. Hastie, R. Tibshirani, and J. Friedman, "Springer series in statistics the elements of statistical learning data mining," *Inference, and Prediction*, 2009.

[239] S. Lemm, B. Blankertz, T. Dickhaus, and K.-R. Müller, "Introduction to machine learning for brain imaging," *Neuroimage*, vol. 56, no. 2, pp. 387–399, 2011.

[240] K. DP and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of the 3rd International Conference for Learning Representations (ICLR)*, 2015.

[241] F. Chollet *et al.*, "keras," 2015.

[242] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[243] M. Murad, M. Bilal, A. Jalil, A. Ali, K. Mehmood, and B. Khan, "Efficient reconstruction technique for multi-slice cs-mri using novel interpolation and 2d sampling scheme," *IEEE Access*, vol. 8, pp. 117452–117466, 2020.

[244] S. Kobayashi, T. B. Kane, and C. Paton, "The privacy and security implications of open data in healthcare," *Yearbook of medical informatics*, vol. 27, no. 01, pp. 041–047, 2018.

[245] I. Kandel and M. Castelli, "The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset," *ICT express*, vol. 6, no. 4, pp. 312–315, 2020.

[246] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*. Mit Press, 2008.

[247] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern recognition*, vol. 45, no. 1, pp. 521–530, 2012.

[248] L. A. Celi, L. Citi, M. Ghassemi, and T. J. Pollard, "The plos one collection on machine learning in health and biomedicine: Towards open code and open data," *PloS one*, vol. 14, no. 1, p. e0210232, 2019.

[249] J. Reunanen, "Overfitting in making comparisons between variable selection methods," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1371–1382, 2003.

[250] S. Alam, G.-R. Kwon, J.-I. Kim, and C.-S. Park, "Twin svm-based classification of Alzheimer's disease using complex dual-tree wavelet principal coefficients and lda," *Journal of healthcare engineering*, vol. 2017, 2017.

[251] X. Liu, D. Tosun, M. W. Weiner, N. Schuff, A. D. N. Initiative, *et al.*, "Locally linear embedding (lle) for mri based Alzheimer's disease classification," *Neuroimage*, vol. 83, pp. 148–157, 2013.

[252] K. R. Gray, P. Aljabar, R. A. Heckemann, A. Hammers, D. Rueckert, A. D. N. Initiative, *et al.*, "Random forest-based similarity measures for multi-modal classification of Alzheimer's disease," *NeuroImage*, vol. 65, pp. 167–175, 2013.

[253] K. Gunawardena, R. Rajapakse, and N. Kodikara, "Applying convolutional neural networks for pre-detection of Alzheimer's disease from structural mri data," in *2017 24th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, pp. 1–7, IEEE, 2017.

[254] A. Valliani and A. Soni, "Deep residual nets for improved Alzheimer's diagnosis," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 615–615, 2017.

[255] S.-H. Wang, P. Phillips, Y. Sui, B. Liu, M. Yang, and H. Cheng, "Classification of Alzheimer's disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling," *Journal of medical systems*, vol. 42, no. 5, p. 85, 2018.

[256] Y. R. Fung, Z. Guan, R. Kumar, J. Y. Wu, and M. Fiterau, "Alzheimer's disease brain mri classification: Challenges and insights," *arXiv preprint arXiv:1906.04231*, 2019.

[257] Y. Huang, J. Xu, Y. Zhou, T. Tong, X. Zhuang, A. D. N. I. (ADNI, *et al.*, "Diagnosis of alzheimer's disease via multi-modality 3d convolutional neural network," *Frontiers in Neuroscience*, vol. 13, p. 509, 2019.

[258] K. Oh, Y.-C. Chung, K. W. Kim, W.-S. Kim, and I.-S. Oh, "Classification and visualization of Alzheimer's disease using volumetric convolutional neural network and transfer learning," *Scientific Reports*, vol. 9, no. 1, pp. 1–16, 2019.

[259] S. Korolev, A. Safiullin, M. Belyaev, and Y. Dodonova, "Residual and plain convolutional neural networks for 3d brain mri classification," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 835–838, IEEE, 2017.

[260] E. Hosseini-Asl, G. Gimel'farb, and A. El-Baz, "Alzheimer's disease diagnostics by a deeply supervised adaptable 3d convolutional network," *arXiv preprint arXiv:1607.00556*, 2016.

[261] S. Wang, H. Wang, Y. Shen, and X. Wang, "Automatic recognition of mild cognitive impairment and Alzheimer's disease using ensemble based 3d densely connected convolutional networks," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 517–523, IEEE, 2018.

