Deep CNN Based Multi-class Classification of Alzheimer's Disease using MRI

Author

Ammarah Farooq

NUST201464501MCEME35214F

Supervisor

Dr. Saad Rehman

DEPARTMENT OF COMPUTER ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

AUGUST, 2017

Deep CNN Based Multi-class Classification of Alzheimer's Disease using MRI

Author

Ammarah Farooq

NUST201464501MCEME35214F

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Computer Engineering

Thesis Supervisor:

Dr. Saad Rehman

Thesis Supervisor's Signature: ____

DEPARTMENT OF COMPUTER ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY, ISLAMABAD

AUGUST, 2017

Declaration

I certify that this research work titled "*Deep CNN based Multi-class Classification of Alzheimer's Disease using MRI"* is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged / referred.

Signature of Student

Ammarah Farooq

NUST201464501MCEME35214F

Language Correctness Certificate

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical and spelling mistakes. Thesis is also according to the format given by the university.

Signature of Student

Ammarah Farooq

NUST201464501MCEME35214F

Signature of Supervisor

Copyright Statement

Acknowledgements

ALHAMD-U-LILLAH. I am thankful to my Creator ALLAH Subhana-Watala to have guided me throughout this work at every step and for every new thought which YOU setup in my mind to improve it. Indeed I could have done nothing without YOUR priceless help and guidance. Whosoever helped me throughout the course of my thesis, whether my parents or any other individual was YOUR will, so indeed none be worthy of praise but YOU.

I am profusely thankful to my beloved parents who raised me when I was not capable of walking and continued to support me throughout in every department of my life. Their prayers are priceless and are the real secret of my success.

I would also like to express special thanks to my supervisor Dr. Saad Rehman for his support throughout my thesis. He is the one who gave me research insight and I always got confidence after meeting him. He is the one who raised my interest towards research.

I would also like to pay special thanks to Dr. Muhammad Awais for his tremendous support and cooperation. Each time I got stuck in something, he came up with the solution. I am really thankful to him for providing me the world class computing resources. Without his help I wouldn't have been able to complete my thesis. I appreciate his patience and guidance throughout the whole thesis and his support in my hard time.

I would also like to thank Dr. Muhammad Abbas, Dr. Ali Hassan and Dr. Farhan Riaz for being in my thesis guidance and evaluation committee and express my special thanks to Dr. Muhammad Anwar for his support, guidance and cooperation.

Finally, I would like to express my gratitude to my college and all the individuals especially lab members and Dr. Salman and his family who have rendered valuable assistance throughout my masters.

Dedicated to my beloved father whose tremendous support, prayers and cooperation led me to this wonderful accomplishment of my dream

Abstract

In the recent years, deep learning has gained huge fame in solving problems from various fields including medical image analysis. This thesis proposes a deep convolutional neural network based pipeline for the diagnosis of Alzheimer's disease and its stages using magnetic resonance imaging (MRI) scans. Alzheimer's disease causes permanent damage to the brain cells associated with memory and thinking skills. The diagnosis of Alzheimer's in elderly people is quite difficult and requires a highly discriminative feature representation for classification due to similar brain patterns and pixel intensities. Deep learning techniques are capable of learning such representations from data. In this thesis, a 4-way classifier is implemented to classify Alzheimer's (AD), mild cognitive impairment (MCI), late mild cognitive impairment (LMCI) and healthy persons. Experiments are performed using ADNI dataset on a high performance graphical processing unit based system and new state-of-the-art results are obtained for multiclass classification of the disease. Results are examined for two state-of-the-art models i.e. googLeNet and ResNet-152. An optimized and dedicated model is also proposed which is based on incorporating residual learning in shallow networks. Experiments are performed in three phases with both learning from scratch and fine-tuning techniques. The acquired results outperformed other techniques for both binary and multiclass classification. The proposed technique results in a prediction accuracy of 99.9% by using proposed optimized model, which is a significant increase in accuracy as compared to the previous studies and clearly reveals the effectiveness of the proposed method.

Key Words: *Alzheimer's disease; mild cognitive impairment; deep learning; structural MRI; multi-class classification, ResNet, GoogleNet, residual learning, ADNI*

Table of Contents

List of Figures

List of Table

Chapter 1

Introduction

Human brain is the most central and the complex organ in its formation as well as functioning. It is the central processing unit of human body. Each action of a person, simplest or complex, is ruled by the signals and messages which are received and sent by his brain. Besides, it is also responsible for providing the memory bank which can store thoughts and memories of one's life. Different parts of brain are responsible for different activities. The study of neurosciences is thus crucial for dealing with neurological diseases and other issues. As we can understand the importance of memory in our life, in the same way we can imagine the disaster of a term "memory loss".

1.1 Problem Statement

Dementia is a generic medical term which is used for the condition in which mental activity is declining [1] and daily life activities of a patient are becoming disturbed and difficult. There are several types of dementia including permanent and temporary dementia. Number of reasons are behind for causing the situation like accidents, nervous breakdown or natural aging process. These causes are accompanied with structural changes and damages to brain tissues. Most commonly known form of dementia which is caused by increasing age is Alzheimer's disease (AD). Alzheimer's is a chronic and neurodegenerative disease associated with elder people and accounts for about 60 to 80 percent dementia cases [2]. Detailed overview of AD is presented later in this chapter. The diagnosis of such a disease is crucial and demands a treatment ahead of time. Owning to complex formation of brain, natural or trauma-caused degeneration of brain tissues becomes hard to detect and understand. In case of AD, the overall size of brain reduces as compared to healthy brain. Thus, a skillful and expert system with high accuracy is required to diagnose the diseased condition.

1.2 Motivation

Recent advances of deep learning and artificial intelligence (AI) in solving problems from medical field became a motivation for this proposed research work. Future is now being considered to be associated with deep learning and its applications. The finest research groups from all over the world have launched dedicated AI based projects and research centers in every domain including medical, healthcare, robotics, online marketing, advance farming, lifestyle, data science, automatic vehicles and so on. A group of Stanford researchers, recently developed a deep learning based algorithm to diagnose 14 type of cardiac rhythm abnormalities with accuracy of an expert cardiologist [3]. Similarly, number of researches are available which efficiently diagnose breast cancer, prostate cancer, skin cancer and various tumors by applying deep learning. Research group from University of Michigan Medicine developed an imaging technique which detects a brain tumor with about 90% accuracy and still faster than the existing laboratory method (takes minimum 30 to 40 minutes). According to team, their technique takes only 3 minutes to detect the tumor [4]. Such researches inspire us to seek solution to our problem of accurately diagnosing AD using deep learning.

1.3 Overview of Alzheimer's disease

Alzheimer's disease (AD) is an irreversible and progressive neurological disease. It is considered as most common form of dementia. It starts with a little fail to recall things or recent events but it worsens overtime and may lead to total loss of memory as well as ability to perform simple tasks e.g. conversation, properly responding to stimulus and self care. In advance stage of disease, the patient is completely dependent on others. The rate of progression of disease varies from person to person owing to underlying reasons and state. Alzheimer's is the sixth leading cause of death in the United States [5]. Since 2000, mortality rate due to AD has been increased by 89% [5]. According to Alzheimer's Association, in 2017, AD will cost about $259 billion to the nation [5]. Moreover, in case of AD patient, care giving responsibilities and family set-up also face challenges. Although, the disease is commonly being seen in elderly people (age > 65) but it may occur in younger age as well. It may be genetic or a result of any mishap or trauma. Thus the specific cause of AD is not attributed but old age increases the chances of getting into the disease. The progression of disease cannot be stopped by any medicine or treatment, however, it may get slow down temporarily.

Alzheimer's disease accompanies with numerous structural changes of brain. Before moving towards the details of these changes, let us have an idea about the structure of brain. There are three main parts of brain; cerebrum, cerebellum and brain stem. Cerebrum covers the largest area and makes the top frontal portion of brain. It is associated with memory formation, movement of body, logical thinking. Cerebellum is associated with body coordination and balance. Brain stem connects brain to spinal cord and controls the automatic processes of body.

Cerebrum, being the major part, constitutes 85% of the weight of brain. It mainly contains cerebral cortex which is divided into two hemispheres. Cortex is the outer surface of cerebrum having wrinkled form. Different regions of cortex control memory, senses, voluntary activities and thinking process. Hippocampus and ventricles are also present in cerebrum. Hippocampus is responsible for handling short term and spatial memory. Cerebrospinal fluid (CSF) is produced in ventricles of brain. Tissue level anatomy of cerebrum consists of Grey matter (GM) and White matter (WM). Grey matter is formed of numerous cell bodies of neurons and white matter consists of networks of axons and dendrites. Almost 86 billion neurons are present in human grey matter [6].

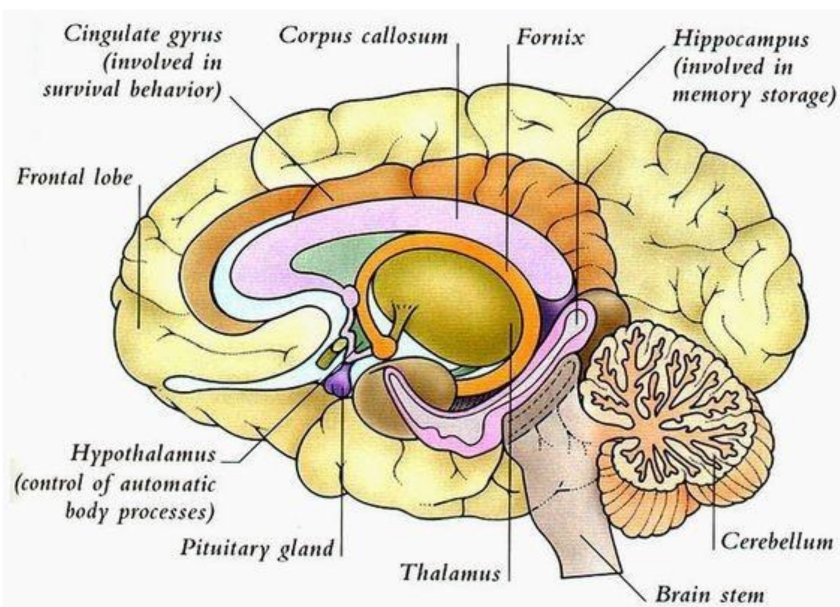

Figure 1.1 - Structure of Human Brain

In Alzheimer's case, cell body of neurons highly reduces with the death of neurons. Tissue matter both GM and WM degenerates and is damaged by protein formed plaques and tangles. Thus size of brain cortex greatly shrinks in AD. Hippocampus also shrinks up and hence AD patients show orientation and short term memory concerns. Brain ventricles get enlarged and CSF chemical composition also changes. Plaques and tangles destroys the cellular connectivity and signaling system. Early difficulties in performing tasks like reading, thinking eventually lead to memory loss and inability to cope with life. Timely diagnosis of disease thus play a vital role and several methods are being used including clinical and imaging techniques.

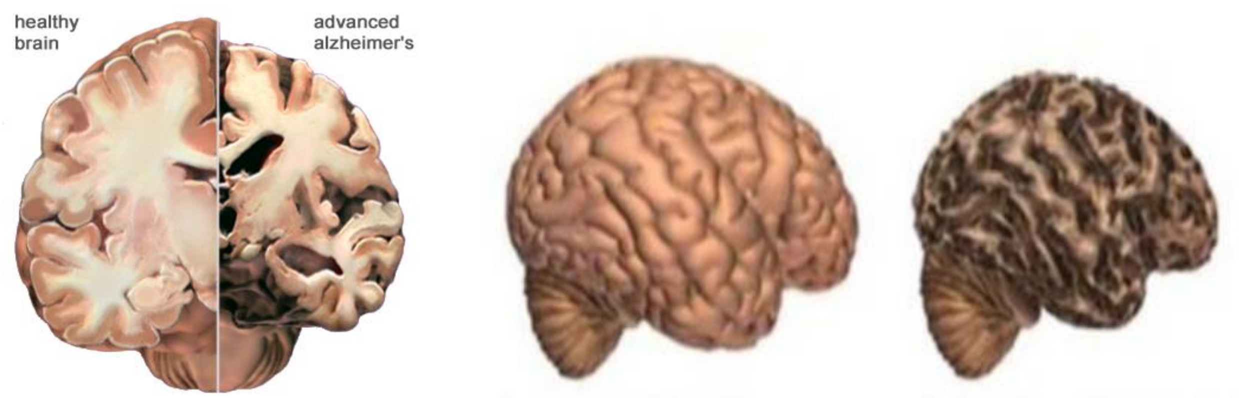

Figure 1.2 - Pictorial view of Normal brain vs Alzheimer's Brain (severe degeneration in AD case)

Clinical methods of diagnosing Alzheimer's are mainly based on several psychiatric assessments which are in the form of questionnaires. These assessments give an idea about cognitive functioning of brain and provide a number to compare against a scale. Mini-mental state examination (MMSE) is the most popular and measures on the scale of 30 where 30 is considered normal and lower values indicate dementia. The questions asked are related to daily life tasks and scored according to patient's ability to cope with them. Clinical dementia rating (CDR). Neuropsychiatric inventory questionnaire (NPI-Q), global deterioration rate (GDR), functional assessment questionnaire (FAQ) are other such assessments and are included as a part of patient's history in formation of standard ADNI [7] dataset. However, modern methods of diagnosing dementia do not rely solely on these assessments and take into account details of information from neuroimaging and biological measures.

Most of the researches, nowadays, are based on Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) scans, mostly complemented with Cerebrospinal fluid analysis (CSF). Cerebrospinal fluid surrounds the brain and spinal cord and acts as a cushion providing the basic protection to central nervous system. This fluid is produced in the ventricles of brain. CSF analysis provides the detailed chemical composition of the fluid. It is considered to be the most accurate bio-marker in diagnosing AD as compared to other fluid bio-markers. AD condition is detected by the presence of senile plaques (composed of a protein A*β)* and a high concentration of tau protein [8]. However, the process of collecting CSF from patient is invasive often combined with a pain as well as difficulty of reproducibility of results and storage of collection limits its use as compared to imaging methods.

