CHARACTERIZATION AND CLASSIFICATION OF ALZHEIMER'S DISEASE BY THE SEGMENTATION OF CORTICAL AND SUBCORTICAL REGIONS USING DEEP LEARNING FRAMEWORK

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

MRI	Magnetic Resonance Imaging
T1-w	T1-Weighted
CNN	Convolutional Neural Network
WM	White Matter
GM	Gray Matter
CSF	Cerebrospinal Fluid

ABSTRACT

Alzheimer's disease is a neurodegenerative disorder that affects aging human population. It causes progressive degeneration of the nervous system, affecting the cognitive ability of human brain. The disorder causes the growth of amyloid plaques and tau tangles to abnormal levels due to excessive synthesis of the constituent proteins. This manifests itself as memory loss, vision loss and overall cognitive impairment leaving the subject dependent on external help for conduct of daily life. The disease is projected to affect a considerable proportion of population in the future. In the context of its wide-ranging effects on broader health system, it is imperative to develop sophisticated methods for early detection and diagnosis of Alzheimer's disease. This can help to delay the onset of the disease. Over the past decade, neuroimaging data from Magnetic Resonance Imaging (MRI) scans have been increasingly used in the study of the brain pathology related to the birth and growth of AD. In the wake of advanced machine learning techniques, recent studies have employed machine learning for detection and classification of AD. Further, deep learning models have also been increasingly employed with varying degree of success. This thesis presents a novel hybrid approach for early detection and classification of AD using structural Magnetic Resonance Imaging (sMRI). Our model employs a unique combination of classical machine learning and deep learning approaches to optimize precision and accuracy of the intelligent detection and classification. The hybrid approach also helps overcome shortcomings of individual approaches regarding computational limitations. The approach combines ML and convolutional neural networks (CNN) where we fuse the image features extracted using classical machine learning techniques with features extracted by a convolutional neural network, before feeding them to classifier. We used a single most widely available modality i.e. sMRI. Our approach advantageously uses processed MRI scans and the extracted features to yield more information. The segmentation of cortical and subcortical regions through our framework helps in breaking spurious correlations in training data, resulting in better generalization. Our approach surpassed multi-modal machine learning algorithms in performance parameters of accuracy, precision, and F1 score. It is pertinent to note that the approach achieved this using only a single modality. Further, the results yielded by the hybrid approach beat Deep Learning models with better performance on validation data sets. To the best of our knowledge, our results in both *AD versus NC* and *sMCI versus pMCI* paradigms defeat the state-of-the-art of the field. Within the *NC versus AD*

paradigm the model we developed achieved 91.84% accuracy on test data. We achieved an accuracy of more than 93% in validation dataset. The model achieved significant improvements in other parameters such as precision and F1 score. It is pertinent to note that the model was trained and tested under two paradigms, namely *NC versus AD* and *sMCI versus pMCI*. These reflect the two binary classifications employed to assess the development of the disease by our model. We conclude in a nutshell the proposed hybrid approach can help bypass bottlenecks of the classical ML and DL techniques and reap better results.

CHAPTER 1.

INTRODUCTION

1.1 Alzheimer's Disease – Overview

Alzheimer's disease (AD) is a brain condition in aging population, which causes progressive degeneration in the nervous system. It consists of a gradual disablement of cognition, usually manifested as early loss of memory in the patients. It usually ends up leaving the patients to require assistance in self-care in severe cases.

The disorder causes the growth of amyloid plaques and tau tangles to abnormal levels due to excessive synthesis of the constituent proteins. Resultantly, the hitherto healthy neurons stop effective functioning and lose connections, eventually leading to their death. Initially, the hippocampus and the entorhinal cortex are damaged. As more neurons die, more parts of the brain are diseased and begin to shrink. Progressively, the neuron deaths begin to affect greater areas of brain causing brain to shrink effectively. The final stage of AD is characterized by a significant shrinkage in brain size and functionality.

Predominant symptoms of the AD include memory loss, vision loss, impaired cognitive ability, difficulty in body's locomotive functions and problems with the sense of smell.

1.1.1 Mild Stage Alzheimer's

The behavioral aspect of the mild Alzheimer's disease include difficulty in performing routine life tasks. This includes loss of memory and a general cognitive impairment. This often leads to wandering, difficulty in handling financial activity and paying bills, redundant speech, sluggishness in daily tasks and general personality erosion.

1.1.2 Moderate Stage Alzheimer's disease

In moderate level of the disease, the brain is affected on multiple fronts. The areas of brain controlling speech, cognition, consciousness and sensory processes are impaired. The loss in memory is aggravated and patient's confusion regarding routine life grows significantly.

1.1.3 Severe Stage Alzheimer's disease

The severe case of the AD is characterized by the spread of protein plaques and tangles to larger areas of the brain. The shrinking of the brain tissues is also increased to severe levels. The patients of severe cases are unable to communicate normally and become completely dependent in the conduct of daily life.

1.2 Magnetic Resonance Imaging (MRI)

The advancement in Machine Learning (ML) methods in the past decades have accelerated its use in medical applications. Over the past decade, neuroimaging data from Magnetic Resonance Imaging (MRI) scans has been increasingly used in the study of the brain pathology related to the birth and growth of AD.

Magnetic resonance imaging (MRI) is an efficient radiological, noninvasive imaging technology to produce detailed images of the body's different internal organs and tissues. MRI scanners obtain 3D images by using radio waves and strong magnetic fields. MRI scans differ from CT scan and X-rays in a way that they do not employ hazardous radiation. MRI has evolved clinically during the past 40 years [10]. MRI is the most powerful tool for diagnosing central nervous system disorders. It is also used to diagnose and evaluate multiple sclerosis [11]. Philips, Hitachi, Siemens, and GE (General Electric) are the most well-known brand names for MRI scanners.

MRI scanners can also be classified based on the strength of the magnet, which is measured in a unit called Tesla (T) and their orientation or construction, which can be closed, open, or upright. MRI scanners use magnetic strength ranging from 0.5T to 7.0T. The most common type of MRI scanners used in the diagnostic process is 1.5T. It provides fast scan times, excellent image quality, and the capability to evaluate how specific organs function in the human body. A 3.0T MRI scanner is twice as powerful as a 1.5T. It is proven to be ideal for visualizing brain, vascular and small bone systems. Current uses of 7.0T MRI increased in clinical trials, and clinical practice as signal-to-noise (SNR) and contrast-to-noise ratio (CNR) are lower than those of the fields [12].

A 3D model of the brain can be acquired in three distinct orientations using brain MRI. These orientations are named as sagittal (from the side), coronal (from the front) and axial (from the top-down). A 3x1 vector determines the resolution of an image, with each axis representing one of the

orientations. A spatial sequence of 2D slices containing an object of interest is used to create an image. A slice is a pixel matrix in x and y coordinates. The z coordinate indicates the slice number.

There are different kinds of MRI modalities. In our work, we used T1-weighted out of three different modalities: T1-weighted MPRAGE (T1-w), T2-weighted (T2-w) and fluid-attenuated inversion recovery (FLAIR). Each modality highlights various healthy and abnormal tissues, specifically:

• **T1-weighted:** This is the best approximation of the MR sequence for the anatomy of the organ of interest. Fluids like CSF have lower intensities. As compared to gray matter, which has intermediate intensity values, white matter is more hyperintense.

• **T2-weighted:** Fluids have high intensities; gray matter, intermediate and white matter have lower intensity values.

• **FLAIR:** In this type of images routine, CSF is aberrated and darkened while the abnormalities appear bright in the images. This sequence is pathologically very sensitive and facilitates differentiating between CSF and abnormality.

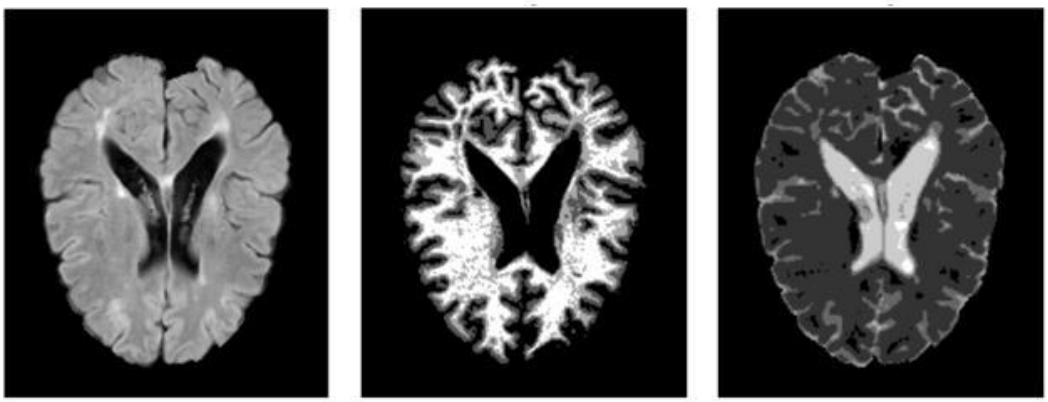

Figure 1 – MRI modalities (a) FLAIR, (b) T1- Weighted, (c) T2- Weighted. (From left to right)

1.3 Motivation

The disease is on the forefront of diseases which cause dementia and its derivatives related to loss of memory [1, 2]. In the United States alone, AD is the sixth leading cause of death and it accounts for 3.6% of the total deaths in the country [3, 4]. It is expected that 1 out 85 people will be affected by the disease globally by 2050. Nevertheless, though advanced medical techniques could delay the onset of the disorder by one year, it is expected to reduce the cases by 9.2 million in 2050 [5]. In past, major efforts have been made to develop strategies for early detection, especially in the pre-symptomatic stage, to delay or prevent the development of the disease [6, 7]. All the aforementioned challenges motivate us to propose an automatic method for early detection and classification of AD.

1.4 Contributions

We present a novel hybrid (Deep Learning + classical Machine Learning) approach for early detection and classification of AD using structural Magnetic Resonance Imaging (sMRI) with the following primary contributions:

- Relying only on single most widely available modality i.e. sMRI
- A hybrid approach to overcome shortcomings of individual approaches
- Taking advantage of more information available (processed MRI scans + Features extracted from these scans)
- Breaking spurious correlations in training data leading to better generalization
- Beating state of the art multi modal machine learning algorithms in terms of performance by using only single modality
- Beating state of the art Deep Learning models with better performance on validation data

1.5 Thesis Outline

This thesis is structured as follows. Chapter 2 describes an overview of the work others have done in classification of Alzheimer's disease. Our proposed methodology and baseline configuration are described in Chapter 3. After that, the results obtained from the proposed approach and its comparison

with state-of-the-art methods are presented in Chapter 4. In chapter 4, we also described the directions for future work.

CHAPTER 2.

LITERATURE REVIEW

In the context of this chapter, we first review the literature of the latest ways employed for classification in medical images using classical Machine Learning based approaches and recent Deep Learning techniques and then we proceed to review challenges faced in each approach.

2.1 Classical Machine Learning based Approaches

For the past decade, the advancements in Machine Learning (ML) have been increasingly employed to classify AD using the available neuroimaging data, yielding promising results for the diagnosis and prognosis of the disease [8-19]. A significant number of studies proposed using predefined features extracted from image preprocessing pipelines and application of various classifiers such as RF and SVM. The features include regional and voxel-based measurements.

Classification using machine learning methods broadly require four essential steps namely feature extraction, selection of features, reducing dimensions of data, and selection of feature-based classification algorithms.

2.1.1 Feature Extraction

For reliable feature extraction, first image preprocessing is performed for all modalities included. The image types and the aims of the preprocessing algorithm greatly influences the nature of the preprocessing procedure. Practically, medical images can be given in formats such as Nifti, DICOM, analyze and Minc.

Common preprocessing steps include anterior commissure (AC)–posterior commissure (PC) correction, N4 bias field correction to correct the intensity of inhomogeneity, skull striping, affine regularization, WM hyperintensity correction and spatial registration to a provided template etc. Later these preprocessed scans are segmented into multiple cortical and sub-cortical regions using an Atlas. Then based on the objective, several measures are taken on these segmented regions such as mean subcortical volumes, densities of gray matter, gray matter volumes, cortical thickness, metabolism of

brain glucose, and accumulation of cerebral amyloid-b etc. These measures are used as features for classification task.

There are several software tools available to perform medical image processing. Following are some tools often used by researchers. Visualization Toolkit (VTK) [41], The Insight Toolkit (ITK) [42], MRIB Software Library (FSL) [43], Statistical Parametric Mapping (SPM) [44], NiftyReg [45], Advances Normalization Tools (ANTS) [46], Medical Imaging Interaction Toolkit (MITK) [47].