[262] J. Rieke, F. Eitel, M. Weygandt, J.-D. Haynes, and K. Ritter, "Visualizing convolutional networks for mri-based diagnosis of Alzheimer's disease," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pp. 24–31, Springer, 2018.

[263] C. Yang, A. Rangarajan, and S. Ranka, "Visual explanations from deep 3d convolutional neural networks for Alzheimer's disease classification," in *AMIA Annual Symposium Proceedings*, vol. 2018, p. 1571, American Medical Informatics Association, 2018.

[264] S. Chakraborty, S. Aich, and H.-C. Kim, "Detection of parkinson's disease from 3t t1 weighted mri scans using 3d convolutional neural network," *Diagnostics*, vol. 10, no. 6, p. 402, 2020.

[265] N. J. Dhinagar, S. I. Thomopoulos, C. Owens-Walton, D. Stripelis, J. L. Ambite, G. Ver Steeg, D. Weintraub, P. Cook, C. McMillan, and P. M. Thompson, "3d convolutional neural networks for classification of Alzheimer's and Parkinson's Disease with t1-weighted brain mri," *bioRxiv*, 2021.

[266] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, M. Chupin, H. Benali, O. Colliot, A. D. N. Initiative, *et al.*, "Automatic classification of patients with Alzheimer's disease from structural mri: a comparison of ten methods using the adni database," *neuroimage*, vol. 56, no. 2, pp. 766–781, 2011.

[267] D. Lu and Q. Weng, "A survey of image classification methods and techniques for improving classification performance," *International journal of Remote sensing*, vol. 28, no. 5, pp. 823–870, 2007.

[268] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 922–928, IEEE, 2015.

[269] H. Zunair, A. Rahman, N. Mohammed, and J. P. Cohen, "Uniformizing techniques to process ct scans with 3d cnns for tuberculosis prediction," in *International Workshop on PRedictive Intelligence In MEdicine*, pp. 156–168, Springer, 2020.

[270] E. Yagis, A. G. S. De Herrera, and L. Citi, "Convolutional autoencoder based deep learning approach for Alzheimer's disease diagnosis using brain mri," in *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 486–491, IEEE, 2021.

[271] M. Habes, R. Pomponio, H. Shou, J. Doshi, E. Mamourian, G. Erus, I. Nasrallah, L. J. Launer, T. Rashid, M. Bilgel, *et al.*, "The brain chart of aging: Machine-learning analytics reveals links between brain aging, white matter disease, amyloid burden, and cognition in the istaging consortium of 10,216 harmonized mr scans," *Alzheimer's & Dementia*, vol. 17, no. 1, pp. 89–102, 2021.

[272] J. Weese and C. Lorenz, "Four challenges in medical image analysis from an industrial perspective," 2016.

[273] C. S. Wickramasinghe, D. L. Marino, and M. Manic, "Resnet autoencoders for unsupervised feature learning from high-dimensional data: Deep models resistant to performance degradation," *IEEE Access*, vol. 9, pp. 40511–40520, 2021.

[274] A. E. Ilesanmi and T. O. Ilesanmi, "Methods for image denoising using convolutional neural network: a review," *Complex & Intelligent Systems*, vol. 7, no. 5, pp. 2179–2198, 2021.

[275] E. Pintelas, I. E. Livieris, and P. E. Pintelas, "A convolutional autoencoder topology for classification in high-dimensional noisy image datasets," *Sensors*, vol. 21, no. 22, p. 7731, 2021.

[276] Y. Sun, H. Mao, Q. Guo, and Z. Yi, "Learning a good representation with unsymmetrical auto-encoder," *Neural Computing and Applications*, vol. 27, no. 5, pp. 1361–1367, 2016.

[277] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," tech. rep., California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[278] Y. Le Cun, "Learning process in an asymmetric threshold network," in *Disordered systems and biological organization*, pp. 233–240, Springer, 1986.

[279] V. Arul, "Deep learning methods for data classification," in *Artificial Intelligence in Data Mining*, pp. 87–108, Elsevier, 2021.

[280] S. S. Kunapuli and P. C. Bhallamudi, "A review of deep learning models for medical diagnosis," *Machine Learning, Big Data, and IoT for Medical Informatics*, pp. 389–404, 2021.

[281] Y. Teganya and D. Romero, "Deep completion autoencoders for radio map estimation," *IEEE Transactions on Wireless Communications*, vol. 21, no. 3, pp. 1710–1724, 2021.