MRI and PET techniques are non-invasive and provide potential information about the brain anatomy and activity. MRI technique uses strong magnetic field and radio waves to generate anatomical scans. It provides tissue level structural details of the brain. Hence, it is termed as structural MRI (sMRI), shown in figure 1.2 on left. One variant of this technique is functional MRI (*f*MRI) shown in middle of figure 1.2, which examines the activity of brain in conjunction with the blood flow. The flow of blood increases in the brain region associated with activity performed by a person. PET (right in figure 1.2) also examines on similar principle of blood flow. It is a nuclear functional imaging technique and detects the metabolic process occurring in the brain. It is very useful in studying the brain activity in case of dementia when structural changes are very little to be distinguished. But it is limited to detect small changes or activity owing to its fast radioactive decaying. Both functional MRI and PET provide colorful interpretation of brain activity. Nevertheless, most of the neuroscience researches are based on CSF, MRI, PET and fusion of these modalities.

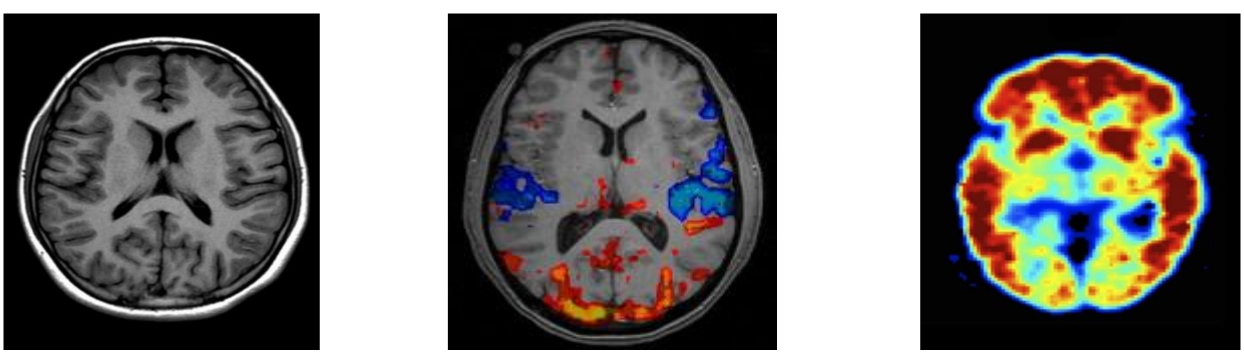

Figure 1.3 - Structural MRI scan (left), Functional MRI scan (middle), PET scan (right)

1.4 Mild Cognitive Impairment

Mild cognitive impairment (MCI) is a brain syndrome which may occur as a transitional phase from normal age decline towards severe forms of dementia. It is also considered as prodromal stage of Alzheimer's if it exhibits symptoms of memory decline. The symptoms of MCI are difficult to be distinguished from the normal aging process of brain and it hardly interferes with the daily life of a person. It may remain stable in some cases for years and thus termed as non-converting MCI (ncMCI). But about 8% to 15% MCI patients are annually converting to dementia condition [9]. Hence, it implies the necessity of diagnosing MCI in its early form and demands the awareness among people to possibly treat or avoid severe conditions i.e. Alzheimer's. MCI requires careful judgement from the clinician due to its unnoticeable yet important to consider nature and after effects. Structural changes are also minor to be differentiated from normal without being an expert.

1.5 Deep Learning

Today's modern age dreams about machine that have intelligence, which can think, act and decide like human. The field of machine learning deals with making machines that can learn and intelligent enough to decide and perform action accordingly. This induced intelligence is termed as artificial intelligence (AI). Traditional pattern recognition and machine learning methods are dependent on the algorithms which can extract useful features from the data. Those features are then fed to a machine learning algorithm to make decision e.g. classify as diseased or healthy. The prediction performance is highly dependent on the learnt features. The AI system based on these algorithms is actually based on the extracted features or representations provided for the decision. The system remains unaware of the actual data and conditions. For better feature extraction, better algorithms are needed which come up with more complexity as well as other factors. The solution for this problem is a learning approach which can learn useful representations automatically and directly from the data. This approach is called Representation Learning [10].

Deep learning is one of the representation learning methods. It is based on artificial neural networks and also termed as hierarchical learning [11]. It is a fresh specialization field of machine learning and an approach to achieve high quality artificial intelligence. The term hierarchical clearly tells us that learning is achieved by moving through the layers of hidden neurons from simple learning representations and moving towards more complex ones. Another important aspect of deep learning is the depth of the network. Each layer represents processing of data to get one level of features. Deeper networks are able to learn very complex representations from data.

Different types of deep learning architectures have been introduced. Research is still active in the field for more and better solution. Deep convolutional neural networks, deep belief networks, deep polynomial networks and recurrent neural networks are some of the introduced architectures. Deep learning has gained huge fame in solving problems of computer vision, natural language processing, medical diagnostics, audio processing and much more. The proposed work is based on convolutional neural networks (CNN) and chapter 3 will provide its detail insight.

1.6 Brief Concept of Proposed Methodology

Section 1.2 gives the main idea that the proposed method will take brain information or scan like MRI and pass it through some deep learnings based system to classify the image into one of the four classes i.e. AD, MCI, later stage of MCI (LMCI) and normal cognitive (NC). Hence, the proposed framework will be a 4-way classifier for diagnosing AD and two stages of MCI and normal. Detailed methodology is discussed in chapter 3.

1.7 Outline of Thesis

The remaining thesis is structured as follows: chapter 2 gives detail literature review on the topic and chapter 3 presents the proposed method in detail. Chapter 4 discusses the experiments and obtained results. Chapter 5 concludes the thesis and discusses the outcomes and future work.

Chapter 2

Literature Review

Diagnosis and classification of Alzheimer's disease remained an active research topic in clinical as well as scientific research groups. This research is extensive and includes methods which use information from various neuroimaging techniques and biological measures, fusion of different information modalities, setting up bio-markers for diagnosis and moves towards complex machine learning systems for higher accuracy. This chapter gives literature summary of different methods of diagnosis and their success in regard of the purpose.

Traditional image processing techniques are based on extracting distinctive features from data. These techniques move from using single modality towards combining various modalities. MRI bio-markers for AD includes hippocampal volume, cortical thickness and grey matter density. Vos et al. [12] believed that combining anatomical measures obtained from MRI would give better diagnosis. They calculated cortical thickness, area, curvature, grey mater (GM) density, subcortical volumes and shape of hippocampus. They took 21 AD patients and 21 healthy. Their work analyzed results using these values as single feature as well as combined features and predicted presence of AD using an elastic net logistic regression method. Their findings demonstrated that combined anatomical measures improved the classification of AD from healthy controls. They achieved an area under the curve (AUC) value of 0.98 by using combination of all the measures. The work of Sorensen et al. [13] is also based on MRI markers. They obtained results for three datasets; ADNI, AIBL and CADDementia challenge test dataset. Two types of data was taken from ADNI; standard MRI set with 504 samples and ADNI HHP set which consisted of 40 manually segmented hippocampus regions. Training set was combined ADNI and AIBL set. Testing accuracy was evaluated on CADDementia test set. The features used are hippocampal shape, texture, volume and cortical thickness. After obtaining features, Z-score transformation was applied to each group according to age. Three way classification was performed using linear discriminant analysis (LDA). They attained the classification accuracy of 63% and AUC of 78.8%

for test set. For ADNI + AIBL, they achieved 62% accuracy for both using full feature representation and selected best 10 features.

Liu et al. [14], recently, proposed a novel method and combined MRI markers based on both ROI and interregional or edge features. They extracted six features i.e. gray matter volume, cortical thickness, surface area of cortex, cortical curvature, folding index, subcortical volume (SV) for both ROI and edge regions. Thus, they got two sets of features having six each. They used feature ranking approach i.e. MKBoost to select and combine features and examine classification performance of each subset. The best performance set gave the required optimal set. They also investigated performance by separately using ROI and edge features with weighted multi kernel learning approach followed by a SVM classifier. Results were obtained for binary classification for four different obtained optimal set of features. Best accuracies achieved for AD vs NC were 92.6, 93.5% and 95.2% by using ROI only, edge features only and combined features respectively. Respective AUC measures were 94.2%, 96.3% and 97.5%. The proposed method gave 86.35% accuracy for MCI vs NC. The proposed method proved powerful for exploiting single modality features.

Previtali et al. [15] extracted features using recently introduced oriented FAST and rotated BRIEF techniques from MRI scans. They also incorporated information about the location of disease and its distribution with in brain. All of the mentioned combined gave outstanding results. ADNI and OASIS dataset for four classes were used and only central slice was selected for the proposed study. Data augmentation was performed by flipping the slice horizontally as well as vertically. Images from both datasets were merged together for training set and test sets were evaluated separately. Feature vector was constructed using image descriptor, histogram, cantor hash, combination of histogram and cantor hash and class label. Classification was performed using SVM by WEKA. Results were obtained by varying number of features and number of pixels in the grid cell. For both datasets, best results for four way classification were acquired at lowest grid size of about 46 pixels. 98% accuracy was achieved for ADNI AND 77% for OASIS. For three way classification, 100% accuracy was achieved for ADNI by using maximum number of features and 0.002% grid cell density and 85.4% for the other set. For AD vs NC, 100% achieved for ADNI and 97% for OASIS. Results were also examined by varying selection of features. Best results were given by selecting combined histogram and cantor hash features. Achieved binary diagnosis accuracy was 99.6 and multi-class accuracy was 98% for ADNI.

Kloppel et al. [16] presented a unique work and investigated the robustness of MRI scans for Alzheimer's diagnosis. They formed four dataset groups with MRI scans from different centers (datasets) and with different age means. Three groups were based for AD vs NC classification. Fourth group was made to examine distinguishing ability of MRI for different diseased groups i.e AD vs frontotemporal lobar degeneration (FTLD). They extracted grey matter from scans and calculated kernel matrix followed by linear SVM classifier. For whole GM image, they obtained 96.4% accuracy for the case where training was performed on group 1 data and group 2 was used as test set. 87.5% accuracy was achieved the other way round. 95.6% accuracy was achieved for combined data of group 1 and 2. 89.2% accuracy was achieved for discriminating AD and FTLD samples. For ROI based GM images, 71.4% and 70% was obtained for alternate train and test sets with group 1 and 2. 94.1% was achieved on combined data set.

Huang et al. [17] claimed that early diagnosis of Alzheimer's was crucial and hence, focused their work on mild cognitive impairment and its prediction for conversion to AD in later stages. They examined longitudinal changes in brain cortex from MRI over time for MCI patients and performed hierarchical classification for cMCI vs ncMCI. Longitudinal changes became noticeable with time if disease was converting to severe dementia. Data was obtained from ADNI for different time stamps; six months to up to 3 years. Thus total 574 MRI volumes were used. Voxel selection was performed in first step followed by hierarchical classifier. Three logical regression classifiers were learnt in stack performing voxel level learning then spatial level and finally image level learning followed by classification. The best accuracy achieved was 79.4% for predicting conversion to Alzheimer's disease.

References [18, 19] dealt with incorporating multi-modal features. Most of the features are derived from sMRI, PET scans and CSF analysis. Zhang et al. [18] extracted features from MRI and PET scan using the standard atlas templates. Linear kernel matrices were then calculated for selected features. Kernel matrix was also obtained CSF features. Three kernels are combined by using kernel combination method. SVM classification was performed on the combined kernel. Diagnosis was performed as AD vs NC and MCI vs NC. 93.2% and76.4% accuracy was achieved for both respectively by using multimodal data. The team also examined results for single modality and obtained 86.2% accuracy for AD vs NC for structural MRI images. By using only imaging data, MRI combined with PET gave 90.6 % classification accuracy. Tong et al. [19] claimed that multi-modal information is not necessarily linearly related to each other. Therefore, they used non-linear graph fusion method instead of linear kernel combination. They used four type of features from for different modalities; volume from MRI scans, tissue intensities from PET scans, CSF measures and genetic information. Similarity graphs were constructed for each of these feature and then combined non-linearly to form a unified graph for classification. Binary and three way classification was performed using random forest approach. The proposed method obtained 60.2% multiclass accuracy and 72.9% AUC measure. 91.8% accuracy was achieved for AD vs NC classification. They also compared results for using single vs multimodality data. However, the best results were obtained for fusing all the four modalities together.

The work discussed so far is dependent on complex feature extraction and selection algorithms. Further discussion is based on deep learning based approaches for Alzheimer's diseases diagnosis. Siqi Liu [20] and his team carried out number of experiments with various techniques. They incorporated stacked autoencoders (SAE) for learning representations from the data. An autoencoder is an artificial neural network with multiple hidden layers. It can learn useful encodings from the data. The output of an autocoder is a sparse representation of the input data. Liu et al. [20] trained the stacked autoencoder network on ADNI sMRI data. Total 83 ROI regions were extracted from scans and passed for training and then applied softmax regression on top of the network as classifier. They also investigated results for single and multi-kernel SVM classification. They classified four classes (AD, ncMCI, cMCI, NC) by one-vs-all strategy. They obtained for 87.7% accuracy for binary classification and average accuracy of 47.7% for multiclass. These results were slightly higher than for the other two classifiers. However, they broadly extended this work in [21]. PET data was also now included. In the first step, ROI was extracted from grey matter images of MRI and PET scans. The extracted information from both modalities was used to train a SAE network. To avoid noise from other modality, zero masking strategy was used. Fine-tuning was performed after putting softmax classifier on top with respect to classifier loss. This work reached about 91.4% accuracy for AD/NC, 82% FOR MCI/NC and 53.7% for four classes as one-vs-all.

In another work by Liu et al. [22], representations were learnt from data in multiple phases. First phase of SAE network learnt features from neuroimaging data i.e. MRI and PET. Second phase fine-tuned the pre-learnt features by incorporating clinical data. MMSE data of patients was used to optimize these features. Linear regression was used to estimate the MMSE features. The trained linear regression layer was put on top instead of softmax classifier. The acquired results demonstrated the usefulness of proposed method over simple SAE and multi-kernel SVM classifiers. The multiphase method achieved 90% accuracy on binary and 59% on three class dataset.