2.1.2 Feature Selection

As a broad definition, feature selection is automated or manual selection of features in an image or other forms of data which aid most to the required prediction variable or output of interest. This implies that irrelevant features in training data can decrease accuracy of the machine learning models.

2.1.3 Dimensionality Reduction

The reliability and accuracy of Machine Learning models are greatly constricted by large numbers of features in a given data set as well as small number of observations. This is often cited as the "curse of dimensionality" and "small-n-large-p problem". To reduce the unwanted experimental noise and redundant features, predictor variable reduction techniques such as subset selection and dimensionality reductions are employed.

2.1.4 Feature-based Classification Algorithms

Several algorithms such as decision trees, support vector machines, K-nearest neighbor, ensemble linear discriminant, boosted trees, and random forests etc. are used for classification using selected features.

Overall workflow of classical machine learning approaches is given in Figure 2.

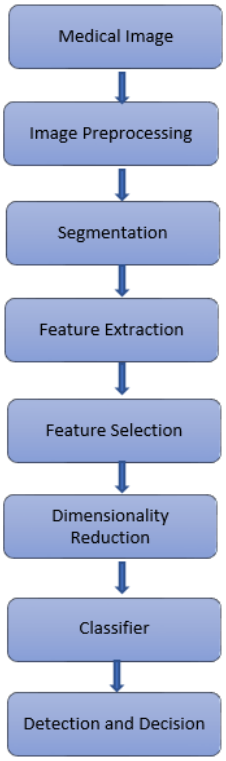

Figure 2 – Overall flow of classical machine learning algorithms

Keeping in view the classical machine learning approaches, we reviewed the state-of-the-art from two perspectives. First, we specifically studied the types and number of modalities used in the selected literature. This was compounded with studying the feature selection criteria by these studies. Secondly, we reviewed the state-of-the-art literature to explore the types of classifiers used and whether there was any specifics related to dimensionality reduction.

A considerable portion of studies in this literature review used single modality, particularly MRI. In [8], the study differentiated AD patients and elderly individuals with normal cognition, using data from multiple scanners. It used T1-W MRI modality. The study used grey matter volumes as features for classification. Whereas, [9] used single modality i.e. MRI and cortical thickness as features. Nevertheless, the study focused on transforming multivariate data to locally linear space having lesser number of dimensions. Further, [10] used three different features, namely, grey matter, white matter and cerebrospinal fluid. It employed MRI as its single modality, and used different clinical features in addition to that. However, the focus of the study was improving the effectiveness of texture based features. To achieve this the authors used grey-level co-occurrence matrix (GLCM), histogram of

gradient and scale invariant feature transform etc. as bag of visual word. In [11], while there is only a single modality, the study nevertheless used three biomarkers namely cerebrospinal fluid, cognitive score and APOE. The features incorporated in the study are cortical thickness, surface area and grey matter volume. Furthermore, some studies used additional modalities. In [12], multi-modality approach was employed. In addition to MRI, PET modality was employed, primarily helping to improve classification accuracy. It is pertinent to note that the study's fundamental focus remained to determine the influence of different classifiers. In [13], voxel-based morphometry was used for feature extraction and SVM classifier to classify MRI images in AD versus CN binary classification. The focus of [14] was on using Fisher's linear discrimination algorithm for feature selection and they used RF classifier in addition to SVM. In addition to the aforementioned studies, [17] used MRI as its single modality. The focus of the study was using ensemble SVM with bagging without replacement for classification. The studies [16] and [18] used multimodal approach and used different biomarkers. The features selected in these include hippocampal volume, volume of cortical and sub-cortical regions, grey matter volume of cortical and sub-cortical regions etc.

In the second perspective of exploring the state-of-the-art, we observed the dimensionality reduction and types of classifiers used in the studies cited above. As a matter of observation, support vector machine classifier is most popular in the studies reviewed. While almost all the studies used SVM, however, not all of them used it as sole classifier. Moreover, most of the studies focused on the classifier more than the dimensionality reduction. In particular, in [16] and [17] ensemble of classifiers was used in each study. The classifiers included SVM, decision trees, ensemble linear discriminant, boosted trees and K-nearest neighbor etc. Moreover, [9] and [18] were particular in emphasizing dimensionality reduction before feeding to classifier. The study in [9] employed local linear embedding (LLE) for the purpose of reducing dimensions. Further, [18] employed truncated SVD (T-SVD) for the purpose.

It is imperative to note that most of the studies used a single binary classification paradigm, namely AD versus CN. Nevertheless [10], [11], [14], [18] used multi-class classification paradigms in their work.

We have reviewed ML based studies so far. However, before moving on to Deep Learning based studies, it is important to reiterate the core methodology of ML based solutions. ML based

classification studies are commonly executed in four broad steps: extraction of features, selection of features, dimensionality reduction, and selection of algorithms for feature-based classification. The process requires specialized skillset and needs optimization, which in total is time intensive activity. Moreover, the reproducibility of this approach is a persistent problem [20]. For instance, in the process of feature selection, features related to AD are chosen from different modalities of neuroimaging so as to obtain meaningful combinatorial measures which may include "mean subcortical volumes, gray matter densities, cortical thickness, brain glucose metabolism, and cerebral amyloid-b accumulation in regions of interest (ROIs), such as the hippocampus" [21].

Machine Learning models using features from single modality often underperform. To overcome this problem, people often use multiple biomarkers and modalities to obtain more meaningful combinatorial measures. These biomarkers and modalities may include

- structural magnetic resonance imaging (sMRI) scans
- Positron emission tomography (PET) scans
- cerebrospinal fluid (CSF)
- cognitive score
- APOE ε4 allele status
- clinical features

All of these biomarkers and modalities may not be readily available for all patients, making it difficult to use multi-modality-based models in production.

ML models are severely impeded by large number of variables or features and low number of observations. This is known as the curse of dimensionality. The redundant variables and accompanying experimental noise are reduced using various techniques of feature reduction. This is an imperative of ML before training a model so that overfitting is avoided. Nevertheless, this dimensionality reduction may result in loss of data, faulty linear relationship being developed, both undesirable side effects. Moreover, we may not be certain about the required number of principal components.

2.2 Deep Learning based Approaches

To overcome the difficulties explained above faced in classical machine learning based approaches, deep learning has attracted immense attention in medical imaging. Deep learning is a promising area of machine learning that uses raw data to extract features from neuroimages [22]. It enables computational models consisting of many layers of processing to extract various representations of data with abstraction at multiple levels [19]. In contrast with traditional ML approaches, DL enables the low-to-high level latent feature representations to be extracted automatically. In essence, it can be said that the deep learning methods are less dependent on image preprocessing techniques and they are less reliant on complicated procedures such as feature selection. This makes a relatively more objective approach and less bias-prone approach [19]. It finds complex patterns in big data sets using the backpropagation algorithm to point at how a computational machine should vary its internal parameters used to compute the differentiation in different layers.

As a matter of observation, one of the most important algorithm in available deep learning models is convolutional neural network (CNN) [23]. It is a category of artificial neural networks that has become increasingly popular in the field of computer vision, as it is attracting attention in various disciplines, including radiology [24]. With the help of various building blocks, such as convolution layers, pooling layers and fully connected layers, the CNN automatically as well as adaptively learns features and their spatial hierarchies through backpropagation. Many recent studies have advanced the idea of assisting the diagnosis of AD with the help of CNNs [25-40].

In [25], the study suggested a strategy of augmenting data to the specified regions of sMRI scans to overcome the limited data regarding the needs of CNN. They also advanced a 2-D+ϵ approach, in which training and classification is carried out using only a limited number of consecutive slices.

In [26], the study developed a method of transfer learning which works to train models from sMRI to Diffusion Tensor Imaging in a cross-modal fashion. Pre-trained models on structural MRI with domain-depended data augmentation were used as initialization of network parameters so as to train them on Mean Diffusive Data. The proposed method increased in accuracy of prediction by reduction of the over-fitting as well as considerable increase in learning performance.

In [27], an effective and methodical architecture was proposed with a simple three-dimensional convolutional network. It consists of five convolutional layers for extracting features, and for the task of classification of AD, three fully-connected layers follow.

In [28], the study proposed to develop many 3D CNNs to learn a variety of features from localized images of brain, which were subsequently combined to construct features for resultant classification for AD.

[29] proposed to develop a convolutional neural network (CNN) with multiple levels to sequentially learn and combine the features with multiple modalities for the classification of the disease. The images used in executing this were MRI and PET images.

[30] proposed a deep convolutional neural network for the diagnosis of AD. It executed the proposed framework with multi-class classification using data analysis of MRI image data.

In [33], the study advanced a method of classification formulated with many cluster-dense convolutional neural networks (DenseNets). Learning of various local features of MRI brain images was carried out by the networks and subsequently the localized images were combined to classify AD.

In [37], DL models were trained using slices of MRI and later these were amalgamated to create a fixed model by employing different voting methods. The deep learning models were trained on MRI slices from the datasets. The variations in the data such as contrast, resolution, signal-to-noise is one of the major impediments in the wide usage of CNNs in clinical diagnostics.

2.3 Classification of AD with end-to-end CNNs

2.3.1 2D slice-level CNN

Many studies employed 2D convolutional neural networks and the input data constituted of 2D slices taken from a volume of 3D MRI scans [40]. One of the advantages of this approach is that it allows for using preexisting successful image classification networks such as ResNet [41] and VGGNet [42] by employing transfer learning techniques. Moreover, this also has increased training data since we can manage more slices from 3D images.

2.3.2 Classification based on 3D patch-level Convolutional Neural Network

Other studies centralized 3D patch-classification in their analysis in order to compensate for lack of 3D information in 2D slicing approach. Such frameworks used, input data consisting of 3D patches extracted from images in sets. Fundamentally, this makes a larger sample size possible since the number of patches is equal to the number of samples, instead of being number of subjects. Nevertheless, this possible benefit of this classification is not used in the surveyed papers since the CNNs were trained for each patch independently. Furthermore, patches' memory usage is lower, a crucial factor in limited computing resources. However memory usage is low only when same network is used for all patches.

2.3.3 ROI-based CNN

Most of the patches used in 3D patch-level methods are not always informative as they may consist of sections of the brain unaffected by the disease. This leaves the patch-level method's input data redundant at times. Methods based on regions of interest (ROI) resolve this issue by narrowing down on parts of the images which are known to contain useful information. This helps in reducing the complexity of the framework as lesser number of inputs is used in network training.

2.3.4 3D subject-level CNN

In recent times, with advancement in high-performance computational resources, 3D subject-level approach has taken up steam in recent studies. This approach classifies the images at the subject level and entire MRI is as a single source input. The benefit of this approach is that spatial information is integrated in the input data.

2.4 Summary

This chapter provided an overview of classical machine learning based approaches and novel deep learning approaches for the classification of Alzheimer's disease.

Machine Learning models using features from single modality often underperform. To overcome this problem, people often use multiple biomarkers and modalities to obtain more meaningful combinatorial measures. These multiple biomarkers and modalities may not be readily available for all patients, making it difficult to use multi-modality based models in production. Moreover, the

performance of ML models is severely curtailed by large number of variables or features and low number of observations. The redundant variables and accompanying experimental noise are reduced using various techniques of dimensionality reduction. This is an imperative of ML before training a model so that overfitting is avoided. Nevertheless, this dimensionality reduction may result in loss of data, faulty linear relationship being developed, both undesirable side effects. Moreover, we may not be certain about the required number of principal components.

To overcome this difficulty, deep learning has found immense attention in medical imaging. In contrast with traditional ML approaches, DL enables the low-to-high level latent feature representations to be extracted automatically. In essence, it can be said that the deep learning methods are less dependent on image preprocessing techniques and it is less reliant on complicated procedures such as feature selection.

CNN is one of the most robust algorithm in comparison to different DL models. It is a category of artificial neural networks that has become increasingly popular in the field of computer vision, as it is attracting attention in various disciplines, including radiology.

The problem with deep learning algorithms especially in the absence of sufficient amount of data is that it can learn undesired spurious correlations in training data and it can easily over-fit resulting in poor generalization.

To overcome individual shortcoming of machine learning and deep learning approaches, we have proposed a unique combined (Deep Learning + Machine Learning) approach to enable early diagnosis and classification of AD through the use of structural MRIs with the following primary contributions:

- Relying only on single most widely available modality i.e. sMRI
- A hybrid approach to overcome shortcomings of individual approaches
- Taking advantage of more information available (processed MRI Scan + Features extracted from these scans)
- Breaking spurious correlations in training data leading to better generalization
- Beating cutting edge multi modal machine learning algorithms in terms of performance by using only single modality
- Beating state of the art Deep Learning models with better performance on validation data

CHAPTER 3.