[282] T. Jo, K. Nho, and A. J. Saykin, "Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data," *Frontiers in aging neuroscience*, vol. 11, p. 220, 2019.

[283] F. J. Martinez-Murcia, A. Ortiz, J.-M. Gorriz, J. Ramirez, and D. Castillo-Barnes, "Studying the manifold structure of Alzheimer's disease: a deep learning approach using convolutional autoencoders," *IEEE journal of biomedical and health informatics*, vol. 24, no. 1, pp. 17–26, 2019.

[284] W. G. Rosen, R. C. Mohs, and K. L. Davis, "A new rating scale for Alzheimer's disease.," *The American journal of psychiatry*, 1984.

[285] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""mini-mental state": a practical method for grading the cognitive state of patients for the clinician," *Journal of psychiatric research*, vol. 12, no. 3, pp. 189–198, 1975.

[286] S. Basu, K. Wagstyl, A. Zandifar, D. L. Collins, A. Romero, and D. Precup, "Early prediction of Alzheimer's disease progression using variational autoencoders.," in *MICCAI (4)*, pp. 205–213, 2019.

[287] R. Ferri, C. Babiloni, V. Karami, A. I. Triggiani, F. Carducci, G. Noce, R. Lizio, M. T. Pascarelli, A. Soricelli, F. Amenta, *et al.*, "Stacked autoencoders as new models for an accurate Alzheimer's disease classification support using resting-state eeg and mri measurements," *Clinical Neurophysiology*, vol. 132, no. 1, pp. 232–245, 2021.

[288] R. Mendoza-Léon, J. Puentes, L. F. Uriza, and M. H. Hoyos, "Single-slice Alzheimer's disease classification and disease regional analysis with supervised switching autoencoders," *Computers in biology and medicine*, vol. 116, p. 103527, 2020.

[289] R. A. M. Leon, J. Puentes, F. A. González, and M. H. Hoyos, "Empirical evaluation of general-purpose image features for pathology-oriented image retrieval of Alzheimer's disease cases," in *CARS 2016: 30th International Congress on Computer Assisted Radiology and Surgery In: International Journal of Computer Assisted Radiology and Surgery*, vol. 11, pp. S39–S40, Springer, 2016.

[290] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, pp. 818–833, Springer, 2014.

[291] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[292] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions," *Journal of big Data*, vol. 8, no. 1, pp. 1–74, 2021.

[293] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into imaging*, vol. 9, no. 4, pp. 611–629, 2018.

[294] A. Rosenberg, A. Solomon, V. Jelic, G. Hagman, N. Bogdanovic, and M. Kivipelto, "Progression to dementia in memory clinic patients with mild cognitive impairment and normal β-amyloid," *Alzheimer's research & therapy*, vol. 11, no. 1, pp. 1–12, 2019.

[295] A. Makhzani and B. Frey, "K-sparse autoencoders," *arXiv preprint arXiv:1312.5663*, 2013.

[296] A. A. Cruz-Roa, J. E. A. Ovalle, A. Madabhushi, and F. A. G. Osorio, "A deep learning architecture for image representation, visual interpretability and automated basal-cell

carcinoma cancer detection," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 403–410, Springer, 2013.

[297] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep learning to improve breast cancer detection on screening mammography," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[298] P. Danaee, R. Ghaeini, and D. A. Hendrix, "A deep learning approach for cancer detection and relevant gene identification," in *Pacific symposium on biocomputing 2017*, pp. 219–229, World Scientific, 2017.

[299] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, and Y. Fan, "A deep learning model integrating fcnns and crfs for brain tumor segmentation," *Medical image analysis*, vol. 43, pp. 98–111, 2018.

[300] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical image analysis*, vol. 35, pp. 18–31, 2017.

[301] D. Sułot, D. Alonso-Caneiro, P. Ksieniewicz, P. Krzyzanowska-Berkowska, and D. R. Iskander, "Glaucoma classification based on scanning laser ophthalmoscopic images using a deep learning ensemble method," *Plos one*, vol. 16, no. 6, p. e0252339, 2021.

[302] C. Ju, A. Bibaut, and M. van der Laan, "The relative performance of ensemble methods with deep convolutional neural networks for image classification," *Journal of Applied Statistics*, vol. 45, no. 15, pp. 2800–2818, 2018.

[303] N. Fogel, "Tuberculosis: a disease without boundaries," *Tuberculosis*, vol. 95, no. 5, pp. 527–531, 2015.