Another deep learning based feature learning approach is deep polynomial networks (DPN). Polynomial networks as named indicate predicts the class based on underlying trained polynomial learning. First layer of the network is a linear layer which processes data and send that to higher order layers in a feed forward fashion. DPN provides complex representation of the input data which are useful for class distinction. Shi et al. [23] used stacked polynomial network architecture for their work where one whole polynomial network was put on top of the other. They developed series of results using single as well multi-modal data. For multimodal network, MRI and PET data was passed through separate set of DPNs. Features learnt in first stage were combined and passed to the final stage of SDPN. The output is then fed to the linear or SVM classifier. Results were obtained for different sets of binary classifications. For AD vs NC, they obtained 93% and 94% for single MRI and PET modality respectively using proposed SDPN with linear classifier. Multimodal accuracy achieved is 96.93% with SVM classifier, results were 95.4%, 95.1% and 97% for MRI only, PET only and multimodal respectively. For MCI vs NC, 87.2% and 86.9% accuracy was achieved for proposed multimodal approach with SVM and linear classifier respectively. They also obtained results for ncMCI vs cMCI. Four way multiclass classification was performed for AD vs cMCI vs ncMCI vs AD. Proposed approach achiecved about 57% accuracy for multi classification with SVM classifier. The proposed method proved to be very effective for AD vs NC as compared to multiclass classification.

In [24, 25], Suk et al. claimed that latent features and non-linear patterns were present in the brain lesions. They extracted those features using stacked autoencoder networks. The extracted MRI, PET and CSF features were fed to SAE to learn and tune latent features in semi-supervised manner. Latent features were used combined with low level features like GM regions. At the next stage, feature selection was performed and MMSE scores from clinical history was also incorporated. After learning sparse representation for each modality, kernel matrices were obtained and fused together and passed through SVM classifier. Results were obtained for multiple binary pairs through extensive experimentation. For AD vs NC, proposed method gave 98.8% accuracy and 90.7% for MCI vs NC. Prediction accuracy of 83% was achieved for MCI vs AD and MCI converters vs non-converters.

References [26, 27] implemented hierarchical classifier for Alzheimer's and MCI diagnosis. However, Suk et al [26] used multimodal features and Liu et al. [27] used MRI features only. Suk et al. [] divided images into numerous patches and performed feature learning through deep restricted Boltzmann machine (DBM). Separate feature learning pipelines were used for MRI and PET images. Classifier was designed in hierarchical fashion in which patch level learning was performed in first stage followed by image level learning. Final classification stage contained ensemble of weighted SVM classifier. On patch level stage, lower layers learnt features from each modality separately and concatenated at the top layers. Patch level learning was combined to train bigger patches and finally image level classifier. Reference [27] achieved 90% and 85.3% for AD vs NC and MCI vs NC respectively. However, by multimodal fusion, [26] obtained 95.35% and 85.67% respectively. Prediction accuracy of 75.9% was achieved for MCI vs AD and MCI converters vs non-converters.

In the recent work, Suk and his team [28] used sparse multi-task learning (MTL) for feature selection and classifier learning. Features were concatenated from MRI, PET and CSF analysis. These were passed through deep multi-task learning module. The idea behind was that through recursive application of multi-task learning, only useful features were selected from the high dimensional feature vector and uninformative data was discarded. At the same time, regression coefficients which represented optimal feature selection were used as weights for subsequent iteration of learning. Thus, proposed method was termed as weighted sparse MTL. Results were examined for both binary and multiclass diagnosis. Fusion of three modalities with proposed method gave 95.09%, 78.77% and 73.04% for AD vs NC, MCI vs NC and stable MCI vs progressive MCI respectively. The proposed method achieved 62.93% for 3-way and 53.7% for 4-way classification. As most of the studies are based on MRI data, Suk also investigated results for only MRI with proposed technique. The dataset included total 805 subjects. Achieved results were

90.27%, 70.86%, 73.93%, 57.74% and 47.83% for three binary, three way and four way classification respectively.

Gupta et al [29] was the first to use the concept of convolution for neuroimaging data classification. They obtained feature bases in first step and convolved those bases with test images to get the required results. First step was performed using SAE network for two types of bases; natural image and MRI bases. They used 10,000 patches of natural images and ten times more for MRI data obtained from ADNI. Training images and patches were passed through SAE for 100 learning bases vectors which could distinguish MRI markers. No preprocessing was performed on MRI data. Each of the 100 bases were convolved with test MRI scan followed by sigmoid activation function. The size of feature vector obtained after pooling was still 61200 for each MRI scan. The feature vector was then passed through the feed forward neural network having logistic units for classification. For binary classification, 94.74% and 93.8% for AD vs NC using natural and MRI bases respectively. For MCI vs NC, 86.3% and 83.3% respectively. Their method obtained 85% accuracy for three way classification using natural images. MRI bases gave 78% performance for multi-classification.

Payan et al. [30] used 3D convolution approach using structural MRI volumes. Most of the other works used 2D slices of axial MRI scan. Payan and his team used complete MRI volume information from all the axes i.e. axial, coronal and sagittal. They compared results for 2D vs 3D convolution method. First step was same as previously mentioned work and extracted bases using SAE network. 150 bases or filters were then applied in convolutional neural network. First stage performed 3D convolutions followed by huge size reduction from 64×91×75 to 12×18×15. Next stage was fully connected layer consisted of 800 hidden neurons and softmax classifier was the final output stage. Weights were trained using gradient descent with mini batch approach. Similar experimental setup was developed for 2D convolution technique. Results obtained with 3D convolutions outperformed the other method and showed superiority in capturing more information from the MRI scan. The method obtained 95.3% for AD vs NC, 92.11% for MCI vs NC and 89.5% for three way classification.

Hosseini et al [31] expanded the concept of 3D convolution and proposed a CNN based classification framework which was based on features learnt from a 3D convolutional autoencoder

(CAE). This work is quite comprehensive and examines various aspects of deep learning. The main CNN network consisted of three convolutional layers followed by various fully connected layers and a classifier layer. Rectified linear activations were applied with max-pooling layers. For the three convolutional layers, pre-learnt bases were obtained by using 3D-CAE. Features were learnt for axial and sagittal axes of ADNI MRI data. Convolutional autoencoder allowed weight sharing and thus computationally less expensive then fill connections. First layer of a CAE was convolutional layer which extracted four types of features maps from structural scan; brain size, ventricle size, cortical thickness and hippocampus size. Feature maps from convolution layer were fed to decoding layer and reconstructed 3D image was obtained at the output of CAE. The feature on CAE was performed on CADDemantia set and tops layers were fine-tuned for target ADNI MRI set. Implementation was performed using Theano framework on GPU based systems. The proposed method proved very effective especially for multiclass classification and obtained highest accuracy compared to previous techniques. 99.3% accuracy was achieved for AD vs NC, 94.2% for MCI vs NC and 95.7% for (AD + MCI) vs NC. For distinguishing AD vs MCI vs NC, 94.8% accuracy was achieved. The proposed 3D CNN method is evidently very powerful.

Further discussion is based on the methods which are totally based on deep convolution neural network and where no pre-learnt SAE stage is present. In a recent work, Korolev et al [32] proposed a 3D convolution neural network model VoxCNN inspired by VGG network. They examined results for plain form of the model as well as residual form. In residual architecture, identity connections were added to the network. Results were obtained for six binary classifications. Only 231 images obtained from ADNI for four classes were used for the experiments. Proposed method achieved 79% for plain network and 80% with residual counterpart for AD vs NC. For early MCI vs NC, accuracies were 54% and 56% and for late MCI vs NC, 63% and 61% respectively for plain and residual network.

Sarraf et al. [33] applied famous LeNet and GoogleNet models for AD classification. They used two datasets, one with resting state functional MRI and other one with structural MRI scans. It is the first work dealing with *f*MRI data for neuroimaging diagnosis. Careful preprocessing with various smoothing filters was applied on structural images. Four sets of sMRI data were prepared with different level of smoothing and total 62,335 images were obtained. The data were converted to LMDB format before passing to the CNN network. For *f*MRI data, 100% accuracy was achieved using GoogLeNet and 99.9% with LeNet. However, for structural scans accuracy ranged from 97 to 98.7% using Lenet for different smoothing levels. For GoogLeNet, 84 to 98.8% was achieved. However, obtained results were lower when balanced dataset was used and accuracy ranged from 95.7 to 97.8% using LeNet. The unbalanced set was highly biased towards AD with ratio 5:1. Apart from slice classification, subject level classification was also performed. For all the slices belonging to a specific subject, classifier prediction was noted and majority voted class was assign to the subject. 100% accuracy rate was achieved with spatial smoothing the sMRI images using both models. The team showed more dedication towards functional MRI and suggested that the strict preprocessing was required by the method. However, this work highly shows the potential of incorporating deep CNN models for neuroimaging classification.

The literature discussed above shows good results for binary classification specifically for discriminating Alzheimer's from normal. Moreover, most of the results were developed in separate setup for each classification. In this thesis, focus is put on improving multiclass diagnosis of the disease in single experimental setup and model learning.

TABLE 2.1 -
SUMMARY OF LITERATURE REVIEW

Technique	Modality	AD vs NC (%)	MCI vs NC (%)	Multiclass (%)
Sorensen et al. [13]	MRI	-	-	63.0
Lui et al. [14]	MRI	95.24	86.35	-

Previtali et al. [15]	MRI	100	100	98
Kloppel et al. [16]	MRI	95.6	-	-
Zhang et al. [18]	MRI + PET +CSF	93.2	76.4	-
Tong et al. [19]	MRI + PET + CSF + genetics	91.8	79.5	60.2
Altaf et al. [34]	MRI + clinical	94.6	-	
Liu et al [20]	MRI +PET	87.7	76.9	47.42
Liu et al. [21]	MRI +PET	91.4	82.1	53.79
Liu et al [22]	MRI + PET + clinical	90.11	-	59.19
Shi et al. [23]	MRI + PET	97.13	87.24	57.00
Suk et al. [24]	MRI + PET + CSF + Clinical	95.9	85	-
Suk et al. [25]	MRI + PET + CSF + Clinical	98.8	90.7	-
Technique	**Modality**	**AD vs NC (%)**	**MCI vs NC (%)**	**Multiclass (%)**
Suk et al. [26]	MRI + PET	95.35	85.67	-
Liu et al. [27]	MRI	92	85.3	-
Suk et al. [28]	MRI + PET +CSF	95.09	78.77	62.93 53.77 (4-way)

Suk et al. [28]	MRI	90.27	70.86	57.74 47.83 (4-way)
Gupta et al. [29]	MRI	94.7	93.8	85
Payan et al. [30]	MRI	95.39	92.11	89.47
Hosseini et al. [31]	MRI	99.3	94.2	94.8
Korolev et al. [32]	MRI	80	63	-
Sarraf et al.[33]	rs-*f*MRI	100	-	-
Sarraf et al. [33]	MRI	98.8	-	-

Chapter 3

Convolutional Neural Networks (CNN)

Convolutional neural networks (CNN) or ConvNets are same as simple artificial neural networks having hidden neurons and output class scores but CNN contain single or multiple convolution layers instead of simple matrix multiplication. It is assumed that input of a convNet is an image or 2D data which permits us to use it as image processing tool. ConvNets have gain huge fame in number of applications. It was first introduced by LeCun et al. [35] in 1989 for hand written digit recognition. The presented network is famous as LeNet. LeNet has five layers with two convolution layers. However, the concept of deeper network was not established at that time. With the rise of high end computing systems, graphics processing units, optimization algorithms and efficient backpropagation algorithm, deeper networks came into being with millions of learning parameters. In 2012, AlexNet [36] introduced for ImageNet recognition challenge got tremendous prominence and after then deep learning applications found their place everywhere.

Convolutional networks are inspired by human visual cortex [37]. It learns features in hierarchical manner starting with simple points, then lines and edges and finally objects. Convolution is a linear operation and it supports sparse connections and parameter sharing. Each convolution works on its receptive field and produces response when input matches with the applied kernel. As compared to end to end connection, convolution operation works on the local patch of the input according to size of the filter. In this way, it also achieves the property of shift invariance. Single set of parameters is shared across the all feature maps of one convolution layer. Next section will discuss the important modules of CNN and their working in detail.

3.1 Building Blocks of CNN

Basic CNN architecture has following building blocks:

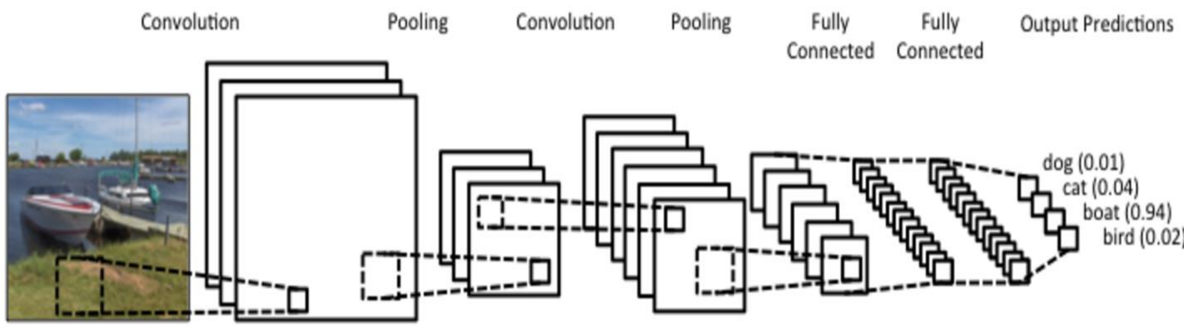

Figure 3.1 – Example of a Convolutional neural network with basic blocks (adapted from [60])

3.1.1 Convolution Layer

It is the basic layer and characterized by number of inputs, number of output maps, kernel size, stride size and padding value. When the kernel is convolved with the input, feature maps are generated. When convolution is performed, size of the input decreases. Stride value tells the step size to kernel while performing the operation. Convolution layer can also be used for dimensionality reduction by using kernel size 1×1 and strides greater than one. It simply discards the pixels according to stride value while keeping the other as such.

3.1.2 Activation Layer or ReLU

Activation layer generates element-wise activation maps according to a threshold or function. This layer has similar task as the firing function for neurons. Once the product of inputs and weights is obtained, activation function looks into values and decides to propagate output signal accordingly. Traditional networks were greatly based on sigmoid activation function. But modern convNets are using Rectified Linear units (ReLU) for activation as it offers faster convergence and easy to implement. If the value is lower than zero, it is set to zero and remains unchanged otherwise. It computes $f(x) = max\ (0, x)$. Other variants of activation functions are also available and depends on choice or hit and trial selection.

3.1.3 Pool Layer

Pooling operation reduces the size of feature maps and simultaneously connects local features to global features. It is applied on the local patch of input and performs down sampling according pooling operation type. Max pooling is most commonly used. It looks for the maximum value in the window and returns it in the downsized map. It is used in between the network after ReLU layers. Average pooling is another type of pooling which replaces the window elements with the average value of all the values in the window. This pooling is usually used at the end of network before fully connected layers.