PROPOSED METHODOLOGY

This chapter describes the details of proposed method for automatic classification of Alzheimer's Disease. We represent the main components of our automatic Alzheimer's Disease classification pipeline; data selection, preprocessing of MRI data, classification models, evaluation metrics and pipeline implementation details.

3.1 Workflow

We present a novel hybrid (Deep Learning + classical Machine Learning) approach where we take the benefit of both classical machine learning and novel deep learning techniques. We first preprocess an MRI and then compute the region-based features such as gray matter volume, white matter volume and CSF volume after performing segmentation/parcellation on MRI given an atlas template. In classical machine learning techniques, this kind of voxel based, surface based or region-based features are directly passed to a classifier after applying feature selection and dimensionality reduction techniques. Whereas in deep learning techniques, a pre-processed raw MRI is passed to a convolutional neural network which auto extracts the features on its own and classification layer uses these features to classify the MRI. We on the other hand, combined these two techniques where we fuse the features extracted using classical machine learning techniques with features extracted by a convolutional neural network. These combines/fused features are then passed to a classifier which gives the final output class/category. A generalized flowchart of our methodology is given in Figure 3.

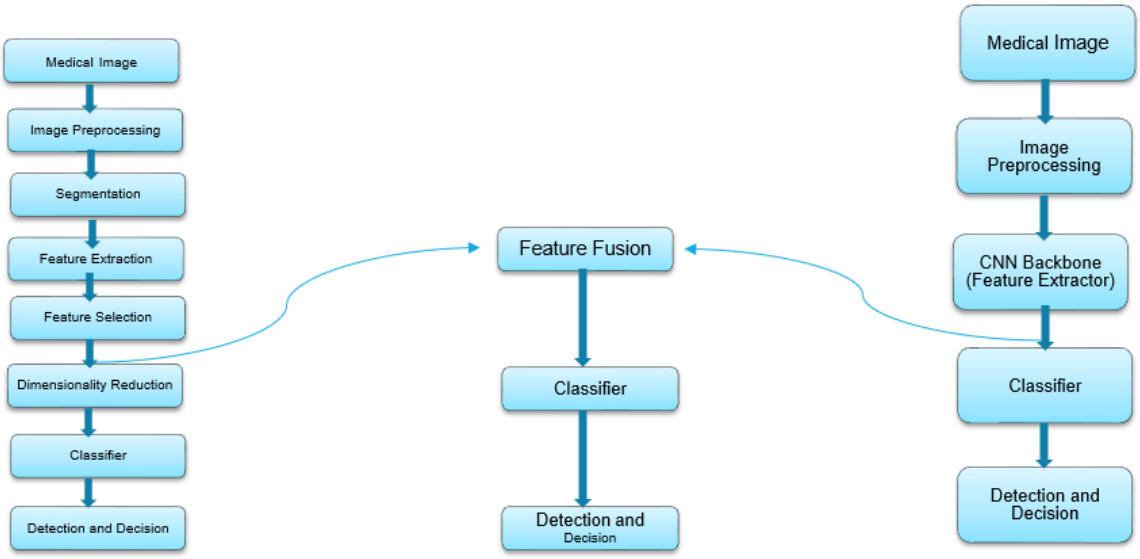

Figure 3 - Generalized flow of the proposed methodology

3.2 Dataset

The data used in our study is from Alzheimer's Disease Neuroimaging Initiative (ADNI) study. "The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD)."

We used ADNI1 screening data which comprises 1075 samples for whom relevant images were available for at least one visit. Four diagnosis groups were considered:

- **CN:** these were subjects who were diagnosed to be completely normal and they remained normal in the subsequent checkups.
- **AD:** these were the subjects initially diagnosed with AD and remained stable in the subsequent follow ups.
- **pMCI:** these are the sessions of patients initially having either MCI, EMCI or LMCI and advanced to full disease in the subsequent 3 years after the first visit.

- **sMCI:** sessions of patients found to have MCI, EMCI and LMCI and did not develop to AD in the next 36 months after the first visit.

Those AD and CN patients whose labels shifted over the passage of time were not included. In addition to that, MCI patients whose labels changed for two or more times were also not included. This exclusion was performed under the assumption that diagnosis of such subjects was not as reliable. Obviously, all pMCI and sMCI classified groups are incorporated under the label of MCI. Since some of the MCI subjects were not able to revert to AD but were not tracked for enough time to check whether they are sMCI, the reverse to the above-mentioned exclusions is false. 32 CN samples which converted from normal to MCI or AD or had multiple conversion were excluded. 2 AD samples which converted from AD to MCI or CN or had multiple conversion were also excluded. Moreover, 31 MCI subjects which had multiple conversions or converted from MCI to normal were also excluded. 2 pMCI, 5 CN, and 7 samples of AD were completely removed since there was preprocessing failure for all these samples.

	Subjects	**Age**	**Gender**	**MMSE**
CN	285	74. 4 ± 5.8 [59.8, 89.6]	137 M / 148 F	29.1 ± 1.1 [24, 30]
sMCI	156	72.3 ± 7.4 [55.0, 88.4]	89 M / 67 F	28.0 ± 1.7 [23, 30]
pMCI	224	73.8 ± 6.9 [55.1, 88.3]	135 M / 89 F	26.9 ± 1.7 [23, 30]
AD	238	75.0 ± 7.8 [55.1, 90.9]	128 M / 110 F	23.2 ± 2.1 [18, 27]

Table 1- Summary of participants' demographics and mini-mental state examination (MMSE)

A sample of this dataset is shown in Figure 4 with three primary imaging planes. Axial plane is on top left, while Sagittal plane and Coronal plane is shown in bottom left and bottom right respectively.

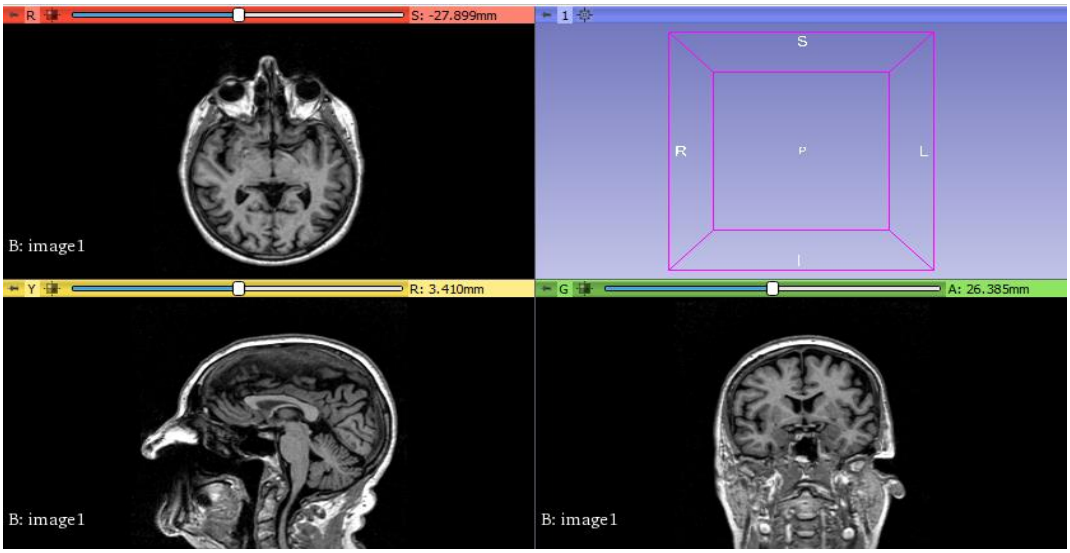

Figure 4 – Sample MRI with axial, sagittal and coronal planes

3.3 Preprocessing

The main objective of data preprocessing pipeline is basic preprocessing of MRI scans so they can be fed to convolutional neural network for automatic feature extraction. And later gray matter volumes of N parcellated/segmented regions are computed and treated as N unique features which are then fused with convolutional features and passed to fully connected and classification layers for classification of AD.

For MRI pre-processing, Computational Anatomy Toolbox (CAT), a matlab based extension to Statistical Parametric Mapping (SPM12) was used. CAT covers basic data pre-processing and diverse morphometric tools such as VBM, surface-based morphometry (SBM), deformation-based morphometry (DBM) and region-based morphometry (RBM). We used CAT for basic MRI pre-processing and whole brain parcellation/segmentation into N cortical and subcortical regions given an atlas template.

1.5-T sMRI T1-weighted images taken from the cited website were used. All downloaded scans were 256 × 256 × 168 resolution with 1mm spacing between each voxel. An overview of registration of MRI with MNI152-T1 and segmentation/parcellation into 246 regions given Brainnetome atlas template is given in Figure 5.

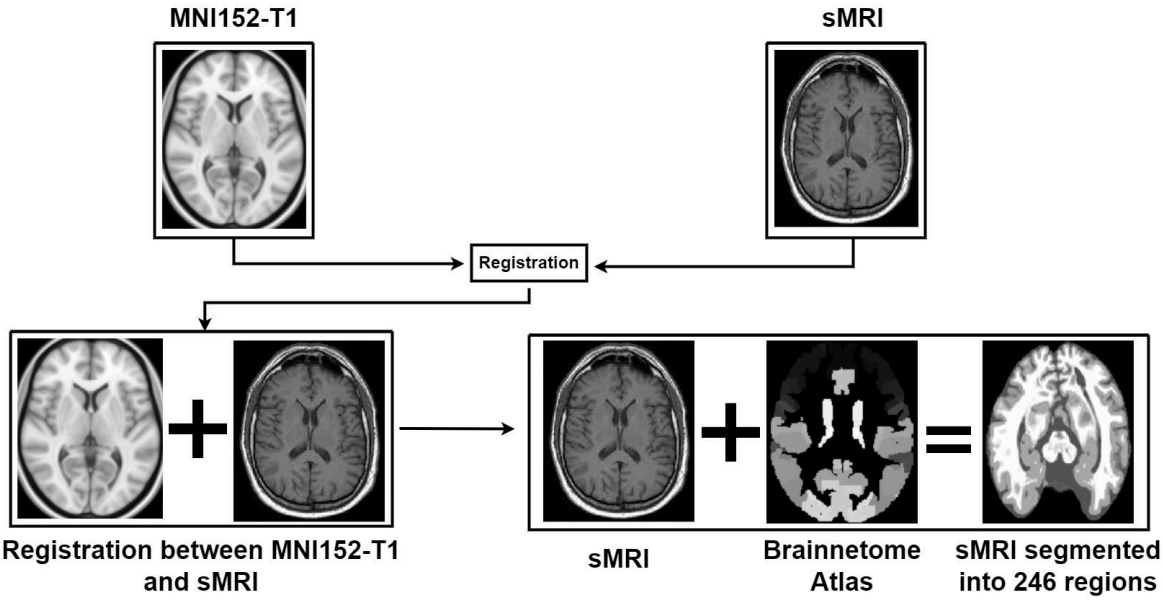

Figure 5 – Overview of registration of MRI with MNI152-T1 and segmentation into 246 regions given Brainnetome atlas template

Preprocessing involved SANLM denoising, N4 bias field correction, skull striping, registration to MNI152 T1-weighted standard image, local intensity correction and anatomical segmentation or parcellation of whole-brain into 246 anatomic regions using 1-mm Brainnetome atlas template which is already segmented into 246 regions, 210 cortical and 36 subcortical regions. An overview of initial voxel-based processing, refined voxel based processing and classification pipeline is given in Figure 6.