[304] N. Bolscher, K. Hoppenbrouwers, and R. Burgmeijer, "Tuberculose," 2007.

[305] I. Barberis, N. Bragazzi, L. Galluzzo, and M. Martini, "The history of tuberculosis: from the first historical records to the isolation of koch's bacillus," *Journal of preventive medicine and hygiene*, vol. 58, no. 1, p. E9, 2017.

[306] "Tuberculosis facts." `https://www.who.int/news-room/fact-sheets/detail/tuberculosis#:~:text=Worldwide%2C%20TB%20is%20one%20of,all%20countries%20and%20age%20groups.`

[307] W. H. O. (WHO), "Global tuberculosis report,2020," pp. 55–56.

[308] P. Pongwittayapanu, T. Anothaisintawee, K. Malathum, and C. Wongrathanandha, "Incidence of newly diagnosed tuberculosis among healthcare workers in a teaching hospital, thailand," vol. 84, no. 3, pp. 342–347.

[309] J. Foulds and R. O'brien, "New tools for the diagnosis of tuberculosis: the perspective of developing countries," *The International Journal of Tuberculosis and Lung Disease*, vol. 2, no. 10, pp. 778–783, 1998.

[310] E. Okur, A. Yilmaz, A. Saygi, A. Selvi, F. Süngün, E. Öztürk, and G. Dabak, "Patterns of delays in diagnosis amongst patients with smear-positive pulmonary tuberculosis at a teaching hospital in turkey," *Clinical microbiology and infection*, vol. 12, no. 1, pp. 90–92, 2006.

[311] A. C. Nachiappan, K. Rahbar, X. Shi, E. S. Guy, E. J. Mortani Barbosa Jr, G. S. Shroff, D. Ocazionez, A. E. Schlesinger, S. I. Katz, and M. M. Hammer, "Pulmonary tuberculosis: role of radiology in diagnosis and management," *Radiographics*, vol. 37, no. 1, pp. 52–72, 2017.

[312] W. H. O. W. G. T. Programme, "Who consolidated guidelines on tuberculosis module 2: Screening – systematic screening for tuberculosis disease," p. 68.

[313] N. Field, J. Murray, M. L. Wong, R. Dowdeswell, N. Dudumayo, L. Rametsi, N. Martinson, M. Lipman, J. R. Glynn, and P. Sonnenberg, "Missed opportunities in tb diagnosis: a tb process-based performance review tool to evaluate and improve clinical care," *BMC Public Health*, vol. 11, no. 1, pp. 1–7, 2011.

[314] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative imaging in medicine and surgery*, vol. 4, no. 6, p. 475, 2014.

[315] H.-P. Chan, R. K. Samala, L. M. Hadjiiski, and C. Zhou, "Deep learning in medical image analysis," *Deep Learning in Medical Image Analysis*, pp. 3–21, 2020.

[316] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers of Computer Science*, vol. 14, no. 2, pp. 241–258, 2020.

[317] R. Logan, B. G. Williams, M. Ferreira da Silva, A. Indani, N. Schcolnicov, A. Ganguly, and S. J. Miller, "Deep convolutional neural networks with ensemble learning and generative adversarial networks for Alzheimer's disease image data classification," *Frontiers in aging neuroscience*, p. 497, 2021.

[318] X. Zheng, J. Shi, Q. Zhang, S. Ying, and Y. Li, "Improving mri-based diagnosis of Alzheimer's disease via an ensemble privileged information learning algorithm," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 456–459, IEEE, 2017.

[319] H.-I. Suk, S.-W. Lee, D. Shen, A. D. N. Initiative, *et al.*, "Deep ensemble learning of sparse regression models for brain disease diagnosis," *Medical image analysis*, vol. 37, pp. 101–113, 2017.

[320] X. Tan, Y. Liu, Y. Li, P. Wang, X. Zeng, F. Yan, and X. Li, "Localized instance fusion of mri data of Alzheimer's disease for classification based on instance transfer ensemble learning," *Biomedical engineering online*, vol. 17, no. 1, pp. 1–17, 2018.

[321] N. An, H. Ding, J. Yang, R. Au, and T. F. Ang, "Deep ensemble learning for Alzheimer's disease classification," *Journal of biomedical informatics*, vol. 105, p. 103411, 2020.

[322] 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*, vol. 11, no. 1, pp. 1–13, 2021.