3.1.4 Fully Connected (FC) Layer

One or more fully connected layers are put on top of the convolutional layers. As the name indicate, FC layers are same as simple multi-layer perceptron with each node connecting to every other node in the next layer. FC layers have dense connections and thus requires more parameters. These layers are used to convert the features into vectorized class scores.

3.1.5 Batch Normalization Layer

In the modern networks, a batch normalization layer id frequently used. It is introduced by Loffe et al. [38] and deals with parameter initialization problem. Bad initialization can lead to wrong results as well as waste of time. However, batch normalization technique force each weight of neuron follow the unit Gaussian distribution. This layer is repeated in the network and updates weights by using batches of data. This is useful as well as differentiable, hence, included as an independent layer.

3.2 Overview of ImageNet Models

ImageNet [39] is a large scale visual recognition challenge which is held annually. About 1.2 million images from 1000 different classes forms the giant ImageNet dataset. It challenges on different tasks including classification, detection and localization. After the success of AlexNet in 2012 for classification, numerous applications are using ImageNet models or their variants. ImageNet models are briefly discussed below.

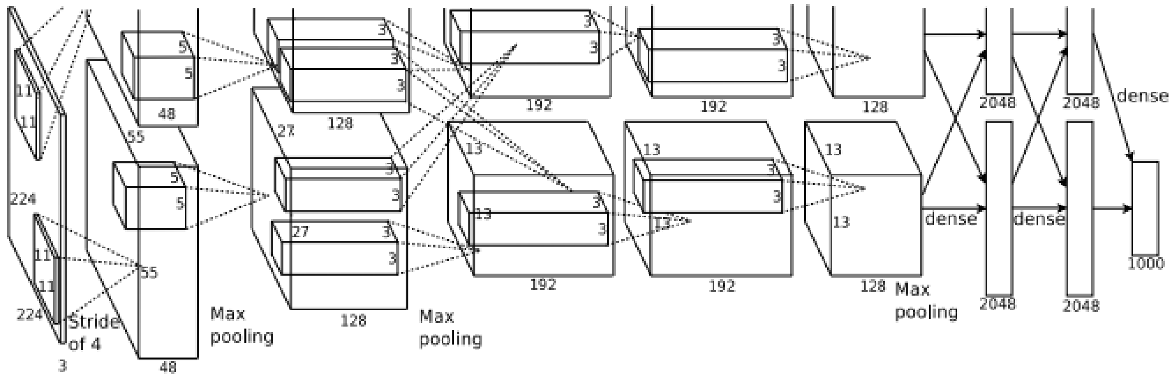

Figure 3.2 – AlexNet architecture (reprinted from [36])

3.2.1 AlexNet

It was introduced by Alex Krizhevsky and his team and obtained first position in classification challenge [36]. Error rate for AlexNet was significantly lower than the runner up. It is simple 8 layered sequential model and hence, most of the new researches are still using this model for simplicity. It used three FC layers and combined the model with the replica of it. It has about 60 million learning parameters. Figure 3.2 shows the AlexNet architecture.

3.2.2 ZF-Net

This model is named after its presenters Matthew Zeiler and Rob Fergus. It is winner of ImageNet 2013 challenge. It is similar to AlexNet and is a kind of its variant with better selection of hyperparameters. This model achieved 11.2% error rate as compared to 15.4% for AlexNet [40].

3.2.3 VGG-Net

This model is the runner up of ImageNet 2014 challenge but still has lower error rate than the previous two models. It is also most widely used specially when considering the visual inputs. Simonyan et al. [41] presented the homogeneous network with using only one fixed size of filters. They introduced 11, 16 and 19 layered sequential VGG nets. Due to fixed smaller kernel size in the lower layers, number of parameters are very large and about 140 million. Thus, this model is computationally expensive. However, it supports the fact that the depth is critical factor to deal with the built-in hierarchy of visual data.

3.2.4 GoogLeNet

It is 22-layered model introduced in 2015 by Google team and was the winner of ILSVRC-2014 [42]. It is the first architecture that differs from traditional CNN architectures i.e., stacking more layers on top sequentially. It is kind of network in network topology considering width more than increasing depth. This model takes care of the computational budget, while allowing processing in depth as well as width. Increasing depth on one hand may produce better results but on other hand it also increases learnable parameters and chances of overfitting in case of small labeled data available. Uniqueness of this network lies in its “Inception module”, formed based on Hebbian principle and multiscale processing as shown in figure 3.3. The idea behind this module is considering maximum information coming from the input patch. Different sizes of filter extract information on fine scale as well as cover large scale, whereas max-pool and 1×1 convolution layers reduce dimensionality. The information is aggregated and passed to the next layer. Overall, the model take control over the number of learnable parameters and has 12 times fewer parameters than AlexNet [42].

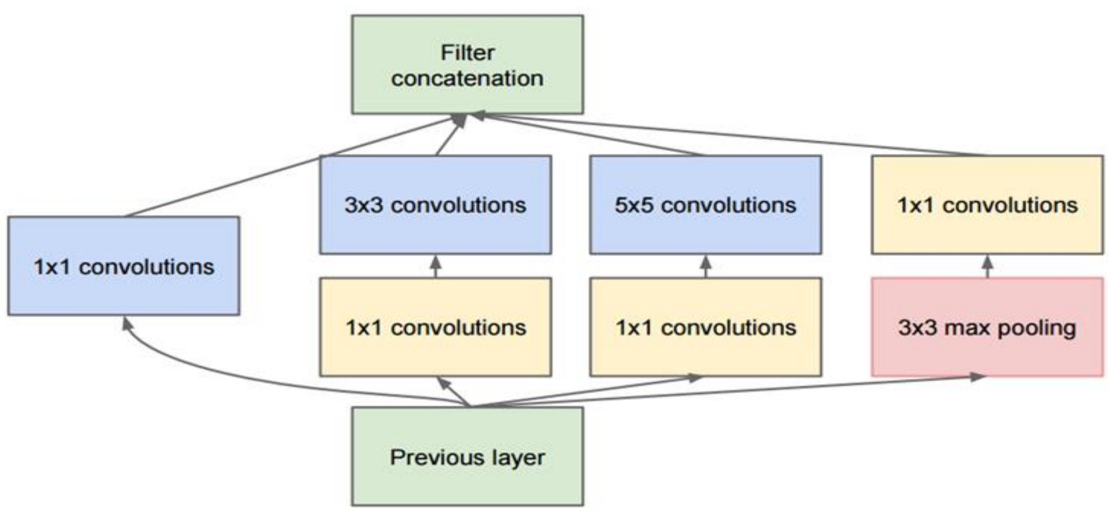

Figure 3.3 – Inception module for GoogLeNet (reprinted from [42])

3.2.5 ResNet

Residual Network (ResNet) was developed by Kaiming et al. [43] and was winner of ILSVRC-2015. The architecture included a central concept of shortcut or skip connections which are added as bypass to convolutional layers of regular feed-forward network; making the block a residual block as shown in figure 3.4. The main idea is that at each step instead of learning features from a function F(x), the features are learnt from F(x) plus original x which makes optimization easier and makes the model to converge faster. For normal feed-forward networks, the prediction accuracy decreases as the depth of network increases. Many factors are responsible for such results including vanishing gradient problem, saturation, size of training data and over-fitting. Residual learning allows network depth to become as deep as more than thousand layers. During backward pass, skip connections make flow of gradient easy and solve vanishing gradient problem. A ResNet with 152 layers is 8 times deeper than VGG network but still have lower computational complexity [43].

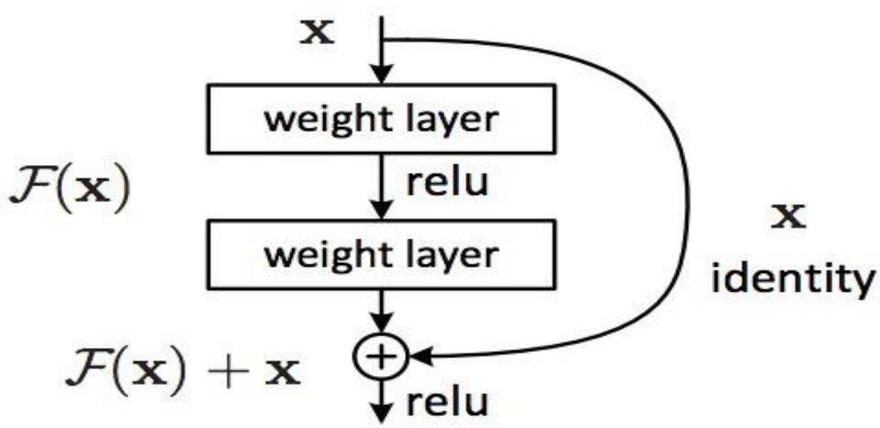

Figure 3.4 – Residual learning block (reprinted from [43])

Chapter 4

Methodology

After getting insight from previous chapters, this chapter presents the comprehensive methodology of the proposed work. It also discusses about the dataset used and experimental set-up for the research. The proposed method is based on single modality data i.e. structural MRI. Extensive experimentation has been performed using variety of models.

The proposed classification framework consists of two main stages; preprocessing and network training using deep convolutional neural networks (CNN). Block diagram is shown in figure 4.1 and details are discussed in the subsequent subsections.

4.1 Data Preprocessing

MRI scans are provided in the form of 3D Nifti volumes containing complete brain scan from three axes; axial, sagittal and coronal. Most of the studies are based on using axial scans only as it provides complete cortical region picture. However, some of the hippocampal region information is also visible through sagittal scans. This work also uses axial scans for investigation. Preprocessing is performed on whole 3D volume. At first, skull stripping and tissue segmentation is performed. Grey matter (GM) and white matter (WM) tissues are segmented out using SPM8 [44] tool. Spatial normalization and bias correction is applied on extracted GM scans. Modulation or warping is performed in final step to check for any discrepancy introduced by normalization step. However, as compared to [33], no spatial smoothing filters are applied in any step. Grey matter volumes are then converted to 2D JPEG images as this work incorporated 2D convolutions only. Slicing is performed using Python Nibabel package. For each axial scan, about 160 slices are obtained separated by 1mm distance. These slices are visually checked for any distortions. Slices from start and end which contain no related information are discarded from the dataset. Thus, about 60 to 75 useful slices are

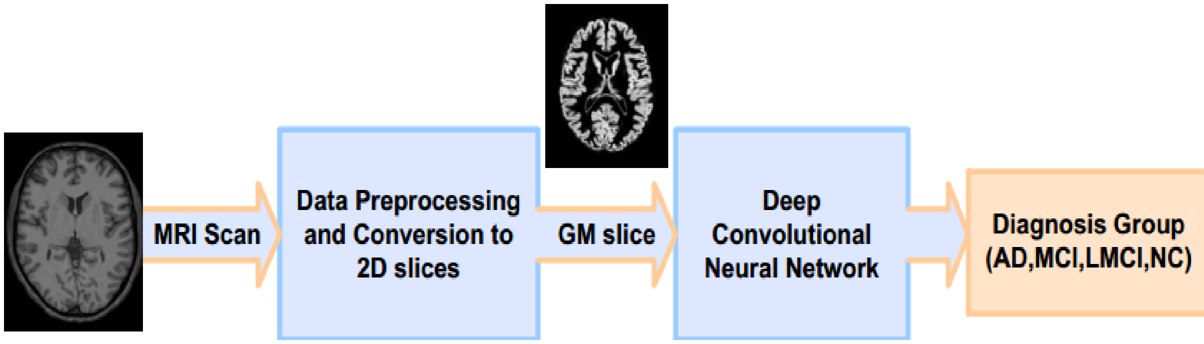

Figure 4.1 – Block Diagram of proposed 4-way classification framework

obtained from one MRI scan. Each slice image has 121 × 145 dimensions which is resized to 256 × 256 before passing to network.

4.2 Network Learning

As mentioned in the previous chapter, network training is based on deep convolutional neural networks. Three architectures are the focus of this work; GoogLeNet, ResNet-152 and the proposed model. Next section will describe the proposed model. Detail experimental set-up with dataset information is given in section 4.4 and 4.5. Figure 4.2 shows the sample images for AD class. Grey matter images for each class are very similar and hence, high level distinctive features are needed for detection. It is anticipated that deep CNN will extract desired latent features from these MRI images.

Figure 4.2 – Grey matter (GM) slices for AD class after pre-processing

4.3 Proposed Model

Aim of the proposed model is to provide optimized model for the AD diagnosis problem. The proposed model is inspired by the power of residual learning. It investigates the effect of residual learning in shallow networks. Residual learning was introduced originally to increase the depth of the network [43]. The proposed architecture is developed on the idea to find shallow solution comparable to AlexNet [36] but simultaneously taking advantage of residual learning. References [45, 46, 47, 48, 49] are based on residual learning and discussed its strength in variety of ways. According to Veit et al. [47], a residual network is not just a single network. It contains multiple paths for gradient flow. It works in the form of ensembles of shallow networks. Thus, the proposed model is motivated to incorporate residual learning with smaller depth.

Figure 4.3 shows the proposed model. It has total 25 independent modules and 8 weight layers with about ~4 million parameters. Batch normalization is performed after each spatial convolution. As compared to AlexNet, only one fully connected layer is used with 512 neurons instead of 4096. Three residual blocks are used with identity connections. Dimensionality is reduced at each shortcut but no extra parameter is introduced. Before the final FC layer, average pooling layer is used instead of max pooling and no dropout layer is used. Number of features maps increases as 64, 128, 256 and 512. Down sampling is performed in first convolution layer of residual block with stride of 2. Input size is reduced from 256×256 to 14×14. Number of weight layers and size of input is comparable to AlexNet.

4.4 ADNI Dataset

The MRI dataset used in the experiment is developed and provided by Alzheimer's disease Neuroimaging Initiative (ADNI) [7, 50]. ADNI study started in 2004 and aims at early diagnosis of AD, improving clinical trial experience with better use of biomarkers. Another objective is to share the neuroimaging data with the medical and engineering researchers to improve health-care sector.