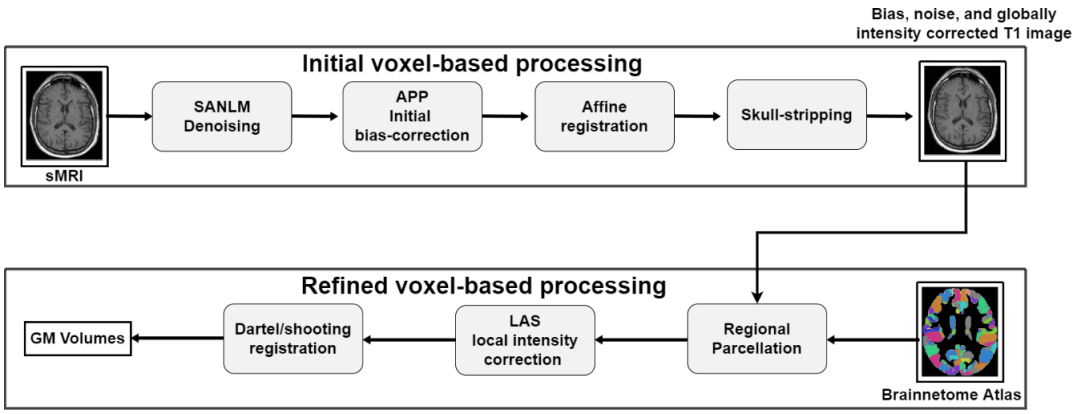

Figure 6 – Overview of initial voxel-based processing, refined voxel based processing and classification pipeline

3.4 Proposed Architecture

We considered the 3D subject-level CNN approach. In deep learning scenario, one of the challenges is finding the "optimal" model with a global minima, including the architecture hyper parameters (e.g. number of layers, filter size/receptive field, dropout, batch normalization etc.) and the training hyper parameters (e.g. learning rate, weight decay, momentum etc.). We have looked into the constructs already proposed in the literature and used the following approach for architecture and hyper parameters selection. The process began by overfitting the model which was heavy since it had more convolutional layers and their feature maps. We iteratively repeated the following operations:

1. Reduce the number of convolutional layers iteratively in block 1 and block 3 keeping the same feature map setting until there is significant drop in validation accuracy. Lock the architecture with as fewer layers as possible with minimal drop in accuracy.
2. Reduce the number of feature maps of remaining convolutional layers in block 1 and block 3 until there is significant drop in validation accuracy.
3. Perform hyper parameters tuning experiments on the selected architecture from above two steps.

Details of experiments performed for architecture selection with number of layers, number of feature maps and resulting train, validation and test accuracy is given in Table 4. In stage 1, we reduced the number of conv layers in Block 1 and Block 3 from five to two with a 0.06% drop in test accuracy and 1.22% drop in validation accuracy. Train accuracy dropped from 99.31% to 97.82% which indicates that initial larger network was overfitting on train data. Its train accuracy dropped by roughly 1.5% with a mere 0.06% dropped in test accuracy indicating same generalization capability even after significantly reducing the number of conv layers. During stage 2, we reduced the number of feature maps or channels of Convolutional layers in Block 1 and Block 3 with final 97.84% test accuracy. Test accuracy drooped only by 0.01% indicating the network still has the capacity to learn given the complexity of training data and it still generalizes well.

Block diagram of final architecture is given in Figure 7. There are 2 convolutional and 2 pooling layers in Block 1, and Block 3, and 6 convolutional layers in Block 2. Block 1, 2 and 3 extract features

from a raw preprocessed MRI. They can also be called feature extractors or backbone of our convolutional neural network. Gray matter volumes are fused with features extracted by the backbone and passed to fully connected layers for classification.

More fine-grained details of our network with each layer type, its output shape, number of learnable parameters, GMacs and total network size in Megabits (MBs) is given in Table 2. First few layers in each block contribute a greater number of trainable parameters and more computational complexity. Only first fully connected layer contributes 37.514% weights of the complete network. Fully connected layers are often very expensive both in number of trainable parameters and computational complexity especially in case of image data. That's why there are only few fully connected layers in practice in a convolutional neural network and that too at the end of the network where feature maps' spatial dimensions are small enough so that fully connected layers are not very expensive. We followed the same approach and used only few fully connected layers after feature extractor or backbone.

- Computational complexity: 76.66 GMac
- Total params: 874,173
- Trainable params: 874,173
- Non-trainable params: 0
- Input size (MB): 27.12
- Forward/backward pass size (MB): 1323.01
- Params size (MB): 3.33
- Estimated Total Size (MB): 1353.46

In addition to the list of experiments mentioned above, we also trained the DL model by passing preprocessed sMRI only as input to see the learning capacity of the proposed CNN. In this case, we removed the two fully connected layers which were taking GM volumes as input and also removed the last concatenation operation. Three set of experiments were conducted by varying the number of conv layers and their parameters, results are given in chapter 4.

We also trained the Support Vector Machine (SVM) classifier on GM volumes only with RBF and polynomial kernels and by exploiting the search space by varying values of gamma and C.

Furthermore, Random Forest (RF) classifier was also trained on GM volumes by varying the max depth and minimum samples split parameters. Results of both SVM and RF classifiers are given in chapter 4.

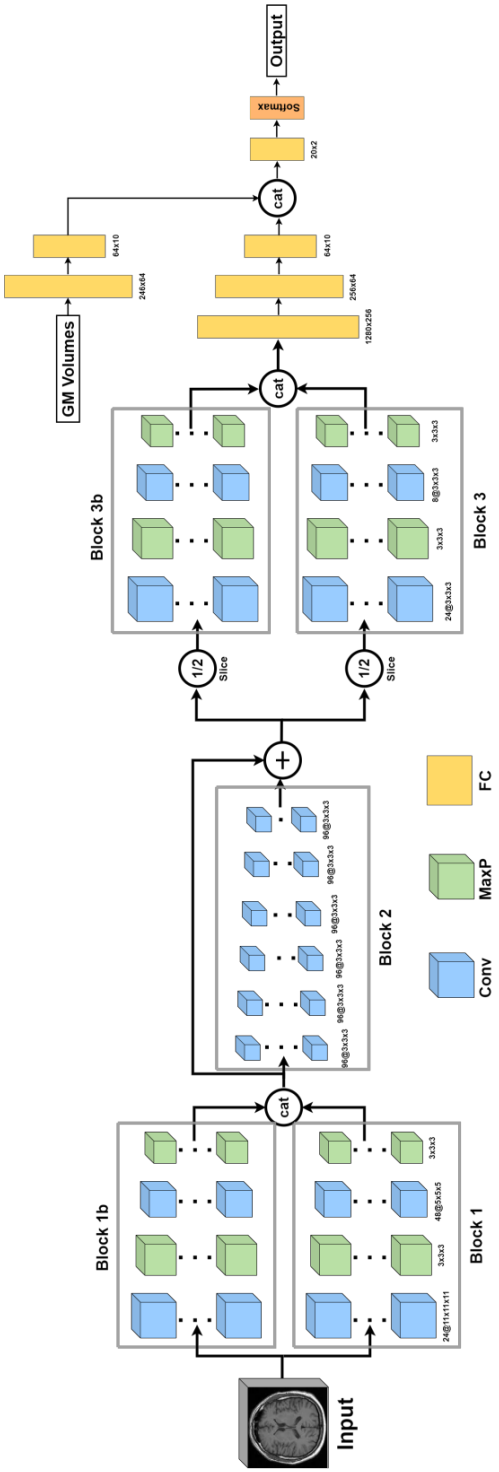

Figure 7 – Block diagram of our convolutional neural network

	Layer (type)	Output Shape	Param #	Param %	GMacs	Computational complexity: 76.66 GMac Total params: 874,173 Trainable params: 874,173 Non-trainable params: 0 ---------- ---------- Input size (MB): 27.12 Forward/backward pass size (MB): 1323.01 Params size (MB): 3.33 Estimated Total Size (MB): 1353.46 ---------- ----------
Block 1	Conv3d-1	[-1, 24, 87, 105, 87]	31,968	3.657	25.406	
	BatchNorm3d-2	[-1, 24, 87, 105, 87]	48	0.005	0.038	
	ELU-3	[-1, 24, 87, 105, 87]	0	0.000	0.019	
	MaxPool3d-4	[-1, 24, 43, 52, 43]	0	0.000	0.019	
	Dropout3d-5	[-1, 24, 43, 52, 43]	0	0.000	0	
	Conv3d-7	[-1, 48, 41, 50, 41]	144,048	16.478	12.107	
	BatchNorm3d-8	[-1, 48, 41, 50, 41]	96	0.011	0.008	
	ELU-9	[-1, 48, 41, 50, 41]	0	0.000	0.004	
	MaxPool3d-10	[-1, 48, 20, 24, 20]	0	0.000	0.004	
	Dropout3d-11	[-1, 48, 20, 24, 20]	0	0.000	0	
Block 1b	Conv3d-13	[-1, 24, 87, 105, 87]	31,968	3.657	25.406	
	BatchNorm3d-14	[-1, 24, 87, 105, 87]	48	0.005	0.038	
	ELU-15	[-1, 24, 87, 105, 87]	0	0.000	0.019	
	MaxPool3d-16	[-1, 24, 43, 52, 43]	0	0.000	0.019	
	Dropout3d-17	[-1, 24, 43, 52, 43]	0	0.000	0	
	Conv3d-19	[-1, 48, 41, 50, 41]	144,048	16.478	12.107	
	BatchNorm3d-20	[-1, 48, 41, 50, 41]	96	0.011	0.008	
	ELU-21	[-1, 48, 41, 50, 41]	0	0.000	0.004	

	MaxPool3d-22	[-1, 48, 20, 24, 20]	0	0.000	0.004	
	Dropout3d-23	[-1, 48, 20, 24, 20]	0	0.000	0	
Block 2	Conv3d-25	[-1, 96, 20, 24, 20]	960	0.110	0.009	
	Conv3d-26	[-1, 96, 20, 24, 20]	27,744	3.174	0.266	
	BatchNorm3d-27	[-1, 96, 20, 24, 20]	192	0.022	0.002	
	ELU-28	[-1, 96, 20, 24, 20]	0	0.000	0.001	
	Dropout3d-29	[-1, 96, 20, 24, 20]	0	0.000	0	
	Conv3d-31	[-1, 96, 20, 24, 20]	960	0.110	0.009	
	Conv3d-32	[-1, 96, 20, 24, 20]	27,744	3.174	0.266	
	BatchNorm3d-33	[-1, 96, 20, 24, 20]	192	0.022	0.002	
	ELU-34	[-1, 96, 20, 24, 20]	0	0.000	0.001	
	Dropout3d-35	[-1, 96, 20, 24, 20]	0	0.000	0	
	Conv3d-37	[-1, 96, 20, 24, 20]	960	0.110	0.009	
	Conv3d-38	[-1, 96, 20, 24, 20]	27,744	3.174	0.266	
	BatchNorm3d-39	[-1, 96, 20, 24, 20]	192	0.022	0.002	
	ELU-40	[-1, 96, 20, 24, 20]	0	0.000	0.001	
	Dropout3d-41	[-1, 96, 20, 24, 20]	0	0.000	0	
Block 3	Conv3d-43	[-1, 24, 20, 24, 20]	31,128	3.561	0.299	
	BatchNorm3d-44	[-1, 24, 20, 24, 20]	48	0.005	0	
	ELU-45	[-1, 24, 20, 24, 20]	0	0.000	0	

	MaxPool3d-46	[-1, 24, 9, 11, 9]	0	0.000	0	
	Dropout3d-47	[-1, 24, 9, 11, 9]	0	0.000	0	
	Conv3d-49	[-1, 8, 9, 11, 9]	5,192	0.594	0.005	
	BatchNorm3d-50	[-1, 8, 9, 11, 9]	16	0.002	0	
	ELU-51	[-1, 8, 9, 11, 9]	0	0.000	0	
	Dropout3d-53	[-1, 8, 4, 5, 4]	0	0.000	0	
Block 3b	Conv3d-55	[-1, 24, 20, 24, 20]	31,128	3.561	0.299	
	BatchNorm3d-56	[-1, 24, 20, 24, 20]	48	0.005	0	
	ELU-57	[-1, 24, 20, 24, 20]	0	0.000	0	
	MaxPool3d-58	[-1, 24, 9, 11, 9]	0	0.000	0	
	Dropout3d-59	[-1, 24, 9, 11, 9]	0	0.000	0	
	Conv3d-61	[-1, 8, 9, 11, 9]	5,192	0.594	0.005	
	BatchNorm3d-62	[-1, 8, 9, 11, 9]	16	0.002	0	
	ELU-63	[-1, 8, 9, 11, 9]	0	0.000	0	
	MaxPool3d-64	[-1, 8, 4, 5, 4]	0	0.000	0	
	Dropout3d-65	[-1, 8, 4, 5, 4]	0	0.000	0	
FC Block 1	Flatten-67	[-1, 1280]	0	0.000	0	
	Linear-68	[-1, 256]	327,936	37.514	0	
	BatchNorm1d-69	[-1, 256]	512	0.059	0	

	ELU-70	[-1, 256]	0	0.000	0	
	Dropout-71	[-1, 256]	0	0.000	0	
	Linear-73	[-1, 64]	16,448	1.882	0	
	BatchNorm1d-74	[-1, 64]	128	0.015	0	
	ELU-75	[-1, 64]	0	0.000	0	
	Dropout-76	[-1, 64]	0	0.000	0	
	Linear-78	[-1, 10]	650	0.074	0	
	BatchNorm1d-79	[-1, 10]	20	0.002	0	
	ELU-80	[-1, 10]	0	0.000	0	
	Dropout-81	[-1, 10]	0	0.000	0	
FC Block 2	Linear-83	[-1, 64]	15,808	1.808	0	
	BatchNorm1d-84	[-1, 64]	128	0.015	0	
	ELU-85	[-1, 64]	0	0.000	0	
	Dropout-86	[-1, 64]	0	0.000	0	
	Linear-88	[-1, 10]	650	0.074	0	
	BatchNorm1d-89	[-1, 10]	20	0.002	0	
	ELU-90	[-1, 10]	0	0.000	0	
	Dropout-91	[-1, 10]	0	0.000	0	
FC Block 3	Linear-93	[-1, 4]	84	0.010	0	
	BatchNorm1d-94	[-1, 4]	8	0.001	0	

	ELU-95	[-1, 4]	0	0.000	0	
	Dropout-96	[-1, 4]	0	0.000	0	
	FCBlock-97	[-1, 4]	0	0.000	0	
	Linear-98	[-1, 1]	5	0.001	0	
Classification Layer	Sigmoid-99	[-1, 1]	0	0.000	0	

Table 2 – Details of our network with layer type, its output shape, number of learnable parameters, and computational complexity in GMacs

We also made a comparison of our architecture with state of the art 2D image classification architectures in Table 3. It is important to note that even with a paradigm of 3D convolutional layers approach and with high dimensional 3D input data, our network contains only 0.874 M parameters, lower more than 100 times in magnitude compared to vgg variants and lower more than 10 times compared to resnet variants.