[323] C. Liu, Y. Cao, M. Alcantara, B. Liu, M. Brunette, J. Peinado, and W. Curioso, "Tx-cnn: Detecting tuberculosis in chest x-ray images using convolutional neural network," in *2017 IEEE international conference on image processing (ICIP)*, pp. 2314–2318, IEEE, 2017.

[324] O. Yadav, K. Passi, and C. K. Jain, "Using deep learning to classify x-ray images of potential tuberculosis patients," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2368–2375, IEEE, 2018.

[325] J. Howard and S. Gugger, "Fastai: a layered api for deep learning," *Information*, vol. 11, no. 2, p. 108, 2020.

[326] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, 2017.

[327] L. Li, H. Huang, and X. Jin, "Ae-cnn classification of pulmonary tuberculosis based on ct images," in *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, pp. 39–42, IEEE, 2018.

[328] M. Norval, Z. Wang, and Y. Sun, "Pulmonary tuberculosis detection using deep learning convolutional neural networks," in *Proceedings of the 3rd International Conference on Video and Image Processing*, pp. 47–51, 2019.

[329] T. Rahman, A. Khandakar, M. A. Kadir, K. R. Islam, K. F. Islam, R. Mazhar, T. Hamid, M. T. Islam, S. Kashem, Z. B. Mahbub, *et al.*, "Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization," *IEEE Access*, vol. 8, pp. 191586–191601, 2020.

[330] M. Rahman, Y. Cao, X. Sun, B. Li, and Y. Hao, "Deep pre-trained networks as a feature extractor with xgboost to detect tuberculosis from chest x-ray," *Computers & Electrical Engineering*, vol. 93, p. 107252, 2021.

[331] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, and S. A. Barman, "An ensemble classification-based approach applied to retinal blood

vessel segmentation," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 9, pp. 2538–2548, 2012.

[332] M. Ayaz, F. Shaukat, and G. Raja, "Ensemble learning based automatic detection of tuberculosis in chest x-ray images using hybrid feature descriptors," *Physical and Engineering Sciences in Medicine*, vol. 44, no. 1, pp. 183–194, 2021.

[333] S. Rajaraman and S. K. Antani, "Modality-specific deep learning model ensembles toward improving tb detection in chest radiographs," *IEEE Access*, vol. 8, pp. 27318–27326, 2020.

[334] A. C. Evans, A. L. Janke, D. L. Collins, and S. Baillet, "Brain templates and atlases," *Neuroimage*, vol. 62, no. 2, pp. 911–922, 2012.

[335] P. Misra and A. S. Yadav, "Impact of preprocessing methods on healthcare predictions," in *Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE)*, 2019.

[336] S. Almuhaideb and M. E. B. Menai, "Impact of preprocessing on medical data classification," *Frontiers of Computer Science*, vol. 10, no. 6, pp. 1082–1102, 2016.

[337] R. H. Sherrier and G. Johnson, "Regionally adaptive histogram equalization of the chest," *IEEE transactions on medical imaging*, vol. 6, no. 1, pp. 1–7, 1987.

[338] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A New Multi-scale Backbone Architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 652–662, Feb. 2021. arXiv: 1904.01169.

[339] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep Layer Aggregation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (Salt Lake City, UT), pp. 2403–2412, IEEE, June 2018.

[340] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," *arXiv:1611.05431 [cs]*, Apr. 2017. arXiv: 1611.05431.

[341] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, pp. 1–15, Springer, 2000.

[342] The MONAI Consortium, "Project MONAI," *Zenodo*, 2020.

[343] B. Ginsburg, P. Castonguay, O. Hrinchuk, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, H. Nguyen, Y. Zhang, and J. M. Cohen, "Stochastic Gradient Methods with Layer-wise Adaptive Moments for Training of Deep Networks," *arXiv:1905.11286 [cs, stat]*, Feb. 2020. arXiv: 1905.11286.

[344] I. Satia, S. Bashagha, A. Bibi, R. Ahmed, S. Mellor, and F. Zaman, "Assessing the accuracy and certainty in interpreting chest x-rays in the medical division," vol. 13.

[345] E. U. Ekpo, N. O. Egbe, and B. E. Akpan, "Radiographers' performance in chest x-ray interpretation: the nigerian experience," vol. 88, no. 1051, p. 20150023.

[346] C. J. Waitt, E. C. Joekes, N. Jesudason, P. I. Waitt, P. Goodson, G. Likumbo, S. Kampondeni, E. B. Faragher, and S. B. Squire, "The effect of a tuberculosis chest x-ray image reference set on non-expert reader performance," vol. 23, no. 9, pp. 2459–2468.