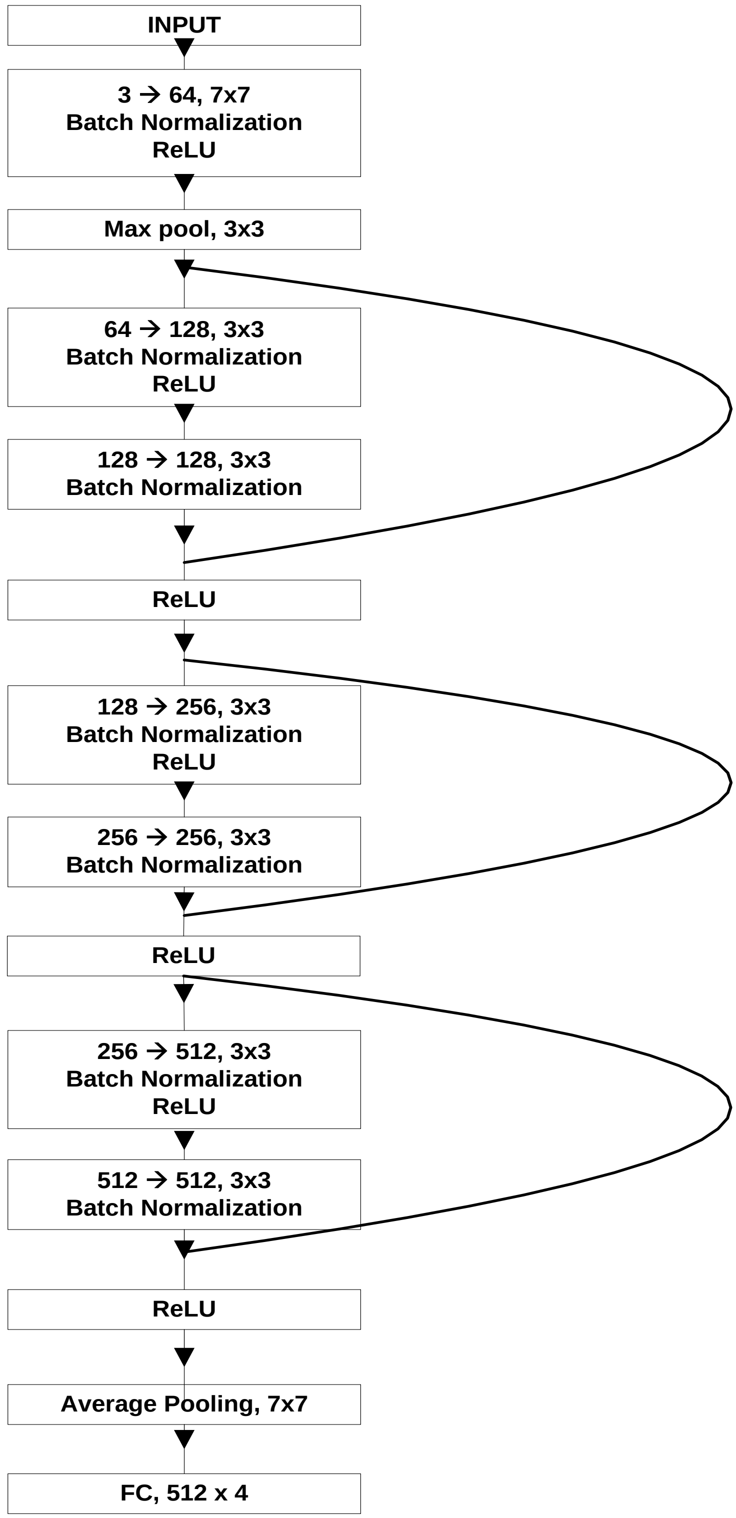

Figure 4.3 – Proposed shallow model with residual learning. Three residual blocks and total 8 weight layers are used, ~ 4M parameters

The selected set of MRI scans is acquired from 3T scanners and belong to ADNI1 phase. To provide good quality imaging data, all MRI scans are T1 weighted and have undergone gradwarping, intensity correction and gradient scaling. Complete 3T data has been used for efficient comparison studies. A subject is scanned at different point of times in different visits i.e., baseline, after one, two and three years. Each such scan is considered as a separate subject in this work. The dataset consists of 33 AD, 22 LMCI, 49 MCI patients and 45 healthy controls which makes a total 355 MRI volumes. The demographic information of the dataset used in proposed work is presented in table 4.1.

TABLE 4.1
DEMOGRAPIC INFORMATION OF DATASET

Diagnosis Group	No. of Subjects	Sex		Age	NPIQ Mean	FAQ Mean
		M	F			
AD	33	11	22	74 ± 8.3	3.33	10.8
LMCI	22	15	7	73 ± 7.4	1.85	3.4
MCI	49	30	19	75 ± 8.2	1.87	4.97
NC	45	18	27	75 ± 3.7	0.57	0.125

Three datasets are prepared for experimentation. Binary set consists of all diseased classes combined as AD and other class is normal. Second set consists of unbalanced multiclass data. Third dataset consists of balanced multiclass data. Balanced data is prepared for classification by using data augmentation for the class which has smaller number of images to get balanced with other classes. Augmentation is done by simply flipping the image along horizontal axis. This operation is valid because of the left and right symmetry of brain regions. The distribution of data for each set is given in table 4.2.

TABLE 4.2
DISTRIBUTION OF DATASET

Dataset	Classes				Total
Binary	17765 AD		9018 NC		26784
Unbalanced Multiclass	4753 AD	5823 MCI	6574 MCI	7518 NC	24668
Balanced Multiclass	9506 each				38024

4.5 Experimental Setup

4.5.1 TORCH7 and GPU

Several dedicated deep learning frameworks are available for implementation including Caffe, Theano, Tensorflow, MXNet, Kearas and Torch. Torch7 [51, 52] framework is used in the implementation of proposed pipeline in this thesis. It is popular among deep learning research groups like Google and Facebook research. It has CUDA bindings for GPU acceleration in the form of the package, cuDNN. It has interfaces for C, C++ and Lua programming languages. The proposed framework is implemented on single NVIDIA GeForce GTX TITAN X GPU. It is one of the most powerful and high speed GPUs available with 12 GB RAM and underlying Maxwell architecture. It has 3072 cores and compute capability value of 5.2.

4.5.2 Optimization Algorithm

In machine learning, the term optimization is quite frequent and associated with the learning process. When the inputs are passed through the network and loss is calculated. This loss is back propagated through the network along with gradients. Then optimization algorithm looks into the information and updates the network parameters in order to minimize the loss function. Optimization gives back the best set of parameters / weights that give minimum loss. Optim package for Torch7 has numerous optimization algorithms available. This work uses the most common one i.e. stochastic gradient descent (SGD) optimization for the purpose. After each backward pass, SGD calculates the gradients of parameters with respect to each layer's input and update the parameters accordingly. Pure stochastic gradient descent means that weights are updated after each single input sample. However, in case of a GPU and huge datasets available, the term SGD is used for mini-batch gradient descent interchangeably. Inputs are passed through the network in the form of small batches and optimization is performed after one complete batch. Batch size is selected according to memory needs and available hardware resources. The most common parameter of SGD algorithm is learning rate (LR). Learning rate is the value of step taken to update the parameters.

4.5.3 Training and Testing

Networks are trained on data for 55 epochs with each epoch consisting of 1000 batches. Batch size used is 32 for all models except ResNet-152 model which used batch size of 16 due to memory constraints. Training is performed on complete global image and no runtime crops or patches are used. All experiments are performed by splitting data into 25% as test and 75% as train data. 10% data from train set is used as validation set. Xavier initialization of weights is adopted to speed up the training of deeper networks as GoogLeNet and ResNet-152 take sufficient time to converge if training starts random weights. It assigns initial weights by drawing them from Gaussian distribution and assumes that the variance of the network is unity. Testing is performed in the similar batch by batch pattern. Loss is calculated using cross entropy criterion function. It is combination of log-softmax function with negative log likelihood (NLL) criterion. Log-softmax gives the log of normalized probability for each class based on output scores.

4.5.4 Training Strategies

Usually it is considered that deep learning only works for large datasets. It is a kind of myth and not true. For large or medium sized data, networks are typically trained with scratch random initial weights and slowly move towards the optimal set of parameters. However, luckily for small sets, we can use transfer learning approach. By this approach, we can learn or train the model on large dataset and then fine-tune those weights on the desired dataset. It is done by chopping off the top FC layer or layers, replacing with new layers and tweaking the weights with lower learning rate accordingly.

4.5.5 Phases of Experiments

Experiments are performed in three phases as follows:

1. Binary Classification
2. Multiclass Classification with scratch training
    a. With unbalanced dataset
    b. With balanced dataset
3. Multiclass Classification with transfer learning approach

Next chapter will discuss all experiments, obtained results for all models and comparisons in detail.

Chapter 5

Experiments and Results

5.1 Experimental Analysis

As mentioned in the previous chapter, preliminary experiments are performed before main investigation. It included comparing outcomes for balanced data versus unbalanced data, preprocessed data vs raw data and checking the effect of selection of MRI slices. For this purpose, ResNet-50 model is used and results are studied for binary classification only. The choice of model is made for its moderate number of layers, neither deep nor shallow. The consequences of analysis are later applied to main experiments.

5.1.1 All Slices versus Cerebrum Slices

This experiment investigates the performance for slice selection. Manual selection or marking of slices is time consuming as well as not precise. On the other hand, not all slices have informative patches of brain. The result of this investigation reveals that selection of slices affects training time. Batch processing time increases with all slices which indicates difficulty in learning representation. It clearly shows that non-informative patches create hindrance in learning.

TABLE 5.1
PERFORMANCE COMPARISON FOR SELECTION OF SLICES

	Training Accuracy (%)	Testing Accuracy (%)	Time (hours)	E/ES/BS	Data Stats
All slices	99.9	88.9	14	55/1000/32	Tr = 17813, Te = 4000 Total = 21813
Selected Slices	100	88	11		Tr = 11188, Te = 2654 Total = 13842

5.1.2 Balanced data versus Un-balanced data

The amount and distribution of data greatly affects the classification performance. The bias of data also disturbs class performances. Accuracy decreases and gets fluctuated as the balanced distribution of data changes. The result of the experiment clearly proves it and 7.5% gain in accuracy is achieved by using balanced data for training.

TABLE 5.2
PERFORMANCE COMPARISON FOR BALANCED AND UNBALANCED DATA

<table>
<tr><th></th><th>Training Accuracy (%)</th><th>Testing Accuracy (%)</th><th>Time (hours)</th><th>E/ES/BS</th><th>Data Stats</th></tr>
<tr><td>Unbalanced</td><td>100</td><td>88</td><td>11</td><td rowspan="2">55/1000/32</td><td>Tr = 11188, Te = 2654
Total = 21813</td></tr>
<tr><td>Balanced</td><td>100</td><td>95.5</td><td>10</td><td>Tr = 16595, Te = 2000
Total = 18595</td></tr>
</table>

5.2 Binary Classification

In the first phase of main experiments, binary classification is performed upon five different models. These models are trained from scratch. Scratch training starts with random initial weights. Total 26784 images are used with 17765 images of diseased class. Diseased class is a combined group of AD, MCI and LMCI. Results are also obtained for VGG network. Results are provided in the table 5.3 and figure 5.1. Proposed method achieved the best classification accuracy of 99.18%, GoogLeNet 99.08% and ResNet-152 showed slightly lower accuracy of 98.6%. Although, training accuracies are nearly 100% for all the models. Lower performance of ResNet-152 is due to over fitting and relatively smaller amount training data as compared to large number of layers and parameters.

TABLE 5.3
PERFORMANCE OF PROPOSED FRAMEWORK FOR BINARY CLASSIFICATION

Model	Training Accuracy (%)	Testing Accuracy (%)	Data Stats
GoogLeNet	100	99.08	total 26784 slices 9018 NC 17765 AD
VGG-11	99.5	98.3	
RseNet-50	99.98	98.9	
ResNet-152	99.95	98.6	
Proposed	99.98	99.18	

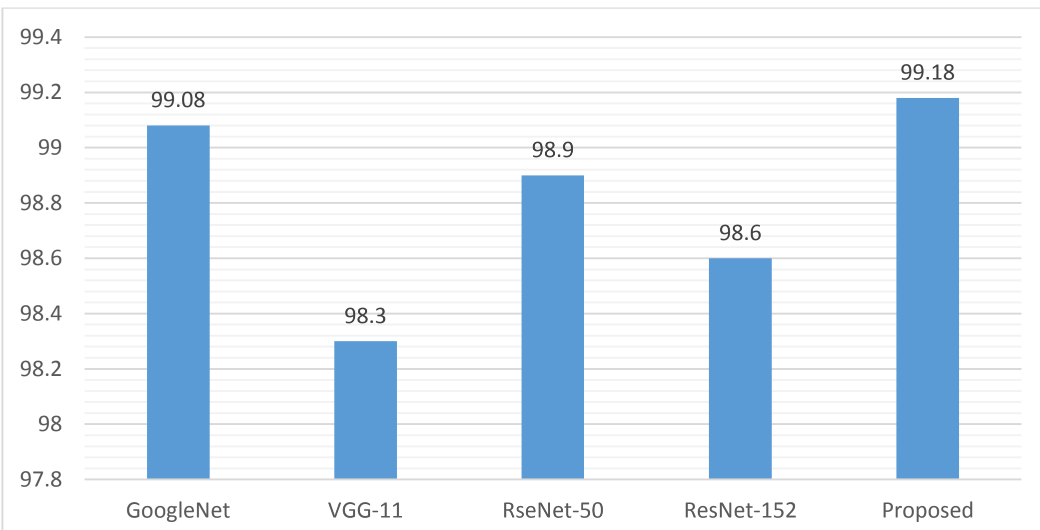

Figure 5.1- Performance of proposed method for Binary Classification

5.2.1 Comparison with Literature on Binary Classification

Detail comparison of proposed method with other method is shown in table 5.4. All of the selected literature uses ADNI dataset. Results clearly shows that the proposed method outperforms other techniques for binary classification. Hosseini et al. [31] achieved the best accuracy for binary classification i.e. 99.3% and our method obtained 99.18% using proposed model and

TABLE 5.4
DETAILED PERFORMANCE COMPARISON OF PROPOSED FRAMEWORK WITH OTHER TECHNIQUES FOR BINARY CLASSIFICATION (AD / NC)

Approach	Technique	Modalities	Accuracy (%)
Lui et al. [21]	SAE-Zeromask	MRI +PET	91.4
Shi et al. [23]	SDPN	MRI +PET	96.93
Lui et al. [22]	MPFR	MRI + PET + Clinical	90.11
Tong et al. [19]	NGF	MRI + PET +CSF	83
Suk et al. [28]	DW-S^2 MTL	MRI + PET + CSF	95.9
Gupta et al. [29]	SAE	MRI	94.7 with natural image bases, 93.8 with MRI bases
Payan et al. [30]	SAE-CNN	MRI	95.39
Zhang et al. [18]	Kernel combination with SVM classifier	MRI + PET + CSF	93.2
Korolev et al. [32]	Vox CNN	MRI	79
Sarraf et al. [33]	GoogLeNet	MRI	98.84
Hosseni et al. [31]	DSA- 3D CNN	MRI	95.7
Proposed	**GoogLeNet**	**MRI**	**99.08**
	VGG-11	**MRI**	**98.3**
	ResNet-50	**MRI**	**98.9**
	ResNet-152	**MRI**	**98.6**
	Proposed	**MRI**	**99.18**

99.08% using GoogleNet. But the proposed method is much convenient than the competing work as no pre-learnt bases or features are required in the proposed CNN method. Figure 5.3 also shows the accuracy comparison. However, main investigation of this thesis is multi classification.