Model	Input Resolution	Params(M)	MACs(G)
alexnet	3x224x224	61.1	0.72
vgg11	3x224x225	132.86	7.63
vgg13	3x224x226	133.05	11.34
vgg16	3x224x227	138.36	15.5
vgg19	3x224x228	143.67	19.67
vgg11_bn	3x224x229	132.87	7.64
vgg13_bn	3x224x230	133.05	11.36

vgg16_bn	3x224x231	138.37	15.53
vgg19_bn	3x224x232	143.68	19.7
resnet18	3x224x233	11.69	1.82
resnet34	3x224x234	21.8	3.68
resnet50	3x224x235	25.56	4.12
resnet101	3x224x236	44.55	7.85
resnet152	3x224x237	60.19	11.58
squeezenet1_0	3x224x238	1.25	0.83
squeezenet1_1	3x224x239	1.24	0.36
densenet121	3x224x240	7.98	2.88
densenet169	3x224x241	14.15	3.42
densenet201	3x224x242	20.01	4.37
densenet161	3x224x243	28.68	7.82
inception_v3	3x224x244	27.16	2.85
This Study	1x181x217x217	0.874	76.66

Table 3 – Comparison of this study with commonly used 2d image classification architectures

3.5 Network Building Blocks

Convolutional Neural Network is a sequence or combination of layers. Using a differentiable function, each layer converts a group of activations to another. We used five main types of layers; convolutional layer, batch norm layer, pooling layer, dropout layer and fully connected layer to build our network. These layers are stacked in a special way along with skip connections to form the final architecture. We used ELU non linearity between the convolutional and pooling layers and Sigmoid before the output layer. A detailed explanation of these layers is given below.

Input: With dimensions of [1x181x217x181], input holds the raw voxel values of the pre-processed MRI. A voxel is a 3-Dimensional unit with a single value whereas a pixel is a 2-Dimensional unit in digital photographs. In this case, a single voxel is of size 1mmx1mmx1mm.

Convolutional Layer: Output of the neurons which are connected to the local region of neurons in previous layer is computed using convolutional layer. An output of a neuron is the dot product between these connected neurons with a filter of weights. There can be N unique filters, resulting in N output feature maps.

Batch norm Layer: A batch norm layer normalizes the activations using batch/mini-batch statistics. It helps an early conversion of a convolutional neural network with better generalization. By scaling/re-scaling the activations to a normal distribution, it smooths out the objective function which in turn improves the performance.

Pooling Layer: Pooling layer down samples the activations along the spatial dimensions with modes of choice including min, max and average pooling etc. Pooling reduces the spatial dimensions of feature maps, thus in result, reduces the numbers of parameters to learn, resulting in low number of computations performed. It also makes the features more robust.

ELU: Exponential Linear Unit that tends to converse cost/loss to zero faster and produce more accurate results. It is very similar to RELU except for the negative numbers. Any non-linear activation function helps to approximate functions which does not follow the linearity. Exponential Linear Unit's mathematical equation and its derivative are given in equation 1 and 2 respectively.

$$f(x) = \begin{cases} z, & z \geq 0 \\ \alpha(e^z - 1), & z < 0 \end{cases} \quad (1)$$

$$f'(x) = \begin{cases} 1, & z \geq 0 \\ \alpha e^z, & z < 0 \end{cases} \quad (2)$$

Dropout Layer: Deep neural networks are likely to quickly overfit the training data when there are few examples in the training set, thus resulting in poor generalization. Dropout layer adds regularization by randomly dropping out the neurons in a feature map. It offers a very computationally cheap and remarkably effective regularization method to reduce overfitting and improves generalizations performance of deep learning models.

Fully Connected Layer: As the name implies, each neuron of a fully connected layer is connected to every neuron of previous layer thus making it fully connected. Fully connected layers are often used at the end of convolutional neural networks and before the classification layer where they compute the class scores. Those class scores are then fed to the classification layer (softmax, sigmoid) to get the output class/category.

Sigmoid: Sigmoid is another non-linear activation function which maps the activations to the range of 0 to 1. It follows the equation given in 3.

$$Sigmoid(z) = \frac{1}{1 + e^z} \quad (3)$$

It is usually used for models where on has to predict the probability as an output as in our case. We have used sigmoid as final/last layer where it predicts the output probabilities and if the predicted

probability is greater than a given threshold, output class is a, otherwise it is b in case of binary classification.

The building blocks of our network will now be explained in detail below. Although our input is 4-Dimensional and we have used 3-Dimensional convolutional and pooling layers, but we will stick to 2-Dimensional convolutional and pooling layers in this section for better understanding of underlying operations. The underlying operations are same in 3-D convolutional and pooling layers as in 2-D except that an extra dimension is added.

3.5.1 Convolutional Layer

The parameters of convolutional layer contain a set of learnable filters. Let's assume, our input has three (RGB) channels with width w and height h. Every filter when convolved with this input will produce a 2-D activation map and K such activation maps are layered to produce the volume of the output.

Let's assume, a filter in first conv layer has a filter size of 7x7x3, 3 because there are three color channels in input. When there is forward pass, we move each filter across spatial dimensions (width and height) of the input and compute the dot product between the entries of the filter and input pixels at that position. The dot product of the filter with single input position/window will produce a single value in output volume. That filter when slided over width and height of input image will produce W_2 x H_2 outputs with W_2 as width and H_2 as height of output volume. And if there are K filters in a conv layer, output volume will be of dimensions K x W_2 x H_2.

Naturally, the network shall learn filters which activates when they see the following: some type of visual features such as an edge of an orientation or a blotch of a color in lower-level convolutional layer and finally entire shape/texture of a chair or cat in higher level convolutional layers.

In dealing with inputs having high number of dimensions, it is not pragmatic to connect each neuron with all neurons in previous layer. Rather a neuron is connected to only a region of neurons in previous layer. This region is local in spatial dimensions and extends to full in depth. The spatial extent of this local connection is called the receptive field of the neuron or filter size in other words. The receptive field is a hyper parameter which needs to be tuned.

We have explained the connectivity of each neuron to the neurons of previous layer, but we have not explained that how the size of output volume is controlled. Four hyper parameters named stride, padding, filter size and depth controls the size of output volume.

1. **Depth**: It is hyper parameter which corresponds to number of filters we would like to use and controls the depth of output volume. Each filter learns to look for something different in input.
2. **Filter Size**: Filter size controls the spatial dimensions of output volume. Small filter sizes focus on more fine-grained features whereas large filter sizes focus on global features. Odd numbered and square filter sizes such as 1x1, 3x3 and 5x5 etc. are preferred as they symmetrically divide the previous layer pixels around the output value. And absence of symmetry causes the distortions across the layers as in case of even numbered filters such as 2x2 and 4x4 etc. Moreover, 1x1 kernel is often employed to reduce the dimensions that target the reduction in number of paths in output volume with an objective to reduce to computational cost of following layers. Most common choice of filter size among deep learning practitioners is 3x3 which is not very compute intensive and looks for fine grained information in input. Some use 1x3 followed by 3x1 which is actually splitting down one convolutional layer to a series of convolutional layers which is quite cost friendly.
3. **Padding**: Often it is useful to cover the volume of the input with zeros around the ends. This padding gives a control of the spatial dimensions of output and sometimes it is used to preserve the spatial dimensions of the input, so that input and output are of exactly same size with same width and height.
4. **Stride**: We slide the filters with stride S over input volume. If stride is 1, we move filter 1 pixel at a time and if stride is 2, we move filter 2 pixels at a time and so on. Stride 1 and 2 are most common in convolutional layers whereas stride 2 is used to down sample the input volume which results in smaller output size.

We can compute the spatial dimensions of output volume as a function of filter size/receptive field F, stride S and amount of zero padding P with input volume size W. The corresponding equation is given in 4.

$$W_{out} = \frac{W - F + 2P}{S} + 1 \quad (4)$$

To summarize, the conv layer

- Accepts a volume of size $W_1 * H_1 * D_1$
- Needs four hyper parameters; number of filters K, filter size F, stride S and amount of zero padding P
- Produces an output volume of size $W_2 * H_2 * D_2$ where
  $W_2 = \frac{W_1 - F + 2P}{S} + 1$, $H_2 = \frac{H_1 - F + 2P}{S} + 1$ and $D_2 = K$
- Introduces $F * F * D_1$ parameters per filter a total of $F * F * D_1 * K$ parameters for one convolutional layer.

3.5.2 Pooling Layer

The periodical insertion of pooling layers between convolutional layers is a matter of common practice. It reduces the spatial dimensions of output volume which in turn reduces the number of trainable parameters. A polling filter slides over the spatial dimensions of input with stride S and filter size F and produces the output volume of size given by equation 5.

$$W_{out} = \frac{W - F}{S} + 1 \quad (5)$$

There are several types of pooling such as max, min and average pooling but max pooling is often used in practice with its proven better performance over average and min pooling. Min pooling is hardly used in practice in convolutional neural networks. A pooling layer with filter size 2x2 and

stride 2, down samples the input feature map by 2 along width and height whereas depth remains the same.

To summarize, a pooling layer

- Accepts two hyper parameters; filter size F and stride S
- Produces an output volume of size $\boldsymbol{W_2 * H_2 * D_2}$ where
  $\boldsymbol{W_2 = \frac{W_1 - F}{S} + 1}$, $\boldsymbol{H_2 = \frac{H_1 - F}{S} + 1}$ and $\boldsymbol{D_2 = K}$
- Introduces zero parameters since it computes a fixed function of the input.

3.5.3 Fully Connected Layer

As the name implies, this layer has full connection to all neurons in previous layer. In other words, every neuron in this layer has connections to all neurons in previous layer. Fully connected layers are often used at the end of convolutional neural networks and before the classification layer where they compute the class scores. Those class scores are then fed to the classification layer (softmax, sigmoid) to get the output class/category.

CHAPTER 4.

EXPERIMENTS AND RESULTS

This chapter provides the detailed explanation of series of experiments performed and their role in selection of final architecture. First, we explore the benefits of our proposed method and then analyze the proposed architecture through tracked evaluation metrics. Lastly, we compare our proposed method with state-of-the-art methods in classification of Alzheimer's disease.

4.1 Evaluation Metrics

Evaluation metrics play a critical role in evaluating the performance of a statistical or machine learning/ deep learning model. Important decisions such as data selection, model/architecture selection, hyper parameters tuning etc. are made by evaluating the model on validation data through tracked evaluation metrics. These metrics also convey the deeper insights of model behavior such as underfitting, overfitting and if it has the enough capacity to learn the complexity of data or if there is any need of more training data for better generalization. In classification problems, most popular evaluation metrics include but are not limited to Accuracy, Precision, Recall, F1 score and Area Under the Receiver Operating Characteristic (ROC) curve (AUC-ROC) etc. We have used all the above-mentioned metrics to evaluate our proposed architecture.

These evaluation metrics are inferred from a confusion metrics. For example, if we want to detect that if a person has Alzheimer's disease or not, the entries of confusion matrix can be defined as

- **True Positives (TP)**: These are the cases where we predicted 'yes' and they do have the Alzheimer's disease
- **True Negatives (TN)**: These are the cases where we predicted 'no' and the actually do not have the Alzheimer's disease
- **False Positives (FP)**: Cases where we predicted 'yes' but they do not have the disease
- **False Negatives (FN)**: These are the case where we predicted 'no' and they do have the disease

A confusion matrix is shown in Figure 8

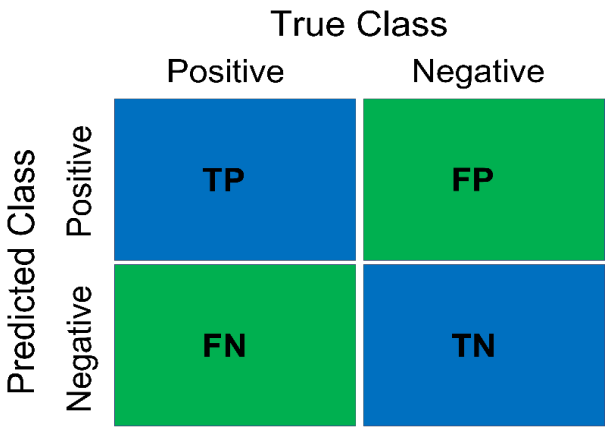

Figure 8 – Confusion matrix for binary classification

4.1.1 Accuracy

In classification, accuracy is a statistical measure of how often model predicts correct output or in other words, how often the classifier is correct. It is the proportion of true results among the total cases examined and given by equation 6.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN}} \quad \textbf{(6)}$$

4.1.2 Precision

It answers the question that what proportion of predicted positives is truly positive. In other words, it tells that how precise is the model with its predictions. Precision penalizes false positives and a model with 100% precision ensures that there are no cases when model's prediction was false positives. It is given by equation 7.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \quad \textbf{(7)}$$

4.1.3 Recall

It answers a different question that what proportion of actual positives is truly classified. Recall penalizes false negatives and a model with 100% recall ensures that model did not miss any positive. It is given by equation 8.

$$\boldsymbol{Recall = \frac{TP}{TP + FN}} \tag{8}$$

4.1.4 F1 Score

In those cases where precision and recall both are important at the same time, F_1 is often used. F_1 score is the harmonic mean of precision and recall and is given by equation 9. It uses harmonic mean instead of arithmetic mean because harmonic mean penalizes extreme values more thus ensuring that a good F_1 score will have a good precision and recall.

$$\boldsymbol{F_1 = \frac{2 * precision * recall}{precision + recall}} \tag{9}$$

A more generic F score is $\boldsymbol{F_\beta}$, where recall is β times more important than precision and β is a positive number. It is given by equation 10.

$$\boldsymbol{F_\beta = \frac{(1 + \beta)^2 * precision * recall}{(\beta^2 * precision) + recall}} \tag{10}$$

4.2 Training Strategy

The training of various ML and DL models are often limited by the factor of available computational facility. In addition, there is a need to expedite the time frame and enhance performance metrics described above at the same time. This implies a trade-off when there's limited computational power at disposal to execute massive training datasets. Therefore, our training strategy was chalked keeping these key considerations in account.

The cornerstone of managing the optimized training method are two key schedulers, namely learning rate scheduler (LR Scheduler) and Warmup Scheduler. Our learning rate scheduler has been optimized such that LR is fast enough to reach to train the model at an optimum rate without missing the global minima of the data set. The positive aspect of the scheduler being faster after the initialization and then gradually slowing down is that in the initial zone, the probability of it missing the global minima is negligible. Therefore, the model can move through the data set quickly. However, in the step-wise reduction in the rate, we ensure that global minima are not missed. Conversely, low learning rate in the initial phase may result in model being stuck at a local minima without surveying the entire range of data.

4.3 Results and Evaluation

Table 4 enlists the experiments performed and respective results. The experiments delineated in the table have been performed with NC versus AD criteria. The experiments are categorized in two stages. In the first stage, six experiments are carried out. The experiments 7 till 9 constitute the second stage of experimentation.

The experiments varied in employing convolutional layers in the architectural blocks, namely block1 and block 3. The effect of variation in number of layers on the test, validation and training accuracies was observed to optimize the number of layers and their relative combination. The columns labelled “Block 1” and “Block 3” represent the variation across all experiments.

As a matter of observation, the training accuracy increased with increase in number of convolutional layers, peaking at 99.31% in experiment 1. Similar trend was observed with validation and test accuracies. The 9th experiment was the final experiment conducted. Its test accuracy stood at 91.84% in NC vs AD classification paradigm with a drop of only 0.01% compared to experiment 1. The training accuracy for the final experiment was computed at 97.54%, while the validation accuracy was 93.62%.

The 9th experiment was selected as the representative for further analysis. The results are described in the section below.

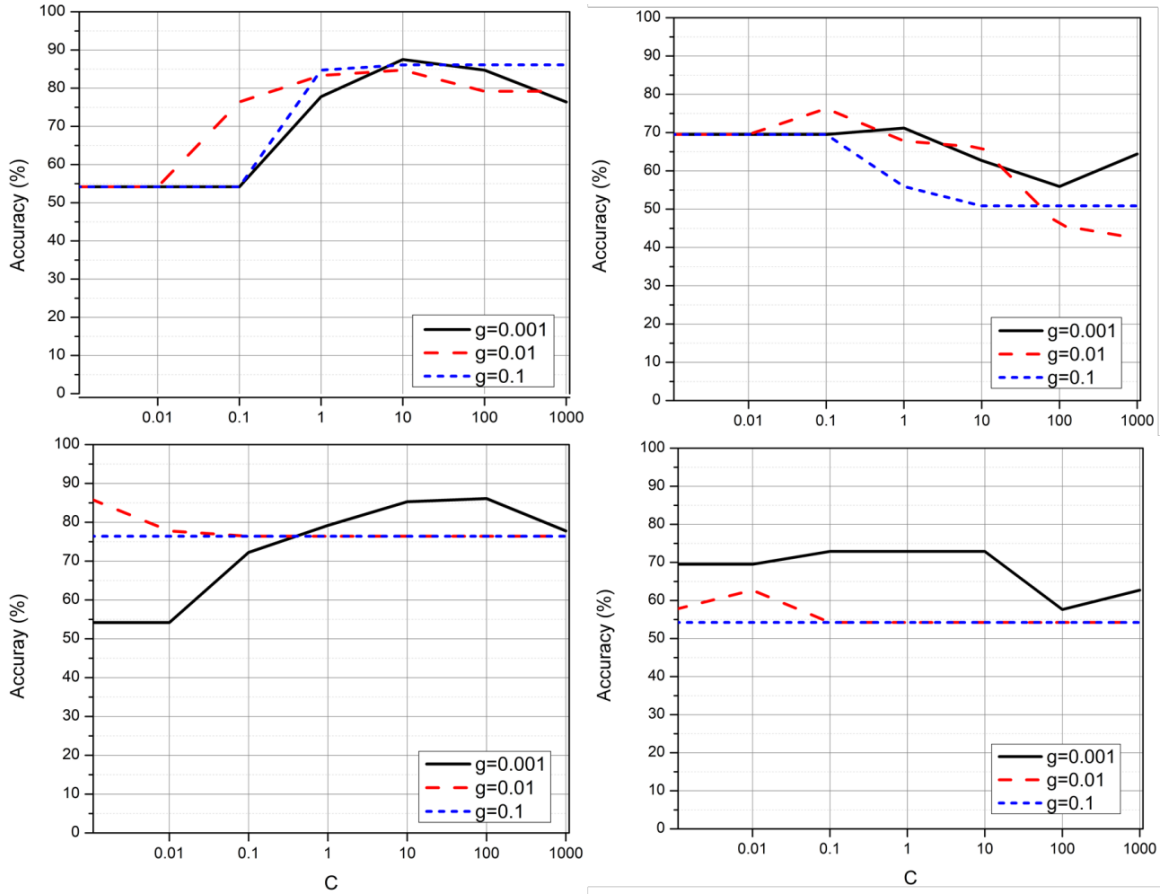

Figure 9 – Validation accuracy curves of SVM classifier with RBF and polynomial kernels for both CN vs AD and sMCI vs pMCI classification. (a) CN vs AD with RBF kernel (b) sMCI vs pMCI with RBF kernel (c) CN vs AD with polynomial kernel (d) sMCI vs pMCI with polynomial kernel

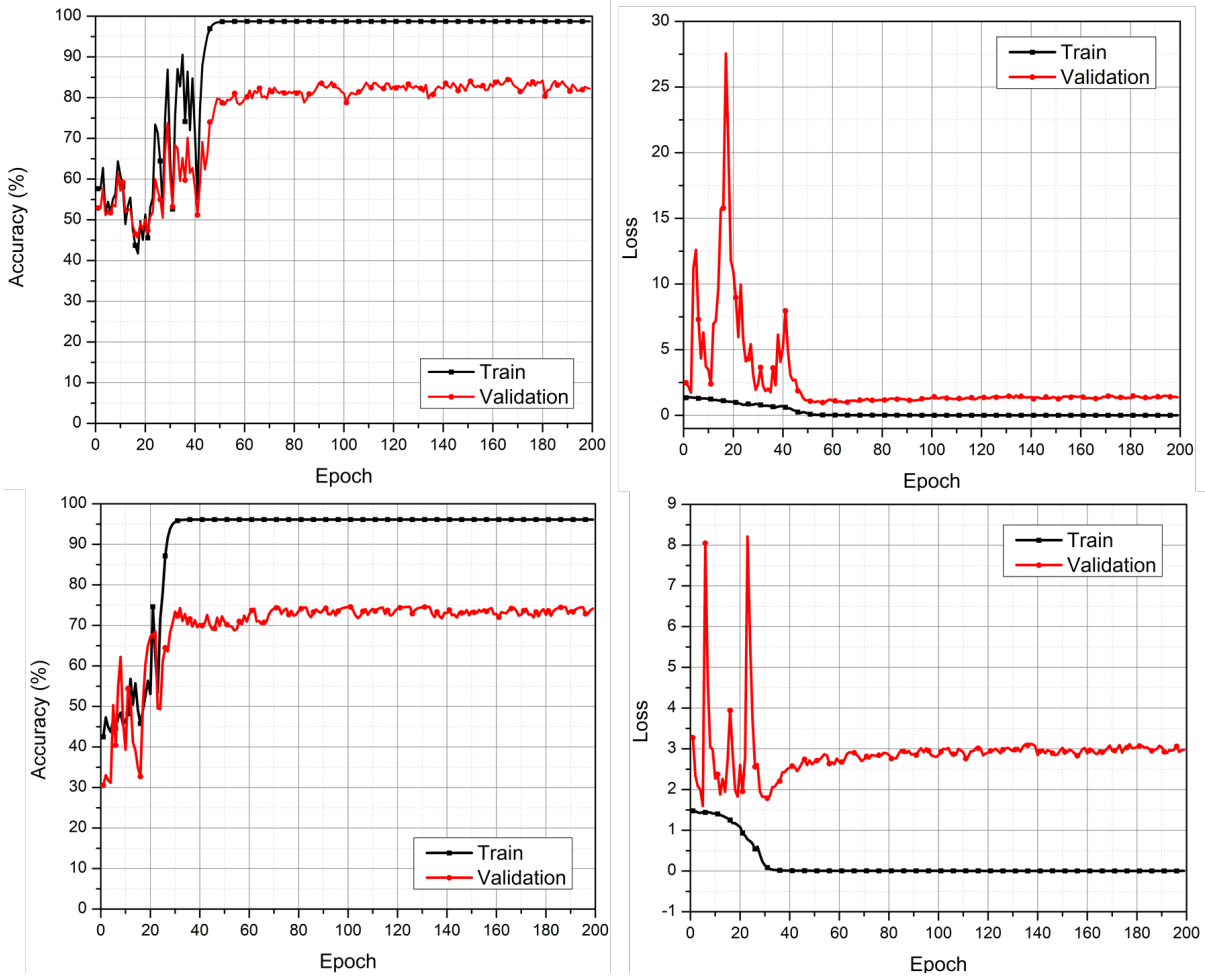

Figure 10- Validation accuracy and loss curves of DL model with sMRI input only. (a) CN vs AD accuracy (b) CN vs AD loss (c) sMCI vs pMCI accuracy (d) sMCI vs pMCI loss