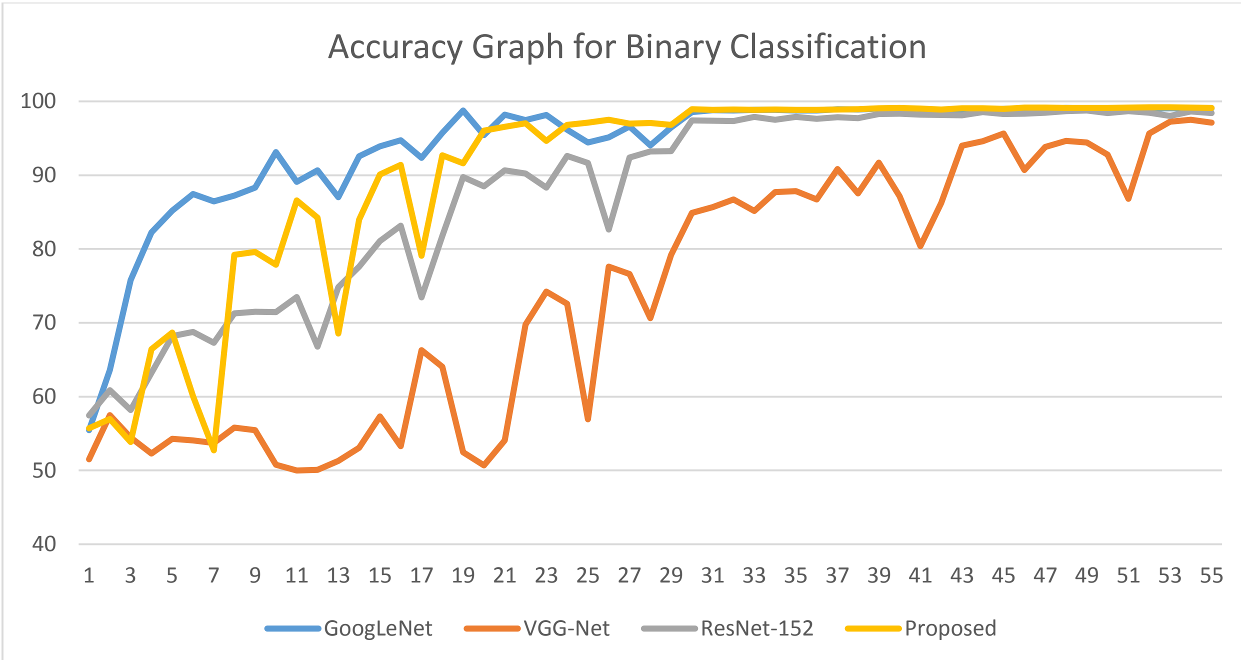

Figure 5.2 – Accuracy curves for testing Binary Classification (trained and tested for 55 epochs)

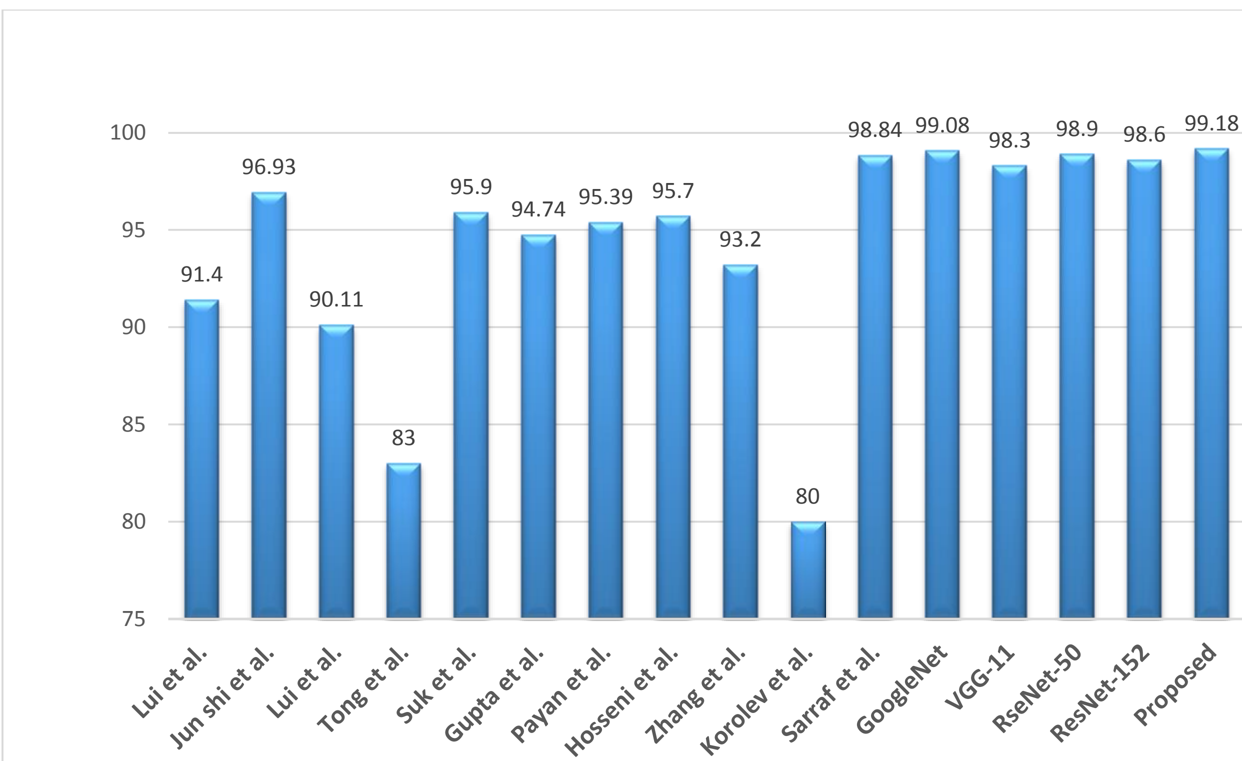

Figure 5.3- Accuracy Comparison for Binary Classification with other methods

5.3 Multi-class Classification

Four way classification with scratch training is performed in second phase of experiments. Results are obtained here for both unbalanced and balanced dataset and are presented in the following subsections.

5.3.1 Training from Scratch with Unbalanced Data

This experiment included total 24668 images with 4753 AD, 5823 LMCI, 6574 MCI and 7518 NC images. Maximum accuracy of 97.2% is achieved by the proposed model. All the three models performed well for unbalanced dataset but showed difference in training and testing accuracy. This shows that models are facing over fitting and thus unable to generalize well on testing data. However, the distribution of data affects the performance of the proposed framework.

TABLE 5.5
PERFORMANCE OF PROPOSED FRAMEWORK FOR MULTICLASS CLASSIFICATION WITH TRAINING FROM SCRATCH ON UNBALANCED DATASET

Model	Training Accuracy (%)	Testing Accuracy (%)	Data Stats
GoogLeNet	100	96	total 24668 images; 4753 AD, 5823 LMCI, 6574 MCI and 7518 normal
ResNet-152	99.95	95.7	
Proposed	99.98	97.2	

5.3.2 Training from Scratch with Balanced Data

Table 5.6 shown below presents the results of 4-way classification with balanced data. Each class has 9506 samples and total 38024 images are used in this experiment. The acquired results are higher as compared to results for unbalanced data. The element of difference in training and testing accuracies is also reduced.

TABLE 5.6
PERFORMANCE OF PROPOSED FRAMEWORK FOR MULTICLASS CLASSIFICATION WITH TRAINING FROM SCRATCH ON BALANCED DATASET

Model	Training Accuracy (%)	Testing Accuracy (%)	Data Stats
GoogLeNet	99.95	98.88	Total 38024 9506 each
ResNet-152	99.94	98.14	
Proposed	99.98	98.01	

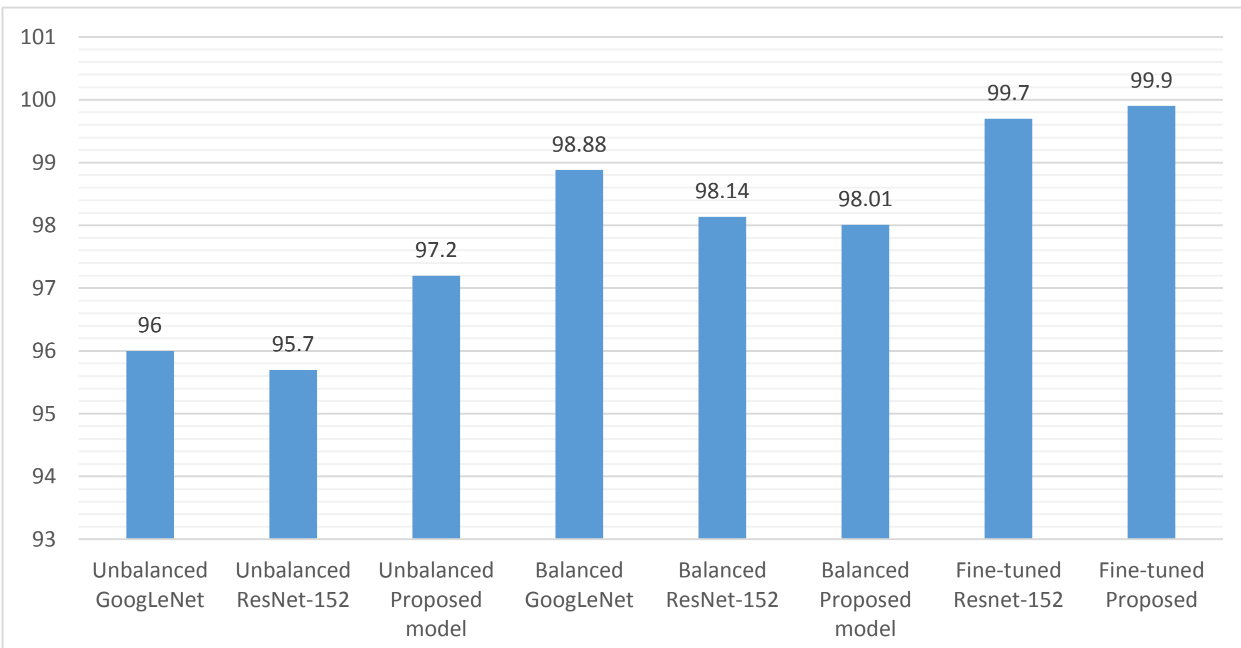

Figure 5.4 - Performance of proposed method for Multiclass Classification

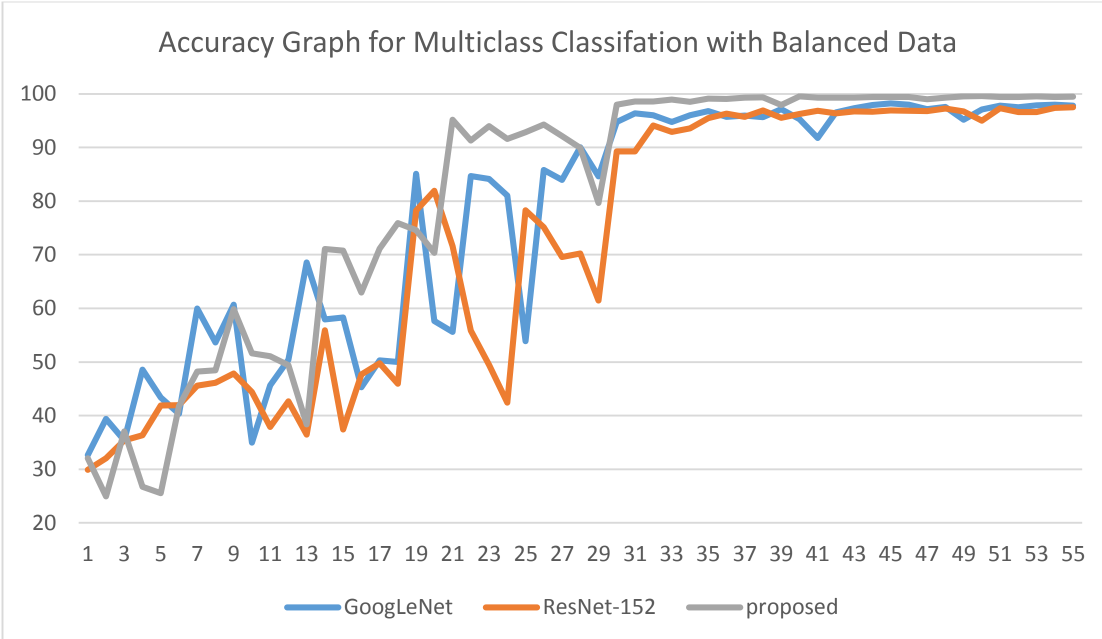

Figure 5.5 – Accuracy curves for testing Multiclass classification with Balanced data (trained and tested for 55 epochs)