Stage	Experiment no	Block1	Block3	Train Accuracy (%)	Validation Accuracy (%)	Test Accuracy (%)
1	1	Conv1: 24@11x11x11 Conv2: 48@5x5x5 Conv3: 96@5x5x5 Conv4: 192@3x3x3 Conv5: 384@3x3x3	Conv1: 384@3x3x3 Conv2: 192@3x3x3 Conv3: 96@3x3x3 Conv4: 48@3x3x3 Conv5: 24@3x3x3	99.31	94.65	91.91
	2	Conv1: 24@11x11x11 Conv2: 48@5x5x5 Conv3: 96@5x5x5 Conv4: 192@3x3x3	Conv1: 384@3x3x3 Conv2: 192@3x3x3 Conv3: 96@3x3x3 Conv4: 48@3x3x3 Conv5: 24@3x3x3	99.26	94.73	91.23
	3	Conv1: 24@11x11x11 Conv2: 48@5x5x5 Conv3: 96@5x5x5	Conv1: 384@3x3x3 Conv2: 192@3x3x3 Conv3: 96@3x3x3 Conv4: 48@3x3x3 Conv5: 24@3x3x3	99.09	94.04	91.86
	4	Conv1: 24@11x11x11 Conv2: 48@5x5x5	Conv1: 384@3x3x3	99.13	94.28	91.49

			Conv2: 192@3x3x3 Conv3: 96@3x3x3 Conv4: 48@3x3x3 Conv5: 24@3x3x3			
	5	Conv1: 24@11x11x11 Conv2: 48@5x5x5	Conv1: 384@3x3x3 Conv2: 192@3x3x3 Conv3: 96@3x3x3 Conv4: 48@3x3x3	98.87	93.86	91.9
	6	Conv1: 24@11x11x11 Conv2: 48@5x5x5	Conv1: 384@3x3x3 Conv2: 192@3x3x3	97.82	93.43	91.85
2	7	Conv1: 24@11x11x11 Conv2: 48@5x5x5	Conv1: 192@3x3x3 Conv2: 96@3x3x3	97.79	93.64	91.81
	8	Conv1: 24@11x11x11 Conv2: 48@5x5x5	Conv1: 96@3x3x3 Conv2: 48@3x3x3	97.71	93.55	91.84
	9	Conv1: 24@11x11x11 Conv2: 48@5x5x5	Conv1: 48@3x3x3 Conv2: 24@3x3x3	97.54	93.62	91.84

Table 4 – Results of the Experiments

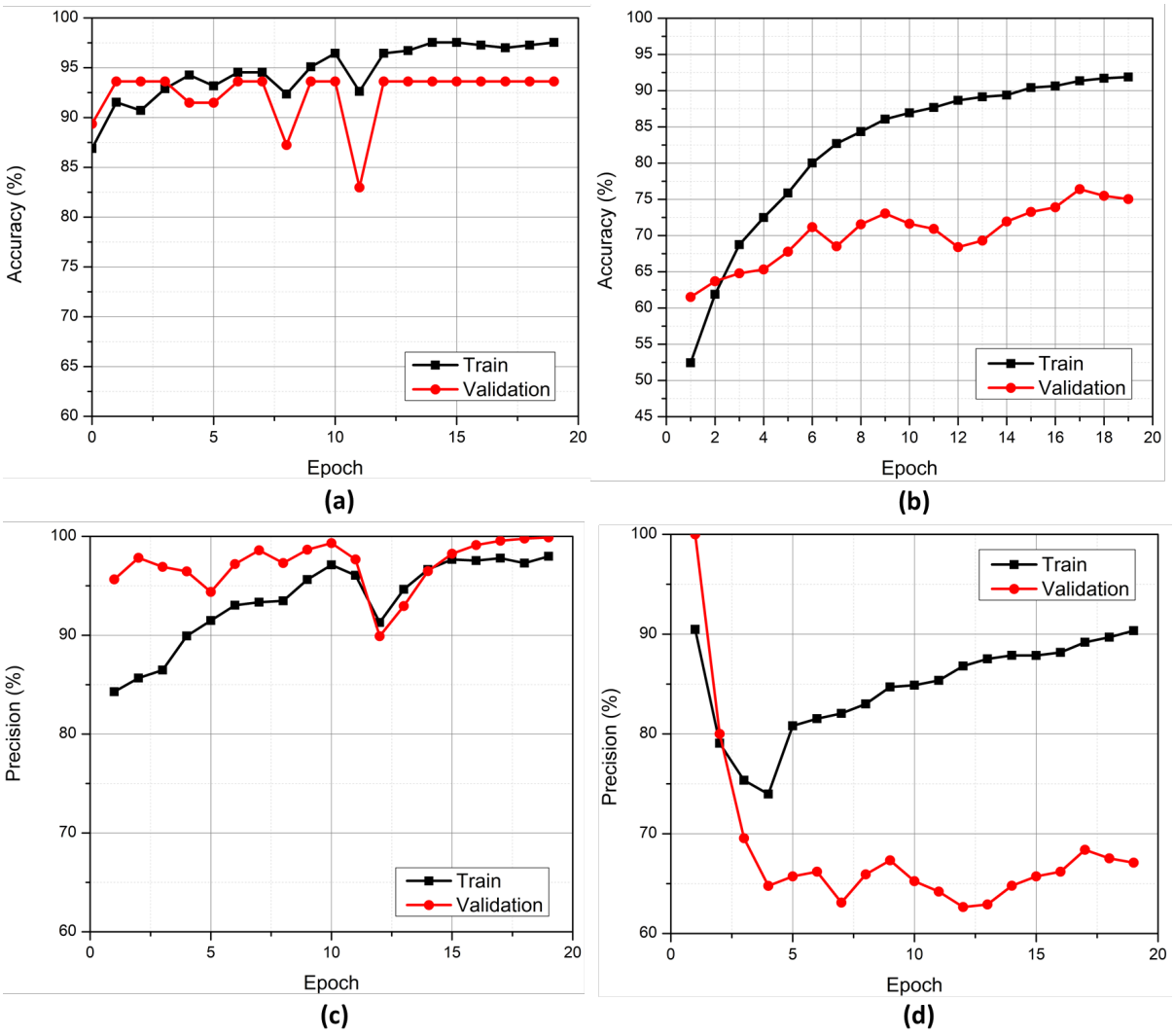

Figure 11 – Accuracy and precision curves of the proposed architecture on train and validation sets. (a) CN vs AD (b) sMCI vs pMCI (c) CN vs AD (d) sMCI vs pMCI

The experiment 11 was selected as representative of this study. The accuracy of the experiment plotted against the epochs is depicted in figure 11 (a). The training accuracy peaked close to 98 percent while the validation accuracy sustained the peak above 93 percent.

The sMCI vs pMCI paradigm showed similar success. Accuracy in this paradigm is shown in figure 11 (b). The train accuracy peaked around 92% and validation accuracy peaked around 75% in this classification paradigm.

The precision of the experiment, plotted against the epochs is shown in figure 11 (c). The training precision of the experiment peaked to 100 percent. The precision of the model on validation data is also close to 99 percent. Similarly, in the sMCI versus pMCI paradigm, training precision peaked at 90% as shown in figure 11 (d).

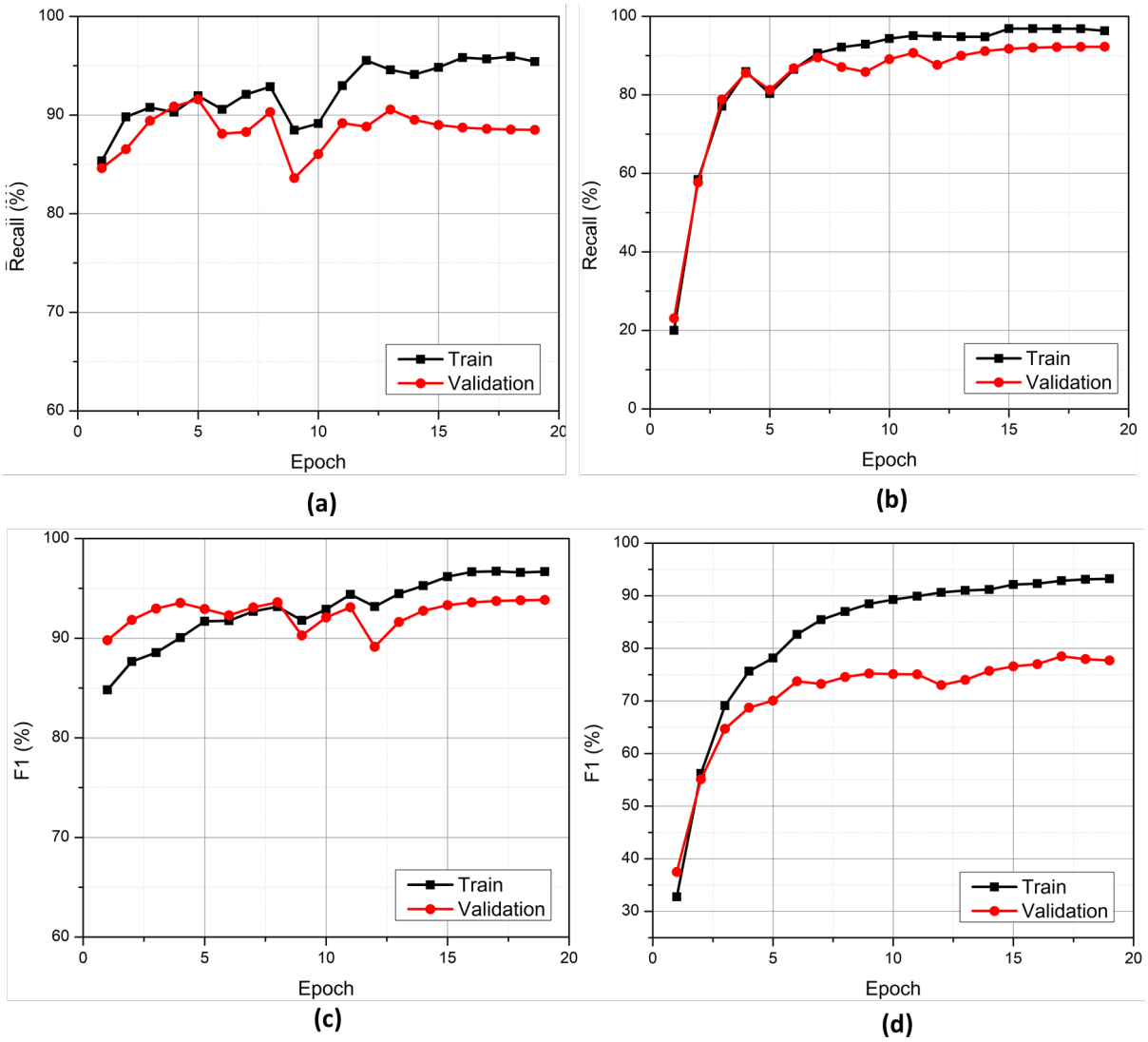

Figure 12 - Recall and F1 score curves of the proposed architecture on train and validation sets. (a) CN vs AD (b) sMCI vs pMCI (c) CN vs AD (d) sMCI vs pMCI

The graphs for binary classifications in both paradigms for its F1 score is depicted in figure 12 (c) and figure 12 (d) respectively. In the training data, the F1 score peaked beyond 96 percent, while the validation F1 score peaks at around 94 percent for AD vs CN classification. Recall curves for both AD vs CN and sMCI vs pMCI final trainings is shown in figure 12 (a) and 12 (b) respectively.

The exact values of results on test data are described in table 5 and 6. In the AD vs CN paradigm, the accuracy reached 91.84 percent. The precision attained by the model valued around 93.48. In addition to the above metrics, the F1 score of the experiment was also evaluated. The F1 score of the representative experiment reached 91.49. Moreover, the Recall attained 89.58 percent for the experiment selected. Furthermore, the specificity attained 94% in the AD vs CN paradigm for the given experiment.

In the sMCI vs pMCI paradigm, accuracy was 90.85%, precision reached around 80%, recall to 88.88% and F1 score reached around 84.21%, while specificity was at 70%.

NC vs AD Results on Test Set				
Accuracy	Precision	Recall	F1	Specificity
0.9184	0.9348	0.8958	0.9149	0.94

Table 5 – NC vs AD Results on the Test Data

sMCI vs pMCI Results on Test Set				
Accuracy	Precision	Recall	F1	Specificity
0.8085	0.8	0.8888	0.8421	0.7

Table 6 – sMCI vs pMCI Results on the Test Data

4.3.2 Comparison of the results with other studies

This section briefly touches upon the comparative assessment of the results of the present study with those in the previous state of the art studies. Table 7 summarizes the comparison drawn.

The results of the study were evaluated for two types of classification namely, AD vs CN and sMCI vs pMCI and are given in table 6. In general, the accuracies remained higher in binary classification of AD vs CN. The test accuracy of the AD vs CN classification was 91.84 whereas the sMCI vs pMCI classification's accuracy was 80.85. Similar trend was observed in validation and training accuracies of the model, with the latter binary classification performing significantly lower.

Compared to previous studies, our study performed better in both binary classifications in general. In the AD vs CN paradigm, our study gave the highest accuracy compared to all the previous studies considered in 3D subject level paradigm. The accuracy in our study was 91.84 whereas the closest accuracies reached 73 percent, with 3D subject level approach.

In the sMCI vs pMCI binary classification paradigm, our study's performance was again highest in 3D subject level paradigm. Our accuracy in this paradigm was 80.85 whereas the second best accuracy in the previous studies was of Wen et al. which reached 73 percent with 3D subject level approach.

Binary Classification Results Achieved in this Study			
Task	**Train Accuracy**	**Validation Accuracy**	**Test Accuracy**
AD vs NC	97.54	93.62	91.84
sMCI vs pMCI	92.07	73.9	80.85

Table 7 – Results of both binary classification paradigms

Comparison of the results with studies performing classification of AD using CNNs on anatomical MRI				
Study	**AD vs CN**	**sMCI vs pMCI**	**Approach**	**Data**
(U. Khatri and G.-R. Kwon, 2020)	ACC = 93.31	ACC = 83.38	SVM	MRI + CSF + APOE + MMSE
(Shmulev et al., 2018)	-	ACC = 0.62	3D subject-level	sMRI
(Wen et al., 2020)	BA = 0.73	BA = 0.73	3D subject-level	sMRI
(This study)	ACC = 91.84	ACC = 80.85	3D subject-level	sMRI

Table 8 – Comparison of the results with previous studies

CHAPTER 5.

CONCLUSION

This thesis set out with an aim to develop a model which could avoid the tedious bottlenecks of the classical ML and Deep Learning approaches, while simultaneously improving the performance in terms of precision and accuracy. We successfully developed a hybrid approach which involved fusing the features extracted by image processing pipeline with extracted by CNN backbone. The approach surpassed the accuracy of the reviewed literature in 3D subject level approach, to the best of our knowledge. The fused features from MRI images appear to be immensely potent to improve the accuracy and precision of the DL models considering the data constraints and computational limitations.

The thesis' main objective of improving accuracy (using only single modality) was met, both in AD versus CN and sMCI versus pMCI classification paradigms. As a future road map, we feel confident that hybrid approach can bring immense benefits for scientists working with limited computational facility and targeting better accuracy levels for the biomedical image classification. Future studies in this direction, with additional classification methods is also viable area to probe.

References

[1] R. S. Wilson, E. Segawa, P. A. Boyle, S. E. Anagnos, L. P. Hizel, and D. A. Bennett, "The natural history of cognitive decline in Alzheimer's disease," Psychol. Aging, vol. 27, no. 4, pp. 1008–1017, 2012.

[2]W. W. Barker et al., "Relative frequencies of Alzheimer disease, Lewy body, vascular and frontotemporal dementia, and hippocampal sclerosis in the State of Florida Brain Bank," Alzheimer Dis. Assoc. Disord., vol. 16, no. 4, pp. 203–212, 2002.

[3]M. Heron, "Deaths: Leading causes for 2014," *Natl. Vital Stat*. Rep., vol. 65, no. 5, pp. 1–96, 2016.

[4]Alzheimer's Association, "2017 Alzheimer's disease facts and figures," *Alzheimers. Dement*., vol. 13, no. 4, pp. 325–373, 2017.

[5]R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arrighi, "Forecasting the global burden of Alzheimer's disease," *Alzheimers. Dement*., vol. 3, no. 3, pp. 186–191, 2007.

[6]J. E. Galvin, "Prevention of Alzheimer's disease: Lessons learned and applied," J. Am. *Geriatr. Soc.*, vol. 65, no. 10, pp. 2128–2133, 2017.

[7] M. W. Schelke et al., "Mechanisms of risk reduction in the clinical practice of Alzheimer's disease prevention," *Front. Aging Neurosci*., vol. 10, p. 96, 2018

[8] S. Kloppel *et al.*, "Automatic classification of MR scans in Alzheimer's disease," *Brain*, vol. 131, no. 3, pp. 681–689, 2008

[9] X. Liu, D. Tosun, M. W. Weiner, N. Schuff, and Alzheimer's Disease Neuroimaging Initiative, "Locally linear embedding (LLE) for MRI based Alzheimer's disease classification," Neuroimage, vol. 83, pp. 148–157, 2013.

[10] T. Altaf, S. M. Anwar, N. Gul, M. N. Majeed, and M. Majid, "Multi-class Alzheimer's disease classification using image and clinical features," *Biomed. Signal Process. Control*, vol. 43, pp. 64–74, 2018.

[11] U. Khatri and G.-R. Kwon, “An efficient combination among sMRI, CSF, cognitive score, and APOE ε4 biomarkers for classification of AD and MCI using extreme learning machine,” *Comput. Intell. Neurosci.*, vol. 2020, p. 8015156, 2020.

[12] B. Naik, A. Mehta, and M. Shah, “Denouements of machine learning and multimodal diagnostic classification of Alzheimer’s disease,” *Vis Comput Ind Biomed Art*, vol. 3, no. 1, p. 26, 2020.

[13] A. H. Syaifullah, A. Shiino, H. Kitahara, R. Ito, M. Ishida, and K. Tanigaki, “Machine learning for diagnosis of AD and prediction of MCI progression from brain MRI using brain anatomical analysis using diffeomorphic deformation,” *Front. Neurol.*, vol. 11, p. 576029, 2020.

[14] A. Gupta and B. Kahali, “Machine learning-based cognitive impairment classification with optimal combination of neuropsychological tests,” *Alzheimers Dement. (N. Y.)*, vol. 6, no. 1, p. e12049, 2020.

[15] X. Liu, K. Chen, T. Wu, D. Weidman, F. Lure, and J. Li, “Use of multimodality imaging and artificial intelligence for diagnosis and prognosis of early stages of Alzheimer’s disease,” *Transl. Res.*, vol. 194, pp. 56–67, 2018.

[16] A. Ezzati *et al.*, “Optimizing machine learning methods to improve predictive models of Alzheimer’s disease,” *J. Alzheimers. Dis.*, vol. 71, no. 3, pp. 1027–1036, 2019.

[17] L. Sørensen and M. Nielsen, “Ensemble support vector machine classification of dementia using structural MRI and mini-mental state examination,” *J. Neurosci. Methods*, vol. 302, pp. 66–74, 2018.

[18] Y. Gupta, R. K. Lama, G.-R. Kwon, and Alzheimer’s Disease Neuroimaging Initiative, “Prediction and classification of Alzheimer’s disease based on combined features from Apolipoprotein-E genotype, cerebrospinal fluid, MR, and FDG-PET imaging biomarkers,” *Front. Comput. Neurosci.*, vol. 13, p. 72, 2019.

[19] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[20] J. Samper-González *et al.*, “Reproducible evaluation of classification methods in Alzheimer’s disease: Framework and application to MRI and PET data,” *Neuroimage*, vol. 183, pp. 504–521, 2018.

[21] B. C. Riedel *et al.*, “Uncovering biologically coherent peripheral signatures of health and risk for Alzheimer’s disease in the aging brain,” *Front. Aging Neurosci.*, vol. 10, p. 390, 2018.

[22] S. M. Plis *et al.*, “Deep learning for neuroimaging: a validation study,” *Front. Neurosci.*, vol. 8, p. 229, 2014.

[23] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, “Object recognition with gradient-based learning,” in *Shape, Contour and Grouping in Computer Vision*, Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 319–345.

[24] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, “Convolutional neural networks: an overview and application in radiology,” *Insights Imaging*, vol. 9, no. 4, pp. 611–629, 2018.

[25] K. Aderghal, M. Boissenin, J. Benois-Pineau, G. Catheline, and K. Afdel, “Classification of sMRI for AD diagnosis with convolutional neuronal networks: A pilot 2-D+ ε study on ADNI,” in *MultiMedia Modeling*, Cham: Springer International Publishing, 2017, pp. 690–701.

[26] K. Aderghal, A. Khvostikov, A. Krylov, J. Benois-Pineau, K. Afdel, and G. Catheline, “Classification of Alzheimer disease on imaging modalities with deep CNNs using cross-modal transfer learning,” in *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, 2018.

[27] K. Backstrom, M. Nazari, I. Y.-H. Gu, and A. S. Jakola, “An efficient 3D deep convolutional network for Alzheimer’s disease diagnosis using MR images,” in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018.

[28] D. Cheng, M. Liu, J. Fu, and Y. Wang, “Classification of MR brain images by combination of multi-CNNs for AD diagnosis,” in *Ninth International Conference on Digital Image Processing (ICDIP 2017)*, 2017.

[29] D. Cheng and M. Liu, "CNNs based multi-modality classification for AD diagnosis," in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2017.

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

[31] 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)*, 2017, pp. 835–838.

[32] F. Li, D. Cheng, and M. Liu, "Alzheimer's disease classification based on combination of multi-model convolutional networks," in *2017 IEEE International Conference on Imaging Systems and Techniques (IST)*, 2017, pp. 1–5.

[33] F. Li, M. Liu, and Alzheimer's Disease Neuroimaging Initiative, "Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks," *Comput. Med. Imaging Graph.*, vol. 70, pp. 101–110, 2018.

[34] C. Lian, M. Liu, J. Zhang, and D. Shen, "Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 880–893, 2020.

[35] J. Liu *et al.*, "Applications of deep learning to MRI images: A survey," *Big Data Min. Anal.*, vol. 1, no. 1, pp. 1–18, 2018.

[36] M. Liu, the Alzheimer's Disease Neuroimaging Initiative, D. Cheng, K. Wang, and Y. Wang, "Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis," *Neuroinformatics*, vol. 16, no. 3–4, pp. 295–308, 2018.

[37] S. Qiu, G. H. Chang, M. Panagia, D. M. Gopal, R. Au, and V. B. Kolachalama, "Fusion of deep learning models of MRI scans, Mini-Mental State Examination, and logical memory test enhances diagnosis of mild cognitive impairment," *Alzheimers Dement. (Amst.)*, vol. 10, no. 1, pp. 737–749, 2018.

[38] U. Senanayake, A. Sowmya, and L. Dawes, “Deep fusion pipeline for mild cognitive impairment diagnosis,” in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018, pp. 1394–1997.

[39] Y. Shmulev, the Alzheimer’s Disease Neuroimaging Initiative, and M. Belyaev, “Predicting conversion of mild cognitive impairments to Alzheimer’s disease and exploring impact of neuroimaging,” in *Lecture Notes in Computer Science*, Cham: Springer International Publishing, 2018, pp. 83–91.

[40] 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 - ACM-BCB ’17*, 2017.

[41] C. Pettit, I. Bishop, V. Sposito, J.-P. Aurambout, and F. Sheth, “Developing a multi-scale visualisation framework for use in climate change response,” *Landsc. Ecol.*, vol. 27, no. 4, pp. 487–508, 2012.

[42] H. J. Johnson, M. McCormick, and L. Ibáñez, “The ITK Software Guide Book 1: Introduction and development guidelines (volume 1),” 2015.

[43] M. Jenkinson, C. F. Beckmann, T. E. J. Behrens, M. W. Woolrich, and S. M. Smith, “FSL,” *Neuroimage*, vol. 62, no. 2, pp. 782–790, 2012.

[44] Elsevier, *Statistical Parametric Mapping: The analysis of functional brain images*, 1st ed. San Diego, CA: Academic Press, 2006; vii, 647.

[45] “CMIC,” *Ucl.ac.uk*. [Online]. Available:http://cmictig.cs.ucl.ac.uk/wiki/index.php/Main_Page. [Accessed: 19-Apr-2022].

[46] B. Avants, N. J. Tustison, and G. Song, “Advanced Normalization Tools: V1.0,” *Insight J.*, 2009.

[47] M. Nolden *et al.*, “The Medical Imaging Interaction Toolkit: challenges and advances : 10 years of open-source development: 10 Years of open-source development,” *Int. J. Comput. Assist. Radiol. Surg.*, vol. 8, no. 4, pp. 607–620, 2013.