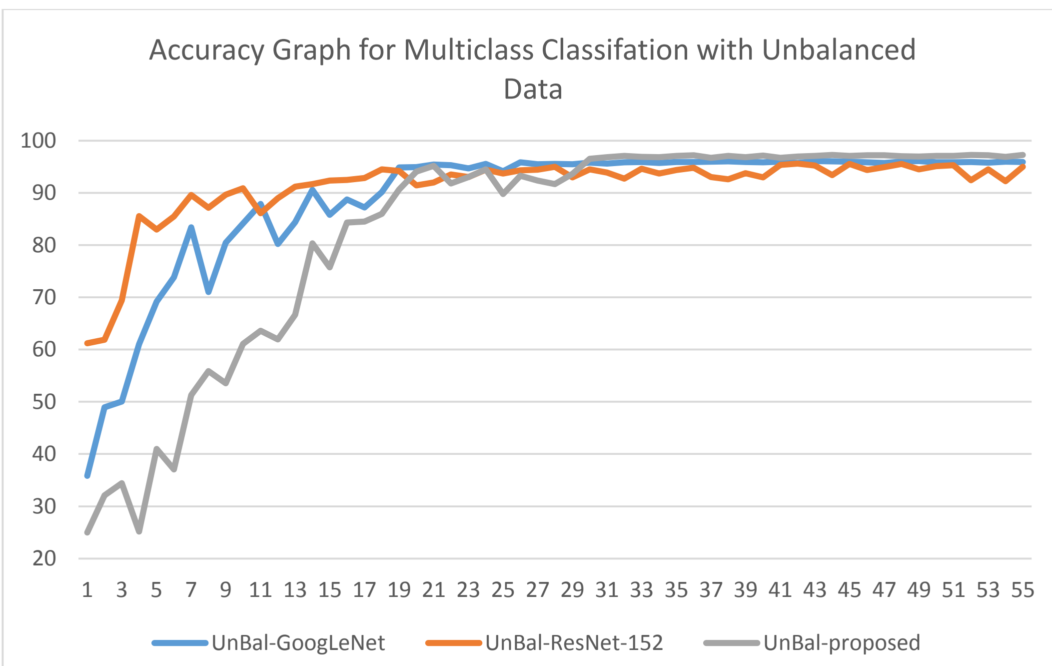

Figure 5.6 – Accuracy curves for testing Multiclass classification with Unbalanced data (trained and tested for 55 epochs)

5.4 Multiclass Classification using Transfer Learning Approach

Transfer learning is useful in case of small dataset. It suggests to train a model on one dataset and tune that model on the desired dataset for training and testing. In all of the above mentioned experiments, ResNet-152 showed slightly lower performance as compared to other models. This model is significantly deeper than other models and thus requires large amount of data as well as more learning information from the input data. However, the amount of dataset used in this work is quite moderate but still very small as compared to huge ImageNet dataset. Transfer learning approach is applied in the third phase of experiments. The models trained on binary dataset are used to fine-tune multiclass models for ResNet-152 and the proposed model. The obtained results are incredible and reach to 99.9% accuracy of diagnosis using the proposed model. Table 5.7 shows the results and figure 5.4 shows performance for all multiclass experiments.

TABLE 5.7
PERFORMANCE OF PROPOSED FRAMEWORK FOR MULTICLASS CLASSIFICATION WITH TRANSFER LEARNING APPROACH

Model	Training Accuracy (%)	Testing Accuracy (%)	Data Stats
ResNet-152	99.98	99.7	Total 38024 9506 each
Proposed	99.94	99.94	

Transfer learning also helps in training and learning process. Accuracy and loss curves for both scratch training and fine-tuning are shown in figure 5.6 below. Training from scratch passes through numerous ups and downs before reaching convergence. But training through fine-tuning process is relatively smooth and shows faster convergence. It is evident from the curves that transfer learning is useful our problem and effective for dealing medical imaging related problems.

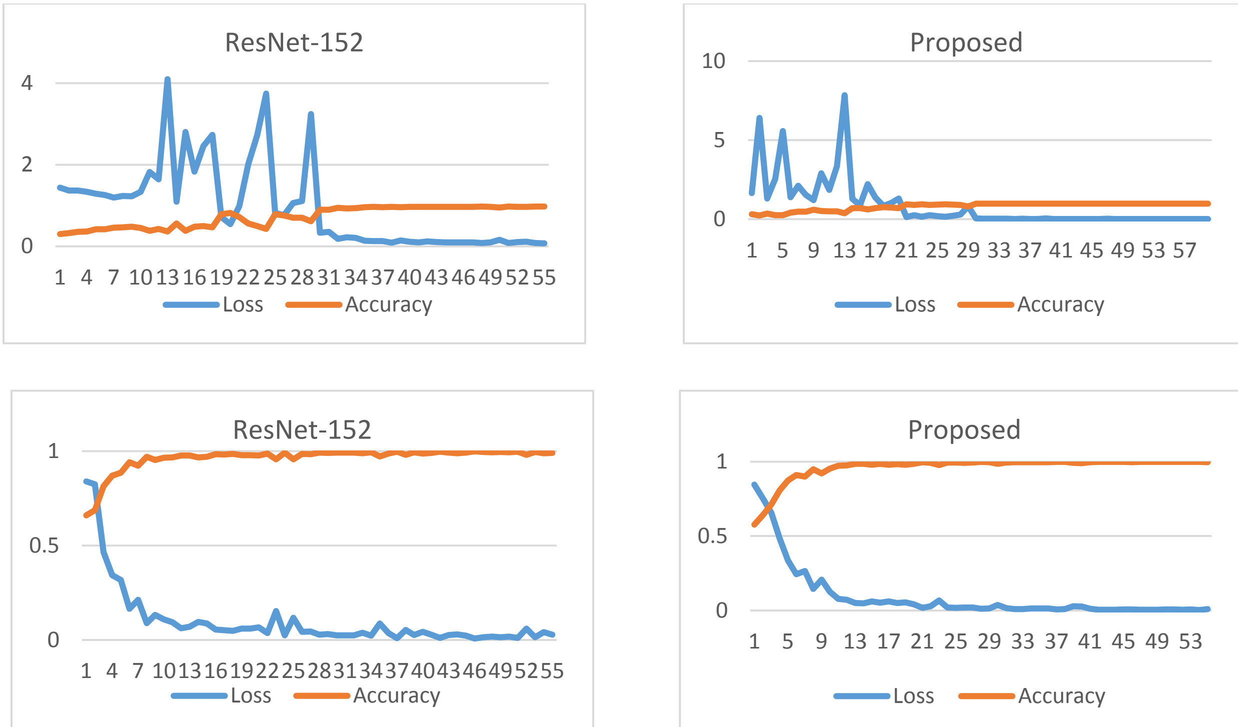

Figure 5.7 - Accuracy and Loss curves for Scratch training (above) and Transfer Learning approach (below)

5.5 Comparison with Literature in Multiclass Classification

Table 5.8 presents the detailed comparison of the proposed framework with literature. In most of the works, multiclass classification is performed as pairs of binary classification i.e. AD/NC, MCI/NC and AD/MCI. Those results are not considered here. The considered results mostly came from three way classifiers (AD/MCI/NC). References [21, 23] have performed four way classification with two types of MCI i.e. converting (cMCI) and non-converting (ncMCI), depending upon whether MCI condition would turn to AD or remain in safe state. Figure 5.7 shows the accuracy comparison graph. Results of the proposed CNN based framework clearly outperforms all other techniques with significant increase in accuracy. The proposed framework is convenient, depends on single modality data and features are learnt directly from raw data without any pre-trained bases.

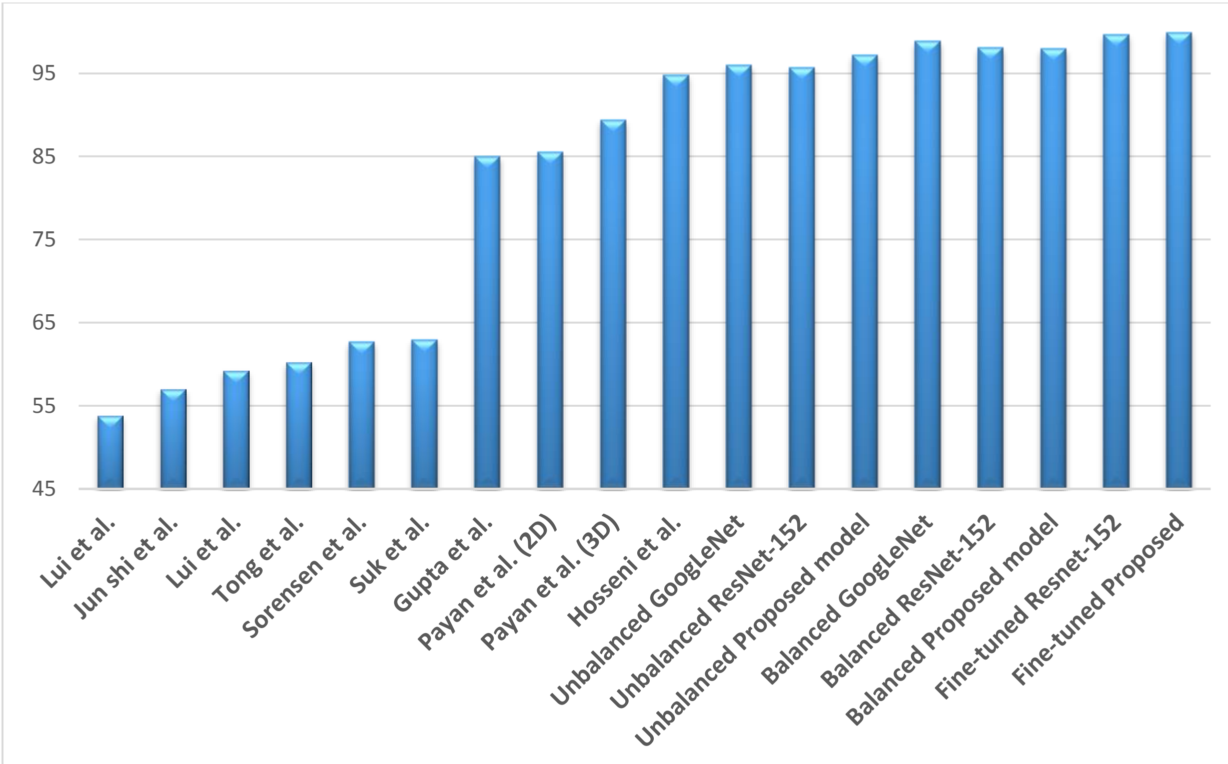

Figure 5.8 - Accuracy Comparison for Multiclass Classification with other literature

5.6 Class-wise Performance

In this section, class-wise performance is discussed. In most of the medical diagnosis work, class-wise performance is evaluated in terms of sensitivity (SEN), specificity (SPE) and positive predictive value (PPV) for each class. Sensitivity tells us the probability of the predicted test class to be the specific class for which evaluation is performed. Similarly, specificity gives probability of how well the negative class is rejected. PPV is the ratio of true predictions to the total predictions made for the specified class. The class-specific performance for each tested model is presented in the table 5.9 below. All of the values for specificity, sensitivity and PPV are greater than 90% which demonstrates that all of the classes are differentiated with high accuracies.

TABLE 5.8
DETAILED PERFORMANCE COMPARISON OF PROPOSED FRAMEWORK WITH OTHER TECHNIQUES FOR MULTICLASS CLASSIFICATION (AD / LMCI / MCI / NC)

Approach	Technique	Modalities	Classification	Accuracy (%)
Lui et al. [21]	SAE-Zeromask	MRI +PET	4 way (AD/cMCI/ncMCI/NC)	53.8
Shi et al. [23]	SDPN	MRI +PET	4 way (AD/cMCI/ncMCI/NC)	57
Lui et al. [22]	MPFR	MRI + PET + Clinical	3 way (AD/MCI/NC)	59.19
Tong et al. [19]	NGF	MRI + PET + CSF	3 way	60.2
Sorensen et al. [13]	Combined biomarkers	MRI	3 way	62.7
Suk et al. [28]	DW-S^2 MTL	MRI + PET + CSF	3way	62.93
Gupta et al. [29]	SAE	MRI	3 way	85 with natural image bases, 78.2 with MRI bases
Payan et al. [30]	SAE-CNN	MRI	3 way	89.4 with 3D convolution 85.53 with 2D convolution
Hosseni et al. [31]	DSA- 3D CNN	MRI	3 way	94.8
Proposed	**GoogLeNet (Unbalanced)**	MRI	4 way (AD/MCI/LMCI/NC)	**96**
	Proposed (Unbalanced)			**97.2**
	ResNet-152 (Unbalanced)			**95.7**
	GoogLeNet			**98.88**
	Proposed			**98.01**
	ResNet-152			**98.14**
	Proposed (fine-tuned)			**99.9**
	ResNet-152 (fine-tuned)			**99.7**

TABLE 5.9
DETAILED CLASS-SPECIFIC PERFORMANCE OF PROPOSED FRAMEWORK FOR DIFFERENT MODELS

Models	AD			LMCI			MCI			NC			ACC
	SPE	SEN	PPV	SPE	SEN	PPV	SPE	SEN	PPV	SPE	SEN	PPV	(%)
GoogLeNet (Unbalanced)	0.99	0.94	0.98	0.99	0.95	0.99	0.98	0.97	0.94	0.98	0.99	0.95	96
Proposed (Unbalanced)	0.99	0.94	0.99	0.99	0.96	0.99	0.98	0.97	0.95	0.97	0.99	0.93	97.2
ResNet-152 (Unbalanced)	0.99	0.91	0.98	0.99	0.94	0.97	0.97	0.97	0.93	0.97	0.99	0.93	95.7
GoogLeNet	0.99	0.97	0.97	0.99	0.99	0.99	0.99	0.97	0.98	0.98	0.96	0.97	98.88
Proposed	0.99	0.97	0.98	0.99	0.99	0.99	0.99	0.96	0.99	0.97	0.97	0.93	98.01
ResNet-152	0.99	0.97	0.98	0.99	0.99	0.99	0.99	0.97	0.99	0.98	0.97	0.95	98.14
Fine-tuned Proposed	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	99.9
Fine-tuned ResNet-152	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	99.7

The comparison of class-wise performance for AD, MCI and NC classes is presented in table 5.10. References [29, 31] showed good sensitivity for AD class as compared to other two classes. Lower sensitivity of normal class is not acceptable as it signifies that most of the normal cases are predicted as diseased class. Similarly, lower PPV for diseased class shows the same situation that only a small portion of the predicted samples actually belong to that class. Table 5.10 clearly shows that the proposed method is equally improving the performance for all of the classes and significantly higher than other methods.

TABLE 5.10
CLASS-SPECIFIC PERFORMANCE COMPARISON OF PROPOSED FRAMEWORK WITH OTHER METHODS

Models	AD			MCI			NC		
	SPE	SEN	PPV	SPE	SEN	PPV	SPE	SEN	PPV
Lui et al. [22]	-	-	0.61	-	-	0.61	-	-	0.49
Sorensen et al. [13]	-	0.4	-	-	0.57	-	-	0.79	-
Gupta et al. [29]	0.91	0.95	-	0.92	0.74	-	0.91	0.87	-
Hosseini et al. [31]	-	1	1	-	0.80	0.60	-	0.47	0.70
GoogLeNet	0.99	0.97	0.97	0.99	0.97	0.98	0.98	0.96	97.6
Proposed	0.99	0.97	0.98	0.99	0.96	0.99	0.97	0.97	93.94
ResNet-152	0.99	0.97	0.98	0.99	0.97	0.99	0.98	0.97	95.4
Fine-tuned Proposed	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Fine-tuned ResNet-152	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99

TABLE 5.11
CONFUSION MATRIX FOR MULTICLASS CLASSIFICATION (SCRATCH) WITH PROPOSED MODEL

	AD	LMCI	MCI	NC
AD	2274	0	0	26
LMCI	0	2300	0	0
MCI	0	1	2288	11
NC	5	0	3	2292

TABLE 5.12
CONFUSION MATRIX FOR MULTICLASS CLASSIFICATION (FINE-TUNING) WITH PROPOSED MODEL

	AD	LMCI	MCI	NC
AD	2297	0	3	0
LMCI	0	2300	0	0
MCI	1	0	2298	1
NC	0	0	0	2300

5.7 Comparison of Architectural Details of Models

This sections discusses the architectural details of all the models used in this work. The proposed model is an optimized solution for the presented problem in terms of performance and computational complexity. Table 5.13 shows the comparison for training time, number of parameters and performance. VGG nets is not considered as ResNet-152 is already 8 times less complex than VGG nets [43]. Although this work is based on deeper networks but AlexNet is considered to compare with the smaller number of layers of the proposed model.

TABLE 5.13
COMPARISON OF IMAGENET MODELS WITH PROPOSED MODEL

Model	No. of Weight Layers	Training Time	No. of Parameters	Minimum batch size / GPU	Accuracy (%)
AlexNet	8	7 hours	~56 M (56442244)	32	99
GoogLeNet	22	2 days	~17 M (17359144)	32	98.88
ResNet-152	152	2 days	~58 M (58152004)	16	98.14
Proposed	8	7 hours	~4 M (4833860)	64	99

Comparison clearly reveals that proposed model is offering an optimized solution to us in terms of time as well as computational complexity. The proposed design has 11 times lesser parameters than AlexNet and supports larger batch size for processing. Training is significantly faster than deeper models with better results. The discussed results demonstrates the power of residual learning and convolutional neural networks for dealing with neuroimaging detection and diagnosis problems.

Chapter 6

Conclusion

In this thesis, we presented a convolutional network based framework for classifying structural MRI images to diagnose Alzheimer's disease, its prodromal stages MCI and LMCI and normal controls. Experimental data was obtained from ADNI and total 355 volumes from 149 subjects were used. MRI scans were first preprocessed to get GM images which were then passed to CNN network. Networks were trained and tested using deep GoogLeNet, ResNet and proposed optimized model. The proposed model incorporated residual learning in shallow networks. The model is optimized in terms of performance as well as learnable parameters. The proposed model has 11 times fewer parameters as compared to famous AlexNet model.

Experiments were performed in three phases. Binary classification was performed in the first phase which classified AD+MCI+LMCI vs NC. The results obtained were comparable to other available literature but convenient in terms single modality data and easy training.

Second phase of experiments dealt with multiclass classification by training the network from scratch. Results were obtained for both balanced and unbalanced dataset. All the acquired results outperformed the other methods for both cases. However, accuracies were slightly low for unbalanced dataset which showed that learning process depends on the amount and distribution of data. Maximum achieved accuracy is 98.88% using GoogLeNet and 98.14% using ResNet-152 and 98.01% using proposed model. However, complexity and training time of proposed model was significantly lower than deeper models.

Transfer learning approach was applied in the third phase. The models trained on binary dataset were tuned on multiclass data. This technique proved to very effective and gave 99.9% accurate diagnosis using proposed models and 99.7% for ResNet-152 model.

Overall, results for all the experiments outperformed other methods of literature in regard of multiclass classification. Significant increase in classification accuracy is achieved by using deep convolutional neural networks. As compared to most of the previous works, pre-trained feature learning is not required anymore and still the network can accurately predict the classes. Class specific performance gain is also achieved improving performance for all classes in terms of sensitivity, specificity and positive predictive value. Hence, it demonstrates the potential of incorporating deep convolutional network models for learning distinctive features from neuroimaging data and has high-level of implications for medical as well as neuro image processing.

Future recommendations include clinical trial of the proposed method to investigate and improve the shortfalls. Also, investigating the effects of combining data from different imaging modalities or combining clinical data with imaging data in the cases where structural data are limited. It is also recommended to explore and incorporate 3D convolutions in networks to get rid of the process of creating and selecting 2D slices and directly pass 3D MRI volumes for processing. It also recommends to develop a generalized model which can analyze structural MRI images for various structural biomarkers related to different brain diseases in single setup.

References

[1] Alzheimer's Association, "what is Dementia," 25 July, 2017. http://www.alz.org/what-is-dementia.asp

[2] Alzheimer's Association, "Alzheimer's Disease," 25 July, 2017. http://www.alz.org/alzheimers_disease_what_is_alzheimers.asp

[3] Nvidia News, "Diagnose heart disease with AI," 25 July, 2017. https://news.developer.nvidia.com/diagnose-heart-disease-with-ai/

[4] Nvidia News, "Diagnosing brain tumor quicker and with higher accuracy," 25 July, 2017. https://news.developer.nvidia.com/diagnosing-brain-tumors-quicker-and-with-higher-accuracy/

[5] Alzheimer's Association, "2016 Alzheimer's disease facts and figures," Alzheimer's & Dementia, vol. 12(4), pp. 459-509, 2016.

[6] Herculano-Houzel, Suzana. "The human brain in numbers: a linearly scaled-up primate brain." Frontiers in human neuroscience 3 (2009).

[7] Wyman, Bradley T., et al. "Standardization of analysis sets for reporting results from ADNI MRI data," Alzheimer's & Dementia, vol. 9(3), pp: 332-337, 2013.

[8] Anoop, A., et al. "CSF biomarkers for Alzheimer's disease diagnosis." International Journal of Alzheimer's Disease 2010 (2010).

[9] Petersen, Ronald C. "Mild cognitive impairment." New England Journal of Medicine 364.23 (2011): 2227-2234.

[10] Goodfellow et al, "Deep Learning," MIT press, 2016. http://www.deeplearningbook.org

[11] Wikipedia contributors, "Deep learning," Wikipedia, The Free Encyclopedia, 11 August, 2017. https://en.wikipedia.org/w/index.php?title=Deep_learning&oldid=794535188

[12] de Vos, F., Schouten, T.M., Hafkemeijer, A., Dopper, E.G., van Swieten, J.C., de Rooij, M , van der Grond, J. and Rombouts, S.A., “Combining multiple anatomical MRI measures improves Alzheimer's disease classification,” Human brain mapping, 37(5), pp.1920-1929, 2016.

[13] L. Sørensen et al., "Differential diagnosis of mild cognitive impairment and Alzheimer's disease using structural MRI cortical thickness, hippocampal shape, hippocampal texture, and volumetry," NeuroImage: Clinical, vol. 13, pp. 470-482, 2017.

[14] Liu, Jin, et al. "Improving Alzheimeres Disease Classification by Combining Multiple Measures." IEEE/ACM Transactions on Computational Biology and Bioinformatics 2017.

[15] Previtali, Fabio, et al. "A novel method and software for automatically classifying Alzheimer’s disease patients by magnetic resonance imaging analysis." Computer Methods and Programs in Biomedicine, vol. 143, pp: 89-95, 2017.

[16] Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack Jr, C.R., Ashburner, J. and Frackowiak, R.S., “Automatic classification of MR scans in Alzheimer's disease,” Brain, 131(3), pp.681-689, 2008.

[17] Huang, Meiyan, et al. "Longitudinal measurement and hierarchical classification framework for the prediction of Alzheimer’s disease." Scientific reports, vol. 7, 2017.

[18] D. Zhang et al., "Multimodal classification of Alzheimer's disease and mild cognitive impairment," NeuroImage, vol. 55(3), pp. 856-867, 2011.

[19] T. Tong, K. Gray, Q. Gao, L. Chen, D. Rueckert, Alzheimer's Disease Neuroimaging Initiative, “Multi-modal classification of Alzheimer's disease using nonlinear graph fusion,” Pattern Recognition, vol. 63, pp.171-181, 2017.

[20] Liu, Siqi, et al. "Early diagnosis of Alzheimer's disease with deep learning." Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on. IEEE, 2014.

[21] S. Liu, S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, D. Feng, M.J. Fulham, “Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer’s disease,” IEEE Transactions on Biomedical Engineering, vol. 62(4), pp. 1132-1140, 2015.

[22] S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, D. Feng, “Multi-phase feature representation learning for neurodegenerative disease diagnosis,” Australasian Conference on Artificial Life and Computational Intelligence, pp. 350–359, 2015.

[23] J. Shi, X. Zheng, Y. L, Q. Zhang, S. Ying, "Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease," IEEE Journal of Biomedical and Health Informatics , vol. PP(99), 2017.

[24] Suk, Heung-Il, and Dinggang Shen. "Deep learning-based feature representation for AD/MCI classification." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Berlin, Heidelberg, 2013.

[25] Suk, Heung-Il, et al. "Latent feature representation with stacked auto-encoder for AD/MCI diagnosis." Brain Structure and Function, vol. 220(2), pp. 841-859, 2015.

[26] Suk, Heung-Il, et al. "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis." NeuroImage, vol.5, pp. 69-582, 2014.

[27] Liu, Manhua, Daoqiang Zhang, and Dinggang Shen. "Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis," Human brain mapping, vol. 35(4), pp: 1305-1319, 2014.

[28] Suk, Heung-Il, et al. "Deep sparse multi-task learning for feature selection in Alzheimer’s disease diagnosis." Brain Structure and Function, vol. 221(5), pp: 2569-2587, 2016.

[29] A. Gupta, M. Ayhan, A. Maida, “Natural image bases to represent neuroimaging data,” International Conference on Machine Learning, pp. 987–994, 2013.

[30] A. Payan, G. Montana, “Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks,” arXiv preprint, arXiv: 1502.02506, 2015.

[31] E. Hosseini-Asl, G. Gimel'farb, A. El-Baz, “Alzheimer's disease diagnostics by a deeply supervised adaptable 3D convolutional network,” arXiv preprint, arXiv: 1607.00556, 2016.

[32] Korolev, Sergey, et al. "Residual and Plain Convolutional Neural Networks for 3D Brain MRI Classification," arXiv:1701.06643, 2017.

[33] S. Sarraf, G. Tofighi, “DeepAD: Alzheimer′s disease classification via deep convolutional neural networks using MRI and fMRI,” bioRxiv preprint, bioRxiv: 070441, 2016

[34] T. Altaf, S. M. Anwar, “Multi-class Alzheimer Disease Classification using Hybrid Features,” IEEE Future Technologies Conference, in press, 2017.

[35] LeCun, Yann, et al. "Backpropagation applied to handwritten zip code recognition," Neural computation, vol. 1(4), pp: 541-551, 1989.

[36] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[37] Y. LeCunn, Y. Bengio, G. Hinton, "Deep learning," Nature, vol. 521(7553), pp. 436-444, 2015.

[38] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." International Conference on Machine Learning. 2015.

[39] O. Russakovsky et al., "Imagenet large scale visual recognition challenge," International Journal of Computer Vision, vol. 115(3), pp. 211-252, 2015.

[40] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." European conference on computer vision. Springer, Cham, 2014.

[41] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556, 2014.

[42] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going deeper with convolutions," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-9, 2015.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv:1512.03385, 2015.

[44] SPM8, Wellcome Trust Centre for Neuroimaging, Institute of Neurology, UCL, London UK. http://www.fil.ion.ucl.ac.uk/spm

[45] Huang, Gao, et al. "Deep networks with stochastic depth." European Conference on Computer Vision. Springer International Publishing, 2016.

[46] He, Kaiming, et al. "Identity mappings in deep residual networks." European Conference on Computer Vision. Springer International Publishing, 2016.

[47] Veit, Andreas, Michael J. Wilber, and Serge Belongie. "Residual networks behave like ensembles of relatively shallow networks," Advances in Neural Information Processing Systems, pp: 550-558, 2016.

[48] Liao, Qianli, and Tomaso Poggio. "Bridging the gaps between residual learning, recurrent neural networks and visual cortex," arXiv: 1604.03640, 2016.

[49] Chen, Hao, et al. "Voxresnet: Deep voxelwise residual networks for volumetric brain segmentation," arXiv:1608.05895, 2016.

[50] Jack, Clifford R., et al. "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," Journal of magnetic resonance imaging, vol. 27(4), pp: 685-691, 2008.

[51] Ronan Collobert, Samy Bengio, and Johnny Marithoz. Torch: A Modular Machine Learning Software Library. 2002.

[52] R. Collobert, K. Kavukcuoglu, and C. Farabet, “Torch7: A MATLAB-like environment for machine learning,” BigLearn, NIPS Workshop, 2011.

[53] I. O. Korolev, "Alzheimer’s disease: a clinical and basic science review," Medical Student Research Journal, vol. 4, pp. 24-33, 2014.

[54] G. M. McKhann et al., “The diagnosis of dementia due to Alzheimer’s disease: recommendations from the National Institute on Aging-Alzheimer’s Association workgroups on diagnostic guidelines for Alzheimer’s disease,” Alzheimers and Dementia, vol. 7(3), pp. 263-269, 2011.

[55] S. Vieira, W. H. L. Pinaya, A. Mechelli, "Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications," Neuroscience & Biobehavioral Reviews, 2017.

[56] E. Arvesen, "Automatic Classification of Alzheimer’s Disease from Structural MRI," 2015.

[57] Fayao Liu, and Chunhua Shen, "Learning Deep Convolutional Features for MRI Based Alzheimer's Disease Classification," arXiv: 1404.3366, 2014.

[58] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks, vol. 61, pp.85-117, 2015.

[60] WildML, "Understanding convolutional neural networks for NLP," 11 August, 2017. http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/